Sparse Spectral Bayesian Permanental Process with Generalized Kernel

Jeremy Sellier
Department of Statistical Science
University College London, UK

Petros Dellaportas
Department of Statistical Science
University College London, UK
Department of Statistics,
Univ. of Econ. and Business, Athens, Greece
and The Alan Turing Institute, UK

Abstract

We introduce a novel scheme for Bayesian inference on permanental processes which models the Poisson intensity as the square of a Gaussian process. Combining *generalized kernels* and a Fourier features-based representation of the Gaussian process with a Laplace approximation to the posterior, we achieve a fast and efficient inference that does not require numerical integration over the input space, allows kernel design and scales linearly with the number of events. Our method builds and improves upon the state-of-the-art Laplace Bayesian point process benchmark of Walder and Bishop (2017), demonstrated on both synthetic, real-world temporal and large spatial data sets.

1 INTRODUCTION

The Poisson process is commonly used in a variety of real-world point pattern applications including seismic activity (Gardner and Knopoff, 1974), epidemiology (Diggle, 2003; Banerjee et al., 2003), neuroscience (Cunningham, Yu, Shenoy and Sahani, 2008) and crime incident locations (Grubesic and Mack, 2008; Flaxman et al., 2019). A flexible generalization, incorporating a stochastic Poisson intensity (Cox, 1955) is achieved by placing a non-parametric Gaussian process (GP) prior over the intensity function, resulting in the popular Gaussian Cox process model.

Inference with the Gaussian Cox process model is challenging due to the *doubly-intractable* likelihood, requiring integration of an infinite-dimensional random function over the input domain. Furthermore, a positive transformation must be applied to the GP prior to ensure the intensity function remains non-negative. Different choices of transformation function are found in the literature. A first classical approach relies on the exponential transformation, resulting in the *log Gaussian Cox process* proposed by Møller et al. (1998). This model usually requires numerical approximation of the integral by discretization over the input space (Møller et al., 1998; Diggle et al., 2013), a computationally-intensive procedure which scales poorly with the dimensionality of the input domain. A second approach uses a sigmoid transformation and an input space augmentation via *thinning*, to construct an exact Markov Chain Monte Carlo sampler (Adams et al., 2009; Gunter et al., 2014), eliminating the need for likelihood integration. In practice however, this approach is computationally intractable for large problems. Other works include the use of the relu (Ko and Seeger, 2016) and softplus (Seeger and Bouchard, 2012; Park et al., 2014) as transformation functions. Finally, Lopez-lopera et al. (2019) introduce a finite approximation where positiveness conditions is imposed directly on the GP.

Another approach exploits the so-called *permanental process*, defining the Poisson process intensity in terms of the square of a GP (McCullagh and Møller, 2006; Lloyd et al., 2015). It enables analytical computation of the intensity integral when coupled with a variational inference scheme with inducing points (similar to Titsias 2009) and has received considerable recent attention (Lian et al., 2015; Flaxman et al., 2017; John and Hensman, 2018).

Walder and Bishop (2017) propose a Laplace Bayesian point process (LBPP) method, a fast alternative for the permanental process that relies on the Mercer decomposition of the Gaussian process kernel and a Laplace approximation to the intensity posterior. They show significant speed improvement compared to variational Bayesian inference. Inference based on the Laplace approximation has already been proposed in the context of a Gaussian Cox process by Cunningham, Shenoy and Sahani (2008), Illian et al. (2012), and Flaxman et al. (2015).

However, the tractability properties of the permanental pro-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

cess used by Lloyd et al. (2015) and Walder and Bishop (2017) only holds for certain standard types of kernels such as the squared exponential kernel, which encodes restrictive assumptions about the form of the function we are modelling. In general, the choice of kernel determines almost all the generalization properties of a Gaussian process model and profoundly affects its performance on a given task (Rasmussen and Williams, 2005). Approaches have been proposed in recent years to achieve more expressible kernels either by a composition of simple analytical forms (Duvenaud et al., 2011, 2013) or more flexibly through a spectral representation (Lázaro-Gredilla et al., 2010; Wilson and Adams, 2013; Samo and Roberts, 2015*a*).

In this paper, we build on the LBPP approach of Walder and Bishop (2017), introducing an alternative fast Laplace-based inference exploiting spectral representation of kernels and random Fourier features (RFFs). Our approach, the Sparse Spectral Permanental Process (SSPP), retains the tractability properties of the permanental process, whilst being able to adapt to a broader range of stationary kernels. Furthermore, our method works with *generalized stationary spectral kernels* (Samo and Roberts, 2015*a*), to our knowledge, the most general class of expressible spectral kernels, that can approximate any stationary kernels to arbitrary precision.

Following John and Hensman (2018), we also include a mean constant to mitigate the effect of *nodal lines* observed for the permanental process, resulting from the non-injective nature of the squared transformation. Our approach shows systematic improvement in accuracy in synthetic and real-world data sets.

2 PRELIMINARIES

2.1 Gaussian Cox Process

A Poisson process (see Daley and Vere-Jones, 2003) models a random sequence of points occurring on a continuous domain X. Inference involves estimation of an intensity function $\lambda(x) : X \to \mathbb{R}^+$, that can be interpreted heuristically as the instantaneous probability of occurrence of a point around a location $x \in X$, i.e. $\lambda(x) := \lim_{\mu(dx) \to 0^+} \mathbb{E}[N(dx)]/\mu(dx)$, where dx is a small neighborhood around x with measure $\mu(dx)$ and $N(A)$ is the random number of points within a sub-region $A \subseteq X$.

We focus our attention on Gaussian Cox process models, where λ is defined as $\lambda(\cdot) := l \circ f(\cdot)$, for a non-negative transformation $l : \mathbb{R} \to \mathbb{R}^+$ and a function $f \sim GP(0, k(x, x'))$, with $k : X \times X \to \mathbb{R}$ being the positive-definite covariance function for f. Assuming $\mathbf{X} = \{x_i\}_{i=1}^N$ to be a realization of N observations in X, the sample likelihood is

$$p(\mathbf{X}|f) = \exp\left(\int_X l(f(x))\,dx\right) \prod_{i=1}^N l(f(x_i)) \quad (1)$$

and a latent posterior $p(f|\mathbf{X})$ is

$$\frac{\exp\left(\int_X l(f(x))\,dx\right) \prod_{i=1}^N l(f(x_i))\; p(f)}{\exp\left(\int_X l(f(x))\,dx\right) \prod_{i=1}^N l(f(x_i))\; p(f)\,df} \quad (2)$$

where f represents the infinite-dimensional object corresponding to $f(x)$. Equations (2) is often described as "doubly-intractable". Inference for f requires evaluating integral terms that cannot be calculated explicitly.

2.2 Permanental Process

The permanental process (McCullagh and Møller, 2006) is obtained by defining the intensity in Equation (1) as the square of a Gaussian process i.e. setting $\lambda(\cdot) = f(\cdot)^2$. This transformation is advantageous, in that the density of the resulting permanental process is available analytically. To express the density we make use of the spectral decomposition of the Gaussian process covariance function, which we shortly review in this section.

2.2.1 Integral Expression via Mercer Theorem

The Gaussian process covariance function k has a Mercer decomposition on (X, μ), if it can be written as

$$k(x, x') = \sum_{i=1}^{\infty} \sigma_i \phi_i(x) \phi_i(x') \quad \text{for } x, x' \in X \quad (3)$$

where $\{\sigma_i\}_{i=1}^{\infty}$ is a sequence of summable, non-negative, non-increasing *eigenvalues*, and $\{\phi_i(\cdot)\}_{i=1}^{\infty}$ is a set of mutually-orthogonal, unit-norm *eigenfunctions* with respect to the inner product $\langle u, v \rangle = \int_X u(x)v(x)d\mu(x)$. Then in a similar manner to McCullagh and Møller (2006), Flaxman et al. (2017) and Walder and Bishop (2017), $f(x) \sim GP(0, k(x, x'))$ can be reformulated as an equivalent linear form

$$f(x) = \sum_{i=0}^{\infty} w_i \phi_i(x) \quad (4)$$

where $w = (w_1, w_2, \ldots)^\top \sim N(0, \Sigma)$ and Σ is a diagonal covariance matrix with entries σ_i; $i = 1, 2, \ldots$. Further, it can be shown that $\mathrm{Cov}(f(x), f(x')) = \phi(x)^\top \Sigma \phi(x') = k(x, x')$, where $\phi(\cdot)$ is a vector with entries $\phi_i(\cdot)$; $i = 1, 2, \ldots$. The integral of the intensity can then be expressed as

$$\begin{aligned}\int_X f(x)^2\, d\mu(x) &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} w_i w_j \langle \phi_i(x), \phi_j(x) \rangle \\ &= \sum_{i=0}^{\infty} w_i^2. \end{aligned} \quad (5)$$

2.2.2 Approximate Bayesian Inference

In our case, to make the reformulation of the integral $\int_S f(x)^2 dx$ possible as in Equation (5), the kernel k for f requires an explicit Mercer representation with respect to x and μ defined as the Lebesgue measure; this is not available for most choices of kernel. In such cases, the Nyström method can be used to approximate the eigenfunctions and eigenvalues of Equation (3). Both Flaxman et al. (2017) and Walder and Bishop (2017) adopt the Nyström approach in the context of the Permanental Cox process with Gaussian kernel. Walder and Bishop (2017) further propose a Bayesian inference scheme based on a Laplace approximation for a non-GP likelihood. A more detailed review of the Laplace approach together with a new result for the integral in Equation (5) with the Nyström method is presented in part A of the Appendix.

3 MODEL

Motivated in part by the shortcomings of the Nyström approach proposed by Walder and Bishop (2017), we now present an alternative LBBP approach to inference for the permanental process. In contrast to the Mercer approach, it is based on a sparse spectral representation of a GP, exploiting random Fourier features (RFFs, Rahimi and Recht, 2007) for reduced-rank kernel expression. As a result, it provides a tractable expression for the integral of the intensity over the input domain.

Our spectral approach works for any bounded, continuous and shift-invariant kernel $k(x, x') := k(x - x')$ that satisfies the condition of Bochner's theorem (see Theorem 3.1) and admits a finite dimensional feature space representation or approximation. In contrast, the variational inference approach of Lloyd et al. (2015) and the LBPP with Nyström, yield an analytical integral expression for a limited choice of kernels, like the Gaussian kernel. Furthermore, we are able to adapt our method to generalized stationary spectral kernels (Samo and Roberts, 2015) which generalize two other classes of expressible spectral kernels, the sparse spectrum kernels (Lázaro-Gredilla et al., 2010) and the mixture spectral kernels (Wilson and Adams, 2013). These two kernels have been proven to be able to approximate any bounded continuous stationary kernels to arbitrary precision.

We also address the issue of modal lines discussed in John and Hensman (2018). This problem arises since the inverse link function $\lambda(\cdot) := f(\cdot)^2$ is not injective, with $f(\cdot)$ producing the same intensity. Therefore, regions of negative and positive f must exhibit zero-crossings, where the intensity is artificially forced to zero, despite the underlying intensity being positive. Following John and Hensman (2018), we add an offset parameter α to the intensity function $\lambda(\cdot) := (f(\cdot) + \alpha)^2$ corresponding to an initial value for the prior mean of the GP, to alleviate the problem.

3.1 Sparse Spectral Kernels

In this section, we briefly present two families of spectral kernels, sparse spectrum kernels (Lázaro-Gredilla et al., 2010) and mixture spectral kernels (Wilson and Adams, 2013) that have been proposed in recent years for kernel design. They are both known to be dense in the family of stationary kernels, implying that they can approximate any stationary kernel to an arbitrary precision given sufficient spectral components.

Spectral kernels are constructed via the Bochner's theorem (Bochner, 1932), which states that any bounded, continuous and shift-invariant kernel $k(x, x') := k(\tau)$ with $\tau = x - x'$, is the inverse Fourier transform of a bounded positive measure.

Theorem 3.1. (Bochner) An integrable function $k : \mathbb{R}^d \to \mathbb{C}$ is the covariance function of a weakly stationary mean square continuous random process on $\mathbb{R}^d$ if and only if it can be represented as

$$k(\tau) = \int_{\mathbb{R}^d} \exp(i z^\top \tau) \, d\mu(z) \tag{6}$$

where $\mu(z)$ is a positive definite measure.

