Universidade Federal do Rio de Janeiro

# *Gaussian Processes and Multi-fidelity*

Ivani Ivanova Ivanova

Rio de Janeiro

Outubro de 2019

**Universidade Federal do Rio de Janeiro**

# *Gaussian Processes and Multi-fidelity*

Ivani Ivanova Ivanova

Dissertação de Mestrado apresentada ao Programa de Pós-graduacao em Matemática Aplicada, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática Aplicada.

Orientador: Prof. Fábio Antônio Tavares Ramos.

**Rio de Janeiro**
**Outubro de 2019**

**Universidade Federal do Rio de Janeiro**

*Gaussian Processes and Multi-fidelity*

Ivani Ivanova Ivanova

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática Aplicada, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática Aplicada.

Aprovada por:

_____

Presidente, Prof. Fábio Antônio Tavares Ramos

_____

Prof. Bernardo Freitas Paulo da Costa

_____

Prof. Carlos Tomei

_____

Prof. Hugo Tremonte de Carvalho

_____

Prof. Luca Roberto Augusto Moriconi

**Rio de Janeiro**
**Outubro de 2019**

Ivanova, Ivani Ivanova.
..... Gaussian Processes and Multi-fidelity/
Ivani Ivanova Ivanova. - Rio de Janeiro:
UFRJ/ IM, 2019.
v, 93f.: il.; 29,7 cm.
Orientador: Fabio Antônio Tavares Ramos.
Dissertação (mestrado) — Universidade Federal do
Rio de Janeiro, Instituto de Matemática, Programa de
Pós-Graduação em Matemática, 2019.
Referências Bibliográficas: f. 90-93.
1. Introduction. 2. Gaussian Process Regression.
3. Multi-Fidelity. 4. Conclusion and Perspectives. 5. Appendix.
I. Ramos, Fabio Antônio
Tavares. II. Universidade Federal do Rio de
Janeiro,Programa de Pós-Graduação em Matemática. III.
Gaussian Processes and Multi-Fidelity.

## Acknowledgments

*The following acknowledgments are written in portuguese*

Antes de tudo, agradeço aos meus pais, Kolka e Ivan, pelo amor infinito e apoio ao longo de toda essa trajetória. Obrigada também por me criarem sempre com imenso carinho e por nunca desistirem de mim.

Agradeço à Professora Alexandra que com suas aulas de Inferência I despertou em mim todo esse interesse pela estatística que não parece estar se extinguindo. Por isso também, agradeço ao Claudio, que me incentivou muito a assistir à matéria. Foi grande o seu apoio e carinho durante todo o período da minha jornada acadêmica, desde o dia em que começamos a falar sobre números cardinais.

Agradeço muito aos meus amigos, principalmente, Daphne, Bangu, Otávio, Aloizio, Turano, Leandro e Cecília, que estiveram do meu lado durante todo esse trabalho. Obrigada por me ouvirem em momentos difíceis e por me fazeram ver o lado bom das coisas.

Agradeço aos professores Luiz Wagner, Cesar e Bernardo por me inspirarem imensamente a aprender coisas novas. Seus conhecimentos e aulas foram motivadores para meu aprendizado em inúmeras áreas diferentes.

Não poderia esquecer dos professores, colegas e amigos Heudson, Hugo, Reinaldo, Iago, Pedro, Danilo e Douglas por terem aguentado todos os seminários de machine learning comigo, muitos dos quais eu é que os estava cansando, e por terem participado de discussões importantes para a minha compreensão de vários temas nesta e em áreas afins.

Também agradeço ao Roberto, que, além de discutir comigo várias coisas relacionadas a essa dissertação, revisou grande parte do texto com extremo cuidado.

Por fim, mas certamente não menos importante, agradeço ao Fábio. Não somente por ter sido um ótimo orientador e por ter me apresentado a modelos de multifidelidade, assunto fascinante que veio a se tornar tema desta dissertação, mas também pela amizade e por todo o apoio durante a escrita deste texto.

Ivani Ivanova Ivanova
Outubro de 2019

# Resumo

*Gaussian Processes and Multi-Fidelity*

Ivani Ivanova Ivanova

Resumo da dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

***Resumo:*** Vários fenômenos de interesse, originados na física, engenharia, biologia, meteorologia, finanças e muitas outras áreas importantes, podem ser explicados por uma grande quantidade de modelos matemáticos. Tais modelos podem ter diferentes ordens de acurácia quando usados para descrever o fenômeno analisado, assim como as muitas simulações que podem ser realizadas baseadas em cada um deles. Não é incomum que simulações computacionais precisas sejam bastante custosas, tornando, desse modo, inviável obter observações de resposta suficientes. Em tais casos, se modelos mais baratos do mesmo fenômeno estiverem disponíveis, suas respostas, assim como as observações de alta fidelidade, podem ser usadas para uma melhor predição e estimativa do fenômeno estudado. Isto é precisamente o objetivo de modelos de multi-fidelidade: integrar informação de alta e baixa fidelidade.
Nesta dissertação, estudamos uma abordagem Bayesiana para o design de multi-fidelidade, em que os outputs de cada nível de fidelidade são modelados por um processo Gaussiano e tais níveis são combinados de uma maneira auto-regressiva. Além disso, exploramos desenvolvimentos recentes nessa técnica que proporcionam custos computacionais mais baixos para predição e validação cruzada.

**Palavras–chave.** Inferência Bayesiana, Processos gaussianos, Modelagem de multi-fidelidade.

**Rio de Janeiro**
**Outubro de 2019**

## *Abstract*

*Gaussian Processes and Multi-Fidelity*

Ivani Ivanova Ivanova

*Abstract* da dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática.

