Conditionally Independent
Multiresolution Gaussian Processes

Jalil Taghia
Systems and Control Division
Department of Information Technology, Uppsala University
Uppsala, Sweden
jalil.taghia@it.uu.se

Thomas B. Schön
Systems and Control Division
Department of Information Technology, Uppsala University
Uppsala, Sweden
thomas.schon@it.uu.se

Abstract

We propose a multiresolution Gaussian process (GP) model which assumes conditional independence among GPs across resolutions. The model is built on the hierarchical application of predictive processes using a particular representation of the GP via the Karhunen-Loève expansion with a Bingham prior model, where each basis vector of the expansion consists of an axis and a scale factor, referred to as the basis axis and the basis-axis scale. The basis axes have unique characteristics: They are zero-mean by construction and live on the unit sphere. These properties allow us to further assume that the axes are shared across all resolutions while their scales remain resolution specific. The properties of the Bingham distribution make it the natural choice when it comes to modeling the axes. We drive a fully Bayesian inference for the model using a structured variational inference with a partially factorized mean-field approximation which learns a joint Gaussian-Bingham posterior distribution over the basis-axis scales and the basis axes. Relaxing the full independence assumption enables the construction of models which are robust to overfitting in the sense of sensitivity to the chosen resolution and predictions that are smooth at the boundaries. Our new model and inference algorithm are compared against current state of the art on 2 synthetic and 9 real-world datasets.

1 Introduction

There is rich literature on methods designed to avoid the computational bottleneck incurred by the vanilla Gaussian process (GP) solution, including sub-sampling [33], low rank approximations [9], covariance tapering [14], inducing variables [32,35], predictive processes [3], and multiresolution models [34,31], to name just a few. Here, we focus mainly on the low rank approximations.

A GP typically assume certain smoothness properties that can undesirably soften abrupt local changes of interest. 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 traditional GPs could further get pronounced in approximate GP methods in general and rank-reduced approximations in particular [39]. A way to overcome the limitations of low rank approximations is to recognize that the long-range dependencies tend to be of lower rank than short-
range dependencies. This idea has been explored in the context of hierarchical matrices [16; 4] and in multiresolution models [34; 31; 22].

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; 32]. Application of such models has recently been exploited in spatial statistics [34; 31; 22] for modeling large spatial datasets. Refer to [12] and [22] for an overview of each category.

Current multiresolution models based on the predictive processes, although effective in terms of computational complexity, assume full independence among resolutions. The independence assumption however results in models which are inherently susceptible to the chosen resolution and approximations which are non-smooth at the boundaries. The latter one stems from the fact that the multiresolution framework, e.g., [22], recursively splits each region at each resolution into a set of subregions. As discussed by Katzfuss and Gong [23], since the remainder process is assumed to be independent between these subregions, this can lead to discontinuities at the region boundaries. A heuristic solution based on tapering functions is proposed in [23] which uses Kanter’s function as the modulating function to address this limitation. The former one, 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 as 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 naturally addressed by allowing uncertainty to propagate between resolutions, specifically, 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 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 has 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 learnt successively in a Bayesian recursive fashion. We consider a fully Bayesian treatment for the proposed model and derive a structured variational inference based on the partially factorized mean-field approximation.

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 data analysis.

Our main contribution is to develop the conditionally independent multiresolution GP model and to derive a variational inference algorithm to learn this model from data. The Bingham distribution [6] is an important distribution in directional and axial statistics commonly used for shape analysis where the inference is typically based on MLE [25], MAP [30], and MCMC [27]. Hence, our use of the Bingham prior model and the corresponding variational inference solution for this prior model could also appeal to the researchers in directional statistics [29].

Finally, this work mainly describes the theoretical derivations in general settings independent of the choice of the basis functions and the spectral densities of the covariance functions used in the series expansion. Details of derivations and experiments are presented in the supplementary material[17]. The resulting model in its standard form can be useful for analysis of large data. It can also be directly used as a building block in more complex models, for example, nonlinear state-space models [40; 13] and deep Gaussian processes [10]. We illustrate key advantages of the model on both synthetic data and real datasets.

1 An implementation of the model will be made available via GitHub.
2 Karhunen-Loève representation of the Gaussian process

Consider a minimalistic model of GP regression, \( y_t = f(x_t) + b + e_t, \forall t \in T = \{1, \ldots, n\} \), where \( f \sim \mathcal{GP}(\cdot) \) denotes a zero-mean GP prior, \( b \) denotes the overall bias, \( e_t \sim \mathcal{N}(0, \gamma^{-1}I) \) denotes Gaussian noise with zero mean and variance \( \gamma^{-1} \), \( x_t \in \mathbb{R}^{d_x} \) denotes input variables, and \( y_t \in \mathbb{R}^{d_y} \) denotes the measurements, \( d_x, d_y \in \mathbb{N}_{\geq 1} \). 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 Karhunen-Loève expansion theorem.

**Gaussian model.** For a \( d_x \)-dimensional input variable \( x_t \) on the interval \([-L_1, L_1] \times \ldots \times [-L_{d_x}, L_{d_x}] \subset \mathbb{R}^{d_x} \), the GP can be represented using the Karhunen-Loève expansion according to \([27]\), \( f(x_t) \approx \sum_{p=1}^{p} w_p \phi_p(x_t, \tau), \forall w_p \sim \mathcal{N}(0, S(\sqrt{\lambda_p(\tau)})I) \), where \( w_i = (w_{i1}, \ldots, w_{id_i})^\top \) denotes the basis vectors of the series expansion, \( \tau = (\tau_1, \ldots, \tau_{d_\text{d}})^\top \) denotes the basis intervals such that \( \tau_d > L_d, \forall d \in \{1, \ldots, d_x\} \), \( \phi_p(x_t, \tau) \) denotes the orthogonal eigenfunctions (basis functions) with corresponding eigenvalues \( \lambda_p(\tau) \), and \( S(\cdot) \) denotes the spectral density of the covariance function. Note that, unlike the minimalistic representation used by Solin and Särkkä \([37]\), we have explicitly included the basis interval \( \tau \) in the representation, which are treated as random variables and are optimized using maximum likelihood estimation.

To ensure that the representation satisfies the dual orthogonality requirement of the Karhunen-Loève expansion, all the basis vectors \( w_i \) must be zero-mean. Normally, we would assign a zero-mean Gaussian distribution over \( w_i \), or alternatively one could assign a zero-mean matrix-normal distribution over \( W = (w_1, \ldots, w_p) \) as was done by Svensson and Schön \([40]\). The choice of zero-mean Gaussian priors over the basis vectors would lead to Gaussian posteriors with non-zero means. In our multisresolution model, as we shall see later in Sec.\([3]\) the basis vector posterior needs to be learnt 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, we would then have had a multisresolution model where all GPs must be fully independent.

To address this issue, we now break the basis vectors into two parts: *basis axes* and *basis-axis scales*. The basis axes vectors are defined to be *antipodally symmetric* 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 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 \([29]\) make for a perfect fit in modeling these axes. In the following we consider a particular choice of prior model, the *Bingham prior model*, which conveniently allows for the design of a conditionally independent multisresolution model.

**Bingham model.** Let \( S^{d-1} = \{z \in \mathbb{R}^d : z^\top z = 1, d \in \mathbb{N}_{\geq 1}\} \) denote the unit sphere. Furthermore, let \( u_t := a_t u, \) such that \( u_t = \frac{u_t}{||u_t||} \in S^{d_x-1} \) and \( a_t = ||u_t|| \) denote the basis axes and the basis-axis scales, respectively. Without loss of generality, we can now express the noisy measurements as

\[
y_t = \sum_{i=1}^{p} a_i u_i \phi_i(x_t, \tau) + b + e_t, \quad \forall t \in T.
\]

(1)

The basis axes \( U = (u_1, \ldots, u_p) \) are modeled as *Bingham distributions* \([6]\) according to \( p(U) = \prod_{i=1}^{p} B(u_i ; B_i) \), where \( B(B_i) \) denotes the Bingham distribution parameterized with a real-symmetric matrix \( B_i \). It is straightforward to show that \( u_t \) satisfies the Karhunen-Loève expansion requirements. Importantly, the Bingham distribution is antipodally symmetric, meaning that \( p(u_t) = p(-u_t) \), which in turn implies that \( E[u_t] = 0 \) by construction \([29]\) Ch. 9.4. We can then assign zero-mean Gaussian distributions as priors over the basis-axis scale variables \( \{a_t\}_{t=1}^{p} \). Assuming \( e_t \sim \mathcal{N}(0, \gamma^{-1}I) \), and using \( ||u_t|| = 1 \), this choice of prior over \( u_t \) and \( a_t \) is conveniently conjugate to the data likelihood.

