On the Bayesian Treed Multivariate Gaussian Process with Linear Model of Coregionalization

Bledar Konomi, Georgios Karagiannis and Guang Lin

Pacific Northwest National Laboratory, Washington, United States

Abstract

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.

Keywords: multivariate Gaussian process, linear model of coregionalization, Bayesian treed Gaussian process, Markov chain Monte Carlo

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 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, 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 et al. (2014) extended these models to the multivariate Gaussian process based on the separable cross-covariance function suggested by Mardia and Goodall (1993) and O'Hagan et al. (1999). Conditional on a treed partition, the prediction of the BTMGP model is done independently within each subregion. We follow the same setting to generalize the BTMGP model proposed in Konomi et al. (2014) with LMC cross-covariance function. In the discussion below we present the likelihood and priors of the BTMGP.

2.2. Likelihood based on the LMC cross-covariance

We consider the multivariate GP regression model:

$$\boldsymbol{\eta}(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{w}(\boldsymbol{x}) + \boldsymbol{\epsilon}(\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{h}(\boldsymbol{x})$ is the $m \times 1$ vector of the basis functions at $\boldsymbol{x}$, $\boldsymbol{\beta}$ is the linear regression coefficient of dimension $m \times q$ and $\boldsymbol{\epsilon}(\boldsymbol{x})$ is the measurement error process (nugget error).

A crucial part of the model is the zero mean multivariate GP $\boldsymbol{w}(\boldsymbol{x}) = (w_1(\boldsymbol{x}), \ldots, w_q(\boldsymbol{x}))^T$, which captures dependences both within measurements at a given site and across the sites. The

cross-covariance matrix function of $\boldsymbol{w}(\boldsymbol{x})$ is defined as $\boldsymbol{c}_w(\boldsymbol{x},\boldsymbol{x}') = [\text{cov}(\boldsymbol{w}_r(\boldsymbol{x}),\boldsymbol{w}_{r'}(\boldsymbol{x}'))]_{\boldsymbol{r},\boldsymbol{r}'=1}^{q}$. For any integer n and any collection of input sites $\boldsymbol{X}$, we denote the multivariate realization $\boldsymbol{w} = (\boldsymbol{w}^T(\boldsymbol{x}_1),\ldots,\boldsymbol{w}^T(\boldsymbol{x}_n))$ which follows an $nq \times 1$ multivariate normal distribution $\boldsymbol{w} \sim MVN(0,\boldsymbol{C_w})$, where $\boldsymbol{C_w}$ is an $nq \times nq$ matrix.

In general, the LMC model can be written as $\boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{v}(\boldsymbol{x})$, where $\boldsymbol{A}$ is a $q \times r$, with $r \leq q$, non-singular transformation matrix that explains the association among the q output variables, and $\boldsymbol{v}(\boldsymbol{x})$ is a vector of r independent, zero mean, unit variance GPs with correlation functions $\rho_1(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\lambda}_1),\ldots,\rho_r(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\lambda}_r)$, and hyperparameters $\boldsymbol{\lambda}_i$. The most basic coregionalization model is the intrinsic specification from Matheron (1982) where $\boldsymbol{A}$ is a $q \times q$ full rank matrix and $\boldsymbol{v}(\boldsymbol{x})$ are i.i.d. spatial processes. This model is equivalent to the separable cross-covariance model proposed by Mardia and Goodall (1993).

The correlation function of $\boldsymbol{v}_j$ is of particular importance as it defines the smoothness of the random field. Different choices can be made here. The Matérn and power exponential correlation families are the two more general and popular choices. In the Matérn correlation function $\rho_j(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\lambda}_j) \propto \left(\sum_{k=1:k_x} ||\boldsymbol{x}_k - \boldsymbol{x}'_k||/\lambda_{j,k}\right)^{\nu/2} K_\nu\left(\sum_{k=1:k_x} ||\boldsymbol{x}_k - \boldsymbol{x}'_k||/\lambda_{j,k}\right)$ where K_ν is a modified Bessel function of order ν and $\lambda_{j,k}$ is the correlation strength. For the power exponential family $\rho_j(\boldsymbol{x},\boldsymbol{x}';\boldsymbol{\lambda}_j) = \exp\left(-\frac{1}{2}\sum_{k=1:k_x} \frac{||\boldsymbol{x}_k - \boldsymbol{x}'_k||^\nu}{\lambda_{j,k}^\nu}\right)$ where ν is a value in the interval $(0,2]$. In any case we define $\boldsymbol{R}_j \in \mathbb{R}^{n\times n}$ as the correlation matrix generated by $\boldsymbol{X}$ and $\rho_j(\,\cdot\,,\,\cdot\,;\lambda_{x,j})$.

A special case of the coregionalization model is the conditional representation (Wackernagel, 2003; Gelfand et al., 2004; Banerjee et al., 2004), which is also referred to by (Royle and Berliner, 1999) as a hierarchical modeling approach. The conditional LMC representation is equivalent with the assumption of lower triangular $\boldsymbol{A}$ in the joint linear model of coregionalization (i.e., the Cholesky decomposition of $\boldsymbol{\Sigma}$), which usually gives similar results to the more general $\boldsymbol{A}$ (Gelfand et al., 2004). The conditional model is written as:

$$\begin{aligned}
\eta_1(\boldsymbol{x})|\boldsymbol{\theta}_1 &= \tilde{\boldsymbol{h}}_1(\boldsymbol{x})^T\boldsymbol{\beta}_1 + \sqrt{\sigma_1^2}v_1(\boldsymbol{x}) + \sqrt{\boldsymbol{g}_1}u_1(\boldsymbol{x}),\\
&\vdots \qquad (2)\\
\eta_q(\boldsymbol{x})|\eta_1(\boldsymbol{x}),\ldots,\eta_{q-1}(\boldsymbol{x}),\boldsymbol{\theta}_q &= \tilde{\boldsymbol{h}}_q(\boldsymbol{x})^T\boldsymbol{\beta}_q + \alpha^{q|1}\eta_1(\boldsymbol{x}) + \cdots + \alpha^{q|q-1}\eta_{q-1}(\boldsymbol{x}) + \sqrt{\sigma_q^2}v_q(\boldsymbol{x}) + \sqrt{\boldsymbol{g}_q}u_q(\boldsymbol{x}),
\end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_q)$ are the parameters of the conditional representation of the corregionalization model, $\tilde{\boldsymbol{h}}_j(\boldsymbol{x})$ are the basis functions of the *linear regression model* of the input in $\eta_j(\boldsymbol{x})$, σ_j^2 is the model variance, and $\boldsymbol{g}_q$ is the nugget variance. To facilitate the representation we denote $\boldsymbol{h}_j(\boldsymbol{x})^T = [\tilde{\boldsymbol{h}}_j(\boldsymbol{x}),\eta_1(\boldsymbol{x}),\ldots,\eta_{j-1}(\boldsymbol{x})]$, for $j = 1,\ldots,q$. We also denote by $\boldsymbol{H}_j^T = (\boldsymbol{h}_j(\boldsymbol{x}_1),\ldots,\boldsymbol{h}_j(\boldsymbol{x}_N))$ the basis matrix, by $\boldsymbol{B}_j = (\boldsymbol{\beta}_j,\boldsymbol{\alpha}_j)$ the linear parameter associated with the $\boldsymbol{h}_j$ basis functions, and by m_j the total number of basis functions. The basis function in each of the conditional GPs introduces dependency between the multivariate output data. If we assume $\alpha^{q|q-1} = \cdots = \alpha^{2|1} = 0$ then the above model is equivalent to the independent multivariate model.

Let $\boldsymbol{Y}^j = (\eta_j(\boldsymbol{x}_1),\ldots,\eta_j(\boldsymbol{x}_N))^T$ denote the response vector of the j^{th} conditional representation in Eq. 2, for $j = (1,\ldots,q)$. Each likelihood of the conditional representation $f(\boldsymbol{Y}^j|\boldsymbol{\theta}_j)$ has an $n \times n$ covariance function and the likelihood of $\boldsymbol{Y}$ is:

$$f(\boldsymbol{Y};\boldsymbol{\theta}) = \prod_{j=1}^{q} f(\boldsymbol{Y}^j|\boldsymbol{Y}^1,\ldots,\boldsymbol{Y}^{j-1};\boldsymbol{\theta}_j).$$

In order to enable agreement between the conditional and marginal specifications, we require a common covariate ($\tilde{\boldsymbol{h}}_1(\boldsymbol{x})^T = \cdots = \tilde{\boldsymbol{h}}_q(\boldsymbol{x})^T$) and $u_1(\boldsymbol{x}) = \cdots = u_{q-1}(\boldsymbol{x}) = 0$. This means that all but one of the processes are purely spatial without error. In practice when we have numerical instabilities and for the better fit we can also introduce a nugget effect $\boldsymbol{g}$ similar to Gramacy and Lee (2012) for every conditional representation, which will give $\boldsymbol{\psi} = (\boldsymbol{\lambda}, \boldsymbol{g})$ parameters for the correlation function.

As it is pointed out in (Gelfand et al., 2004; Banerjee et al., 2004) sampling from the full conditional of $\boldsymbol{\Sigma}$ is a challenging task and the MCMC sampler can have a prohibitive slow convergence. Different strategies have been proposed to overcome the computational difficulties in the above work. However, most of them still have problems and lead to slow convergence of the MCMC. These problems have been observed in our examples, especially when we used reversible jump MCMC moves in the Bayesian tree. Therefore, in the following discussion we will concentrate only on the conditional representation, where we can further simplify the computations.

2.3. Priors for BTMGP with LMC cross-covariance

Let us consider a partition $\{\mathcal{X}_1,\ldots,\mathcal{X}_D\}$ of disjoint subregions of the input domain $\mathcal{X}$, such that $\mathcal{X} = \bigcup_{i=1}^{D} \mathcal{X}_i$, that corresponds to a tree structure $\mathcal{T}$ with D external nodes. We model each partition $\{\mathcal{X}_i\}$ with a multivariate GP and likelihood $f_i(\boldsymbol{Y}_i|\boldsymbol{\theta}_i)$ defined in section 2.2, where $\boldsymbol{\theta}_i = (\tilde{\boldsymbol{B}}_i, \boldsymbol{\sigma}_i^2, \boldsymbol{\lambda}_i, \boldsymbol{g}_i)$ denotes the parameters of the LMC conditional representation of the i^{th} external node. Given the partitions and the parameters, the likelihood has a step function form.

