Flexible Regression Models with Gaussian Process Prior

Anis Fradi, Tran Tien Tam, C. Samir

Flexible Regression Models with Gaussian Process Prior

A. Fradi, T.-T. Tran, C. Samir

University of Clermont Auvergne, France

Abstract

In this paper, we introduce a set of novel data-driven regression models with low complexities. We address the challenges of infeering and learning from a substantial number of observations ($N >> 1$) with Gaussian process prior. We propose a flexible construction of well adapted covariances originally derived from specific differential operators.

Keywords: Gaussian process regression, Computational complexity, Orthogonal polynomials, Differential operators.

1. Introduction

Gaussian processes are powerful and flexible statistical models that have gained significant popularity in the field of econometrics, shape analysis, signal processing, data science, machine learning, etc [1, 2, 3, 4, 5]. They provide a non-parametric approach for modeling complex relationships and uncertainty estimation in data [6]. The core idea of Gaussian processes is the assumption that any finite set of data points can be jointly modeled as a multivariate Gaussian distribution [7]. Rather than explicit formulations Gaussian processes allow for the incorporation of prior knowledge and inference of a nonparametric function f that generates the Gaussian process for given a training dataset $(t_i, y_i)_{i=1}^N$ where $y_i = f(t_i) + \tau_i$; with $t_i \in \Omega \subseteq \mathbb{R}^d$ and noisy measurements $y_i \in \mathbb{R}$. If f is modeled with a Gaussian process prior then it can be fully characterized by a mean $m(.)$ and a covariance function $k(.,.)$ satisfying

$$m(t) = \mathbb{E}\big(f(t)\big); \quad t \in \Omega \tag{1}$$

$$k(t,s) = \mathbb{E}\big((f(t) - m(t))(f(s) - m(s))\big); \quad t, s \in \Omega \tag{2}$$

The mean function is usually assumed to be zero $\big(m(t) = 0\big)$ whereas the covariance $k(t, s)$ gives the dependence between two instances t and s. Gaussian processes can be applied for various tasks, including regression [8], classification [9], and time series analysis [10]. In regression, Gaussian processes can capture complex and non-linear patterns in data, while in classification, they enable probabilistic predictions and can handle imbalanced dataset effectively [11]. Additionally, Gaussian processes have been successfully employed in optimization, experimental design, and reinforcement learning, among other areas [12].

Organization. The paper is organized as follows. Section 2 provides background information on Gaussian processes regression. In Section 3, we discuss the low complexity Gaussian processes and highlight their main advantages in terms of computational complexity. Section 4 presents the proposed solutions for several differential operators with orthogonal polynomial bases.

2. Canonical Gaussian processes regression

A Gaussian process (GP) particularly defined on an univariate index set $\Omega \subseteq \mathbb{R}$ is a stochastic process in which the marginal variables for any finite set in Ω follows a Gaussian distribution. In a regression task, a nonparametric function f is assumed to be a realization of a stochastic GP prior whereas the likelihood term holds from observations corrupted by a noise term according to the canonical form

$$\begin{cases} y_i = f(t_i) + \tau_i; \quad i = 1, \ldots, N \\ f \sim \mathcal{GP}(0, k(t, s)) \end{cases} \tag{3}$$

where $\tau_i \sim \mathcal{N}(0, \sigma_n^2)$ is a Gaussian noise. Given a training dataset $\mathcal{D} = (\mathbf{t}, \mathbf{y}) = (t_i, y_i)_{i=1}^N$ the posterior distribution over $\mathbf{f} = f(\mathbf{t}) = (f(t_1), \ldots, f(t_N))^T$ is also Gaussian: $\mathbb{P}(\mathbf{f}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. From Bayes' rule, we state that the mean and the covariance posterior are expressed as

$$\boldsymbol{\mu} = \mathbf{K}(\mathbf{K} + \sigma_n^2 \mathcal{I}_N)^{-1}\mathbf{y} \tag{4}$$

$$\boldsymbol{\Sigma} = (\mathbf{K}^{-1} + \frac{1}{\sigma_n^2}\mathcal{I}_N)^{-1} \tag{5}$$