The main limitation of our choice of using the Bingham prior model is the implicit requirement of \( d_y > 1 \), as the Bingham density is defined on \( S^{d_y-1} \). Other prior models should be considered for the special case of \( d_y = 1 \). One possible choice is constituted by the one-parameter variant of the
The conditional distribution of the observations is ex-

sively in the generative model. As before, let \( f(\cdot) \) be the stochastic process of interest. Once the process is observed at \( x_{\tau} \), it gives rise to the noisy observations \( y_{\tau} \). By making use of a Gaussian process as the prior over \( f(\cdot) \), the observations \( y_{\tau} \) at resolution \( j = 0 \) are modeled according to \( f(\cdot) \). In a multiresolution setting based on the hierarchical application of predictive processes, we approximate \( f(\cdot) \) according to \( f(\cdot) = \hat{f}(0)(\cdot) + \hat{f}(1)(\cdot) \), where \( \hat{f}(0) \) is the approximate predictive process at resolution \( j = 0 \), and \( \hat{f}(1)(\cdot) \) is the so-called remainder process. Let \( z_{1,l}^{(1)} \) indicate the noisy instantiations of the latent process \( f(1)(\cdot) \) at \( x_{\tau} \). We will treat \( z_{1,l}^{(1)} \) as a latent variable, and model it using a conditionally independent GP prior, \( z_{1,l}^{(1)} = \sum_{i=1}^{p} a_{i,l}^{(1)} u_{i}(x_{l}, \tau_{i}^{(1)}) + \epsilon_{l}^{(1)} \), \( \forall x_{l} \in x_{\tau}^{(1)} \), where the basis axes \( u_{i} \) 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 \( f(1)(\cdot) \) is in turn approximated by \( f(1)(\cdot) = \hat{f}(1)(\cdot) + \hat{f}(2)(\cdot) \). In general, for resolution \( j \) we have \( f(j)(\cdot) = \hat{f}(j)(\cdot) + \hat{f}(j+1)(\cdot) \), where

\begin{align*}
T^{(0)} : & \begin{bmatrix} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 & t_8 \\ 1 \end{bmatrix} \\
T^{(1)} : & \begin{bmatrix} t_1^{(1)} & t_2^{(1)} & t_3^{(1)} & t_4^{(1)} & t_5^{(1)} & t_6^{(1)} & t_7^{(1)} & t_8^{(1)} \\ 1 \end{bmatrix} \\
T^{(2)} : & \begin{bmatrix} t_1^{(2)} & t_2^{(2)} & t_3^{(2)} & t_4^{(2)} & t_5^{(2)} & t_6^{(2)} & t_7^{(2)} & t_8^{(2)} \\ 1 \end{bmatrix}
\end{align*}

Figure 1: (a) Recursive partitioning of the index set by a factor of 2 for a model with resolution \( m = 2 \). (b) Graphical representation of the proposed model with resolution \( m = 2 \), using the conventional plate notation. The boxes indicate \( |T^{(j)}| \) replications and the arrows show the dependency between variables. Note that, for better readability, we have only shown noise and bias variables explicitly for \( j = 2 \). A similar structure should be assumed for other resolutions, as indicated in (2) and (3).

Bingham model \cite{24} for modeling axes concentrated asymmetrically near a small circle. Here, we restrict ourselves 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 \( q \), such that \( \mathcal{T} \) is first partitioned into \( q \) regions, each of which is then partitioned into \( q \) subregions. The partitioning continues until resolution \( m \) where the index sets at various resolution are denoted by

\[ \mathcal{T}^{(0)} := \mathcal{T}, \quad \mathcal{T}^{(1)} = \{T^{(0)}_1, \ldots, T^{(0)}_q\}, \quad \cdots, \quad \mathcal{T}^{(m)} = \{T^{(m-1)}_1, \ldots, T^{(m-1)}_q\}, \]

where \( |\mathcal{T}^{(0)}| = 1, |\mathcal{T}^{(1)}| = q, \) and \( |\mathcal{T}^{(m)}| = q^m \). An example of such partitioning by a factor of \( q = 2 \) is shown in Fig. 1(a). As a convention, we will use the notation \( T^{(j)}_l \) to indicate the \( l \)-th element of the set \( \mathcal{T}^{(j)} = \{T^{(j)}_1, \ldots, T^{(j)}_l\} \), which corresponds to the index set related to region \( l \) at resolution \( j \). We also define \( x_{\tau}^{(0)} := x_{\tau} \) and \( x_{\tau}^{(j)} = \{x_{\tau}^{(j)}\}_{l=1}^{|\mathcal{T}^{(j)}|} \), where \( x_{\tau}^{(j)} = \{x_{t} \mid \forall t \in \mathcal{T}^{(j)}_l\} \).

Generative model. As before, let \( f(\cdot) \) be the stochastic process of interest. Once the process is observed at \( x_{\tau} \), it gives rise to the noisy observations \( y_{\tau} \). By making use of a Gaussian process as the prior over \( f(\cdot) \), the observations \( y_{\tau} \) at resolution \( j = 0 \) are modeled according to \( f(\cdot) \). In a multiresolution setting based on the hierarchical application of predictive processes, we approximate \( f(\cdot) \) according to \( f(\cdot) = \hat{f}(0)(\cdot) + \hat{f}(1)(\cdot) \), where \( \hat{f}(0) \) is the approximate predictive process at resolution \( j = 0 \), and \( \hat{f}(1)(\cdot) \) is the so-called remainder process. Let \( z_{1,l}^{(1)} \) indicate the noisy instantiations of the latent process \( f(1)(\cdot) \) at \( x_{\tau} \). We will treat \( z_{1,l}^{(1)} \) as a latent variable, and model it using a conditionally independent GP prior, \( z_{1,l}^{(1)} = \sum_{i=1}^{p} a_{i,l}^{(1)} u_{i}(x_{l}, \tau_{i}^{(1)}) + \epsilon_{l}^{(1)} \), \( \forall x_{l} \in x_{\tau}^{(1)} \), where the basis axes \( u_{i} \) 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 \( f(1)(\cdot) \) is in turn approximated by \( f(1)(\cdot) = \hat{f}(1)(\cdot) + \hat{f}(2)(\cdot) \). In general, for resolution \( j \) we have \( f(j)(\cdot) = \hat{f}(j)(\cdot) + \hat{f}(j+1)(\cdot) \), where

\footnote{Note that, for \( d_{y} = 1 \), if we simply assume \( u_{i} = 1 \), the model reduces to a multiresolution architecture with fully independent GPs.}
We denote the latent function instantiations at where the pair of

The prior model parameter in (2) is factorized as

The recursive procedure continues until resolution
sponding graphical model is shown in Fig. 1-(b).
captures global patterns and finer details are captured at higher resolutions.
tive processes from all resolutions,
l may not be simply removed as a part of the preprocessing step, as the bias at each resolution carries

could assume the noise to be only a resolution-specific variable. In a multiresolution model, bias

The complete joint distribution of observations and all variables is expressed as

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 \( \hat{f}(\cdot) \) as the sum of predictive processes from all resolutions, \( \hat{f}(\cdot) = f^{(m+1)}(\cdot) + \sum_{j=0}^{m} \hat{f}^{(j)}(\cdot) \approx \sum_{j=0}^{m} \hat{f}^{(j)}(\cdot) \), where \( \hat{f}^{(0)} \) captures global patterns and finer details are captured at higher resolutions.

4 Bayesian inference

Notation. Let \( y_{T(0)} := y_T \) where \( y_T = \{ y_t \mid \forall t \in T \} \) denote the set of noisy observations, and \( z_{T(j)} = \{ z_{T(i)} \}_{i=1}^{\tau(j)} \) denote the set of latent variables for \( j \geq 1 \), where \( z_{T(j)} = \{ z_{T(i)} \mid \forall t \in T(j) \} \).

We denote the latent function instantiations at \( x_{T(j)} \) by \( f_{T(j)} = \{ f_{T(j)}(x_t) \mid \forall x_t \in x_{T(j)} \} \).

Similarly, let \( f_{T(j)} = \{ f_{T(j)}^{(l)} \}_{l=1}^{m} \). Furthermore, to keep the notation uncluttered, let:

Joint distribution. The complete joint distribution of observations and all variables is expressed as

where the pair of \( \Gamma \) and \( r \) are hierarchical parameters which shall be discussed shortly. The corresponding graphical model is shown in Fig. 1-(b).