According to the Bayesian framework, we assign a prior distribution on the parameter $(\mathcal{T}, \boldsymbol{\theta})$, such as:

$$\pi(\mathcal{T},\boldsymbol{\theta}) = \pi(\mathcal{T}) \prod_{i=1:D} \pi(\boldsymbol{\theta}_i) = \pi(\mathcal{T}) \prod_{i=1:D} \prod_{j=1:q} \pi(\boldsymbol{B}_{i,j}) p(\boldsymbol{\sigma}_{i,j}^2) p(\boldsymbol{\lambda}_{i,j}) p(\boldsymbol{g}_{i,j}).$$

where j represents the j^{th} conditional representation. The marginal prior distribution of the binary tree $\pi(\mathcal{T})$ is defined according to the tree generating process suggested by Chipman et al. (1998). The multivariate GP parameters $(\boldsymbol{B}_i, \boldsymbol{\lambda}_i, \boldsymbol{g}_i, \boldsymbol{\sigma}_i^2)$ are apriori independent between different partitions and independent of each other within the partitions of the input domain. For each of the q conditional models in Eq. 2 , we assign independent priors, $\pi(\boldsymbol{\theta}_i) = \prod_{j=1}^{q} \pi(\theta_{i,j}) = \prod_{i=1}^{D} \prod_{j=1}^{q} \pi(\boldsymbol{\beta}_{i,j}, \boldsymbol{\alpha}_{i,j}, \sigma_{i,j}^2, \boldsymbol{\lambda}_{i,j}, \boldsymbol{g}_{i,j})$. The problem can be seen as q different Gaussian processes, each of which can be fitted separately.

Standard prior distributions can be considered for the parameters in the model. For simplicity, conjugate prior distributions can be used for the parameters associated with the mean $\boldsymbol{B}_{i,j} = (\boldsymbol{\beta}_{i,j}, \boldsymbol{\alpha}_{i,j})$ and variance $\boldsymbol{\sigma}_{i,j}^2$. We suggest non-informative priors for correlation hyperparameters. The joint prior distribution of $\boldsymbol{B}_{i,j}$ and $\boldsymbol{\sigma}_{i,j}^2$ is chosen as $p(\boldsymbol{\beta}_{i,j}, \boldsymbol{\alpha}_{i,j}, \sigma_{i,j}^2) \propto \sigma_{i,j}^{-2}$ which leads to a marginal posterior distribution of $\boldsymbol{\lambda}_i$ in closed form as described by Oakley and O'Hagan (2002). This is the equivalent of using inverse Wishart with diagonal parameter matrix for $\boldsymbol{\Sigma}$ (Banerjee et al., 2004) in the joint representation of the coregionalization model.

In order to ensure positive support for the values of $\boldsymbol{\lambda}_{i,j}$ and $\boldsymbol{g}_{i,j}$ we assign exponential prior distributions with parameters depending on the problem. The posterior distribution of

the above parameters can be derived using methods that are similar in computational cost to the separable model.

3. Bayesian computations

Because the resulting posterior distribution is intractable, we use MCMC methods to carry out inference. A blockwise MCMC sampler (Gelfand and Smith, 1990) is used to simulate each component of $\mathcal{T}|\boldsymbol{\theta}, \boldsymbol{Y}$ and $\boldsymbol{\theta}|\mathcal{T}, \boldsymbol{Y}$ from $p(\mathcal{T}, \boldsymbol{\theta}|\boldsymbol{Y})$.

3.1. MCMC simulation given the tree

For each of the external nodes of the Bayesian tree we independently use the same setting and follow the same MCMC sampling strategies. For brevity's sake we will give the posterior inference steps without specifying the external node.

The conjugate prior for $\boldsymbol{B}_j, \sigma_j^2$, which is $p(\boldsymbol{B}_j, \sigma_j^2) \propto \sigma_j^{-2}$, when combined with the likelihood of the j^{th} conditional LMC Gaussian process, leads to further computational simplifications.

The posterior distribution of $\boldsymbol{B}_j|\boldsymbol{Y}, \sigma_j^2, \boldsymbol{\lambda}_j, \boldsymbol{g}_j \sim \mathrm{N}(\boldsymbol{H}_j\hat{\boldsymbol{B}}_j, \sigma_j^2(\boldsymbol{H}_j^T\boldsymbol{R}_j^{-1}\boldsymbol{H}_j))$ and $\boldsymbol{\sigma}_j|\boldsymbol{Y}, \boldsymbol{\lambda}_j, \boldsymbol{g}_j \sim \mathrm{InvGam}\big[\frac{N-1}{2}, \frac{(N-m_j-2)\hat{\sigma}_j^2}{2}\big]$, where $\hat{\boldsymbol{B}}_j$ and $\hat{\sigma}_j^2$ denote the generalized least squares GLS estimators of $\boldsymbol{B}$ and σ_j^2 correspondingly.

After integrating out $\boldsymbol{B}_j$ and σ_j^2 from the posterior of $\boldsymbol{\lambda}_j, \boldsymbol{g}_j, \sigma_j^2, \boldsymbol{B}_j|\boldsymbol{Y}$, it can be shown that:

$$p(\boldsymbol{\lambda}_j, \boldsymbol{g}_j|\boldsymbol{Y}) \propto \pi(\boldsymbol{\lambda}_j)\pi(\boldsymbol{g}_j)|\boldsymbol{R}_j|^{-\frac{1}{2}}|\boldsymbol{H}_j^T\boldsymbol{R}_j^{-1}\boldsymbol{H}_j|^{-\frac{1}{2}}(\hat{\sigma}_j^2)^{\frac{N-m_j}{2}}, \tag{3}$$

where $\boldsymbol{R}_j$ and $\hat{\sigma}_j^2$ depend on $\boldsymbol{\lambda}_j$ and $\boldsymbol{g}_j$. The above posterior distribution is intractable and the inference is carried out with MCMC computational techniques. Integrating over $\boldsymbol{B}_j$ and $\boldsymbol{\sigma}_j^2$ can improve the mixing of the MCMC (Berger et al., 2001). This is crucial since the MCMC we applied in our problem is a combination of Metropolis-Hasting within Gibbs sampling (Mueller, 1993; Gelman et al., 2004) which requires a lot of iterations. For more details see Appendix (A). The computational cost for the conditional model is q times more expensive than for the separable model. The MCMC update for each θ_i of different partitions can be done in parallel which may lead to further reduction of the computational cost.

3.2. MCMC simulation updating the Bayesian tree

The structure of the binary tree of the GP model is updated through a random scan MCMC sweep that includes as updates the *Change*, *Swap*, *Rotate*, and *Grow* & *Prune* operations introduced by Chipman et al. (1998) and Gramacy and Lee (2008). The first three operations are Metropolis-Hastings updates operating on fixed dimensional spaces, while the last two are a reversible jump pair of moves (Green, 1995) that performs changes to the dimension of the parameter space. We perform a random scan MCMC of *Change*, *Grow*, *Prune*,*Swap*, *Rotate* operations with rates 0.4, 0.25, 0.25, 0.05 and 0.05 correspondingly. Although the same types of operations are used in the LMC and the separable models, they differ in the posterior of the model and the proposals. Below we give a brief review on the Bayesian tree operations.

The *Change* operation is a Metropolis-Hastings update whose proposal randomly picks an internal node and changes the associated splitting rule by drawing a new value uniformly such that no overlap occurs to the corresponding partitions. The tree structure and the parameters of the multivariate GP remain fixed, however the likelihood changes since the proposals cause changes to the limit boundaries of regions below the chosen node. Thus, the acceptance ratio of

this operation reduces to a likelihood ratio. In the *Swap* operation, a swap between the splitting rules of two randomly selected interior parent-child nodes is proposed. The *Rotation* operation is a Metropolis-Hastings update with proposals designed according to the left/right rotation operation in binary trees. This operation is discussed in detail by Gramacy and Lee (2008) where it is suggested as counterpart of the *Swap* operation because when *Swap* is performed on parent-child (internal) nodes that split on the same variables the *Swap* operation can present problematic behavior.

Given that the current state is at the binary tree $\mathcal{T}$, the *Grow* operation performs as follows. We randomly select an external node ζ_{i_0} that corresponds to a subregion $\mathcal{X}_{i_0}$ with data $\{\boldsymbol{X}_{i_0}, \boldsymbol{Y}_{i_0}\}$ and a multivariate GP model with parameters $\boldsymbol{\theta}_{i_0} = (\boldsymbol{B}_{i_0}, \boldsymbol{\lambda}_{i_0}, \boldsymbol{g}_{i_0}, \boldsymbol{\Sigma}_{i_0})$. We propose the node ζ_{i_0} to split into two new child nodes ζ_{i_1} and ζ_{i_2} according to the splitting rule P_{rule} used in the priors, and we denote the proposed tree as $\mathcal{T}'$. We consider that nodes ζ_{i_1} and η_{i_2} correspond to disjoint subregions $\mathcal{X}_{i_1}$ and $\mathcal{X}_{i_2}$, the union of which is $\mathcal{X}_{i_0}$, with data $\{\boldsymbol{X}_{j_1}, \boldsymbol{Y}_{j_1}\}$ and $\{\boldsymbol{X}_{i_2}, \boldsymbol{Y}_{i_2}\}$, respectively. Let $\boldsymbol{\theta}_{i_1} = (\boldsymbol{B}_{i_1}, \boldsymbol{\lambda}_{i_1}, \boldsymbol{g}_{i_1}, \boldsymbol{\Sigma}_{i_1})$ and $\boldsymbol{\theta}_{i_2} = (\boldsymbol{B}_{i_2}, \boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2}, \boldsymbol{\sigma}_{i_2})$ denote the parameter vectors of the multivariate GP associated to the new nodes ζ_{i_1} and ζ_{i_2}. A newly formed child, let us say ζ_{i_1}, is randomly chosen to receive values for $(\boldsymbol{\lambda}_{i_1}, \boldsymbol{g}_{i_1})$ from the parent one such that $(\boldsymbol{\lambda}_{i_1}, \boldsymbol{g}_{i_1}) = (\boldsymbol{\lambda}_{i_0}, \boldsymbol{g}_{i_0})$, while for the other, $(\boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2})$, we generate values from a proposal $Q(\boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2})$. $Q(\boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2})$ can be the prior distribution of $(\boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2})$. We generate proposals for $(\boldsymbol{B}_{i_1}, \boldsymbol{\Sigma}_{i_1})$ and $(\boldsymbol{B}_{i_2}, \boldsymbol{\sigma}_{i_2})$ from the posterior conditional distributions $p(\boldsymbol{B}_{i_1}, \boldsymbol{\sigma}_{i_1} | \boldsymbol{Y}_{i_1}, \boldsymbol{\lambda}_{i_1}, \boldsymbol{g}_{i_1})$ and $p(\boldsymbol{B}_{i_2}, \boldsymbol{\sigma}_{i_2} | \boldsymbol{Y}_{i_2}, \boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2})$. Let G and P' denote the set of the growable nodes of $\mathcal{T}$ and prunable nodes of $\mathcal{T}'$, respectively. The *Grow* operation is accepted with probability $\min\{1, A\}$ where