where $\mathbf{K} = [k(t_i, t_j)]_{i,j=1}^{N}$ is the prior covariance matrix and $\mathcal{I}_N$ is the $N \times N$ identity matrix. The predictive distribution at any test input $t_\star$ can be computed in closed-form as $f(t_\star)|\mathcal{D}, t_\star \sim \mathcal{N}\big(\bar{f}_\star, v(f_\star)\big)$, with

$$\bar{f}_\star = \mathbf{k}(t_\star)^T(\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}\mathbf{y} \tag{6}$$

$$v(f_\star) = k(t_\star, t_\star) - \mathbf{k}(t_\star)^T(\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}\mathbf{k}(t_\star) \tag{7}$$

where $\mathbf{k}(t_\star) = [k(t_i, t_\star)]_{i=1}^{N}$.

The covariance function $k(.,.)$ usually depends on a set of hyperparameter denoted θ_k that needs to be estimated from the training dataset. The log marginal likelihood for GP regression serves as an indicator of the degree to which the selected model accurately captures the observed patterns. The log marginal likelihood is typically used for model selection and optimization. Let $\theta = (\theta_k, \sigma_n^2)$ denote the set of all model parameters then the log marginal likelihood $\log \mathbb{P}(\mathbf{y}|\mathbf{t}, \theta)$ is given by

$$l(\theta) = -\frac{1}{2}\log|\mathbf{K} + \sigma_n^2\mathcal{I}_N| - \frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi) \tag{8}$$

Here, $|.|$ denotes the determinant. The goal is to estimate θ that maximizes the log marginal likelihood. This can be achieved using different methods, such as gradient-based algorithm [26], where the gradient vector w.r.t. θ is

$$\frac{\partial l(\theta)}{\partial \theta_k} = \frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}\frac{\partial \mathbf{K}}{\partial \theta_k}(\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}\mathbf{y} - \frac{1}{2}\mathrm{tr}((\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}\frac{\partial \mathbf{K}}{\partial \theta_k}) \tag{9}$$

$$\frac{\partial l(\theta)}{\partial \gamma^2} = \frac{1}{2}\mathbf{y}^T(\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}(\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}\mathbf{y} - \frac{1}{2}\mathrm{tr}((\mathbf{K} + \sigma_n^2\mathcal{I}_N)^{-1}) \tag{10}$$

The weakness of inferring the posterior mean or the mean prediction or even learning the hyperparameter from the log marginal likelihood is the need to invert the $N \times N$ Gram matrix $\mathbf{K} + \sigma_n^2\mathcal{I}_N$. This operation costs $\mathcal{O}(N^3)$ which limits the applicability of standard GPs when the sample size N increases significantly. Furthermore, the memory requirements for GP regression scale with a computational complexity of $\mathcal{O}(N^2)$.

A covariance function $k(t, s)$ is said to be stationary (homogeneous) if it is invariant to translation, i.e., a function of $t - s$ only. Two commonly used stationary covariance functions for GP regression are the Squared Exponential

(SE) and Matérn-ν kernels defined by

$$k(t,s) = \sigma^2 e^{-\varepsilon^2(t-s)^2}; \quad t,s \in \Omega = \mathbb{R} \tag{11}$$

$$k(t,s) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \big(\varepsilon\sqrt{2\nu}|t-s|\big)^{\nu} K_{\nu}\big(\varepsilon\sqrt{2\nu}|t-s|\big); \quad t,s \in \Omega = \mathbb{R} \tag{12}$$

respectively, where σ^2 is the variance parameter controlling the amplitude of the covariance, ε is the shape parameter and $\nu = k + 1/2$; $k \in \mathbb{N}$ is the half integer smoothness parameter controlling its differentiability. Here, Γ is the gamma function and K_ν is the modified Bessel function of the second kind. Both the SE and Matérn covariance functions have hyperparameter that needs to be estimated from the data during the model training process. A GP with a Matérn-ν covariance is $\lceil \nu \rceil - 1$ times differentiable in the mean-square sense. The SE covariance is the limit of Matérn-ν as the smoothness parameter ν approaches infinity. When choosing between the SE and Matérn covariance functions, it is often a matter of balancing the trade-off between modeling flexibility and computational complexity. The SE covariance function is simpler and more computationally efficient but may not capture complex patterns in data as well as the Matérn covariance function with an appropriate choice of smoothness parameter.