The prior model parameter in (2) is factorized as

(3)

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

\[ p(y_T | f_{T(1)}, x_T, \theta^{(0)}) = \prod_{k \in \{0,1\}} \prod_{l=1}^{\tau(j)} p(y_l | f_{T(j)}^{(l)}, x_{T(j)}^{(l)}, \theta^{(j)}(l), \Gamma, r) \left[ \prod_{j=1}^{m} p(f_{T(j)}^{(l)} | z_{T(j)}^{(l)}) \right] p(f_{T(m+1)}, \Gamma, r) \]

\[ = \prod_{k \in \{0,1\}} \prod_{l=1}^{\tau(j)} \left[ 1 - 3_{k,j}^{(j)} \right] \times \prod_{d \in T^{(0)}} \mathcal{N}(y_d; \mathbf{b} + \sum_{i=1}^{p} a_i u_i \phi_{i}^{(0)}(x_t, T^{(0)}), \gamma^{-1}) \]
and the conditional distribution of the latent variables $z_{T_l^{(j)}}, \forall j$, is expressed by

$$
p(\mathbf{z}_{T_l^{(j)}}, \mathbf{f}_{T_l^{(j)}}, \mathbf{x}_{T_l^{(j)}}, \theta_{T_l^{(j)}}) = \prod_{k \in \{j+1\}} \prod_{l=1}^{\mid T_l^{(j)} \mid} \prod_{x_{T_l^{(j)}}} \mathcal{N}(\mathbf{z}_{T_l^{(j)}}, \gamma_{T_l^{(j)}}, 1) \mathcal{N}(\mathbf{f}_{T_l^{(j)}}, \gamma_{T_l^{(j)}}, 1),
$$

where $z_{T_l^{(j)}}, \forall j \geq 1$, is defined as: $z_{T_l^{(j)}} = \sum_{i=0}^{j-1} f_{T_l^{(j)}} + b_{T_l^{(j)}} + \sum_{i=1}^{p} a_{T_l^{(j)}} u_{T_l^{(j)}}(x_{T_l^{(j)}}, \tau_{T_l^{(j)}})$, and $p_0(f_i)$ approaches the Dirac point density $\delta(f_i)$, while the exact form of this density is not critical.

As mentioned earlier, in the expression of the joint distribution (2) we have introduced hierarchical parameters $\Gamma = [\Gamma_{ik}], i, k \in \{1, \ldots, p\}$ and $r = (r_1, \ldots, r_p)^T$, which are not explicit in the generative model, Fig. 1(b). The parameters $r_i$ represent the precision of the basis-axis scale parameter $a_{T_l^{(j)}}$ 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 at 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 axes of $U$ to be the same across resolutions. More precisely, we shall learn the posterior distribution over $U$ 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 $\{u_i\}_{i=1}^p$ might get totally arbitrary indexing at each resolution. To formally handle the axis-index ambiguity across resolutions, we have introduced a latent sparse matrix $\Gamma$ of binary indicator variables to account for the possible index permutation between prior and posterior of basis axes in transition 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 $i$ of the current resolution $j$. In defining the prior distributions, Eq. (A.1) and Eq. (A.2), we have conditioned both $u_i$ and $r$ on $\Gamma$ to ensure accumulation of “aligned prior beliefs” of these parameters across resolutions (see (3) and Fig. 1(b).

The explicit form of the prior distributions over all variables in (3) is discussed in details in App. A.

Variational inference. Using variational inference (21), the goal is to find a tractable approximation to the true posterior distribution. Here, we consider a variational posterior in the form of:

$$q(\mathbf{z}_T, \mathbf{f}_T, \mathbf{f}_{T^{(m+1)}}; \mathbf{U}, \mathbf{a}, \Gamma, r) = \prod_{j=1}^{\mid T \mid} \prod_{l=1}^{\mid T \mid} q(\theta_{T_l^{(j)}}, \Gamma, r) q(\mathbf{f}_{T_l^{(j)}}, \mathbf{x}_{T_l^{(j)}}, \theta_{T_l^{(j)}}) p(\mathbf{f}_{T^{(m+1)}}),$$

where further using a partially factorized mean-field approximation:

$$q(\theta_{T_l^{(j)}}, \Gamma, r) = q(b_{T_l^{(j)}}) q(\gamma_{T_l^{(j)}}) q(\alpha_{T_l^{(j)}}) q(u_j) q(U) q(r) q(\Gamma).$$

We then take $p(\mathbf{f}_{T^{(m+1)}})$ and $p(\mathbf{f}_{T^{(j)}}, \mathbf{y}_{T^{(j)}})$ to match the ones in the prior model of the joint expression (4) allowing a tractable solution. A similar approach was used by Frigola et al. (13) and Damianou and Lawrence (10). Furthermore, notice the difference in factorization of the prior (5) and the posterior (4). In particular, we have considered a joint posterior over basis axes and their scales, $q(u_j) q(a_{T_l^{(j)}} | u_j)$. The joint posterior allows us to conveniently use the posterior $q(u_j)$ 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

$$\mathcal{L} = \mathcal{L}_{y_T} + \sum_{j=1}^{m} \mathcal{L}_{z_{T_l^{(j)}}},$$

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

$$\mathcal{L}_{y_T} = \left\langle \log p(y_T | \mathbf{x}_T, \mathbf{f}_{T^{(1)}}, \theta^{(0)}) \right\rangle_{q(\theta^{(0)} p(\mathbf{f}_{T^{(1)}}))} - \left\langle \log \frac{q(\theta^{(0)} | \Gamma, r)}{p(\theta^{(0)} | \Gamma, r)} \right\rangle_{q(\theta^{(0)} | \Gamma, r)}.$$
The notation $\left< \cdot \right>_q$ is used to denote the expectation of its variable with respect to its variational posterior distribution. Similarly $\mathcal{L}_{z,T^{(j)}}$ 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,

$$
\mathcal{L}_{z,T^{(j)}} = \left< \log p(z_{T^{(j)}} | x_{T^{(j)}}, f_{T^{(j+1)}}) \right>_q - \left< \log \frac{q(\theta^{(j)}, \Gamma, \tau)}{p(\theta^{(j)}, \Gamma, \tau)} \right>_q + \left< \log \frac{q(U)}{p(U | \Gamma)} \right>_q + \left< \log \frac{q(r)}{p(r | \Gamma)} \right>_q,
$$

where using (3) and (4), the second term factorizes as:

$$
\left< \log \frac{q(\theta^{(j)}, \Gamma, \tau)}{p(\theta^{(j)}, \Gamma, \tau)} \right>_q = \left< \log \frac{q(b^{(j)}, \gamma^{(j)})}{p(b^{(j)}, \gamma^{(j)})} \right>_q + \left< \log \frac{q(U)}{p(U | \Gamma)} \right>_q + \left< \log \frac{q(r)}{p(r | \Gamma)} \right>_q.
$$

Taking into account the convenient form of (5), 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.

5 Experiments

Throughout of this section, we consider spectral densities of the Matérn class of covariance functions (order 1.5 and length scale 1), [33, ch. 4], and we consider eigenfunctions of the Laplace operator as the basis functions across all resolutions. Thus, for $d_x$-dimensional input variable $x_t$, we choose the basis functions, $\phi_{i,d}^{(j)}(x_t, \tau_t^{(j)}) = \prod_{d=1}^{d_x} (1/\tau_d^{(j)}) \sin(\pi i(x_{t,d} + \tau_d^{(j)})/2\tau_d^{(j)}), \forall x_t \in x_{\tau^{(j)}},$ with corresponding eigenvalues $\lambda_i^{(j)}(\tau_t^{(j)}) = \sum_{d=1}^{d_x} (\pi i/2\tau_d^{(j)})^2, \forall i, l, j.$ 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, conditionally independent and fully independent models, namely ciMRGP and iMRGP. Note that iMRGP here is obtained from ciMRGP by forcing all GPs across resolutions to be independent. For simplicity, we consider uniform subdivision of the index set by a factor of $q = 2$. 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 demonstrates 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 performance of ciMRGP and iMRGP at various resolutions on synthetic and real data. Figure 2-(a) presents a regression task of identifying (2-dimensional) latent functions from 32 noisy measurements on ToyData dataset, App. F.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 $m = 1$, both models ciMRGP1 and iMRGP1 perform comparatively. However, with increasing resolution, these two perform very differently. In particular, notice the non-smooth boundaries in the case of full independence at the highest resolution, iMRGP5, 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, $|T_l^{(5)}| = 1, \forall l$. Also notice that iMRGP5 is closely following these data points, exhibiting signs of overfitting. The overfitting issue associated with iMRGP is partly due to the unconstrained flexibility of 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. F.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 estimated by each method are shown with solid lines. At $m = 1$, both methods ciMRGP1 and iMRGP1 perform equally well but with the increase of resolution to $m = 2$, iMRGP2 begins to fail which worsens as the resolution increases, while ciMRGP family of methods remain intact and comparative at all resolutions.