***Abstract:*** Many phenomena of interest, originated in physics, engineering, biology, meteorology, finance, and many other important fields, can be explained by a large number of mathematical models. These models can have different order of accuracy when they are used to describe the targeted phenomenon, and so do the many different computer simulations one can perform based on each one of them. It is not uncommon for precise computer simulations to be very expensive, thus making obtaining enough response observations not viable. In such cases, if cheaper models of the same phenomenon are available, their responses, in addition to the high-fidelity observations, can be used for better prediction and estimation of the underlying phenomenon. This is precisely the objective of multi-fidelity models, to integrate high and low-fidelity information. In this dissertation, we study a Bayesian approach to multi-fidelity design, where the output of each level of fidelity is modeled by a Gaussian process and the fidelity levels are combined in an autoregressive manner. Furthermore, we explore recent developments of this technique that provide lower computational cost for prediction and cross-validation procedures.

**Keywords.** Bayesian inference, Gaussian processes, Multi-fidelity modeling.

**Rio de Janeiro**
**October 2019**

# List of Symbols

| | |
|---|---|
| $\prod_{i=m}^{n} a_i$ | product of the terms $a_i$, with $i = m, \ldots, n$, and the convention that $\prod_{i=m}^{n} a_i$ is equal to 1 if it defines a product of 0 terms ($m > n$) |
| $A \odot B$ | Hadamard product (entrywise) of matrices, or vectors, $A$ and $B$ |
| $\bigodot_{i=m}^{n} A_i$ | Hadamard product of matrices, or vectors, $A_i$, with $i = m, \ldots, n$ |
| $v_\zeta$ | vector containing the entries of the vector $v$ with of indices in the set $\zeta$ |
| $v_{-\zeta}$ | vector containing the entries of the vector $v$ with indices not in the set $\zeta$ |
| $[A]_{[\zeta,\gamma]}$ | submatrix of $A$ containing the rows of indices in the set $\zeta$, and columns of indices in the set $\gamma$ |
| $[A]_{[-\zeta,\gamma]}$ | submatrix of $A$ containing the rows of indices not in the set $\zeta$, and columns of indices in the set $\gamma$ |
| $[A]_{[\zeta,-\gamma]}$ | submatrix of $A$ containing the rows of indices in the set $\zeta$, and columns of indices not in the set $\gamma$ |
| $[A]_{[-\zeta,-\gamma]}$ | submatrix of $A$ containing the rows of indices not in the set $\zeta$, and columns of indices not in the set $\gamma$ |
| $\mathbf{1}_n$ | vector of ones of length $n$ |
| $\mathbf{0}_{n\times m}$ | $n$ by $m$ matrix of zeros |

# Contents

# Chapter 1

# Introduction

## 1.1 Once upon a time, in a golden land...

Danie Gerhardus Krige (1919-2013) was a South African statistician and mining engineer who studied exploitation data of several orebodies, particularly gold ores. Specifying accurately the tonnage and grade of ores in a mine is of extreme importance to carry out appropriate selective mining when using these estimates. This means that for stoping (the removal of the desired ore from an underground mine), only parcels of ore which contain a sufficient amount of gold to cover the costs of extraction are selected and parcels with insufficient amount of gold are left intact. Krige observed in [Krige '51] that the method used when trying to estimate the gold content of a block of ore was simply to take the mean of the limited available observations on its boundary and sought ways to determine the reliability and improve the accuracy of the estimate, see [Krige '51] and [Chilès & Desassis '18].

D. G. Krige's contributions to ore grade estimation inspired the French mathematician and civil engineer of mines Georges Matheron (1930-2000) in his development of a linear unbiased predictor, a technique which he named *kriging* after Krige in [Matheron '63]. Matheron was also interested in inferring the grade of a panel using a weighted average of available samples.

In the context of Matheron's work, we suppose that the spatial data $Z(x_1), \ldots, Z(x_n)$ are observations of a process $Z(x)$ with $x \in \mathcal{X} \subset \mathbb{R}^d$ with $d = 2$ or 3 at the locations $x_1, \ldots, x_n$. Furthermore, we write the expression of this process as a sum of a known mean $m(x)$ and a deviation factor $\delta(x)$:

$$Z(x) = m(x) + \delta(x).$$

The deviation $\delta(x)$ is a zero-mean stochastic process with known covariance function

$$k(x, x') = \mathrm{Cov}\{\delta(x), \delta(x')\} = \mathrm{Cov}\{Z(x), Z(x')\},\ x, x' \in \mathcal{X}.$$

This covariance function is not known a priori and must be defined or estimated using the available data and their variability.

Matheron proposed the *simple kriging* predictor in this setting when looking for an unbiased predictor at a location $x_0$ that is a linear combination of the available samples:

$$\sum_{i=1}^{n} l_i Z(x_i) + c,$$

with $l_i$ for $i = 1, \dots n$ and $c$ unknown constants. The predictor is chosen by minimizing the mean-squared prediction error

$$\mathbb{E}\left[\left(Z(x_0) - \left(\sum_{i=1}^{n} l_i Z(x_i) + k\right)\right)^2\right].$$

Thus, the minimum is obtained when

$$l^T = k^T K^{-1} \quad \text{and} \quad c = m(x_0) - l^T m(X),$$

with $l = (l_1, \dots, l_n)^T$, $k = (k(x_0, x_1), \dots, k(x_0, x_n))^T$, $K_{ij} = k(x_i, x_j)$ and $m(X) = (m(x_1), \dots, m(x_n))^T$. Matheron's best linear unbiased predictor is then

$$p_{SK}(x_0) = m(x_0) + k^T K^{-1}(Z(X) - m(X)),$$

with $Z(X) = (Z(x_1), \dots, Z(x_n))^T$, and this predictor has a mean-squared prediction error given by

$$\mathbb{E}[(Z(x_0) - p_{SK}(x_0))^2] = k(x_0, x_0) - k^T K^{-1} k.$$

Other forms of kriging on which Matheron worked are known as *ordinary kriging* and *univeral kriging*, cases when the mean function $m(x)$ is not assumed as known, and they yield other optimal predictors linear in the data, but with larger prediction errors than simple kriging, see [Cressie '93]. For these methods, predictive expressions with simple dependences on the variogram, defined as

$$2\gamma(x, x') = \mathrm{Var}[Z(x) - Z(x')] = \mathbb{E}[((Z(x) - m(x)) - (Z(x') - m(x'))^2],$$

for a stochastic process $Z(x)$ with mean function $m(x)$, at the locations $x$ and $x'$, are found, and a sample variogram could be used when determining the covariance structure [Cressie '90].

