Conditionally Independent Multiresolution Gaussian Processes

Jalil Taghia
Department of Information Technology
Uppsala University, Sweden

Thomas B. Schön
Department of Information Technology
Uppsala University, Sweden

Abstract

The multiresolution Gaussian process (GP) has gained increasing attention as a viable approach towards improving the quality of approximations in GPs that scale well to large-scale data. Most of the current constructions assume full independence across resolutions. This assumption simplifies the inference, but it underestimates the uncertainties in transitioning from one resolution to another. This in turn results in models which are prone to overfitting in the sense of excessive sensitivity to the chosen resolution, and predictions which are non-smooth at the boundaries. Our contribution is a new construction which instead assumes *conditional independence* among GPs across resolutions. We show that relaxing the full independence assumption enables robustness against overfitting, and that it delivers predictions that are smooth at the boundaries. Our new model is compared against current state of the art on 2 synthetic and 9 real-world datasets. In most cases, our new conditionally independent construction performed favorably when compared against models based on the full independence assumption. In particular, it exhibits little to no signs of overfitting.

1 INTRODUCTION

There is a rich literature on methods designed to avoid the computational bottleneck incurred by the vanilla Gaussian process (GP), including sub-sampling [30], low rank approximations [9], covariance tapering [14], inducing variables [29; 32], predictive processes [3], and multiresolution models [31; 28], to name just a few. Here, we focus mainly on the low rank approximations. Many existing GP models assume certain smoothness properties which can be counterproductive when it comes to representing abrupt local changes. Although some less smooth kernel choices can be helpful at times, they assume stationary processes that do not adapt well to varying levels of smoothness. The undesirable smoothness characteristic of the traditional GPs could further get pronounced in approximate GP methods in general and rank-reduced approximations in particular [36]. A way to overcome the limitations of low rank approximations is to recognize that the long-range dependencies tend to be of lower rank when compared to short-range dependencies. This idea has previously been explored in the context of hierarchical matrices [16; 4; 2] and in multiresolution models [31; 28; 21].

Multiresolution GPs, seen as hierarchical models, connect collections of smooth GPs, each of which is defined over an element of a random nested partition [15; 12; 11]. The long-range dependencies are captured by the GP at the top of hierarchy while the bottom-level GPs capture the local changes. We can also view the multiresolution GPs as a hierarchical application of predictive processes—approximations of the true process arising from conditioning the initial process on parts of the data [3; 29]. The use of such models has recently been exploited in spatial statistics [31; 28; 21] for modeling large spatial datasets. Refer to [12] and [21] for overviews of these applications.

The existing multiresolution models are based on predictive processes and event though they are efficient in terms of computational complexity, they do assume full independence across the different resolutions. This independence assumption results in models which are inherently susceptible to the chosen resolution and approximations which are non-smooth at the boundaries. The latter problem stems from the fact that the multiresolution framework, e.g., [21], recursively split each region at each resolution into a set of subregions. As discussed by Katzfuss and Gong [22], since the remainder process is assumed to be independent between these subregions, which can give rise to discontinuities at the region boundaries. A heuristic solution based on

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

tapering functions is proposed in [22] which employs Kanter's function as the modulating function to address this limitation. The sensitivity to the chosen resolution is partly due to the nature of the remainder process and the unconstrained representative flexibility of the GPs which manifests itself most noticeably at higher resolutions. As the size of the region under consideration decreases when the resolution increases, the remainder process may inevitably include certain aspects of data which might not be the patterns of interest. When all GPs are forced to be independent, there is no natural mechanism to constrain the representative flexibility of the GPs.

These limitations can be addressed naturally by allowing the uncertainty to propagate across the different resolutions. We achieve this by conditioning the GPs on each other. Thus, here, we propose a new model which unlike the previous models that impose full independence among resolutions, instead assumes *conditional independence*. Relaxing the full independence assumption is shown to result in models that are robust to overfitting in the sense of reduced sensitivity to the chosen resolution—that is regardless of the extra computational complexity, arbitrary increasing the resolution only has a small effect on the optimal model performance. Furthermore, it results in predictions which are smooth at the boundaries. This is facilitated by constructing a low-rank representation of the GP via a Karhunen-Loève expansion with the Bingham prior model that consists of basis axes and basis-axis scales. Our multiresolution model ties all GPs, across all resolutions, to the same set of basis axes. These axes are learned successively in a Bayesian recursive fashion. We consider a fully Bayesian treatment of the proposed model and derive a structured variational inference based on a partially factorized mean-field approximation[1].

The idea of using conditional independence in the context of multiresolution GPs has previously been studied by Fox and Dunson [12]. The two models differ in their underlying generative models and in their inference. While the computational complexity of the proposed model scales linearly with respect to the number of samples, Fox & Dunson's model scales cubically and relies on MCMC inference which may further limit its application to large datasets.

Our main *contribution* is to develop the conditionally independent multiresolution GP model and to derive a variational inference method to learn this model from data. The Bingham distribution [6] is an important distribution in directional statistics [26] where it is commonly used for shape analysis where the inference is typically based on MLE [24], MAP [27], and MCMC [25]. Hence, our use of the Bingham distribution and the corresponding variational inference solution for this model might also appeal to researchers in directional statistics.

[1] An implementation of the model is available at: https://github.com/jtaghia/ciMRGP

2 KARHUNEN-LOÈVE REPRESENTATION OF THE GP

Consider a minimalistic model of GP regression, $\boldsymbol{y}_t = \boldsymbol{f}(\boldsymbol{x}_t) + \boldsymbol{b} + \boldsymbol{e}_t, \forall t \in \mathcal{T} = \{1, \ldots, n\}$, where $\boldsymbol{f} \sim \mathcal{GP}(\cdot)$ denotes a zero-mean GP prior, $\boldsymbol{b}$ denotes a constant bias, $\boldsymbol{e}_t \sim \mathcal{N}(\boldsymbol{0}, \gamma^{-1}\mathbf{I})$ denotes Gaussian noise with zero mean and variance γ^{-1}, $\boldsymbol{x}_t \in \mathbb{R}^{d_x}$ denotes the input variables, and $\boldsymbol{y}_t \in \mathbb{R}^{d_y}$ denotes the measurements, $d_x, d_y \in \mathbf{N}_{\geq 1}$. The standard solution involves inversion of a Gram matrix which is an $\mathcal{O}(n^3)$ operation in general. In the following, we consider low rank representations of the GP enabled via the Karhunen-Loève expansion theorem.

Gaussian Model For a d_x-dimensional input variable $\boldsymbol{x}_t$ on the interval $[-L_1, L_1] \times \ldots \times [-L_{d_x}, L_{d_x}] \in \mathbb{R}^{d_x}$, the GP can be represented using the (truncated) Karhunen-Loève expansion according to [34],

$$\boldsymbol{f}(\boldsymbol{x}_t) \approx \sum_{i=1}^{p} \boldsymbol{w}_i \phi_i(\boldsymbol{x}_t, \boldsymbol{\tau}), \;\; \forall \; \boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{0}, S(\sqrt{\lambda_i(\boldsymbol{\tau})})\mathbf{I}), \quad (1)$$

where $\boldsymbol{w}_i = (w_{i1}, \ldots, w_{id_y})^\top$ denotes the basis vectors of the series expansion, $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{d_x})^\top$ denotes the basis intervals such that $\tau_d > L_d, \forall d \in \{1, \ldots, d_x\}$, $\phi_i(\boldsymbol{x}_t, \boldsymbol{\tau})$ denotes the orthogonal eigenfunctions (basis functions) with the corresponding eigenvalues $\lambda_i(\boldsymbol{\tau})$, and $S(\cdot)$ denotes the spectral density of the covariance function. Note that, unlike the minimalistic representation used by Solin and Särkkä [34], we have explicitly included the basis intervals $\boldsymbol{\tau}$ in the representation, which are treated as random variables. Their specific values are found using maximum likelihood estimation.