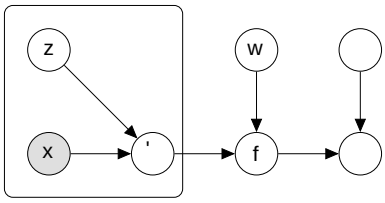
Sparse spectrum kernels can be obtained by setting in Equation (6) to be a positive discrete symmetric measure $\mu_{ss} = \sum_{k=1}^{K} \frac{a_k}{2} (\delta_{\omega_k} + \delta_{-\omega_k})$[1] where $a_k > 0$ and δ_{ω_k} denotes the Dirac measure centred at the point spectral frequencies $\omega_k \in \mathbb{R}^d$ for $k = 1, \dots, K$. Note that as such, μ_{ss} is singular with respect to the Lebesgue measure and does not admit a density. Through Equation (6), we obtained the sparse spectrum kernel spanned by the trigonometric functions $\{\cos(\omega_k^\top x)\}_{k=1}^K$. A major challenge is that a direct optimization of the linear coefficients $\{a_k\}_{k=1}^K$ and the frequencies $\{\omega_k\}_{k=1}^K$ often leads to over-fitting as illustrated by Lázaro-Gredilla et al. (2010) in the context of GP regression.

Wilson and Adams (2013) consider the case where μ is absolutely continuous with respect to the Lebesgue measure and admits a spectral density $S(\cdot)$. In that case, $S(\cdot)$ and the kernel function k are Fourier duals of each other. Mixture spectral kernels model the spectral density S as a mixture of independent Gaussian densities with non-zero mean. Since mixtures of Gaussians are dense in the set of all distribution functions (Plataniotis and Hatzinakos, 2001), the resulting dual of this set is dense in the family of continuous stationary kernels.

Random Fourier Features Random Fourier features (Rahimi and Recht, 2007) are closely related to sparse spectrum kernel. From Bochner's theorem, a kernel function k

[1] Note that positive finite discrete measures are weakly dense in the space of all positive finite measure (Hu and Papageorgiou, 2013).

Figure 1: Illustration of the model, where the arrows directions suggest directions of influence.



can be rewritten as

$$k(x - x') \approx \frac{2}{r} \sum_{k=1}^{r} \exp(i z_k^\top (x - x')) \quad (7)$$

where $z_1, \ldots, z_r$ in $\mathbb{R}^d$ are independent samples from the distribution with density $S(\cdot)$, for some integer $r > 0$ and > 0. Here, we have assumed that in Equation (6) is absolutely continuous with respect to the Lebesgue measure and has a spectral density $S(\cdot)$. Equation (6) is then approximated using Monte Carlo integration. We also treat the scale parameter of the kernel function separately for convenience.

We thus obtain a kernel approximation

$$k(x - x') \approx \phi^{(r)}(x)^\top \phi^{(r)}(x') \quad (8)$$

where $\phi^{(r)}$ is an explicit feature mapping $\phi^{(r)} : X \to \mathbb{R}^r$ such that

$$\phi^{(r)}(x) = \frac{}{\sqrt{r}} \exp(i z_1^\top x), \ldots, \exp(i z_r^\top x)^\top : \quad (9)$$

We may obtain a $2r$-sized real-valued mapping that satisfies Equation (7) using

$$\phi^{(r)}(x) = \frac{}{\sqrt{r}} [\cos(z_1^\top x), \ldots, \cos(z_r^\top x), \sin(z_1^\top x), \ldots, \sin(z_r^\top x)]^\top \quad (10)$$

where $z \sim S(z)$.

The derivation of Equation (10) is provided in part C of the Appendix. Some common kernels and their corresponding spectral densities $S(\cdot)$ are presented in part B of the Appendix. Inspecting Equation (10), we see RFF as a special case of sparse spectrum kernel where the frequencies $\{ _k \}$ are sampled at random from some distribution rather than optimised. However, RFF methods do not address the need for flexibly learning the spectral measure from the data.

3.2 Generalized Stationary Kernels

One advantage of our method is that it can work with the generalized stationary kernels (Samo and Roberts, 2015) that are dense in the family of stationary kernel and admits sparse spectrum kernels and mixture spectral kernels as special cases. Generalized kernels can also account for different degree of differentiability of the latent function. Sparse spectrum kernels and mixture spectral kernels are more limited in a sense that, when used as covariance functions, they yield infinite differentiability of the corresponding stochastic process, which might be unrealistic for certain learning tasks (Stein, 1999).

Definition 3.2. (Generalized stationary kernel) Let g be a stationary kernel $g : \mathbb{R}^d \to \mathbb{R}$ such that $g(0) = 1$. A generalized kernel k_{GS} with $K \in \mathbb{N}^+$ components takes the form

$$k_{GS}( ) = \sum_{k=1}^{K} {}_k^2 g({}_k) \cos(! {}_k^\top) \quad (11)$$

where $! _k \in \mathbb{R}^d$, $ _k \in \mathbb{R}^{+d}$, $ _k > 0$ for $k = 1, \ldots, K$ and denotes the element-wise Hadamard product.

The parameters $\{ _k\}_{k=1}^K$ are used as inverse input scales. When $\{ _k\}_{k=1}^K$ are set to zero, we retrieve the sparse spectrum kernels. The spectral mixture kernel corresponds to a special case where $g( ) = \exp(-\| \|_2^2/2) = \sqrt{2}$.

The degree of smoothness of a zero-mean GP with kernel k_{GS} is determined by the kernel. Samo (2017) proposes learning the differentiability of the underlying latent function, by setting g to be a Matérn kernel with different parameter values from $\frac{1}{2} + i; i = 0, \ldots, 2$. The case of $= 0$ corresponds to continuity and the case of > 0 to i times differentiability.

Finite-dimensional Feature Space Approximation For our methodology, we are interested in having a reduced rank representation for k_{GS} similar to Equation (10). Any consistent RFF approximation of in Equation (11) such that $g(x - x') \approx \phi_g^{(r)}(x)^\top \phi_g^{(r)}(x')$ where $\phi_g^{(r)}$ is an explicit feature mapping $\phi_g : X \to \mathbb{R}^r$, would results in a finite-dimensional feature space approximation for k_{GS} (Samo, 2017). In that case,

$$k_{GS}(x; x') \approx \sum_{k=1}^{K} h_k(x)^\top h_k(x')$$

with

$$h_k(x) = {}_k \phi_g^{(r)}(x - {}_k) \begin{matrix} \cos(! {}_k^\top x) \\ \sin(! {}_k^\top x) \end{matrix} \quad (12)$$

for $k = 1, \ldots, K$, where denotes the Kronecker product.

To be consistent with previous notations, we define $k_{GS}(x; x') \approx \phi^{(r)}(x)^\top \phi^{(r)}(x')$ where $\phi^{(r)}(x)$ is a feature mapping $\phi^{(r)}(x) : X \to \mathbb{R}^{4rK}$ satisfying

$$\phi^{(r)}(x) = [h_1(x)^\top, \ldots, h_K(x)^\top]^\top : \quad (13)$$

3.3 Sparse Spectral Permanental Process (SSPP)

Using RFFs for the GP approximation leads to a so-called sparse spectrum GP, first proposed by Lázaro-Gredilla et al.

(2010) in the context of GP regression. Sparse spectrum GPs are GPs de ned with the kernel induced by the feature map in Equations (10) or (13), $k^{(r)}(x; x') = \phi^{(r)}(x)^\top \phi^{(r)}(x')$.

The resulting approximate Gaussian process can be written in terms of a new r-size latent vector

$$f^{(r)}(x) \quad w^{(r)\top} \phi^{(r)}(x) \text{ where } w^{(r)} \quad N(0; I_r): \quad (14)$$

We de ne a permanental process with the spectral approximation of Equation (14), in which the intensity vector follows a similar linear form with feature vector given by Equations (10) or (13). We refer to it as the Sparse Spectral Permanental Process (SSPP) when using the feature map (10) or Generalized Sparse Spectral Permanental Process (GSSPP) when using the feature map (13).

In both cases, a tractable expression can be obtained for the integral term in the likelihood, now de ned as $\int_X (x)\,dx := \int_X (f(x) +)^2\,dx$, as follows:

Proposition 3.3. (with proof in parts C and D of the Appendix) Under the GP approximation of Equation (14), the integral expression $\int_X (x)\,dx$ can be expressed as

$$\int_X (x)\,dx = w^{(r)\top} M^{(r)} w^{(r)} + 2 \quad w^{(r)\top} m^{(r)} + \quad ^2 |X| \quad (15)$$

where $M^{(r)}$ is an $r \quad r$ matrix with $i; j$ entries de ned as

$$M^{(r)}_{i;j} := \int_{X^2} \phi^{(r)}_i(x) \phi^{(r)}_j(x)\,dx \quad (16)$$

for $i; j = 1; \quad ; r$ and $m^{(r)}$ is a r-vector with entries

$$m^{(r)}_i := \int_X \phi^{(r)}_i(x)\,dx: \quad (17)$$

for $i = 1; \quad ; r$. Final expressions for $M^{(r)}$ and $m^{(r)}$ are provided in the Appendix

The solution of Proposition 3.3 shares similarities with Warren et al. (2022) approach, who utilized Random Fourier Features (RFF) in Bayesian Quadrature (BQ). However, our result also covers the calculation of $\int f^2(x)dx$ and applies to the feature map of a generalized kernel, with the feature map of standard RFF being a speci c instance.

4 INFERENCE

Adopting the sparse spectral GP, $f^{(r)}$ assumes the linear form of Equation (14) for weight vector $w^{(r)}$ with independently-distributed standard Gaussian elements. Moreover, the integral in Equation (15) reduces to a quadratic form. We also de ne the model hyperparameters to be the parameters of the kernel function together with the offset term . More precisely, for SSPP, the hyperparameters consist of $= (\; ; `; \;)$, while for GSSPP, $:= (\; f \; _k g_{k=1}^K ; f \; ! _k g_{k=1}^K ; f \; _k g_{k=1}^K ; \;)$.

The (non-log) likelihood function in Equation (1) therefore becomes

$$\log p(X | w^{(r)};) = \quad w^{(r)\top} M^{(r)} w^{(r)} \quad 2 \quad w^{(r)\top} m^{(r)} \quad ^2 |X| + \sum_{i=1}^N \log(|w^{(r)\top} \phi^{(r)}(x_i) + \quad |^2): \quad (18)$$

where $M^{(r)}$ and $m^{(r)}$ are the matrix and vector terms from proposition 3.3. We can compute the log-likelihood function in $O(r^2 N)$, i.e. linearly in N. The log of the joint-distribution over $w^{(r)}$ and X is then

$$\log p(w^{(r)}; X |) = \log \quad p(X | w^{(r)};) + \log \quad p(w^{(r)}) = \log p(X | w^{(r)};) \quad \frac{1}{2} w^{(r)\top} w^{(r)} + C \quad (19)$$

for some constant C, where $p(w^{(r)}) = N(0; I_r)$ denotes the prior distribution over $w^{(r)}$.

4.1 Laplace Approximation

The latent posterior $p(w^{(r)} | X;)$ induced from Equation (19) is approximated using Laplace's method. A second order Taylor expansion of $\log p(w^{(r)} | X;)$ around the maximum of the posterior, yields a Gaussian approximation

$$p(w^{(r)} | X;) \quad N(w^{(r)} | \hat{w}^{(r)}; Q) := q(w^{(r)} | X;) \quad (20)$$

where $\hat{w}^{(r)} := \arg\max_{w^{(r)}} p(w^{(r)} | X;)$ is the mode of the latent posterior and Q is chosen to be the negative inverse Hessian of the true posterior at that point.

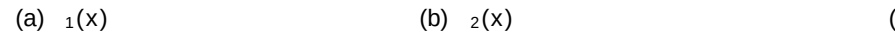
The gradient and the Hessian of the true posterior with respect to $w^{(r)}$ are

$$r_{w^{(r)}} \log p(w^{(r)} | X;) = \quad (2M^{(r)} + I_r) w^{(r)} \quad 2 \quad m^{(r)} + 2 \sum_{i=1}^N \frac{\phi^{(r)}(x_i)}{w^{(r)\top} \phi^{(r)}(x_i) +}$$

$$r^2_{w^{(r)}} \log p(w^{(r)} | X;) = \quad (2M^{(r)} + I_r) \quad 2 \sum_{i=1}^N \frac{\phi^{(r)}(x_i) \phi^{(r)}(x_i)^\top}{(w^{(r)\top} \phi^{(r)}(x_i) +)^2}:$$

The mode $\hat{w}^{(r)}$ must satisfy the stationary constraint

$$r_{w^{(r)}} \log p(w^{(r)} | X;) \; |_{w^{(r)} = \hat{w}^{(r)}} = 0;$$



(a) $\lambda_1(x)$ (b) $\lambda_2(x)$ (c) $\lambda_3(x)$

Figure 2: Mean predictive intensity of the three toy intensity functions λ_1, λ_2 and λ_3 defined as in Adams et al. (2009). Solid colored lines represent the predictive mean. The solid black lines shows the ground truth. The shaded areas are the 80% credible region of the SSPP model.

that implies

$$\hat{w}^{(r)} = \left(M^{(r)} + \frac{1}{2} I_r\right)^{-1} \left(\sum_{i=1}^{N} \frac{\phi^{(r)}(x_i)}{w^{(r)\top}\phi^{(r)}(x_i) + \epsilon} - m^{(r)}\right). \quad (21)$$

Equation (21) cannot be solved analytically. Instead, we estimate $\hat{w}^{(r)}$ iteratively using Newton-Raphson method, with step

$$w^{(r)\,new} = w^{(r)} - \left(\nabla^2_{w^{(r)}} \log p(w^{(r)})\right)^{-1} \nabla_{w^{(r)}} \log p(w^{(r)}). \quad (22)$$

The precision matrix is Q^{-1} is then given by $-\nabla^2_{w^{(r)}} \log p(w^{(r)} | X, \theta)\big|_{w^{(r)} = \hat{w}^{(r)}}$.