3. Low complexity Gaussian processes

One of the main advantages of a GP is that it can be represented as a series expansion involving a complete set of deterministic basis functions with corresponding random coefficients. Let the inner product in $\mathbb{L}^2(\Omega, \rho)$ be defined as

$$\big\langle \phi, \psi \big\rangle = \int_{\Omega} \phi(t)\psi(t)\rho(t)dt \tag{13}$$

where $\rho(t)$ is a positive weight function such that $\int_\Omega \rho(t)dt < \infty$. Consider a linear integral operator $\mathcal{K} : \mathbb{L}^2(\Omega, \rho) \mapsto \mathbb{L}^2(\Omega, \rho)$, expressed in terms of the inner product, as

$$\mathcal{K}\phi = \int_{\Omega} k(.,t)\phi(t)\rho(t)dt \tag{14}$$

Theorem 1 (Spectral theorem)**.** *Suppose A is a compact self-adjoint opera-*

tor on a Hilbert space $\mathcal{V}$. Then there is an orthonormal basis of $\mathcal{V}$ consisting of eigenfunctions of A with real eigenvalues.

According to our case, the operator $\mathcal{K}$ is self-adjoint with respect to the inner product defined in (13) since $\langle \mathcal{K}\phi, \psi \rangle = \langle \mathcal{K}\psi, \phi \rangle$ allowing to apply the spectral theorem for $\mathcal{V} = \mathbb{L}^2(\Omega, \rho)$. Consequently, there exists an orthonormal set of basis functions $\{\phi_j\}_{j=1}^{\infty}$ in the weighted space $\mathbb{L}^2(\Omega, \rho)$, that is,

$$\int_{\Omega} \phi_j(t)\phi_l(t)\rho(t)dt = \delta_{jl} \tag{15}$$

and a set of real eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$. If further $\mathcal{K}$ is positive and bounded then it admits absolutely summable positive eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. By the Mercer' theorem the covariance function has the series expansion

$$k(t, s) = \sum_{j=1}^{\infty} \lambda_j \phi_j(t)\phi_j(s) \tag{16}$$

The eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ and eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$ can be obtained from the integral operator and the solution is provided by the Fredholm integral equation

$$\mathcal{K}\phi_j(t) = \lambda_j \phi_j(t) \tag{17}$$

Theorem 2 (Karhunen Loève)**.** *Let f be a nonparametric function defined on Ω modeled with a GP of a covariance function $k(.,.)$. Then, for all $t \in \Omega$ the function f can be written as*

$$f(t) = \sum_{j=1}^{\infty} a_j \phi_j(t), \;\; \textit{with } a_j \overset{ind}{\sim} \mathcal{N}\big(0, \lambda_j\big) \tag{18}$$

where $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\phi_j\}_{j=1}^{\infty}$ are eigenvalues and eigenfunctions of the integral operator $\mathcal{K}$ defined in (14).

In order to avoid the inversion of the $N \times N$ Gram matrix $\mathbf{K} + \sigma_n^2 \mathcal{I}_N$ we use the approximation scheme presented above and project the GP to a truncated set of M basis functions. The truncated version of f at an arbitrary

order $M \in \mathbb{N}^*$ is given by

$$f_M(t) = \sum_{j=1}^{M} a_j \phi_j(t) \tag{19}$$

with an approximation error $e_M(t) = \sum_{j=M+1}^{\infty} a_j \phi_j(t)$. The canonical GP regression model (3) adapted to the truncated GP prior becomes

$$\begin{cases} y_i = f_M(t_i) + \tau_i; \quad i = 1, \ldots, N \\ f_M \sim \mathcal{GP}(0, k_M(t,s)) \end{cases} \tag{20}$$

with a covariance function approximated by $k_M(t,s) = \sum_{j=1}^{M} \lambda_j \phi_j(t) \phi_j(s)$.

Proposition 1.

