Journal of Machine Learning Research 26 (2025) 1-34 Submitted 10/24; Revised 3/25; Published 4/25

Linear cost and exponentially convergent approximation of Gaussian Matérn processes on intervals

David Bolin [*] DAVID.BOLIN@KAUST.EDU.SA
CEMSE Division, statistics program
King Abdullah University of Science and Technology (KAUST),
Thuwal 23955-6900, Kingdom of Saudi Arabia

Vaibhav Mehandiratta VAIBHAV.MEHANDIRATTA@KAUST.EDU.SA
CEMSE Division, statistics program
King Abdullah University of Science and Technology (KAUST),
Thuwal 23955-6900, Kingdom of Saudi Arabia

Alexandre B. Simas ALEXANDRE.SIMAS@KAUST.EDU.SA
CEMSE Division, statistics program
King Abdullah University of Science and Technology (KAUST),
Thuwal 23955-6900, Kingdom of Saudi Arabia

Editor: Brian Kulis

Abstract

The computational cost for inference and prediction of statistical models based on Gaussian processes with Matérn covariance functions scales cubically with the number of observations, limiting their applicability to large data sets. The cost can be reduced in certain special cases, but there are no generally applicable exact methods with linear cost. Several approximate methods have been introduced to reduce the cost, but most lack theoretical guarantees for accuracy. We consider Gaussian processes on bounded intervals with Matérn covariance functions and, for the first time, develop a generally applicable method with linear cost and a covariance error that decreases exponentially fast in the order m of the proposed approximation. The method is based on an optimal rational approximation of the spectral density and results in an approximation that can be represented as a sum of m independent Gaussian Markov processes, facilitating usage in general software for statistical inference. Besides theoretical justifications, we demonstrate accuracy empirically through carefully designed simulation studies, which show that the method outperforms state-of-the-art alternatives in accuracy for fixed computational cost in tasks like Gaussian process regression.

Keywords: Gaussian process, Gaussian Markov random field, inference, prediction

1. Introduction

Gaussian stochastic processes with Matérn covariance functions (Matérn, 1960) are important models in statistics and machine learning (Porcu et al., 2023), and in particular in areas such as spatial statistics (Stein, 1999; Lindgren et al., 2022), computer experiments (Santner et al., 2003; Gramacy, 2020) and Bayesian optimization (Srinivas et al., 2009). Although the Matérn covariance often is used for spatial data, it is also commonly used for temporal

[*] Authors are listed in alphabetical order.

data, in particular in areas such as functional data analysis (Yang et al., 2016), longitudinal data analysis (Asar et al., 2020), and growth rate modeling (Swain et al., 2016). A Gaussian process u on $\mathbb{R}$ has a Matérn covariance function if

$$\mathrm{Cov}(u(s), u(t)) = \varrho(|s-t|; \nu, \kappa, \sigma^2), \quad \varrho(h; \nu, \kappa, \sigma^2) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^{\nu} K_{\nu}(\kappa h), \tag{1}$$

where $K_\nu(\cdot)$ is a modified Bessel function of the second kind of order ν, and $\Gamma(\cdot)$ denotes the gamma function. The three parameters $\kappa, \sigma^2, \nu > 0$, determine the practical correlation range, variance and smoothness of the process, respectively. We introduce the practical correlation range $\rho = \sqrt{8\nu}/\kappa$ as the distance at which the correlation is approximately 0.1, and a new smoothness parameter $\alpha = \nu + 1/2$, which will be used throughout this text.

However, the approach of defining Gaussian processes via the covariance function incurs a significant computational cost for inference and prediction, mainly due to the requirement of factorising dense covariance matrices. To address this so called "Big n" problem, referring to the $\mathcal{O}(n^3)$ computational cost required for problems with n observations, several methods have been developed in recent years to reduce the cost. In this work, we consider the case where the data is observed on an interval $I \subset \mathbb{R}$. The processes have Markov properties when $\alpha \in \mathbb{N}$, and this has been used to derive exact methods with a $\mathcal{O}(n)$ cost for these particular cases, see for example the Kernel packet method (Chen et al., 2022) or the state-space methods of Hartikainen and Särkkä (2010); Särkkä and Hartikainen (2012); Särkkä et al. (2013). If the data is evenly spaced in $\mathbb{R}$, Toeplitz methods (Wood and Chan, 1994) can also be used to reduce the computational cost to $\mathcal{O}(n(\log n)^2)$ (Ling, 2019; Ling and Lysy, 2022). Outside these particular cases, there are currently no exact methods that can reduce the computational cost, and one instead needs to rely on approximations. One popular method is the random Fourier features approach of Rahimi and Recht (2007) which gives a cost $\mathcal{O}(m^2 + mn)$ when using m features, and an accuracy of $\mathcal{O}(m^{-1/2})$. This is one of the few methods that have theoretical guarantees for how quickly the approximation error decreases as the order m increases. Unfortunately, the theoretical rate is low and as we will later see, the method provides poor approximations for Matérn processes on $\mathbb{R}$.

Another widely used method is the SPDE approach of Lindgren et al. (2011), where the Matérn process is represented as a solution to a stochastic differential equation (SDE) which is approximated via a finite element method (FEM) approximation. This was originally proposed for the case $\nu - 1/2 \in \mathbb{N}_0$ (i.e., $\alpha \in \mathbb{N}$) and was later extended by Bolin et al. (2020, 2024b) to general $\nu > 0$, where the authors also derived explicit rates of convergence of the approximation in terms of the finite element mesh width. The approach is computationally efficient, but has the disadvantage that the SDE has certain boundary conditions which makes the approximation converge to a non-stationary covariance which is only similar to the Matérn covariance away from the boundary of the computational domain. The rate of convergence is better than that of the random Fourier features method, but does not decrease exponentially fast. Other approaches, which are generally applicable, but without theoretical rates of convergence of the covariance function approximation are covariance tapering (Furrer et al., 2006), Vecchia approximation (Vecchia, 1988; Gramacy and Apley, 2015; Datta et al., 2016), low rank methods (Higdon, 2002; Cressie and Johannesson, 2008), and multiresolution approximations (Nychka et al., 2015). The state-space methods have also been extended to general $\nu > 0$ in Karvonen and Särkkä (2016) and Tronarp et al. (2018)

through spectral transformation methods and the approximation of the Matérn kernel by a finite scale mixture of squared exponential kernels, respectively. However, the accuracy of the resulting approximated covariance function has been demonstrated only through numerical experiments and no theoretical analysis of the rate has been provided.

In this work, we develop a new method for Gaussian Matérn processes on intervals, which has at most $\mathcal{O}(nm^3\lceil\alpha\rceil^3)$ computational cost when used for statistical inference, prediction and sampling, where m is the order of the approximation. We prove in Section 2 that the error of the covariance function converges exponentially fast in m, which in practice means that m can be chosen very low. Specifically, the error is $O\big(\exp(-2\pi\sqrt{\{\alpha\}m})\big)$ where $\{\alpha\}$ is the fractional part of $\alpha \notin \mathbb{N}$. If $\alpha \in \mathbb{N}$, the method is exact and has a cost of $\mathcal{O}(n\lceil\alpha\rceil^3)$. The approach is based on a rational approximation of the spectral density, which is thus similar in spirit to Karvonen and Särkkä (2016); Roininen et al. (2018). The difference is, however, that we have theoretical guarantees for the error. Because the method is based on a rational approximation, it can be used in combination with state-space methods for efficient inference. We, however, derive, in Section 3, a direct representation of the approximated process as a sum of Gaussian Markov random fields with sparse precision (inverse covariance) matrices, which in fact are band matrices. This representation has several advantages, and perhaps the most important is that it can directly be incorporated in general software for Bayesian inference, such as `R-INLA` (Lindgren and Rue, 2015). Another important feature is that the linear cost is applicable to working with the process and its derivatives jointly. This is useful in several applications that arise in the natural sciences where observations of the derivatives are available (see, e.g. Solak et al., 2002; Padidar et al., 2021; Roos et al., 2021; Yang et al., 2018) or when derivatives are of direct interest (Swain et al., 2016).

To validate the effectiveness and accuracy of the method, we compare it in Section 4 to the covariance-based rational approximation method of Bolin et al. (2024b) as well as the Vecchia approximations (Datta et al., 2016). These methods demonstrated superior performance among several alternatives for approximating Gaussian Matérn fields in specific geostatistical test problems, as shown in the comparative study by Hong et al. (2023). We also compare with the state-space approach of Karvonen and Särkkä (2016), the random Fourier features method of Rahimi and Recht (2007) and the covariance tapering approach of Nychka et al. (2015). We show that the method outperforms the alternatives in terms of accuracy for a fixed cost when used for Gaussian process regression. The comparison also includes a principal component analysis (PCA) (Wang, 2008) approach, which serves as an "optimal" low-rank method. As the proposed method outperforms this, it means that it also would outperform any other low-rank method, such as fixed rank kriging (Cressie and Johannesson, 2008) or process convolutions (Higdon, 2002). Extensions of the method beyond stationary Matérn processes on intervals, as well as concluding remarks, are given in Section 5. The proposed method is implemented in the R package `rSPDE` (Bolin and Simas, 2023) available on CRAN, and all code for the comparisons, as well as a Shiny application with further results can be found in `https://github.com/vpnsctl/MarkovApproxMatern/`. Proofs and technical details are provided in two technical appendices.

2. Exponentially convergent rational approximation

The proposed method can be obtained through two equivalent formulations. Either through a rational approximation of the spectral density of the Gaussian process, or through a rational approximation of the covariance operator of the process. The latter formulation enables extensions which we will explore in Section 5. In this section, we outline the idea through the spectral density approach, which is less technical.

Let u be a centered Gaussian Process on an interval $I \subset \mathbb{R}$ with covariance function (1) and $\alpha = \nu + 1/2 \notin \mathbb{N}$ (which is the case for which no exact and efficient methods exist). This process has spectral density $f_\alpha(w) = A\sigma^2(\kappa^2 + w^2)^{-\alpha}$, where $A = \sqrt{2}\kappa^{2\nu}\Gamma(\nu + 1/2)\Gamma(\nu)^{-1}$ (Lindgren, 2012). We define a rational approximation of the process u as a Gaussian process with spectral density

$$f_{m,\alpha}(w) = A\kappa^{-2\alpha}\frac{\sigma^2}{(1+\kappa^{-2}w^2)^{\lfloor\alpha\rfloor}}\frac{P_m(1+\kappa^{-2}w^2)}{Q_m(1+\kappa^{-2}w^2)}, \tag{2}$$

where $P_m(x) = \sum_{i=0}^m a_i x^{m-i}$ and $Q_m = \sum_{i=0}^m b_i x^{m-i}$ are polynomials derived from the optimal rational approximation of order m for the real-valued function $f(x) = x^{\{\alpha\}}$ on the interval $[0,1]$, with respect to the supremum norm, where we recall that $\{\alpha\}$ is the fractional part of α. More precisely, we use the approximation $x^{-\{\alpha\}} = \left(x^{-1}\right)^{\{\alpha\}} \approx P_n(x)/Q_n(x)$, where the coefficients $\{a_i\}_{i=0}^m$ and $\{b_i\}_{i=0}^m$ are such that the approximation

$$x^{\{\alpha\}} \approx \frac{\sum_{i=0}^m a_i x^i}{\sum_{i=0}^m b_i x^i} \tag{3}$$

is the best with respect to the supremum norm on $[0,1]$. The coefficients $\{a_i\}_{i=0}^m$ and $\{b_i\}_{i=0}^m$ in the rational approximation are unique (Lorentz et al., 1996, Chapter 7.2) and can be obtained via the second Remez algorithm (Remez, 1934) or by using the recent, and more stable, BRASIL algorithm (Hofreither, 2021). Note that the polynomials $P_m(x)$ and Q_m in (2) are written in terms of x^{m-i} instead of x^i, using the coefficients from (3), because the rational approximation is applied to $(x^{-1})^{\{\alpha\}}$.

The true and approximate spectral densities for $\nu = 0.4$, 1.4 and 2.4, along with their absolute errors, are shown in Figure 1 for different orders of rational approximation, where we obtained the coefficients using the BRASIL algorithm. Here, we consider $\sigma = 1$ and $\rho = 2$, where $\rho = \sqrt{8\nu}/\kappa$ represents the practical correlation range. We can observe that the approximation is very good even for low orders, with an almost perfect match already for order 2. A plot showing relative absolute errors exhibits very similar patterns, thus we chose not to include it in the paper, but it is available in the shiny application.

Let $r_{m,\alpha}$ be the corresponding covariance function obtained from the rational approximation of order m of f_α given by (2), more precisely, let

$$r_{m,\alpha}(t) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{iwt} f_{m,\alpha}(w)\,dw, \quad t \in \mathbb{R}. \tag{4}$$

An explicit expression for $r_{m,\alpha}(\cdot)$ will be obtained in the next section.

The following result demonstrates that the approximated covariance function $r_{m,\alpha}(\cdot)$ converges exponentially fast to the true Matérn covariance function with respect to both

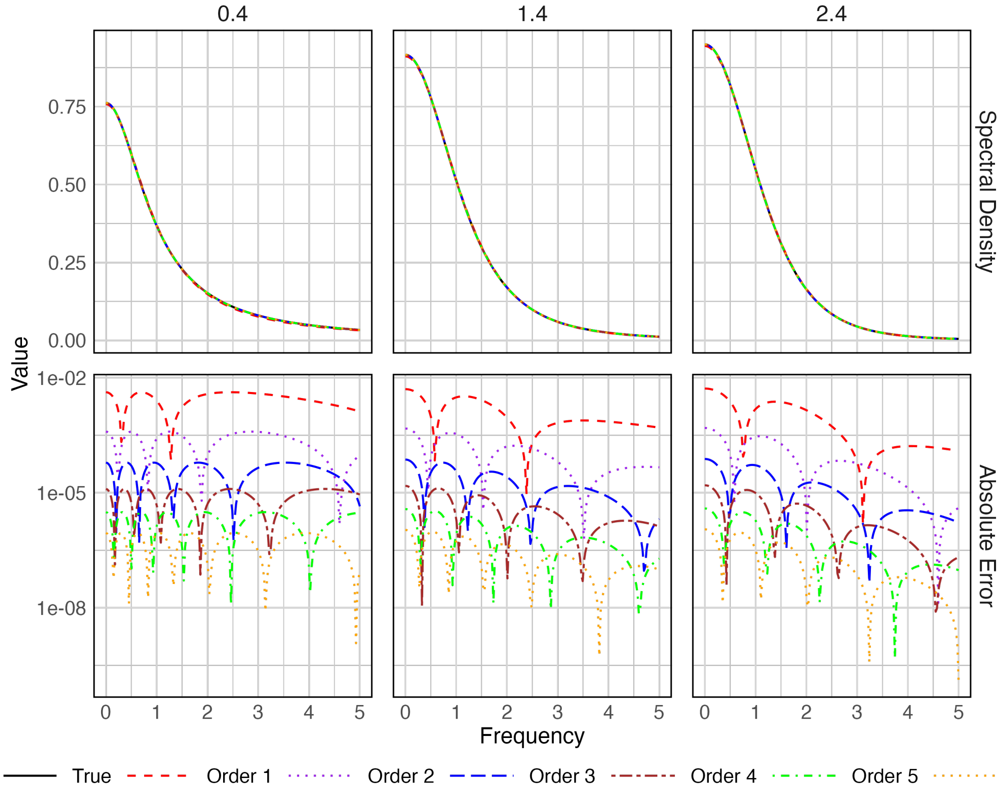

Figure 1: True and approximate spectral densities for $\nu = 0.4$, 1.4 and 2.4, and for different rational approximation orders (top row), and corresponding absolute errors for the rational approximations (bottom row).

the $L_2(I\times I)$-norm and the supremum norm. For $f \in L_2(I\times I)$, the $L_2(I\times I)$-norm is defined as $\|f\|^2_{L_2(I\times I)} = \int_{I\times I} |f(s,t)|^2\, ds\, dt$, where $L_2(I\times I)$ is the space of equivalence classes of real-valued square-integrable functions on $I \times I$. For $f \in C(I \times I)$, the supremum norm is defined as $\|f\|_{C(I\times I)} = \sup_{(s,t)\in I\times I} |f(s,t)|$, where $C(I \times I)$ is the space of continuous real-valued functions on $I \times I$.

