Scalable Gaussian Process Separation for Kernels with a Non-Stationary Phase

Jan Graßhoff [1] **Alexandra Jankowski** [1] **Philipp Rostalski** [1]

Abstract

The application of Gaussian processes (GPs) to large data sets is limited due to heavy memory and computational requirements. A variety of methods has been proposed to enable scalability, one of which is to exploit structure in the kernel matrix. Previous methods, however, cannot easily deal with mixtures of non-stationary processes. This paper investigates an efficient GP framework, that extends structured kernel interpolation methods to GPs with a non-stationary phase. We particularly treat the separation of non-stationary sources, which is a problem that commonly arises e.g. in spatio-temporal biomedical datasets. Our approach employs multiple sets of non-equidistant inducing points to account for the non-stationarity and retrieve Toeplitz and Kronecker structure in the kernel matrix allowing for efficient inference and kernel learning. Our approach is demonstrated on numerical examples and large spatio-temporal biomedical problems.

1. Introduction

Gaussian processes (GPs) provide interpretable models for solving regression, classification and prediction problems in a huge number of scientific domains. GP regression originates from early work in geostatistics, where the technique was known under the name of kriging (Rasmussen & Williams, 2005). Since then, its great potential of discovering intricate structure in data has been shown empirically in numerous problems. Still, due to $\mathcal{O}(n^3)$ computational and $\mathcal{O}(n^2)$ storage cost in the number of training points n, scalability to large datasets remains as the main limiting factor in many practical applications and restricts GPs to datasets containing at most a few thousand observations.

Different approaches to scalable GP regression have been proposed (Liu et al., 2018), one of which is based on computing low-rank approximations of the kernel matrix by using sparse inducing point sets (Snelson & Ghahramani, 2006; Quinonero Candela & Rasmussen, 2005). Inducing point methods are particularly useful when the data are densely sampled compared to the characteristic length-scale of the underlying process. However, short-scale variability requires a large amount of inducing points, which in such cases diminishes the performance. Also, these methods scale poorly on long time-series data (and spatio-temporal data) as the extending domain has to be filled up with inducing points (Solin et al., 2018).

Another orthogonal line of research deals with the exploitation of structure in the kernel matrix: among the most promising approaches are (1) state-space representation methods (Hartikainen & Särkkä, 2010; Solin & Särkkä, 2014; Särkkä & Hartikainen, 2012) and (2) methods exploiting Toeplitz and Kronecker matrix structure (Cunningham et al., 2008; Saatci, 2011). The state-space approach enables highly scalable $\mathcal{O}(n)$ inference and marginal likelihood evaluation for spatio-temporal data – though, it might become slow if the gaps between data points are very uneven. We focus on the second structure exploiting approach, namely Toeplitz/Kronecker matrix structure for fast matrix-vector multiplications (MVMs) which requires that samples are distributed on a multidimensional lattice and that the kernel is stationary/separable. The restriction to lattice structures was later lifted through an approach called structured kernel interpolation (SKI), introduced by Wilson & Nickisch (2015). It employs a structured set of inducing points and a sparse interpolation matrix enabling fast MVMs with the kernel matrix without requiring any special data structure.

In this paper we are concerned with solving the source separation and model learning problem for additive mixtures of non-stationary processes. We show, that SKI can be naturally extended for mixtures of non-stationary kernels, more specifically kernels with a non-stationary phase in an approach we call warpSKI. We propose to use multiple non-equispaced sets of inducing points to recover structure in the kernel matrix. This permits us to solve new classes of important problems via GPs, in particular temporal/spatio-temporal source separation and regression problems on large biomedical datasets, which is the main contribution of this work. We focus on the case that the

[1]Institute for Electrical Engineering in Medicine, Universität zu Lübeck, Germany. Correspondence to: Jan Graßhoff <j.grasshoff@uni-luebeck.de>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

non-stationary phase can be directly extracted as a feature from the data. It was also proposed to learn non-linear transformations of the inputs using neural networks (Wilson et al., 2016) or a second GP (Plagemann et al., 2008; Heinonen et al., 2016) – for the herein considered biomedical problems it can be beneficial to use mixtures of predetermined warping functions, e.g. accounting for known fluctuations in the respiratory or heart rate, to extract highly interpretable structure from the data. We combine this approach with stochastic trace estimation (Dong et al., 2017) to efficiently evaluate the marginal likelihood of the full mixture model. We demonstrate scalability of our proposed method to $n \geq 10^5$ points on a numerical example and on openly available biomedical datasets. As in standard SKI, storage complexity is reduced to $\mathcal{O}(n+m)$ and computational complexity to $\mathcal{O}(n+g(m))$ for inference/learning, where $g(m) \leq m \log m$ with m being the number of inducing points. Code implementations of the proposed warpSKI in MATLAB are provided as an extension to the GPML 4.2 toolbox (Rasmussen & Nickisch, 2010) and are available under `github.com/ime-luebeck/non-stationary-phase-gp-mod`.

2. Background

2.1. Gaussian Process Regression

This section provides a brief overview of Gaussian process regression and model learning, the interested reader is referred to Rasmussen & Williams (2005) for more details. A Gaussian process $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ with $\mathbf{x} \in \mathbb{R}^D$ encodes the prior belief in the distribution of the function values $f(\mathbf{x})$ and is fully specified by a kernel function $k(\mathbf{x}, \mathbf{x}')$, parameterized in a (usually low) number of hyperparameters $\boldsymbol{\theta}$. The choice of the kernel allows to define the properties of the function f, e.g. its noise color or periodicity. A GP can formally be understood as an infinite-dimensional generalization of the normal distribution. For any finite set of points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^D$, the function $f(\mathbf{x}_i)$ evaluated at those points has a multivariate Gaussian distribution $\mathbf{f} = [f(\mathbf{x}_1), \ldots f(\mathbf{x}_n)]^\top \sim \mathcal{N}(\mathbf{0}, K_{XX})$, where the entries of the kernel matrix $(K_{f,XX})_{i,j}$ are formed by evaluating the kernel function for all pairs $\mathbf{x}_i$ and $\mathbf{x}_j$. In most GP applications it is assumed, that only noisy measurements of the latent function are available, which we denote as a vector of targets $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Here we assume that the measurements are subject to additive noise $y = f(\mathbf{x}) + \epsilon(\mathbf{x})$ which is in turn a Gaussian process $\epsilon \sim \mathcal{GP}(0, k_\epsilon(\mathbf{x}, \mathbf{x}'))$. In this specific case, the predictive distribution at n_* test points X_* has a closed form solution given by