To ensure that the representation satisfies the dual orthogonality requirement of the Karhunen-Loève expansion, all the basis vectors $\boldsymbol{w}_i$ must be zero-mean. Normally, we would assign a zero-mean Gaussian distribution over $\boldsymbol{w}_i$, or alternatively we could assign a zero-mean matrix-normal distribution over $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_p)$ as was done by Svensson and Schön [37]. The choice of zero-mean Gaussian priors over the basis vectors would lead to Gaussian posteriors with *non-zero* means. In our multiresolution model, as we shall see later in Sec. 3, the basis vector posterior needs to be learned in a recursive fashion such that the posterior from the current resolution is used as the prior for the resolution in the next level of the hierarchy. Now, as the expansion requires the prior to be zero-mean, we

would then need a posterior over basis vectors which is zero-mean by construction. If we were going to use Gaussian priors, the result would be a multiresolution model where all GPs must be fully independent.

To address this issue, we now separate the basis vectors into two parts: *basis axes* and *basis-axis scales*. The basis axis vectors are defined to be *antipodally symmetric*—meaning that for a random variable $\boldsymbol{\vartheta}$, $p(\boldsymbol{\vartheta}) = p(-\boldsymbol{\vartheta})$—and thus zero-mean by construction. They primarily carry information about the direction and we can for that reason without loss of generality assume them to be on the unit sphere. The axes will be shared across resolutions such that given the axes, all GPs are independent. Although the GPs are tied to the same set of axes, they will be scaled by resolution-specific variables, namely the basis-axis scales. The axial distributions from directional statistics [26] make for a perfect fit in modeling these axes. In the following we consider a very specific choice of prior model, namely the *Bingham distribution*, since it conveniently allows for the design of a conditionally independent multiresolution model.

Bingham Model Let $\mathcal{S}^{d-1} = \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{z}^\top \boldsymbol{z} = 1, d \in \mathbb{N}_{>1}\}$ denote the unit sphere. Furthermore, let $\boldsymbol{w}_i := a_i \boldsymbol{u}_i$ such that $\boldsymbol{u}_i = \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|} \in \mathcal{S}^{d_y-1}$ and $a_i = \|\boldsymbol{w}_i\|$ denote the basis axes and the basis-axis scales, respectively. Without loss of generality, we can now express the noisy measurements in (1) as

$$\boldsymbol{y}_t = \sum_{i=1}^{p} a_i \boldsymbol{u}_i \phi_i(\boldsymbol{x}_t, \boldsymbol{\tau}) + \boldsymbol{b} + \boldsymbol{e}_t, \qquad \forall t \in \mathcal{T}. \tag{2}$$

The basis axes $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_p)$ are modeled as *Bingham distributions* [6] according to

$$p(\boldsymbol{U}) = \prod_{i=1}^{p} p(\boldsymbol{u}_i), \qquad \forall\ \boldsymbol{u}_i \sim \mathcal{B}(\mathbf{B}_i),$$

where $\mathcal{B}(\mathbf{B}_i)$ denotes the Bingham distribution parameterized with a real-symmetric matrix $\mathbf{B}_i$—the matrix $\mathbf{B}_i$ is often presented using the notion of an eigendecomposition as: $\mathbf{B}_i = \mathbf{M}_i \times \text{diag}[\boldsymbol{\kappa}_i] \times \mathbf{M}_i^\top$ with $\mathbf{M}_i$ and $\boldsymbol{\kappa}_i$ being the eigenvectors and the eigenvalues of the decomposition. It is straightforward to show that $\boldsymbol{u}_i$ satisfies the Karhunen-Loève expansion requirements. Importantly, the Bingham distribution is antipodally symmetric, which in turn implies that $\mathrm{E}[\boldsymbol{u}_i] = 0$ by construction [26, Ch. 9.4]. We can then assign zero-mean Gaussian distributions as priors over the basis-axis scale variables $\{a_i\}_{i=1}^p$. Assuming $\boldsymbol{e}_t \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$, and using $\|\boldsymbol{u}_i\| = 1$, this choice of prior over $\boldsymbol{u}_i$ and a_i is conveniently *conjugate* to the data likelihood.

The main constraint enforced by our choice of the Bingham prior model is the implicit requirement of $d_y > 1$, as the Bingham density is defined on $\mathcal{S}^{d_y-1}$. For the case of $d_y = 1$, if we assume $\boldsymbol{u}_i = 1$, the Bingham model reduces to a multiresolution architecture with fully independent GPs. Other prior models should be considered for the special case of $d_y = 1$. One possible choice is provided by the one-parameter version of the Bingham model [23] for modeling axes concentrated asymmetrically near a small circle. As the objective of this work is to show the advantage of the conditional independence over the full independence, we restrict our theoretical discussion to the Bingham prior model and cases where $d_y > 1$.

3 MODEL

Notation Consider a recursive partitioning of the index set $\mathcal{T} = \{1, \ldots, n\}$ across m resolutions. At each resolution $j \in \{1, \ldots, m\}$, $\mathcal{T}$ is partitioned into a number of non-overlapping regions. The partitioning of $\mathcal{T}$ can be structured or random. Without loss of generality, consider a uniform subdivision of the index set across resolutions by a factor of $\mathfrak{q}$, such that $\mathcal{T}$ is first partitioned into $\mathfrak{q}$ regions, each of which is then partitioned into $\mathfrak{q}$ subregions. The partitioning continues until resolution m where the index sets at various resolution are denoted by $\mathcal{T}^{(0)} := \mathcal{T}$, $\mathcal{T}^{(1)} = \{\mathcal{T}_1^{(0)}, \ldots, \mathcal{T}_{\mathfrak{q}}^{(0)}\}$, and similarly by $\mathcal{T}^{(m)} = \{\mathcal{T}_1^{(m-1)}, \ldots, \mathcal{T}_{\mathfrak{q}}^{(m-1)}\}$, where $|\mathcal{T}^{(0)}| = 1$, $|\mathcal{T}^{(1)}| = \mathfrak{q}$, and $|\mathcal{T}^{(m)}| = \mathfrak{q}^m$. An example of such a partitioning by a factor of $\mathfrak{q} = 2$ is shown in Fig. 1-a. As a convention, we will use the notation $\mathcal{T}_l^{(j)}$ to indicate the l-th element of the set $\mathcal{T}^{(j)} = \{\mathcal{T}_l^{(j)}\}_{l=1}^{|\mathcal{T}^{(j)}|}$, which corresponds to the index set related to region l at resolution j. We also define $\boldsymbol{x}_{\mathcal{T}^{(0)}} := \boldsymbol{x}_{\mathcal{T}}$ and $\boldsymbol{x}_{\mathcal{T}^{(j)}} = \{\boldsymbol{x}_{\mathcal{T}_l^{(j)}}\}_{l=1}^{|\mathcal{T}^{(j)}|}$, where $\boldsymbol{x}_{\mathcal{T}_l^{(j)}} = \{\boldsymbol{x}_t \mid \forall t \in \mathcal{T}_l^{(j)}\}$.