$$A = \frac{1 - a(1 + d_{\zeta_{i_0}})^{-b}}{a(1 + d_{\zeta_{j_0}})^{-b}(1 - a(2 + d_{\zeta_{j_0}})^{-b})^2} \frac{|G|}{|P'|} \frac{p(\boldsymbol{\lambda}_{i_1}, \boldsymbol{g}_{i_1} | \boldsymbol{Y}_{i_1}) p(\boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2} | \boldsymbol{Y}_{i_2})}{p(\boldsymbol{\lambda}_{i_0}, \boldsymbol{g}_{i_0} | \boldsymbol{Y}_{i_0}) q(\boldsymbol{\lambda}_{i_2}, \boldsymbol{g}_{i_2})}, \tag{4}$$

where G and P' denote the set of the growable nodes of $\mathcal{T}$ and prunable nodes of $\mathcal{T}'$, respectively. The Jacobian of the ratio is $|J| = 1$.

The *Prune* operation is the reverse analog of the *Grown*, from tree $\mathcal{T}'$ to $\mathcal{T}$, and is designed so that the detailed balanced condition is satisfied. Given the notation above, we randomly select a parent ζ_{i_0} of two external nodes ζ_{i_1}, ζ_{i_2} and turn it into a terminal node by collapsing the nodes below it. We randomly select a child node, let us say ζ_{i_1}, in order to pass the values of the parameters $(\boldsymbol{\lambda}_{i_0}, \boldsymbol{g}_{i_0}) = (\boldsymbol{\lambda}_{i_1}, \boldsymbol{g}_{i_1})$ and generate $(\boldsymbol{B}_{i_0}, \boldsymbol{\sigma}_{i_0})$ from the conditional posterior $p(\boldsymbol{B}_{i_0}, \boldsymbol{\sigma}_{i_0} | \boldsymbol{\lambda}_{i_0}, \boldsymbol{g}_{i_0}, \boldsymbol{Y}_{i_0})$. The operation is accepted with probability $\min\{1, 1/A\}$.

Note that here there is no need to propose linear coefficient and variances, $\boldsymbol{B}_i$ or $\boldsymbol{\sigma}_i$, for the *Grow/ Prune* operations. This can lead to more acceptable MCMC moves, as discussed by Godsill (2001), and creates simpler acceptance ratios at relatively low computational cost. However, $\boldsymbol{B}_i$ and $\boldsymbol{\sigma}_i$ can be updated after the operations have been performed, if necessary.

If we used the joint LMC for the cross-covariance instead of the conditional LMC we would not be able to integrate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{A}$. This makes the use of the joint LMC difficult in practice, where we have to propose all the parameters involved in the model. However, the difference between the conditional LMC and the separable models is noticeable. The reversible jump proposed parameters in the conditional LMC are qk_x while in the separable model only k_x as explained in Konomi et al. (2014).

3.3. Prediction and sampling

The Bayesian predictive density function $\boldsymbol{\eta}(\cdot)|\boldsymbol{Y}$ is calculated through Bayesian Model Averaging, which can recover a smooth representation of the prediction, $\boldsymbol{\eta}(x')$, around the limits of

the partitions $\{\mathcal{X}_i\}$. The proposed method allows the computation of the predictive distribution of any function of $\boldsymbol{\eta}$ for every input (or spatial) point.

For computational reasons one may be interested in sampling strategies of the input space given the proposed model. We give details on a possible extension of well known sequential sampling design in see Appendix B. However, we avoid a direct use of this extension since it is out of the scope of this paper.

4. Illustrations

In this section, we conduct a number of simulation studies to illustrate the performance of the BTMGP with conditional LMC cross-covariance (BTMGPC) which is introduced in this paper and compare it to BTMGP with independent cross-covariance (BTMGPI), multiple independent BTGP (Gramacy and Lee, 2008) and the BTMGP with separable cross-covariance (BTMGPS) (Konomi et al., 2014). For simplicity, we will use these abbreviations throughout this section. We design the simulation studies so that they involve multivariate output with discontinuities and localized features. The parameters in the prior distribution of the tree are $\alpha = 0.6$ and $\beta = 2$ as in Chipman et al. (1998). For simplicity we chose to work with constant mean in each external leaf (subregion) of the BTMGP.

4.1. 1-input and 3-output simulations

Our first example involves three different simple functions in one-dimensional input space. The functions are chosen such that different input subregions have functions with different dependencies. Specifically, we chose synthetic sinusoidal functions as:

$$\mathbf{f}_1(x) = \begin{cases} \sin\left(\frac{\pi x}{5}\right) + \cos\left(\frac{4\pi x}{5}\right), & x < 10 \\ x/10 - 1, & \text{otherwise} \end{cases}$$

$$\mathbf{f}_2(x) = \sin\left(\frac{\pi x}{5}\right) + \frac{1}{3}\cos(\pi x),$$

$$\mathbf{f}_3(x) = \begin{cases} \exp\left\{\sin\left(\frac{\pi x}{5}\right) + \cos\left(\frac{4\pi x}{5}\right)\right\}, & x < 10 \\ 1, & \text{otherwise} \end{cases}.$$

The first function is used by Gramacy and Lee (2008) and the two others are variational forms of this function. In the first half input interval ($x \in [0, 10]$) the three functions have similar spatial variation, while in the second half interval the second function has a completely different spatial variation. From the modeling point of view, the separable model is not appropriate for the second half input interval. The simulation study which follows will show this lack of agreement. We assume that the data have a small independent nugget error with variance 0.04 to ensure that the results of the joint LMC and the conditional LMC are similar.

We follow the Bayesian inference described in Section 3 to sample from the posterior of $\boldsymbol{\theta} = (\mathcal{T}, \boldsymbol{\sigma}, \boldsymbol{B}, \boldsymbol{\lambda}, \boldsymbol{g})$. The priors of $\boldsymbol{\lambda}$ and $\boldsymbol{g}$ are chosen independently as described in Section 2.3. A mixture of gammas priors for each spatial correlation parameter λ is a good choice when we deal with the Bayesian tree and the reversible jump moves (Gramacy and Lee, 2008). Specifically, we chose: $\pi(\lambda) = [G(\lambda|\alpha_G = 1, \beta_G = 20) + G(\lambda|\alpha_G = 10, \beta_G = 10)]/2$. An exponential distribution with mean 10^{-4} is chosen as a prior for g. The use of the RJ-MCMC in the Bayesian tree is relatively easy because only three spatial correlation parameters are

proposed from the prior and the *Grow* operation is done in a one-dimensional input. The algorithm is not very sensitive to the nugget proposed variances.

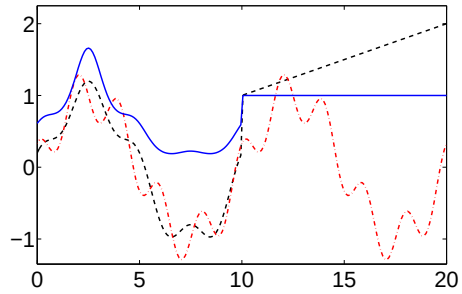
We train BTMGPC for $n = 25$, and 30 observations using Latin hypercube samples (LHS) and compare it to the BTMGPI, BTGP and BTMGPS. The LHS of size n is drawn as follows: the domain is partitioned into n disjoint and equally spaced intervals and from each of those intervals a value is drawn randomly, and they are randomly permuted if the domain has dimension bigger than 1. The real functions and the computed predicted mean functions with the 90% confidence interval for sample sizes 30 and different methods are shown in Fig. 1. We show the predicted mean functions and the 90% confidence intervals, using (b) BTMGPI,(c) BTGP, (d) BTMGPS, and (e) the proposed BTMGPC.

Most of the differences between BTMGPC and BTMGPI are observed in the first half input interval $[0, 10]$ where the dependences and the variation of the three functions are similar. This means that the conditional LMC cross-covariance is able to borrow information in this subregion, while in the other half it is not. The use of the conditional LMC in the BTMGP improved the prediction results in all three different functions. Instead, most of the differences between BTMGPC and BTMGPS are observed in the second half input interval $[10, 20]$, where the dependences and the variation of the three different functions are noticeably different. The independent error is captured as a part of systematic function deviation. In the fist half input interval $[0, 10]$ we observe more similarities. The BTMGPC gives similar predictions to BTMGPS for the first and third output function ($\mathbf{f}_1$ and $\mathbf{f}_3$) in the overall input region. However, better predictions for the second function in the second half are observed. Overall the BTMGPC gives better predictions.