Theorem 1 *Let r_α be the covariance function*

$$r_\alpha(s,t;\kappa,\sigma^2) = \varrho(|s-t|;\alpha - 1/2, \kappa, \sigma^2), \quad s,t, \in \mathbb{R}, \tag{5}$$

where $\varrho(\cdot)$ is the Matérn covariance function defined in (1) *and $\alpha = \nu + 1/2$. Further, let $r_{m,\alpha}$ be the rational approximation given in* (4)*. If $\alpha > 1/2$, then*

$$\|r_{m,\alpha} - r_\alpha\|_{L_2(I\times I)} \leq 2\sqrt{\pi}(b-a)A\sigma^2\kappa^{-2\alpha}\min\{1, M_{\lfloor\alpha\rfloor,\kappa}\}C_{\{\alpha\}}\mathbb{I}_{\alpha\notin\mathbb{N}}e^{-2\pi\sqrt{\{\alpha\}m}}, \tag{6}$$

where $\{\alpha\}$ denotes the fractional part of α, $M_{n,\kappa} = \kappa\pi\Gamma(2n-1)/(4^{n-1}\Gamma(n)^2)$, $n \geq 1$ and $M_{0,\kappa} = \infty$. Additionally, $C_{\{\alpha\}} \in (0,\infty)$ is a constant that depends only on $\{\alpha\}$. Further, if

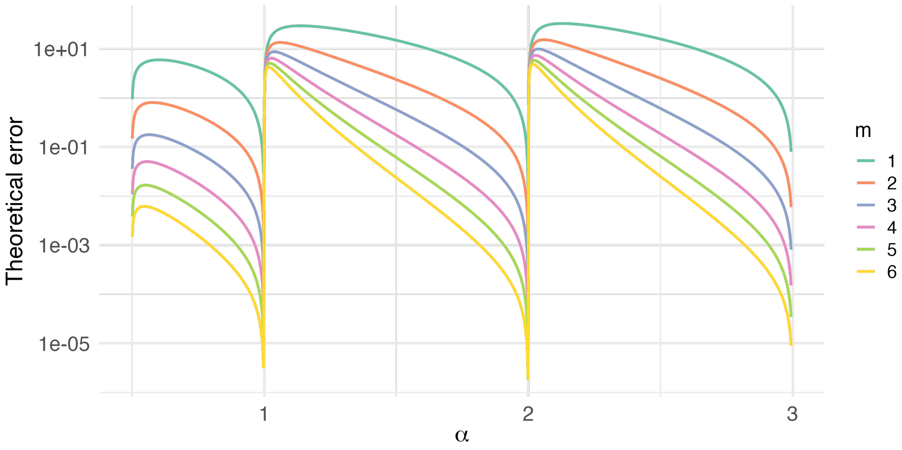

Figure 2: Theoretical error bound from (6) in Theorem 1 using the approximation for $C_{\{\alpha\}}$ given in Remark 2.

$\alpha > 1$, then a constant $K_\alpha \in (0,\infty)$ that only depends on α exists such that

$$\|r_{m,\alpha} - r_\alpha\|_{C(I\times I)} \leq A\sigma^2\kappa^{-2\alpha}K_\alpha \mathbb{I}_{\alpha\notin\mathbb{N}} e^{-2\pi\sqrt{\{\alpha\}m}}. \tag{7}$$

Remark 2 *The constant $C_{\{\alpha\}}$ in (6) comes from the error of the rational approximation*

$$\sup_{x\in[0,1]} \left| x^{\{\alpha\}} - \frac{p_m(x)}{q_m(x)} \right| \leq C_{\{\alpha\}} e^{-2\pi\sqrt{\{\alpha\}m}},$$

where $C_{\{\alpha\}}$ is independent of m. Further, for $\alpha \in (0,1)$, by Saff and Stahl (1995, Theorem 2), there is an approximate expression for this uniform error that is valid for large m:

$$\sup_{x\in[0,1]} \left| x^{\{\alpha\}} - \frac{p_m(x)}{q_m(x)} \right| = 4^{\{\alpha\}+1}|\sin(\pi\{\alpha\})| e^{-2\pi\sqrt{\{\alpha\}m}}(1+o(1)), \quad \text{as } m\to\infty.$$

This expression provides an intuition on how the constant $C_{\{\alpha\}}$ behaves.

The theoretical error (6) using the approximate expression of $C_{\{\alpha\}}$ given in Remark 2 is shown in Figure 2 for $\rho = 2$, $\sigma = 1$, and different values of ν and m on the interval $I = [0, 50]$. We can observe that the errors do not decay monotonically as $\{\alpha\}$ increases, but instead, they have a more complex behavior with respect to α. However, they decay monotonically as m increases. One should note that the true error bounds may look a bit different compared to those in Figure 2 as they are based on using $4^{\{\alpha\}+1}|\sin(\pi\{\alpha\})|$ in place of $C_{\{\alpha\}}$. See Figure 3 for the actual covariance errors for this example.

3. Linear cost inference

Our goal now is to use the rational approximation in (2) to obtain a linear cost approximation of the covariance function of the Gaussian process u. This approximation will based on a partial fractions decomposition of the rational function in the approximate spectral density. Therefore, we first derive an important property of such partial fractions decompositions.

Proposition 3 *Fix $\alpha \in (0,1)$ and $m \in \mathbb{N}$. Let the coefficients $\{a_i\}_{i=0}^m$ and $\{b_i\}_{i=0}^m$ be such that the best rational approximation of x^α on $[0,1]$ is given by (3). Further, let $P_m(x) = \sum_{i=0}^m a_i x^{m-i}$ and $Q_m = \sum_{i=0}^m b_i x^{m-i}$. Then, we have the following partial fractions decomposition of $P_m(x)/Q_m(x)$:*

$$\frac{P_m(x)}{Q_m(x)} = k + \sum_{i=1}^m \frac{c_i}{x - p_i},$$

where $k, c_i > 0$ and $p_i < 0$ for $i = 1, \dots, m$.

Remark 4 *Proposition 3 fills a theoretical gap left in Bolin et al. (2024b), where such a decomposition, along with the signs of k, c_i, and p_i for $i = 1, \dots, m$, was verified numerically.*

In view of Proposition 3, we can perform a partial fraction decomposition of the rational function $P_m(x)/Q_m(x)$ in (2) to obtain that the spectral density of the approximation is

$$\begin{aligned} f_{m,\alpha}(w) &= A\sigma^2\kappa^{-2\alpha}\left[\frac{k}{(1+\kappa^{-2}w^2)^{\lfloor\alpha\rfloor}} + \sum_{i=1}^m c_i \frac{1}{(1+\kappa^{-2}w^2)^{\lfloor\alpha\rfloor}(1+\kappa^{-2}w^2 - p_i)}\right] \\ &=: \left[f_{m,0,\alpha}(w) + \sum_{i=1}^m f_{m,i,\alpha}(w)\right], \end{aligned} \tag{8}$$

where $k, c_i > 0$ and $p_i < 0$ for $i = 1, \dots, m$, and $A = \sqrt{2}\kappa^{2\nu}\Gamma(\nu + 1/2)\Gamma(\nu)^{-1}$. Thus, we obtain that $f_{m,\alpha}(\cdot)$ can be decomposed as a sum of valid spectral densities. Hence, we can write $r_{m,\alpha}(\cdot,\cdot)$, given in (4), as a sum of covariance functions.

Let $\varrho_{m,\alpha}(\cdot)$ be defined as $\varrho_{m,\alpha}(t-s) = r_{m,\alpha}(t,s)$. Based on (8), we obtain the following explicit expression of the approximated covariance function $\varrho_{m,\alpha}(\cdot)$.

Proposition 5 *Let u be a Gaussian process with spectral density (8). Then, it has covariance function*

$$\varrho_{m,\alpha}(h) = \varrho_{m,0,\alpha}(h) + \sum_{i=1}^m \varrho_{m,i,\alpha}(h), \tag{9}$$

where

$$\varrho_{m,0,\alpha}(h) = k\sigma^2 \cdot \begin{cases} \frac{c_\alpha\sqrt{4\pi}}{\kappa} 1_{[h=0]} & 0 < \alpha < 1, \\ \varrho\left(h; \lfloor\alpha\rfloor - \frac{1}{2}, \kappa, \frac{c_\alpha}{c_{\lfloor\alpha\rfloor}}\right) & \alpha \geq 1, \end{cases}$$

and

$$\varrho_{m,i,\alpha}(h) = c_i\sigma^2 \cdot \begin{cases} \varrho\left(h; \frac{1}{2}, \kappa_i, \frac{c_\alpha\sqrt{\pi}}{\sqrt{1-p_i}}\right) & 0 < \alpha < 1, \\ \frac{1}{p_i^{\lfloor\alpha\rfloor}}\varrho\left(h; \frac{1}{2}, \kappa_i, \frac{c_\alpha\sqrt{\pi}}{\sqrt{1-p_i}}\right) - \sum_{j=1}^{\lfloor\alpha\rfloor} \frac{1}{p_i^{\lfloor\alpha\rfloor+1-j}}\varrho\left(h; j - \frac{1}{2}, \kappa, \frac{c_\alpha}{c_j}\right) & \alpha \geq 1. \end{cases}$$

Here $c_a := {}^{\Gamma(a)}/_{\Gamma(a-1/2)}$, $\kappa_i = \kappa\sqrt{1-p_i}$, and ϱ is the Matérn covariance (1).

Since the spectral density in (8) is a sum of valid spectral densities $f_{m,0,\alpha}$ and $f_{m,i,\alpha}$ for $i = 1, \ldots, m$, a Gaussian process with the spectral density given by (8) can be expressed as a sum of independent Gaussian processes

$$u(x) = u_0(x) + u_1(x) + \cdots + u_m(x), \tag{10}$$

where u_0 has spectral density $f_{m,0,\alpha}$ and each u_i has spectral density $f_{m,i,\alpha}$ for $i = 1, 2, \ldots, m$. All these spectral densities are reciprocals of polynomials, implying that each process u_i, for $i = 0, \ldots, m$, is a Gaussian Markov process (Pitt, 1971, Theorem 10.1). Specifically, u_0 is a Markov process of order $\max(\lfloor\alpha\rfloor, 1)$, and it is $\max(\lfloor\alpha\rfloor - 1, 0)$ times differentiable in the mean-squared sense. Moreover, for $i > 0$, each u_i is a Markov process of order $\lceil\alpha\rceil$, and it is $\lfloor\alpha\rfloor$ times differentiable in the mean-squared sense. As a result, the multivariate process $\mathbf{u}_0(t) = \left(u_0(t), u_0'(t), \ldots, u_0^{(\max(\lfloor\alpha\rfloor-1,0))}(t)\right)$ is a first-order Markov process with a multivariate covariance function

$$\mathbf{r}_0 : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{\lfloor\alpha\rfloor\times\lfloor\alpha\rfloor}, \quad \mathbf{r}_0(s,t) = \left[\frac{\partial^{i-1}}{\partial s^{i-1}}\frac{\partial^{j-1}}{\partial t^{j-1}}\varrho_{m,0,\alpha}(s-t)\right]_{i,j\in\{1,\ldots,\max(\lfloor\alpha\rfloor,1)\}}.$$

Similarly, the multivariate process $\mathbf{u}_i(t) = \left(u_0(t), u_0'(t), \ldots, u_0^{(\lfloor\alpha\rfloor)}(t)\right)$, for $i \in \{1, \ldots, m\}$, is a first-order Markov process with a multivariate covariance function

$$\mathbf{r}_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^{\lceil\alpha\rceil\times\lceil\alpha\rceil}, \quad \mathbf{r}_i(s,t) = \left[\frac{\partial^{k-1}}{\partial s^{k-1}}\frac{\partial^{\ell-1}}{\partial t^{\ell-1}}\varrho_{m,i,\alpha}(s-t)\right]_{k,\ell\in\{1,\ldots,\lceil\alpha\rceil\}}.$$

Since these multivariate processes are first-order Markov, we can derive the following result regarding their finite-dimensional distributions.

Proposition 6 *Consider a set of unique locations $t_1, \ldots, t_n \in I$. For $j = 1, \ldots, n$, define the vectors*

$$\mathbf{u}_{0,j} = \left[u_0(t_j), u_0'(t_j), \ldots, u_0^{(\max(\lfloor\alpha\rfloor-1,0))}(t_j)\right], \quad \textit{and} \quad \mathbf{u}_{i,j} = \left[u_i(t_j), u_i'(t_j), \ldots, u_i^{(\lfloor\alpha\rfloor)}(t_j)\right],$$

for $i \in \{1, \ldots, m\}$. Then, the concatenated vector $\mathbf{u}_i = [\mathbf{u}_{i,1}, \ldots, \mathbf{u}_{i,n}]$ is a centered Gaussian random variable with a block tridiagonal precision matrix given by

$$\mathbf{Q}_i = \begin{bmatrix} \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} & & & & \\ \mathbf{Q}_{2,1} & \mathbf{Q}_{2,2} & \mathbf{Q}_{2,3} & & & \\ & \mathbf{Q}_{3,2} & \mathbf{Q}_{3,3} & \mathbf{Q}_{3,4} & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathbf{Q}_{n-1,n-2} & \mathbf{Q}_{n-1,n-1} & \mathbf{Q}_{n-1,n} \\ & & & & \mathbf{Q}_{n,n-1} & \mathbf{Q}_{n,n} \end{bmatrix}, \tag{11}$$

for $i = 0, 1, \ldots, m$. The first set of blocks, $\mathbf{Q}_{1,1}, \mathbf{Q}_{1,2}$, and $\mathbf{Q}_{2,1}$, are obtained as

$$\begin{bmatrix} \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} \\ \mathbf{Q}_{2,1} & \mathbf{M} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_i(t_1,t_1) & \mathbf{r}_i(t_1,t_2) \\ \mathbf{r}_i(t_2,t_1) & \mathbf{r}_i(t_2,t_2) \end{bmatrix}^{-1}, \tag{12}$$

where $\mathbf{M}$ is a dummy variable that is discarded. For $j \in \{2, \dots, n-1\}$, the subsequent blocks $\mathbf{Q}_{j,j}, \mathbf{Q}_{j,j+1}, \mathbf{Q}_{j+1,j}$, and $\mathbf{Q}_{j+1,j+1}$ are obtained as

$$\begin{bmatrix} \mathbf{M}_1 & \mathbf{M}_2 & \mathbf{0} \\ \mathbf{M}_2 & \mathbf{Q}_{j,j} & \mathbf{Q}_{j,j+1} \\ \mathbf{0} & \mathbf{Q}_{j+1,j} & \mathbf{Q}_{j+1,j+1} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_i(t_{j-1}, t_{j-1}) & \mathbf{r}_i(t_{j-1}, t_j) & \mathbf{r}_i(t_{j-1}, t_{j+1}) \\ \mathbf{r}_i(t_j, t_{j-1}) & \mathbf{r}_i(t_j, t_j) & \mathbf{r}_i(t_j, t_{j+1}) \\ \mathbf{r}_i(t_{j+1}, t_{j-1}) & \mathbf{r}_i(t_{j+1}, t_j) & \mathbf{r}_i(t_{j+1}, t_{j+1}) \end{bmatrix}^{-1}, \tag{13}$$

where $\mathbf{M}_1$ and $\mathbf{M}_2$ are dummy variables that are also discarded.

The inverses in (12) and (13) can be computed analytically. However, since these matrices have sizes $2\max(\lfloor\alpha\rfloor, 1) \times 2\max(\lfloor\alpha\rfloor, 1)$ and $3\max(\lfloor\alpha\rfloor, 1) \times 3\max(\lfloor\alpha\rfloor, 1)$, respectively, for the process $\mathbf{u}_0$, and sizes $2\lceil\alpha\rceil \times 2\lceil\alpha\rceil$ and $2(\lceil\alpha\rceil + 1) \times 2(\lceil\alpha\rceil + 1)$, respectively, for the process $\mathbf{u}_i$ with $i > 0$, their numerical inversion is computationally trivial. As a result, the benefit of deriving analytical expressions is negligible.