Generative Model As before, let $\boldsymbol{f}(\cdot)$ be the stochastic process of interest. Once the process is observed at $\boldsymbol{x}_{\mathcal{T}}$, it gives rise to the noisy observations $\boldsymbol{y}_t$. By making use of a Gaussian process as the prior over $\boldsymbol{f}(\cdot)$, the observations $\boldsymbol{y}_t$ at resolution $j = 0$ are modeled according to (2). In a multiresolution setting based on the hierarchical application of predictive processes, we approximate $\boldsymbol{f}(\cdot)$ according to

$$\boldsymbol{f}(\cdot) = \widehat{\boldsymbol{f}}^{(0)}(\cdot) + \boldsymbol{f}^{(1)}(\cdot),$$

where $\widehat{\boldsymbol{f}}^{(0)}$ is the approximate predictive process at resolution $j = 0$, and $\boldsymbol{f}^{(1)}(\cdot)$ is the so-called *remainder process*. Let $\boldsymbol{z}_{t,l}^{(1)}$ indicate the noisy instantiations of the latent process $\boldsymbol{f}^{(1)}(\cdot)$ at $\boldsymbol{x}_{\mathcal{T}^{(1)}}$. We will treat $\boldsymbol{z}_{t,l}^{(1)}$ as a *latent variable*, and model it using a conditionally independent GP prior, for all $\boldsymbol{x}_t \in \boldsymbol{x}_{\mathcal{T}_l^{(1)}}$,

$$\boldsymbol{z}_{t,l}^{(1)} = \sum_{i=1}^{p} a_{i,l}^{(1)} \boldsymbol{u}_i \phi_i^{(1)}(\boldsymbol{x}_t, \boldsymbol{\tau}_l^{(1)}) + \boldsymbol{b}_l^{(1)} + \boldsymbol{e}_{t,l}^{(1)},$$

where the basis axes $\boldsymbol{u}_i$ are shared among all the processes while the basis-axis scales $a_{i,l}^{(1)}$ are region specific. At the higher resolution, $j=2$, the latent process $\boldsymbol{f}^{(1)}(\cdot)$ is in turn approximated by $\boldsymbol{f}^{(1)}(\cdot)=\widehat{\boldsymbol{f}}^{(1)}(\cdot)+\boldsymbol{f}^{(2)}(\cdot)$. In general, for resolution j we have

$$\boldsymbol{f}^{(j)}(\cdot)=\widehat{\boldsymbol{f}}^{(j)}(\cdot)+\boldsymbol{f}^{(j+1)}(\cdot),$$

where $\boldsymbol{f}^{(j+1)}(\cdot)$ is the remainder process at resolution $j+1$ whose noisy instantiations on $\mathcal{T}^{(j+1)}$ are modeled according to, $\forall \boldsymbol{x}_t \in \boldsymbol{x}_{\mathcal{T}_l^{(j+1)}}$:

$$\boldsymbol{z}_{t,l}^{(j+1)}=\sum_{i=1}^{p} a_{i,l}^{(j+1)}\boldsymbol{u}_i\phi_i^{(j+1)}(\boldsymbol{x}_t,\boldsymbol{\tau}_l^{(j+1)})+\boldsymbol{b}_l^{(j+1)}+\boldsymbol{e}_{t,l}^{(j+1)}.$$

Throughout, $\boldsymbol{u}_i$ has been written without indexing w.r.t. l and j. This is to emphasize that these are shared across all resolutions and regions such that in transition from one resolution to another, the axes of the basis vectors remain the same but they may be scaled differently via a region-specific and resolution-specific variable $a_{i,l}^{(j)}$. The noise variable is indexed w.r.t. both l and j, but we could alternatively assume the noise to be a resolution-specific variable. In a multiresolution model, bias may not be simply removed as a part of the preprocessing step, as the bias at each resolution carries uncertainties from the previous resolutions. These parameters are expressed using indexing on both j and l. We have indicated the basis functions with indexing on j, as generally one might consider a different choice of basis functions at different resolutions. The basis interval variables $\boldsymbol{\tau}_l^{(j)}$ are learned from data and expressed with both j and l.

The recursive procedure continues until resolution $j=m$ is reached. By assuming that the latent remainder process at $j=m+1$ approaches zero, we can approximate $\boldsymbol{f}(\cdot)$ as the sum of the predictive processes from all resolutions,

$$\boldsymbol{f}(\cdot)=\boldsymbol{f}^{(m+1)}(\cdot)+\sum_{j=0}^{m}\widehat{\boldsymbol{f}}^{(j)}(\cdot)\approx\sum_{j=0}^{m}\widehat{\boldsymbol{f}}^{(j)}(\cdot),$$

where $\widehat{\boldsymbol{f}}^{(0)}$ captures global patterns and finer details are captured at higher resolutions.

4 BAYESIAN INFERENCE

Notation Let $\boldsymbol{y}_{\mathcal{T}^{(0)}} := \boldsymbol{y}_{\mathcal{T}}$ where $\boldsymbol{y}_{\mathcal{T}}=\{\boldsymbol{y}_t \mid \forall t\in\mathcal{T}\}$ denote the set of noisy observations, and $\boldsymbol{z}_{\mathcal{T}^{(j)}}=\{\boldsymbol{z}_{\mathcal{T}_l^{(j)}}\}_{l=1}^{|\mathcal{T}^{(j)}|}$ denote the set of latent variables for $j\geq 1$, where $\boldsymbol{z}_{\mathcal{T}_l^{(j)}}=\{\boldsymbol{z}_{t,l}^{(j)} \mid \forall t\in\mathcal{T}_l^{(j)}\}$. We denote the latent function instantiations at $\boldsymbol{x}_{\mathcal{T}_l^{(j)}}$ by $\mathbf{f}_{\mathcal{T}_l^{(j)}}=\{\mathbf{f}_l^{(j)}(\boldsymbol{x}_t)\equiv\mathbf{f}_{l,t}^{(j)} \mid \forall \boldsymbol{x}_t\in\boldsymbol{x}_{\mathcal{T}_l^{(j)}}\}$. Similarly, let $\mathbf{f}_{\mathcal{T}^{(j)}}=\{\mathbf{f}_{\mathcal{T}_l^{(j)}}\}_{l=1}^{|\mathcal{T}^{(j)}|}$. Furthermore, to keep the notation uncluttered, let:

$$\begin{aligned}
&\underline{\boldsymbol{z}_{\mathcal{T}}}=\{\boldsymbol{z}_{\mathcal{T}^{(j)}}\}_{j=1}^{m},\\
&\underline{\boldsymbol{x}_{\mathcal{T}}}=\{\boldsymbol{x}_{\mathcal{T}^{(j)}}\}_{j=0}^{m},\\
&\underline{\mathbf{f}_{\mathcal{T}}}=\{\mathbf{f}_{\mathcal{T}^{(j)}}\}_{j=0}^{m},\quad \widetilde{\underline{\mathbf{f}}}^{(j)}=\{\mathbf{f}_{\mathcal{T}^{(j')}}\}_{j'=0}^{j-1},\forall j\geq 1,\\
&\underline{\boldsymbol{a}}=\Big\{\{\boldsymbol{a}_l^{(j)}\}_{l=1}^{|\mathcal{T}^{(j)}|}\Big\}_{j=0}^{m},\ \boldsymbol{a}^{(0)}\equiv\boldsymbol{a},\\
&\underline{\gamma}=\Big\{\{\gamma_l^{(j)}\}_{l=1}^{|\mathcal{T}^{(j)}|}\Big\}_{j=0}^{m},\ \gamma^{(0)}\equiv\gamma,\\
&\underline{\boldsymbol{b}}=\Big\{\{\boldsymbol{b}_l^{(j)}\}_{l=1}^{|\mathcal{T}^{(j)}|}\Big\}_{j=0}^{m},\ \boldsymbol{b}^{(0)}\equiv\boldsymbol{b},\\
&\underline{\boldsymbol{\theta}}=\Big\{\{\boldsymbol{\theta}_l^{(j)}\}_{l=1}^{|\mathcal{T}^{(j)}|}\Big\}_{j=0}^{m},\ \boldsymbol{\theta}_l^{(j)}=\{\boldsymbol{a}_l^{(j)},\boldsymbol{U},\boldsymbol{b}_l^{(j)},\gamma_l^{(j)}\}.
\end{aligned}$$

We first discuss the design of a fully independent model and its limitation. We then introduce the case of the conditionally independent model.