Regression on multiple datasets. We now compare 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 methods are summarized in Table 1, and are described in more details in App. 8. The performance is evaluated in terms of the root-mean-square error (RMSE) and the mean log-likelihood (MLL) on test sets, shown respectively in Table 2 and Table 3. The method ciMRGP8 is only applied to the datasets with larger data samples. The main results are summarized as follows. In the case of ciMRGP, generally increasing 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, iMRGP showed instabilities in particular at the higher resolutions $m \geq 2$. In other cases, it only resulted in marginal improvements over the base model, MRGP0. In compression to sparse GP family of methods, ciMRGP at the higher resolutions performed competitively in terms of RMSE but resulted in noticeably higher MLL scores. Generally, in cases with larger data samples, we found beneficial to increase the resolution to higher values. Consider two cases of scm1d and scm20d datasets. 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 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 be safely 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 data analysis.

A Prior model

This section describes our choice of the prior model parameters, and details of their initializations.
We assign a product of zero-mean Gaussian densities conditional on the basis-axis scale-precision $S$-covariance functions at different resolutions. Similarly, there are various choices of basis functions.

| Dataset      | Source | $d_x$ | $d_y$ | $n_{\text{train}}$ | $n_{\text{test}}$ | Note | Method      | Source | Note |
|--------------|--------|-------|-------|---------------------|-------------------|------|------------|--------|------|
| oes10        | [38]   | 298   | 16    | 302                 | 100               | E.11 | MRGP0      | this paper | $m = 0, q = 2$ |
| oes97        | [38]   | 263   | 16    | 250                 | 83                | E.11 | ciMRGP1    | this paper | $m = 1, q = 2$ |
| atp1d        | [38]   | 411   | 6     | 303                 | 33                | E.11 | ciMRGP2    | this paper | $m = 2, q = 2$ |
| atp7d        | [38]   | 411   | 6     | 221                 | 74                | E.11 | ciMRGP3    | this paper | $m = 3, q = 2$ |
| scml1d-a     | [38]   | 828   | 16    | 2499                | 750               | E.11 | ciMRGP8    | this paper | $m = 8, q = 2$ |
| scml1d       | [38]   | 280   | 16    | 7352                | 2450              | E.11 | iMRGP1     | this paper | $m = 1, q = 2$ |
| scml20d      | [38]   | 61    | 16    | 6724                | 2241              | E.11 | iMRGP2     | this paper | $m = 2, q = 2$ |
| naval        | [8]    | 16    | 2     | 8951               | 983               | E.11 | iMRGP3     | this paper | $m = 3, q = 2$ |
| vicon        | [20]   | 3     | 3     | 8806               | 8806              | E.11 | SGPMC      | 19       | E.3.1 |
| hrtf         | [11]   | 8     | 200   | 29                 | 8                 | E.11 | SVGP       | 36       | E.3.2 |
| nengo        | [5]    | 1     | 7     | 1211               | 403               | E.11 | SVIGP      | 17       | E.3.5 |
| lorenz96     | synthetic | 1  | 20    | 1000               | $10^5$            | E.13 |           |          |      |

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

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

| Dataset      | Source | $d_x$ | $d_y$ | $n_{\text{train}}$ | $n_{\text{test}}$ | Note | Method      | Source | Note |
|--------------|--------|-------|-------|---------------------|-------------------|------|------------|--------|------|
| oes10        | [38]   | 0.784 | 0.757 | 0.758               | 0.758             | 0.785 | 0.788      | 0.799 | 0.775 | 0.774 | 0.775 |
| oes97        | [38]   | 0.702 | 0.699 | 0.697               | 0.696             | 0.703 | 0.707      | 0.720 | 0.705 | 0.705 | 0.705 |
| atp1d        | [38]   | 1.334 | 1.297 | 1.293               | 1.291             | 1.313 | 1.312      | 1.309 | 1.039 | 1.039 | 1.039 |
| atp7d        | [38]   | 1.228 | 1.231 | 1.229               | 1.232             | 1.226 | 1.222      | 1.217 | 1.005 | 1.005 | 1.006 |
| scml1d-a     | [38]   | 0.887 | 0.884 | 0.882               | 0.880             | 0.871 | large      | large | large | 0.994 | 1.001 | 1.002 |
| scml1d       | [38]   | 1.073 | 1.052 | 1.047               | 1.041             | 1.021 | large      | large | large | 1.018 | 1.018 | 1.021 |
| scml20d      | [38]   | 1.053 | 1.051 | 1.048               | 1.042             | 0.990 | large      | large | large | 0.996 | 0.996 | 0.997 |
| naval        | [38]   | 0.009 | 0.006 | 0.005               | 0.005             | 0.005 | 0.004      | 0.531 | 0.011 | 0.011 | 0.019 |
| vicon        | [38]   | 0.019 | 0.018 | 0.018               | 0.018             | 0.017 | 0.026      | 0.326 | 0.325 | 0.326 | 0.326 |
| hrtf         | [38]   | 0.015 | 0.014 | 0.014               | 0.014             | 0.015 | 0.016      | 0.019 | 0.014 | 0.014 | 0.014 |
| nengo        | [38]   | 0.593 | 0.574 | 0.564               | 0.561             | 0.552 | 0.594      | 0.591 | 0.603 | 0.813 | 0.812 | 0.824 |
| lorenz96     | [38]   | 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.

A.1 Prior over basis-axis scales

We assign a product of zero-mean Gaussian densities conditional on the basis-axis scale-precision variables as the prior over basis-axis scales,

$$p(a | r) = \prod_{j=0}^{m} \prod_{l=1}^{q} \prod_{i=1}^{p} \mathcal{N}(a_{j,l}^{(i)}; 0, r_{l}^{-1} S^{(j)}(\sqrt{\lambda_{l}^{(j)}(r_{l}^{(j)})^{-1}})),$$

where $S^{(j)}(\cdot)$ is the spectral density of the covariance function and $\lambda_{l}^{(j)}(r_{l}^{(j)})$ is the eigenvalue of the basis function $\phi_{l}^{(j)}(\cdot)$ at resolution $j$. There are various choices of covariance functions. Among them, we are interested in those for which $S(\nu) \to 0$ for all $\nu \to \infty$, that is the case for most classes of covariance functions, including Matérn and exponentiated quadratic covariance functions. We have indicated spectral densities with indexing on $j$, as in general, we are free to choose different covariance functions at different resolutions. Similarly, there are various choices of basis functions
which are interpretable as GPs. As discussed in the paper, the choice of basis functions can in general be resolution-specific.

Our choice of prior implies that $\alpha_{i,l}^{(j)}$ are resolution-region specific, which means that regardless of the resolution or the region the prior must be initialized with zero-mean even though the posterior mean is non-zero.

### A.2 Prior over basis axes

Considering the possible index permutation across resolutions, we assign a product of independent Bingham densities \[^6, 29\], conditional on the binary index-mapping matrix $\mathbf{Z}$, as the prior over basis axes

$$ p(U \mid \Gamma) = \prod_{i=1}^{p} \prod_{k=1}^{p} \left[ B(u_i; \mathbf{B}_k') \right]_{\Gamma, ik}^{\Gamma, ik} = \prod_{i=1}^{p} \prod_{k=1}^{p} \left[ \frac{1}{\mathcal{C}(\kappa_k')} \exp \left( u_i^\top \mathbf{B}_k' u_i \right) \right]_{\Gamma, ik}^{\Gamma, ik}. \quad (A.1) $$

Here, $\mathbf{B}_k' = \mathbf{M}_k' \times \text{diag}[\kappa_k'] \times \mathbf{M}_k'^\top$, and the pair of $\mathbf{M}_k' = (\mu_k', \ldots, \mu_{kd_y}')$, $\mu_k' \in \mathbb{S}^{d_y-1}$, $\kappa_k' = (\kappa_{k1}', \ldots, \kappa_{kd_y}')^\top$ are given by the eigendecomposition of $\mathbf{B}_k'$ and $\mathcal{C}(\kappa_k')$ is the Bingham normalization factor, which is algebraically problematic, but the saddle-point approximation \[^{26}\] provides an accurate numerical result.

Notice that, at resolution $j > 0$, $\mathbf{B}_k'$ is given by the posterior hyper-parameter from the previous resolution $j - 1$. At resolution $j = 0$, we set simply $\mathbf{B}_k' = \mathbf{0} = 0$.