Remark 7 *If $\alpha \in \mathbb{N}$, there is no need for a rational approximation since the process itself is Markov. In this case, we can derive the precision matrix for the process and its derivatives using the same strategy as for $\mathbf{u}_i$ in Proposition 6 but where the covariance function $\mathbf{r}_i$ is replaced by the multivariate covariance function of u and its derivatives. We do not go into more details as there already are exact methods for $\alpha \in \mathbb{N}$. It should, however, be noted that the costs of using this exact method are the same as those we discuss below with $m = 1$.*

We now demonstrate how these expressions can be utilized for computationally efficient sampling, inference, and prediction. Let $t_1, \dots, t_n$ be a set of locations in I, and define the vector $\mathbf{U}_i = [\mathbf{u}_i(t_1), \dots, \mathbf{u}_i(t_n)]$ for $i = 0, \dots, m$. Since these multivariate processes are independent, the vector $\bar{\mathbf{U}} = \left[\mathbf{U}_0^\top, \dots, \mathbf{U}_m^\top\right]^\top$ is a centered multivariate Gaussian random variable of dimension $N = n\left(m\lceil\alpha\rceil + \max(\lfloor\alpha\rfloor, 1)\right)$, with a block diagonal precision matrix $\mathbf{Q} = \operatorname{diag}(\mathbf{Q}_0, \mathbf{Q}_1, \dots, \mathbf{Q}_m)$, where each block is obtained using Proposition 6.

Because $\mathbf{Q}$ is a band matrix, the following result provides the computational cost of computing its Cholesky factor.

Proposition 8 *Computing the Cholesky factor $\mathbf{R}$ of $\mathbf{Q}$, $\mathbf{Q} = \mathbf{R}^\top\mathbf{R}$, requires $\mathcal{O}(nm\lceil\alpha\rceil^3)$ floating point operations, given that $\alpha \ll n$. Further, solving $\mathbf{Y} = \mathbf{R}^{-1}\mathbf{X}$ for some vector $\mathbf{X} \in \mathbb{R}^N$ requires $\mathcal{O}(nm\lceil\alpha\rceil^2)$ floating point operations.*

Thus, the computational cost for sampling $\bar{\mathbf{U}}$ by first computing the Cholesky factor $\mathbf{R}$ of $\mathbf{Q}$ and then solving $\bar{\mathbf{U}} = \mathbf{R}^{-1}\mathbf{Z}$ for a vector $\mathbf{Z}$ with independent standard Gaussian elements can be done in $\mathcal{O}(nm\lceil\alpha\rceil^3)$ computational cost. Moreover, introduce the sparse matrix $\mathbf{A} = [\mathbf{A}_0, \dots, \mathbf{A}_m]$ where $\mathbf{A}_0$ is the sparse $n \times n\max(\lfloor\alpha\rfloor, 1)$ matrix that extracts the values $u_0(t_1), \dots, u_0(t_n)$ from the vector $\mathbf{U}_0$ (i.e., a matrix where each row has one 1 and the rest of the values equal to 0) and where $\mathbf{A}_i$ for $i > 0$ is the $n \times n\lceil\alpha\rceil$ matrix that extracts the values $u_i(t_1), \dots, u_i(t_n)$ from the vector $\mathbf{U}_i$. We then have that $\mathbf{u} = [u(t_1), \dots, u(t_n)]^\top = \mathbf{A}\bar{\mathbf{U}}$. Hence, it follows that $\mathbf{u} \sim \mathcal{N}(0, \mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^\top)$, and we can sample $\mathbf{u}$ in $\mathcal{O}(nm\lceil\alpha\rceil^3)$ cost by first sampling $\bar{\mathbf{U}}$ and then computing $\mathbf{u} = \mathbf{A}\bar{\mathbf{U}}$.

Partitioning $\bar{\mathbf{U}} = (\bar{\mathbf{U}}_A^\top \, \bar{\mathbf{U}}_B^\top)^\top$ for some $A \cup B = \{1, \dots, N\}$ with $A \cap B = \emptyset$, we have that $\bar{\mathbf{U}}_A | \bar{\mathbf{U}}_B \sim \mathcal{N}(-\mathbf{Q}_{A,A}^{-1}\mathbf{Q}_{A,B}\bar{\mathbf{U}}_B, \mathbf{Q}_{A,A}^{-1})$. This means that conditional distributions also can be computed efficiently, and in particular, the conditional mean $\boldsymbol{\mu}_{A|B} = -\mathbf{Q}_{A,A}^{-1}\mathbf{Q}_{A,B}\bar{\mathbf{U}}_B$ can be computed in $\mathcal{O}(|A|(2\lfloor\alpha\rfloor+1)^2)$ computational cost.

These costs do not include the cost of building $\mathbf{Q}$; however, it turns out that one can directly construct an LDL factorization $\mathbf{Q}_i = \mathbf{L}_i^\top\mathbf{D}_i\mathbf{L}_i$ at a slightly lower computional cost than of computing $\mathbf{Q}_i$ through the following result.

Proposition 9 *Using the same notation as in Proposition 6, the precision matrix of $\mathbf{u}_i$ can be constructed as $\mathbf{Q}_i = \mathbf{L}_i^\top\mathbf{D}_i\mathbf{L}_i$. Here $\mathbf{D}_i$ is a diagonal matrix with positive diagonal entries and $\mathbf{L}_i$ is a lower triangular block matrix with ones on the main diagonal. Specifically, $\mathbf{L}_i$ has the form*

$$\mathbf{L}_i = \begin{bmatrix} \mathbf{L}_{1,1} & & & & \\ \mathbf{L}_{2,1} & \mathbf{L}_{2,2} & & & \\ & \mathbf{L}_{3,2} & \mathbf{L}_{3,3} & & \\ & & \ddots & \ddots & \\ & & & \mathbf{L}_{n,n-1} & \mathbf{L}_{n,n} \end{bmatrix}, \tag{14}$$

Here all blocks are of the same size as $\mathbf{u}_{i,j}$ and the matrices $\mathbf{L}_i$ and $\mathbf{D}_i$ can be constructed in $\mathcal{O}(n\max(\lfloor\alpha\rfloor,1)^4)$ and $\mathcal{O}(n\lceil\alpha\rceil^4)$ computational cost for $i=0$ and $i>0$, respectively, through the method in Appendix A.

It should be noted that the costs $\mathcal{O}(n\max(\lfloor\alpha\rfloor,1)^4)$ and $\mathcal{O}(n\lceil\alpha\rceil^4)$ can likely be reduced slightly by taking advantage of that several redundant calculations are performed in the construction. Forming $\mathbf{L} = diag(\mathbf{L}_0, \mathbf{L}_1, \dots, \mathbf{L}_m)$ and $\mathbf{D} = diag(\mathbf{D}_0, \mathbf{D}_1, \dots, \mathbf{D}_m)$, where the blocks are obtained by using Proposition 9, we can now, for example, sample $\bar{\mathbf{U}}$ as $\bar{\mathbf{U}} = \mathbf{L}^{-1}\mathbf{D}^{-1/2}\mathbf{Z}$ at $\mathcal{O}(nm\lceil\alpha\rceil^2)$ computational cost.

Next, consider the situation of a Gaussian process regression where the stochastic process is observed under Gaussian measurement noise. That is, suppose that we have observations $\mathbf{y} = [y_1, y_2, \dots, y_n]$ obtained as $y_i|u(\cdot) \sim \mathcal{N}(u(t_i), \sigma_e^2)$. Then, $\mathbf{y}|\bar{\mathbf{U}} \sim \mathcal{N}(\mathbf{A}\bar{\mathbf{u}}, \sigma_e^2\mathbf{I})$ and $\bar{\mathbf{U}} \sim \mathcal{N}(0, \mathbf{Q}^{-1})$, where the matrix $\mathbf{A}$ and the precision matrix $\mathbf{Q}$ are same as defined above. The goal is typically to estimate the latent process by computing the posterior mean $\boldsymbol{\mu}_{\mathbf{u}|\mathbf{y}} = \mathbb{E}(\mathbf{u}|\mathbf{y})$ of the vector $\mathbf{u} = (u(t_1), \dots, u(t_n))^\top$.

By standard results for latent Gaussian models, we obtain $\bar{\mathbf{U}}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_{\bar{\mathbf{U}}|\mathbf{y}}, \mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}}^{-1})$, where

$$\boldsymbol{\mu}_{\bar{\mathbf{U}}|\mathbf{y}} = \frac{1}{\sigma_e^2}\mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}}^{-1}\mathbf{A}^\top\mathbf{y} \quad \text{and} \quad \mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}} = \mathbf{Q} + \frac{1}{\sigma_e^2}\mathbf{A}^\top\mathbf{A},$$

and $\boldsymbol{\mu}_{\mathbf{u}|\mathbf{y}} = \mathbf{A}\boldsymbol{\mu}_{\bar{\mathbf{U}}|\mathbf{y}}$. The sparsity structure of $\mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}}$ is different from $\mathbf{Q}$; however, we still have linear cost for computing its Cholesky factor through the following proposition.

Proposition 10 *Computing the Cholesky factor $\mathbf{R}_{\bar{\mathbf{U}}|\mathbf{y}}$ of $\mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}}$ requires $\mathcal{O}(nm^3\lceil\alpha\rceil^3)$ floating point operations, given that $m, \alpha \ll n$. Further, solving $\mathbf{Y} = \mathbf{R}^{-1}\mathbf{X}$ for some vector $\mathbf{X} \in \mathbb{R}^N$ requires $\mathcal{O}(nm^2\lceil\alpha\rceil^2)$ floating point operations.*

Thus, the cost of obtaining the Cholesky factor $\mathbf{R}_{\bar{\mathbf{U}}|\mathbf{y}}$ and computing the posterior mean $\mu_{\bar{\mathbf{U}}|\mathbf{y}}$ is $\mathcal{O}(nm^3\lceil\alpha\rceil^3)$. Finally, the log-likelihood of $\mathbf{y}$ is

$$2\ell(\mathbf{y}) = \log|\mathbf{Q}| - 2n\log(\sigma_e) - \log|\mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}}| - \mu_{\bar{\mathbf{U}}|\mathbf{y}}^\top\mathbf{Q}\mu_{\bar{\mathbf{U}}|\mathbf{y}} - \frac{1}{\sigma_e^2}\|\mathbf{y} - \mathbf{A}\mu_{\bar{\mathbf{U}}|\mathbf{y}}\|^2 - \log(2\pi).$$

Computing this requires computing the Cholesky factors of $\mathbf{Q}$ and $\mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}}$, after which we obtain the log-determinants and the posterior mean in linear cost. Therefore, the total cost for evaluating the log-likelihood, and thus, performing likelihood-based inference is $\mathcal{O}(nm^3\lceil\alpha\rceil^3)$. To conclude, all relevant tasks for applying the proposed rational approximation in statistical inference require at most $\mathcal{O}(nm^3\lceil\alpha\rceil^3)$ computational cost, and are thus linear in n.

Remark 11 *Because the rational approximation of the previous section results in a Gaussian process with a spectral density that is a rational function, an alternative to the Markov representations above is to directly apply the state-space methods of Karvonen and Särkkä (2016) to perform inference and prediction at a linear cost $\mathcal{O}(M^3n)$, where M is the approximation order for the state-space method. However, the Markov approximation can directly be used in general Bayesian inference software such as R-INLA (Rue et al., 2009), which is not possible for the state-space methods.*

4. Numerical results

In this section, we illustrate the performance and accuracy of the proposed rational approximation method by comparing it with several alternative approaches. Specifically, we compare our method (referred to as *Rational Approximation*) with the state-space method of Karvonen and Särkkä (2016), the nearest-neighbor Gaussian process approximation (referred to as *nnGP*) of Datta et al. (2016), the random Fourier features method (referred to as *Fourier*) from Rahimi and Recht (2007), a principal component analysis approach (referred to as *PCA*) to serve as a lower bound for low-rank methods, the tapering method (referred to as *Taper*) of Furrer et al. (2006), and the covariance-based rational SPDE approach (referred to as *FEM*) from Bolin et al. (2024b).

We compare the accuracy of the proposed method with the alternative methods by carrying out three tasks. First, we measure and compare the accuracy in terms of the accuracy of the covariance function, and then consider the cost of Gaussian process regression and measure the quality in terms of the accuracy of the posterior mean $\boldsymbol{\mu}_{\mathbf{u}|\mathbf{y}}$ and posterior standard deviation $\boldsymbol{\sigma}_{\mathbf{u}|\mathbf{y}}$. Finally, we compare the accuracy of the process approximation as a whole by computing Kullback–Leibler (KL) divergences and coverage probabilities of joint confidence bands in a Gaussian process regression. In all these cases, we consider the accuracy for a fixed computational cost (as explained below) to make the comparison completely fair. Further, all methods were implemented in R, using the same methods for sparse matrices, matrix solves, and other tasks. The results were obtained using a MacBook Pro Laptop with an M3 Max processor and 128Gb of memory, without using any parallel computations to make the comparison as fair as possible.

4.1 Setup for the first comparison

We first consider the case of a dataset with 5000 observations, $\mathbf{y} = [y_1, y_2, \ldots, y_n]$, with the observation locations being evenly spaced over the interval $I = [0, 50]$. Each observation is generated as $y_i \mid u(\cdot) \sim \mathcal{N}(u(t_i), \sigma_e^2)$, where t_i are the observation locations, and u is a centered Gaussian process with the Matérn covariance function given by (1). We set $\sigma = 1$ (as it is merely a scaling parameter) and explore two different noise levels: $\sigma_e = 0.1$ and

Table 1: Asymptotic costs for different methods. For nnGP, m is the number of neighbors and for the low rank methods, m is the rank. For the state-space method, the "preconditioned" approximation of Karvonen and Särkkä (2016) is used in combination with our Markov approach, see Remark 11. All costs are in "Big O" assuming n is much larger than α and m. For FEM, N is the number of mesh nodes.

Method	Construction	Sampling	Prediction
Proposed	$nm\lceil\alpha\rceil^4$	$nm\lceil\alpha\rceil^2$	$nm^3\lceil\alpha\rceil^3$
state-space (Karvonen and Särkkä, 2016)	$n(m+\lfloor\alpha\rfloor)\lceil\alpha\rceil^4$	$n(m+\lfloor\alpha\rfloor)\lceil\alpha\rceil^2$	$n(m+\lfloor\alpha\rfloor)^3\lceil\alpha\rceil^3$
nnGP (Datta et al., 2016)	nm^3	nm	nm^2
Fourier (Rahimi and Recht, 2007)	nm	nm	nm^2
PCA (Wang, 2008)	n^3	nm	nm^2
Tapering (Furrer et al., 2006)	nm	nm	nm^2
FEM (Bolin et al., 2024b)	$Nm\lceil\alpha\rceil+n$	$Nm\lceil\alpha\rceil^2+n$	$Nm^2\lceil\alpha\rceil^2+n$

$\sigma_e = \sqrt{0.1}$. Additionally, we vary the smoothness parameter ν over the interval $(0, 2.5)$. For each value of ν, we choose $\kappa = \sqrt{2\nu}$, ensuring that the practical correlation range, $\rho = \sqrt{8\nu}/\kappa$, remains fixed at 2.

This choice of range, along with the interval length, ensures numerical stability across all methods compared. Specifically, a practical correlation range of 2 was the largest range for which all methods, including the most sensitive (nnGP), remained stable. While a practical correlation range of 2 on a domain of size 50 may seem small, stability and accuracy of the predictions depend on the number of observations within the correlation range at any given point, rather than the size of the domain itself. Other ranges, and other numbers of observation and prediction locations, are explored in the accompanying Shiny app. For example, one scenario considers 10000 evenly spaced prediction locations and 5000 observation locations randomly selected without replacement from the prediction locations.