1. *The approximation $k_M(.,.)$ converges uniformly to $k(.,.)$ when $M \to \infty$, i.e.,*

$$\lim_{M \to \infty} \Big(\sup_{t,s \in \Omega} |k(t,s) - \sum_{j=1}^{M} \lambda_j \phi_j(t) \phi_j(s)| \Big) = 0 \tag{21}$$

2. *The mean integrated squared error (MISE) of f_M tends to 0 as $M \to \infty$.*

Proof. The proof of 1) was provided in [23], while here we solely present that of 2). The MISE of f_M also known as the $\mathbb{L}^2$ risk function is given by

$$\begin{aligned} MISE &= \mathbb{E}\big(||f - f_M||^2_{\mathbb{L}^2}\big) \\ &= \mathbb{E}\big(||e_M||^2_{\mathbb{L}^2}\big) \\ &= \mathbb{E}\Big(\int_\Omega \big(\sum_{j=M+1}^{\infty} a_j \phi_j(t) \big)^2 dt \Big) \\ &= \mathbb{E}\Big(\sum_{j=M+1}^{\infty} a_j^2 \int_\Omega \phi_j(t)^2 dt \Big) \\ &= \mathbb{E}\Big(\sum_{j=M+1}^{\infty} a_j^2 \Big) \\ &= \sum_{j=M+1}^{\infty} \lambda_j \end{aligned} \tag{22}$$

which tends to 0 as $M \to \infty$ since λ_j are absolutely summable. ■

The convergence of the Mercer' series hardly depends on the eigenvalues and the differentiability of the covariance function. [27] showed that the speed of the uniform convergence varies in terms of the decay rate of eigenvalues and demonstrated that for a 2β times differentiable covariance $k(.,.)$ the truncated covariance $k_M(.,.)$ approximates $k(.,.)$ as $\mathcal{O}\big((\sum_{j=M+1}^{\infty}\lambda_j)^{\frac{\beta}{\beta+1}}\big)$. For infinitely differentiable covariances the latter is $\mathcal{O}\big((\sum_{j=M+1}^{\infty}\lambda_j)^{1-\epsilon}\big)$ for any $\epsilon > 0$. To summarize, smoother covariance functions tend to exhibit faster convergence, while less smooth or non-differentiable covariance functions may exhibit slower or no convergence.

The resulting covariance fall into the class of reduced-rank approximations based on approximating the covariance matrix $\mathbf{K}$ with a matrix $\tilde{\mathbf{K}} = [k_M(t_i,t_j)]_{i,j=1}^N = \mathbf{\Phi\Gamma\Phi}^T$, where $\mathbf{\Gamma}$ is a $M \times M$ diagonal matrix eigenvalues such that $\mathbf{\Gamma}_{jj} = \lambda_j$ and $\mathbf{\Phi}$ is a $N \times M$ matrix eigenfunctions such that $\mathbf{\Phi}_{ij} = \phi_j(t_i)$. Note that the approximate covariance matrix $\tilde{\mathbf{K}}$ is ill-conditioned if λ_1/λ_M is large or if the observation points t_i are too closed to each other [28]. This leads to much numerical errors when inverting $\tilde{\mathbf{K}}$.

Definition 3.1. *A covariance function $k(\cdot,\cdot)$ is said to be positive semi-definite on Ω if for all $N \in \mathbb{N}^*$, $t_i \in \Omega$ and $b_i \in \mathbb{R}$, $i = 1,\ldots,N$, we have*

$$\sum_{i=1}^{N}\sum_{l=1}^{N} b_i b_l k(t_i,t_l) \geq 0 \tag{23}$$

Proposition 2. *Let $M \in \mathbb{N}^*$ be the order of truncation. Let λ_j and ϕ_j be eigenvalues and eigenfunctions of the integral operator $\mathcal{K}$, for $j = 1,\ldots,M$. If $k(.,.)$ is positive semi-definite then $k_M(.,.)$ is also positive semi-definite.*

Proof. Let $N \in \mathbb{N}^*, \{t_1,\ldots,t_N\}$ and $\{b_1,\ldots,b_N\}$ be as in Definition 3.1.