4.1 Fully Independent MRGP

Joint Distribution The joint distribution of all observations and all latent variables is expressed as

$$\begin{aligned}
&p\big(\boldsymbol{y}_{\mathcal{T}},\underline{\boldsymbol{z}_{\mathcal{T}}},\underline{\mathbf{f}_{\mathcal{T}}},\mathbf{f}_{\mathcal{T}^{(m+1)}},\underline{\boldsymbol{x}_{\mathcal{T}}},\underline{\boldsymbol{\theta}}\big)\\
&\qquad=p(\boldsymbol{y}_{\mathcal{T}}\,|\,\mathbf{f}_{\mathcal{T}_l^{(1)}},\boldsymbol{x}_{\mathcal{T}},\boldsymbol{\theta}^{(0)})p(\boldsymbol{\theta}^{(0)})\\
&\times\left[\prod_{j=1}^{m}\prod_{l=1}^{|\mathcal{T}^{(j)}|}p(\boldsymbol{z}_{\mathcal{T}_l^{(j)}}\,|\,\mathbf{f}_{\mathcal{T}_l^{(j+1)}},\widetilde{\underline{\mathbf{f}}}_l^{(j)},\boldsymbol{x}_{\mathcal{T}_l^{(j)}},\boldsymbol{\theta}_l^{(j)})p(\boldsymbol{\theta}_l^{(j)})\right]\\
&\qquad\times\left[\prod_{j=1}^{m}\prod_{l=1}^{|\mathcal{T}^{(j)}|}p(\mathbf{f}_{\mathcal{T}_l^{(j)}}\,|\,\boldsymbol{z}_{\mathcal{T}_l^{(j)}})\right]p(\mathbf{f}_{\mathcal{T}^{(m+1)}}).\quad(3)
\end{aligned}$$

The corresponding graphical representation of the model is shown in Fig 1-b, for the special case of $m=2$.

Variational Inference Using variational inference [20; 7], the goal is to find a tractable approximation of the true posterior distribution. Consider a variational posterior in the form of:

$$\begin{aligned}
q(\underline{\boldsymbol{z}_{\mathcal{T}}},\underline{\mathbf{f}_{\mathcal{T}}},\mathbf{f}_{\mathcal{T}^{(m+1)}},\underline{\boldsymbol{\theta}})&=\left[\prod_{j=0}^{m}\prod_{l=1}^{|\mathcal{T}^{(j)}|}q(\boldsymbol{\theta}_l^{(j)})\right]\\
&\times\left[\prod_{j=1}^{m}\prod_{l=1}^{|\mathcal{T}^{(j)}|}q(\boldsymbol{z}_{\mathcal{T}_l^{(j)}})q(\mathbf{f}_{\mathcal{T}_l^{(j)}}\,|\,\boldsymbol{z}_{\mathcal{T}_l^{(j)}})\right]q(\mathbf{f}_{\mathcal{T}^{(m+1)}}).\quad(4)
\end{aligned}$$

Using the mean-field assumption and choosing conjugate priors, it is possible to find tractable expressions for $q(\boldsymbol{\theta}_l^{(j)})$ and $q(\boldsymbol{z}_{\mathcal{T}_l^{(j)}})$. However, $q(\mathbf{f}_{\mathcal{T}^{(m+1)}})$ and $q(\mathbf{f}_{\mathcal{T}_l^{(j)}}\mid\boldsymbol{z}_{\mathcal{T}_l^{(j)}})$ can still be intractable. Following a similar approach as in [13] and [10], we can

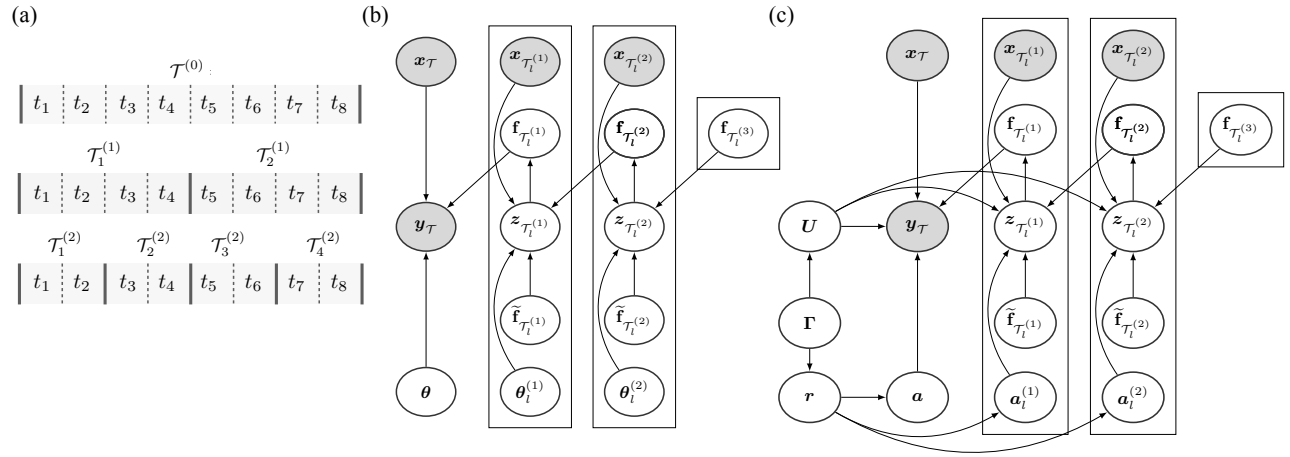

Figure 1: (a) Recursive partitioning of the index set by a factor of 2 for a model with resolution $m = 2$. (b) The graphical representation of the fully independent MRGP (fiMRGP) model using the conventional plate notation. The boxes indicate $|\mathcal{T}^{(j)}|$ replications and the arrows show the dependency between variables. (c) The graphical representation of the conditionally independent MRGP (ciMRGP) model. Note that, for better readability, we have not shown noise and bias variables as indicated in (6).

take $q(\mathbf{f}_{\mathcal{T}^{(m+1)}})$ and $q(\mathbf{f}_{\mathcal{T}_l^{(j)}} \,|\, \boldsymbol{z}_{\mathcal{T}_l^{(j)}})$ to match the prior model. These difficult-to-compute terms would then effectively cancel in the optimization when computing the Kullback-Leibler divergence between the prior and posterior. This simplifying assumption, in particular for $q(\mathbf{f}_{\mathcal{T}_l^{(j)}} \,|\, \boldsymbol{z}_{\mathcal{T}_l^{(j)}})$, makes the inference tractable but it comes with the price of severely underestimating uncertainties which ultimately causes overfitting in terms of sensitivity to the chosen resolution.

To reduce the implications of this simplification while maintaining a tractable solution, we will allow the GPs to share part of the parameter space $\underline{\boldsymbol{\theta}}$. In the following, we discuss this model alternative.

4.2 Conditionally Independent MRGP

Joint Distribution The joint distribution of all observations and all latent variables is given by

$$
\begin{aligned}
&p\big(\boldsymbol{y}_{\mathcal{T}}, \underline{\boldsymbol{z}_{\mathcal{T}}}, \underline{\mathbf{f}_{\mathcal{T}}}, \mathbf{f}_{\mathcal{T}^{(m+1)}}, \underline{\boldsymbol{x}_{\mathcal{T}}}, \boldsymbol{U}, \underline{\boldsymbol{a}}, \underline{\boldsymbol{b}}, \underline{\gamma}, \boldsymbol{\Gamma}, \boldsymbol{r}\big) \\
&\qquad = p(\boldsymbol{y}_{\mathcal{T}} \,|\, \mathbf{f}_{\mathcal{T}_l^{(1)}}, \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{\theta}^{(0)}) p(\boldsymbol{\theta}^{(0)} \,|\, \boldsymbol{\Gamma}, \boldsymbol{r}) \\
&\times \left[\prod_{j=1}^{m} \prod_{l=1}^{|\mathcal{T}^{(j)}|} p(\boldsymbol{z}_{\mathcal{T}_l^{(j)}} \,|\, \mathbf{f}_{\mathcal{T}_l^{(j+1)}}, \widetilde{\underline{\mathbf{f}}}_l^{(j)}, \boldsymbol{x}_{\mathcal{T}_l^{(j)}}, \boldsymbol{\theta}_l^{(j)}) p(\boldsymbol{\theta}_l^{(j)}, \boldsymbol{\Gamma}, \boldsymbol{r}) \right] \\
&\qquad \times \left[\prod_{j=1}^{m} \prod_{l=1}^{|\mathcal{T}^{(j)}|} p(\mathbf{f}_{\mathcal{T}_l^{(j)}} \,|\, \boldsymbol{z}_{\mathcal{T}_l^{(j)}}) \right] p(\mathbf{f}_{\mathcal{T}^{(m+1)}}), \quad (5)
\end{aligned}
$$

where the pair of $\boldsymbol{\Gamma}$ and $\boldsymbol{r}$ are hierarchical parameters which will be discussed shortly. The corresponding graphical model is shown in Fig. 1-c.

