STAGES M2

Description Scientifique

ACRONYME et titre du projet :
Gaussian **P**rocesses with **H**igh-**D**imensional **I**nputs **(GPHDI)**

Noms et coordonnées des porteurs :
Prof. Julien Bect, julien.bect@centralesupelec.fr
Prof. Emmanuel Vazquez, emmanuel.vazquez@centralesupelec.fr
Dr. Xujia Zhu, xujia.zhu@centralesupelec.fr

Laboratoires ou équipes :
Laboratoire des Signaux et Systèmes (L2S, UMR 8506)

Durée et dates envisagées du projet :
6 mois, avril – septembre 2025

1. Context and objectives

Driven by advances in computational infrastructure and the increasing demand for modern engineering systems, computational models have rapidly progressed in recent years. These high-fidelity simulators empower engineers to perform experiments *in silico* and to conduct studies that answer questions such as determining the optimum of a quantity of interest. Another key question is how uncertainties in the input parameters propagate through the model to affect the output quantities of interest. Uncertainty quantification (UQ) is a critical area of research aimed at measuring and possibly controlling these uncertainties, yielding precise uncertainty characterization and ensuring more reliable and robust designs.

Conducting UQ analysis is typically computationally intensive, often entailing costly Monte Carlo simulations. Consequently, this is intractable for expensive simulators. To alleviate the computational burden, surrogate models are usually constructed and evaluated as a proxy of the original model. Over the past few decades, advancements in artificial intelligence, especially in statistical learning, have accelerated the development of surrogate models that are both sample-efficient and accurate.

Among these models, Gaussian processes (GPs) stand out as one of the most widely employed surrogates across various types of UQ analyses. GPs are a Bayesian, nonparametric approach that can flexibly approximate the simulator and provide uncertainty estimates for their predictions (see Figure 1). However, due to their reliance on kernels, GPs are susceptible to the so-called curse of dimensionality. More precisely, results from scattered data approximation theory [1, 2] provide theoretical bounds on the convergence of GPs, showing that the approximation error scales as $\epsilon(n,d) \sim C \cdot n^{-\alpha/d}$ where n is the number of data points, d is the dimension, and α reflects the regularity of the underlying function. This demonstrates that as the dimensionality increases, the convergence rate deteriorates exponentially. This limitation has significantly constrained the direct application of GPs to high-dimensional problems.

Driven by the growing demand for efficient surrogate models, this project aims to contribute to the development of GPs capable of handling high-dimensional inputs with quantifiable prediction errors.

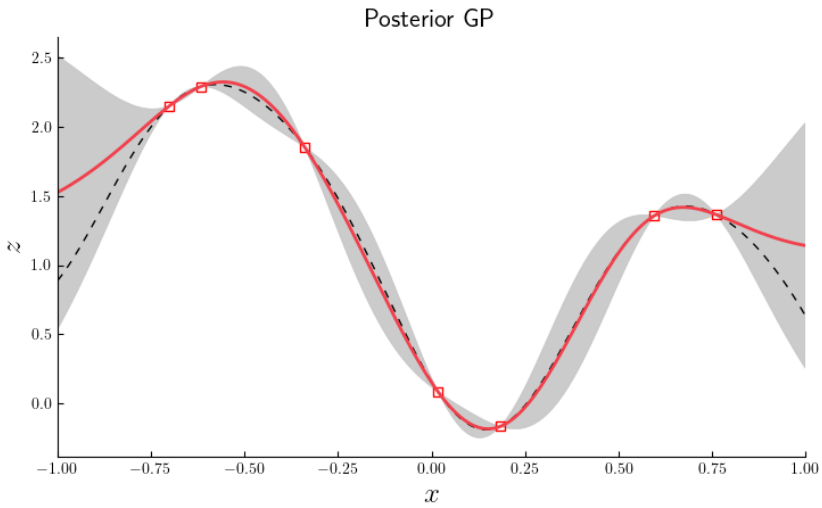

Figure 1: Illustration of a GP with one-dimensional input. The red squares are observations (e.g., results of a computer simulation). The red line corresponds to the posterior mean, and the gray areas represent uncertainty derived from the posterior variance of the GP. In realistic applications, the input dimension is typically much higher. In some cases, it could be on the order of 10,000, which poses significant challenges in practice.

2. Methodology

The problem of Gaussian processes in high-dimensional settings has seen significant advances [3, 4]. However, there is still room for improvement, especially by adopting a Bayesian approach to tackle these challenges more effectively. In high-dimensional problems, it is often observed that the quantities of interest depend on only a small subset or specific combinations of the input variables, which motivates the use of dimensionality reduction techniques. These techniques aim to identify and focus on the "active" variables—those that have a meaningful influence on the model's output—while disregarding inactive ones.

In this context, an anisotropic Gaussian process model can associate inactive variables with very large length-scale parameters, indicating that their contribution to the model is negligible. Estimating these length-scales via maximum likelihood helps detect which variables are active. However, as the dimension of the problem increases, there is often insufficient information in the data to reliably estimate the parameters of highly complex models, leading to the usual "p > n" problem in statistics, where the number of parameters exceeds the number of observations. This typically results in a flat likelihood surface, making parameter estimation difficult.

By adopting a Bayesian framework, we propose to introduce informative priors on the number of active variables. These priors act as regularizers, guiding the model toward more robust parameter estimation in high-dimensional settings. One of the key objectives of this internship will be to explore the effect of such priors on the convergence properties of Gaussian processes in high dimensions.

In addition, we will investigate promising approaches, such as the matching pursuit algorithm for sparse Gaussian process regression and embeddings [5], which can further enhance model performance in high-dimensional spaces.

3. Significance

The scientific significance of this project lies in its potential to advance the field of high-dimensional GPs, which play a crucial role in UQ and complex data modeling. The methodological development will be implemented within **GPmp** [6], which is an open-source Python package. Therefore, the scientific outcomes of this project can be readily disseminated to a larger community, benefiting high-dimensional practical applications.

Reference

[1] Wendland, H. (2004). *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.
[2] Vazquez, E. and J. Bect (2011). Sequential search based on Kriging: convergence analysis of some algorithms. In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland.
[3] Binois, M. and N. Wycoff (2022). A survey on high-dimensional Gaussian process modeling with application to Bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization 2* (2).
[4] Hvarfner, C., E. Hellsten, and L. Nardi (2024). Vanilla Bayesian optimization performs great in high dimensions. In *Proceedings of the 41st International Conference on International Conference on Machine Learning*.
[5] Garnett, R., M. A. Osborne, and P. Hennig (2014). Active learning of linear embeddings for Gaussian processes. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, UAI'14, Arlington, Virginia, USA, pp. 230–239.
[6] Vazquez, E., GPmp: the Gaussian process micro package (2024), https://github.com/gpmp-dev/gpmp