Outside of the spatial context of kriging, similar approaches were proposed before the work of Matheron. But, for an equivalent situation, using spatial data, we only know that Russian meteorologist Lev Semenovich Gandin proposed a similar approach in the field of meteorology. In his 1965 book *"Objective Analysis of Meteorological Fields"*, concerned with both theory and application, Gandin presents analyses of spatial prediction and design. There, simple kriging is called *optimal interpolation* and ordinary kriging is called *optimal interpolation with normalization of weighting factors* [Cressie '90], [Chilès & Desassis '18].

## 1.2 It's all about that Gauss(ian)

The previously presented method, kriging, is concerned with estimating one part of a random process using information about other parts. Kriging is widely used for spatial data in geostatistics; however, a predictive procedure need not be restricted to these cases. In a more general case, a predictor $p(x_0)$ for $Z(x_0)$ is one that minimized an error characterized by a loss function $\mathcal{L}(a, b)$. Then, the optimal predictor is given by minimizing the expected value of the loss when estimating $Z(x_0)$ using $p(x_0)$:

$$\mathbb{E}[\mathcal{L}(Z(x_0), p(x_0))].$$

The previous expression does not include the information of the available observations $Z(x_1), \ldots, Z(x_n)$. This is done by conditioning the loss function on the data:

$$\mathbb{E}[\mathcal{L}(Z(x_0), p(x_0))|Z(X)].$$

When the squared error loss is used for $\mathcal{L}(a, b)$, we have

$$\mathbb{E}[(Z(x_0) - p(x_0))^2|Z(x_1), \ldots, Z(x_n)].$$

By minimizing this expected error, the resulting best predictor is given by

$$p(x_0) = \mathbb{E}[Z(x_0)|Z(X)].$$

So why is this all about *Gaussian processes*???

The conditional distribution $Z(x_0)|Z(x_1), \ldots Z(x_n)$ is needed to determine the predictor $p(x_0)$, however, the estimation of this distribution may not be available unless some simplifying model assumptions are made. The most simple of these assumptions is that $Z(x)$ is a Gaussian random process (this means that the joint distribution of $Z(x)$ at a finite number of locations is a multivariate Gaussian distribution, a precise definition will be given in Section 2.1). In this case, when assuming Gaussian data, $\mathbb{E}[Z(x_0)|Z(X)]$ has a closed form expression and the predictor $p(x_0)$ is exactly the best linear unbiased predictor $p_{SK}(x_0)$ given by the simple kriging procedure. Also, both have the same predictive error [Cressie '93]. Indeed... kriging and Gaussian processes are equivalent techniques.

And why the name "Gaussian Processes"? In his work *"Theoria motus corporum coelestium in sectionibus conicis solem ambientium"*, German physicist and mathematician Karl Friedrich Gauss (1777-1855) introduced several novel mathematical objects, such as *least squares*, *maximum likelihood* and the famous *normal (Gaussian) distribution*

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Gaussian processes take their name from Gauss simply because their construction is entirely based on the *Gaussian distribution*.

Other famous researchers who contributed to the theory of Gaussian processes were Soviet mathematician Andrey Nikolaevich Kolmogorov (1903-1987), and American mathematician and philosopher Norbert Wiener (1894-1964). In his work on turbulence

developed in the 40's, Kolmogorov also assumed the existence of the variogram, but considered an equally spaced timeseries with known mean. He was interested in the limits of mean-squared prediction errors for interpolation and extrapolation of data. During World War II, Wiener obtained similar equations to Kolmogorov's in a spatial setting. When he was studying the prediction of enemy aircraft movements from known radar measurements. Their approach was not well adaptable for the sparse and irregular spatial data found in the geological context [Chilès & Desassis '18], [Rasmussen & Williams '05].

More recently, the modern Gaussian process theory was developed in [O'Hagan '78], where O'Hagan presents Gaussian process prediction in a general Bayesian regression setting and uses this theory in a number of illustrative examples.

And why modern Gaussian process theory?

The need to obtain accurate predictions using known observations is a common topic throughout science. When we have noisy data of the form

$$y = f(x) + \epsilon,$$

and the function $f(x)$ must be specified, a regression procedure is usually performed. Classical regression models, such as linear regression, are not flexible enough when dealing with generic functions $f(x)$ and may lead to overfitting if a large number of basis functions are used in order to accommodate many possible latent functions.

A more flexible tool to tackle this kind of problem is precisely Gaussian processes (GP), which can be seen as a kind of nonparametric regression. This regression technique can be understood as carrying out Bayesian inference on the function space by placing a prior over the functions and incorporating the information of observations through Bayes theorem to obtain a posterior over all possible functions. The flexibility of the GP is achieved through the covariance function $k(x, x')$, which is used to translate the variability of the data throughout space and encodes other not so explicit features of the possible latent functions, such as smoothness and length-scale.

This adaptability of GP explains its expanding use in the many areas of science. Particularly in engineering, Gaussian processes are used for interpolation of data that are responses of expensive computational experiments. In this situation, a few response observations are obtained and a surrogate model such as a Gaussian process is employed to interpolate and predict the outcome of the computational experiment at unknown design points.

## 1.3 Gotta stack 'em all!

Many phenomena of interest, originating in physics, engineering, biology, meteorology, finance, and many other important fields, can be explained by a large number of mathematical models. These models can have different levels of accuracy when compared to the phenomenon, and so do the many different computer simulations one can do based on each one of them. As an example, take differential equations and versions with linearizations and/or approximations, their finite element simulation on a refined or coarser grid. Unfortunately, in many cases, obtaining a large amount of responses of more precise simulations of a complex model is computationally expensive, and only a small

quantity of outputs can be acquired in a reasonable amount of time. This makes describing the underlying modeled phenomenon difficult. Methods for interpolation, regression and prediction that work with the available simulation results of the computer codes may be used to infer the desired quantities in input regions where there is no response data, but this creates large errors in areas of no information and can lead to misleading conclusions, especially when in high-dimensional input spaces.