The goal is to compute the posterior mean $\boldsymbol{\mu}_{\mathbf{u}|\mathbf{y}}$ with elements $(\boldsymbol{\mu}_{\mathbf{u}|\mathbf{y}})_j = \mathbb{E}(u(p_j)|\mathbf{y})$, and the posterior standard deviation $\boldsymbol{\sigma}_{\mathbf{u}|\mathbf{y}}$ with elements $(\boldsymbol{\sigma}_{\mathbf{u}|\mathbf{y}})_j = \sqrt{\mathrm{Var}(u(p_j)|\mathbf{y})}$, where p_j are the prediction locations. To make the comparison simple, we initially choose $p_j = t_j$ evenly spaced in the interval. Because the locations are evenly spaced, we could also include the Toeplitz method of Ling and Lysy (2022) in the comparison. However, we do not include that here as we only consider generally applicable methods.

4.2 Calibration of the computational costs

The asymptotic costs of the different methods are summarized in Table 1. However, we calibrate the methods to have the same total runtime for assembling all matrices (the construction cost) and computing the posterior mean (the prediction cost). This ensures fair comparisons based on actual performance rather than theoretical complexity, which can be affected by the constants in the cost expression and the size of the study. Specifically, for a given set of parameters (κ, σ, ν), and a fixed value of m for the proposed method, we calibrate the values of m for the other methods to ensure that the total computation time is the same. The total prediction times were averaged over 100 samples to obtain the calibrations.

Table 2: The choice of m for the different methods that result in equal computational cost as for the proposed method. The cost for the Fourier method is the same as for PCA, SS denotes the state-space method and Tap the tapering method.

	$\nu < 0.5$				$0.5 < \nu < 1.5$				$1.5 < \nu < 2.5$			
m	nnGP	PCA	SS	Tap	nnGP	PCA	SS	Tap	nnGP	PCA	SS	Tap
2	1	308	2	1	2	473	1	2	30	810	1	210
3	1	355	3	1	13	561	2	3	37	945	1	342
4	1	406	4	1	21	651	3	62	45	1082	2	376
5	1	433	5	1	27	708	4	124	51	1205	3	405
6	1	478	6	1	31	776	5	166	54	1325	4	501

Table 3: Number of mesh nodes N for the FEM method for different values of m.

m	2	3	4	5	6
$\nu < 0.5$	10495	15494	20493	20493	20493
$0.5 < \nu < 1.5$	25492	35490	40489	45488	50487
$1.5 < \nu < 2.5$	25492	20493	15494	15494	10495

The final calibration results are shown in Table 2. We now provide some details regarding this calibration procedure.

Some of the methods have costs which depend on the smoothness parameter ν. The calibration is therefore done separately for the ranges $0 < \nu < 0.5$, $0.5 < \nu < 1.5$, and $1.5 < \nu < 2.5$. For the taper method, the taper function was chosen as in Furrer et al. (2006) depending on the value of ν, and the taper range was chosen so that each observation, on average, had m neighbors within the range, and the value of m was then chosen to ensure that the total computational cost matches that of the rational approximation. For $\nu < 0.5$, the calibration was not possible for nnGP and taper, because the rational approximation remained faster even with $m = 1$. For these cases, we set $m = 1$.

Given the number of basis functions, the PCA and Fourier methods have the same computational cost for prediction. Thus, the value of m for the Fourier method (the number of basis functions) was set to match the value obtained for the PCA method. The PCA method was calibrated disregarding the construction cost, which is equivalent to assuming that we know the eigenvectors of the covariance matrix. This is not realistic in practice, but makes the method act as a theoretical lower bound for any low-rank method.

As described in Remark 11, the state-space method provides an alternative Markov representation for which we could use the same computational methods as for the proposed method. Its value of m was therefore chosen according to Table 1 as $m - \lfloor \alpha \rfloor$.

To minimize boundary effects of the FEM method, we extended the original domain $[0, 50]$ to $[-4\rho, 50 + 4\rho]$, where ρ is the practical correlation range. This extension ensures accurate approximations of the Matérn covariance at the target locations. We refer the reader to Khristenko et al. (2019, Theorem 3.2) and Bolin et al. (2024b, Proposition 2) for a theoretical justification of this commonly used procedure. Because the FEM method uses the

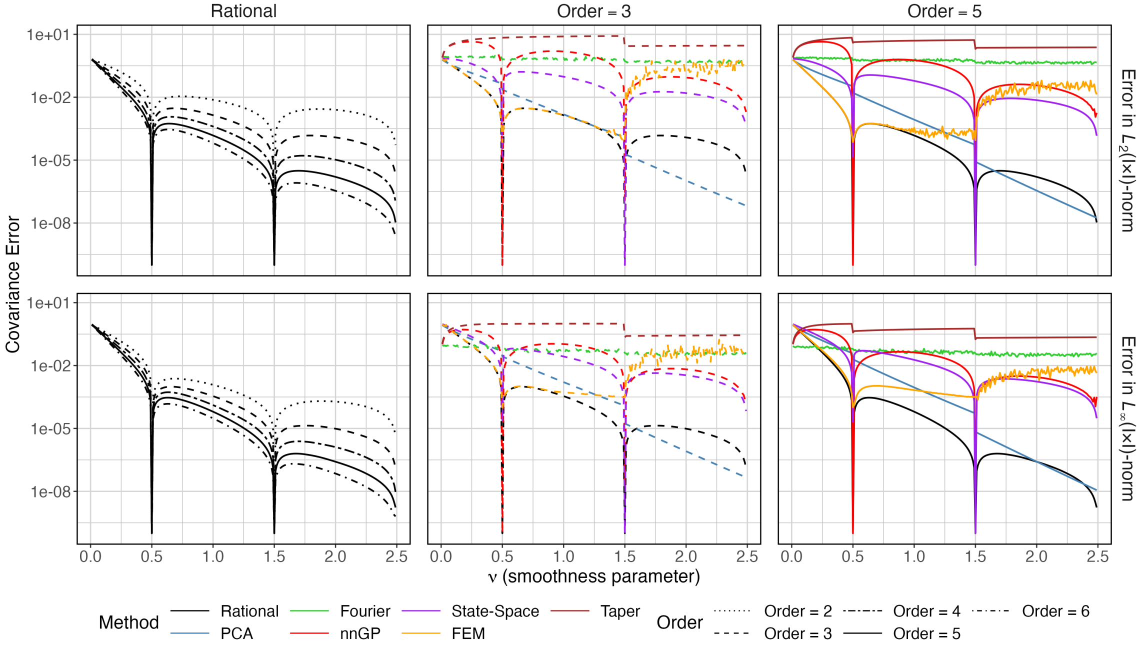

Figure 3: Covariance errors for different values of ν and different orders of approximation.

same type of rational approximation as the proposed method, we fixed the value of m for the FEM method to be equal to m for the proposed method. The calibration was then instead performed on the finite element mesh. Specifically, a mesh with $N = kn + 500 - (k+3)$ locations in the extended domain used, where 500 locations were in the extensions and kn locations in the interior $[0, 50]$, and the $-(k+3)$ term appears to ensure that the regular mesh contains the observation locations. These kn locations were chosen equally spaced to include the observation locations and $k \in \mathbb{N}$ was calibrated to make the total computational cost match that of the proposed method for $\nu < 1.5$, and for $1.5 < \nu < 2.5$ it was chosen as the largest values that yielded a stable prediction, as the value which matched the computational cost yielded unstable predictions. The resulting values of N are shown in Table 3.

4.3 Covariance errors

We measure the quality of the approximations by calculating the L_2 error and the L_∞ error of the respective covariance function approximation, computed as

$$\|r_\alpha - \hat{r}_\alpha\|^2_{L_2(I\times I)} \approx \frac{1}{n^2}\sum_{i,j=1}^{n}(r(t_i,t_j) - \hat{r}(t_i,t_j))^2, \quad \|r - \hat{r}\|_{C(I\times I)} \approx \max_{i,j}|r(t_i,t_j) - \hat{r}(t_i,t_j)|,$$

where r_α denotes the Matérn covariance given in (5), $I = [0, 50]$ is the domain we are considering, and $\hat{r}_\alpha$ represents the approximation obtained using the methods under consideration. Figure 3 illustrates the resulting L_2 and supremum norm errors in the covariance approximation for the different methods when $n = 5000$. The left panel illustrates the error for our method with varying choices of m. Notably, the covariance error decreases rapidly to zero as m increases, with the rate of decrease becoming faster for larger values of ν. Moreover, one can observe that, especially for ν away from zero, each increase in m leads to improvements on the scale of orders of magnitude. In the remaining two panels, we show the error for $m = 3$ and $m = 5$ and compare these errors to the error from the competing methods, calibrated to have the same cost of prediction as described above. We can note that the competing methods are less accurate and in fact, one can observe that the nnGP method becomes slightly unstable for large values of smoothness parameter ν, whereas the FEM method becomes very unstable due to the calibrated mesh being very fine. It is also noteworthy that the mesh is fine enough that when the method is stable (which mean $\nu < 0.9$ for $m = 3$ and $\nu < 0.5$ for $m = 5$), we can only see the rational approximation error, as both FEM and rational methods overlap. This shows that the FEM method is an excellent choice when stable, but that the rational is better because it is more stable and does not require choosing a mesh or a boundary extension. The PCA method is more accurate for large values of ν, but one should recall that it is a theoretical lower bound for low rank methods, and not a practically competitive method, as it has an $\mathcal{O}(n^3)$ construction cost unless the eigenvectors are known explicitly. Any practically useful low rank method, such as the process convolution approach (Higdon, 2002) or fixed rank kriging (Cressie and Johannesson, 2008), will have larger errors, as can clearly be seen when considering the errors of the random Fourier features method. Additionally, it is important to note that for the values $\nu = 0.5, 1.5, 2.5$ within the considered interval $(0, 3)$, the proposed method is exact. Further, Figure 3 highlights an important observation regarding the theoretical bounds and numerical results: The numerical errors behave similar to the (approximate) theoretical bounds in Figure 2, where the error is monotonic in m but not monotonic in $\{\alpha\}$. Observe that the tapering method had the worst performance in terms of covariance errors. This is expected since our benchmark uses the Matérn covariance with the true parameters, and tapering methods are not specifically designed to approximate the original covariance function but rather to induce sparsity. One could also compare with compactly supported covariance functions (Bevilacqua et al., 2019); however, we expect their performance to be similar to that of tapering methods for the same reason. While other tapering approaches exist, such as those proposed in Bolin and Wallin (2016) and Stein (2013), they would likely yield similar results.

4.4 Prediction errors

To compute the prediction error, we first compute the true posterior mean $\boldsymbol{\mu}_{\mathbf{u}|\mathbf{y}}$, the posterior mean $\hat{\boldsymbol{\mu}}_{\mathbf{u}|\mathbf{y}}$ under each approximate model and the corresponding L_2 errors $\|\boldsymbol{\mu}_{\mathbf{u}|\mathbf{y}} - \hat{\boldsymbol{\mu}}_{\mathbf{u}|\mathbf{y}}\|_{L_2}$, where $\|\mathbf{v}\|_{L_2}^2 = \frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^2$ for $\mathbf{v} \in \mathbb{R}^n$. Similarly, we also compute the true posterior standard deviation $\boldsymbol{\sigma}_{\mathbf{u}|\mathbf{y}}$ and the posterior standard deviation $\hat{\boldsymbol{\sigma}}_{\mathbf{u}|\mathbf{y}}$ under each approximate model and the corresponding L_2 error $\|\boldsymbol{\sigma}_{\mathbf{u}|\mathbf{y}} - \hat{\boldsymbol{\sigma}}_{\mathbf{u}|\mathbf{y}}\|_{L_2}$.

The prediction errors under the noise levels $\sigma_e = 0.1$ and $\sigma_e = \sqrt{0.1}$ are shown in Figure 4 and Figure 5, respectively. The left columns of these figures show that the L_2 error of both the posterior means and posterior standard deviations decrease consistently and rapidly to zero for the rational approximation as m increases. The other two columns of the figures compare the errors of the different methods when calibrated to have the same cost as the rational approximation method with $m = 3$ and $m = 5$. These results show

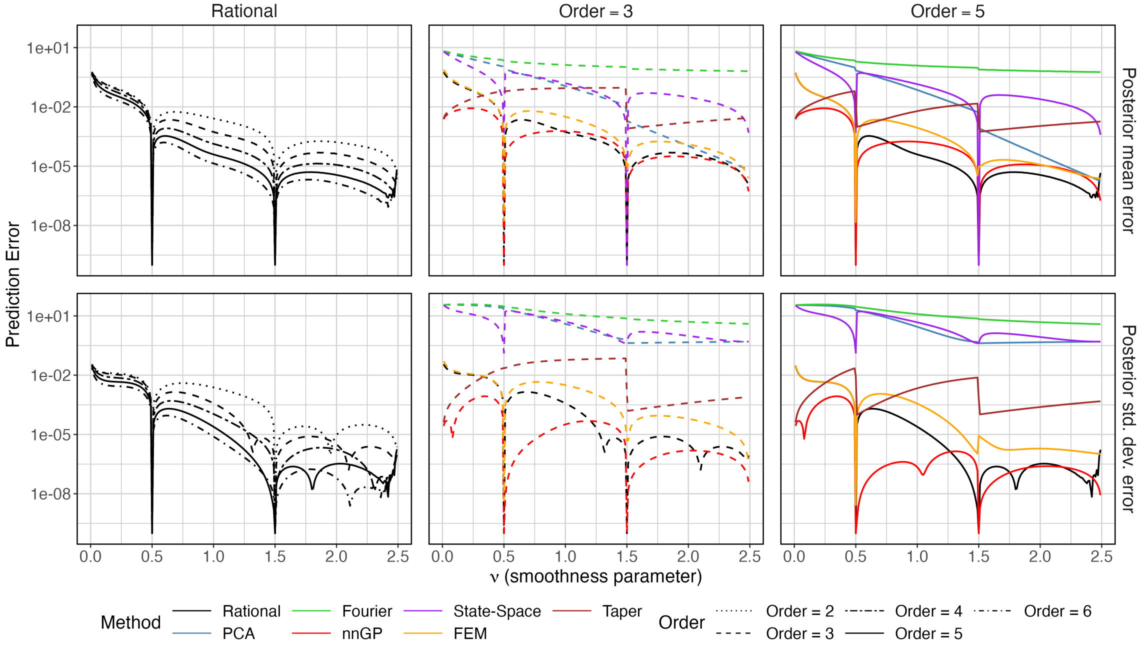

Figure 4: Prediction errors, measured as mean squared errors of the posterior means and posterior standard deviations, evaluated on the interval $I = [0, 50]$ for $\sigma_e = 0.1$, $\rho = 2$, various values of ν, and different orders of approximation.

that state-space method and the Fourier method are not competitive. The PCA method performs poorly for the posterior standard deviations, and also for the posterior mean unless ν is large. Additionally, using PCA at this cost requires prior knowledge of the eigenvectors, so it is typically not a feasible option. Further, the tapering method also performs poorly except for very small values of ν (essentially $\nu < 0.25$ where it is not calibrated). Notably, despite being the worst method in terms of covariance approximation, the tapering method outperforms the Fourier method in prediction errors and, in some cases, even performs better than the PCA and state-space methods. This demonstrates that the accuracy of covariance approximation does not directly translate to prediction accuracy. A method with poor covariance approximation can still yield reasonable prediction performance, which aligns with findings in Furrer et al. (2006) and Stein (2013) which showed that tapering can yield asymptotic optimality of prediction. As with covariance errors, we expect other tapering approaches and compactly supported covariance functions to behave similarly, potentially performing better than our chosen tapering method but unlikely to improve by orders of magnitude. This leaves three methods which are competitive: the rational approximation, the FEM method and nnGP.