4.2 Model Selection

We first derive a marginal likelihood approximation similar to Walder and Bishop (2017, Section 4.1.6).

$$\begin{aligned}
&\log p(X|\theta) \\
&= \log p(\hat{w}^{(r)}, X|\theta) - \log p(\hat{w}^{(r)}|X, \theta) \\
&\quad - \log p(\hat{w}^{(r)}, X|\theta) + \log q(\hat{w}^{(r)}|X, \theta) \\
&= -\hat{w}^{(r)\top} M^{(r)} \hat{w}^{(r)} + 2\, \hat{w}^{(r)\top} m^{(r)} - \epsilon |X| \\
&\quad + \sum_{i=1}^{N} \log(|\hat{w}^{(r)\top}\phi^{(r)}(x_i) + \epsilon|) \\
&\quad - \frac{1}{2}\hat{w}^{(r)\top}\hat{w}^{(r)} + \frac{1}{2}\log|Q| + \frac{N}{2}\log(2\pi) \quad (23)
\end{aligned}$$

since the quadratic term of $\log q(\hat{w}^{(r)}|X, \theta)$ cancels out.

We tune the hyperparameters by maximizing Equation (23). The model selection is facilitated by the fact that the gradient of the marginal likelihood in Equation (23) with respect to θ can be easily expressed. The terms $M^{(r)}$, $m^{(r)}$ and $\phi^{(r)}(\cdot)$ are functions of the hyperparameters The mode $\hat{w}^{(r)}$ is also a function of θ.

The partial derivatives of the marginal likelihood with respect to θ_i is obtained using the chain rule,

$$\nabla_{\theta_i} \log p(X|\theta) = \frac{\partial \log p(X|\theta)}{\partial \theta_i}\Big|_{\text{explicit}} + \sum_{j=1}^{r} \frac{\partial \log p(X|\theta)}{\partial \hat{w}_j} \frac{\partial \hat{w}_j}{\partial \theta_i}. \quad (24)$$

Expressions for the terms $\frac{\partial \log p(X|\theta)}{\partial \theta_j}$, $\frac{\partial \log p(X|\theta)}{\partial \hat{w}_j}$ and $\frac{\partial \hat{w}^{(r)}}{\partial \theta_i}$ above are given in part F of the Appendix, requiring a full mode search within each iterative hyperparameters update. In the current work, we note that assuming $\frac{\partial \hat{w}^{(r)}}{\partial \theta_i} = 0$ and alternating independent updates for the mode in Equation (22) and the hyperparameters in Equation (24) provides faster and yet acceptable results. The algorithms are presented in part G of the Appendix.

5 PREDICTIVE DISTRIBUTION

To form predictive distributions, we assume that the latent posterior is approximated by $q(w^{(r)}|X, \theta)$ as in Equation (20).

5.1 Predictive Intensity Distribution

For some $x \in \mathcal{X}$, the predictive distribution of $f(x)$ can be deduced from Equations (14) and (20) to be

$$f(x)|X, \theta \sim \mathcal{N}(\mu(x), \sigma(x)) \quad (25)$$

where

$$\mu(x) := \hat{w}^{(r)\top}\phi^{(r)}(x) \quad (26)$$

and

$$\sigma(x) := \phi^{(r)}(x)^\top Q^{-1}\phi^{(r)}(x). \quad (27)$$

Given $\lambda(\cdot) = (f(\cdot) + \epsilon)^2$ and Equation (25), we can also derive the predictive distribution of the intensity function i.e.

$$\lambda(x)|X, \theta \sim \text{Gamma}(a, b) \quad (28)$$

with parameters a and b expressed in part E of the Appendix.

5.2 Predictive Expected Log-likelihood

For a training set $X = \{x_i\}_{i=1}^N$ and an held-out test set $X' = \{x'_i\}_{i=1}^{N'}$, we can derive from Equation (1), an approximation for the expected predictive log-likelihood

$$E[\log p(X' \mid X)] \approx E_{w^{(r)}} \left[\int_X (w^{(r)\top} \cdot \phi^{(r)}(x) + \beta)^2 dx \right] + \sum_{i=1}^{N'} \left[E_{w^{(r)}} \log(w^{(r)\top} \cdot \phi^{(r)}(x'_i) + \beta)^2) \right]$$

where $w^{(r)} \sim q(w^{(r)} \mid X; \theta)$. The expectation over the integral term can be solved analytically. The sum-of-expectation can be expressed using Pochhammer series, that we approximate in practise by interpolation of a look-up table of precomputed values. This is very similar to Lloyd et al. (2015, section 4.3). We provide more details in part E of the Appendix.

6 EXPERIMENTS

We benchmark the SSPP scheme introduced in Sections 4 and 5 against a Nyström-based implementation of the LBPP scheme of Walder and Bishop (2017), a frequentist kernel smoothing approach with edge correction (KS) (Diggle, 1985) and a variational inference scheme for point processes proposed by Lloyd et al. (2015), referred to as the Variational Bayesian Point Process (VBPP). We test the algorithms on three 1D synthetic data sets and three real data sets (one in 1D and two in 2D).

6.1 Benchmarks Settings

Our KS implementation uses standard kernel density estimation with truncated normal kernels to account for domain knowledge. The kernel bandwidth parameter is estimated via grid search using the leave-one-out log average likelihood objective of Lloyd et al. (2015). We used a publicly-available implementation of VBPP (https://github.com/st–/vbpp). We adopt fixed inducing points on a grid over $|X|$. For consistency, we also used a constant offset for both LBPP and VBPP implementations.

6.2 Performance Metrics

The average test expected log-likelihood $L_{\text{test}} := E[\log p(X' \mid X)]$ is used as an evaluation metric. This is generally difficult to compute for point process models, but is available for SSPP and LBPP (see Section 5). For the synthetic experiment we also consider the normalized L_2 norm to the known ground truth intensity function i.e. $L_2 := \frac{1}{|X|} \int_X E[(\lambda(x) - \lambda_{\text{truth}}(x))^2] dx \Big]^{\frac{1}{2}}$ where $\lambda(x)$ denotes the predictive intensity.

Table 1: Performance of GSSPP, SSPP, KS, VBPP and LBPP schemes on three samples of synthetic data. Values in bold-face refer to best performance, which corresponds to lower values of L_2, but higher values of L_{test}.

	$\lambda_1(x)$			$\lambda_2(x)$			$\lambda_3(x)$		
	L_2	L_{test}	time(s)	L_2	L_{test}	time(s)	L_2	L_{test}	time(s)
GSSPP-SE	**0.74**	**105.92**	1.32	**0.83**	**52.62**	1.59	**1.82**	**842.05**	2.94
GSSPP-m12	**0.71**	**104.23**	1.88	**0.86**	**48.58**	1.42	**1.98**	**838.17**	2.54
GSSPP-m32	**0.68**	**106.05**	1.79	**0.78**	**51.65**	1.56	**1.75**	**840.80**	2.53
GSSPP-m52	**0.69**	**106.12**	1.71	**0.84**	**52.77**	1.59	**1.68**	**841.77**	2.32
SSPP	**0.78**	**105.19**	**0.63**	**0.74**	**56.01**	**0.69**	**1.60**	**835.35**	**1.05**
KS	1.10	102.49	0.09	0.89	58.07	0.08	3.22	834.68	0.19
VBPP	0.72	104.65	1.98	0.85	51.11	1.15	1.63	838.65	1.52
LBPP	0.81	103.29	0.06	0.96	51.76	0.07	1.87	833.31	0.13

6.3 Synthetic Dataset

Three 1D simulated examples from Adams et al. (2009) are considered. The corresponding intensities are defined as $\lambda_1(x) = 2 \exp(-x/15) + \exp(-((x-25)/10)^2)$ on the interval $[0, 50]$ for approximately 47 events per sample, $\lambda_2(x) = 5 \sin(x^2) + 6$ on the interval $[0, 5]$ for approximately 36 events per sample and $\lambda_3(x)$ is a piecewise linear function shown in Figure 2 on the interval $[0, 100]$ for approximately 225 events per sample.

These intensity functions have been considered previously in the context of Gaussian Cox process by Samo and Roberts (2015b), Donner and Opper (2018), John and Hensman (2018) and Aglietti et al. (2019). We train the models on 10 independent samples generated from the ground truth, and evaluate the performance of each using 50 test sets sampled independently from the ground truth. We use the acronyms GSSPP-SE and GSSPP-(m) to refer to the generalized kernel variants of SSPP as in Equation (11) with ψ set to be the Gaussian kernel and Matérn kernel with parameter ν respectively.

We report optimal performance across models for sets of spectral points or inducing points of size (denoted by p) ranging from 15 to 100. Results are given in Table 1 and the mean predictive intensities displayed in Figure 2. GSSPP and SSPP outperform the other methods in terms of both L_{test} and L_2 for λ_1 and λ_2. Compared to LBPP, GSSPP and SSPP perform better consistently, but with slightly increase execution times. In a similar manner to findings for VBPP, GSSPP and SSPP fitting remains up to three orders of magnitude faster than alternative MCMC-based methods (Adams et al., 2009), see Lloyd et al. (2015) for comparison.

(a) GSSPP-m12 (b) SSPP (c) VBPP (d) LBPP

Figure 3: Heat map of the predictive mean intensity for the Taxi data set scaled to a unit square. The black dots are the input data points.

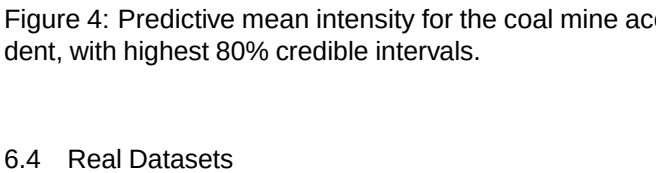


Figure 4: Predictive mean intensity for the coal mine accident, with highest 80% credible intervals.

6.4 Real Datasets

The classic coal mine accidents data set consists of the dates of 191 coal-mining accidents with fatalities in Britain between 15 March 1875 and 22 March 1962 (Jarrett, 1979). For this data set, we evaluate predictive performance for the competing inference schemes using 100 random partitions of the sample into train and test subsets (X_{train} and X_{test}) of approximately equal size. Figure 4 shows the predicted mean intensity with credible intervals. Results are presented in Table 2.

The bei data set is comprised of the locations of 3605 trees in the tropical rainforest on Barro Colorado Island (Hubbell and Foster, 1983). For these data, we evaluate predictive performance using 100 random partitions of the original sample into train and test subsamples of approximately equal size, with p now ranging from 15 to 150. Table 2 presents the results. Figure 6 in Appendix H, provides an illustration of a single fit. Additionally, Figure 5 in Appendix H shows the performances per number of spectral points or inducing points.

As a third dataset, we consider the Porto taxi dataset (Moreira-Matias et al., 2013) which contains 1.7×10^6 trajectories of taxi journeys in the years 2013-2014 in the Portuguese city of Porto. We consider the pick-up locations as observations of a point process. As in Aglietti et al. (2019), we restrict the analysis to 7000 events selected with (latitude, longitude) pairs bounded by $(41.147, -8.58)$ and $(41.18, -8.65)$. We select 1400 events at random as training set and use the rest as testing set. We set p ranging from 15 to 200. Table 2 presents the results. Figure 3 illustrates a single fit to the full data set for four models. Figure 5 in Appendix H shows the performances per number of spectral points or inducing points.

Table 2: Results on real-world data experiments with standard errors in brackets.