Then... what can be done?

If cheaper computational models are available, we can use the data they provide to aid us in estimating the desired quantities or to profile the phenomenon. In other words, we can integrate the information of a large number of computationally cheap data that captures a few important features, even if less precise, and a small number of computationally expensive and accurate data to achieve better estimates than only using the "good" data. This integration of various levels of accuracy is the idea behind multi-fidelity modeling. A great number of multi-fidelity models are available, see [Fernández-Godino et al. '16], and we will examine a particular class of models that uses Gaussian processes as a tool for prediction.

Kennedy and O'Hagan propose an autoregressive multi-fidelity model in [Kennedy & O'Hagan '98] for combining the information of data obtained from deterministic computational codes. In their work, the fidelity levels are sorted by increasing level of fidelity (accuracy) and a Gaussian process prior is used for each one of them. Even though it is a powerful model, difficulties arise when working with a large number of data points, which is exactly what is desired as it means more information. Since Gaussian process prediction requires the inversion of a matrix that has a number of lines and columns equal to the number of data points, this process quickly becomes expensive. In fact, in this original model the dimension of the matrix that needs to be inverted is the sum of the number of observations of all levels of fidelity. As we wish to work with a large number of low-fidelity data and several evaluations of the model for selection of parameters and further elements, this can easily become an intractable problem.

The more recent work of Le Gratiet [Le Gratiet & Garnier '14], [Le Gratiet '13], [MuFiCokriging] explores a simple idea that reduces the computational cost by rewriting the model of Kennedy and O'Hagan in a smart way that allows to break down the problem of inverting the big matrix into several smaller inversion problems. Both models, in fact, offer the same predictive distributions, which shows the importance of the model improvements made by Le Gratiet. Furthermore, this work provides formulas for a fast cross-validation procedure that does not require several model fittings when performing model selection.

# Chapter 2

# Gaussian Process Regression

## 2.1 Basics

We begin with the definition of the main object that will permeate the next few chapters and that is the fundamental tool in *multi-fidelity modeling via Gaussian Processes*.

**Definition 2.1.** A *Gaussian process* is a collection of random variables, any finite number of which have a joint Gaussian distribution.

A Gaussian process is completely specified by its mean function and covariance function, see [Rasmussen & Williams '05] and [Adler '09]. We define the mean function $m(x)$ and the covariance function $k(x, x')$ of a real process $Z(x)$ as

$$\begin{aligned} m(x) &= \mathbb{E}[Z(x)], \\ k(x, x') &= \mathbb{E}[(Z(x) - m(x))(Z(x') - m(x'))], \end{aligned} \tag{2.1}$$

and denote the Gaussian process $Z(x)$ as

$$Z(x) \sim \mathcal{GP}(m(x), k(x, x')). \tag{2.2}$$

In this case, the mentioned random variables represent values of the function $Z(x)$ at a location $x$, with the Gaussian process being defined, for example, over time or space. In this dissertation, we will use Gaussian processes for $x \in \mathcal{X} \subseteq \mathbb{R}^D$.

Usually, for simplicity, the mean function is taken to be zero, since it is usually unknown and would require a parametrization and subsequent estimation of a larger set of heyperparameters.

The covariance function may take many forms, as will be disscussed later, but for a first example, consider the squared exponential covariance function, given by

$$k_{SE}(x, x') = \exp\left\{ -\frac{||x - x'||^2}{2l^2} \right\},$$

where $l$ is a length-scale parameter. This basic covariance function, the most widely used, is based on the Gaussian function of Figure 2.1.

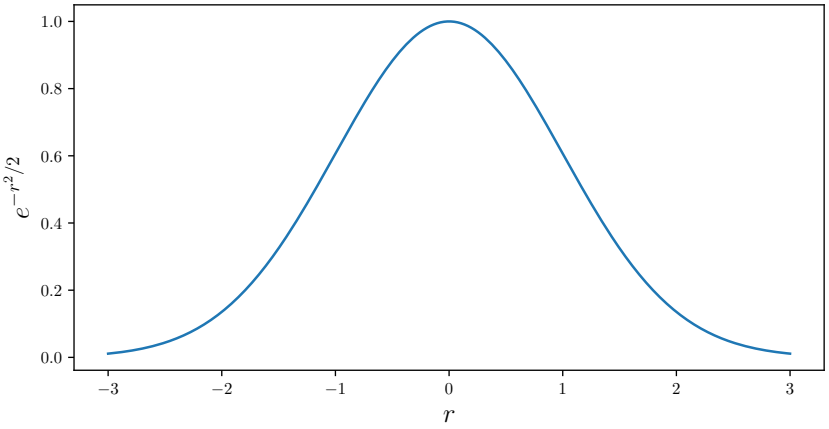



Figure 2.1: The Gaussian function.

## 2.2 Prediction

For training the model, we have a training data set $\mathcal{D}$ with $n$ observations of a latent function $z(x)$, $\mathcal{D} = \{(x_i, z_i)\}_{i=1,\dots,n}$, where $x_i \in \mathbb{R}^D$ denotes an input vector of dimension $D$ and $z_i \in \mathbb{R}$ the associated scalar output called target. The $n$ inputs are aggregated in a $n \times D$ matrix $X$, and the outputs in a $n$-dimensional vector $y$. We want to make inference about the output value for any given input. For this, first we will obtain the necessary expressions for the zero-mean case, with which it is simple to generalize for an arbitrary $m(x)$.

### 2.2.1 Noise-free observations

For the noise-free case, our observations are of the form

$$z = z(x),$$

as in Figure 2.2. We model the latent function as a Gaussian process $Z(x)$,

$$Z(x) \sim \mathcal{GP}(0, k(x, x')), \tag{2.3}$$

such that the output $z_i = z(x_i)$ stands for a realization of the the random process at the location $x_i$.