The accuracy of the FEM method is essentially equal to that of the rational approximation for $\nu \leq 0.5$, whereas it is slightly worse for $\nu > 0.5$. The nnGP method has the highest accuracy for $\nu < 1$ and $\sigma_e = 0.1$, but one should recall that the method is not calibrated for

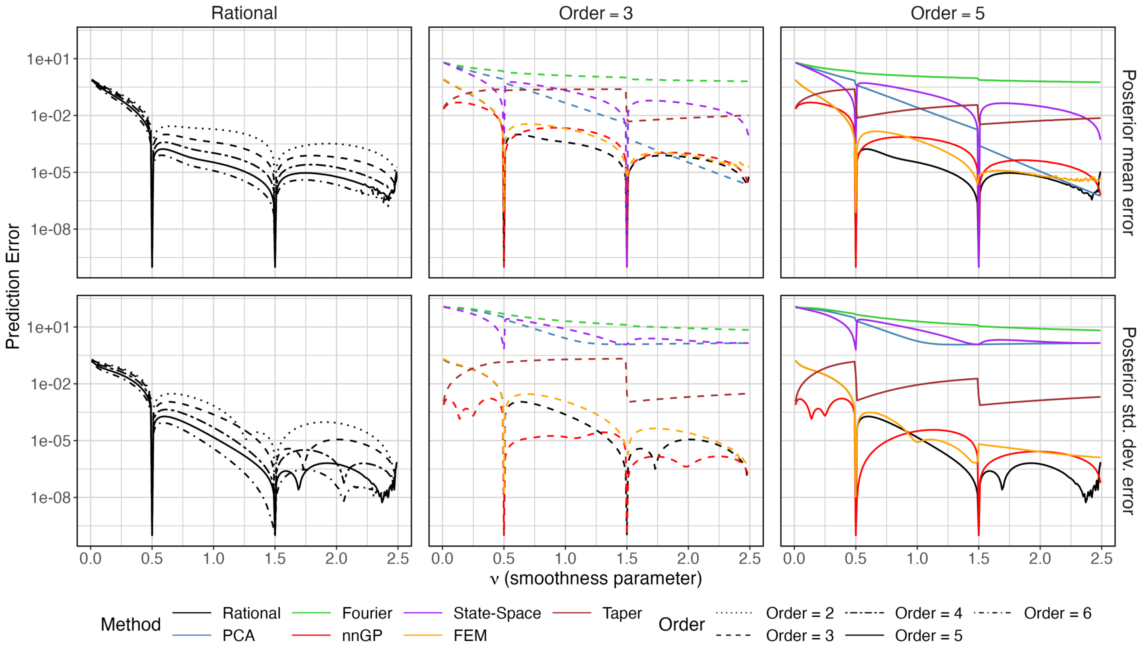

Figure 5: Prediction errors, measured as mean squared errors of the posterior means and posterior standard deviations, evaluated on the interval $I = [0, 50]$ for $\sigma_e = \sqrt{0.1}$, $\rho = 2$, various values of ν, and different orders of approximation.

$\nu < 0.5$ so the comparison is not completely fair. For $\nu > 1$ and $\sigma_e = 0.1$, the two methods have similar performance but the rational approximation is slightly better for the posterior mean. For $\sigma_e = \sqrt{0.1}$, the rational approximation outperforms nnGP and all other methods if $\nu > 1$, whereas for $\nu \in (0.5, 1)$, the rational approximation has the best accuracy for the posterior mean but nnGP has the best accuracy for the standard deviation.

Further comparisons for other choices of the number of prediction locations n, the number of observation locations n_{obs}, and for different practical correlation ranges can be found in the accompanying Shiny application. The general conclusions for these choices align closely with those presented above. The main difference is that, for prediction tasks with a very small practical correlation range and an equal number of prediction and observation locations, nnGP tends to perform the best. This scenario is particularly favorable for nnGP, as all locations coincide with support points. Additionally, when $n > n_{obs}$, the rational approximation outperforms all methods, except for $\nu > 1.5$, where nnGP achieves the best performance in prediction tasks.

4.5 Process approximation

As observed in Section 4.3, no alternative method comes close to the proposed method in terms of the stability and accuracy of the covariance approximation. However, in terms of prediction accuracy, the nnGP and FEM methods exhibit similar performance in certain scenarios. Specifically, for the FEM method, similar accuracies are observed for small values of ν, such as $\nu < 1$ for order 3 and $\nu < 0.5$ for order 5. Also, for certain cases, the nnGP method has outperformed the proposed method in terms of prediction accuracy. However, one important difference of the nnGP method is that the approximation is strongly dependent on the support points, which typically is chosen as the observation locations in applications (Datta et al., 2016). In this section, we illustrate that this results in a poor approximation of the stochastic process, even though the approximation of the finite dimensional distribution at the support points is accurate. To this end, we consider the task of computing a joint confidence band for the latent process in two different scenarios. In both scenarios, we assume that we have data $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ generated as $y_i|u(\cdot) \sim \mathcal{N}(u(t_i), \sigma_e^2)$, where u is a centered Gaussian process with the Matérn covariance (1). As before, we consider the two noise levels $\sigma_e = 0.1$ and $\sigma_e = \sqrt{0.1}$. Based on these observations, we perform prediction at locations $p_1, \ldots, p_n$, and use the `excursions` package (Bolin and Lindgren, 2018) to compute the upper and lower bounds, $c_l(s)$ and $c_u(s)$, of a joint confidence band so that $P(c_l(s) < u(s) < c_u(s)|\mathbf{y}, s \in \{p_1, \ldots, p_n\}) = 0.9$.

In Scenario 1, we consider a forecasting setting where we construct a regular mesh of 1501 points in the interval $[0, 15]$. The first 1001 points, evenly spaced in $[0, 10]$, serve as observation locations. For prediction, we use all observation locations plus the next $n \in \{0, 1, \ldots, 10, 20, 30, 40, 50, 75, 100, 125, \ldots, 500\}$ consecutive points from the mesh as prediction locations. In Scenario 2, we instead use 250 observation locations randomly sampled uniformly in the interval $[0, 10]$, with a minimum spacing of 10^{-3} between locations ensured through resampling if needed, to ensure stability for nnGP. For an increasing sequence of values n between 0 and 3000, we consider n evenly spaced prediction locations in the interval. The combined set of observation and prediction locations is processed to maintain the minimum spacing requirement while preserving all observation locations. To account for the randomness in the selection of observation locations, we repeat this scenario 10 times and average the results.

For each scenario, we computed the approximate coverage probabilities for nnGP, FEM, and the proposed method: $\tilde{p} = \tilde{P}(c_l(s) < u(s) < c_u(s)|\mathbf{y}, s \in \{p_1, \ldots, p_n\})$, where $\tilde{P}(\cdot|\cdot)$ denotes the posterior probability under the approximate model. If the process approximation is accurate, we expect that $\tilde{p} \approx 0.9$. Additionally, we evaluate the accuracy using the Kullback-Leibler (KL) divergence between the approximate and true posterior distributions.

We choose $\nu = 1$ and three values of the practical correlation range: $\rho = 0.5, 1$, and 2. For each scenario and each value of n, the order m of the nnGP approximation is chosen so that the computational cost for evaluating the posterior mean at the prediction locations is the same as for the proposed method, and the support points for the nnGP approximation is kept fixed at the observation locations. For $n < 1000$, we use the calibration cost from $n = 1000$, as the observed cost in smaller cases primarily reflects computational overhead.

The error $0.9 - \tilde{p}$ and KL divergences for the rational, nnGP and FEM methods are shown as functions of n for Scenario 1 and noise level $\sigma_e = 0.1$ in Figure 6 and for Scenario 2 and noise level $\sigma_e = \sqrt{0.1}$ in Figure 7. The results for Scenario 1 with $\sigma_e = \sqrt{0.1}$ and Scenario 2 with $\sigma_e = 0.1$ are very similar, so we do not include them here. However, the results for those cases can be seen in the accompanying Shiny app.

As expected, the error of the nnGP method is small for $n = 0$ but increases quickly with n and has poor performance both in terms of probability approximation and KL divergence

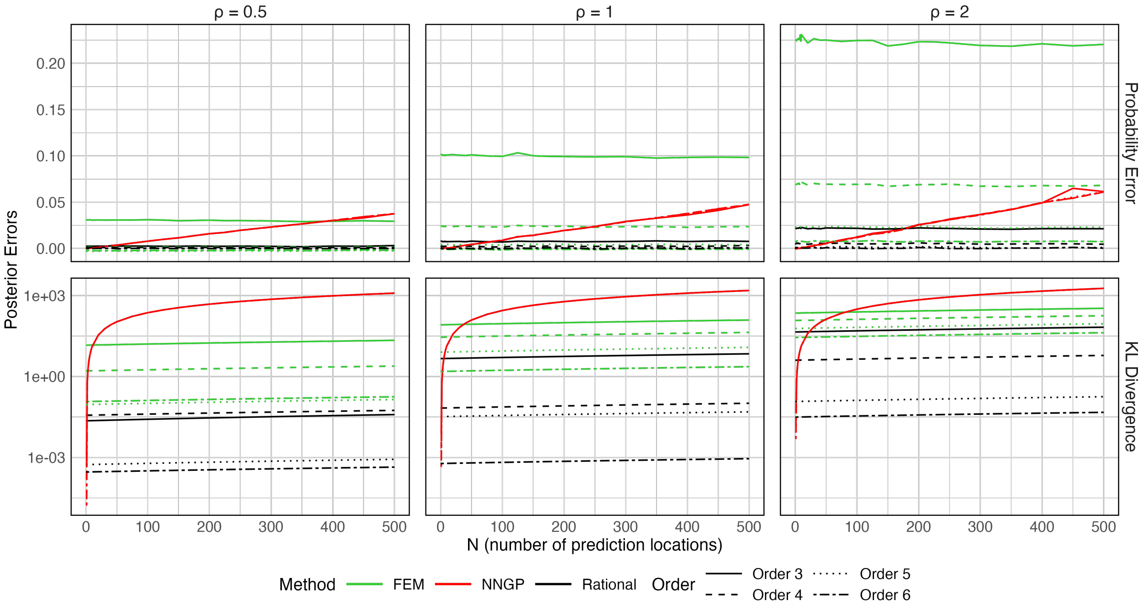

Figure 6: Posterior probability errors and KL divergence under Scenario 1 with $\sigma_e = 0.1$ for $\nu = 1$, orders m from 3 to 6, and three different practical correlation ranges ρ.

even for a very small number of prediction locations. The results for the proposed approximation are much more stable, which supports the claim that the proposed approximation is a better process approximation than the nnGP method. Further, the approximation is also better than the corresponding (calibrated) ones from the FEM method.

As a final study, we consider Scenario 1, in which we do forecasting, and compute the KL divergence between the true distribution of the field $(u(p_1), \dots, u(p_n))$, where $u(\cdot)$ is a Gaussian process with the Matérn covariance function 1, and their approximations by the rational, nnGP and FEM methods. We follow the approach of Scenario 1 and consider the first 1001 points as support points for nnGP. We then compute the KL divergence between the true distribution of the field and the approximations by the different methods. The results are shown in Figure 8, and we observe a similar situation as in the posterior distributions study. The proposed method outperforms the FEM method consistently, and the nnGP method has good performance if all locations are support points. However, the nnGP error increases rapidly as the number of non-support points increases, and as soon as we have a few non-support points, the proposed method also outperforms nnGP.

Remark 12 *We did not include Scenario 2 in the prior distribution study because the nnGP method is not stable for this scenario. The reason for this instability is that the nnGP method is highly dependent on the ordering of the support points and prediction points as shown in (Schäfer et al., 2021, Section 4.2.1). For the posterior calculations we were able to overcome this instability by using a Markov property for performing prediction for nnGP (Datta et al.,*

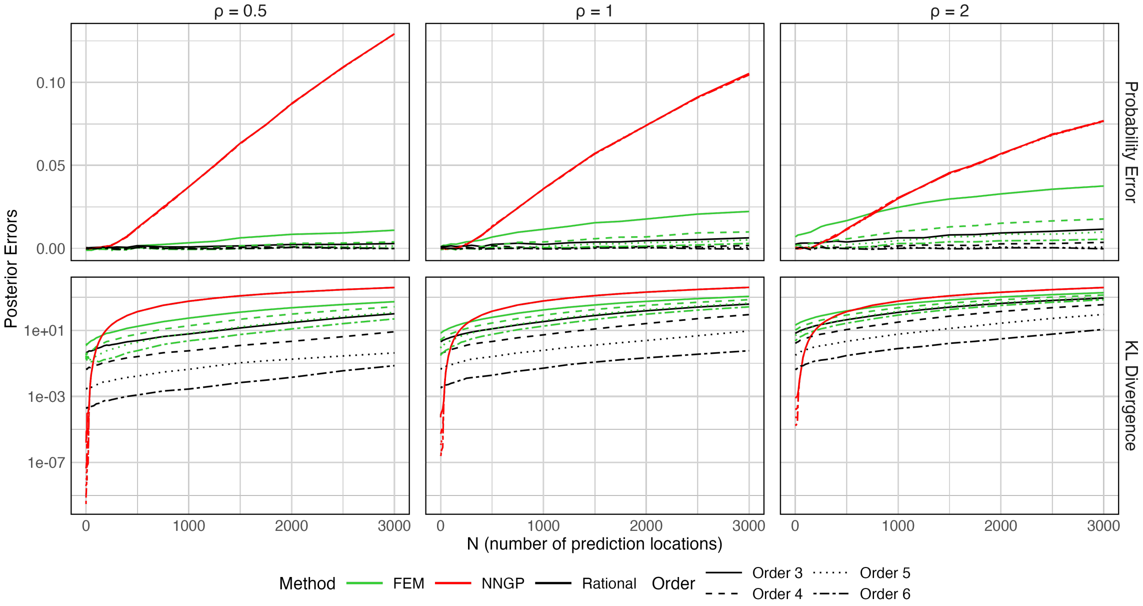

Figure 7: Posterior probability errors and KL divergence under Scenario 2 with $\sigma_e = \sqrt{0.1}$ for $\nu = 1$, orders m from 3 to 6, and three different practical correlation ranges ρ.

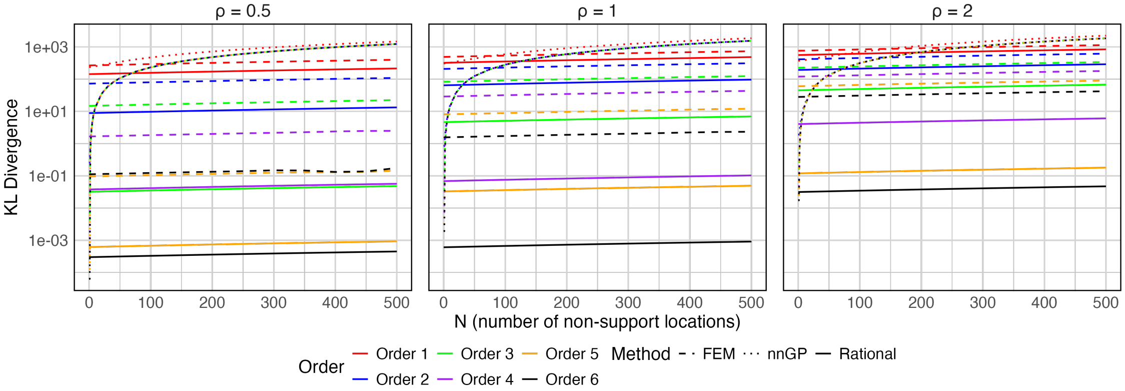

Figure 8: KL divergences (in log scale) under Scenario 1 for N non-support locations, $\nu = 1$, orders m from 1 to 6, and three different practical correlation ranges ρ.

2016), which stabilized the posterior distributions. However, as seen in Figures 6 through 8, it is unlikely that the conclusions would be different for prior distributions under Scenario 2.

5. Discussion

We have introduced a broadly applicable method for working with Gaussian processes (GPs) with Matérn covariance functions on bounded intervals, facilitating statistical inference at linear computational cost. Unlike existing approaches, which often lack theoretical guarantees for accuracy, our method achieves exponentially fast convergence of the covariance approximation. This guarantees both computational efficiency and high accuracy. Furthermore, the method yields a Markov representation of the approximated process, which simplifies its integration into general-purpose software for statistical inference.