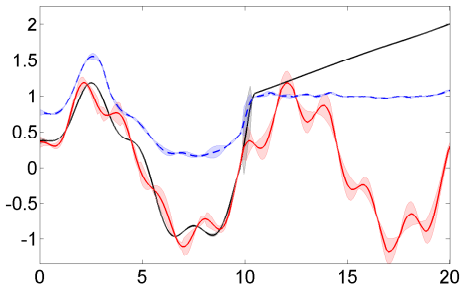
The differences of the BTMGPC in comparison with BTGP vary depending on the function. For the first and third functions ($\mathbf{f}_1$ and $\mathbf{f}_3$) the BTMGP seems to give better results. However, for the smooth function $\mathbf{f}_2$ we observe that BTGP gives better predictions. This is supported from the fact that, when using the BTGP we compute separately and independently for each variable a BTGP. When the function is smooth the results of the BTGP are similar to those of the GP model. However, when we use BTMGP we partition the input space regardless the output. The same noticeable differences are also true when we compare BTGP with BTMGPS.

The mean square errors (MSPEs) and presented in Table 1. For each model and function the MSPE is calculated as the mean square of the real value minus the prediction mean. As in Fig. 1, we compare the results of BTMGP with different cross-covariances and the multiple BTGP. As with the figure we conclude that the conditional LMC in the BTMGP improved the prediction results in two different sample sizes. BTMGPC seems to be closer to the BTMGPS when the functions have similar variability and closer to BTMGPI when functions have different variability. For the first ($\mathbf{f}_1$) and third function ($\mathbf{f}_3$) the MSPE of BTMGPC is similar to the MSPE of BTMGPS while for the second function ($\mathbf{f}_2$) the BTMGPC improves the prediction. Moreover, BTMGPC have smaller MSPE compare to BTGP for the two functions with discontinuity ($\mathbf{f}_1$ and $\mathbf{f}_3$). However, the MSPE of the continuous function ($\mathbf{f}_2$) is smaller when we use BTMGP. This a logical conclusion, since the BTMGP split functions without making any exception. The BTMGPC improved the overall prediction while maintaining computational complexity similar to that of BTMGPI.

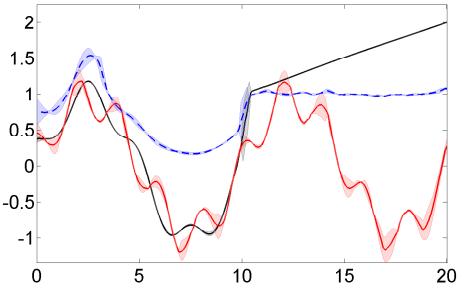
Summarizing, the proposed method is more flexible than BTMGP when the multivariate outputs are not dependent while outperform the BTGP when the outputs are dependent.



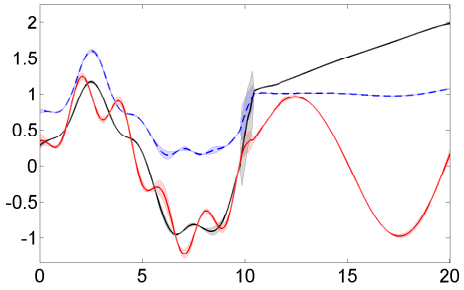
(a) Exact



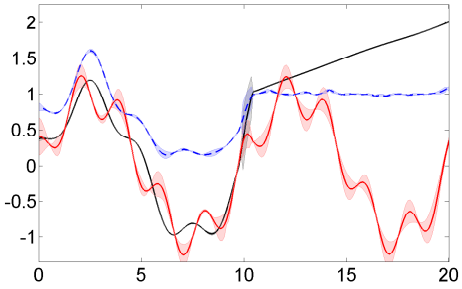
(b) BTMGPI $n = 30$



(c) BTGP $n = 30$



(d) BTMGPS $n = 30$



(e) BTMGPC $n = 30$

Figure 1: The real and the mean prediction using BMA (solid line) and the 90% confidence interval (CI) of the mean (shaded areas) of the three different functions four different cross-covaraiances in Bayesian treed models. For all the plots the black, red and blue lines and shaded area represent the first, second and third equation respectively. In specific : (a) exact functions, (b) BMA result and the 90% CI using the BTMGPI, (c) BMA result and the 90% CI using the BTGP, (d) BMA result and the 90% CI using the BTMGPS, and (e) BMA results for nd the 90% CI using the proposed BTMGPC.

4.2. 2-input and 2-output simulations

Our second example involves a two-dimensional input space $\boldsymbol{x} \in [-2, 6]^2$ and two-dimensional output functions problem which have been used in Konomi et al. (2014):

$$\mathbf{f}_1(\boldsymbol{x}) = x_1 \exp\left(-x_1^2 - x_2^2\right) + \epsilon_1,$$

and

$$\mathbf{f}_2(\boldsymbol{x}) = \sqrt{|x_1|} \exp\left(-x_1^2 - x_2^2\right) + \epsilon_2,$$

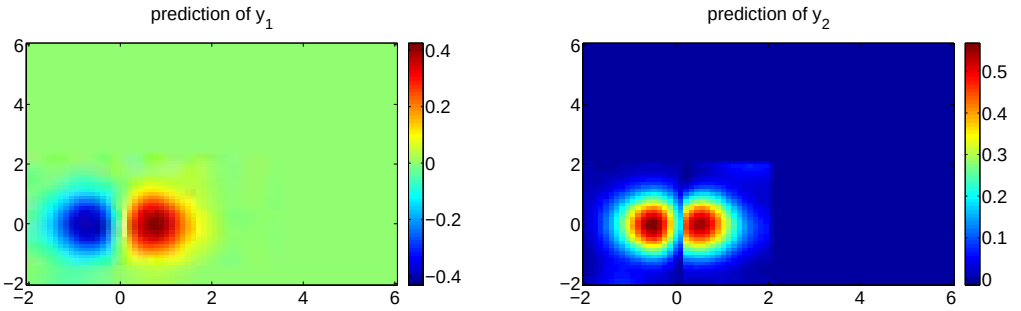
Table 1: MSPE of three different functions for different covariance model (conditional LMC, independent and separable model) in the BTMGP and different sample size.

		Sample size	
Model	Function	$n = 25$	$n = 30$
BTMGPI	$\mathbf{f}_1$	0.0374	0.0190
	$\mathbf{f}_2$	0.1998	0.1164
	$\mathbf{f}_3$	0.0503	0.0424
BTGP	$\mathbf{f}_1$	0.0408	0.0191
	$\mathbf{f}_2$	0.1028	0.0758
	$\mathbf{f}_3$	0.0491	0.0422
BTMGPS	$\mathbf{f}_1$	0.0323	0.0166
	$\mathbf{f}_2$	0.1432	0.1363
	$\mathbf{f}_3$	0.0445	0.0381
BTMGPC	$\mathbf{f}_1$	0.0332	0.0168
	$\mathbf{f}_2$	0.1128	0.0793
	$\mathbf{f}_3$	0.0455	0.0381

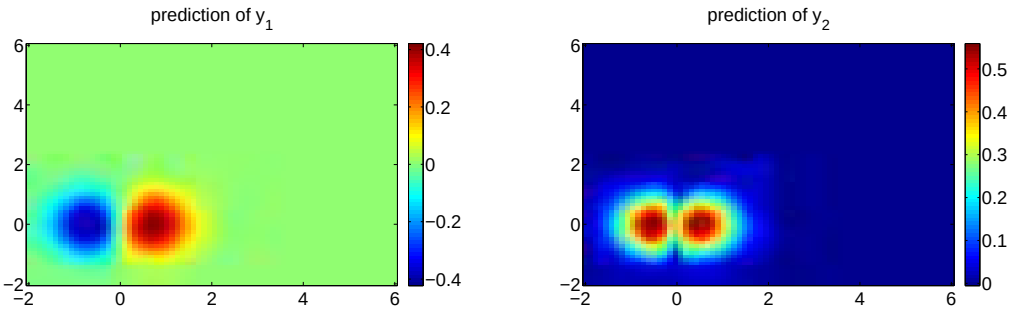
where $\epsilon_1 \equiv \epsilon_2 \sim N(0, \sigma = 0.001)$. Both functions have localized features in the box $[-2, 2] \times [-2, 2]$, while they are practically zero everywhere else. In subregion $[-2, 0] \times [-2, 2]$ the two output functions have negative correlation while for input subregion $[0, 2] \times [-2, 2]$ the two output functions have positive correlation. Moreover, both functions have similar variation over space. We utilize Latin hypercube samples (LHS) with two independent $Beta(\alpha_B, \beta_B; \min = -2, \max = 6)$ distributions and parameters $\alpha_B = 1.5$ and $\beta_B = 2.5$, which give higher probability density inside the box $[-2, 2] \times [-2, 2]$.

We train our model for $n = 65, 75$, and 85 observations by sampling the posterior of $\boldsymbol{\theta} = (\mathcal{T}, \boldsymbol{\sigma}, \boldsymbol{B}, \boldsymbol{\lambda}, \boldsymbol{g})$ following the MCMC procedure described in Sec. 3. Given an MCMC sample drawn, we made predictions by using BMA on a grid of 120×120. The MSPEs for the different Bayesian tree models and functions are calculated as it is described in the first simulation study and are reported in Table 2. Namely we use BTMGPS, BTGP and BTMGPC models as described above. Similar MSPEs are observed for both BTMGPS and BTMGPC models. The separable model is good enough to model the data. The values of $\boldsymbol{\lambda}_i$ in each MCMC iteration are very similar for all the conditional representations of the LMC. This is supported by the fact that the variations of the two output functions over the input space are similar. Instead the BTGP gives different results in the first function and second function. The first function ($\mathbf{f}_1$) seems to be slightly better predicted using BTGP while the second function ($\mathbf{f}_2$) is predicted better using BTMGP. Overall, modeling the dependence in the output helped us to predict the functions more accurately. For visualization purposes we also give the BMA response surface for both functions computed with BTMGPS and BTMGPC Fig. 2.