Let our observations (training data) be $\{(x_i, z_i)\}_{i=1,\dots,n}$ with $x_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$ for $i = 1, \dots, n$, and the test inputs, for which we wish to make inference, $\{x_{*i}\}_{i=1,\dots,n_*}$ with $x_{*i} \in \mathbb{R}^D$ for $i = 1, \dots, n_*$. We aggregate the training inputs and test inputs as rows of matrices $X$ and $X_*$, respectively, and aggregating in the same way, we have $Z(X) = (Z(x_1), \dots, Z(x_n))^T \in \mathbb{R}^n$ and $Z(X_*) = (Z(x_{*1}), \dots, Z(x_{*n}))^T \in \mathbb{R}^{n_*}$. Let $K(X, X_*)$ be the $n \times n_*$ covariance matrix of the process evaluated at all pairs of points in the training and test sets,

$$K(X, X_*)_{ij} = k(x_i, x_{*j}),$$

and consider the equivalent expressions for the matrices $K(X_*, X), K(X, X)$ and $K(X_*, X_*)$. Then, the joint distribution of the training and test outputs, according to the specified Gaussian process prior (2.3) is

$$\begin{bmatrix} Z(X) \\ Z(X^*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right). \tag{2.4}$$

When observing $Z(X) = y$, we can condition the joint Gaussian distribution on the observations using (A.2) to obtain

$$Z(X_*)|X, y \sim \mathcal{N}(\bar{z}_*, \mathrm{Cov}[\bar{z}_*]),$$

with

$$\bar{z}_* = K(X_*, X)K(X, X)^{-1}y, \tag{2.5}$$

and

$$\mathrm{Cov}[z_*] = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*). \tag{2.6}$$

Notice that, since $K(X, X)$ represents a covariance matrix, it is positive semi-definite and, therefore, $K(x_*, X)K(X, X)^{-1}K(X, x_*) \geq 0$ for every $x_*$, which implies that, when we condition on observed values, the predictive variance at any possible input $x_*$ decreases when compared to the prior covariance $K(x_*, x_*) = k(x_*, x_*)$. Also, on points used for training, $\mathrm{Cov}[z_*] = 0$, since we assumed that the exact value of the function is obtained in the observations. Compare the prior and posterior distribution for the data of Figure 2.2 displayed in Figures 2.3 and 2.4. The decrease in variance on any input can be observed in Figure (2.4a), where the hyperparameters are not optimized and have the same value as in the prior model. In Figure (2.4b), the parameters are optimized via maximum likelihood, this will be clarified in Section 2.7.1.

### 2.2.2 Noisy observations

In more realistic situations, we do not have access to the values of the desired function, but to noisy versions of them,

$$y = z(x) + \varepsilon,$$

where $\varepsilon$ denotes an additive independent and identically distributed Gaussian random variable with mean 0 and variance $\sigma_n^2$, see Figure 2.5.

In this case, we have

$$\mathrm{Cov}\{y_i, y_j\} = k(x_i, x_j) + \sigma_n^2\delta_{ij} \implies \mathrm{Cov}[y] = K(X, X) + \sigma_n^2 I,$$

where $\delta_{ij}$ denotes the Kronecker delta. The joint distribution of $y$ and $Z(X_*)$ then becomes

$$\begin{bmatrix} y \\ Z(X_*) \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right), \tag{2.7}$$

with the noise only affecting the diagonal of the covariance submatrix relative to the observations $X$ only, since, for $X_*$, we wish to make inference on the latent function itself

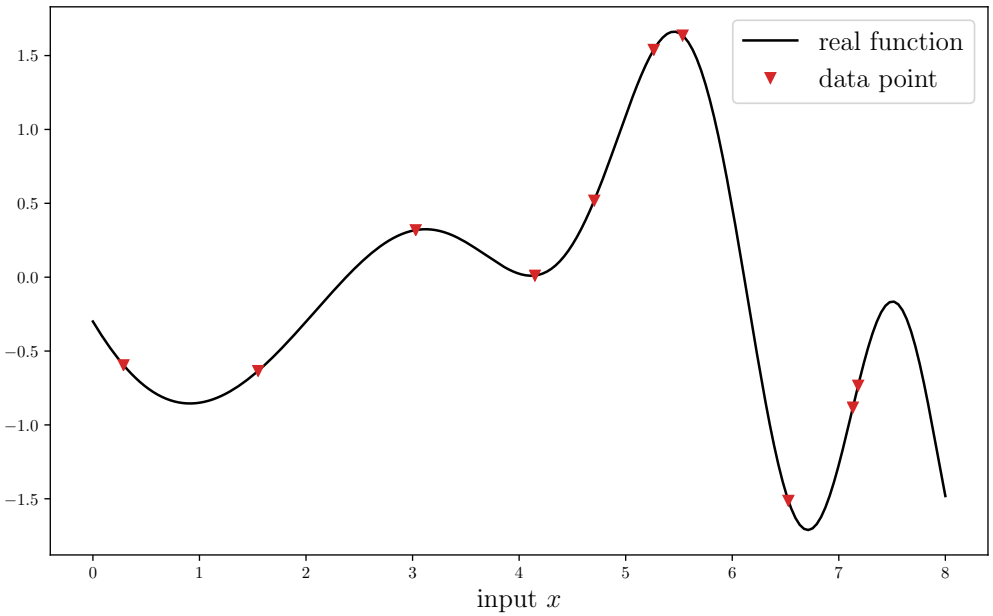



Figure 2.2: The latent function and noiseless observations.