### A.3 Prior over basis-axis scale-precision

Considering the possible index permutation across resolutions, we express the prior over precision of the basis scales as conditional on the binary index-mapping matrix $\Gamma$ using Gamma densities

$$ p(r \mid \Gamma) = \prod_{i=1}^{p} \prod_{k=1}^{p} \left[ \mathcal{G}(r_i; \alpha_k', \beta_k') \right]_{\Gamma, ik}^{\Gamma, ik}, \quad (A.2) $$

where $\alpha_k'$ and $\beta_k'$ are the Gamma densities shape and inverse scale hyper-parameters. At resolution $j > 0$, $\alpha_k'$ and $\beta_k'$ are the posterior hyper-parameters computed from resolution $j - 1$. At $j = 0$, in absence of prior data, non-informative distributions may be assigned with $\alpha_k' \to 0$, but $\beta_k'$ may still be assigned an informative value. Values of $\beta_k'$ for which $\alpha_k' / \beta_k' \to 0$ reduces the overall influence of the prior toward a non-regularized basis function expansion.

### A.4 Prior over basis-axis index mapping

As discussed earlier, the index-mapping binary matrix $\Gamma$ has exactly one element $\Gamma_{ik} = 1$ in each row and each column, indicating that the basis axis identified by index $k$ in the previous resolution $j - 1$ is identical to the basis axis denoted by index $i$ at the current resolution $j$. The prior probability mass for these index-mapping variables is assigned as totally non-informative, except for the uniqueness requirement

$$ \sum_{k=1}^{p} p(\Gamma_{ik} = 1) = 1, \quad \forall i \in \{1, \ldots, p\}, \quad (A.3a) $$

$$ \sum_{i=1}^{p} p(\Gamma_{ik} = 1) = 1, \quad \forall k \in \{1, \ldots, p\}. \quad (A.3b) $$

### A.5 Prior over overall bias and residual noise precision

We assign product of Gaussian-Gamma densities over the joint distribution of the overall bias and the residual noise precision as

$$ p(b, \gamma) = \prod_{j=0}^{m} \prod_{l=1}^{[T^{(j)}]} \mathcal{N}(b_{l}^{(j)}; \nu_{0l}^{(j)}, \frac{1}{\eta_{0l}^{(j)} \gamma_{l}^{(j)}}) \mathcal{G}(\gamma_{l}^{(j)}; \upsilon_{0l}^{(j)}, \delta_{0l}^{(j)}). $$
In the absence of prior information, a non-informative prior must be applied by setting \( \nu_{o_l}^{(j)} = 0 \) and \( \beta_{o_l}^{(j)} \to 0 \). The hyper-parameters \( \epsilon_{o_l}^{(j)} \) and \( \delta_{o_l}^{(j)} \) are shape and inverse scale parameters of the corresponding Gamma distributions. In the absence of prior information, a noninformative distribution is assigned by \( \epsilon_{o_l}^{(j)} \to 0 \), but \( \delta_{o_l}^{(j)} \) may still be assigned an informative value to indicate the most likely value (mode), \( \delta_{o_l}^{(j)}/(\epsilon_{o_l}^{(j)} + 1) \), for the residual variance which has an inverse-gamma distribution.

### B Posterior model

In this section, we summarize the optimized posterior distribution which is obtained by maximizing the lower bound \( \mathcal{L} \) in [5]. For ease of notation, we use: \( \langle \cdot \rangle_{q(\cdot)} \equiv \langle \cdot \rangle \) wherever possible. Descriptive statistics of the posterior distributions are summarized in Appendix C.

#### B.1 Conditional posterior over basis-axis scales

Optimized conditional posterior distribution of \( q(a_i | U) \) is given by the following product of Gaussian densities

\[
q(a_i | U) = \prod_{j=0}^{m} \prod_{l=1}^{p} \prod_{i=1}^{|\mathcal{T}^{(j)}|} \mathcal{N}(a_{i,l}^{(j)}; m_{i,l}^{(j)}(u_i), v_{i,l}^{(j)})
\]

with the mean value \( m_{i,l}^{(j)}(u_i) \) as the function of the basis axis vector \( u_i \) and the precision \( v_{i,l}^{(j)} \) given by

\[
v_{i,l}^{(j)} = \frac{\langle v_i^{(j)} \rangle}{S^{(j)}\left(\sqrt{\lambda_i^{(j)}(\hat{T}_i^{(j)})^2}\right)} + \langle \gamma_i^{(j)} \rangle \sum_{t \in \mathcal{T}_i^{(j)}} \left( \phi_i^{(j)}(x_t, \hat{T}_i^{(j)}) \right)^2,
\]

\[
m_{i,l}^{(j)}(u_i) = s_{i,l}^{(j)} u_i^\top z_{i,l}^{(j)},
\]

where we have defined

\[
s_{i,l}^{(j)} = \frac{\langle \gamma_i^{(j)} \rangle}{\langle v_i^{(j)} \rangle},
\]

\[
z_{i,l}^{(j)} = \begin{cases} y_i - \sum_{t \in \mathcal{T}_i^{(j)}} \phi_i^{(j)}(x_t, \hat{T}_i^{(j)}) \left( y_t - \tilde{y}_i \right), & \forall j = 0 \\ \sum_{t \in \mathcal{T}_i^{(j)}} \phi_i^{(j)}(x_t, \hat{T}_i^{(j)}) \left( z_t^{(j)} - \tilde{z}_{i,t,l}^{(j)} \right), & \forall j \geq 1 \end{cases},
\]

\[
\tilde{y}_i = \langle b \rangle + \sum_{k \neq i} \langle a_k u_k \rangle \phi_k^{(0)}(x_i, \hat{T}_i^{(0)}),
\]

\[
\tilde{z}_{i,t,l}^{(j)} = \left( \sum_{j' = 0}^{j-1} \tilde{z}_{i,t,l}^{(j')} \right) + \langle b_{i,l}^{(j)} \rangle + \sum_{k \neq i} \langle a_{k,i} u_k \rangle \phi_k^{(j)}(x_i, \hat{T}_i^{(j)}).
\]

The seemingly complicated form of this result makes intuitively good sense: The conditional expected value of the basis-axis scale variables, given by \( m_{i,l}^{(j)}(u_i) \), is determined by the mean predictions from the previous resolution plus the remaining part of the observed (or latent for \( j > 0 \)) vector that is not already explained by its components along the other basis axes. The conditional expected value is scaled by \( s_{i,l}^{(j)} \), which is the currently estimated proportion of the variance of the observed data (or latent variables for \( j > 0 \)) that is explained by the basis-axis scale variables in the \( i \)th axis, in relation to the total variance that also includes the residual noise component along this axis.
Where the probability parameters are normalized using scale factors \( \eta \).

Where as before the pair of \( \kappa_i \) and \( M_i \) are eigenvalues and the corresponding eigenvectors of

\[
B_i = \left( \sum_{k=1}^{p} (\Gamma_{ik}) B_k^i \right) + \sum_{i=1}^{\left| T^{(j)} \right|} \frac{\gamma^{(j)}}{2} \dot{\zeta}_{i,l}^{(j)} \dot{z}_{i,l}^{(j)} \dot{z}_{i,l}^{(j)}
\]

Note the first term where Bingham’s posterior hyper-parameter from the previous resolution, \( B_k^i \), has been weighted by \( \langle \Gamma_{ik} \rangle \). This ensures that the axis indices remain aligned throughout and hence allows for recursive (successive) learning of these parameters.

### B.3 Posterior over basis-scale precision

The optimized posterior distribution of the latent variables \( r \) is given by the product of Gamma densities as

\[
q(r) = \prod_{i=1}^{p} \mathcal{G}(r_i; \alpha_i, \beta_i),
\]

where \( \alpha_i \) and \( \beta_i \) are the shape and inverse scale posterior hyper-parameters of Gamma density given by

\[
\alpha_i = \left( \sum_{k=1}^{p} (\Gamma_{ik}) \alpha_k^i \right) + \frac{\left| T^{(j)} \right|}{2},
\]

\[
\beta_i = \left( \sum_{k=1}^{p} (\Gamma_{ik}) \beta_k^i \right) + \frac{1}{2} \sum_{l=1}^{\left| T^{(j)} \right|} \frac{\langle \alpha_{i,l}^{(j)} \rangle^{2}}{S^{(j)} \langle \dot{\alpha}_{i,l}^{(j)} \rangle \langle \dot{\tau}_{i,l}^{(j)} \rangle},
\]

where \( \alpha_k^i \) and \( \beta_k^i \) are posterior hyper-parameters from the previous resolution, \( k \), weighted by the posterior mean of the basis-axis index-mapping variable, \( \langle \Gamma_{ik} \rangle \).