To better demonstrate the prediction difference and similarities of the two types of cross-covariances in BTMGP, we compute the prediction mean densities of ($\eta_1(\boldsymbol{x}_1)$ and $\eta_2(\boldsymbol{x}_1)$) for sample size $n = 85$ in three input values $\boldsymbol{x}_1 = \{(-1, 0.5), (-1.5, 1.5), (-0.25, 2.25)\}$. Histograms for the different equations and BTMGP models are constructed in Fig. 3 using the last $15,000$ MCMC samples. Each column corresponds to the same input point and each row corresponds to different BTMGP models and output functions. The red star in each histogram represents the true value. We can see that both methods appear to have similar accuracy, agreeing with the



(a) $\mathbf{f}_1$ Separable $n = 85$ (b) $\mathbf{f}_2$ Separable $n = 85$



(c) $\mathbf{f}_1$ LMC $n = 85$ (d) $\mathbf{f}_2$ LMC $n = 85$

Figure 2: The BMA mean prediction using the separable model and the LMC cross-covariance in the Bayesian tree.

general results of the MSPEs presented in Table 2. However, this accuracy does not necessarily translate to the same posterior distributions. In some areas the BTMGPS model gives the narrowest prediction distribution, in others the BTMGPC. From these prediction distributions and others which we construct, there is not a clear answer as to which model produce the narrowest and less-biased prediction distributions in this case.

The predicted probability distributions are non-normal and non-symmetric, e.g. Fig. 3 shows non-symmetric prediction distributions. The BTMGP with both cross-covariance functions is eligible to model a non-normal random field despite the assumption of normality for each distinct MCMC iteration. Given the MGP parameters, the BTMGP can model a mixture of normal predicted probability distributions. When the data are sparse, such that the Bayesian tree cannot contribute on the non-stationary and normal model assumptions, the nugget effect will maintain the good statistical properties for the emulator. Finally, we evaluated the joint probability density of the prediction by drawing $15,000$ samples from the posterior and then building a two-dimensional kernel density estimator. For $\boldsymbol{x}_1 = (-1.5, 1.5)$ the two dimensional probability density for the mean is shown in 4. Both models give dependent prediction distribution for the multivariate output. Notice also that the posterior's densities are very narrow.

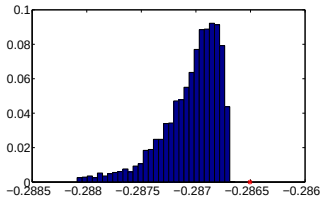
Summarizing, this example we show that for dependent outputs the proposed BTMGPC outperforms the BTGP and gives relatively better results than the BTMGPS.

4.3. Carbon capture regenerator unit

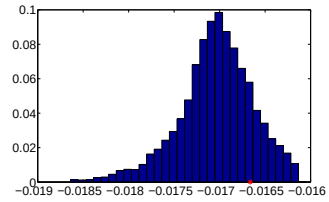
A typical carbon capture unit consists of two devices: the adsorber and the regenerator. A sorbent medium capable of reversibly reacting with carbon dioxide (CO_2) is looped through

Table 2: MSPE of the two functions for different sample size using three different models (BTMGPS, BTGP and BTMGPC).

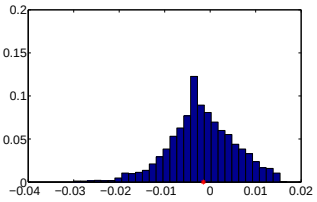
		Sample size		
Model	Function	$n = 65$	$n = 75$	$n = 85$
BTMGPS	$\mathbf{f}_1$	0.0033	0.0024	0.0014
	$\mathbf{f}_2$	0.0043	0.0036	0.0027
BTGP	$\mathbf{f}_1$	0.0028	0.0023	0.0014
	$\mathbf{f}_2$	0.0079	0.0059	0.0046
BTMGPC	$\mathbf{f}_1$	0.0029	0.0023	0.0014
	$\mathbf{f}_2$	0.0043	0.0036	0.0024



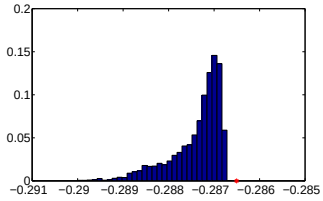
(a) $\mathbf{f}_1$ Separable $\boldsymbol{x} = (-1, 0.5)$



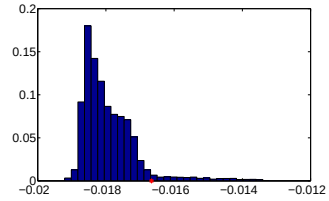
(b) $\mathbf{f}_1$ Separable $\boldsymbol{x} = (-1.5, 1.5)$



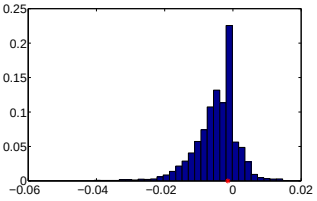
(c) $\mathbf{f}_1$ Separable $\boldsymbol{x} = (-0.25, 2.25)$



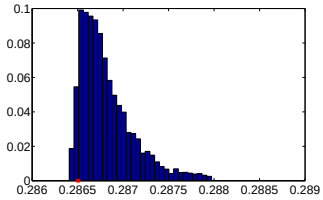
(d) $\mathbf{f}_1$ LMC $\boldsymbol{x} = (-1, 0.5)$



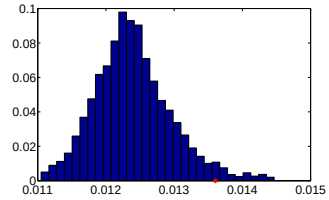
(e) $\mathbf{f}_1$ LMC $\boldsymbol{x} = (-1.5, 1.5)$



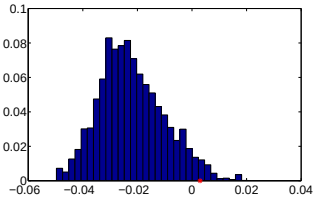
(f) $\mathbf{f}_1$ LMC $\boldsymbol{x} = (-0.25, 2.25)$



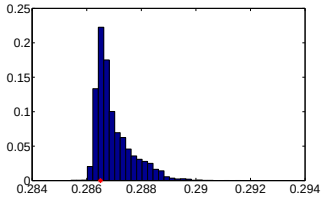
(g) $\mathbf{f}_2$ Separable $\boldsymbol{x} = (-1, 0.5)$



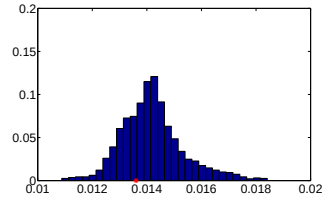
(h) $\mathbf{f}_2$ Separable $\boldsymbol{x} = (-1.5, 1.5)$



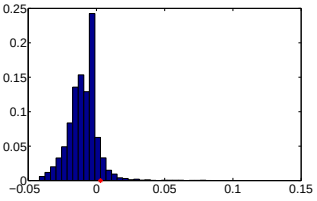
(i) $\mathbf{f}_2$ Separable $\boldsymbol{x} = (-0.25, 2.25)$



(j) $\mathbf{f}_2$ LMC $\boldsymbol{x} = (-1, 0.5)$

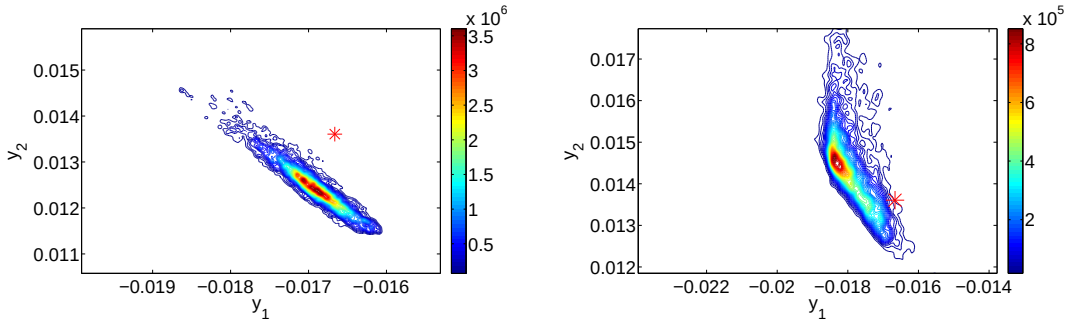


(k) $\mathbf{f}_2$ LMC $\boldsymbol{x} = (-1.5, 1.5)$



(l) $\mathbf{f}_2$ LMC $\boldsymbol{x} = (-0.25, 2.25)$

Figure 3: The prediction distribution using the separable and the LMC at three different inputs: the first column corresponds to $\boldsymbol{x} = (-1, 0.5)$, the second to $\boldsymbol{x} = (-1.5, 1.5)$, and the third to $\boldsymbol{x} = (-0.25, 2.25)$. Each row depicts the estimated PDF of prediction distribution for a particular function and BTMGP cross-covariance.



(a) Separable (b) LMC

Figure 4: Two dimensional contour plot of the prediction distribution for $\boldsymbol{x}_1 = (-1.5, 1.5)$ using (a) the separable cross-covariance and (b) the LMC cross-covariance. The red stars denote the real values.

the two devices. In the adsorber, fresh sorbent medium reacts and traps the CO_2 from the exhaust flue gas. The depleted sorbent is then transferred to the regenerator, where the reverse chemical reaction releases the carbon dioxide back into the gaseous phase. The CO_2 released in the regenerator is liquefied for sequestration and the regenerated sorbents are recycled back to the adsorber.

The bulk of the energy penalty is associated with the regenerator (MacDowell et al., 2010) and therefore efforts to increase capture plant efficiency should begin with optimizing the regenerator performance. Regenerator efficiency is maximized if the solid fraction throughout the regenerator is homogeneous and close to the optimal value. Clustering behavior can result in significant reduction of the overall chemical kinetics of gas-solid fluidized bed reactors (Holloway and Sundaresan, 2012). Cluster formation results in segregation of the reacting particles and gas, which is detrimental to regenerator efficiency. We are interested in investigating the dependence of sorbent distribution on two operating conditions: the particle diameter d_p (expressed in micro meters, μm), and the superficial gas velocity v_g scaled by the minimum particle fluidization velocity u_{mf}, which we denote by K.