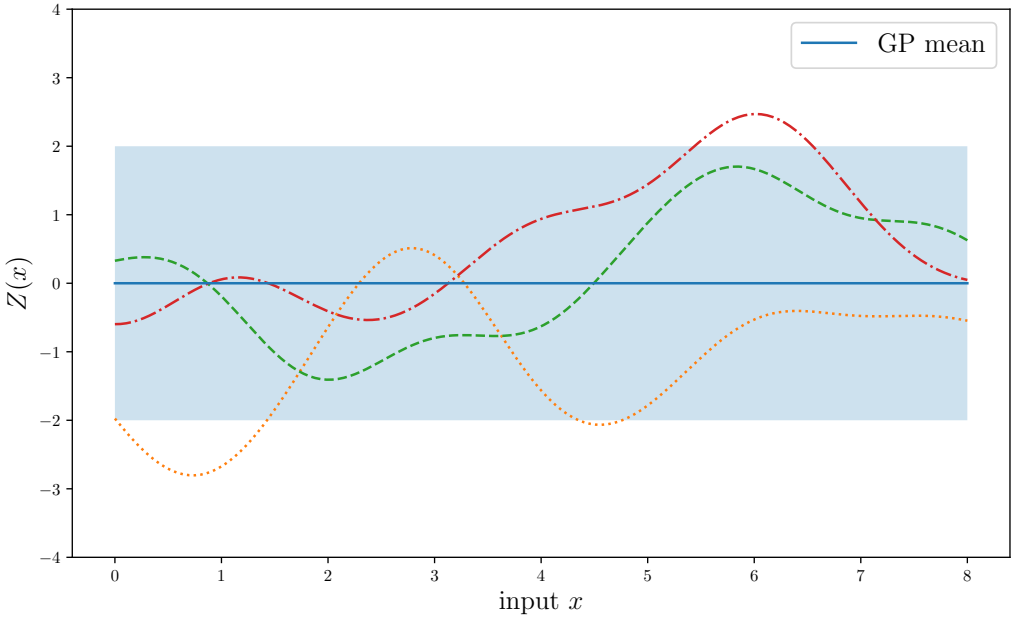
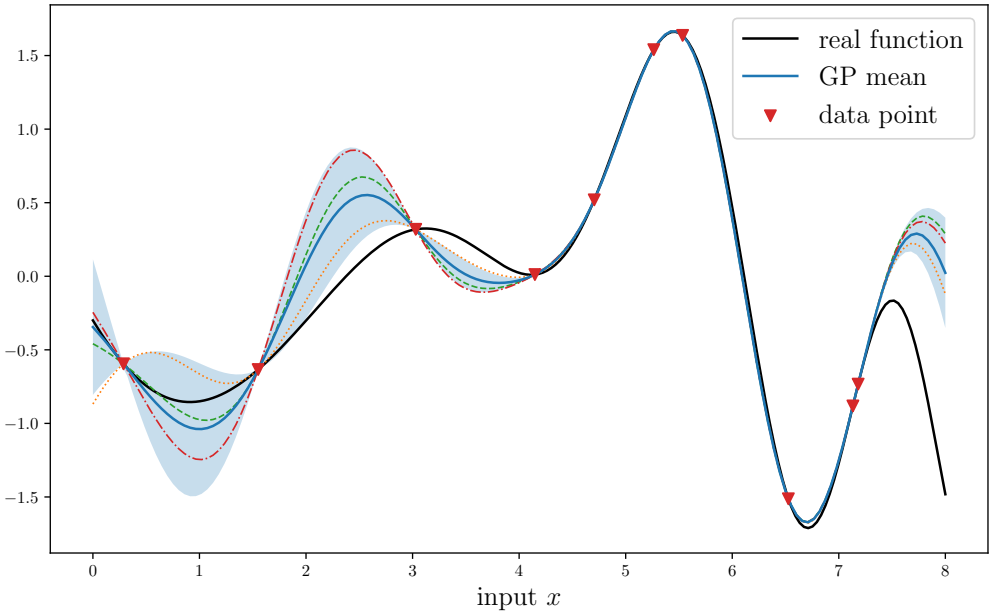
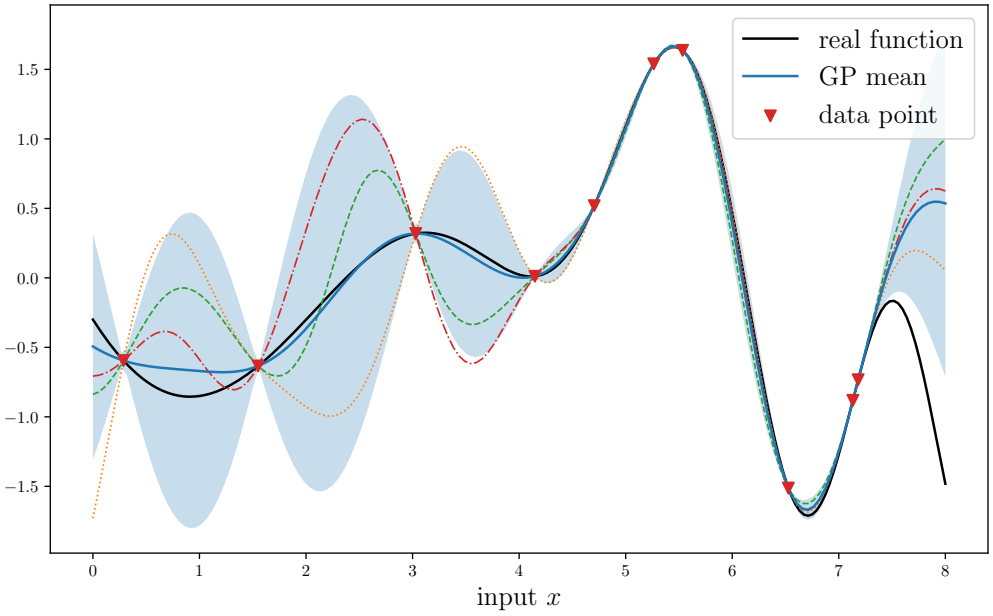



Figure 2.3: Prior samples of a GP modeled by a squared exponential kernel $k_{SE}(x, x') = \sigma^2 \exp\{-\frac{||x-x'||^2}{2l^2}\}$ with hyperparameters $\sigma^2 = 1$ and $l = 1$. 95% confidence intervals are shown.




(a) Fixed hyperparameters $\sigma^2 = 1$ and $l = 1$.