### B.4 Posterior over basis-axis index mapping

The optimized posterior distribution of the latent variables \( q(\Gamma) \) is given by \( q(\Gamma) = \prod_{i=1}^{p} \prod_{k=1}^{p} \omega_{ik}^{\Gamma_{ik}} \),

where the probability parameters are normalized using scale factors \( \eta_i \) and \( \eta_k \) as

\[
\omega_{ik} = \eta_i \eta_k \omega_{ik},
\]

such that:

\[
\sum_{k=1}^{p} \omega_{ik} = 1, \quad \forall i \in \{1, \ldots, p\},
\]

\[
\sum_{i=1}^{p} \omega_{ik} = 1, \quad \forall k \in \{1, \ldots, p\},
\]

to satisfy the prior requirements, Eq. (A.3), with

\[
\log \omega_{ik} = \langle u_i^\top B_k^i u_i \rangle - \log C(B_k^i) + \alpha_k^i \log \beta_k^i - \log F(\alpha_k^i) + (\alpha_k^i - 1) \langle \log r_i \rangle - \beta_k^i \langle r_i \rangle,
\]

where \( F(\cdot) \) denotes the so-called digamma function. We may view \( \log \omega_{ik} \) as a logarithmic similarity measure between the \( k \)th prior axes at the previous resolution and \( i \)th posterior axes at the current resolution.
B.5 Posterior distribution of the latent remainder term

The optimal posterior distribution of \( q(\mathbf{z}_T) \) is given by

\[
q(\mathbf{z}_T) = \prod_{j=1}^{m} \prod_{l=1}^{T(0)} \mathcal{N}(\mathbf{z}_{t,l}; \langle \mathbf{z}_{t,l}^{(j)} \rangle, \langle \gamma^{(j)} \rangle^{-1}),
\]

\[
\langle \mathbf{z}_{t,l}^{(j)} \rangle = \sum_{j'=0}^{j-1} \hat{\mathbf{f}}_{t,l}^{(j')} + \langle \mathbf{b}_l^{(j)} \rangle + \sum_{i=1}^{p} \langle a_{i,l} \mathbf{u}_i \rangle \phi_i^{(j)}(\mathbf{x}_l, \hat{\mathbf{r}}_{t,l}^{(j)}).
\]

B.6 Posterior distribution of overall bias and residual noise precision

The optimized posterior of the joint distribution of the mean vector and the residual noise is given by

\[
q(\mathbf{b}, \gamma) = \prod_{j=0}^{m} \prod_{l=1}^{T(l)} \mathcal{N}(\mathbf{b}_l^{(j)}; \nu_l^{(j)}, \frac{1}{\hat{\gamma}_l^{(j)}}) \mathcal{G}(\gamma_l^{(j)}; \zeta_l^{(j)}, \delta_l^{(j)}),
\]

with the posterior hyper-parameters given by

\[
\hat{\gamma}_l^{(j)} = \gamma_l^{(j)} + |T_l^{(j)}|,
\]

\[
\nu_l^{(j)} = \frac{1}{\hat{\gamma}_l^{(j)}} \left( \nu_l^{(j)} + \nu_l^{(j)} + \hat{\nu}_l^{(j)} \right),
\]

\[
\zeta_l^{(j)} = \zeta_l^{(j)} + \frac{d_y}{2} |T_l^{(j)}|,
\]

\[
\delta_l^{(j)} = \delta_l^{(j)} + \frac{1}{2} \delta_l^{(j)},
\]

where \( \hat{\nu}_l^{(0)} \) and \( \bar{\delta}_l^{(0)} \) are given by

\[
\hat{\nu}_l^{(0)} = \sum_{t \in T_l^{(0)}} \left( \mathbf{y}_t - \sum_{i=1}^{p} \phi_i^{(0)}(\mathbf{x}_l, \hat{\mathbf{r}}_l^{(0)}) \right),
\]

\[
\bar{\delta}_l^{(0)} = \hat{\nu}_l^{(0)} ||\nu_l^{(0)}||^2 - \hat{\gamma}_l^{(0)} ||\nu_l^{(0)}||^2 +
\sum_{t \in T_l^{(0)}} \left( ||\mathbf{y}_t - \sum_{i=1}^{p} \phi_i^{(0)}(\mathbf{x}_l, \hat{\mathbf{r}}_l^{(0)}) \langle a_i \mathbf{u}_i \rangle ||^2 + \sum_{i=1}^{p} \phi_i^{(0)}(\mathbf{x}_l, \hat{\mathbf{r}}_l^{(0)})^2 \langle ||a_i \mathbf{u}_i - \langle a_i \mathbf{u}_i \rangle ||^2 \rangle \right).
\]

and similarly \( \hat{\nu}_l^{(j)} \) and \( \bar{\delta}_l^{(j)} \), \( \forall j \geq 1 \), are given by

\[
\hat{\nu}_l^{(j)} = \sum_{t \in T_l^{(j)}} \left( \langle \mathbf{z}_{t,l}^{(j)} \rangle - \sum_{j'=0}^{j-1} \hat{\mathbf{f}}_{t,l}^{(j')} - \sum_{i=1}^{p} \langle a_{i,t} \mathbf{u}_i \rangle \phi_i^{(j)}(\mathbf{x}_l, \hat{\mathbf{r}}_{t,l}^{(j)}) \right),
\]

\[
\bar{\delta}_l^{(j)} = \hat{\nu}_l^{(j)} ||\nu_l^{(j)}||^2 - \hat{\gamma}_l^{(j)} ||\nu_l^{(j)}||^2 +
\sum_{t \in T_l^{(j)}} \left( ||\langle \mathbf{z}_{t,l}^{(j)} \rangle - \sum_{i=1}^{p} \phi_i^{(j)}(\mathbf{x}_l, \hat{\mathbf{r}}_{t,l}^{(j)}) \langle a_{i,t} \mathbf{u}_i \rangle - \sum_{j'=0}^{j-1} \hat{\mathbf{f}}_{t,l}^{(j')} ||^2 \right.
\]

\[
+ \langle ||\mathbf{z}_{t,l}^{(j)} - \langle \mathbf{z}_{t,l}^{(j)} \rangle ||^2 \rangle + \sum_{j'=0}^{j-1} \mathbb{E} \left[ ||\hat{\mathbf{f}}_{t,l}^{(j')} - \hat{\mathbf{f}}_{t,l}^{(j')} ||^2 \right]
\]

\[
+ \sum_{i=1}^{p} \phi_i^{(j)}(\mathbf{x}_l, \hat{\mathbf{r}}_{t,l}^{(j)})^2 \langle ||a_{i,t} \mathbf{u}_i - \langle a_{i,t} \mathbf{u}_i \rangle ||^2 \rangle \right).
\]

Estimating the noise precision at resolution \( j \geq 1 \) also includes the second central moments of the predictive processes and the latent remainder terms at the previous resolutions.
C Descriptive statistics

Descriptive statistics of the posterior distributions \( q(r_i), q(a_{i,j}^{(j)} | u_i) \), and \( q(z_{\mathcal{T}^{(j)}}) \) are conveniently given by the known statistics of the Gamma and Gaussian distributions. For \( q(\Gamma) \), we have the standard result of \( (\Gamma_{ik}) = \omega_{ik} \). With a special notational treatment for \( j = 0 \), the required statistics for the joint posterior \( q(u_i, a_{i,j}^{(j)}), \forall j \), are summarized as

\[
\langle u_i u_i^T \rangle = \sum_{d=1}^{d_u} \rho_{id}(k_i) \mu_{id} \mu_{id}^T,
\]
\[
\langle a_{i,j}^{(j)} u_i \rangle = \zeta_{i,j} \langle u_i u_i^T \rangle \tilde{z}_{i,j}^T,
\]
\[
\left( \langle a_{i,j}^{(j)} \rangle^2 \right) = \frac{1}{\nu_{i,j}} + \left( \zeta_{i,j} \right)^2 \left( \langle u_i u_i^T \rangle \right) \langle a_{i,j}^{(j)} \rangle^T \left( \langle u_i u_i^T \rangle - \langle a_{i,j}^{(j)} \rangle \langle a_{i,j}^{(j)} u_i \rangle \right) \tilde{z}_{i,j}^T,
\]
\[
\left( \| a_{i,j}^{(j)} u_i - \langle a_{i,j}^{(j)} \rangle u_i \| \right)^2 = \frac{1}{\nu_{i,j}} + \zeta_{i,j}^2 \left( \langle a_{i,j}^{(j)} \rangle \right)^T \left( \langle u_i u_i^T \rangle - \langle a_{i,j}^{(j)} \rangle \langle a_{i,j}^{(j)} u_i \rangle \right) \tilde{z}_{i,j}^T,
\]

where \( \rho_{id}(k_i) \) is the \( d \)-th element of \( \rho_i(k_i) \) given by

\[
\rho_i(k_i) = \frac{\partial \log \mathcal{C}(k_i)}{\partial k_i}, \quad \forall i \in \mathcal{P}.
\]