From (16), we have

$$\begin{aligned}\sum_{i=1}^{N}\sum_{l=1}^{N} b_i b_l k_M(t_i,t_l) &= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{l=1}^{N} b_i b_l \lambda_j \phi_j(t_i)\phi_j(t_l) \\ &= \sum_{j=1}^{M}\lambda_j \sum_{i=1}^{N}\sum_{l=1}^{N} b_i b_l \phi_j(t_i)\phi_j(t_l) \\ &= \sum_{j=1}^{M}\lambda_j \left(\sum_{i=1}^{N} b_i \phi_j(t_i)\right)^2 \geq 0\end{aligned}$$

In the above equality, we have used the fact that if $k(.,.)$ is positive semi-definite then all eigenvalues λ_j are nonnegative. ■

Now, we show how our regression model that utilizes Gaussian processes decomposition technique is able to achieve low complexity. We write down the expressions needed for both inference and hyperparameter learning and discuss the computational requirements. Applying the matrix inversion lemma [29] we re-rewrite the predictive distribution (6–7) as

$$\bar{f}_\star \approx \phi_\star^T(\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Gamma}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{24}$$

$$v(f_\star) \approx \sigma^2\phi_\star^T(\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Gamma}^{-1})^{-1}\phi_\star \tag{25}$$

where $\phi_\star$ is an M-dimensional vector with the j-th entry being $\phi_j(t_\star)$. When the number of observations is much higher than the number of required basis functions ($N >> M$) the use of this approximation is advantageous. Thus, any prediction mean evaluation is dominated by the cost of constructing $\mathbf{\Phi}^T\mathbf{\Phi}$, which means that the method has an overall asymptotic computational complexity of $\mathcal{O}(NM^2)$.

The approximate log marginal likelihood adapted to the model (20) sat-

isfies

$$
\begin{aligned}
l(\theta) \approx & -\frac{1}{2}\log|\mathbf{\Phi\Gamma\Phi}^T + \sigma_n^2\mathcal{I}_N| - \frac{1}{2}\mathbf{y}^T(\mathbf{\Phi\Gamma\Phi}^T + \sigma_n^2\mathcal{I}_N)^{-1}\mathbf{y} - \frac{N}{2}\log(2\pi) \\
= & -\frac{1}{2}(N-M)\log\sigma_n^2 - \frac{1}{2}\log|\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Gamma}^{-1}| - \frac{1}{2}\sum_{j=1}^{M}\log\lambda_j \qquad (26) \\
& - \frac{1}{2\sigma_n^2}(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Gamma}^{-1})^{-1}\mathbf{\Phi}^T\mathbf{y}) - \frac{N}{2}\log(2\pi)
\end{aligned}
$$

After the initial cost needed for inferring the prediction mean (24) evaluating the approximate log marginal likelihood has $\mathcal{O}(M^3)$ complexity needed to inverse the $M \times M$ matrix $\mathbf{\Phi}^T\mathbf{\Phi} + \sigma_n^2\mathbf{\Gamma}^{-1}$. In practice, if the sample size N is large it is preferable to cache the result of $\mathbf{\Phi}^T\mathbf{\Phi}$ causing a memory requirement scaling as $\mathcal{O}(M^2)$.

4. Explicit solutions for low complexity Gaussian processes

In this section, we describe explicit solutions of the low complexity GP (LCGP) with covariances derived from differential operators. This paper focuses on the construction of covariance functions that incorporate orthogonal polynomials as eigenfunctions for two main reasons: i) On the one hand, polynomials can approximate a wide range of functions with various degrees of complexity. They can be adjusted to predict different data patterns and can capture both linear and nonlinear relationships [30]. ii) On the other hand, polynomial regression is a well-established technique that extends linear regression by incorporating polynomial terms. It allows for more flexible modeling and can capture complex relationships between all predictors and the response variable.