Our experimental results demonstrate that the proposed method outperforms state-of-the-art alternatives, offering significant improvements in accuracy. The method has been implemented in the `rSPDE` package, which is compatible with popular tools such as `R-INLA` (Lindgren and Rue, 2015) and `inlabru` (Bachl et al., 2019), facilitating seamless integration into Bayesian hierarchical models. Further, all codes and a Shiny application can be found at `https://github.com/vpnsctl/MarkovApproxMatern/`. The Shiny application not only contains all plots presented in this paper but also additional plots with different numbers of observations, as well as plots of relative errors.

The two most competitive methods besides the proposed method were nnGP and FEM. The nnGP model provides accurate predictions, particularly for lower values of the smoothness parameter ν, as illustrated in Figure 4. However, it falls short in terms of process approximation. Specifically, the covariance errors associated with nnGP are significantly larger than those from the rational approximation method, as shown in Figure 3. Additionally, the probability approximations deteriorate markedly when observations are not contained in the support points, evidenced by Figures 6 through 8. This limitation suggests that while nnGP may provide good results in contexts focused on prediction, such as those commonly found in machine learning, it is less competitive in broader statistical applications which require accurate process approximations. The main advantage of the proposed method compared to the FEM method is a generally higher accuracy and stability (in particular for large values of ν), and the fact that no mesh needs to be chosen.

To ensure a fair comparison, we assumed all parameters to be known throughout the simulation studies, providing a clear baseline using the true covariance function. We expect similar results also in cases with unknown parameters, such as when performing likelihood-based statistical inference, as iterative numerical procedures rely on intermediate approximations of the true covariance. However, this remains to be investigated.

An exciting direction for future work involves extending our method to more general domains. As mentioned in Section 2, the method can also be derived through a rational approximation of the covariance operator of the process. More precisely, consider a Gaussian process u on a domain $\mathcal{D}$ defined as the solution to

$$L^{\beta}(\tau u) = \mathcal{W}, \quad \text{on } \mathcal{D}, \tag{15}$$

where $\mathcal{W}$ is Gaussian white noise on $\mathcal{D}$, $\beta > 0$ is a real number, and L is a densely defined self-adjoint operator on $L_2(\mathcal{D})$, so that the covariance operator of u is given by $L^{-2\beta}$. For example, if we have $L = (\kappa^2 - \Delta)$ on $\mathbb{R}$, we obtain a Gaussian process with a stationary Matérn covariance function (Whittle, 1963). Following the approach in Bolin et al. (2024b), one can approximate the covariance operator as

$$L^{-2\beta} \approx L^{-\lfloor 2\beta \rfloor} \left(\sum_{i=1}^{m} c_i (L - p_i I)^{-1} + kI \right), \tag{16}$$

where $\{c_i\}_{i=1}^m$, $\{p_i\}_{i=1}^m$ and k are real numbers, satisfying $p_i \leq 0$, $c_i > 0$, and $k > 0$ by Proposition 3, and $I : L_2(\mathcal{D}) \to L_2(\mathcal{D})$ is the identity operator. If suitable conditions on L hold, this approximation expresses $L^{-2\beta}$ as a linear combination of covariance operators. We will refer to this approach as the covariance-based approach.

By computing the finite-dimensional distributions corresponding to the operators $L^{-\lfloor 2\beta \rfloor}$ and $L^{-\lfloor 2\beta \rfloor}(L - p_i I)^{-1}$ for $i = 1, \ldots, m$, we can approximate the finite-dimensional distributions of $L^{-2\beta}$ and thus the solution to (15). Assume further that these covariance operators are induced by covariance functions. Let $\varrho_\beta(\cdot, \cdot)$ denote the covariance function associated with $L^{-2\beta}$, $\widehat{\varrho}_\beta(\cdot, \cdot)$ the covariance function corresponding to $kL^{-\lfloor 2\beta \rfloor}$, and $\widehat{\varrho}_{\beta,i}(\cdot, \cdot)$ the covariance function associated with $c_i L^{-\lfloor 2\beta \rfloor}(L - p_i I)^{-1}$ for $i = 1, \ldots, m$. Then, (16) implies that

$$\varrho_\beta(x, y) \approx \widehat{\varrho}_\beta(x, y) + \sum_{i=1}^{m} \widehat{\varrho}_{\beta,i}(x, y), \quad x, y \in \mathcal{D}, \tag{17}$$

which corresponds exactly to the expression in Proposition 5, demonstrating the equivalence between the spectral and covariance-based approaches.

It is important to note, however, that both approaches—the covariance-based and the spectral approach—yield approximations that do not converge to the true covariance operator when $\mathcal{D} = \mathbb{R}^d$ for $d \geq 1$. Nonetheless, one can instead consider compact domains, $\mathcal{D}$, where convergence of the approximation can generally be established.

This formulation offers a pathway to extending our method to more complex (compact) domains. For instance, starting with $L = (\kappa^2 - \Delta)$ on a circle, one can define rational approximations for periodic Matérn fields. More generally, the approach can be adapted to network-like domains, following recent developments in Gaussian Whittle–Matérn fields on metric graphs (Bolin et al., 2023, 2024a). These extensions, including their theoretical and computational aspects, will be explored in future work. Finally, if $\mathcal{D}$ is a d-dimensional manifold (which also includes compact subsets of $\mathbb{R}^d$), where $d \geq 2$, this approach is not expected to provide sparse approximations unless $\mathcal{D}$ is a Cartesian product of intervals, and L is a separable differential operator, that is, $L = \sum_{i=1}^d L_i$, where L_i is a differential operator acting only on x_i, where $x = (x_1, \ldots, x_d) \in \mathcal{D}$. This implies that, while the approximation may still be applicable for manifolds of dimension greater than one, it is generally not practical unless the model is separable. Such separable models, which have been explored in previous work (see, e.g., Chen et al. (2022)), are rarely suitable for modeling spatial data.

Another advantage with the covariance-based formulation is that it can be combined with FEM approximations, and because by taking advantage of the fact that (17) is stationary, one could use this to avoid having to rely on boundary extensions to remove boundary effects for the FEM approach. As this is a common problem with the FEM approach that increases computational cost, it represents an interesting topic to investigate in future work.

Appendix A. LDL Algorithm and proof of Proposition 9

In this section, we explain how to construct the matrices $\mathbf{L}_i$ and $\mathbf{D}_i$ of Proposition 9. To simplify the notation, let p denote the size of the vector $\mathbf{u}_{i,j}$ (which is either $\max(\lfloor\alpha\rfloor, 1)$ if $i = 0$ or $\lceil\alpha\rceil$ if $i > 0$). Introduce the matrices $\mathbf{B}^1 = \mathbf{L}_{1,1}$ and $\mathbf{B}^k = [\mathbf{L}_{k,k-1}\, \mathbf{L}_{k,k}], k > 1$, corresponding to the part of $\mathbf{L}_i$ that is related to the kth location. Similarly, let $\mathbf{F}^k$ denote the diagonal matrix corresponding to the part of $\mathbf{D}^{-1}$ that is related to the kth location. That is, $\mathbf{D} = \text{diag}(\mathbf{F}^1, \dots, \mathbf{F}^n)^{-1}$. These are low-dimensional matrices with a dimension $2p \times p$ and $p \times p$, respectively. Further, introduce the covariance matrices $\mathbf{\Sigma}^1 = \mathbf{r}_i(t_1, t_1)$ and

$$\mathbf{\Sigma}^k = \begin{bmatrix} \mathbf{r}_i(t_1, t_1) & \mathbf{r}_i(t_1, t_2) \\ \mathbf{r}_i(t_2, t_1) & \mathbf{r}_i(t_2, t_2) \end{bmatrix}, \quad k = 2, \dots, n,$$

and for an $n \times n$ matrix $\mathbf{M}$, let $\mathbf{M}_{a:b,c:d}$, for natural numbers $a \leq b \leq n$ and $c \leq d \leq n$ denote the submatrix obtained by extracting rows $a, \dots, b$ and columns $c, \dots, d$ from $\mathbf{M}$.

We are now ready to construct the elements in the matrices. We set $\mathbf{F}^1_{1,1} = 1/\mathbf{\Sigma}^1_{1,1}$ and

$$\begin{aligned} \mathbf{F}^1_{j,j} &= \mathbf{\Sigma}^1_{j,j} - \mathbf{\Sigma}^1_{j,1:(j-1)}(\mathbf{\Sigma}^1_{1:(j-1),1:(j-1)})^{-1}\mathbf{\Sigma}^1_{1:(j-1),j}, & j &= 2, \dots, p \\ \mathbf{F}^k_{j,j} &= \mathbf{\Sigma}^k_{p+j,p+j} - \mathbf{\Sigma}^k_{p+j,1:(p+j-1)}(\mathbf{\Sigma}^k_{1:(p+j-1),1:(p+j-1)})^{-1}\mathbf{\Sigma}^k_{1:(p+j-1),j}, & j &= 1, \dots, p, k > 1. \end{aligned}$$

Further, we set $\mathbf{B}^1_{j,j} = 1$ and $\mathbf{B}^1_{j,j+i} = 0$ for $i > 0$ and

$$\mathbf{B}^1_{j,1:(j-1)} = -\mathbf{\Sigma}^1_{j,1:(j-1)}(\mathbf{\Sigma}^1_{1:(j-1),1:(j-1)})^{-1}, \quad i = 2, \dots, p.$$

Finally, we set $\mathbf{B}^k_{j,p+j} = 1$ and $\mathbf{B}^k_{j,p+j+i} = 0$ for $i > 0$ and

$$\mathbf{B}^k_{j,1:(j-1)} = -\mathbf{\Sigma}^k_{j,1:(p+j-1)}(\mathbf{\Sigma}^1_{1:(p+j-1),1:(p+j-1)})^{-1}, \quad i = 1, \dots, p, k > 1.$$

These expressions follow directly from using the fact that $\mathbf{u}_i$ is a first order Markov process and then writing the joint density of $\mathbf{u}_i$ as

$$\pi(\mathbf{u}_i) = \pi(\mathbf{u}_{i,1}) \prod_{k=w}^{n} \pi(\mathbf{u}_{i,k}|\mathbf{u}_{i,k-1}).$$

After this, standard results for conditional Gaussian densities give the expressions above (see, for example Datta et al., 2016). To compute the elements in each block, we need to perform p solves of matrices of size $p, p+1, \dots, 2p$, the total cost of this is $\mathcal{O}(p^4)$ and since we have n blocks, the total cost is thus $\mathcal{O}(np^4)$.

Appendix B. Collected proofs

We start with a simple technical lemma needed for the proof of Theorem 1.

Lemma 13 *Let $h : \mathbb{R} \to \mathbb{R}$ be a measurable function. Then, the following inequality holds:*

$$\int_a^b \int_a^b h(x-y)^2 dx\, dy \leq (b-a) \int_{a-b}^{b-a} h(x)^2 dx,$$

where $a, b \in \mathbb{R}, a < b$.

Proof We have, by the change of variables $u = x - y$, that

$$\begin{aligned}\int_a^b \int_a^b h(x-y)^2 dx\, dy &= \int_a^b \int_{a-y}^{b-y} h(u)^2 du\, dy \\ &\leq \int_a^b \int_{a-b}^{b-a} h(u)^2 du\, dy = (b-a) \int_{a-b}^{b-a} h(u)^2 du,\end{aligned}$$

where we used the fact that $a \leq y \leq b$. ■

Proof [of Theorem 1] Our idea is to use the exponential convergence of the best rational approximation with respect to the L_∞ norm. To such an end, Stahl (2003, Theorem 1), gives us that for every $\alpha \in (0,1)$, there exist polynomials of degree m, $p_m(\cdot)$ and $q_m(\cdot)$ such that the following exponential bound holds:

$$\sup_{x\in[0,1]} \left| x^\alpha - \frac{p_m(x)}{q_m(x)} \right| \leq C_\alpha e^{-2\pi\sqrt{\alpha m}}, \tag{18}$$

where $C_\alpha > 0$ is a constant that only depends on α. Now, let $g(w) = (1 + \kappa^{-2}w^2)^{-1}$ and observe that for every $w \in \mathbb{R}$, we have $g(w) \in [0,1]$. Further, let $\alpha = \nu + 0.5$ and $f_\alpha(\cdot)$ be the spectral density of a stationary Gaussian process with a Matérn covariance function (1). Then, $f_\alpha(w) = A\sigma^2\kappa^{-2\alpha}(g(w))^\alpha$, where $A = \sqrt{2}\kappa^{2\nu}\Gamma(\nu + 1/2)\Gamma(\nu)^{-1}$. Observe that we have the following decompositions:

$$f_\alpha(w) = f_{\lfloor\alpha\rfloor}(w)(g(w))^{\{\alpha\}} \quad \text{and} \quad f_{m,\alpha}(w) = f_{\lfloor\alpha\rfloor}(w)\frac{p_m(g(w))}{q_m(g(w))}.$$

We also have, for any $\alpha > 1/2$, that

$$\sup_{w\in\mathbb{R}} \left| (g(w))^{\{\alpha\}} - \frac{p_m(g(w))}{q_m(g(w))} \right| = \sup_{x\in[0,1]} \left| x^{\{\alpha\}} - \frac{p_m(x)}{q_m(x)} \right| \leq C_{\{\alpha\}} e^{-2\pi\sqrt{\{\alpha\}m}},$$

where we used (18) and that $0 \leq x \leq 1$. Because $\sup_{w\in\mathbb{R}} |f_{\lfloor\alpha\rfloor}(w)| = |f_{\lfloor\alpha\rfloor}(0)| = A\sigma^2\kappa^{-2\alpha}$,

$$\begin{aligned}\sup_{w\in\mathbb{R}} |f_\alpha(w) - f_{m,\alpha}(w)| &\leq \sup_{w\in\mathbb{R}} |f_{\lfloor\alpha\rfloor}(w)| \sup_{w\in\mathbb{R}} \left| (g(w))^{\{\alpha\}} - \frac{p_m(g(w))}{q_m(g(w))} \right| \\ &\leq A\sigma^2\kappa^{-2\alpha} C_{\{\alpha\}} e^{-2\pi\sqrt{\{\alpha\}m}}.\end{aligned} \tag{19}$$

Now, let $I = [a,b] \subset \mathbb{R}$ be an interval in $\mathbb{R}$ and $1_I : \mathbb{R} \to \mathbb{R}$ denote the indicator function such that $1_I(x) = 1$ if $x \in I$ and $1_I(x) = 0$ otherwise. We then have by Plancherel's theorem (e.g. McLean, 2000, Corollary 3.13), the convolution theorem (e.g. McLean, 2000, p.73), Lemma 13, and the bound (19) above, that

$$\begin{aligned}\|r_{m,\alpha}-r_\alpha\|^2_{L_2(I\times I)} &\leq (b-a)\|\varrho_{m,\alpha}-\varrho_\alpha\|^2_{L_2(a-b,b-a)} = (b-a)\|(\varrho_{m,\alpha}-\varrho_\alpha)1_{[a-b,b-a)]}\|^2_{L_2(\mathbb{R})}\\ &= (b-a)\left\|(f_{m,\alpha}-f_\alpha)*2(b-a)\frac{\sin((b-a)w)}{(b-a)w}\right\|^2_{L_2(\mathbb{R})}\\ &\leq (b-a)\sup_{w\in\mathbb{R}}\left((g(w))^{\{\alpha\}}-\frac{p_m(w)}{q_m(w)}\right)^2\left\|2(b-a)\frac{\sin((b-a)w)}{(b-a)w}\right\|^2_{L_2(\mathbb{R})}\\ &\leq 4(b-a)^3A^2\sigma^4\kappa^{-4\alpha}C^2_{\{\alpha\}}e^{-4\pi\sqrt{\{\alpha\}m}}\left\|\frac{\sin((b-a)w)}{(b-a)w}\right\|^2_{L_2(\mathbb{R})}\\ &= \pi(b-a)^2A^2\sigma^4\kappa^{-4\alpha}C^2_{\{\alpha\}}e^{-4\pi\sqrt{\{\alpha\}m}},\end{aligned}$$

where $*$ denotes the convolution and we used the fact that $\left\|\frac{\sin((b-a)w)}{(b-a)w}\right\|^2_{L_2(\mathbb{R})} = \pi/(b-a)$. Let us now consider another strategy that helps us obtain another bound for $\alpha > 1$, that gets better as α increases. First, observe that by Young's convolution inequality (Bogachev, 2007, Theorem 3.9.4), we have $\|f*g\|_{L_2(\mathbb{R})} \leq \|f\|_{L_1(\mathbb{R})}\|g\|_{L_2(\mathbb{R})}$, which yields,

$$\begin{aligned}\left\|f_{\lfloor\alpha\rfloor}(w)*\frac{\sin((b-a)w)}{(b-a)w}\right\|_{L_2(\mathbb{R})} &\leq \|f_{\lfloor\alpha\rfloor}(w)\|_{L_1(\mathbb{R})}\left\|\frac{\sin((b-a)w)}{(b-a)w}\right\|_{L_2(\mathbb{R})}\\ &= A\sigma^2\kappa^{-2\alpha}M_{\lfloor\alpha\rfloor,\kappa}\sqrt{\frac{\pi}{b-a}},\end{aligned}$$

where $M_{0,\kappa}=\infty$ and $M_{n,\kappa} = \|(g(w))^n\|_{L_1(\mathbb{R})} = \kappa\pi\Gamma(2n-1)/(4^{n-1}\Gamma(n)^2), n\geq 1$. Therefore, by proceeding as in the first case, and using the above inequality, we have

$$\begin{aligned}\|r_{m,\alpha}-r_\alpha\|^2_{L_2(I\times I)} &\leq (b-a)\left\|(f_{m,\alpha}-f_\alpha)*2(b-a)\frac{\sin((b-a)w)}{(b-a)w}\right\|^2_{L_2(\mathbb{R})}\\ &\leq 4(b-a)^3\sup_{w\in\mathbb{R}}\left((g(w))^{\{\alpha\}}-\frac{p_m(w)}{q_m(w)}\right)^2\left\|f_{\lfloor\alpha\rfloor}(w)*\frac{\sin((b-a)w)}{(b-a)w}\right\|^2_{L_2(\mathbb{R})}\\ &\leq 4\pi(b-a)^2A^2\sigma^4\kappa^{-4\alpha}M^2_{\lfloor\alpha\rfloor,\kappa}C^2_{\{\alpha\}}e^{-4\pi\sqrt{\{\alpha\}m}}.\end{aligned}$$