$$\begin{aligned} \mathbf{f}_* | X, X_*, \mathbf{y}, \boldsymbol{\theta}, \sigma^2 &\sim \mathcal{N}(\bar{\mathbf{f}}_*, \operatorname{cov}(\mathbf{f}_*)), \\ \bar{\mathbf{f}}_* &= K_{X_*X}[K_{XX} + K_{\epsilon,XX}]^{-1}\mathbf{y}, \\ \operatorname{cov}(\mathbf{f}_*) &= K_{X_*X_*} - K_{X_*X}[K_{XX} + K_{\epsilon,XX}]^{-1}K_{XX_*}, \end{aligned} \tag{1}$$

where $\mathbf{f}_*$ is the vector of function values evaluated at the test points and matrices of the form $K_{X_iX_j}$ denote cross-covariances between respectively two sets of points X_i and X_j. In the standard GP regression setting, ϵ is assumed to be a Gaussian white noise process with variance σ^2, which corresponds to choosing $K_{\epsilon,XX} = \sigma^2 I$.

The hyperparameters $\boldsymbol{\theta}$ of the kernel are usually learned directly from the data by optimizing the negative log marginal likelihood

$$\begin{aligned} &-\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) \propto \\ &\quad \mathbf{y}^\top (K_{XX} + K_{\epsilon,XX})^{-1}\mathbf{y} + \log|K_{XX} + K_{\epsilon,XX}|, \end{aligned} \tag{2}$$

which can be done e.g. by gradient-based minimization or sampling. Computing the inverse and the log determinant of $K_{XX} + K_{\epsilon,XX}$ are the main bottlenecks for GP inference and model learning. Both involve computing the Cholesky factorization which leads to an overall complexity of $\mathcal{O}(n^3)$. The storage complexity is determined by the need to store the full kernel matrix, leading to $\mathcal{O}(n^2)$.

2.2. Kronecker and Toeplitz Methods

A growing line of research investigates the exploitation of structure in the kernel matrix to achieve scalable GP inference and model learning. When input points are on a multidimensional rectilinear lattice (not necessarily equispaced) and the kernel is separable along input dimensions, $k(\mathbf{x}, \mathbf{x}') = \prod_{d=1}^{D} k^{(d)}(\mathbf{x}^{(d)}, \mathbf{x}'^{(d)})$, the kernel matrix has Kronecker structure, i.e. it can be written as $K = K_1 \otimes \cdots \otimes K_D$. This enables fast MVMs in $\mathcal{O}(Dn^{\frac{D+1}{D}})$ time (Wilson et al., 2014) and the eigendecomposition of the full matrix can be efficiently calculated by separately taking the eigendecompositions of the smaller matrices K_i. Thus, in the GP regression setting subject to Gaussian white noise, the solution to the linear system $(K_{XX}+\sigma^2 I)^{-1}\mathbf{y}$ and the log determinant $\log|K_{XX}+\sigma^2 I|$ can be evaluated efficiently (Wilson & Nickisch, 2015).

Cunningham et al. (2008) proposed another, orthogonal method namely the exploitation of Toeplitz structure (constant diagonals of the kernel matrix), which arises when the input points are placed equidistantly in $\mathbb{R}$ and the kernel is stationary (that is, $k(x, x') = k(\tau)$, with $\tau = x - x'$). Toeplitz structure allows for fast MVMs using Fourier transforms (Cunningham et al., 2008; Wilson & Nickisch, 2015), thus the matrix inverse can be computed by conjugate gradients in $\mathcal{O}(n \log n)$. Kronecker and Toeplitz methods complement each other in the sense that they exploit multidimensional and 1D structure, respectively. The supplementary material contains a comprehensive overview of fast Kronecker/Toeplitz methods in the GP setting.

2.3. Structured Kernel Interpolation

Kronecker and Toeplitz methods are restricted to a few highly specialized problems due to the requirement, that input points are either equispaced or on a multidimensional lattice. The highest performance gains are reached when both of these requirements are met. Wilson & Nickisch (2015) presented a general purpose inference framework, that exploits structure even for partial-grid/unstructured data by placing inducing points on a multidimensional equispaced lattice. Inducing point methods have long been used throughout the literature for large-scale GP applications and reduce runtime cost to $\mathcal{O}(m^3 + m^2 n)$ and storage cost to $\mathcal{O}(mn + m^2)$ (Quinonero Candela & Rasmussen, 2005), where m is the number of inducing points. Usually inducing point methods perform best when the data are densely sampled and few inducing points suffice, i.e. $m \ll n$. Yet, in long temporal/spatio-temporal regression tasks these methods suffer from the need to sample the extending input domain with a high amount of inducing points. Addressing these problems, Wilson & Nickisch (2015) introduced an approximation called structured kernel interpolation (SKI) of the form $\tilde{K}_{XU} = W K_{UU}$, where $K_{UU} \in \mathbb{R}^{m \times m}$ is the exact kernel matrix evaluated for inducing points U and $W \in \mathbb{R}^{n \times m}$ is an interpolation weight matrix. The interpolation matrix can be made very sparse, consisting of only four non-zero entries per row. The full approximate kernel matrix can thus be written as

$$K_{XX} \approx W K_{UU} W^\top := K_{\text{SKI}}. \tag{3}$$

MVMs with W can be computed in $\mathcal{O}(n)$ time and, when placing the inducing points on a lattice, MVMs with K_{UU} can exploit Kronecker and Toeplitz structure, with worst-case cost of $\mathcal{O}(m \log m)$ for only Toeplitz structure – the total runtime for MVMs is thus $\mathcal{O}(n + m \log m)$ (Wilson & Nickisch, 2015). When both Toeplitz structure and Kronecker structure are exploited, the total runtime for MVMs becomes $\mathcal{O}(n + g(m))$, where $g(m) < m \log m$ and in many cases we can assume a quasi-linear complexity $g(m) \approx m$, (Wang et al., 2019). Storage costs are reduced to $\mathcal{O}(n+m)$. Thus, SKI significantly relaxes restrictions on the number of inducing points, allowing even for $m \approx n$.