The connection between a differential operator denoted $\mathcal{L}$ and the linear integral operator $\mathcal{K}$ have been largely used, see for example [31]. We follow the same idea and we define the Green's function G of the differential operator $\mathcal{L}$ as its "right inverse", i.e.,

$$\mathcal{L}G(t-s) = \delta(t-s); \quad t, s \in \Omega \qquad (27)$$

where $\delta(.)$ denotes the Kronecker delta function. If $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\phi_j\}_{j=1}^{\infty}$ refer to the eigenvalues and eigenfunctions of the integral operator $\mathcal{K}$ and the Green's function acts as a stationary covariance function, i.e., $G(t-s) =$

$k(t,s) = k(t-s)$, we have

$$\begin{aligned} \lambda_j \mathcal{L}\phi_j(t) &= \mathcal{L}\mathcal{K}\phi_j(t) \\ &= \int_\Omega \mathcal{L}k(t-s)\phi_j(s)\rho(s)ds \\ &= \int_\Omega \delta(t-s)\phi_j(s)\rho(s)ds \\ &= \phi_j(t) \end{aligned} \tag{28}$$

Finally, we get

$$\mathcal{L}\phi_j(t) = \frac{1}{\lambda_j}\phi_j(t) \tag{29}$$

This shows that the eigenvalues of $\mathcal{K}$ correspond to reciprocal eigenvalues of $\mathcal{L}$, while the corresponding eigenfunctions still the same [32, 33]. At this stage, we compute eigenvalues and eigenfunctions of $\mathcal{L}$, from which we deduce the Mercer' decomposition of $k(t,s)$ given in (16) replacing λ_j by $\gamma_j = \frac{1}{\lambda_j}$. This task is available for a wide range of differential operators with positive and bounded corresponding integral operators.

Operator	$\mathcal{L}$	Ω	ρ	γ_j	γ_j^2	ϕ_j	$\|\|\phi_j\|\|_{\mathbb{L}^2}$	MISE
Matérn[1]	$\sigma^{-2}\left(\varepsilon - \frac{d^2}{dt^2}\right)^{\alpha}$	$[0,1]$	1	$\sigma^{-2}\left(\varepsilon + j^2\pi^2\right)^{\alpha}$	-	$\sqrt{2}\sin(j\pi t)$	1	$\sigma^2 \sum_{j=M+1}^{\infty}\left(\varepsilon + j^2\pi^2\right)^{-\alpha}$
Legendre	$-(1-t^2)\frac{d^2}{dt^2} + 2t\frac{d}{dt}$	$[-1,1]$	1	$j(j+1)$	-	$\frac{1}{2^j j!}\frac{d^j}{dt^j}(t^2-1)^j$	$\sqrt{\frac{2}{2j+1}}$	$\frac{1}{M+1}$
Laguerre	$t\frac{d^2}{dt^2} + (1-t)\frac{d}{dt}$	$[0,\infty)$	e^{-t}	$-j$	j^2	$\frac{e^t}{j!}\frac{d^j}{dt^j}\left(e^{-t}t^j\right)$	1	$\frac{\pi^2}{6} - \sum_{j=1}^{M}\frac{1}{j^2}$
Hermite	$\frac{d^2}{dt^2} - 2t\frac{d}{dt}$	$\mathbb{R}$	e^{-t^2}	$-2j$	$4j^2$	$(-1)^j e^{t^2}\frac{d^j}{dt^j}e^{-t^2}$	$\sqrt{\sqrt{\pi}2^j j!}$	$\frac{1}{4}\left(\frac{\pi^2}{6} - \sum_{j=1}^{M}\frac{1}{j^2}\right)$
Chebyshev	$(1-t^2)\frac{d^2}{dt^2} - t\frac{d}{dt}$	$[-1,1]$	$\frac{1}{\sqrt{1-t^2}}$	$-j^2$	j^4	$\cos(j\arccos t)$	$\sqrt{\frac{\pi}{2}}$	$\frac{\pi^4}{90} - \sum_{j=1}^{M}\frac{1}{j^4}$

[1] *Matérn hyperparameters: σ^2 for variance, ε for shape and α for smoothness: $\alpha = \nu + 1/2 = k + 1$; $k \in \mathbb{N}$*

In this paper and without loss of generality, we choose a list of some interesting and useful differential operators that act on $\mathbb{L}^2(\Omega,\rho)$: Matérn, Legendre, Laguerre, Hermite and Chebyshev from which we explicitly find the corresponding decompositions. Table **??** summarizes each class of $\mathcal{L}$, the