The prior model parameter in (5) is factorized as

$$
\begin{aligned}
p(\boldsymbol{\theta}_l^{(j)}, \boldsymbol{\Gamma}, \boldsymbol{r}) &= p(\boldsymbol{b}_l^{(j)} \,|\, \gamma_l^{(j)}) p(\gamma_l^{(j)}) p(\boldsymbol{U} \,|\, \boldsymbol{\Gamma}) \\
&\quad \times p(\boldsymbol{a}_l^{(j)} \,|\, \boldsymbol{r}) p(\boldsymbol{r} \,|\, \boldsymbol{\Gamma}) p(\boldsymbol{\Gamma}). \quad (6)
\end{aligned}
$$

To facilitate expressions of the conditional distributions, let $\mathfrak{Z}_k^{(j)}, \forall k \in \{j, j+1\}$, indicate a binary switch parameter such that $\mathfrak{Z}_k^{(j)} = 1$ when $k = j$ and $\mathfrak{Z}_k^{(j)} = 0$ when $k = j + 1$. The conditional distribution of the observations is expressed by

$$
\begin{aligned}
p(\boldsymbol{y}_{\mathcal{T}} \,|\, \mathbf{f}_{\mathcal{T}^{(1)}}, \boldsymbol{x}_{\mathcal{T}}, \boldsymbol{\theta}) &= \prod_{k \in \{0,1\}} \left[\prod_{l=1}^{|\mathcal{T}^{(1)}|} \prod_{t \in \mathcal{T}_l^{(1)}} p_0(\mathbf{f}_t) \right]^{1 - \mathfrak{Z}_k^{(j)}} \\
&\times \left[\prod_{t \in \mathcal{T}^{(0)}} \mathcal{N}\big(\boldsymbol{y}_t; \boldsymbol{b} + \sum_{i=1}^{p} a_i \boldsymbol{u}_i \phi_i^{(0)}(\boldsymbol{x}_t, \boldsymbol{\tau}^{(0)}), \gamma^{-1}\big) \right]^{\mathfrak{Z}_k^{(j)}},
\end{aligned}
$$

and the conditional distribution of the latent variables $\boldsymbol{z}_{\mathcal{T}_l^{(j)}}, \forall j$, is expressed by

$$
\begin{aligned}
&p(\boldsymbol{z}_{\mathcal{T}_l^{(j)}} \,|\, \mathbf{f}_{\mathcal{T}^{(j+1)}}, \widetilde{\underline{\mathbf{f}}}_l^{(j)}, \boldsymbol{x}_{\mathcal{T}_l^{(j)}}, \boldsymbol{\theta}_l^{(j)}) \\
&\qquad = \prod_{k \in \{j, j+1\}} \left[\prod_{l=1}^{|\mathcal{T}^{(j+1)}|} \prod_{t \in \mathcal{T}_l^{(j+1)}} p_0(\mathbf{f}_t) \right]^{1 - \mathfrak{Z}_k^{(j)}} \\
&\qquad\qquad \times \left[\prod_{t \in \mathcal{T}_l^{(j)}} \mathcal{N}\big(\boldsymbol{z}_{t,l}^{(j)}; \bar{\boldsymbol{z}}_{t,l}^{(j)}, {\gamma_l^{(j)}}^{-1}\big) \right]^{\mathfrak{Z}_k^{(j)}},
\end{aligned}
$$

where $\bar{\boldsymbol{z}}_{t,l}^{(j)}, \forall j \geq 1$, is defined as:

$$
\bar{\boldsymbol{z}}_{t,l}^{(j)} = \sum_{j'=0}^{j-1} \mathbf{f}_{t,l}^{j'} + \boldsymbol{b}_l^{(j)} + \sum_{i=1}^{p} a_{i,l}^{(j)} \boldsymbol{u}_i \phi_i^{(j)}(\boldsymbol{x}_t, \boldsymbol{\tau}_l^{(j)}),
$$

and $p_0(\mathbf{f}_t)$ approaches the Dirac point mass $\delta(\mathbf{f}_t)$.

Role of Hierarchical Parameters As mentioned earlier, in the expression for the joint distribution (5) we have introduced hierarchical parameters

$\boldsymbol{\Gamma} = [\Gamma_{ik}], i, k \in \{1, \ldots, p\}$ and $\boldsymbol{r} = (r_1, \ldots, r_p)^\top$, which are not explicit in the generative model, Fig. 1-b.

The parameters r_i represent the precision of the basis-axis scale parameter $a_{i,l}^{(j)}$ and are shared across resolutions and regions. These parameters will enable automatic determination of the effective number of basis axes, as the posterior will approach zero for axes that are effectively not used. Thus at each resolution and in each region, only a subset of the basis axes will be used and others will have little to no influence.

Furthermore, our recursive framework requires the indexing of the axes of $\boldsymbol{U}$ to be the same across resolutions. More precisely, we shall learn the posterior distribution over $\boldsymbol{U}$ in a Bayesian recursive fashion such that the posterior from the previous resolution is used as the prior for the current resolution. A complication is that the indexing of $\{\boldsymbol{u}_i\}_{i=1}^p$ might end up being completely arbitrary at each resolution. This is because $\boldsymbol{u}_i$ is distributed according to a Bingham distribution as $\boldsymbol{u}_i \sim \mathcal{B}(\mathbf{B}_i)$, where $\mathbf{B}_i$ is expressed via a set of eigenvectors and eigenvalues, $\mathbf{B}_i = \mathbf{M}_i \times \text{diag}[\boldsymbol{\kappa}_i] \times \mathbf{M}_i^\top$. The complication is that the indexing of these eigenvectors can be completely arbitrary, implying that the necessary one-to-one correspondence between the eigenvectors representing the prior and those representing the posterior is lost. Our sequential (recursive) learning however requires a unique one-to-one correspondence. We might consider to sort the eigenvectors (axes) based on their corresponding eigenvalues. However, that would result in sub-optimal performance.

To formally handle the axis-index ambiguity across resolutions, we have introduced a latent sparse matrix $\boldsymbol{\Gamma}$ of binary indicator variables to account for the possible index permutation between the prior and the posterior of the basis axes in transitioning from resolution $j-1$ to j. A matrix element $\Gamma_{ik} = 1$ indicates that the axis identified by index k in the posterior model of resolution $j-1$ is identical to the axis denoted by index i in the current resolution j. In defining the prior, Eq. (A.2) and Eq. (A.3), we have conditioned both $\boldsymbol{u}_i$ and $\boldsymbol{r}$ on $\boldsymbol{\Gamma}$ to ensure accumulation of "aligned prior beliefs" of these parameters across resolutions (see (6) and Fig. 1-c).

The explicit form of the prior distributions over all variables in (6) is discussed in detail in App. A.