	Coal data (1D)		Bei data (2D)		Taxi data (2D)	
	L_{test}	time(s)	L_{test} [$\cdot 10^1$]	time(s)	L_{test} [$\cdot 10^2$]	time(s)
GSSPP-SE	224.44 (0.57)	1.56	763.49 (3.81)	20.48	278.54 (1.64)	23.91
GSSPP-m12	220.80 (0.85)	1.84	760.82 (4.31)	20.31	283.23 (1.11)	23.86
GSSPP-m32	224.25 (0.55)	1.58	764.73 (2.60)	20.85	280.48 (0.72)	19.60
GSSPP-m52	223.84 (0.54)	1.25	763.82 (1.00)	20.55	281.18 (1.43)	23.94
SSPP	221.23 (0.86)	0.64	751.50 (2.55)	17.54	268.32 (0.65)	12.13
KS	219.50 (0.33)	0.11	735.78 (1.49)	4.22	262.13 (0.26)	2.55
VBPP	221.19 (1.34)	1.75	757.95 (3.14)	28.41	281.02 (0.63)	20.45
LBPP	218.68 (0.87)	0.16	711.72 (1.35)	1.35	254.45 (0.17)	1.04

For each of the real-data applications, GSSPP performs best. The average fitting time using GSSPP is comparable to that of VBPP. Figure 4 shows the effect of different choice of GSSPP kernel function g. When choosing g to be a Matérn kernel, with parameter ν, spectral points z are drawn from a Student-t distribution with 2ν degree of freedom as discussed in part C of the Appendix. Compared to GSSPP-SE, these models tend to produce feature mappings with larger coefficients for trigonometric components; hence the resulting intensity appears less smooth.

7 CONCLUSION

We introduce a novel Bayesian framework to infer the intensity function of a permanental process. Our approach uses a Laplace-based inference exploiting generalized kernels and random Fourier features (RFFs). The approach requires no discretization of the domain, allows kernel designs, and provides better predictive accuracy than the alternative Laplace-based approach of Walder and Bishop (2017). The performance of our scheme also compares favorably with other standard methods on both real temporal and large spatial data sets.

Acknowledgements

We also thank the anonymous reviewers for useful comments during the review process. Jeremy Sellier was supported under an EPSRC CASE studentship award in conjunction with Shell Research Ltd.

References

Adams, R. P., Murray, I. and MacKay, D. J. (2009), Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities, in `Proceedings of the 26th Annual International Conference on Machine Learning', pp. 9–16.

Aglietti, V., Bonilla, E. V., Damoulas, T. and Cripps, S. (2019), Structured variational inference in continuous Cox process models, in `Advances in Neural Information Processing Systems', Vol. 32.

Banerjee, S., Carlin, B. P., Gelfand, A. E. and Banerjee, S. (2003), Hierarchical Modeling and Analysis for Spatial Data, Chapman and Hall/CRC. 1 st edition.

Bochner, S. (1932), `Vorlesungen über fouriersche integrale', Akademische Verlagsgesellschaf.

Cox, D. (1955), `Some statistical methods connected with series of events', Journal of the Royal Statistical Society Series B.17, 129–164.

Cunningham, J. P., Shenoy, K. V. and Sahani, M. (2008), Fast Gaussian process methods for point process intensity estimation, in `Proceedings of the 25th International Conference on Machine Learning', ICML '08, Association for Computing Machinery, p. 192–199.

Cunningham, J. P., Yu, B. M., Shenoy, K. V. and Sahani, M. (2008), Inferring neural firing rates from spike trains using gaussian processes, in `Advances in Neural Information Processing Systems', Vol. 20.

Daley, D. J. and Vere-Jones, D. (2003), An introduction to the theory of point processes. Vol. I. Probability and its Applications, Springer, New York.

Diggle, P. J. (1985), `A Kernel method for smoothing point process data', Journal of The Royal Statistical Society Series C-applied Statistics 34, 138–147.

Diggle, P. J. (2003), Statistical analysis of spatial point patterns, Hodder Education. 2nd edition.

Diggle, P. J., Moraga, P., Rowlingson, B. and Taylor, B. M. (2013), `Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm', Statistical Science 28(4), 542 – 563.

Donner, C. and Opper, M. (2018), `Ef cient Bayesian inference of Sigmoidal Gaussian Cox processes', J. Mach. Learn. Res. 19(1), 2710–2743.

Duvenaud, D. K., Nickisch, H. and Rasmussen, C. (2011), Additive Gaussian processes, in `Advances in Neural Information Processing Systems', Vol. 24, Curran Associates, Inc.

Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B. and Ghahramani, Z. (2013), Structure discovery in nonparametric regression through compositional kernel search, in `Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28', ICML'13.

Flaxman, S., Chirico, M., Pereira, P. and Loef er, C. (2019), `Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: a winning solution to the NIJ "real-time crime forecasting challenge"', Annals of Applied Statistics.

Flaxman, S., Teh, Y. W. and Sejdinovic, D. (2017), `Poisson intensity estimation with reproducing kernels', Electronic Journal of Statistics 11(2), 5081–5104.

Flaxman, S., Wilson, A., Neill, D., Nickisch, H. and Smola, A. (2015), Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods, in `Proceedings of the 32nd International Conference on Machine Learning', Vol. 37 of Proceedings of Machine Learning Research, pp. 607–616.

Gardner, J. K. and Knopoff, L. (1974), `Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?', Bulletin of the Seismological Society of America 64(5), 1363–1367.

Grubesic, T. H. and Mack, E. A. (2008), `Spatio-temporal interaction of urban crime', Journal of Quantitative Criminology 24, 285–306.

Gunter, T., Lloyd, C., Osborne, M. and Roberts, S. (2014), `Ef cient Bayesian nonparametric modelling of structured point processes', Uncertainty in Arti cial Intelligence - Proceedings of the 30th Conference, UAI 2014

Hu, S. and Papageorgiou, N. (2018), Handbook of Multivalued Analysis: Volume I: Theory, Mathematics and Its Applications, Springer US.

Hubbell, S. and Foster, R. (1983), `Diversity of canopy trees in a neotropical forest and implications for conservation', Tropical Rain Forest: Ecology and Management 11, 25–41.

Illian, J. B., Sørbye, S. H. and Rue, H. (2012), `A toolbox for tting complex spatial point process models using integrated nested Laplace approximation (INLA)', The Annals of Applied Statistics 6(4), 1499–1530.

Jarrett, R. G. (1979), `A note on the intervals between coal-mining disasters', Biometrika 66(1), 191–193.

John, S. and Hensman, J. (2018), Large-scale Cox process inference using variational Fourier features, in `International Conference on Machine Learning', pp. 2362–2370.

Ko, Y.-J. and Seeger, M. W. (2016), Expectation propagation for recti ed linear poisson regression, in `Asian Conference on Machine Learning', Vol. 45 of Proceedings of Machine Learning Research, Hong Kong, pp. 253–268.

Lázaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E. and Figueiras-Vidal, A. R. (2010), `Sparse spectrum Gaussian process regression', The Journal of Machine Learning Research 11, 1865–1881.

Lian, W., Henao, R., Rao, V., Lucas, J. p. and Carin, L. (2015), A multitask point process predictive model, in `Proceedings of the 32nd International Conference on Machine Learning', Vol. 37 of ICML'15, pp. 2030–2038.

Lloyd, C., Gunter, T., Osborne, M. and Roberts, S. (2015), Variational inference for Gaussian process modulated Poisson processes, in `International Conference on Machine Learning', pp. 1814–1822.

Lopez-lopera, A. F., John, S. and Durrande, N. (2019), Gaussian process modulated cox processes under linear inequality constraints, in `Proceedings of the Twenty-Second International Conference on Arti cial Intelligence and Statistics', Vol. 89 of Proceedings of Machine Learning Research, pp. 1997–2006.

McCullagh, P. and Møller, J. (2006), `The permanental process', Advances in Applied Probability 38(4), 873–888.

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J. and Damas, L. (2013), `Predicting taxi–passenger demand using streaming data', IEEE Transactions on Intelligent Transportation Systems 14(3), 1393–1402.

Møller, J., Syversveen, A. R. and Waagepetersen, R. P. (1998), `Log Gaussian Cox processes', Scandinavian Journal of Statistics 25(3), 451–482.

Park, M., Weller, J., Horwitz, G. and Pillow, J. (2014), `Bayesian active learning of neural ring rate maps with transformed gaussian process priors', Neural computation 26, 1–23.

Plataniotis, K. N. and Hatzinakos, D. (2001), Gaussian Mixtures and their Applications to Signal Processing, Vol. Advanced Signal Processing Handbook, CRC Press, Springer-Verlag.

Rahimi, A. and Recht, B. (2007), Random features for large-scale kernel machines, in `Proceedings of the 20th International Conference on Neural Information Processing Systems', p. 1177–1184.

Rasmussen, C. E. and Williams, C. K. I. (2006), Gaussian processes for machine learning, MIT Press.

Samo, Y.-L. K. (2017), Advances in kernel methods : towards general-purpose and scalable models, PhD thesis, University of Oxford.

Samo, Y.-L. K. and Roberts, S. (2015), `Generalized spectral kernels', arxiv:1506.02236

Samo, Y.-L. K. and Roberts, S. (2015), Scalable nonparametric Bayesian inference on point processes with Gaussian processes, in `Proceedings of the 32th International Conference on Machine Learning', Vol. 37 of ICML'15, pp. 2227–2236.

Seeger, M. and Bouchard, G. (2012), Fast variational bayesian inference for non-conjugate matrix factorization models, in `Proceedings of the 15th International Conference on Arti cial Intelligence and Statistics', Vol. 22 of Proceedings of Machine Learning Research, pp. 1012–1018.

Stein, M. L. (1999), Interpolation of spatial data, Springer Series in Statistics, Springer-Verlag, New York.

Titsias, M. (2009), Variational learning of inducing variables in sparse Gaussian processes, in `Proceedings of the 20th International Conference on Arti cial Intelligence and Statistics', Vol. 5 of Proceedings of Machine Learning Research, pp. 567–574.

Walder, C. J. and Bishop, A. N. (2017), Fast Bayesian intensity estimation for the permanental process, in `Proceedings of the 34th International Conference on Machine Learning', Vol. 70 of Proceedings of Machine Learning Research, pp. 3579–3588.

Warren, H., Oliveira, R. and Ramos, F. (2022), Generalized bayesian quadrature with spectral kernels, in `Proceedings of the 38th Conference on Uncertainty in Arti cial Intelligence', Vol. 180 of Proceedings of Machine Learning Research, pp. 2085–2095.

Wilson, A. and Adams, R. (2013), Gaussian process kernels for pattern discovery and extrapolation, in `Proceedings of the 30th International Conference on Machine Learning', pp. 1067–1075.

A LBPP WITH NYSTRÖM METHOD

We provide a quick review of the LBPP model using Nyström method proposed by Walder and Bishop (2017). We also provide in Proposition A.1, an expression for the integral term $\int_{\mathcal{X}} f(x)^2 dx$ under Nyström approximation that is not directly available in Walder and Bishop (2017).

A.1 Method Review

The Nyström method (see Rasmussen and Williams, 2005, Chap. 4.3, 8) provides a reduced-rank approximation for $n \ll N$ data points $X^{(n)} = \{x_i^{(n)}\}_{i=1}^n$ sampled uniformly from the original data. The eigenvalues and eigenfunctions in Equation (3) of the main text are approximated using the eigenvectors $u_i^{(n)}$ and eigenvalues $\lambda_i^{(n)}$ of $K_{n,n}$, the Gram matrix with i,j entry $k(x_i^{(n)}, x_j^{(n)})$. Thus, f can be approximated by

$$f(x) \approx \sum_{i=1}^{n} \hat{\lambda}_i^{\frac{1}{2}} w_i^{(n)} \hat{\phi}_i(x) \tag{29}$$

where $w^{(n)} = (w_1^{(n)}, \dots, w_n^{(n)})^\top \sim \mathcal{N}(0, I_n)$ and

$$\hat{\lambda}_i := \frac{N}{n} \lambda_i^{(n)} \tag{30}$$

$$\hat{\phi}_i(\cdot) := \sqrt{\frac{n}{N}} \frac{1}{\lambda_i^{(n)}} k(\cdot, X^{(n)}) u_i^{(n)}. \tag{31}$$

Substituting Equations (30) and (31) into Equation (29) yields a linear Gaussian latent formulation as follows

$$\begin{aligned} f(x) &\approx k(x, X^{(n)}) \sum_{i=1}^{n} \sqrt{\frac{N}{n}} \frac{w_i^{(n)}}{\lambda_i^{(n)}} u_i^{(n)} \\ &= k(x, X^{(n)}) U^{(n)} \Lambda^{(n)\,-\frac{1}{2}} w^{(n)} \\ &:= w^{(n)\top} \psi^{(n)}(x) \end{aligned} \tag{32}$$

and $\psi^{(n)}(\cdot) := \left[k(\cdot, X^{(n)}) U^{(n)} \Lambda^{(n)\,-\frac{1}{2}} \right]^\top$ denotes the new features vector.

A.2 Integral Calculation

In Proposition A.1 we express the integral term $\int_{\mathcal{X}} f(x)^2 dx$ under the Nyström approximation both because it is not available in Walder and Bishop (2017) and to demonstrate the similarities with the corresponding derivation of our proposed method in Proposition 3.3.