The saddle-point approximation of Kume and Wood [26] is used to calculate the derivatives above. The mean and the second central moment of the predictive processes can be computed using Eq. (D.1c) and (D.1d),

\[
E \left[ \tilde{f}_{i,j}^{(j)} \right] = \bar{f}_{i,j}^{(j)} = \hat{f}_{j}^{(j)}(x_i),
\]
\[
E \left[ \| \tilde{f}_{i,j}^{(j)} - \hat{f}_{j}^{(j)} \|^2 \right] = \text{trace} \left[ \tilde{F}_{i,j}^{(j)}(x_i) \right], \quad \forall x_i \in \mathcal{T}_i^{(j)}.
\]

D Predictive process

For a new test input \( x^* \), we shall first determine if we know to which region it belongs in each resolution. If such information is available the required statistics of the approximate predictive process at \( x^* \) can be computed from the sum of their contributions across all resolutions, as

\[
E \left[ p(f(x^*) | x^*, x_{\mathcal{T}}, y_{\mathcal{T}}, z_{\mathcal{T}}) \right] = \sum_{j=0}^{m} \bar{f}_{i,j}^{(j)}(x^*),
\]
\[
\text{Cov} \left[ p(f(x^*) | x^*, x_{\mathcal{T}}, y_{\mathcal{T}}, z_{\mathcal{T}}) \right] = \sum_{j=0}^{m} \tilde{F}_{i,j}^{(j)}(x^*),
\]

where \( \bar{f}_{i,j}^{(j)}(x^*) \) and \( \tilde{F}_{i,j}^{(j)}(x^*) \) are given by

\[
\bar{f}_{i,j}^{(j)}(x^*) = \langle b_{i,j}^T \rangle + \sum_{i=1}^{p} \langle a_{i,j}^{(j)} u_i \rangle \phi_{i,j}^{(j)}(x^*, \tau_{i,j}^{(j)}),
\]
\[
\tilde{F}_{i,j}^{(j)}(x^*) = \langle b_{i,j} b_{i,j}^T \rangle - \langle b_{i,j} \rangle \langle b_{i,j}^T \rangle + \sum_{i=1}^{p} \left( \phi_{i,j}^{(j)}(x^*, \tau_{i,j}^{(j)}) \right)^2 \times \left[ \left( \langle a_{i,j}^{(j)} \rangle^2 \right) - \langle a_{i,j}^{(j)} u_i \rangle \langle a_{i,j}^{(j)} u_i \rangle^T \right].
\]

In many applications however we may indeed not know the position of \( x^* \) in the training index sets, \( \mathcal{T}_i^{(j)}, \forall j \),—in other words we may not know to which region \( x^* \) belongs at a given resolution. In such
cases, since the basis axes are shared across all resolutions and learnt in a group fashion, predictions are made only from $j = 0$,
\begin{align}
E \left[ p(f(x^*) \mid x^*, x_T, y_T, z_T) \right] &= \tilde{F}^{(0)}(x^*), \quad \text{(D.2a)} \\
\text{Cov} \left[ p(f(x^*) \mid x^*, x_T, y_T, z_T) \right] &= F^{(0)}(x^*). \quad \text{(D.2b)}
\end{align}
We emphasize that, among others, this is one of the advantages of the conditional independence over models with full independence.

## E Optimization of basis interval variables

The basis interval variables $\tau^{(j)}_d = (\tau^{(j)}_{d_1}, \ldots, \tau^{(j)}_{d_D})^T$ are optimized using maximum likelihood estimation, as an analytical solution within our standard variational inference may not exist in general form for various choices of basis functions and spectral densities. The optimized point estimate values are given from
\[
\hat{\tau}^{(j)}_d = \arg \max_{\tau^{(j)}_d} h(\tau^{(j)}_d),
\]
s.t.
\[
L^{(j)}_{d_{s_d}} < \tau^{(j)}_d < L^{(j)}_{d_{s_d} + 1} + \frac{p}{L^{(j)}_{d_{s_d}}},
\]
where $L^{(j)}_{d_{s_d}}$ is the input range at $(j, l)$, and
\[
h(\hat{\tau}^{(j)}_d) \propto h_{\text{prior}}(\tau^{(j)}_d) + h_{\text{likelihood}}(\tau^{(j)}_d),
\]
where $h_{\text{prior}}(\tau^{(j)}_d)$ includes all relevant terms from the prior,
\[
h_{\text{prior}}(\tau^{(j)}_d) = -\frac{1}{2} \sum_{i=1}^p \left( \log \left( \mathcal{E}_{d,l,i}^{(j)}(\tau^{(j)}_d) \right) - \frac{(\tau^{(j)}_d - a^{(j)}_{l,i})^2}{2\mathcal{E}_{d,l,i}^{(j)}(\tau^{(j)}_d)} \right),
\]
and $h_{\text{likelihood}}(\tau^{(j)}_d)$ includes all relevant terms in the likelihood term,
\[
h_{\text{likelihood}}(\tau^{(j)}_d) = -\left( \gamma^{(j)}_l \right) \sum_{t=1}^{T^{(j)}} \sum_{i=1}^p \left( \| \mathcal{X}(x^{(j)}_{t,d,i}, \tau^{(j)}_d) \|^2 + 2 \left( \phi^{(j)}_{d,l,t,i} \right)^T \mathcal{X}(x^{(j)}_{t,d,i}, \tau^{(j)}_d) \right)
\]
\[
+ \left( \left\| a^{(j)}_{t,i} \mathbf{u}_i - \left( a^{(j)}_{t,i} \right) \right\|^2 \left( \phi^{(j)}_{d,l,t,i} \theta^{(j)}_{i,d,l,t} \phi^{(j)}_{t} (x^{(j)}_{t,d,i}, \tau^{(j)}_d) \right) \right),
\]
where we have defined
\[
\phi^{(j)}_{t} = \left( \phi^{(j)}_{t} \right) + \sum_{j'=0}^{j-1} \mathbf{f}^{(j')} - \frac{1}{2} \mathbf{z}^{(j)},
\]
\[
\mathcal{E}_{d,l,i}^{(j)}(\tau^{(j)}_d) = S^{(j)} \left( \sqrt{\lambda^{(j)}_i (\tau^{(j)}_d) + \bar{\lambda}^{(j)}_{i,d,l}} \right),
\]
\[
\mathcal{X}(x^{(j)}_{t,d,i}, \tau^{(j)}_d) = \left( a^{(j)}_{t,i} \mathbf{u}_i \right) \phi^{(j)}_{d,l,t,i} \phi^{(j)}_{t} (x^{(j)}_{t,d,i}, \tau^{(j)}_d),
\]
\[
\lambda^{(j)}_{i,d,l} = \sum_{k \neq d} \lambda^{(j)}_i (\tau^{(j)}_{k,l}),
\]
\[
\phi^{(j)}_{d,l,t,i} = \prod_{k \neq d} \phi^{(j)}_{t} (x^{(j)}_{t,k}, \tau^{(j)}_{k,l}),
\]
where $\tau^{(j)}_{k,l}$ are the previous optimized values. The optimization problem is solved numerically.
F  Experiment details

This section provides further details on the experiments in Sec. 5.

F.1  Datasets

F.1.1  oes10 and oes97

The datasets oes10 and oes97 were obtained from [38]. The Occupational Employment Survey (OES) datasets contain records from the years of 1997 (OES97) and 2010 (OES10) of the annual Occupational Employment Survey compiled by the US Bureau of Labor Statistics. As described in [38], "each row provides the estimated number of full-time equivalent employees across many employment types for a specific metropolitan area". We selected the same 16 target variables as listed in [38, Table 5]. The remaining 298 and 263 variables serve as the inputs in the case of oes10 and oes97, respectively. Data samples were randomly divided into training and test sets (refer to Table 1).

F.1.2  atp1d and atp7d

The datasets atp1d and atp7d were obtained from [38]. The Airline Ticket Price (ATP) dataset includes the prediction of airline ticket prices. As described in [38], the target variables are either the next day price, atp1d, or minimum price observed over the next seven days atp7d for 6 target flight preferences listed in [38, Table 5]. There are 411 input variables in each case. The inputs for each sample are values considered to be useful for prediction of the airline ticket prices for a specific departure date, for example, the number of days between the observation date and the departure date, or the boolean variables for day-of-the-week of the observation date. Data samples were randomly divided into training and test sets (refer to Table 1).