Variational Inference Here, we consider a variational posterior in the form of:

$$q(\underline{\boldsymbol{z}}_{\mathcal{T}}, \underline{\mathbf{f}}_{\mathcal{T}}, \mathbf{f}_{\mathcal{T}^{(m+1)}}, \boldsymbol{U}, \underline{\boldsymbol{a}}, \boldsymbol{\Gamma}, \boldsymbol{r}) = \left[\prod_{j=0}^{m} \prod_{l=1}^{|\mathcal{T}^{(j)}|} q(\boldsymbol{\theta}_l^{(j)}, \boldsymbol{\Gamma}, \boldsymbol{r})\right] \times \left[\prod_{j=1}^{m} \prod_{l=1}^{|\mathcal{T}^{(j)}|} q(\boldsymbol{z}_{\mathcal{T}_l^{(j)}}) p(\mathbf{f}_{\mathcal{T}_l^{(j)}} \,|\, \boldsymbol{z}_{\mathcal{T}_l^{(j)}})\right] p(\mathbf{f}_{\mathcal{T}^{(m+1)}}),$$

where the use of a *partially factorized* mean-field approximation results in

$$q(\boldsymbol{\theta}_l^{(j)}, \boldsymbol{\Gamma}, \boldsymbol{r}) = q(\boldsymbol{b}_l^{(j)} \mid \gamma_l^{(j)}) q(\gamma_l^{(j)}) \times q(\boldsymbol{a}_l^{(j)} \,|\, \boldsymbol{U}) q(\boldsymbol{U}) q(\boldsymbol{r}) q(\boldsymbol{\Gamma}). \quad (7)$$

We then take $p(\mathbf{f}_{\mathcal{T}^{(m+1)}})$ and $p(\mathbf{f}_{\mathcal{T}^{(j)}} \mid \boldsymbol{y}_{\mathcal{T}^{(j)}})$ to match the ones in the prior model of the joint expression (5) allowing a tractable solution. Furthermore, notice the difference in factorization of the prior (6) and the posterior (7). In particular, we have considered a joint posterior over basis axes and their scales, $q(\boldsymbol{u}_i) q(a_{i,l}^{(j)} \mid \boldsymbol{u}_i)$. The joint posterior allows us to conveniently use the posterior $q(\boldsymbol{u}_i)$ as the prior in the factorized prior for the sequential (recursive) learning procedure.

Given the joint distribution and our choice of the variational posterior distribution, the variational lower bound is expressed by

$$\mathcal{L} = \mathcal{L}_{\boldsymbol{y}_{\mathcal{T}}} + \sum_{j=1}^{m} \mathcal{L}_{\boldsymbol{z}_{\mathcal{T}^{(j)}}}, \quad (8)$$

where $\mathcal{L}_{\boldsymbol{y}_{\mathcal{T}}}$ can be written as the sum of the likelihood and the negative Kullback-Leibler divergence (KLD) between the posterior and the prior,

$$\mathcal{L}_{\boldsymbol{y}_{\mathcal{T}}} = \Big\langle \log p(\boldsymbol{y}_{\mathcal{T}} \,|\, \boldsymbol{x}_{\mathcal{T}}, \mathbf{f}_{\mathcal{T}^{(1)}}, \boldsymbol{\theta}^{(0)}) \Big\rangle_{q(\boldsymbol{\theta}^{(0)}) p(\mathbf{f}_{\mathcal{T}^{(1)}})} - \left\langle \log \frac{q(\boldsymbol{\theta}^{(0)}, \boldsymbol{\Gamma}, \boldsymbol{r})}{p(\boldsymbol{\theta}^{(0)}, \boldsymbol{\Gamma}, \boldsymbol{r})} \right\rangle_{q(\boldsymbol{\theta}^{(0)}, \boldsymbol{\Gamma}, \boldsymbol{r})}.$$

The notation $\langle \cdot \rangle_{q(\cdot)}$ is used to denote the expectation with respect to its variational posterior distribution. Similarly $\mathcal{L}_{\boldsymbol{z}_{\mathcal{T}^{(j)}}}$ can be expressed as the sum of the likelihood and the negative KLD between the posterior and the prior plus the posterior entropy of the remainder term,

$$\mathcal{L}_{\boldsymbol{z}_{\mathcal{T}^{(j)}}} = \Big\langle \log p(\boldsymbol{z}_{\mathcal{T}^{(j)}} \mid \boldsymbol{x}_{\mathcal{T}^{(j)}}, \mathbf{f}_{\mathcal{T}^{(j+1)}}, \widetilde{\underline{\mathbf{f}}}^{(j)}, \boldsymbol{\theta}^{(j)}) \Big\rangle_{q(\cdot)p(\cdot)} - \left\langle \log \frac{q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Gamma}, \boldsymbol{r})}{p(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Gamma}, \boldsymbol{r})} \right\rangle_{q(\boldsymbol{\theta}^{(j)}, \boldsymbol{\Gamma}, \boldsymbol{r})} - \langle \log q(\boldsymbol{z}_{\mathcal{T}^{(j)}}) \rangle_{q(\boldsymbol{z}_{\mathcal{T}^{(j)}})},$$

where $q(\cdot)p(\cdot) := q(\boldsymbol{z}_{\mathcal{T}^{(j)}}) q(\boldsymbol{\theta}^{(j)}) p(\widetilde{\underline{\mathbf{f}}}^{(j)}) p(\mathbf{f}_{\mathcal{T}^{(j+1)}})$. Taking into account the convenient form of (8), the optimal posterior distribution can now be obtained by maximizing the lower bound using standard variational inference.

The explicit forms of the optimized variational posterior distributions are derived in App. B. Descriptive statistics of the posterior distributions are summarized in App. C. The predictive process is discussed in App. D. The optimization of the basis interval parameters is discussed in App. E. Finally, an algorithmic presentation of the model is described in App. F.

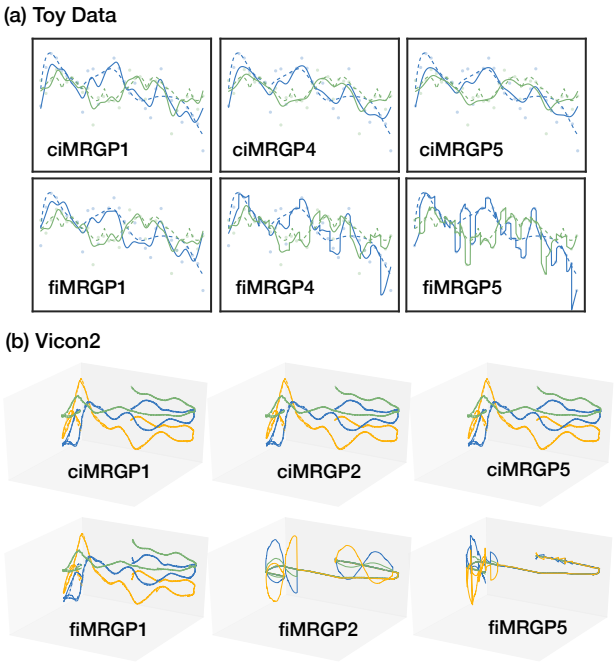

Figure 2: Illustrative comparison of ciMRGP and fiMRGP at various resolutions on (a) the synthetic dataset ToyData, App. G.2.1, and (b) the real dataset vicon2, App. G.2.2. See the text for details.