(b) Optimized hyperparameters $\sigma^2 = 0.897$ and $l = 0.626$.

Figure 2.4: Posterior samples of a GP with squared exponential kernel $k_{SE}(x, x') = \sigma^2 \exp\{-\frac{||x-x'||^2}{2l^2}\}$. 95% confidence intervals are shown.

and not on it's noisy version. This, as in the noiseless case, gives rise to the predictive distribution

$$\big(Z(X_*)|X, y\big) \sim \mathcal{N}(\bar{z}_*, \mathrm{Cov}[z_*]),$$

with

$$\bar{z}_* = K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} y,$$

and

$$\mathrm{Cov}[z_*] = K(X_*, X_*) - K(X_*, X)[K(X, X) + \sigma_n^2 I]^{-1} K(X, X_*).$$

As in the previous case, we know that $[K(X, X) + \sigma_n^2 I]$ is positive definite, thus, we also have a decrease in the variance when conditioning the Gaussian process on a set of observations. It is interesting to remark that the predictive covariance $\mathrm{Cov}[z_*]$ does not depend on the observed values, but only on the variances associated to training and test locations.

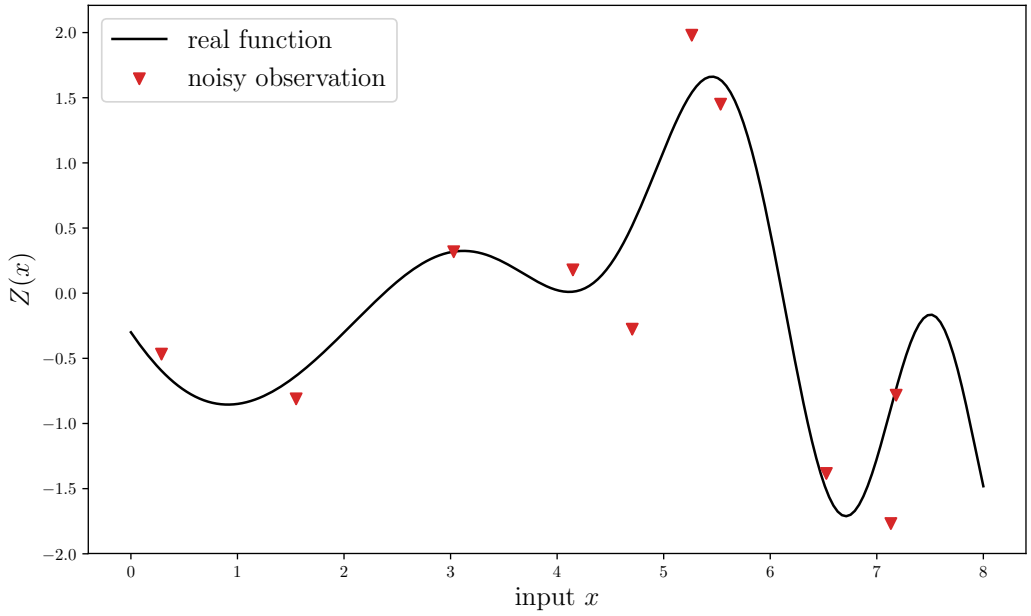



Figure 2.5: The latent function and noisy observations obtained using $\sigma_n^2 = 0.49$.

### 2.2.3 Non-zero mean

For an arbitrary mean function $m(x)$, we can obtain the predictive mean and variance of $Z(x) \sim \mathcal{GP}(m(x), k(x, x'))$ simply by noting that $Z(x) - m(x) \sim \mathcal{GP}(0, k(x, x'))$. Therefore, if $K_y$ denotes the covariance matrix at the location of the observations, being equal to $K(X, X)$ or $K(X, X) + \sigma_n^2 I$ for the noiseless and noisy case, respectively, then

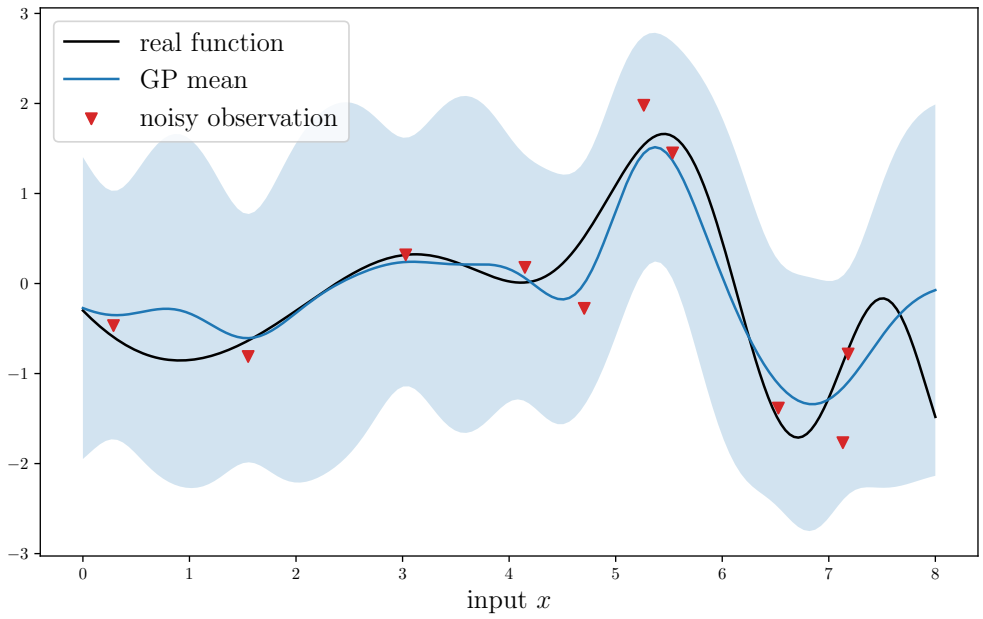



Figure 2.6: Posterior GP of the noisy observations modeled by the kernel $\sigma^2 k_{SE}(x, x') + \sigma_n^2 \delta(x, x')$ with optimized hyperparameters $\sigma^2 = 0.805$, $l = 0.416$ and $\sigma_n^2 = 0.272$.