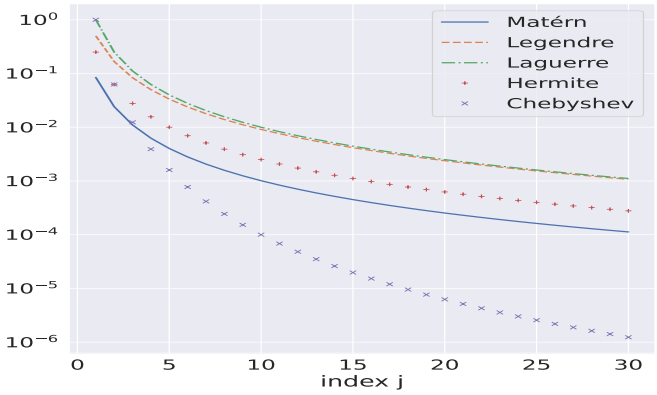

Figure 1: The eigenvalues λ_j of different operators using a base-logarithmic scale. For Matérn we consider: $\varepsilon = 2$ and $\alpha = 1$.

index set Ω, the weight function ρ, the eigenvalues γ_j, the eigenfunctions as polynomials ϕ_j and the resulting MISE. Note that for Laguerre, Hermite and Chebyshev polynomials the eigenvalues γ_j of $\mathcal{L}$ are negatives. Therefore, we consider the iterated operator $\mathcal{L}^2$ with eigenvalues γ_j^2 and unchanged eigenfunctions ϕ_j. Besides, for Legendre, Hermite and Chebyshev we remark that $||\phi_j||_{\mathbb{L}^2} \neq 1$ which means that ϕ_j should be normalized to produce an orthonormal basis allowing to build the truncated covariance $c_M(.,.)$. Figure 1 shows the behavior of the eigenvalues $\lambda_j = \frac{1}{\gamma_j}$ of $\mathcal{K}$ when varying the index j between 1 to 30. It can be observed that, for all truncated covariances, an order as low as $M = 30$ yields a very good convergence, while for Chebyshev a smaller M is enough. This is due the smoothness of the true covariance when M tends to infinity. Along the rest of the paper, we will use $M = 30$ in order to perform inference with LCGP.

5. Conclusion

In this paper, we have introduced a novel regression model with a Gaussian process prior. This nonparametric model is designed for inferring, predicting, and learning. The proposed methods are derived from specific differential operators. We study and test different configurations with will adapted eigenfunctions' bases, enabling straightforward implementations with closed-form expressions.

References

References

[1] A. Fradi, Y. Feunteun, C. Samir, M. Baklouti, F. Bachoc, J.-M. Loubes, Bayesian regression and classification using Gaussian process priors indexed by probability density functions, Information Sciences 548 (2021) 56–68.

[2] A. Fradi, C. Samir, Bayesian cluster analysis for registration and clustering homogeneous subgroups in multidimensional functional data, Communications in Statistics - Theory and Methods 49 (2020) 1–17.

[3] K. R. Ulrich, D. E. Carlson, K. Dzirasa, L. Carin, GP kernels for cross-spectrum analysis, in: Advances in Neural Information Processing Systems, MIT Press, Montreal, Canada, 2015, p. 1999–2007.

[4] P. A. Alvarado, M. A. Alvarez, D. Stowell, Sparse Gaussian process audio source separation using spectrum priors in the time-domain, in: 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Brighton, UK, 2019, pp. 995–999.

[5] C. E. Rasmussen, C. K. I. Williams, Gaussian processes for machine learning, The MIT Press, Cambridge, London, 2006.

[6] R. P. Adams, I. Murray, D. J. C. MacKay, The Gaussian process density sampler, in: Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS), Curran Associates, Inc., Vancouver, British Columbia, Canada, 2008, pp. 9–16.

[7] T. Muschinski, G. J. Mayr, T. Simon, N. Umlauf, A. Zeileis, Cholesky-based multivariate Gaussian regression, Econometrics and Statistics (2022) 2452–3062.

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

[9] C. K. I. Williams, D. Barber, Bayesian classification with Gaussian processes, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1342–1351.

[10] G. Corani, A. Benavoli, M. Zaffalon, Time series forecasting with Gaussian processes needs priors, in: ECML/PKDD (4), Lecture Notes in Computer Science, Springer, Bilbao , Spain, 2021, pp. 103–117.