5 EXPERIMENTS

Throughout this section, we consider spectral densities of the Matérn class of covariance functions (order 1.5 and length scale 1), [30, ch. 4], and we consider eigenfunctions of the Laplace operator as the basis functions across all resolutions. Thus, for a d_x-dimensional input variable $\boldsymbol{x}_t$, we choose the basis functions, $\forall \boldsymbol{x}_t \in \boldsymbol{x}_{\mathcal{T}_l^{(j)}}$,

$$\phi_{i,l}^{(j)}(\boldsymbol{x}_t, \boldsymbol{\tau}_l^{(j)}) = \prod_{d=1}^{d_x} (1/\sqrt{\tau_{d,l}^{(j)}}) \sin(\pi i (x_{t,d} + \tau_{d,l}^{(j)})/2\tau_{d,l}^{(j)}),$$

with $\lambda_i^{(j)}(\boldsymbol{\tau}_l^{(j)}) = \sum_{d=1}^{d_x} (\pi i/2\tau_{d,l}^{(j)})^2$, $\forall i, l, j$. The number of basis functions is set to $p = \min\{n, 100\}$.

In all experiments, we compare the performance of two different multiresolution model architectures, the conditionally independent and the fully independent models, namely ciMRGP and fiMRGP. Note that fiMRGP here is obtained from ciMRGP by forcing the GPs across all resolutions to be independent (refer to Fig. 1). For simplicity, we consider uniform subdivision of the index set by a factor of $\mathfrak{q} = 2$. Finally, for instance, the notation ciMRGP4 is used to refer to ciMRGP of resolution $m = 4$.

Conditional Independence versus Full Independence We begin with an illustrative experiment which demonstrate some limitations of the full independence assumption, non-smooth boundaries and overfitting in the sense of sensitivity to the chosen resolution. For this demonstration, we compare the performance of ciMRGP and fiMRGP at various resolutions on synthetic data and real data. Figure 2-a presents a regression task of identifying (2-dimensional) latent functions from 32 noisy measurements on the ToyData dataset, App. G.2.1. The dotted lines show the ground-truth and the solid lines indicate the predictions at 10^5 test locations within the input range. At resolution $m = 1$, the two models ciMRGP1 and fiMRGP1 perform comparatively. However, with increasing resolution, these models perform very differently. In particular, notice the non-smooth boundaries in the case of fully independent model at the highest resolution, fiMRGP5, which are almost non-existing in ciMRGP5. Given that the training set includes $n = 32$ data samples, at $m = 5$ practically every single data point is a region, $|\mathcal{T}_l^{(5)}| = 1, \forall l$. Also notice that fiMRGP5 is closely following these data points, exhibiting signs of overfitting. The overfitting issue associated with fiMRGP is partly due to the unconstrained flexibility of the GPs which manifest itself at the higher resolutions where the size of the regions under consideration becomes increasingly smaller. In our experiments on real data, however, the overfitting even happened at the lower resolutions. An example on the vicon2 dataset, a subset of data recorded from a magnetic field, App. G.2.2, is shown in Fig. 2-b. The 3-dimensional noisy measurements are shown by dotted lines and the predicted strength of the magnetic fields at three different heights is estimated by each method and shown with solid lines. At $m = 1$, both models (ciMRGP1 and fiMRGP1) perform equally well, but with the increase of resolution to $m = 2$, fiMRGP2 begins to fail which worsens as the resolution is further increased, while the ciMRGP family of models remain intact and comparative at all resolutions.

Regression on Multiple Datasets We now compare the performance of various MRGP models on a number of datasets in a more structured fashion. As baselines, we include other scalable GP methods in this comparison. Key features of the datasets and models are summarized in Table 1, and they are described in more details in App. G. The performance is evaluated in terms of the root-mean-square error (RMSE) and the mean log-likelihood (MLL) on test sets, shown in Table 2 and Table 3, respectively. The model ciMRGP8 is only applied to the datasets with larger data samples. The main results are summarized as follows. In the case of ciMRGP, increasing the resolution from $m = 0$ to the higher resolutions, $m \geq 1$, resulted in noticeable improvements in terms of MLL scores. The advantage is noticeable to a lesser degree in terms of the RMSE scores. In some cases, fiMRGP showed instabilities in particular at the higher resolutions $m \geq 2$. In other cases, it only resulted in marginal improvements over

Table 1: Summary of datasets and methods used in the comparison.

Dataset							Method		
Name	Source	d_x	d_y	n_{train}	n_{test}	Note	Name	Source	Note
oes10	[35]	298	16	302	100	F.1.1	MRGP0	this paper	$m = 0$, $\mathfrak{q} = 2$
oes97	[35]	263	16	250	83	F.1.1	ciMRGP1	this paper	$m = 1$, $\mathfrak{q} = 2$
atp1d	[35]	411	6	303	33	F.1.2	ciMRGP2	this paper	$m = 2$, $\mathfrak{q} = 2$
atp7d	[35]	411	6	221	74	F.1.2	ciMRGP3	this paper	$m = 3$, $\mathfrak{q} = 2$
scm1d-a	[35]	280	16	2249	750	F.1.3	ciMRGP8	this paper	$m = 8$, $\mathfrak{q} = 2$
scm1d	[35]	280	16	7352	2450	F.1.3	fiMRGP1	this paper	$m = 1$, $\mathfrak{q} = 2$
scm20d	[35]	61	16	6724	2241	F.1.3	fiMRGP2	this paper	$m = 2$, $\mathfrak{q} = 2$
naval	[8]	16	2	8951	983	F.1.4	fiMRGP3	this paper	$m = 3$, $\mathfrak{q} = 2$
vicon	[19]	3	3	8806	8806	F.1.5	SGPMC	[18]	F.3.1
hrtf	[1]	8	200	29	8	F.1.6	SVGP	[33]	F.3.2
nengo	[5; 38]	1	7	1211	403	F.1.7	SVIGP	[17]	F.3.3
lorenz96	synthetic	1	20	1000	10^5	F.1.8			

Table 2: Average test RMSE for all methods across five repetitions.

Dataset	MRGP0	ciMRGP1	ciMRGP2	ciMRGP3	ciMRGP8	fiMRGP1	fiMRGP2	fiMRGP3	SGPMC	SVGP	SVIGP
oes10	0.784	**0.757**	**0.757**	0.758	—	0.785	0.788	0.799	0.775	0.774	0.775
oes97	0.702	0.699	0.697	**0.696**	—	0.703	0.707	0.720	0.705	0.705	0.705
atp1d	1.334	1.297	1.293	1.291	—	1.313	1.312	1.309	**1.039**	**1.039**	**1.039**
atp7d	1.228	1.231	1.229	1.232	—	1.226	1.222	1.217	**1.005**	**1.005**	1.006
scm1d-a	0.887	0.884	0.882	0.880	**0.871**	large	large	large	0.994	1.001	1.002
scm1d	1.073	1.052	1.047	1.041	1.021	large	large	large	**1.018**	**1.018**	1.021
scm20d	1.053	1.051	1.048	1.042	**0.990**	large	large	large	0.996	0.996	0.997
naval	0.009	0.006	0.005	0.005	0.005	**0.004**	0.531	large	0.011	0.011	0.019
vicon	0.019	0.018	0.018	0.018	**0.017**	0.026	large	large	0.326	0.325	0.326
hrtf	0.015	**0.014**	**0.014**	**0.014**	—	0.015	0.016	0.019	**0.014**	**0.014**	**0.014**
nengo	0.593	0.574	0.564	0.561	**0.552**	0.594	0.591	0.603	0.813	0.812	0.824
lorenz96	0.361	**0.329**	**0.329**	0.330	0.330	0.433	large	large	4.142	4.018	4.121

Table 3: Average test MLL for all methods across five repetitions.