3. Scalable GPs with a Non-Stationary Phase

In this work, we are concerned with mixtures of non-stationary Gaussian processes of the form

$$\begin{aligned} f_{\text{m}}(\mathbf{x}) &= \sum_i f_{i,\text{warp}}(\mathbf{x}) \\ \text{with } f_{i,\text{warp}} &\sim \mathcal{GP}(0, k_i(\phi_i(\mathbf{x}), \phi_i(\mathbf{x}'))), \end{aligned} \tag{4}$$

where k_i are product separable along input dimensions and stationary (that is, $k_i(\mathbf{x}, \mathbf{x}') = k_i(\boldsymbol{\tau})$, with $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$). The functions $\phi_i : \mathcal{D}_{\text{in}} \to \mathcal{D}_i$ are invertible space warping functions with $\mathcal{D}_{\text{in}} \subseteq \mathbb{R}^D$, $\mathcal{D}_i \subseteq \mathbb{R}^D$ and do not have singularities in the input domain $\mathcal{D}_{\text{in}}$. The full kernel corresponding to (4) is given by $k_{\text{m}}(\mathbf{x}, \mathbf{x}') = \sum_i k_{i,\text{warp}}(\mathbf{x}, \mathbf{x}') = \sum_i k_i(\phi_i(\mathbf{x}), \phi_i(\mathbf{x}'))$. In our experiments we will show that this kernel is applicable to a large number of important problems. The kernel property would be preserved also for non-invertible space warping functions, but for reasons that will become clear later, we focus on invertible functions. Note, that product separability of the kernel is lost due to the warping function.

The particular use case, that is investigated in this paper, is the separation of non-stationary GPs. In the GP framework, source separation can be achieved by extracting the function corresponding to the jth source of the mixture via the posterior $\mathbf{f}_{j,\text{warp}}|\mathbf{y}$. Again, this posterior has a closed form solution (Liutkus et al., 2011), which is given by equation (1) replacing k with $k_{j,\text{warp}}$, k_ϵ with $\sum_{i \neq j} k_{i,\text{warp}}$ and choosing $X = X_*$.

3.1. WarpSKI

We wish to reduce the time and storage costs for marginal likelihood evaluations and for the source separation problem when the GP is a mixture of processes with a non-stationary phase as in (4). We do not want to introduce any restrictions on the structure of input points, in particular partial grid structures or missing values must be supported. At first, structure in kernels with a non-stationary phase cannot easily be exploited: previously proposed methods fail in this case because they either assume stationarity of the kernel or rely on its separability. One of the main difficulties in using (4) in temporal GP regression is that Toeplitz structure is lost due to the warping functions. To make things worse, the summation structure in (4) alone leads to the loss of the Kronecker product property for the whole kernel k_{m} even when all ϕ_i are linear functions.

In the following we apply structured kernel interpolation to mixtures of non-stationary functions in a variant we call warpSKI. We propose to employ multiple sets of inducing points to recover structure in the kernel of the GP in (4) and enable scalable GP regression.

As a first step we consider a single warped kernel $k_{\text{warp}}(\mathbf{x}, \mathbf{x}') = k(\phi(\mathbf{x}), \phi(\mathbf{x}'))$, corresponding to one of the summands in (4). It is easily verified, that SKI can be applied to the stationary and separable kernel k when using the warped input points $\mathbf{z} = \phi(\mathbf{x})$ (Wilson et al., 2016). This leads to an approximate form for the warped kernel matrix

$$K_{\text{warp},XX} \approx W_Z K_{UU} W_Z^\top, \tag{5}$$

where we have used $\tilde{K}_{ZU} = W_Z K_{UU}$ and W_Z interpolates between the inducing points and the warped points

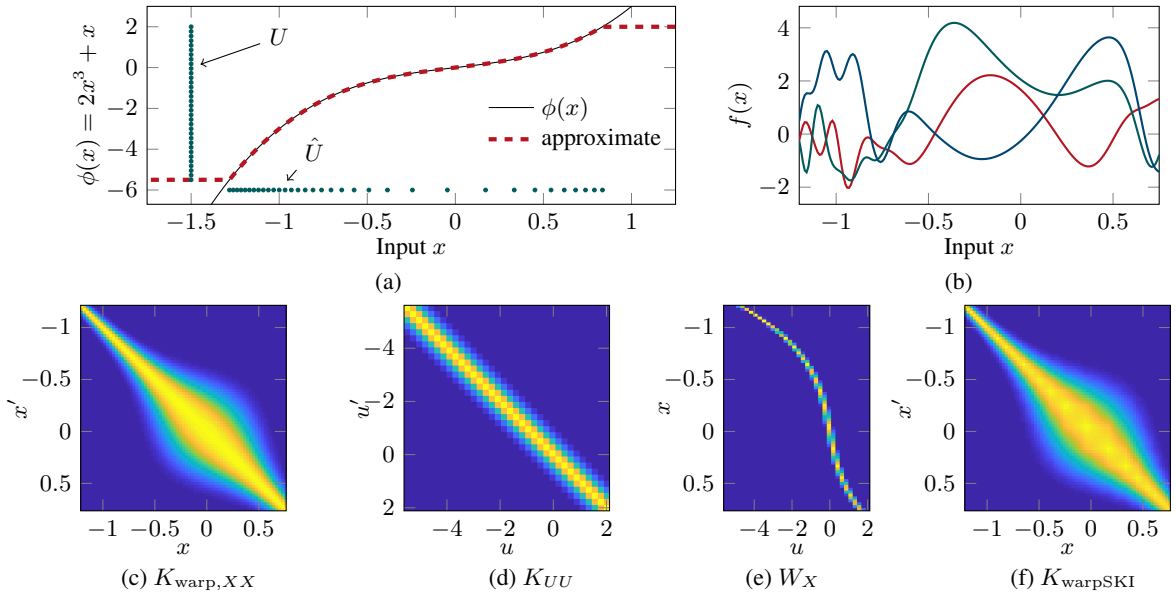


Figure 1: Illustrative example of warpSKI for squared exponential kernel $k_{\text{SE}}(\phi(x), \phi(x'))$ with $\phi(x) = 2x^3 + x$. In (a), the warping function ϕ and the equidistant inducing points U as well as the non-equidistant inducing points $\hat{U}$ are depicted. The approximate warping function induced by U and $\hat{U}$ closely reflects the true function within the support domain of U and $\hat{U}$. In (b), samples from the warped kernel are depicted. Figures (c) to (f) show the different matrices involved in warpSKI.