Proposition A.1. Under the GP approximation (32), the integral expression $\int_{\mathcal{X}} f(x)^2 \, dx$ can be written as

$$\int_{\mathcal{X}} f(x)^2 \, dx = w^{(n)\top} M^{(n)} w^{(n)}$$

where $M^{(n)}$ is a $n \times n$ matrix defined as

$$M^{(n)} := \Lambda^{(n)\,-\frac{1}{2}} \left[U^{(n)\top} \Lambda^{(n)} U^{(n)} \right] \Lambda^{(n)\,-\frac{1}{2}}$$

and $\Lambda^{(n)}$ is a $n \times n$ matrix given in the proof below.

Proof. Let f be approximated by the Nyström-based approach defined in Equation (32) i.e. $f(x) \approx w^{(n)\top} \psi^{(n)}(x)$. The

integral expression $\int_X (x)^2\,dx$ can be written as

$$\begin{aligned}
\int_X f(x)^2 dx &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_i^{(n)} w_j^{(n)} \int_X \phi_i^{(n)}(x)\,\phi_j^{(n)}(x)\,dx \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_i^{(n)} w_j^{(n)}}{\sqrt{\lambda_i^{(n)}\lambda_j^{(n)}}} \int_X k(x;X^{(n)})u_i^{(n)}\; k(x;X^{(n)})u_j^{(n)}\,dx \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{w_i^{(n)} w_j^{(n)}}{\sqrt{\lambda_i^{(n)}\lambda_j^{(n)}}}\, u_i^{(n)\top} \int_X k(X^{(n)};x)k(x;X^{(n)})\,dx\; u_j^{(n)} \\
&= w^{(n)\top} \underbrace{\left[\tfrac{1}{2}U^{(n)\top}\right]\Lambda^{(n)}\left[U^{(n)}\tfrac{1}{2}\right]}_{:=M^{(n)}} w^{(n)}
\end{aligned}$$

where $\Psi^{(n)} = \int_X k(X^{(n)};x)k(x;X^{(n)})\,dx$ is the integral statistic already defined in Lloyd et al. (2015) and John and Hensman (2018). In particular, for the separable Gaussian kernel defined in Equation (35),

$$\begin{aligned}
\Psi_{i,j} &= \int_{\mathcal{X}} \prod_{k=1}^{d} \exp\left(\frac{(x_{k;i}^{(n)} - x_{k;j}^{(n)})^2}{4\ell_k^2}\right) \exp\left(\frac{(x_k - x_{k;i,j}^{(n)})^2}{\ell_k^2}\right) dx \\
&= \prod_{k=1}^{2} \frac{\ell_k\sqrt{\pi}}{d} \exp\left(\frac{(x_{k;i}^{(n)} - x_{k;j}^{(n)})^2}{4\ell_k^2}\right) \left[\operatorname{erf}\left(\frac{x_{k;i,j}^{(n)} - \mathcal{X}_k^{\min}}{\ell_k}\right) - \operatorname{erf}\left(\frac{x_{k;i,j}^{(n)} - \mathcal{X}_k^{\max}}{\ell_k}\right)\right]
\end{aligned}$$

where ℓ and $\boldsymbol{\ell} := (\ell_1, \ell_2)$ are respectively the scaling and length-scale parameters of the covariance function, $x_{k,i}^{(n)}$ is the kth coordinate of the ith Nyström-sampled point $x_i^{(n)}$ and $x_{k;i;j}^{(n)} := (x_{k;i}^{(n)} + x_{k;j}^{(n)})/2$.

Offset Term Adding an offset term α to the intensity i.e $\lambda(\cdot) = (f(\cdot) + \alpha)^2$ yields

$$\int_X (f(x) + \alpha)^2\,dx = \int_X f(x)^2 dx + 2\alpha \int_X f(x)\,dx + \alpha^2 |X|$$

with

$$\int_X f(x)\,dx \int_X k(x;X^{(n)})\,dx \quad U^{(n)} \tfrac{1}{2} w^{(n)} = \Psi^{(n)\top} U^{(n)} \Psi^{(n)} \tfrac{1}{2} w^{(n)}$$

where $\Psi^{(n)} := \int_X k(X^{(n)};x)\,dx$ is an n-vector such that in the separable Gaussian kernel case above, we have

$$\Psi_i^{(n)} = \prod_{d=1}^{2} \frac{\ell_d\sqrt{\pi}}{2} \left[\operatorname{erf}\left(\frac{x_{d;i}^{(n)} - \mathcal{X}_d^{\max}}{\ell_d\sqrt{2}}\right) - \operatorname{erf}\left(\frac{x_{d;i}^{(n)} - \mathcal{X}_d^{\min}}{\ell_d\sqrt{2}}\right)\right] :$$

□

A.3 Inference with Laplace Approximation

For a fixed set of GP hyperparameters (i.e. parameters of the covariance function), the intensity function is determined by the latent n-vector $w^{(n)}$. Walder and Bishop (2017) use a Laplace approximation for a non-Gaussian posterior $p(w^{(n)}|X;\cdot)$. The latent posterior is approximated by $p(w^{(n)}|X;\cdot) \approx \mathcal{N}(w^{(n)}|\hat{w}^{(n)}, Q)$ where $\hat{w}^{(n)}$ is the mode of the posterior and Q is chosen to be the inverse Hessian of the true posterior at $\hat{w}^{(n)}$.

B COMMON KERNELS SPECTRAL DENSITIES

We present in Table (3) below some common stationary distance-dependent kernels mentionned in the main text and their corresponding spectral densities $S(\cdot)$.

Table 3: Stationary distance-dependent kernels and their duals, where $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is a modified Bessel function.

NAME	KERNEL FUNCTION $k(\tau)$	SPECTRAL DENSITY $S(z)$
Gaussian	$\exp\left(-\frac{\|\tau\|^2}{2}\right)$	$\frac{\prod_{i=1}^{d}\ell_i}{(2\pi)^{d/2}}\exp\left(-\frac{\|z\ell\|^2}{2}\right)$
Matérn(ν)	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\tau\right)^{\nu} K_\nu\left(\sqrt{2\nu}\tau\right)$	$\frac{\Gamma(\nu+d/2)(2\nu)^{\nu}}{\Gamma(\nu)\pi^{d/2}}\left(1+\frac{1}{2\nu}\|z\|^2\right)^{-(\nu+d/2)}$

C PROOF OF PROPOSITION (3.3) : Integral Expression via RFF

Let f be approximated by a RFF-based approach as defined in Equation (14) i.e. $f(x) \approx w^{(r)\top}\varphi^{(r)}(x)$ where the feature map $\varphi^{(r)}$ follows Equation (9).

C.1 Real Valued Feature Mapping

We first detail the derivation of the real valued Fourier features described in Equation (10) of the main text. The imaginary part of Equation (9) of the main text can be discarded as follows

$$
\begin{aligned}
k(x - x') &= \sigma^2 \mathbb{E}_z\left[\exp(-iz^\top(x-x'))\right] \\
&= \sigma^2 \mathbb{E}_z\left[\cos(z^\top(x-x')) + i\sin(z^\top(x-x'))\right] \qquad (33)\\
&= \sigma^2 \mathbb{E}_z\left[\cos(z^\top(x-x'))\right] \\
&= \sigma^2 \mathbb{E}_z\left[\cos(z^\top x)\cos(z^\top x') + \sin(z^\top x)\sin(z^\top x')\right] \qquad (34)\\
&\approx \frac{\sigma^2}{r}\sum_{i=1}^{r}\cos(z_i^\top x)\cos(z_i^\top x') + \sin(z_i^\top x)\sin(z_i^\top x') \\
&= \frac{\sigma^2}{r}\varphi^{(r)}(x)^\top\varphi^{(r)}(x')
\end{aligned}
$$

where $z_1, \dots, z_r$ are independent samples with density $S(z)$ and the explicit feature mapping $\varphi^{(r)}(\cdot)$ is defined as

$$
\varphi^{(r)}(x) := \sqrt{\frac{\sigma^2}{r}}\begin{bmatrix}\cos(z_1^\top x)\\ \vdots \\ \cos(z_r^\top x)\\ \sin(z_1^\top x)\\ \vdots \\ \sin(z_r^\top x)\end{bmatrix}.
$$

Gaussian Kernel Specifically, without approximation, for a Gaussian kernel k_g with $\mathcal{X} = \mathbb{R}^d$ and where

$$
k_g(x - x') = \sigma^2\prod_{i=1}^{d}\exp\left(-\frac{(x_i - x_i')^2}{2\ell_i^2}\right) \qquad (35)
$$

with scaling parameter and length-scale vector $\ell = [\ell_1, \dots, \ell_d]^\top$, the corresponding spectral density $S(z)$ is a multivariate normal $\mathcal{N}(0, \mathrm{diag}(\Lambda))$ with $\Lambda := [1/\ell_1^2, \dots, 1/\ell_d^2]^\top$.

Matérn Kernel For a Matérn class of kernel function k_m such that

$$k_m(x - x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}(x - x')}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}(x - x')}{\ell} \right) \tag{36}$$

where $\nu \in \mathbb{R}^+$, $\ell \in \mathbb{R}^+$, $\sigma^2 \in \mathbb{R}^+$ and K_ν is a modified Bessel function, the corresponding spectral density $S(z)$ is a d dimension multivariate Student-t distribution $\mathcal{MVT}(0, \Sigma, 2\nu)$ with covariance function $\Sigma = (1/\ell) I_d$ and degree of freedom 2ν. The spectral locations z are sampled as

$$z = \sqrt{u/2\nu}\, G \quad \text{where} \quad u \sim \chi^2(2\nu) \tag{37}$$

and G is a $d \times r$ matrix of of i.i.d. standard normal random variables.

C.2 Integral Calculation

We now detail the integral expression of proposition (3.3) for the real valued Fourier features in Equation (10) of the main text. We consider without loss of generality the spatial case where $X = [-a, a]^2$. $x_{d,i}$ refers to the dth coordinate of the ith training input x_i for $i = 1, \ldots, N$ and $z_{d,i}$ to the dth coordinate of the ith spectral point z_i for $i = 1, \ldots, r$. The integral of f over X becomes

$$\int_{[-a,a]^2} f(x)^2 dx = \sum_{i,j} w_i^{(r)} w_j^{(r)} \int_{[-a,a]^2} \varphi_i^{(r)}(x) \varphi_j^{(r)}(x)\, dx \tag{38}$$

where

$$\varphi_i^{(r)}(x) \varphi_j^{(r)}(x) = \frac{2}{r} \begin{cases} \cos(z_i^\top x) \cos(z_j^\top x) & \text{if } (i,j) \in [1,r]^2 \\ \sin(z_i^\top x) \sin(z_j^\top x) & \text{if } (i,j) \in [r+2, 2r]^2 \\ \cos(z_i^\top x) \sin(z_j^\top x) & \text{if } (i,j) \in [1,r] \times [r+2, 2r] \\ \sin(z_i^\top x) \cos(z_j^\top x) & \text{if } (i,j) \in [r+2, 2r] \times [r, r]. \end{cases} \tag{39}$$

Thus, $\int_X f(x)^2 dx = w^\top M^{(r)} w$, where $M^{(r)}$ is the matrix with i,j entry obtained by integrating Equation (39). The 'cos', 'sin' and 'cos-sin' expressions can be written as

$$\int_{[-a,a]^2} \cos(z_i^\top x) \cos(z_j^\top x)\, dx = \frac{1}{2} \int_{[-a,a]^2} \cos((z_i - z_j)^\top x) + \cos((z_i + z_j)^\top x)\, dx$$

$$\int_{[-a,a]^2} \sin(z_i^\top x) \sin(z_j^\top x)\, dx = \frac{1}{2} \int_{[-a,a]^2} \cos((z_i - z_j)^\top x) - \cos((z_i + z_j)^\top x)\, dx$$

and

$$\int_{[-a,a]^2} \cos(z_i^\top x) \sin(z_j^\top x)\, dx = \frac{1}{2} \int_{[-a,a]^2} \sin((z_i - z_j)^\top x) + \sin((-z_i + z_j)^\top x)\, dx = 0.$$

Thus, since the off-diagonal blocks of $M^{(r)}$ are null, we can rewrite Equation (38) as

$$\begin{aligned} \int_X f(x)^2 dx &= \frac{2}{r} \left[w_{:r}^{(r)\top} (A + B) w_{:r}^{(r)} + w_{r:}^{(r)\top} (A - B) w_{r:}^{(r)} \right] \\ &= \frac{2}{r} w^{(r)\top} \begin{bmatrix} D_l^\top (A + B) D_l & D_l^\top (A - B) D_r \\ \cdots & \cdots \end{bmatrix} w^{(r)} \end{aligned} \tag{40}$$

where $w_{:r}^{(r)} := [w_1^{(r)}, \ldots, w_r^{(r)}]^\top$, $w_{r:}^{(r)} := [w_{r+1}^{(r)}, \ldots, w_{2r}^{(r)}]^\top$, $D_l := \begin{bmatrix} I_r \\ 0 \end{bmatrix}$, $D_r := \begin{bmatrix} 0 \\ I_r \end{bmatrix}$ and A and B are two $r \times r$ matrices defined as

$$A_{i,j} = \frac{1}{2} \int_{[-a,a]^2} \cos((z_i - z_j)^\top x)\, dx \quad \text{and} \quad B_{i,j} = \frac{1}{2} \int_{[-a,a]^2} \cos((z_i + z_j)^\top x)\, dx.$$