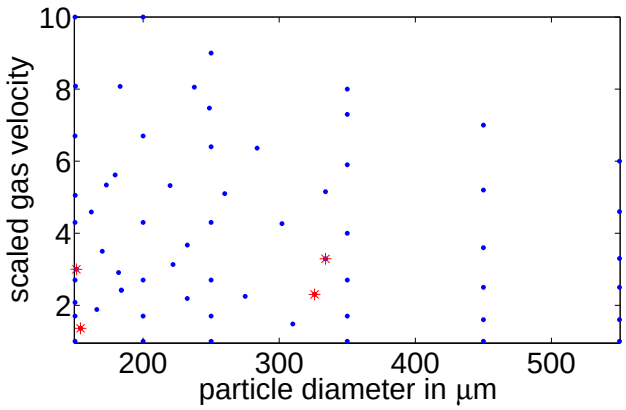


Figure 5: Input of the observed data (blue dot) and the input of cross-validation data (red stars).

For our purposes, six bins of the empirical solid fraction distribution are considered sufficient enough to distinguish between the dilute, intermediate, and dense region in the flow.

The full solid fraction range of 0.0 to 0.6 is subdivided into six bins of fixed length, given by $[0, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5]$, and $(0.5, 0.6]$. The frequency distribution of the number of cells (i.e., regenerator bed volume) lying in each bin is calculated as the response.

The above problem can be summarized as two input parameters with six responses. Konomi et al. (2014) developed a model based on the BTMGPS to predict and investigate this data. Here, we utilize these data to compare the existing BTMGPI and BTMGPS with BTMGPC model. To better investigate the performance of the different forms of cross-covariance functions inside the Bayesian tree we sample the computer code 4 more times as it is shown in Fig. 5 and compute the six bin empirical solid fraction distribution. The new observations have different degrees of distance from the previous observations in order to better see the efficiency of the model. These values will serve as cross-validation to our application and as a comparison of different BTMGP models.

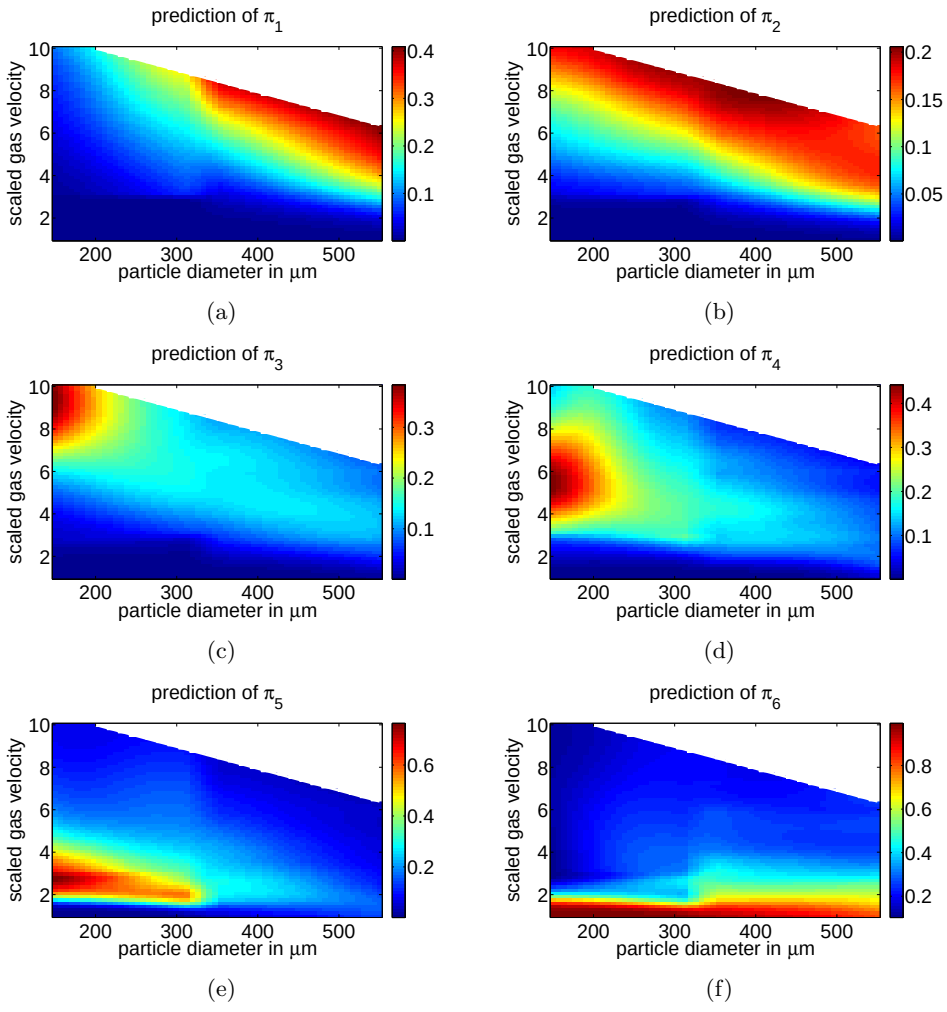


Figure 6: BMA prediction surface of the six different probabilities using BTMGP with conditional LMC cross-covariance.

The *Grow* operation in the case of BTMGPC and BTMGPI becomes extremely cumbersome in comparison to the BTMGPS. This is because we have to propose 12 new parameters from

the prior specification in comparison with only 2 of the separable model. This makes the mixing slow and not very efficient in these type of problems. However, after running the algorithm for $25,000$ MCMC iterations we observe satisfactory results. The posterior and the predictions of the BTMGP with different cross-covariances are mostly similar. However, predictions show improvements when we use BTMGPC.

Fig. 6 shows the prediction surface of the six different probabilities π_i in a grid of (70×70) using BMA in the BTMGPC. Operating conditions lying in the upper-right corner are not of interest and therefore no simulations were performed for those values. If good regeneration is expected for an intermediate solid fraction range of, say, 0.3 to 0.4, the area of interest would be regions where π_4 is large. From Figure 6(d), the region where π_4 is large is given by $d_p \in (150\,\mu\text{m}, 250\,\mu\text{m})$ and scaled gas velocity $v_g/u_{mf} \in (4.0, 8.0)$. These results are similar to the results driven from the BTMGPS which are presented in (Konomi et al., 2014).

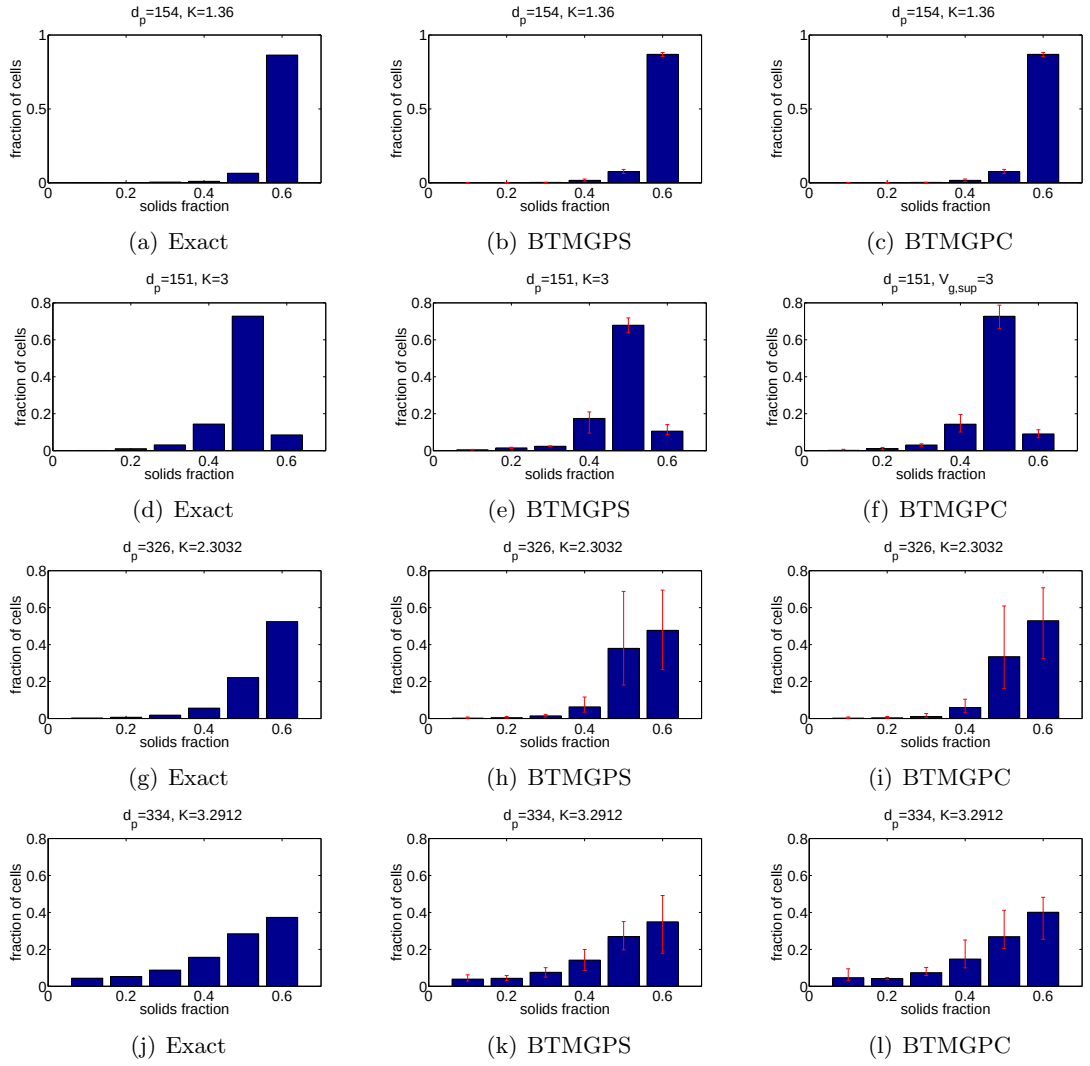
To better evaluate the prediction abilities of the different models a cross-validation analysis should be considered. We compute and compare the predicted solid fraction distribution of four input values using BTMGPS and BTMGPC with the solid fraction distribution computed directly from the computer code runs which are shown in the first column of Fig. 7. The closest the predictions are the better the model. The second and the third column of Fig. 7 shows the mean prediction solid fraction distribution and their 95% confidence interval, of the four observed computer experments, computed by the two differnt cros-covariance in the BTMGP. Each row reprent one of the four differnt combination of different combinations of particle diameter d_p and scaled gas velocity K where we have observations. The second column shows the mean prediction solid fraction distribution and their 95% confidence interval computed by BTMGPS. The third shows the mean prediction solid fraction distribution and their 95% confidence interval computed by BTMGPC.