$Z = \{\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_n)\} \coloneqq \Phi(X)$. The matrix K_{UU} now only depends on the stationary/separable kernel k and thus structure (Toeplitz and Kronecker) can be imposed by placement of U. Interestingly, the non-stationarity of k_{warp} is now fully embedded in the sparse matrix W_Z.

We propose a reinterpretation for the case that ϕ is invertible: Instead of using interpolation between the warped points $Z = \Phi(X)$ and the inducing points U (placed on an equispaced rectilinear grid), one can equivalently interpolate between the original input points X and a set of inducing points $\hat{U} = \Phi^{-1}(U)$. This leads to the approximation

$$\begin{aligned} K_{\text{warp},XX} &\approx W_X K_{\text{warp},\hat{U}\hat{U}} W_X^\top = (5) \\ &= W_X K_{UU} W_X^\top \coloneqq K_{\text{warpSKI}}, \end{aligned} \tag{6}$$

where we have used $\tilde{K}_{X\hat{U}} = W_X K_{\text{warp},\hat{U}\hat{U}}$ and we exploit the fact that $K_{\text{warp},\hat{U}\hat{U}} = K_{UU}$. The sparse matrix W_X interpolates between the inducing points $\hat{U}$ and the input points X. Here, in contrast to standard SKI, the inducing point set $\hat{U}$ does have a warped (non-equidistant) lattice structure and thereby accounts for the warping function. Using this consideration, we now have an approximate form $W_X K_{UU} W_X^\top$ for the warped kernel k_{warp}, that is fully specified in the structure of an inducing point set $\hat{U}$ and a stationary and separable kernel k. This can be seen as an extension to SKI, that is able to generate rich kernel structure by placing non-equidistant/warped inducing points $\hat{U}$. The herein presented idea is related to the deep kernel learning method (Wilson et al., 2016), where a single set of equidistant inducing points is placed in the warped space – when ϕ is invertible this directly leads to working with non-equidistant inducing points, whether this is done explicitly or implicitly. See Figure 1 as an example for all matrices involved in warpSKI.

We continue by specifying the approximate form for the full non-stationary kernel in (4):

$$K_{\text{m},XX} \approx \sum_i K_{i,\text{warpSKI}} = \sum_i W_{i,X} K_{i,U_iU_i} W_{i,X}^\top, \tag{7}$$

where we use multiple sets of non-equispaced/warped inducing point sets $\hat{U}_i$ and $W_{i,X}$ interpolating between those points and the input points X. Note, that the space warping functions are now fully embedded in the matrices $W_{i,X}$ through placement of $\hat{U}_i$ and structure (Toeplitz/Kronecker) is imposed on K_{i,U_iU_i} by placing U_i. Fast MVMs are possible and the inference problem can be solved by linear conjugate gradients requiring $j \ll n$ steps for convergence up to machine precision leading to $\mathcal{O}(n + g(m))$ runtime with $g(m) \leq m \log m$.

3.2. Spatio-Temporal Gaussian Processes

Next, we consider the case, that the space warping function ϕ is an elementwise function, which means that it can be written as a vector of one dimensional functions

$$\phi(\mathbf{x}) = \left[\phi^{(1)}(\mathbf{x}^{(1)}), \ldots, \phi^{(D)}(\mathbf{x}^{(D)})\right]^\top. \tag{8}$$

When the space warping functions ϕ_i in (4) are elementwise functions, the summands $k_{i,\text{warp}}$ become product separable and the whole kernel can be written as

$$k_{\text{m}}(\mathbf{x}, \mathbf{z}) = \sum_i \prod_d k_i^{(d)} \left(\phi_i^{(d)}(\mathbf{x}^{(d)}), \phi_i^{(d)}(\mathbf{z}^{(d)}) \right). \quad (9)$$

An important special case of (9) is given by the spatio-temporal covariance function

$$k_{\text{m}}(\mathbf{s}, \mathbf{s}', t, t') = \sum_i k_{i,\mathbf{s}}(\mathbf{s}, \mathbf{s}') k_{i,\text{t}}(\phi_i(t), \phi_i(t')), \quad (10)$$

where non-stationarity is only assumed for the temporal domain – in this case $\phi_i(t)$ is called a time warping function (Müller, 2007). When $k_{i,\text{t}}$ is a periodic kernel, $\phi_i(t)$ can be considered to be a 'phase warping function' and maps from the time domain to multiples of 2π, thus specifying the period length. Such models often arise in biomedical applications due to superposition of physiological processes such as cardiac or respiratory activity, which are inherently non-stationary in time. In many practical large-scale biomedical problems, this necessitates efficient solutions to the corresponding inference and model learning task.

Note, that the kernel in (9) enables Kronecker structure MVMs also for standard SKI (with equidistant lattice inducing points), but Toeplitz structure is lost. In contrast, warpSKI (with non-equidistant lattice inducing points) does recover Kronecker and Toeplitz structure. We argue, that Toeplitz structure is particularly important in temporal and spatio-temporal regression problems due to the possibly long time axis and that Kronecker structure exploitation can quickly become prohibitive with respect to the temporal domain. Therefore, warpSKI offers important advantages and poses no restrictions on the amount of inducing points placed over the temporal axis. A comparison between standard SKI methods and the warpSKI variant can be found in the supplementary material.

3.3. Fast Source Separation

Having approximated the solution to the linear problem $\boldsymbol{\alpha} \approx \tilde{\boldsymbol{\alpha}} = \left[\sum_i K_{i,\text{warpSKI}} + \sigma^2 I\right]^{-1} \mathbf{y}$ via conjugate gradients, we continue to consider the source separation problem (Liutkus et al., 2011), i.e. our goal is to approximate the mean of the posterior $\mathbf{f}_{j,\text{warp}}|\mathbf{y}$ corresponding to the jth source in the mixture. The standard GP solution to the posterior mean is given by $K_{j,\text{warp},X_*X}\tilde{\boldsymbol{\alpha}}$ and would require $\mathcal{O}(n^2)$ time for n test points $X_* = X$. For the source separation problem we can again exploit kernel structure using the approximation

$$\mathbb{E}\left[\mathbf{f}_{j,\text{warp}}|\mathbf{y}\right] \approx K_{j,\text{warpSKI}}\tilde{\boldsymbol{\alpha}} = W_{j,X} K_{j,U_jU_j} W_{j,X}^\top \tilde{\boldsymbol{\alpha}}. \quad (11)$$

The complexity for source separation, once $\tilde{\boldsymbol{\alpha}}$ was obtained, is $\mathcal{O}(n + m \log m)$ for Toeplitz structure in K_{j,U_jU_j} and quasi-linear $\mathcal{O}(n + g(m))$ (with $g(m) \approx m$) for Kronecker/Toeplitz structure.