A and B can be evaluated as follows

Case 1: $z_i \neq z_j$ We define again $z_{d;i;j} := z_{d;i} + z_{d;j}$. Then,

$$A_{i;j} = \frac{\cos[a(z_{1;i;j} - z_{2;i;j})] - \cos[a(z_{1;i;j} + z_{2;i;j})]}{z_{1;i;j}\, z_{2;i;j}} = \frac{2}{z_{1;i;j}\, z_{2;i;j}} (\sin[az_{1;i;j}] \sin[az_{2;i;j}]) :$$
$$B_{i;j} = \frac{\cos[a(z_{1;i;j} - z_{2;i;j})] - \cos[a(z_{1;i;j} + z_{2;i;j})]}{z_{1;i;j}\, z_{2;i;j}} = \frac{2}{z_{1;i;j}\, z_{2;i;j}} (\sin[az_{1;i;j}] \sin[az_{2;i;j}]) :$$

Case 2: $z_i = z_j$

$$A_{i;i} = 2a^2 :$$
$$B_{i;i} = \frac{\cos[2a(z_{1;i} - z_{2;i})] - \cos[2a(z_{1;i} + z_{2;i})]}{4 z_{1;i}\, z_{2;i}} = \frac{1}{2 z_{1;i}\, z_{2;i}} (\sin[2az_{1;i}] \sin[2az_{2;i}]) :$$

Offset Term For the offset term, we need to compute the integral of f that is obtained from

$$\int_{[-a;a]^2} f(x)dx = \frac{p}{r} \sum_{i=1}^{r} w_i^{(r)} \int_{[-a;a]^2} \cos(z_i^\top x)\, dx + \frac{p}{r} \sum_{i=r+1}^{2r} w_i^{(r)} \underbrace{\int_{[-a;a]^2} \sin(z_i^\top x)\, dx}_{=0}$$
$$= \frac{p}{r} w^{(r)\top} m^{(r)}$$

where m is a r-vector such that

$$m_i^{(r)} = \frac{2\cos[a(z_{1;i} - z_{2;i})] - 2\cos[a(z_{1;i} + z_{2;i})]}{z_{1;i}\, z_{2;i}} = \frac{4}{z_{1;i}\, z_{2;i}} (\sin[az_{1;i}] \sin[az_{2;i}]) :$$

D PROOF OF PROPOSITION (3.3) : Integral Expression for Generalized Kernel

We assume k_{GS} to be a Generalized kernel given in Equation (11) in the main text, with a kernel g that admits a consistent RFF representation such that $g(x - x') \approx \varphi_g^{(r)}(x)^\top \varphi_g^{(r)}(x')$ where $\varphi_g^{(r)}$ is an explicit feature mapping $g_g : X \to \mathbb{R}^r$.

D.1 Real Valued Feature Mapping

The Generalized kernel k_{GS} becomes

$$k_{GS}(\cdot) \approx \sum_{k=1}^{K} \frac{2}{k} \varphi_g^{(r)}(x - \cdot_k)^\top \varphi_g^{(r)}(x' - \cdot_k)\, \cdot_k(x^\top \cdot\!\!\cdot_k)^\top \cdot_k(x'^\top \cdot\!\!\cdot_k)$$

where $\cdot_k(x)$ is a map $\cdot_k : X \to \mathbb{R}^2$ such that $\cdot_k(x) = \begin{pmatrix} \cos(x^\top \cdot\!\!\cdot_k) \\ \sin(x^\top \cdot\!\!\cdot_k) \end{pmatrix}$ for $k = 1; \dots; K$ and $\forall x \in X$ so that $\cdot_k(x^\top \cdot\!\!\cdot_k)\, \cdot_k(x'^\top \cdot\!\!\cdot_k) = \cos((x - x')^\top \cdot\!\!\cdot_k)$ for all $\cdot\!\!\cdot_k \in \mathbb{R}^d$ and $\forall x; x' \in X$. Thus,

$$k_{GS}(x; x') \approx \sum_{k=1}^{K} h_k(x)^\top h_k(x')$$

with

$$h_k(x) = \cdot_k \varphi_g^{(r)}(x - \cdot_k) \begin{pmatrix} \cos(\cdot\!\!\cdot_k^\top x) \\ \sin(\cdot\!\!\cdot_k^\top x) \end{pmatrix}$$

for $k = 1; \dots; K$, where $\otimes$ denotes the Kronecker product.

In particular, when $\varphi_g^{(r)}$ follows Equation (10) in the main text,

$$h_k(x) = \frac{2}{p} \frac{k}{r} \begin{pmatrix} \cos(z_1^\top (x - \cdot_k)) \begin{pmatrix} \cos(\cdot\!\!\cdot_k^\top x) \\ \sin(\cdot\!\!\cdot_k^\top x) \end{pmatrix} \\ \vdots \\ \sin(z_r^\top (x - \cdot_k)) \begin{pmatrix} \cos(\cdot\!\!\cdot_k^\top x) \\ \sin(\cdot\!\!\cdot_k^\top x) \end{pmatrix} \end{pmatrix} \tag{41}$$

where $z_1, \ldots, z_r$ are independent samples from $S_g(z)$ the spectral density of g.

The resulting approximate Gaussian process with generalized kernel can be written, in terms of a $4Kr$-size latent vector as follow

$$f(x) \approx w^{(r)\top} \phi^{(r)}(x) \quad \text{with } w^{(r)} \sim N(0, I_{4Kr}):$$

where

$$\phi^{(r)}(x) = [h_1(x)^\top, \ldots, h_K(x)^\top]^\top$$

with $h_k(x)$ defined as in Equation (41) for $k = 1, \ldots, K$.

D.2 Integral Calculation

The integral of f^2 over X becomes

$$\int_{[-a,a]^2} f(x)^2 dx = \sum_{i,j} w_i^{(r)} w_j^{(r)} \int_{[-a,a]^2} \phi_i^{(r)}(x) \phi_j^{(r)}(x)\, dx \tag{42}$$

Thus, $\int_{[-a,a]} f(x)^2 dx = w^{(r)\top} M^{(r)} w^{(r)}$, where $M^{(r)}$ is the matrix with i,j entry obtained by integrating Equation (42). The computation of $M^{(r)}$ can be split into different cases expressed below as `cos', `sin' and `cos-sin' terms.

Cos Terms The `cos' i,j terms can be written as

$$\begin{aligned} &\int_{[-a,a]^2} \cos(z_i^\top (x - \mu_i)) \cos(x^\top \omega_i)) \cos(z_j^\top (x - \mu_j)) \cos(x^\top \omega_j))\, dx \\ &= \frac{1}{8} \sum_{k=1}^{8} \int_{[-a,a]^2} \cos(x^\top \alpha_{i,j}^{(k)})\, dx \end{aligned} \tag{43}$$

where

$$\begin{aligned} \alpha_{i,j}^{(1)} &= (z_i - \mu_i) + \omega_i + (z_j - \mu_j) + \omega_j; & \alpha_{i,j}^{(5)} &= (z_i - \mu_i) - \omega_i + (z_j - \mu_j) - \omega_j; \\ \alpha_{i,j}^{(2)} &= (z_i - \mu_i) + \omega_i + (z_j - \mu_j) - \omega_j; & \alpha_{i,j}^{(6)} &= (z_i - \mu_i) - \omega_i + (z_j - \mu_j) + \omega_j; \\ \alpha_{i,j}^{(3)} &= (z_i - \mu_i) + \omega_i - (z_j - \mu_j) - \omega_j; & \alpha_{i,j}^{(7)} &= (z_i - \mu_i) - \omega_i - (z_j - \mu_j) + \omega_j; \\ \alpha_{i,j}^{(4)} &= (z_i - \mu_i) + \omega_i - (z_j - \mu_j) + \omega_j; & \alpha_{i,j}^{(8)} &= (z_i - \mu_i) - \omega_i - (z_j - \mu_j) - \omega_j: \end{aligned}$$

Integrating the left hand integrants in Equation (43) yields

$$\int_{[-a,a]^2} \cos(x^\top \alpha_{i,j}^{(k)})\, dx = \begin{cases} 2a^2; & \text{if } i = j \text{ and } k \in \{3, 7\} \\ \frac{1}{2} \frac{1}{\alpha_{1;i,j}^{(k)} \alpha_{2;i,j}^{(k)}} \sin[a\, \alpha_{1;i,j}^{(k)}] \sin[a\, \alpha_{2;i,j}^{(k)}]; & \text{otherwise} \end{cases} : \tag{44}$$

Sin Terms The `sin' i,j terms are

$$\begin{aligned} &\int_{[-a,a]^2} \sin(x^\top (z_i - \mu_i)) \sin(x^\top \omega_i)) \sin(x^\top (z_j - \mu_j)) \sin(x^\top \omega_j))\, dx \\ &= \frac{1}{8} \sum_{k=1}^{8} (-1)^k \int_{[-a,a]^2} \cos(x^\top \alpha_{i,j}^{(k)})\, dx \end{aligned} \tag{45}$$

The left hand integrants in Equation (45) integrate alike to Equation (44) up to $(-1)^k$ factor.

Cos-sin Terms The `cos-sin'$i; j$ terms can be evaluated as follows

$$\int_{[-a;a]^2} \sin(x^\top (z_i - \cdot_i)) \sin(x^\top \cdot\!! \cdot_i)) \cos(x^\top (z_j - \cdot_j)) \cos(x^\top \cdot\!! \cdot_j)) \, dx$$
$$= \frac{1}{8} \sum_{k=1}^{8} (-1)^{m(k)} \int_{[-a;a]^2} \cos(x^\top \cdot_{i;j}^{(k)}) \, dx \qquad (46)$$

where $m(k) = 1$ if $k = 1; \ldots; 4$ and 0 else. The left hand integrants in Equation (46) integrate alike to Equation (44) up to a $m(k)$ factor.

The remaining terms, yields sums of integrals of the type $\int_{[-a;a]} \sin(x^\top \cdot) dx$ with $\cdot \in \mathbb{R}^d$, that equal zero.

Offset Term For the offset term , we need to compute the integral of, that is $\int f(x) dx = w^{(r)\top} m^{(r)}$ where $m^{(r)}$ is a $4^K r$ -vector such that

$$m_i^{(r)} = \int_{[-a;a]^2} \cdot_i^{(r)}(x) \, dx$$

The computation of $m^{(r)}$ can be split into two cases : the `cos' terms

$$\int_{[-a;a]^2} \cos(z_i^\top (x - \cdot_i)) \cos(x^\top \cdot\!! \cdot_i)) \, dx =$$
$$\frac{1}{\cdot_{1;i}^{(1)} \cdot_{2;i}^{(1)}} \sin[a(\cdot_{1;i}^{(1)})] \sin[a(\cdot_{1;i}^{(1)})] + \frac{1}{\cdot_{1;i}^{(2)} \cdot_{2;i}^{(2)}} \sin[a(\cdot_{1;i}^{(2)})] \sin[a(\cdot_{1;i}^{(2)})]$$

and the `sin' terms

$$\int_{[-a;a]^2} \sin(z_i^\top (x - \cdot_i)) \sin(x^\top \cdot\!! \cdot_i)) \, dx =$$
$$\frac{1}{\cdot_{1;i}^{(1)} \cdot_{2;i}^{(1)}} \sin[a(\cdot_{1;i}^{(1)})] \sin[a(\cdot_{1;i}^{(1)})] - \frac{1}{\cdot_{1;i}^{(2)} \cdot_{2;i}^{(2)}} \sin[a(\cdot_{1;i}^{(2)})] \sin[a(\cdot_{1;i}^{(2)})]$$

where

$$\cdot_i^{(1)} = (z_i - \cdot_i) + \cdot\!! \cdot_i; \qquad \cdot_{i;j}^{(2)} = (z_i - \cdot_i) - \cdot\!! \cdot_i :$$

The remaining `cos-sin'terms equal zero.

E PREDICTIVE DISTRIBUTION

To form predictive distributions, we assume that the latent posterior is approximated by $q(\mathbf{w}^{(r)}|\mathbf{X}; \ )$ as in Equation (20).