[11] C. K. I. Williams, M. Seeger, The effect of the input density distribution on kernel-based classifiers, in: Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, p. 1159–1166.

[12] P.-Y. Chen, R.-B. Chen, Y.-S. Chen, W. K. Wong, Numerical methods for finding a-optimal designs analytically, Econometrics and Statistics (2022) 2452–3062.

[13] A. Melkumyan, F. Ramos, A sparse covariance function for exact Gaussian process inference in large datasets, in: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann Publishers Inc., Pasadena, California, USA, 2009, p. 1936–1942.

[14] E. Snelson, Z. Ghahramani, Sparse Gaussian processes using pseudo-inputs, in: Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, USA, 2005, p. 1257–1264.

[15] A. Damianou, M. Titsias, N. Lawrence, Variational Gaussian process dynamical systems, in: Advances in Neural Information Processing Systems, Curran Associates, Inc., Granada, Spain, 2011, pp. 2510–2518.

[16] P. Jorgensen, F. Tian, Decomposition of Gaussian processes, and factorization of positive definite kernels, Opuscula Math 39 (2019) 497–541.

[17] P. Deheuvels, G. Martynov, A Karhunen-Loève decomposition of a Gaussian process generated by independent pairs of exponential random variables, Journal of Functional Analysis 255 (2008) 2363–2394.

[18] C. K. I. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, in: Advances in Neural Information Processing Systems, Vol. 13, MIT Press, 2000, pp. 585–591.

[19] J. Fritz, W. Nowak, I. Neuweiler, Application of FFT-based algorithms for large-scale universal kriging problems, Mathematical Geosciences 51 (2009) 199–221.

[20] A. Solin, S. Särkkä, Explicit link between periodic covariance functions and state space models, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research (PMLR), Reykjavik, Iceland, 2014, pp. 904–912.

[21] R. E. Kalman, A new approach to linear filtering and prediction problems, Journal of Basic Engineering 82 (1960) 35.

[22] A. Solin, S. Särkkä, Hilbert space methods for reduced-rank Gaussian process regression, Statistics and Computing 30 (2020) 419–446.

[23] P. Greengard, M. O'Neil, Efficient reduced-rank methods for Gaussian processes with eigenfunction expansions (2022). `arXiv:2108.05924`.

[24] R. G. Ghanem, P. D. Spanos, Stochastic finite elements: a spectral approach, Springer-Verlag, Berlin, Heidelberg, 1991.

[25] N. I. Akhiezer, I. M. Glazman, Theory of linear operators in Hilbert space, Dover Books on Mathematics, Dover Publications, New York, USA, 2013.

[26] J. Barzilai, J. M. Borwein, Two-point step size gradient methods, IMA Journal of Numerical Analysis 8 (1988) 141–148.

[27] R. Takhanov, On the speed of uniform convergence in Mercer's theorem, Journal of Mathematical Analysis and Applications 518 (2023) 126718.

[28] R. Cavoretto, G. Fasshauer, M. McCourt, An introduction to the Hilbert-Schmidt SVD using iterated Brownian bridge kernels, Numerical Algorithms 68 (2014) 393–422.

[29] G. H. Golub, C. F. Van Loan, Matrix computations, 3rd Edition, The Johns Hopkins University Press, Baltimore, MD, 1996.

[30] T. S. Chihara, An introduction to orthogonal polynomials, Ellis Horwood series in mathematics and its applications, Gordon and Breach, New York, USA, 1978.

[31] G. E. Fasshauer, Green's functions: taking another look at kernel approximation, radial basis functions, and splines, in: Approximation Theory XIII, Springer, New York, 2012, pp. 37–63.

[32] E. Aristidi, Representation of signals as series of orthogonal functions, in: EAS Publications Series, Mathematical Tools for Instrumentation & Signal Processing in Astronomy, Nice, France, 2016, pp. 99–126.

[33] D. J. Griffiths, D. F. Schroeter, Introduction to quantum mechanics, 3rd Edition, Cambridge University Press, Cambridge, 2018.

[34] Z. Stednick, Machine-Learning-with-R-datasets, `https://github.com/stedy/Machine-Learning-with-R-datasets` (2017).