In general the BTMGPC gives better predictions for the four distributions. For example, in the first, second and third row the prediction probability of π_5 and π_6 using LMC is closer to the real value of the computer code. Only in the case of input $d_p = 334$ and $K = 2.29$ does the BTMGP with separable cross-covariance give predictions closer to the real computer code. Despite these differences, the two models give approximately similar results. Cross-validation values close to observations give smaller confidence intervals. Also, probabilities close to 0 or 1 give also small confidence intervals. For example, the point $(154, 1.36)$, gives smaller prediction confidence intervals than all the other the rest cross-validation points. The LMC model in the BTMGP is a relatively better but we should usually account for the computational complexity these model is introducing.

Note: The prediction differences in this application appear to be minor due to the dependencies between the outputs. This application is closer in spirit to the 2-input and 2-output simulation study.

5. Concluding remarks and extensions

We developed a Bayesian treed multivariate Gaussian process (BTMGP) based on the linear model of coregionalization (LMC). The conditional representation of the LMC cross-covariance simplifies the form of the inverse and determinant of the covariance matrix involved in the MCMC updates. Moreover, the *Grow* and *Prune* operations of the Bayesian tree are facilitated in the conditional LMC cross-covariance by integrating out the linear model parameters and the variance. Only the parameters of the correlation function need to be updated in each MCMC iteration for prediction purposes.



(a) Exact (b) BTMGPS (c) BTMGPC

(d) Exact (e) BTMGPS (f) BTMGPC

(g) Exact (h) BTMGPS (i) BTMGPC

(j) Exact (k) BTMGPS (l) BTMGPC

Figure 7: Prediction probabilities and their 95% confidence intervals using different models for four different combinations of particle diameter d_p and scaled gas velocity K. First column gives the probabilities as they are computed form the computer code, second column gives the prediction probabilities and their 95% confidence intervals using BTMGPS and the third column gives the prediction probabilities and their 95% confidence intervals using BTMGPC.

The proposed BTMGP with conditional LMC cross-covariance (BTMGPC) is compared to the BTMGP with independent cross-covariance (BTMGPI), multiple BTGP (Gramacy and Lee, 2008) and the newly developed BTMGP with separable cross-covariance (BTMGPS) (Konomi et al., 2014). These comparisons allow us to see in practice the strength and the weakness of each model. The prediction results of BTMGPC are similar to those of BTMGPS when dependence between different dimensions of the output is present (second simulation study), but more robust when the output functions are independent to each other (first simulation study). In comparison with the BTGP the proposed model shows better results when there is a dependency between outputs and similar when the dependence assumption is violated.

In cases where the number of the input/output variables is large, the *Grow* and *Prune*

operation may not perform well in BTMGP with LMC. This is mainly because the reversible jump MCMC involves moves between states with large differences in dimensionality, and thus it may present high rejection rates. To address this issue, more sophisticated variations of the Reversible Jump MCMC may be used, such as those in Brooks et al. (2003); Karagiannis and Andrieu (2013). The examples presented in the paper provide encouraging results for further work. Finally, Konomi et al. (2013) can be used to simplify the computations when dealing with huge amount of datasets or link the cross-covariance for different tree subregions. However, we leave these type of extensions for future work.

Acknowledgments

The research at Pacific Northwest National Laboratory (PNNL) was supported by the Department of Energy Carbon Capture Simulation Initiative. PNNL is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830. The authors thank Dr. Avik Sarkar and Dr. Xin Sun for helping with the computer simulations in carbon capture regenerator.

Appendix A. Metropolis within Gibbs sampler for $(\boldsymbol{\lambda}_j, \boldsymbol{g}_j)$ in a partition

Let us denote $\boldsymbol{\psi}_j = (\boldsymbol{\lambda}_j, \boldsymbol{g}_j) = (\lambda_{j,1}, \ldots, \lambda_{j,k_x}, \boldsymbol{g}_j)$, the joint vector of the parameters of the correlation function, where $\boldsymbol{\psi}_j \in (0,\infty)^{k_x+1}$, for $j = 1, \ldots, q$. To sample from the posterior distribution $p(\boldsymbol{\psi}_j|\boldsymbol{Y}_j)$, we apply a Metropolis within Gibbs as in (Mueller, 1993), that performs recursively Metropolis-Hastings updates in a component-wise manner. At time t, let $\psi_{j,k}^{(t)}$ be the k-th component of $\boldsymbol{\psi}_j^{(t)} = (\psi_{j,1}^{t+1}, \ldots, \psi_{j,k-1}^{t+1}, \psi_{j,k}^{t}, \ldots, \psi_{j,(k_x+1)}^{t})$ and $\boldsymbol{\psi}_{j,(-k)}^{(t)} = (\psi_{j,1}^{t+1}, \ldots, \psi_{j,k-1}^{t+1}, \psi_{j,k+1}^{t}, \ldots, \psi_{j,(k_x+1)}^{t})$. Given that at time t, the chain is at state $\boldsymbol{\psi}_j^{(t)}$, the algorithm works as follows.

For $k = 1, \ldots, (k_x + 1)$:

1. Generate $\psi_{j,k}^{*} \sim q_k(\psi_{j,k}|\psi_{j,k}^{(t)}) \equiv \text{logN}(\psi_{j,k}|\psi_{j,k}^{(t)}, s)$ where logN is the log-Normal distribution and $s \in (0,\infty)$ is a user defined scaling parameter.
2. Compute:

$$r_j = \frac{p(\psi_{j,k}^{*}|\boldsymbol{Y}_j, \boldsymbol{\psi}_{j,(-k)}^{(t)}) q_k(\psi_{j,k}^{(t)}|\psi_{j,k}^{*})}{p(\psi_{j,k}^{(t)}|\boldsymbol{Y}_j, \boldsymbol{\psi}_{j,(-k)}^{(t)}) q_k(\psi_{j,k}^{*}|\psi_{j.k}^{(t)})} = \frac{p(\psi_{j,k}^{*}|\boldsymbol{Y}_j, \boldsymbol{\psi}_{j,(-k)}^{(t)})\psi_{j,k}^{*}}{p(\psi_{j,k}^{(t)}|\boldsymbol{Y}_j, \boldsymbol{\psi}_{j,(-k)}^{(t)})\psi_{j,k}^{(t)}}. \tag{A.1}$$

   For $\psi_{j,k} = \lambda_{j,k}$ the posterior $p(\psi_{j,k}|\boldsymbol{Y}_j, \boldsymbol{\psi}_{j,(-k)}^{(t)}) = p(\lambda_{j,k}|\boldsymbol{Y}_j, \boldsymbol{g}_j^{(t)})$ and for $\psi_{j,(k_x+1)} = \boldsymbol{g}_j$ the posterior $p(\psi_{j,(k_x+1)}|\boldsymbol{Y}_j, \boldsymbol{\psi}_{j,(-(k_x+1))}^{(t)}) = p(\boldsymbol{g}_j|\boldsymbol{Y}_j, \boldsymbol{\lambda}_j^{(t+1)})$.
3. Set $\psi_{j,k}^{(t+1)} = \psi_{j,k}^{*}$ with probability $\min(1, r_j)$ and $\psi_{j,k}^{(t+1)} = \psi_{j,k}^{(t)}$ with the remaining probability.

In this algorithm, the MH step is performed only once at each iteration. Chen and Schmeiser (1998) note that multiple MH steps are not necessary. A precise approximation of the conditional probability does not necessary lead to a better approximation of the join distribution, and a single step may be beneficial for the speed of the sampler.

Appendix B. Active Learning for BTMGP with conditional LMC cross-covariance

The sequential experimental update of the subset, usually called *active learning*, is proven to be a good choice to find the best possible subset. Two main approaches of active learning are: *Active Learning MacKay* (ALM) and *Active Learning Chon* (ALC). A detail description of these techniques for the univariate case can be found in Seo et al. (2000) and Gramacy and Lee (2009). Here we give a brief description on how we can be extend the ALC to the BTMGP with conditional LMC cross-covariance.

The ALC approach sequentially selects a subset of data by maximizing the expected reduction in mean square error:

$$\Delta\tilde{\sigma}^2(\tilde{\boldsymbol{x}}) = \int_{\mathcal{X}} \Delta\hat{\sigma}^2_{\tilde{\boldsymbol{x}}}(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} = \int_{\mathcal{X}} (\hat{\sigma}^2(\boldsymbol{z}) - \hat{\sigma}^2_{\tilde{x}}(\boldsymbol{z}))p(\boldsymbol{z})d\boldsymbol{z}. \tag{B.1}$$

where $\Delta\hat{\sigma}^2_{\tilde{\boldsymbol{x}}}(\boldsymbol{x})$ is the reduction of variance of the output in location $\boldsymbol{z}$ when we add an observation in location $\tilde{\boldsymbol{x}}$ which will make the total observation $\boldsymbol{X}_{N+1} = [\boldsymbol{X}_N, \tilde{\boldsymbol{x}}]$. Also, $p(\boldsymbol{z})$ is the input variable density function which can be considered as a prior of the input space (a generalized *Beta* or truncated *Normal* distribution is usually a good choice in practice), $\sigma^2(\boldsymbol{z})$ is the variance mean of output in location $\boldsymbol{z}$ without observing the output in location $\tilde{\boldsymbol{x}}$ and $\hat{\sigma}^2_{\tilde{\boldsymbol{x}}}(\boldsymbol{z})$ is the variance mean at location $\boldsymbol{z}$ when we have an observation at location $\tilde{\boldsymbol{x}}$. The variance mean is the mean of each conditional representation of the LMC. Because of the independence assumption of the Bayesian tree, if $\boldsymbol{z}$ and $\tilde{\boldsymbol{x}}$ belong in two different external nodes, we take $\Delta\hat{\sigma}_{\tilde{\boldsymbol{x}}}(\boldsymbol{z}) = 0$.