3.4. Fast Model Learning

The hyperparameters of all kernels in the sum can be learned by jointly optimizing the marginal likelihood. Different structure exploiting approximations have been proposed. Unfortunately, the model specified in (4) does not allow to use the highly efficient scaled eigenvalue method (Wilson et al., 2014) due to its summation structure. Therefore, we will consider the recently introduced stochastic trace estimation approach (Dong et al., 2017), which approximates the log determinant $\log|K_{XX} + \sigma^2 I|$ and its derivative w.r.t. the hyperparameters also via an iterative MVM scheme. In its core, the method uses $\log|\hat{K}| = \text{trace}(\log(\hat{K})) = \mathbb{E}[\mathbf{z}^\top \log(\hat{K})\mathbf{z}]$, where $\mathbf{z}$ is commonly chosen to be a vector with Rademacher random variables as entries. Usually, few probe vectors suffice to get a good approximation – leaving us with the task of calculating $\log(\hat{K})\mathbf{z}$. For this task, a Lanczos decomposition approach has been proposed by Dong et al. (2017). The stochastic trace estimation approach accesses the kernel matrix only through MVMs and is thus compatible with the proposed approximate form in (7).

4. Experiments

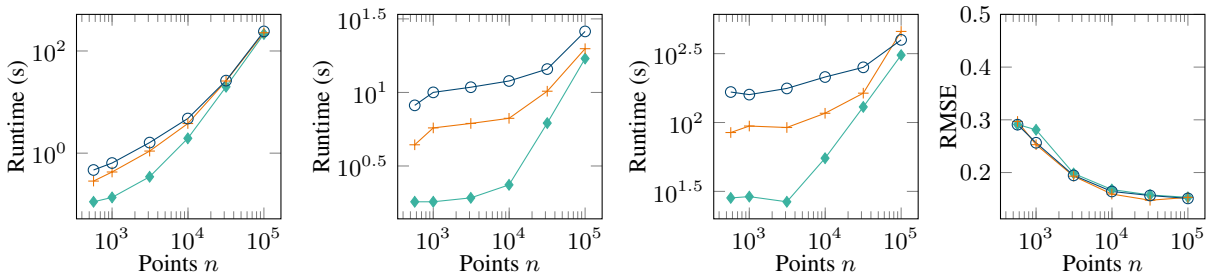
We evaluate the non-stationary GP framework on a numerical dataset and then we aim to motivate the usefulness of the method in relevant large-scale biomedical applications, covering both temporal and spatio-temporal problems.

WarpSKI was implemented in MATLAB as an extension to the GPML 4.2 library, all experiments were carried out on a workstation with an INTEL Core i7-6700K CPU. In all experiments, L-BFGS (Liu & Nocedal, 1989) was used for hyperparameter learning, respectively with a maximum of 100 optimization steps.

4.1. Numerical Data

As a first test, we apply warpSKI to a mildly non-linear separable warping function on a numerical 2D example. Samples are generated from a warped squared exponential kernel of the form $k_{\text{SE}}(\phi(\mathbf{x}), \phi(\mathbf{x})')$, where ϕ is given by $\phi([x_1, x_2]^\top) = [2x_1^3 + x_1, x_2]^\top$ and the hyperparameters were set to $\ell_{\text{SE}} = 0.4$ and $\sigma_{\text{SE}} = 1.5$. The input point positions are sampled from a uniform distribution within a rectangular area (spanning $[-1.2, 0.75] \times [-2.5, 2.5] \subset \mathbb{R}^2$). To account for the non-stationarity of the kernel, one non-equidistant inducing point set $\hat{U}$ is used.

The generation of high-dimensional GP samples is itself a non-trivial task as it requires Cholesky decomposition of the full kernel matrix. Therefore, to generate samples with up to 10^5 input points, we exploit the Kronecker structure



(a) Inference (b) Marg. Likelihood Eval. (c) Learning (d) Error

Figure 2: Results of warpSKI for numerical data with variable numbers of inducing/input points: Inference time is in (a), time for marginal likelihood evaluations is in (b), hyperparameter learning runtime in (c) and RMSE against true GP sample in (d). The number of inducing points is $m = 9933$ (◆), $m = 49\,824$ (+) and $m = 74\,836$ (○). We report numerical values and standard deviations in the supplementary material.

of the inducing point kernel matrix and use a high number of inducing points to warp the samples with high accuracy. We then apply additive Gaussian white noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$ to form the final regression targets $\mathbf{y}$. In Figure 2, we show the inference and model optimization times (here, σ, σ_{SE} and ℓ_{SE} were learned) as well as the root-mean-square-error (RMSE) for different inducing/input point sizes. The reported numbers are averages over five independent runs with respectively different random seeds. The tolerance for conjugate gradients was set to 10^{-1}, and marginal likelihood evaluations were done using 20 probe vectors in the stochastic trace estimation.

As this example does not include mixtures of non-stationary warping functions, it enables direct comparison of the proposed method to previous publications because it could be transformed back to an equivalent standard SKI task as discussed in Section 3.1. Therefore the results obtained for this space-warped GP match earlier results, compare for instance to Wilson & Nickisch (2015). In particular, note that SKI is inexpensive with respect to the number of inducing points.