Finally, by combining both inequalities above, we obtain

$$\|r_{m,\alpha}-r_\alpha\|_{L_2(I\times I)} \leq 2\sqrt{\pi}(b-a)A\sigma^2\kappa^{-2\alpha}\min\{1,M_{\lfloor\alpha\rfloor,\kappa}\}C_{\{\alpha\}}e^{-2\pi\sqrt{\{\alpha\}m}}.$$

This proves (6).

Now, for the $L_\infty(I\times I)$ estimate. Recall the notation from the first part of the proof. Also, recall the definition of the Sobolev space $H^s(\mathbb{R})$ for $s\geq 0$ (see, e.g. McLean, 2000, p.75-76). Then, for $s > 1/2$, we have by definition of the Sobolev norm that

$$\begin{aligned}\|\varrho_\alpha-\varrho_{m,\alpha}\|_{H^s(\mathbb{R})} &= \|(1+w^2)^{s/2}(f_\alpha-f_{m,\alpha})\|_{L_2(\mathbb{R})}\\ &\leq \sup_{w\in\mathbb{R}}\left|(g(w))^{\{\alpha\}}-\frac{p_m(w)}{q_m(w)}\right|\|f_{\lfloor\alpha\rfloor}(1+w^2)^{s/2}\|_{L_2(\mathbb{R})}\\ &\leq A\sigma^2\kappa^{-2\alpha}C_{\{\alpha\}}e^{-2\pi\sqrt{\{\alpha\}m}}\|(1+w^2)^{s/2-\lfloor\alpha\rfloor}\|_{L_2(\mathbb{R})},\end{aligned}$$

Now, because $\alpha > 1$, we have that $\lfloor\alpha\rfloor \geq 1$. Further, $\|(1+w^2)^{s/2-\lfloor\alpha\rfloor}\|_{L_2(\mathbb{R})} < \infty$ if, and only if, $2\lfloor\alpha\rfloor - s > 1/2$. In particular, since $\lfloor\alpha\rfloor \geq 1$ we can find $s_\alpha > 1/2$ such that $\|(1+w^2)^{s_\alpha/2-\lfloor\alpha\rfloor}\|_{L_2(\mathbb{R})} < \infty$. Therefore, if $\alpha > 1$, we can find $s_\alpha > 1/2$ such that

$$\|\varrho_\alpha - \varrho_{m,\alpha}\|_{H^{s_\alpha}(\mathbb{R})} \leq A\sigma^2\kappa^{-2\alpha}\breve{C}_\alpha C_{\{\alpha\}} e^{-2\pi\sqrt{\{\alpha\}m}},$$

where the constant $\breve{C}_\alpha$ depends on s_α, which only depends on α. Now, from Sobolev embedding, see, e.g., (McLean, 2000, Theorem 3.26), since $s_\alpha > 1/2$, we have that there exists $C_{sob,\alpha} > 0$, depending only on s_α, thus only on α, such that for every $x \in \mathbb{R}$, $|\varrho_\alpha(x) - \varrho_{m,\alpha}(x)| \leq C_{sob,\alpha}\|\varrho_\alpha - \varrho_{m,\alpha}\|_{H^{s_\alpha}(\mathbb{R})}$, so that

$$\sup_{x\in\mathbb{R}} |\varrho_\alpha(x) - \varrho_{m,\alpha}(x)| \leq C_{sob,\alpha}\|\varrho_\alpha - \varrho_{m,\alpha}\|_{H^s(\mathbb{R})} \leq A\sigma^2\kappa^{-2\alpha}C_{sob,\alpha}\breve{C}_\alpha C_{\{\alpha\}} e^{-2\pi\sqrt{\{\alpha\}m}}.$$

Therefore,

$$\|r_{m,\alpha} - r_\alpha\|_{L_\infty(I\times I)} \leq \sup_{x\in\mathbb{R}} |\varrho_\alpha(x) - \varrho_{m,\alpha}(x)| \leq A\sigma^2\kappa^{-2\alpha}K_\alpha e^{-2\pi\sqrt{\{\alpha\}m}},$$

where $K_\alpha > 0$ is a constant that only depends on α. This proves (7). ■

We now move to the proof of Proposition 3. To this end, we start by recalling the following result (Saff and Stahl, 1995, Lemma 2.1):

Lemma 14 *Let* $0 < \alpha < 1$.

1. *The best uniform rational approximation* $R_m^*(x)$ *of the function* x^α *on* $[0,1]$*, among rational functions whose numerator has degree at most* m*, is such that both the numerator and denominator have degree exactly* m*.*

2. *All* m *zeros* $\tilde{z}_1, \ldots, \tilde{z}_m$ *and poles* $\tilde{p}_1, \ldots, \tilde{p}_m$ *of* $R_m^*(\cdot)$ *lie on the negative half-axis and are interlacing; i.e., with an appropriate numbering,*

$$0 > \tilde{z}_1 > \tilde{p}_1 > \tilde{z}_2 > \tilde{p}_2 > \cdots > \tilde{z}_m > \tilde{p}_m > -\infty. \tag{20}$$

We are now in a position to prove Proposition 3.

Proof [of Proposition 3] Define $\widetilde{P}_m(x) = \sum_{i=0}^m a_i x^i$ and $\widetilde{Q}_m(x) = \sum_{i=0}^m b_i x^i$ so that the best rational approximation of x^α on $[0,1]$ is $R_m^*(x) = \widetilde{P}_m(x)/\widetilde{Q}_m(x)$. Then, by Lemma 14, $\widetilde{P}_m(x) = m_{\widetilde{P}}(x-\tilde{z}_1)(x-\tilde{z}_2)\cdots(x-\tilde{z}_m)$, and $\widetilde{Q}_m(x) = m_{\widetilde{Q}}(x-\tilde{p}_1)(x-\tilde{p}_2)\cdots(x-\tilde{p}_m)$, where $m_{\widetilde{P}}$ and $m_{\widetilde{Q}}$ are constants and $\{\tilde{z}_i\}_{i=1}^m, \{\tilde{p}_i\}_{i=1}^m$ satisfy (20). Let $\operatorname{sgn}(x) = |x|/x$ denote the sign of $x \neq 0$. Since $\tilde{z}_i, \tilde{p}_i < 0$ for $i = 1, \ldots, m$, we have that for every $x \in (0,1)$, $\operatorname{sgn}(\widetilde{P}_m(x)) = \operatorname{sgn}(m_{\widetilde{P}})$ and $\operatorname{sgn}(\widetilde{Q}_m(x)) = \operatorname{sgn}(m_{\widetilde{Q}})$. Since $x^{\{\alpha\}} > 0$ for $x \in (0,1)$, $\widetilde{P}_m(x)$ and $\widetilde{Q}_m(x)$ must have the same sign for $x \in (0,1)$, so that

$$\operatorname{sgn}(m_{\widetilde{P}}) = \operatorname{sgn}(m_{\widetilde{Q}}). \tag{21}$$

Now, observe that $P_m(x) = x^m \widetilde{P}_m(x^{-1})$ and $Q_m(x) = x^m \widetilde{Q}_m(x^{-1})$. Thus,

$$\begin{aligned} P_m(x) &= x^m \widetilde{P}_m(x^{-1}) = m_{\widetilde{P}}(1 - x\tilde{z}_1)\cdots(1 - x\tilde{z}_m) \\ &= m_{\widetilde{P}}\tilde{z}_1\cdots\tilde{z}_m(-1)^m \left(x - \frac{1}{\tilde{z}_1}\right)\cdots\left(x - \frac{1}{\tilde{z}_m}\right) \\ &= m_P(x - z_1)\cdots(x - z_m), \end{aligned} \tag{22}$$

where $m_P = m_{\widetilde{P}}\tilde{z}_1\cdots\tilde{z}_m(-1)^m$ and $z_i = 1/\tilde{z}_i$ for $i = 1, \ldots, m$. Similarly, we have that

$$\begin{aligned} Q_m(x) &= x^m \widetilde{Q}_m(x^{-1}) = m_{\widetilde{Q}}(1 - x\tilde{p}_1)\cdots(1 - x\tilde{p}_m) \\ &= m_{\widetilde{Q}}\tilde{p}_1\cdots\tilde{p}_m(-1)^m \left(x - \frac{1}{\tilde{p}_1}\right)\cdots\left(x - \frac{1}{\tilde{p}_m}\right) \\ &= m_Q(x - p_1)\cdots(x - p_m), \end{aligned} \tag{23}$$

where $m_Q = m_{\widetilde{Q}}\tilde{p}_1\cdots\tilde{p}_m(-1)^m$ and $p_i = 1/\tilde{p}_i$ for $i = 1, \ldots, m$. Now, observe that

$$\operatorname{sgn}(m_P) = (-1)^m \operatorname{sgn}(\tilde{z}_1\cdots\tilde{z}_m)\operatorname{sgn}(m_{\widetilde{P}}) = (-1)^{2m}\operatorname{sgn}(m_{\widetilde{P}}) = \operatorname{sgn}(m_{\widetilde{P}}),$$

where we used (20). Similarly, we have that $\operatorname{sgn}(m_Q) = \operatorname{sgn}(m_{\widetilde{Q}})$ so that (21) implies that

$$\operatorname{sgn}(m_P) = \operatorname{sgn}(m_Q). \tag{24}$$

Furthermore, (20) implies

$$0 > p_n > z_n > p_{n-1} > z_{n-1} > \cdots > p_1 > z_1 > -\infty. \tag{25}$$

Therefore, $P_m(\cdot)$ and $Q_m(\cdot)$ have no common roots and do not have multiple roots. In particular, the function $R_m(x) = P_m(x)/Q_m(x)$ can be decomposed in partial fractions as

$$R_m(x) = k + \sum_{i=1}^{m} \frac{c_i}{x - p_i}. \tag{26}$$

Observe that we already have by (25) that $p_i < 0$ for $i = 1, \ldots, m$. It remains to show that $k > 0$ and $c_i > 0$ for $i = 1, \ldots, m$. Let us start by handling k. Note that by (24)

$$k = \lim_{x\to\infty} R_m(x) = \lim_{x\to\infty} \frac{P_m(x)}{Q_m(x)} = \lim_{x\to\infty} \frac{m_P\left(1 - \frac{z_1}{x}\right)\cdots\left(1 - \frac{z_m}{x}\right)}{m_Q\left(1 - \frac{p_1}{x}\right)\cdots\left(1 - \frac{p_m}{x}\right)} = \frac{m_P}{m_Q} > 0.$$

To conclude the proof it remains to show that $c_i > 0$ for $i = 1, \ldots, m$. To this end, observe that $p_1, \ldots, p_m$ are simple poles of $R_m(x)$ and that c_i is the residual of R_m associated to p_i. Indeed, by (26), we have that

$$\lim_{x\to p_i}(x - p_i)R_m(x) = \lim_{x\to p_i}\left[c_i + k(x - p_i) + (x - p_i)\sum_{j\neq i}\frac{c_j}{x - p_j}\right] = c_i.$$

Furthermore, recall that $Q_m(p_i) = 0$ for $i = 1, \dots, m$, so that

$$c_i = \lim_{x \to p_i} (x - p_i) \frac{P_m(x)}{Q_m(x)} = P_m(p_i) \lim_{x \to p_i} \frac{x - p_i}{Q_m(x)} = P_m(p_i) \lim_{x \to p_i} \frac{x - p_i}{Q_m(x) - Q_m(p_i)}$$
$$= P_m(p_i) \lim_{x \to p_i} \frac{1}{\frac{Q_m(x) - Q_m(p_i)}{x - p_i}} = \frac{P_m(p_i)}{Q_m'(p_i)}.$$

So $c_i = P_m(p_i)/Q_m'(p_i) > 0$ for $i = 1, \dots, m$. Let us first study the sign of $P_m(p_i)$. By (25), we have $m - i$ values in $\{z_i\}_{i=1}^m$ that are larger than p_i, so that

$$\operatorname{sgn}(P_m(p_i)) = \operatorname{sgn}\Big(m_P \prod_{j=1}^m (p_i - z_j) \Big) = \operatorname{sgn}(m_P)(-1)^{m-i}.$$

Let us now study the sign of $Q_m'(p_i)$. We have that

$$Q_m'(x) = m_Q \sum_{j=1}^m \prod_{k \neq j} (x - p_k) \quad \text{so that} \quad Q_m'(p_i) = m_Q \prod_{j \neq i} (p_i - p_j).$$

By (25) we have $m - i$ values in $\{p_j\}_{j=1}^m$ that are strictly larger than p_i, so that

$$\operatorname{sgn}(Q_m'(p_i)) = \operatorname{sgn}\Big(m_Q \prod_{j \neq i} (p_i - p_j) \Big) = \operatorname{sgn}(m_Q)(-1)^{m-i}.$$

Therefore, by (24), we have that

$$\operatorname{sgn}(c_i) = \frac{\operatorname{sgn}(P_m(p_i))}{\operatorname{sgn}(Q_m'(p_i))} = \frac{\operatorname{sgn}(m_P)(-1)^{m-i}}{\operatorname{sgn}(m_Q)(-1)^{m-i}} = 1,$$

for $i = 1, \dots, m$. This shows that $c_i > 0$ for $i = 1, \dots, m$ and concludes the proof. ■

Proof [of Proposition 5] Note that

$$\frac{1}{(\kappa^2 p)^k (\kappa^2(1-p) + w^2)} - \sum_{j=1}^k \frac{1}{(\kappa^2 p)^{k+1-j} (\kappa^2 + w^2)^j}$$
$$= \frac{1}{(\kappa^2 p)^k (\kappa^2(1-p) + w^2)} - \frac{1}{(\kappa^2 p)^{k+1}} \sum_{j=1}^k \left(\frac{\kappa^2 p}{\kappa^2 + w^2} \right)^j = \frac{1}{(\kappa^2 + w^2)^k (\kappa^2(1-p) + w^2)},$$

where we used the closed-form expression of the geometric sum to arrive at the final identity. Based on this, we can rewrite the spectral density (8) as

$$f_{m,\alpha}(w) = A\sigma^2 \kappa^{-2\alpha} \Bigg[\frac{k\kappa^{2\lfloor\alpha\rfloor}}{(\kappa^2 + w^2)^{\lfloor\alpha\rfloor}}$$
$$+ \sum_{i=1}^m c_i \kappa^{2\lfloor\alpha\rfloor + 2} \Bigg(\sum_{j=1}^{\lfloor\alpha\rfloor} \frac{1}{(\kappa^2 p)^{\lfloor\alpha\rfloor} (\kappa^2(1 - p_i) + w^2)} - \frac{1}{(\kappa^2 p)^{\lfloor\alpha\rfloor + 1 - j} (\kappa^2 + w^2)^j} \Bigg) \Bigg].$$

Finally, the explicit expression for the covariance function (9) thus directly follows by taking the inverse Fourier transform of spectral density $f_{m,\alpha}$ and using the linearity of the inverse Fourier transform. ■

Proof [of Proposition 6] We begin by recalling that if $v(\cdot)$ is a stationary Gaussian process on $\mathbb{R}$ with spectral density h, then by (Pitt, 1971, Theorem 10.1), the process v will be a Markov process of order p, $p \in \mathbb{N}$, on $\mathbb{R}$, if, and only if, the spectral density of v, namely h, is a reciprocal of a polynomial of degree $2p$.