The above integral is usually analytically intractable and as such we compute it numerically by choosing a predetermine subset of gridded input data $\boldsymbol{X}$. For each of the j^{th} conditional representations of the LMC we can write:

$$\hat{\sigma}^2_j(\boldsymbol{z}) = \mathrm{tr}\big\{(\boldsymbol{r}_{j,N}(\boldsymbol{z},\boldsymbol{z}) - \boldsymbol{r}_j^T(\boldsymbol{X}_N,\boldsymbol{z})\boldsymbol{R}^{-1}_{j,N}\boldsymbol{r}_{j,N}(\boldsymbol{z},\boldsymbol{z}))\sigma_j^2\big\} + \tau_j^2(\boldsymbol{z}), \tag{B.2}$$

$$\hat{\sigma}^2_{j,\tilde{\boldsymbol{x}}}(\boldsymbol{z})) = \mathrm{tr}\big\{(\boldsymbol{r}_j(\boldsymbol{z},\boldsymbol{z}) - \boldsymbol{r}_j^T(\boldsymbol{X}_{N+1},\boldsymbol{z})\boldsymbol{R}^{-1}_{j,(N+1)}\boldsymbol{r}_j(\boldsymbol{X}_{N+1},\boldsymbol{z}))\sigma_j^2\big\} + \tau^2_{j,\tilde{\boldsymbol{x}}}(\boldsymbol{z}). \tag{B.3}$$

where

$$\begin{aligned}\tau_j^2 &= (\hat{\alpha}^{j|1})^2\hat{\sigma}_1^2(\boldsymbol{z}) + \cdots + (\hat{\alpha}^{j|j-1})^2\hat{\sigma}^2_{j-1}(\boldsymbol{z}) \\ &\quad + \hat{\alpha}^{j|1}\hat{\alpha}^{j|2}\,\mathrm{cov}(\eta_1(\boldsymbol{x}'),\eta_2(\boldsymbol{z})|\boldsymbol{X}_N) + \cdots + \hat{\alpha}^{j|(j-2)}\hat{\alpha}^{j|(j-1)}\,\mathrm{cov}(\eta_{(j-2)}(\boldsymbol{z}),\eta_{(j-1)}(\boldsymbol{z})|\boldsymbol{X}_N)\end{aligned}$$

and

$$\begin{aligned}\tau^2_{j,\tilde{\boldsymbol{x}}} &= (\hat{\alpha}^{j|1})^2\hat{\sigma}^2_{1,\tilde{\boldsymbol{x}}}(\boldsymbol{z}) + \cdots + (\hat{\alpha}^{j|j-1})^2\hat{\sigma}^2_{j-1,\tilde{\boldsymbol{x}}}(\boldsymbol{z}) \\ &\quad + \hat{\alpha}^{j|1}\hat{\alpha}^{j|2}\,\mathrm{cov}(\eta_1(\boldsymbol{z}),\eta_2(\boldsymbol{z})|\boldsymbol{X}_{N+1}) + \cdots + \hat{\alpha}^{j|(j-2)}\hat{\alpha}^{j|(j-1)}\,\mathrm{cov}(\eta_{(j-2)}(\boldsymbol{z}),\eta_{(j-1)}(\boldsymbol{z})|\boldsymbol{X}_{N+1}).\end{aligned}$$

The covariance terms of $\mathrm{cov}(\eta_{(j-2)}(\boldsymbol{z}),\eta_{(j-1)}(\boldsymbol{z})|\boldsymbol{X})$ can be computed through the variance and the coefficient $\hat{\alpha}^{(j-1)|(j-2)}$, e.g. $\mathrm{cov}(\eta_{(2)}(\boldsymbol{z}),\eta_{(1)}(\boldsymbol{z})|\boldsymbol{X}) = \hat{\alpha}^{2|1}(\boldsymbol{r}_j(\boldsymbol{z},\boldsymbol{z}) - \boldsymbol{r}_j^T(\boldsymbol{X},\boldsymbol{z})\boldsymbol{R}_j^{-1}\boldsymbol{r}_j(\boldsymbol{X},\boldsymbol{z}))$. We can express all the terms in Eq. B.2 and Eq. B.3 as a spatial variance. For each of these differences of the spatial variance we can utilize the matrix inversion in low cost proposed by Gramacy and Lee (2009). We follow the same setting to invert $(N+1)\times(N+1)$ covariance matrices in terms of $N\times N$ covariance matrices. More details on how to improve sampling with ALC are given in Gramacy and Lee (2009).

References

Apanasovich, T. V., Genton, M. G., 2010. Cross-covariance functions for multivariate random fields based on latent dimensions. Biometrika 97, 15–30.

Banerjee, S., Carlin, B., Gelfand, A., 2004. Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall-CRC, Boca Raton, FL.

Berger, J. O., Oliveira, V. D., Sanso, B., 2001. Objective Bayesian analysis of spatially correlated data. Journal of the American Statistical Association 96, 1361–1374.

Brooks, S. P., Giudici, P., Roberts, G. O., 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65 (1), 3–39.

Chen, M.-H., Schmeiser, B., 1998. Monte Carlo methods on Bayesian analysis of constrained parameter problems with normalizing constants. Journal of Computational and Graphical Statistics 7, 1–22.

Chipman, H., George, E., McCulloch, R., 1998. Bayesian CART model search. Journal of the American Statistical Association 93, 935–960.

Conti, S., O'Hagan, A., 2010. Bayesian emulation of complex multi-output and dynamic computer models. Journal of Statistical Planning and Inference 140, 640–651.

Cressie, N., 1993. Statistics for Spatial Data. 2nd edition. John Wiley and Sons Inc, New York.

Cressie, N., Wickle, C., 2011. Statistics for Spatial-Temporal Data. John Wiley and Sons Inc, New York.

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1988. A bayesian approach to the design and analysis of computer experiments. Tech. rep., ORNL-6498, Oak Ridge Laboratory.

Gelfand, A., Diggle, P., Fuentes, M., Guttorp, P., 2010. Handbook of Spatial Statistics. Chapman & Hall-CRC, Boca Raton, FL.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., Sirmans, C. F., 2004. Non-stationary multivariate process modeling through spatially varying coregionalization. TEST 13, 263–312.

Gelfand, A. E., Smith, A. F. M., 1990. Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85, 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B., 2004. Bayesian Data Analysis. Boca Raton: Chapman & Hall/CRC.

Godsill, S., 2001. On the relationship between markov chain monte carlo methods for model uncertainty. Journal of Computational and Graphical Statistics 10 (2), 230–248.

Gramacy, R. B., Lee, H. K., 2012. Cases for the nugget in modeling computer experiments. Statistics and Computing 22 (3), 713–722.

Gramacy, R. B., Lee, H. K. H., 2008. Bayesian treed Gaussian process models with an application to computer modeling. Journal of the American Statistical Association 103, 1119–1130.

Gramacy, R. B., Lee, H. K. H., 2009. Adaptive design and analysis of supercomputer experiments. Technometrics 51, 130–145.

Green, P., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82, 711–732.

Grzebyk, M., Grzebyk, M., Wackernagel, H., Wackernagel, H., 1994. Multivariate analysis and spatial/temporal scales: Real and complex models.

Holloway, W., Sundaresan, S., 2012. Filtered models for reacting gas-particle flows. Submitted for review to Chemical Engineering Science.

Karagiannis, G., Andrieu, C., 2013. Annealed importance sampling reversible jump mcmc algorithms. Journal of Computational and Graphical Statistics 22 (3), 623–648.

Konomi, B., Karagiannis, G., Sarkar, A., Lin, G., 2014. Bayesian treed multivariate gaussian process with adaptive design: Application to a carbon capture unit. Technometrics 56, 145–158.

Konomi, B., Sang, H., Mallick, B., 2013. Adaptive bayesian nonstationary modeling for large spatial datasets using covariance approximations. Journal of Computational and Graphical Statistics 1, In press.

MacDowell, N., Florin, N., Buchard, A., Hallett, J., Galindo, A., Jackson, G., Adjiman, C., Williams, C., Shah, N., Fennell, P., 2010. An overview of CO_2 capture technologies. Energy & Environmental Science 3 (11), 1645–1669.

Mardia, K. V., Goodall, C. R., 1993. Spatial temporal analysis of multivariate environmental monitoring data. In Multivariate Environmental Statistics (G. P. Patil and C. R. Rao, eds.), 347–386.

Matheron, G., 1982. Pour une analyse krigeante des données régionalisées. Tech. rep., Centre de Gostatistique, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, France.

Mueller, P., 1993. Alternatives to the gibbs sampling scheme. Tech. rep., Institute Statistics and Decision Sciences, Duke University.

Oakley, J., O'Hagan, A., 2002. Bayesian inference for the uncertainty distribution of computer model outputs. Biometrika 89 (4), 769–784.
URL `http://biomet.oupjournals.org/cgi/doi/10.1093/biomet/89.4.769`

O'Hagan, A., Kennedy, M. C., Oakley, J. E., 1999. Uncertainty analysis and other inference tools for complex computer codes (with discussion). In Bayesian Statistics 6, 503–524.

Royle, J. A., Berliner, L. M., 1999. A hierarchical approach to multivariate spatial modeling and prediction. Journal of Agricultural, Biological, and Environmental Statistics 4, 29–56.

Sacks, J., Welch, W. J., Mitchell, T. J., Wynn, H. P., 1989. Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. Statistical Science 4, 409–435.

Schmidt, A. M., O'Hagan, A., 2003. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. Journal of the Royal Statistical Society, Series B 65, 743–758.

Seo, S., Wallat, M., Graepel, T., Obermayer, K., 2000. Gaussian process regression: Active data selection and test point rejection. IEEE. In Proceedings of the International Joint Conference on Neural Networks III (5), 241–246.

Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications, 2nd Edition. Springer, Berlin.