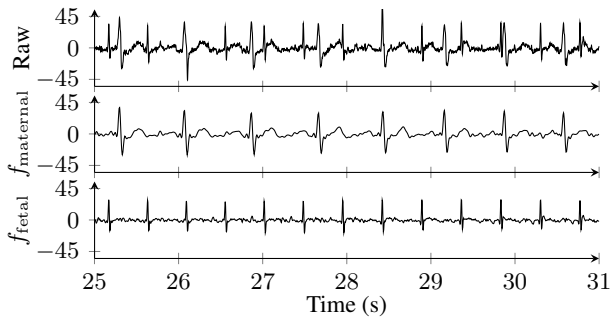
4.2. Fetal ECG Data

Next, we test the proposed fast source separation algorithm based on warpSKI on large biomedical datasets. The application of GPs to biomedical time-series data has been proposed by many authors, but scalability to large datasets has been lacking. In this example we apply GP models to ECG data, specifically we treat the separation of fetal ECG and maternal ECG signals in baseline-free abdominal recordings. This was previously demonstrated on small datasets by Niknazar et al. (2012), where a model as in (4) with two nonlinear warping functions was used and the source separation was solved via the classical batch GP formulation. As opposed to warpSKI this previous approach did not exploit matrix structure and thus could not be applied to larger datasets. We validate the proposed warpSKI on data taken from the Physionet fetal ECG database (Jezewski et al., 2012), using the 4th channel of subject R01 and compare its performance to that of the batch GP.

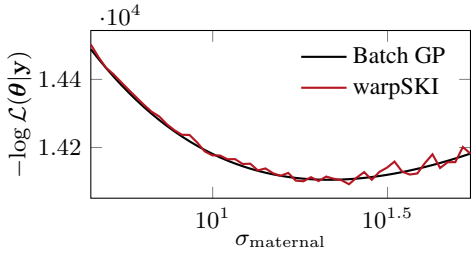
Similar to Niknazar et al. (2012), the kernel is chosen as a mixture of two phase-warped processes, i.e. $k_{\text{m}}(t, t') = k_{\text{maternal}}(\phi_1(t), \phi_1(t')) + k_{\text{fetal}}(\phi_2(t), \phi_2(t'))$, where both k_{maternal} and k_{fetal} are quasi-periodic kernels as defined by Rasmussen & Williams (2005). Note, that vanilla SKI cannot be applied in this example as it does not recover Toeplitz structure.

Following Niknazar et al. (2012), the optimization of hyperparameters in the ECG separation problem should be guided by prior knowledge by either fixing hyperparameters to reasonable estimates or by using strong hyperpriors. In this experiment, we used fixed values for the length-scales of the two quasi-periodic kernel functions (which appears to be beneficial for ECG signals) and optimize for the variance hyperparameters σ_{maternal} and σ_{fetal}. In the considered dataset, the nonlinear warping functions can be extracted directly from the data by detecting fetal and maternal R peaks in the provided reference signals and assuming a constant phase between respectively two R peaks. In Niknazar et al. (2012), it was also proposed to learn the full warping function by optimization of the model likelihood, which however is difficult due to the high number of local minima. The optimization of non-stationary warping functions was discussed in more detail by e.g. Plagemann et al. (2008); Heinonen et al. (2016), but shall not be the focus of this work. Generally, for this type of applications, we recommend to use reference signals for the phase whenever possible or at least use strong priors on the phase.

As a first step, we apply warpSKI to a subset of the data consisting of $n = 5000$ points (corresponding to 10 seconds of data sampled at 500 Hz) and validate it against the batch GP solution. We then do a large-scale stress test using 10^5



(a)



(b)

Figure 3: Results of fetal ECG extraction. In (a), an excerpt from the large-scale (100 s) source separation is depicted. In (b), a comparison between exact and approximate warpSKI marginal likelihoods is shown on the smaller dataset (10 s) for different values of σ_{maternal}.

input points (corresponding to 100 seconds of data sampled at 1000 Hz) to demonstrate scalability. As a measure for the separation success we use the SNR improvement metric that was proposed in the context of ECG denoising by (Bartolo et al., 1996). For the stresstest, again 20 probe vectors were used for stochastic trace estimation. LCG tolerance was set to 10^{-1} and $5 \cdot 10^{-3}$ for parameter learning and source separation, respectively. In Figure 3, an excerpt of the large-scale source separation as well as a comparison of exact/approximate marginal likelihood evaluations is depicted. Performance measures are reported in Table 1.

4.3. Electrical Impedance Tomography Data

Next, we consider a non-stationary spatio-temporal signal separation problem given by electrical impedance tomography (EIT) images of the chest. In this example EIT is used to measure regional changes in the impedance of the lung, which are caused either by changes in the ventilation or perfusion of the lung tissue. The separation of these two effects is a long-standing problem, previous approaches applied pixel-wise Fourier filtering, which however omits the spatial structure of pixels and cannot fully separate the two effects due to overlap in the spectrum (Pikkemaat & Leonhardt, 2010). We show that the separation of the two pulsatile components in EIT images can be posed as a spatio-temporal GP regression problem using a mixture of non-stationary

Table 1: Performance comparison of fetal ECG extraction for different input sizes and methods.

	batch GP	warpSKI	warpSKI
Input points	5000	5000	10^5
Inducing points	–	3400 (m) + 4800 (f)	14300 (m) + 21600 (f)
Time for Inference	2.60 s	0.27 s	4.14 s
Time for learning	28.5 s	47.4 s	462.3 s
σ_{fetal}	5.08	5.95	5.11
σ_{maternal}	21.48	21.42	32.11
SNR improvement	18.6 dB	18.1 dB	18.2 dB

kernels. Inference and model learning can then be solved efficiently via the methods proposed in this paper.

As a model for the two superposed effects we use the spatio-temporal kernel

$$k_{\text{m}}(\mathbf{s}, \mathbf{s}', t, t') = k_{\text{vent,SE}}(\mathbf{s}, \mathbf{s}') k_{\text{vent,QP}}(\phi_1(t), \phi_1(t')) + k_{\text{perf,SE}}(\mathbf{s}, \mathbf{s}') k_{\text{perf,QP}}(\phi_2(t), \phi_2(t')),$$

where a squared exponential kernel is assumed for the spatial domain and quasi-periodicity for the temporal domain in both signal components (ventilation and perfusion).

The considered dataset of a spontaneously breathing neonate is taken from Heinrich et al. (2006). Note, that the EIT problem has 'partial-grid' structure, consisting only of pixels within a circular area. We use the first 215 frames to train our model and as a measure of training success we predict the next frame and evaluate the prediction error (using normalized RMSE), see Table 2. As in the previous example, the phase warping function is determined directly from the data – the respiratory phase was extracted from a pixel belonging to the left lung, the cardiac phase was extracted from a pixel between the two lungs. Note, that standard SKI could also be applied for each of the kernels in the sum but, in contrast to warpSKI, would not recover Toeplitz structure and thus does not enable scalability to longer recordings.