Dataset	MRGP0	ciMRGP1	ciMRGP2	ciMRGP3	ciMRGP8	fiMRGP1	fiMRGP2	fiMRGP3	SGPMC	SVGP	SVIGP
oes10	-9.7	-4.8	-3.9	**-3.4**	—	-9.7	-9.8	-10.1	-5.6	-5.7	-10.4
oes97	-5.0	-2.9	-2.6	**-2.5**	—	-5.0	-5.1	-5.3	-4.8	-4.8	-8.7
atp1d	-19.0	-6.7	-3.9	**-2.9**	—	-18.6	-18.5	-18.5	-4.1	-4.1	-7.5
atp7d	-11.3	-4.7	-3.3	**-2.7**	—	-11.2	-11.2	-11.1	-3.9	-3.9	-7.1
scm1d-a	-56.8	-25.5	-16.9	-12.5	**-2.9**	-large	-large	-large	-8.8	-8.8	-large
scm1d	-98.2	-46.3	-30.8	-22.8	**-8.0**	-large	-large	-large	-9.2	-9.1	-large
scm20d	-92.2	-41.5	-27.9	-20.6	**-7.5**	-large	-large	-large	-8.9	-8.8	-large
naval	1.8	2.2	2.3	3.3	3.4	**3.6**	-2497.	-large	1.2	-0.9	-47.2
vicon	-0.5	-1.2	-0.9	-0.3	**-0.2**	-4.8	-large	-large	-14.5	-1.0	-large
hrtf	**0.1**	-0.0	-0.5	-0.9	—	0.0	0.0	-0.2	-0.9	-0.9	-0.3
nengo	-22.6	-7.0	-4.8	-3.6	**-2.3**	-27.0	-26.4	-25.8	-228.	-3.2	-23.4
lorenz96	-139.	-30.6	-16.1	-9.3	**-3.1**	-109.	-large	large	-large	-170.	-359.

the base model, MRGP0. In comparison to the family of sparse GP models, ciMRGP at the higher resolutions performed well in terms of RMSE, but resulted in noticeably higher MLL scores. Generally, in cases with more data samples, we found it beneficial to increase the resolution to higher values. Consider the two datasets scm1d and scm20d. We increased the resolution further to $m = 10$. The resulting models ciMRGP10 improved upon previously achieved scores reaching to RMSE and MLL scores of 0.994 and -6.4 in the case of scm1d, and 0.989 and -4.9 in the case of scm20d. This additional gain of course comes with the cost of a longer computational time which may be justifiable in certain applications and for larger datasets.

6 CONCLUSION

We have derived a multiresolution Gaussian process model which assumes *conditional independence* among the GPs across all resolutions. Relaxing the full independence assumption was shown to result in models robust to overfitting in the sense of reduced sensitivity to the chosen resolution, and predictions which are smooth at the boundaries. Although models with high resolutions may safely be used for small amounts of data, they are most relevant, and computationally justified, when there are large amounts of data. This property, combined with the favorable computational advantages of the low rank representation via the Karhunen-Loève expansion, could make the proposed model appealing for large datasets. We conclude the paper by reiterating that sharing the basis axes is an effective approach toward creating cross-talk between GPs, an approach that could be useful for learning deep GPs with conditional independence across layers.

Acknowledgements

This research is financially supported by The Knut and Alice Wallenberg Foundation (J. Taghia, contract number: KAW2014.0392), and by the Swedish Research Council (VR) via the project *NewLEADS - New Directions in Learning Dynamical Systems* (T. Schön, contract number: 621-2016-06079). We are grateful for the help and equipment provided by the UAS Technologies Lab, Artificial Intelligence and Integrated Computer Systems Division (AIICS) at the Department of Computer and Information Science (IDA), Linköping University, Sweden. The real data set used in this paper has been collected by Arno Solin, Niklas Wahlström, Manon Kok, and Simo Särkkä. We thank them for allowing us to use this data. We also thank Arne Leijon, Andreas Svensson, and Niklas Wahlström for useful feedback on early versions of this paper.

References

[1] R. V. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF Database. In *WASSAP*, 2001.

[2] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil. Fast direct methods for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):252–265, Feb. 2016.

[3] S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society. Series B (Methodological)*, 70(4): 825–848, 2008.

[4] M. Bebendorf. Low-rank approximation of elliptic boundary value problems with high-contrast coefficients. *SIAM Journal on Mathematical Analysis*, 48(2):932–949, 2016.

[5] T. Bekolay, J. Bergstra, E. Hunsberger, T. DeWolf, T. Stewart, D. Rasmussen, X. Choo, A. Voelker, and C. Eliasmith. Nengo: a Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, 7(1), 2014.

[6] C. Bingham. An antipodally symmetric distribution on the sphere. *Annals of Statistics*, 2(6): 1201–1225, 1974.

[7] D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017.

[8] A. Coraddu, L. Oneto, A. Ghio, S. Savio, D. Anguita, and M. Figari. Machine learning approaches for improving condition-based maintenance of Naval propulsion plants. *Journal of Engineering for the Maritime Environment*, 230(1), 2014.

[9] N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society. Series B (Methodological)*, 70(1):209–226, 2008.

[10] A. C. Damianou and N. D. Lawrence. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics(AISTATS)*, 2013.

[11] Y. Ding, R. Kondor, and J. Eskreis-Winkler. Multiresolution kernel approximation for Gaussian process regression. In *Advances in Neural Information Processing Systems (NIPS)*. 2017.

[12] E. B. Fox and D. B. Dunson. Multiresolution Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[13] R. Frigola, Y. Chen, and C. E. Rasmussen. Variational Gaussian process state-space models. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[14] R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, 2006.

[15] R. B. Gramacy and H. K. H. Lee. Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, 103(483):1119–1130, 2008.

[16] W. Hackbusch and B. N. Khoromskij. A sparse h-matrix arithmetic. part II: Application to multidimensional problems. *Computing*, 64(1):21–47, 2000.

[17] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intellegence (UAI)*, 2013.

[18] J. Hensman, A. G. Matthews, M. Filippone, and Z. Ghahramani. MCMC for variationally sparse Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

[19] C. Jidling, N. Wahlström, A. Wills, and T. B. Schön. Linearly constrained Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*. 2017.

[20] M. I. Jordan, Z. Ghahramani, and L. K. Jaakkola, T. S. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233, 1999.

[21] M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214, 2017.

[22] M. Katzfuss and W. Gong. Bmulti-resolution approximations of Gaussian processes for large spatial datasets. *arXiv:1710.08976*, 2017.

[23] D. Kelker and C. W. Langenberg. A mathematical model for orientation data from macroscopic conical folds. *Journal of the International Association for Mathematical Geology*, 14(4):289–307, 1982.

[24] J. T. Kent. The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(2):285–299, 1994.

[25] R. Leu and P. Damien. Bayesian shape analysis of the complex Bingham distribution. *Journal of Statistical Planning and Inference*, 149:183–200, 2014.

[26] K. V. Mardia and P. E. Jupp. *Directional Statistics*. John Wiley & Sons, 2009.

[27] A. C. Micheas, D. K. Dey, and K. V. Mardia. Complex elliptical distributions with application to shape analysis. *Journal of Statistical Planning and Inference*, 136(9):2961–2982, 2006.

[28] D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and D. Sain. A multi-resolution Gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015.

[29] J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[30] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. New York, NY, USA, 2006.

[31] H. Sang and J. Z. Huang. A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society. Series B (Methodological)*, 74(1):111–132, 2012.

[32] A. Schwaighofer and V. Tresp. Transductive and inductive methods for approximate Gaussian process regression. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[33] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems (NIPS)*. 2006.

[34] A. Solin and S. Särkkä. Hilbert space methods for reduced-rank Gaussian process regression. *arXiv:1401.5508*, 2014.

[35] E. Spyromitros-Xioufis, G. Tsoumakas, W. Groves, and I. Vlahavas. Multi-target regression via input space expansion: treating targets as inputs. *Machine Learning*, 104(1):55–98, 2016.

[36] M. L. Stein. Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8(1):1–19, 2014.

[37] A. Svensson and T. B. Schön. A flexible state-space model for learning nonlinear dynamical systems. *Automatica*, 80:189–199, 2017.

[38] J. Taghia, W. Cai, S. Ryali, J. Kochalka, J. Nicholas, T. Chen, and V. Menon. Uncovering hidden brain state dynamics that regulate performance and decision-making during cognition. *Nature Communications*, 9(2505), 2018.