E.1 Predictive Intensity Distribution

For some $\mathbf{x} \ 2 \ X$, the predictive distribution of $f(\mathbf{x} \ )$ can be deduced from Equations (14) and (20) in the main text to be

$$f(\mathbf{x} \ )|\mathbf{X}; \quad N(\quad (\mathbf{x} \ ); \quad (\mathbf{x} \ )) \tag{47}$$

where

$$(\mathbf{x} \ ) := \hat{\mathbf{w}}^{(r)\top} \ '^{(r)}(\mathbf{x} \ ) \tag{48}$$

and

$$(\mathbf{x}) := \ '^{(r)}(\mathbf{x} \ )^{\top} \mathbf{Q} \ '^{(r)}(\mathbf{x} \ ). \tag{49}$$

Given $(\ ) = (f(\ ) + \ )^2$ and Equation (47) above, we can also derive the predictive distribution of the intensity function

$$(\mathbf{x} \ )|\mathbf{X}; \quad Gamma(a \ (\mathbf{x} \ ); b \ (\mathbf{x} \ )) \tag{50}$$

with parameters

$$a \ (\mathbf{x} \ ) = \frac{(\mathbf{x} \ )^2 + \quad (\mathbf{x} \ )^2 \ ^2}{2 \quad (\mathbf{x} \ )^2 (2 \quad (\mathbf{x} \ )^2 + \quad (\mathbf{x} \ )^2)} \text{and} \tag{51}$$

$$b \ (\mathbf{x} \ ) = \frac{(\mathbf{x} \ )^2 + \quad (\mathbf{x} \ )^2}{2 \quad (\mathbf{x} \ )^2 (2 \quad (\mathbf{x} \ )^2 + \quad (\mathbf{x} \ )^2)}. \tag{52}$$

E.2 Predictive Expected Log-likelihood

For a training set $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and an held-out test set $\mathbf{X} \ = \{\mathbf{x}_i\}_{i=1}^N$, we can derive an approximation for the expected predictive log-likelihood

$$\mathbb{E}[\log p(\mathbf{X} \ |\mathbf{X})] \quad \mathbb{E}_{\mathbf{w}^{(r)}} \int_X (\mathbf{w}^{(r)\top} \ '^{(r)}(\mathbf{x}) + \ )^2 d\mathbf{x} \ + \sum_{i=1}^{N} \mathbb{E}_{\mathbf{w}^{(r)}} \left[\log(\mathbf{w}^{(r)\top} \ '^{(r)}(\mathbf{x}_i) + \ )^2) \right]$$

where $\mathbf{w}^{(r)} \quad q(\mathbf{w}^{(r)}|\mathbf{X}; \ )$.

The integral term can be solved as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{w}^{(r)}} \int_X (\mathbf{w}^{(r)\top} \ '^{(r)}(\mathbf{x}) + \ )^2 d\mathbf{x} \\
&= \int_X \mathbb{E}_{\mathbf{w}^{(r)}}[\mathbf{w}^{(r)\top} \ '^{(r)}(\mathbf{x}) + \ ]^2 d\mathbf{x} + \int_X \text{Var}[\mathbf{w}^{(r)\top} \ '^{(r)}(\mathbf{x})] d\mathbf{x} \\
&= \int_X \ '^{(r)}(\mathbf{x})^{\top} \hat{\mathbf{w}}^{(r)} \hat{\mathbf{w}}^{(r)\top} \ '^{(r)}(\mathbf{x}) \ d\mathbf{x} + 2 \int_X \hat{\mathbf{w}}^{(r)\top} \ '^{(r)}(\mathbf{x}) \ d\mathbf{x} + \ ^2 |X| + \int_X \ '^{(r)}(\mathbf{x})^{\top} \mathbf{Q} \ '^{(r)}(\mathbf{x}) \ d\mathbf{x} \\
&= \text{tr} \ (\hat{\mathbf{w}}^{(r)} \hat{\mathbf{w}}^{(r)\top} + \mathbf{Q}) \underbrace{\int_X \ '^{(r)}(\mathbf{x}) \ '^{(r)}(\mathbf{x})^{\top} d\mathbf{x}}_{:= \mathbf{M}^{(r)}} + 2 \ \hat{\mathbf{w}}^{(r)\top} \underbrace{\int_X \ '^{(r)}(\mathbf{x}) d\mathbf{x}}_{:= \mathbf{m}^{(r)}} + \ ^2 |X| \\
&= \hat{\mathbf{w}}^{(r)\top} \mathbf{M}^{(r)} \hat{\mathbf{w}}^{(r)} + \text{tr} \ \mathbf{Q} \mathbf{M}^{(r)} \ + 2 \ \hat{\mathbf{w}}^{(r)\top} \mathbf{m}^{(r)} + \ ^2 |X|
\end{aligned}$$

where $\mathbf{M}^{(r)}$ and $\mathbf{m}^{(r)}$ are defined in Proposition 3.3. Note that we used the cyclical property of the trace in the last two lines. We also used the Tonelli's theorem in the first line to reverse the ordering of the integration over the positive integrand $(\mathbf{w}^{(r)\top} \ '^{(r)}(\mathbf{x}) + \ )^2 q(\mathbf{w}^{(r)})$.

The sum-of-expectations can also be expressed analytically. It takes of form

$$\sum_i \mathbb{E}[\log z_i^2] \quad \text{where} \quad z_i \quad N(\ _i; \ _i) \tag{53}$$

with

$$_i := \mathbf{w}^{(r)\top} '^{(r)}(\mathbf{x}_i) + \qquad \text{and} \qquad _i := '^{(r)}(\mathbf{x}_i)^\top \mathbf{Q}\, '^{(r)}(\mathbf{x}_i). \tag{54}$$

Following Lloyd et al. (2015, section 4.3), each summand can be expressed as

$$\mathrm{E}[\log z_i^2] = \quad G\Big(\ \frac{i}{2\ \frac{2}{i}}\Big) + \log\ \ \frac{\frac{2}{i}}{2} \qquad C$$

where $G(\ )$ is defined as

$$G(z) = 2z \sum_{j=0}^{\infty} \frac{j!z^j}{(2)_j (1=2)_j} \tag{55}$$

with $(\ )_j$ being the rising Pochhammer series. The constant $C \quad 0{:}57721566$ is the Euler Mascheroni constant. $G(\ )$ can in practice be evaluated using a large multi-resolution look-up table of pre-computed values. Accurate evaluation can be obtained by linear interpolation of the values from the table.

F MARGINAL LIKELIHOOD DERIVATIVES

In Equation (23) the marginal likelihood is expressed as

$$\begin{aligned}
\log p(\mathbf{X}j\ \ ) &= \log p(\mathbf{w}^{(r)}; \mathbf{X}j\ \ ) \quad \log p(\mathbf{w}^{(r)}j\mathbf{X};\ \ ) \\
&\quad \log p(\mathbf{w}^{(r)}; \mathbf{X}j\ \ ) \quad \log q(\mathbf{w}^{(r)}j\mathbf{X};\ \ ) \\
&= \quad \mathbf{w}^{(r)\top} \mathbf{M}^{(r)} \mathbf{w}^{(r)} + \sum_{i=1}^{N} \log(j\mathbf{w}^{(r)\top} '^{(r)}(\mathbf{x}_i)j^2) \quad \frac{1}{2}\mathbf{w}^{(r)\top}\mathbf{w}^{(r)} + \frac{1}{2}\log j\mathbf{Q}j + C
\end{aligned}$$

for some constant C, where we assume $\quad = 0$ without loss of generality.

Marginal Likelihood Derivatives We now compute the gradient with respect to the hyperparameters $\ $. Using the chain rule,

$$r_{\ _i} \log p(\mathbf{X}j\ \ ) = \frac{@\log p(\mathbf{X}j\ \ )}{@\ _i}_{explicit} + \sum_{j=1}^{N} \frac{@\log p(\mathbf{X}j\ \ )}{@\mathbf{w}_j} \frac{@\mathbf{w}_j}{@\ _i}. \tag{56}$$

The first term of Equation (56) can be solved as

$$\frac{@\log p(\mathbf{X}j\ \ )}{@\ _j}_{explicit} = \quad \mathbf{w}^{(r)\top} \frac{@\mathbf{M}^{(r)}}{@\ _j}\mathbf{w}^{(r)} + 2\sum_{i=1}^{N} \frac{\mathbf{w}^{(r)\top}\frac{@\, '^{(r)}(\mathbf{x}_i)}{@\ _j}}{\mathbf{w}^{(r)\top} '^{(r)}(\mathbf{x}_i)} + \frac{1}{2}\mathrm{tr}(\mathbf{Q}^{\ 1}\frac{@\mathbf{Q}}{@\ _j}) \tag{57}$$

where $\mathbf{Q}^{\ 1} = \ r^2_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}j\mathbf{X};\ \ )j_{\mathbf{w}^{(r)}=\mathbf{w}^{(r)}}$ is the precision matrix expressed in Section 4.1. The last term of Equation (57) can be expressed as

$$\begin{aligned}
\frac{1}{2}\mathrm{tr}(\mathbf{Q}^{\ 1}\frac{@\mathbf{Q}}{@\ _j}) &= \quad \frac{1}{2}\mathrm{tr}(\frac{@\mathbf{Q}^{\ 1}}{@\ _j}\mathbf{Q}) \\
&= \quad \mathrm{tr}\ \ \frac{@\mathbf{M}^{(r)}}{@\ _j}\mathbf{Q} + \frac{@}{@\ _j}\ \ \sum_{i=1}^{N} \frac{'(\mathbf{x}_i)\, '(\mathbf{x}_i)^\top}{(\mathbf{w}^\top '(\mathbf{x}_i))^2}\ \ \mathbf{Q}\ \ \Bigg]\ \ ! \\
&= \quad \mathrm{tr}\ \ \frac{@\mathbf{M}^{(r)}}{@\ _j}\mathbf{Q}\ \ \sum_{i=1}^{N} \frac{2}{(\mathbf{w}^{(r)\top} '(\mathbf{x}_i))^2}\ \ '(\mathbf{x}_i)^\top\mathbf{Q}\frac{@\ '(\mathbf{x}_i)}{@\ _j} \\
&\qquad + 2\sum_{i=1}^{N} \frac{\mathbf{w}^{(r)\top}\frac{@\, '(\mathbf{x}_i)}{@\ _j}}{(\mathbf{w}^{(r)\top} '(\mathbf{x}_i))^3}\ \ '(\mathbf{x}_i)^\top\mathbf{Q}\ '(\mathbf{x}_i)
\end{aligned} \tag{58}$$

Algorithm 1 Compute the log marginal likelihood derivatives [a].

input: $\mathbf{X}$ (inputs), spectral locations and hyper-parameters $\tilde{}$, posterior mode $\hat{\mathbf{w}}^{(r)}$, precision matrix $\mathbf{Q}^{-1}$, r r "integral" matrix $\mathbf{M}^{(r)}$, N r features matrix $\boldsymbol{\phi}^{(r)}(\mathbf{X})$
compute r $\mathbf{M}^{(r)}$, the dim($\tilde{}$) r r tensor of partial derivatives of $\mathbf{M}^{(r)}$ with respect to $\tilde{}$
compute r $\boldsymbol{\phi}^{(r)}(\mathbf{X})$, the dim($\tilde{}$) N r tensor of partial derivatives of $\boldsymbol{\phi}^{(r)}(\mathbf{X})$ with respect to $\tilde{}$
$\hat{\mathbf{f}} := \boldsymbol{\phi}^{(r)}(\mathbf{X})\hat{\mathbf{w}}^{(r)}$
$\mathbf{L} = \text{cholesky}(\mathbf{Q}^{-1})$. Solve $\mathbf{L}\mathbf{L}^{\top} = \mathbf{Q}^{-1}$
$\mathbf{C}_{\cdot} := \mathbf{L}\, n\, \boldsymbol{\phi}^{(r)}(\mathbf{X})$
$\mathbf{C}_{r\cdot} := \mathbf{L}\, n\, r\, \boldsymbol{\phi}^{(r)}(\mathbf{X})$
$\mathbf{r}_{\cdot} := (\mathbf{C}_{\cdot} \quad \mathbf{C}_{\cdot})\mathbf{1}_r$
$\mathbf{R}_{r\cdot} = (\mathbf{C}_{\cdot} \quad \mathbf{C}_{r\cdot})\mathbf{1}_r$
$\mathbf{S} := r\, \boldsymbol{\phi}^{(r)}(\mathbf{X})\hat{\mathbf{w}}^{(r)}$
$\mathbf{E} := (\mathbf{S} \text{ diag}(\hat{\mathbf{f}}^{2})) \quad \boldsymbol{\phi}^{(r)}(\mathbf{X})$
$\mathbf{s}_{integral} = \hat{\mathbf{w}}^{(r)\top} r\, \mathbf{M}^{(r)}\hat{\mathbf{w}}^{(r)}$
$\mathbf{s}_{data} = 2\,(\mathbf{S} \text{ diag}(\hat{\mathbf{f}}^{1}))\mathbf{1}_N$
$\mathbf{s}_1 = \text{tr}(\mathbf{L}^{\top} n\, (\mathbf{L}\, n\, r\, \mathbf{M}^{(r)}))$
$\mathbf{s}_2 = 2\,(\mathbf{R}_{r\cdot} \text{ diag}(\hat{\mathbf{f}}^{2}))\mathbf{1}_N$
$\mathbf{s}_3 = 2(\mathbf{S} \text{ diag}(\hat{\mathbf{f}}^{3}))\mathbf{r}_{\cdot}$
$\mathbf{v} := \text{diag}(\hat{\mathbf{f}}^{1})\, r\, \boldsymbol{\phi}^{(r)}(\mathbf{X}) \quad \mathbf{1}_N^{\top}\mathbf{E}$. Equation (60)
$\frac{d\mathbf{p}}{d} := \mathbf{s}_{data} \quad \mathbf{s}_{integral} \quad \mathbf{s}_1 \quad \mathbf{s}_2 + \mathbf{s}_3$. Equation (57)
$\frac{d\mathbf{p}}{d\mathbf{w}} := 2\,\boldsymbol{\phi}^{(r)}(\mathbf{X})^{\top}\text{diag}(\hat{\mathbf{f}}^{3})\mathbf{r}_{\cdot}$. Equation (59)
$\frac{d\mathbf{W}}{d} := 2(\,\mathbf{L}\, n\, (\mathbf{L}^{\top} n\, (r\, \mathbf{M}^{(r)}\hat{\mathbf{w}}^{(r)} \quad \mathbf{v})))$. Equation (61)
$\mathbf{g} = \frac{d\mathbf{p}}{d} + \frac{d\mathbf{W}}{d}\frac{d\mathbf{p}}{d\mathbf{w}}$. Equation (56) **return** $\mathbf{g}$ (dim($\tilde{}$)-vector of partial derivatives)