As in the previous example, it is beneficial to use hyperpriors and fix some of the hyperparameters (based on prior knowledge) to guide the optimization. Here, we optimize for the variances σ_{vent}, σ_{perf}, the length-scales of the spatial kernel $\ell_{\text{vent,SE}}$, $\ell_{\text{perf,SE}}$ and the length-scales of the time domain $\ell_{\text{vent,SE-QP}}$, $\ell_{\text{perf,SE-QP}}$ with regularizing lognormal hyperpriors on all of the length-scales. The remaining hyperparameters σ, $\ell_{\text{vent,PE-QP}}$ and $\ell_{\text{perf,PE-QP}}$ were fixed based on features we already found in the data a priori. For this experiment only 15 probe vectors were used for stochastic trace estimation to speed up the optimization. The LCG tolerance was set to 0.25 and 10^{-2} for optimization and source separation, respectively. Figure 4 shows the result of the source separation on the considered dataset.

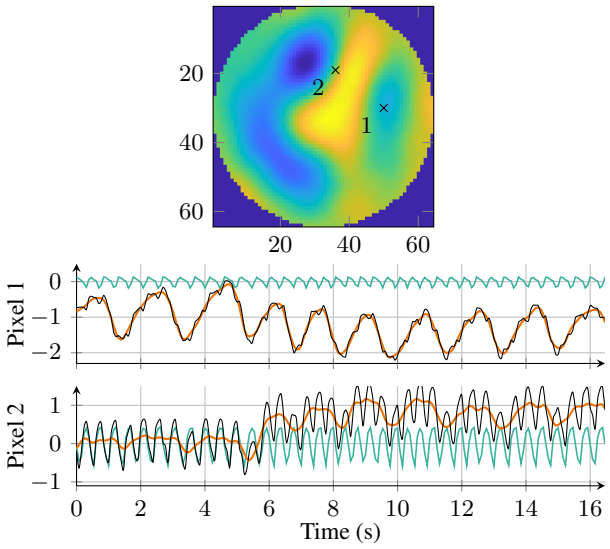


Figure 4: Result of EIT perfusion-ventilation separation. Time traces correspond to the marked pixels and include measured signals (**black**) and the posterior means of perfusion (**green**) ventilation (**orange**) related signals.

Table 2: Runtime and performance of warpSKI for the spatio-temporal experiments.

	EIT	BSP
Input points	699 825	67 256
Inducing points	~3×10^6 (perf) + ~2×10^6 (vent)	~1.9×10^6
Time for inference	159.1 s	5.4 s
Time for learning	~9 h	~1.2 h
nRMSE	0.176	0.217

4.4. Body Surface Potential Data

As a last example we consider a spatio-temporal regression problem given by body surface potentials (BSP) data measured by means of high-density electrode grids. When electrodes are placed on the thorax, these data can be used to obtain detailed electrocardiographic information beyond the classical 12 lead ECG. The measured potential is typically displayed for diagnostic purposes using a two-dimensional map of the unfolded thorax geometry – to this end, the measured BSP is to be interpolated between the electrodes. Different methods for interpolation have been proposed (Schijvenaars et al., 1995), which however only use spatial correlation between electrodes. We show, that the interpolation task can also be solved exploiting temporal structure by using a non-stationary spatio-temporal GP.

As a BSP model we use $k_{\mathrm{SE}}(\mathbf{s}, \mathbf{s}')k_{\mathrm{QP}}(\phi(t), \phi(t'))$ similar to the EIT model. The spatial kernel is defined over the surface of the thorax geometry using the cylindrical mapping $k_{\mathrm{SE}}(\mathbf{s}, \mathbf{s}') = k_{1,\mathrm{SE}}\left(\frac{\mathbf{v}}{\|\mathbf{v}\|}, \frac{\mathbf{v}'}{\|\mathbf{v}'\|}\right) \cdot k_{2,\mathrm{SE}}(z, z')$, where $\mathbf{v} = (x, y)^\top$ and z are the 3D electrode coordinates. As before, the phase warping function ϕ is determined using a linear phase between detected R peaks. We use one non-equidistant inducing point set $\hat{U}$ to account for ϕ and solve the regression problem. The data for this problem were taken from (Aras et al., 2015) and consist of a 3D model and measured electrode potentials. We train our model on 6 s from 62 electrodes sampled at 200 Hz. As a measure of training success, we remove 6 of the anterior electrodes prior to training and report the nRMSE of predicted potentials (Table 2). Figure 5 shows predictions for a single cardiac cycle.

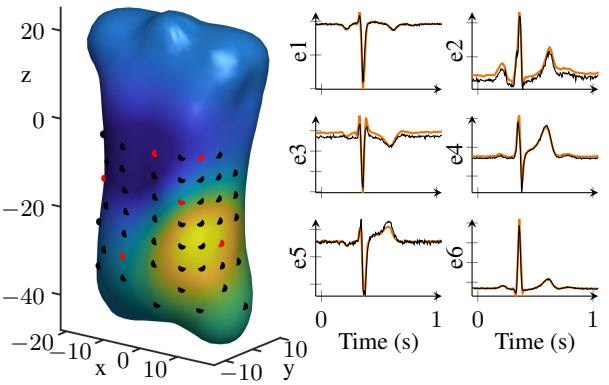


Figure 5: Results for BSP interpolation. The time traces of measured (**black**) and predicted (**orange**) potentials correspond to the six electrodes marked on the 3D geometry (removed prior to training).

Again, it is beneficial to fix some hyperparameters to guide the optimization – here we fix the parameters of the temporal model (based on a priori knowledge about ECG) and only optimize for the spatial parameters, where we learn the length-scales of both $k_{1,\mathrm{SE}}$ and $k_{2,\mathrm{SE}}$. The LCG and optimization parameters were the same as in Section 4.2.

5. Discussion

We have extended Gaussian process structured kernel interpolation to mixtures of kernels with a non-stationary phase. Our approach exploits matrix structure using multiple sets of non-equidistant/warped inducing point sets. We have shown, that this allows to solve large-scale source separation problems, which often arise in biomedical applications due to superposition of non-stationary physiological processes (such as respiratory/cardiac activity). In many biomedical modalities, where GPs could not be applied so far, the proposed method has a high potential of uncovering new and relevant structure.