The spectral density of the process u_0 is $f_{m,0,\alpha}$ and the spectral density of u_i, where $i = 1, \dots, m$, is $f_{m,i,\alpha}$, where $f_{m,0,\alpha}$ and $f_{m,i,\alpha}$ were defined in (8). Moreover, observe that $f_{m,0,\alpha}$ is a reciprocal of a polynomial with degree $2\lfloor\alpha\rfloor$, and $f_{m,i,\alpha}$, $i = 1, \dots, m$, are reciprocals of polynomials of degree $2(\lfloor\alpha\rfloor+1)$, thus u_0 is a markov process of order $\lfloor\alpha\rfloor$ and $u_j, j = 1, \dots, m$ are Markov processes of order $\lfloor\alpha\rfloor + 1$. Now, it is well-known, see Whittle (1963) and (Lindgren, 2012, Theorem 4.7 and 4.8), that u_0 also solves

$$\begin{cases} k(\kappa + \partial_x)(\kappa^2 - \Delta)^{\frac{\lfloor\alpha\rfloor - 1}{2}} u_0 = \mathcal{W} & \text{when } \lfloor\alpha\rfloor \text{ is odd,} \\ k(\kappa^2 - \Delta)^{\frac{\lfloor\alpha\rfloor}{2}} u_0 = \mathcal{W} & \text{when } \lfloor\alpha\rfloor \text{ is even,} \end{cases}$$

in the sense that the solution to the above equation has the same covariance function as u_0. Similarly, we also have that u_i solves

$$\begin{cases} c_i(\kappa\sqrt{1-p_i} + \partial_x)(\kappa + \partial_x)(\kappa^2 - \Delta)^{\frac{\lfloor\alpha\rfloor - 1}{2}} u_i = \mathcal{W}, & \text{when } \lfloor\alpha\rfloor \text{ is odd,} \\ c_i(\kappa\sqrt{1-p_i} + \partial_x)(\kappa^2 - \Delta)^{\lfloor\alpha\rfloor/2} u_i = \mathcal{W}, & \text{when } \lfloor\alpha\rfloor \text{ is even,} \end{cases}$$

on interval I. Now, since each process is stationary, we start with the stationary distribution, i.e., $\mathbf{u}_0 = [u_0(0), u_0'(0), \dots, u_0^{(\lfloor\alpha\rfloor)}(0)]^\top$ and $\mathbf{u}_i = [u_i(0), u_i'(0), \dots, u_i^{(\lceil\alpha\rceil)}(0)]^\top$, $i > 0$, follow the stationary distribution, and thus, the restrictions of these processes from $\mathbb{R}$ to interval I will also be Markov. Therefore, the tridiagonal structure of the precision matrix follows from the fact that if we consider the open set $O_j = (j-1, j+1)$, $j \geq 2$, $j \in \mathbb{N}$, contained in interval I, then, in view of the Markov property we have that for each $i = 0, \dots, m$, $\mathbf{u}_{i,j} \perp \mathbf{u}_{i,k} | \mathbf{u}_{i,j-1}, \mathbf{u}_{i,j+1}$ for $k \in I \backslash \{j-1, j, j+1\}$, where $\mathbf{u}_{0,j} = (u_0(w_j), \dots, u_0^{(\lfloor\alpha\rfloor)}(w_j))$ and $\mathbf{u}_{i,j} = (u_i(w_j), \dots, u_i^{(\lceil\alpha\rceil)}(w_j))$, $i = 1, 2, \dots, m$, $w_j \in O_j$. Finally, by using this conditional independences and standard techniques (see, for example, the computations of Rue and Held (2005) for conditional autoregressive models), we obtain the expressions for the local precision matrices $\mathbf{Q}_{i,j}$, $i, j = 1, \dots, n$. ■

Proof [of Proposition 8.] First, recall that $N = n\left(m\lceil\alpha\rceil + \max(\lfloor\alpha\rfloor, 1)\right)$. According to Algorithm 2.9 in Rue and Held (2005), the computational cost of factorizing an $n \times n$ band matrix with bandwidth p is $n(p^2 + 3p)$, and solving a linear system via back-substitution requires $2np$ floating-point operations. Since $\mathbf{Q}$ is an $N \times N$ matrix with bandwidth $2\lfloor\alpha\rfloor + 1$, the total number of floating-point operations required for the Cholesky factorization is $n\left(m\lceil\alpha\rceil + \max(\lfloor\alpha\rfloor, 1)\right)\left((2\lfloor\alpha\rfloor+1)^2 + 3(2\lfloor\alpha\rfloor+1)\right)$. Similarly, the floating-point operations required for solving the linear system are $2n\left(m\lceil\alpha\rceil + \max(\lfloor\alpha\rfloor, 1)\right)(2\lfloor\alpha\rfloor + 1)$. ■

Proof [of Proposition 10] By reordering $\bar{\mathbf{U}}$ by location, i.e.

$$\bar{\mathbf{U}} = [\mathbf{u}_0(t_1), \mathbf{u}_1(t_1), \ldots, \mathbf{u}_m(t_1), \mathbf{u}_0(t_2), \mathbf{u}_1(t_2), \ldots]^\top,$$

we have that $\mathbf{Q}_{\bar{\mathbf{U}}|\mathbf{y}}$ is an $N \times N$ matrix, with bandwidth $\lceil\alpha\rceil(m+1)$. Following the same strategy as in the proof of Proposition 8 gives that the Cholesky factor can be computed in

$$n(m\lceil\alpha\rceil + \max(\lfloor\alpha\rfloor, 1))(\lceil\alpha\rceil^2(m+1)^2 + 3\lceil\alpha\rceil(m+1))$$

floating point operations, and $2n(m\lceil\alpha\rceil+1)\lceil\alpha\rceil(m+1)$ floating point operations are needed for solving a linear system through back-substitution. Computing the reordering of $\bar{\mathbf{U}}$ and reordering the results back can clearly be done in $\mathcal{O}(N)$ cost as the reordering is explicitly known. ■

References

Ozgur Asar, David Bolin, Peter J. Diggle, and Jonas Wallin. Linear mixed effects models for non-gaussian continuous repeated measurement data. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 69(5):1015–1065, 2020.

Fabian E. Bachl, Finn Lindgren, David L. Borchers, and Janine B. Illian. inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods Ecol. Evol.*, 10:760–766, 2019.

Moreno Bevilacqua, Tarik Faouzi, Reinhard Furrer, and Emilio Porcu. Estimation and prediction using generalized wendland covariance functions under fixed domain asymptotics. *The Annals of Statistics*, 47(2):828–856, 2019.

Vladimir Igorevich Bogachev. *Measure theory*, volume 1. Springer, 2007.

D. Bolin and A. B. Simas. *rSPDE: Rational Approximations of Fractional Stochastic Partial Differential Equations*, 2023. URL `https://CRAN.R-project.org/package=rSPDE`. R package version 2.3.3.

D. Bolin, K. Kirchner, and M. Kovács. Numerical solution of fractional elliptic stochastic PDEs with spatial white noise. *IMA J. Numer. Anal.*, 40(2):1051–1073, 2020.

D. Bolin, A. B. Simas, and J. Wallin. Statistical inference for Gaussian Whittle–Matérn fields on metric graphs. *arXiv preprint arXiv:2304.10372*, 2023.

D. Bolin, A. B. Simas, and J. Wallin. Gaussian Whittle–Matérn fields on metric graphs. *Bernoulli*, 30(2):1611–1639, 2024a.

D. Bolin, A. B. Simas, and Z. Xiong. Covariance–based rational approximations of fractional SPDEs for computationally efficient Bayesian inference. *J. Comp. Graph. Stat.*, 33(1):64–74, 2024b.

David Bolin and Finn Lindgren. Calculating probabilistic excursion sets and related quantities using excursions. *J. Stat. Softw.*, 86(5):1–20, 2018.

David Bolin and Jonas Wallin. Spatially adaptive covariance tapering. *Spatial Statistics*, 18:163–178, 2016.

H. Chen, L. Ding, and R. Tuo. Kernel packet: An exact and scalable algorithm for Gaussian process regression with Matérn correlations. *J. Mach. Learn. Res.*, 23(127):1–32, 2022.

N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 70(1):209–226, 2008.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.*, 111 (514):800–812, 2016.

R. Furrer, M. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *J. Comp. Graph. Stat.*, 15(3):502–523, 2006.

R. B. Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

R. B. Gramacy and D. W. Apley. Local Gaussian process approximation for large computer experiments. *J. Comp. Graph. Stat.*, 24(2):561–578, 2015.

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

D. Higdon. Space and space-time modeling using process convolutions. *Quantitative methods for current environmental issues*, 3754:37–56, 2002.

C. Hofreither. An algorithm for best rational approximation based on barycentric rational interpolation. *Numerical Algorithms*, 88(1):365–388, 2021.

Yiping Hong, Yan Song, Sameh Abdulah, Ying Sun, Hatem Ltaief, David E Keyes, and Marc G Genton. The third competition on spatial statistics for large datasets. *J. Agric. Biol. Environ. Stat.*, 28(4):618–635, 2023.

T. Karvonen and S. Särkkä. Approximate state-space Gaussian processes via spectral transformation. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

U. Khristenko, L. Scarabosio, P. Swierczynski, E. Ullmann, and B. Wohlmuth. Analysis of boundary effects on PDE-based sampling of whittle–matérn random fields. *SIAM J. Uncertainty Quantification*, 7(3):948–974, 2019.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 73(4):423–498, 2011.

F. Lindgren, D. Bolin, and H. Rue. The SPDE approach for Gaussian and non-Gaussian fields: 10 years and still running. *Spatial Statistics*, page 100599, 2022.

Finn Lindgren and Håvard Rue. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.*, 63(19), 2015.

G. Lindgren. *Stationary stochastic processes: theory and applications.* CRC Press, 2012.

Y. Ling. Superfast inference for stationary Gaussian processes in particle tracking microrheology. PhD Thesis. http://hdl.handle.net/10012/15338, 2019.

Y. Ling and M. Lysy. *SuperGauss: Superfast Likelihood Inference for Stationary Gaussian Time Series*, 2022. URL `https://CRAN.R-project.org/package=SuperGauss`. R package version 2.0.3.

George G Lorentz, Manfred von Golitschek, and Yuly Makovoz. *Constructive approximation: advanced problems*, volume 304. Citeseer, 1996.

B. Matérn. Spatial variation. *Meddelanden från statens skogsforskningsinstitut*, 49(5), 1960.

W. McLean. *Strongly elliptic systems and boundary integral equations.* Cambridge University Press, 2000.

D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multiresolution Gaussian process model for the analysis of large spatial datasets. *J. Comp. Graph. Stat.*, 24(2):579–599, 2015.

M. Padidar, X. Zhu, L. Huang, J. Gardner, and D. Bindel. Scaling Gaussian processes with derivative information using variational inference. *Advances in Neural Information Processing Systems*, 34:6442–6453, 2021.

L. D. Pitt. A Markov property for Gaussian processes with a multidimensional parameter. *Archive for Rational Mechanics and Analysis*, 43(5):367–391, 1971.

E. Porcu, M. Bevilacqua, R. Schaback, and C. J. Oates. The Matŕn model: A journey through statistics, numerical analysis and machine learning. *arXiv preprint arXiv:2303.02759*, 2023.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

E. Y. Remez. Sur la détermination des polynômes d'approximation de degré donnée. *Communications of the Kharkov Mathematical Society*, 10(196):41–63, 1934.

L. Roininen, S. Lasanen, M. Orispää, and S. Särkkä. Sparse approximations of fractional Matérn fields. *Scand. J. Stat.*, 45(1):194–216, 2018.

F. D. Roos, A. Gessner, and P. Hennig. High-dimensional Gaussian process inference with derivatives. In *International Conference on Machine Learning*, pages 2535–2545. PMLR, 2021.

H. Rue and L. Held. *Gaussian Markov random fields*, volume 104 of *Monographs on Statistics and Applied Probability.* Chapman & Hall/CRC, Boca Raton, FL, 2005. Theory and applications.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 71:319–392, 2009.

E Saff and H Stahl. Asymptotic distribution of poles and zeros of best rational approximants to xˆα on 0, 1. *Banach Center Publications*, 31(1):329–348, 1995.

T. J. Santner, B. J. Williams, W. I. Notz, and B. J Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003.

S. Särkkä and J. Hartikainen. Infinite-dimensional Kalman filtering approach to spatio-temporal Gaussian process regression. In *Artificial intelligence and statistics*, pages 993–1001. PMLR, 2012.

S. Särkkä, A. Solin, and J. Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

Florian Schäfer, Matthias Katzfuss, and Houman Owhadi. Sparse Cholesky factorization by Kullback–Leibler minimization. *SIAM J. Sci. Comput.*, 43(3):A2019–A2046, 2021.

E. Solak, R. M. Smith, W. E. Leithead, D. Leith, and C. Rasmussen. Derivative observations in Gaussian process models of dynamic systems. *Advances in neural information processing systems*, 15, 2002.

N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

H. R. Stahl. Best uniform rational approximation of x^α on [0, 1]. *Acta mathematica*, 190 (2):241–306, 2003.

M. L. Stein. *Interpolation of spatial data. Springer series in statistics.* Springer New York, 1999.

Michael L Stein. Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics*, 22(4):866–885, 2013.

Peter S. Swain, Keiran Stevenson, Allen Leary, Luis F. Montano-Gutierrez, Ivan B. N. Clark, Jackie Vogel, and Teuta Pilizota. Inferring time derivatives including cell growth rates using gaussian processes. *Nat. Commun.*, 7(1):13766, 2016.

F. Tronarp, T. Karvonen, and S. Särkkä. Mixture representation of the Matérn class with applications in state space approximations and Bayesian quadrature. In *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2018.

A. V. Vecchia. Estimation and model identification for continuous spatial processes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 50(2):297–312, 1988. ISSN 0035-9246.

L. Wang. *Karhunen-Loève expansions and their applications.* London School of Economics and Political Science (United Kingdom), 2008.

P. Whittle. Stochastic processes in several dimensions. *Bull. Internat. Statist. Inst.*, 40: 974–994, 1963.

A. T. Wood and G. Chan. Simulation of stationary Gaussian processes in $[0, 1]^d$. *J. Comp. Graph. Stat.*, 3(4):409–432, 1994.

A. Yang, C. Li, S. Rana, S. Gupta, and S. Venkatesh. Sparse approximation for Gaussian process with derivative observations. In *Australasian Joint Conference on Artificial Intelligence*, pages 507–518. Springer, 2018.

Jingjing Yang, Hongxiao Zhu, Taeryon Choi, and Dennis D. Cox. Smoothing and Mean–Covariance Estimation of Functional Data with a Bayesian Hierarchical Model. *Bayesian Anal.*, 11(3):649 – 670, 2016.