[a] We assume basic operations on $3d$ tensors are performed over its last two dimensions and repeated over its first one. $*$ denotes the element-wise multiplication. $A \setminus B$ where A is a triangular $r \times r$ matrix and B is a $\cdots \times r$ matrix or tensor is performed over the last dimension of B (i.e. r) and repeated over its firsts dimensions.

The $\frac{@\log p(\mathbf{X}j\ )}{@\hat{w}_j}$ terms of Equation (56) is

$$\begin{aligned}
\frac{@\log p(\mathbf{X}j\ )}{@\hat{w}_j} &= \underbrace{\frac{@\log p(\hat{\mathbf{w}}^{(r)};\mathbf{X}j\ )}{@\hat{w}_j}}_{=0} + \frac{1}{2}\text{tr}(\mathbf{Q}^{-1}\frac{@\mathbf{Q}}{@\hat{w}_j}) \\
&= \frac{1}{2}\text{tr}(\frac{@\ \mathbf{Q}^{-1}}{@\hat{w}_{j,,}}\mathbf{Q}) \\
&= \text{tr}\left(\frac{@}{@\hat{w}_j}\sum_{i=1}^{N}\frac{\boldsymbol{\phi}(\mathbf{x}_i)\,\boldsymbol{\phi}(\mathbf{x}_i)^{\top}}{(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\phi}(\mathbf{x}_i))^2}\right)^{\#\ !}\mathbf{Q} \\
&= 2\sum_{i=1}^{N}\frac{\boldsymbol{\phi}_j(\mathbf{x}_i)}{(\hat{\mathbf{w}}^{(r)\top}\boldsymbol{\phi}(\mathbf{x}_i))^3}\quad \boldsymbol{\phi}(\mathbf{x}_i)^{\top}\mathbf{Q}\ \boldsymbol{\phi}(\mathbf{x}_i)
\end{aligned} \tag{59}$$

where, in the first line, we have imposed stationarity using $r_{\mathbf{w}^{(r)}}\ \log p(\mathbf{w}^{(r)}j\mathbf{X};\quad)j_{\mathbf{w}^{(r)}=\hat{\mathbf{w}}^{(r)}} = 0$.

We finally express the last terms $\frac{@\hat{\mathbf{w}}^{(r)}}{@\ _i}$ of Equation (56). From the expression of $\hat{\mathbf{w}}^{(r)}$ in Equation (21) in the main text. We

obtain

$$\begin{aligned}
\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j}\left[\left(\mathbf{M}^{(r)} + \frac{1}{2}\mathbf{I}_r\right)^{-1}\left(\sum_{i=1}^{N}\frac{\phi'^{(r)}(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\phi'^{(r)}(\mathbf{x}_i)}\right)\right] \\
&= \frac{\partial}{\partial \theta_j}\left[\left(\mathbf{M}^{(r)} + \frac{1}{2}\mathbf{I}_r\right)^{-1}\right]\left(\sum_{i=1}^{N}\frac{\phi'(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\phi'(\mathbf{x}_i)}\right) \\
&\quad + \left(\mathbf{M}^{(r)} + \frac{1}{2}\mathbf{I}_r\right)^{-1}\left[\frac{\partial}{\partial \theta_j}\left(\sum_{i=1}^{N}\frac{\phi'(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\phi'(\mathbf{x}_i)}\right) + \frac{\partial}{\partial \hat{\mathbf{w}}^{(r)}}\left(\sum_{i=1}^{N}\frac{\phi'(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\phi'(\mathbf{x}_i)}\right)\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \theta_j}\right] \\
&= \left(\mathbf{M}^{(r)} + \frac{1}{2}\mathbf{I}_r\right)^{-1}\left[-\frac{\partial \mathbf{M}^{(r)}}{\partial \theta_i}\hat{\mathbf{w}}^{(r)} - \boldsymbol{\nu} - \sum_{i=1}^{N}\frac{\phi'(\mathbf{x}_i)\,\phi'(\mathbf{x}_i)^\top}{(\mathbf{w}^\top\phi'(\mathbf{x}_i))^2}\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \theta_i}\right]
\end{aligned}$$

where

$$\boldsymbol{\nu} := \frac{\partial}{\partial \theta_j}\left(\sum_{i=1}^{N}\frac{\phi'(\mathbf{x}_i)}{\hat{\mathbf{w}}^{(r)\top}\phi'(\mathbf{x}_i)}\right) = \sum_{i=1}^{N}\frac{\frac{\partial \phi'(\mathbf{x}_i)}{\partial \theta_j}}{\hat{\mathbf{w}}^{(r)\top}\phi'(\mathbf{x}_i)} - \sum_{i=1}^{N}\frac{\hat{\mathbf{w}}^{(r)\top}\frac{\partial \phi'(\mathbf{x}_i)}{\partial \theta_j}}{(\hat{\mathbf{w}}^{(r)\top}\phi'(\mathbf{x}_i))^2}\phi'(\mathbf{x}_i) \tag{60}$$

Thus

$$\begin{aligned}
\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \theta_i} &= -\left[\left(\mathbf{M}^{(r)} + \frac{1}{2}\mathbf{I}_r + \sum_{i=1}^{N}\frac{\phi'(\mathbf{x}_i)\,\phi'(\mathbf{x}_i)^\top}{(\mathbf{w}^\top\phi'(\mathbf{x}_i))^2}\right)^{-1}\frac{\partial \mathbf{M}^{(r)}}{\partial \theta_i}\hat{\mathbf{w}}^{(r)} - \boldsymbol{\nu}\right] \\
&= -2\mathbf{Q}\frac{\partial \mathbf{M}^{(r)}}{\partial \theta_i}\hat{\mathbf{w}}^{(r)} - \boldsymbol{\nu}
\end{aligned} \tag{61}$$

Implementation Details The implementation of the marginal log likelihood partial derivatives with respect to the hyperparameters $\tilde{}$ is shown in Algorithm 1.

G ALGORITHMS

We present in Algorithm 2 the standard process to compute the mode $\hat{\mathbf{w}}^{(r)}$ and the hyperparameters .

Algorithm 2 Compute the mode $\hat{\mathbf{w}}^{(r)}$ and the hyperparameters

input: data $\mathbf{X}$
initialize $_0$ and $\hat{\mathbf{w}}_0^{(r)}$
for $t = 1, \ldots, T$ **do**
  compute $\mathbf{M}^{(r)}$ and $\phi'^{(r)}(\mathbf{X})$
  $\hat{\mathbf{w}}_t^{(r)} := \text{mode}(p(\mathbf{w}^{(r)}|\mathbf{X}; \ ); \hat{\mathbf{w}}_{t-1}^{(r)})$. locate posterior mode using Equation (22) with initial value $\hat{\mathbf{w}}_{t-1}^{(r)}$
  $\hat{\mathbf{f}} := \phi'^{(r)}(\mathbf{X})\hat{\mathbf{w}}_t^{(r)} +$
  $\mathbf{V} := \text{diag}(\hat{\mathbf{f}}^{-1})\,\phi'^{(r)}(\mathbf{X})$
  $\mathbf{Q}^{-1} := 2\mathbf{M}^{(r)} + \mathbf{I}_r + 2\mathbf{V}^\top\mathbf{V}$. precision matrix $\mathbf{Q}^{-1} = -\nabla^2_{\mathbf{w}^{(r)}} \log p(\mathbf{w}^{(r)}|\mathbf{X}; \ )|_{\mathbf{w}^{(r)}=\hat{\mathbf{w}}^{(r)}}$
  compute the gradient $\mathbf{g}$ from Algorithm 1
  $_t$ update($_{t-1}; \mathbf{g}$)
end for
return $\hat{\mathbf{w}}_T^{(r)}$ (mode) and $_T$ (hyperparameters)

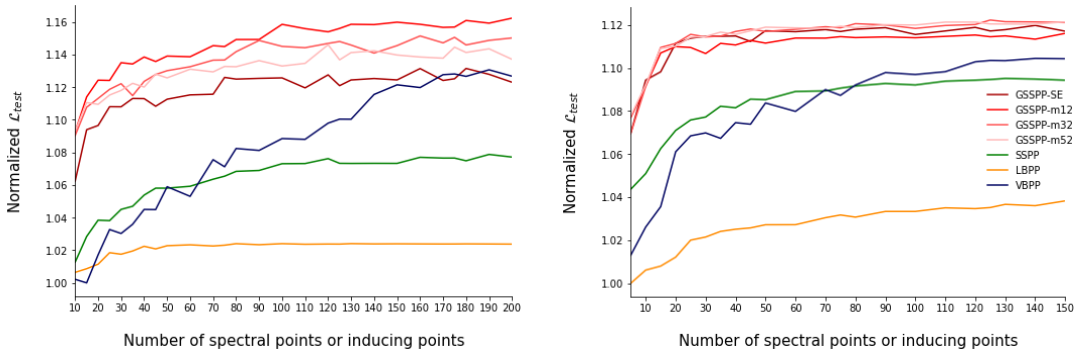
As an alternative (Section 4.2), we assume $\frac{\partial \hat{\mathbf{w}}^{(r)}}{\partial \theta_j} = 0$ and alternate independent update for the mode in Equation (22) and the hyperparameters in Equation (24), which yields Algorithm 3.

Algorithm 3 Compute the mode $\hat{\mathbf{w}}^{(r)}$ and the hyperparameters with independent updates

input: data $\mathbf{X}$
initialize $_0$ and $\hat{\mathbf{w}}_0^{(r)}$
for $t = 1, \ldots, T$ **do**
  compute $\mathbf{M}^{(r)}$ and $'^{(r)}(\mathbf{X})$
  $\hat{\mathbf{w}}_t^{(r)}$ $\hat{\mathbf{w}}_t^{(r)}$. update the posterior mode using one iteration of Equation (22)
  $\hat{\mathbf{f}} := '^{(r)}(\mathbf{X})\hat{\mathbf{w}}_t^{(r)} +$
  $\mathbf{V} := \mathrm{diag}(\hat{\mathbf{f}}^{\ 1})'^{(r)}(\mathbf{X})$
  $\mathbf{Q}^{\ 1} := 2\mathbf{M}^{(r)} + I_r + 2\mathbf{V}^{>}\mathbf{V}$. precision matrix
  compute the gradient $\frac{d\mathbf{p}}{d}$ from Algorithm 1 (line 18)
  $_t$ update($_{t\ 1}; \frac{d\mathbf{p}}{d}$)
end for
return $\hat{\mathbf{w}}_T^{(r)}$ (mode) and $_T$ (hyperparameters)

H SUPPLEMENTARY FIGURES

Figure 5 (below) shows the average test expected log-likelihood as a function of the number of basis functions or inducing points for *bei* data set and *Taxi* data set across models.



(a) Taxi data (b) Bei data

Figure 5: Average normalized test expected log-likelihood (L_{test}) of the different methods on 2D real data, as a function of the number of spectral points or inducing points.