Beyond that, we argue that the placement of non-equidistant

inducing points could be generally used as a tool to account for non-stationarity and build rich kernel structure – this idea might also be extended to other GP frameworks that are compatible with SKI such as Hensman et al. (2013).

It should be mentioned, that non-stationary phase functions can also be implemented via equivalent state-space models using the methods proposed by Hartikainen & Särkkä (2010); Solin & Särkkä (2014); Särkkä & Hartikainen (2012). This, in principle, leads to linear complexity in the number of time steps – in practice, the corresponding state-space models are sometimes high-dimensional in particular for spatio-temporal data. When choosing between the iterative batch approach and the recursive state-space approach one has to trade-off the different factors influencing the runtime for the specific problem at hand.

We see our work as part of a larger push in the recent GP literature that aims to access the kernel matrix only through matrix multiplications (Gardner et al., 2018; Wang et al., 2019) thus enabling highly scalable inference and learning. Exploiting intricate matrix structure for fast MVMs will be key to solving large-scale problems via GPs in the future.

Acknowledgements

The authors would like to thank Hannes Nickisch, Philips Research, Hamburg, Germany, for his constructive comments on the manuscript. Furthermore, they would also like to thank the anonymous reviewers for their valuable input.

References

Aras, K., Good, W., Tate, J. D., Burton, B., Brooks, D. H., Coll-Font, J., Dössel, O., Schulze, W. H. W., Potyagaylo, D., Wang, L., van Dam, P. M., and Macleod, R. Experimental data and geometric analysis repository-edgar. *Journal of electrocardiology*, 48 6:975–81, 2015. Data from http://edgar.sci.utah.edu/.

Bartolo, A., Roberts, C., Dzwonczyk, R. R., and Goldman, E. Analysis of diaphragm emg signals: comparison of gating vs. subtraction for removal of ecg contamination. *Journal of Applied Physiology*, 80(6):1898–1902, 1996.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. Fast gaussian process methods for point process intensity estimation. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pp. 192–199, New York, NY, USA, 2008. ACM.

Dong, K., Eriksson, D., Nickisch, H., Bindel, D., and Wilson, A. Scalable log determinants for gaussian process kernel learning. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 6327–6337, 2017.

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7587–7597, 2018.

Hartikainen, J. and Särkkä, S. Kalman filtering and smoothing solutions to temporal gaussian process regression models. *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 379–384, 2010.

Heinonen, M., Mannerström, H., Rousu, J., Kaski, S., and Lähdesmäki, H. Non-stationary gaussian process regression with hamiltonian monte carlo. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pp. 732–740, 09–11 May 2016.

Heinrich, S., Schiffmann, H., Frerichs, A., Klockgether-Radke, A., and Frerichs, I. Body and head position effects on regional lung ventilation in infants: an electrical impedance tomography study. *Intensive Care Medicine*, 32(9):1392, Jun 2006. Data from eidors3d.sourceforge.net/data_contrib/if-neonate-spontaneous/.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 282–290, Arlington, Virginia, United States, 2013. AUAI Press.

Jezewski, J., Matonia, A., Kupka, T., Roj, D., and Czabanski, R. Determination of fetal heart rate from abdominal signals: Evaluation of beat-to-beat accuracy in relation to the direct fetal electrocardiogram. *Biomedizinische Technik/Biomedical Engineering*, 57, 08 2012. Data from https://physionet.org/content/adfecgdb/1.0.0/.

Liu, D. C. and Nocedal, J. On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, 45(1):503–528, Aug 1989.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. When gaussian process meets big data: A review of scalable gps. 2018.

Liutkus, A., Badeau, R., and Richard, G. Gaussian Processes for Underdetermined Source Separation. *IEEE Transactions on Signal Processing*, 59(7):3155 – 3167, February 2011.

Müller, M. Dynamic time warping. *Information Retrieval for Music and Motion*, 2:69–84, 01 2007.

Niknazar, M., Rivet, B., and Jutten, C. Fetal ecg extraction from a single sensor by a non-parametric modeling. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*, pp. 949–953, 2012.

Pikkemaat, R. and Leonhardt, S. Separation of ventilation and perfusion related signals within EIT-data streams. *Journal of Physics: Conference Series*, 224:012028, 2010.

Plagemann, C., Kersting, K., and Burgard, W. Nonstationary gaussian process regression using point estimates of local smoothness. In *Machine Learning and Knowledge Discovery in Databases*, pp. 204–219, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

Quinonero Candela, J. and Rasmussen, C. E. A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research (JMLR)*, 6:1935–1959, December 2005.

Rasmussen, C. E. and Nickisch, H. Gaussian processes for machine learning (gpml) toolbox. *Journal of Machine Learning Research (JMLR)*, 11:3011–3015, December 2010.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Saatci, Y. *Scalable Inference for Structured Gaussian Process Models*. PhD thesis, University of Cambridge, 2011.

Särkkä, S. and Hartikainen, J. Infinite-dimensional kalman filtering approach to spatio-temporal gaussian process regression. In Lawrence, N. D. and Girolami, M. (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 993–1001, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

Schijvenaars, B. J., Kors, J. A., van Herpen, G., Kornreich, F., and van Bemmel, J. Interpolation of body surface potential maps. *Journal of Electrocardiology*, 28:104 – 109, 1995.

Snelson, E. and Ghahramani, Z. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18 (NeurIPS)*, pp. 1257–1264. 2006.

Solin, A. and Särkkä, S. Explicit link between periodic covariance functions and state space models. In *17th International Conference on Artifcial Intelligence and Statistics (AISTATS)*, 2014.

Solin, A., Hensman, J., and Turner, R. E. Infinite-horizon gaussian processes. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

Wang, K. A., Pleiss, G., Gardner, J. R., Tyree, S., Weinberger, K. Q., and Wilson, A. G. Exact gaussian processes on a million data points. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

Wilson, A. G. and Nickisch, H. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning (ICML)*, pp. 1775–1784, 2015.

Wilson, A. G., Gilboa, E., Nehorai, A., and Cunningham, J. P. Fast kernel learning for multidimensional pattern extrapolation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3626–3634, 2014.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. Deep kernel learning. In *Artificial Intelligence and Statistics*, pp. 370–378, 2016.