F.1.3  scm1d, scm1d-a and scm20d

The datasets scm1d and scm20d were obtained from [38]. The Supply Chain Management (SCM) datasets are derived from the Trading Agent Competition in Supply Chain Management (TAC SCM) tournament from 2010. As described in [38], each row corresponds to an observation day in the tournament. There are 280 input variables in these datasets which are observed prices for a specific tournament day. The datasets contain 16 regression targets, where each target corresponds to the next day mean price scm1d or mean price for 20 days in the future scm20d for each product [38, Table 5]. Dataset scm1d-a is a subset of scm1d which includes the first 3000 samples. Data samples were randomly divided into training and test sets (refer to Table 1).

F.1.4  naval

The dataset naval [8] was obtained from UCI Machine Learning Repository. The input variables are 16-dimensional feature vectors containing the gas turbine (GT) measures at steady state of the physical asset, for example, GT rate of revolutions, and Gas Generator rate of revolutions. The targets are two dimensional vectors measuring GT Compressor decay state coefficients and GT Turbine decay state coefficients. Data samples were randomly divided into training and test sets (refer to Table 1).

F.1.5  vicon

The dataset vicon contains measurements recorded from a magnetic field which maps a 3-dimensional (3D) position to a 3D magnetic field strength [20]. The inputs are \((x, y, z)\)-coordinates and the responses measured at there different heights are the target values. Data samples were randomly divided into training and test sets (refer to Table 1).

---

3http://archive.ics.uci.edu/ml/datasets/condition+based+maintenance+of+naval+propulsion+plants
4More information about data can be found in [20]. The data is available from https://github.com/carji475/linearly-constrained-gaussian-processes
F.1.6 hrtf

The dataset hrtf was obtained from the CIPIC HRTF database \[1\] which is a public-domain database of high-spatial-resolution head-related transfer function (HRTF) measurements.\[2\] We used the datasets of 37 subjects divided into training and test sets (refer to Table 1). Data for each subject includes 200-dimensional measurements of head-related impulse responses (HRIRs) and 8 input variables which are in fact the anthropometric parameters considered to have strong direct physical effect on HRIRs. The objective is to predict the HRIRs of the test subjects given their individualized anthropometric parameters\[3\].

F.1.7 nengo

The dataset nengo for this analysis was generated using The neural engineering object (Nengo) simulator \[5\]. The generated time series data is constructed from a Nengo-based spiking model of action selection in the cortex-basal ganglia-thalamus circuit with timing predictions that are well matched to both single-cell recordings in rats and psychological paradigms in humans. Target measurements here are ensembles of leaky integrate-and-fire neurons comprised from seven nodes of the basal ganglia circuit (namely: globus pallidus internal, globus pallidus external, subthalamic nucleus, striatum D1, striatum D2; thalamus; motor cortex). Measurements from these 7 nodes are the target outputs. The advantage of using the Nengo neural simulator in the regression task is that we also have access to the ground-truth, the function generating the noisy target measurements at each node. Data samples\[4\] were randomly divided into training and test sets (refer to Table 1).

F.1.8 lorenz96

The synthetic dataset lorenz96 was generated using the Lorenz model \[28, Eq. 3.2\]. Using a locally defined notation, consider the Lorenz model of

\[
\frac{dx_k}{dt} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + F, \quad \forall k \in K,
\]

where \(x_k\) represent the state of the system and \(F\) is the forcing constant. In our simulation, we let \(K = 20\) and set \(F = 8\), which cause chaotic behavior. The initial state was set to equilibrium and a small perturbation was given to a randomly selected state. A small amount of noise was added to the resulting \(d_y = 20\) dimensional feature vector. For the input ranging from 0 to 8, 1000 samples were collected on a linear space from the system. The objective is to identify the latent function generating data and perform predictions at \(10^5\) locations in this interval, \([0, 8]\).

F.2 Datasets used in the illustrative experiment in Section 5, Figure 2

F.2.1 ToyData

The synthetic dataset ToyData for the regression task in Figure 2(a) is generated using the following nonlinear functions

\[
f_1(x) = \exp\{\sin(\cos(x)) \sin(\log(1 + |x^2 - 3x|))\},
\]

\[
f_2(x) = \log(|\tan(-2x) \cos(2x) + 1|) \sin(x).
\]

We generated 32 noisy samples for input values in the range of \(x \in [0, 12]\). The objective is to estimate the latent functions and perform predictions at \(10^5\) locations in this interval, \([0, 12]\).

F.2.2 vicon2

The dataset vicon2 is a subset of the vicon dataset \[F.1.5\] which includes 6000 samples from which 5000 randomly selected samples are used in the test set and 1000 samples in the training set. vicon2 is used in our numerical simulation presented in Figure 2(b).

---

\[1\] Details of the database can be found at: https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/.

\[2\] The preprocessed data can be obtained from our GitHub page: <GitHub link to data>.

\[3\] Data can be downloaded from our GitHub page: <GitHub link to data>.
F.3 Methods

F.3.1 SGPMC
MCMC for Variational Sparse Gaussian Processes (SGPMC) model of [19] using GPflow implementation with RBF-ARD kernels, Gaussian likelihood, and 1000 pseudo inputs.

F.3.2 SVGP
The scalable variational Gaussian process (SVGP) model of [18] using a GPflow implementation with RBF-ARD kernels, Gaussian likelihood, and 1000 pseudo inputs.

F.3.3 SVIGP
Stochastic variational GP (SVIGP) model of [17] using a GPy implementation with RBF-ARD kernels, Gaussian likelihood, and 1000 pseudo inputs.

---

8https://github.com/GPflow
9https://github.com/SheffieldML/GPy
Acknowledgements

This research is financially supported by The Knut and Alice Wallenberg Foundation (J. Taghia, contract number: KAW2014.0392), by the Swedish Research Council (VR) via the project NewLEADS - New Directions in Learning Dynamical Systems (T. B. Schön, contract number: 621-2016-06079), and by the Swedish Foundation for Strategic Research (SSF) via the project ASSEMBLE (T. B. Schön, contract number: RIT15-0012). 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 thank Arne Leijon and Andreas Svensson for useful discussions and feedback on the manuscript.

References

[1] R. V. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The CIPIC HRTF Database. In WASSAP ’01, 2001.

[2] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O’Neil. Fast direct methods for Gaussian processes. IEEE Trans. Pattern Anal. Mach. Intell., 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, 2014. ISSN 1662-5196. doi: 10.3389/fninf.2013.00048.

[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, 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), Scottsdale, AZ, USA, 2013.

[11] Y. Ding, R. Kondor, and J. Eskreis-Winkler. Multiresolution kernel approximation for Gaussian process regression. In Neural Information Processing Systems (NIPS) 30. Long Beach, CA, USA, 2017.

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

[13] R. Frigola, Y. Chen, and C. E. Rasmussen. Variational Gaussian process state-space models. In Neural Information Processing Systems (NIPS), Montreal, Canada, 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 multi-dimensional 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 Intelligence*, pages 282–290. auai.org, 2013.

[18] J. Hensman, A. G. de G. Matthews, and Z. Ghahramani. Scalable variational Gaussian process classification. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2015, San Diego, California, USA, May 9-12, 2015*, 2015.

[19] J. Hensman, A. G. Matthews, M. Filippone, and Z. Ghahramani. MCMC for variationally sparse Gaussian processes. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1648–1656, 2015.

[20] C. Jidling, N. Wahlström, A. Wills, and T. B. Schön. Linearly constrained Gaussian processes. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1215–1224, 2017.

[21] 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.

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

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

[24] 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.

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

[26] A. Kume and A. T. A. Wood. Saddlepoint approximations for the Bingham and Fisher-Bingham normalising constants. *Biometrika*, 92(2):465–476, 2005.

[27] R. Leu and P. Damien. Bayesian shape analysis of the complex Bingham distribution. *Journal of Statistical Planning and Inference*, 149:183–200, 2014. ISSN 0378-3758.

[28] E. Lorenz. Predictability: a problem partly solved. In *Seminar on Predictability, 4-8 September 1995*, volume 1, pages 1–18, Shinfield Park, Reading, 1995.

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

[30] 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. ISSN 0378-3758.

[31] 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.

[32] 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.

[33] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. New York, NY, USA, 2006.
[34] 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.

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

[36] E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press, 2006.

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

[38] 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.

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

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