$$Z(X_*)|X, y \sim \mathcal{N}(\bar{z}_*, \mathrm{Cov}[z_*]),$$

with

$$\bar{z}_* = m(X_*) + K(X_*, X)K_y^{-1}(y - m(X)), \tag{2.8}$$

while the predictive covariance remains equal

$$\mathrm{Cov}[z_*] = K(X_*, X_*) - K(X_*, X)K_y^{-1}K(X, X_*). \tag{2.9}$$

Here, we use $m(X) = (m(x_1), \ldots, m(x_n))^T$ and $m(X_*) = (m(x_{*1}), \ldots, m(x_{*n}))^T$

While incorporating a mean function may be useful for interpretability of the model and integration of prior knowledge, setting it to zero does not restrict the model too much, since the posterior is not confined to be zero too. Specifying the mean may be a difficult task, however. A more practical approach is to perform regression, expressing the mean of the Gaussian process as a combination of fixed basis functions. For this, we let $h(x) = (h_1(x), \ldots, h_p(x))^T$ be $p$ fixed basis functions, for example, polynomials up to order $p-1$, $(1, x, x^2, \ldots, x^{p-1})^T$, and $\beta \in \mathbb{R}^p$ a parameter vector, which must inferred from the data. The new model consists of assuming that the observations $y$ are realizations of a process $W(x)$, and

$$W(x) = Z(x) + h^T(x)\beta,$$

with $Z(x) \sim \mathcal{GP}(0, k(x, x'))$ in a noiseless or noisy case.

It is common to put an independent Gaussian prior on the parameters $\beta$, such that $\beta \sim \mathcal{N}(b, B)$. In this case, the prior distribution at any input point $x$ is given by

$$W(x) \sim \mathcal{GP}\big(h^T(x)\beta, k(x,x') + h^T(x)Bh(x')\big).$$

Hence, using equations (2.8) and (2.9), the predictive distribution $W(X_*)|X, W(X) = y$ at the test inputs $X_*$ is given by $\mathcal{N}(\bar{w}_*, \mathrm{Cov}[\bar{w}_*])$ with

$$\bar{w}_* = H_*^T b + (K_*^T + H_*^T BH)(K_y + H^T BH)^{-1}(y - H^T b),$$

and

$$\mathrm{Cov}[w_*] = K(X_*, X_*) + H_*^T BH_* - (K_*^T + H_*^T BH)^T (K_y + H^T BH)^{-1}(K_* + H^T BH_*),$$

where $H$ and $H_*$ are the matrices that collect the vectors $h(x)$ in their rows at the training and test locations, respectively. That is, the $i$ row of $H$ is equal to the vector $h(x_i) = \big(h_1(x_i), \dots h_p(x_i)\big)$, and similarly for $H_*$. Moreover, for simplicity we will denote $K_* = K(X, X_*) = K(X_*, X)^T$. After rearranging the terms [see section (A.5)], the predictive mean and covariance can be rewritten as

$$\bar{w}_* = H_*^T \bar{\beta} + K_*^T K_y^{-1}(y - H^T\bar{\beta}) = \bar{z}_* + R^T\bar{\beta}, \tag{2.10}$$

$$\mathrm{Cov}[w_*] = \mathrm{Cov}[z_*] + R^T(B^{-1} + HK_y^{-1}H^T)^{-1}R, \tag{2.11}$$

with $\bar{\beta} = (B^{-1} + HK_y^{-1}H^T)^{-1}(HK_y^{-1}y + B^{-1}b)$ and $R = H_* - HK_y^{-1}K_*$. We can, now, interpret the predictive mean as the mean linear output $H_*^T\bar{\beta}$ plus the prediction of the Gaussian process for the residuals $K_*^T K_y^{-1}(y - H^T\bar{\beta})$ and the covariance as the sum of the usual covariance and a term $R^T(B^{-1} + HK_y^{-1}H^T)^{-1}R$ with non-negative diagonal entries (uncertainty is added in the predictions when we include uncertainty on $\beta$).

### 2.2.4 Marginal likelihood

The marginal likelihood $p(y|X)$ is obtained when we integrate the latent function at the training locations $Z(X)$ from the likelihood $p(y|X, Z(X)) = p(y|Z(X))$, obtaining just the probability of the outputs given the inputs. This will be important when performing model selection in Section 2.7. Observe that

$$p(y|X) = \int p(y|X, Z(X))p(Z(X)|X)dZ(X).$$

For the noisy zero mean case, we know that $y|Z(X) \sim \mathcal{N}(Z(X), \sigma_n^2 I)$ and $Z(X)|X \sim \mathcal{N}(0, K)$, with $K = K(X, X)$. Since we have a product of two Gaussian functions, using equation (A.3), we easily obtain

$$y|X \sim \mathcal{N}(0, K + \sigma_n^2 I). \tag{2.12}$$

For the non-zero mean case, we have that $y|X, b, B \sim \mathcal{N}(H^T b, K_y + H^T BH)$. We may integrate out $b$ and $B$ if a prior is available or use $b = 0$ and the limit $B^{-1} \to O$ (a matrix of zeros) if the prior is vague, see [Rasmussen & Williams '05].

## 2.3 General properties of Gaussian Processes

**Definition 2.2.** A stochastic process $Z(x)$ is said to be *strictly stationary* if its finite dimensional distributions are invariant under translations in the location $x$. That means that, for any set of points $\tau, x_1, \ldots, x_n \in \mathbb{R}^D$, the joint distribution of $Z(x_1), \ldots, Z(x_n)$ should be the same as the joint distribution of $Z(x_1+\tau), \ldots, Z(x_n+\tau)$. For this type of process, it is evident that the mean function must be constant.

A less restrictive condition than strict stationarity, when dealing with random processes, is to impose the mean $\mathbb{E}[Z(x)]$ to be a constant $m$ and that the covariance function $\mathbb{E}[(Z(x)-m)(Z(x')-m)]$ to be a function of $r = x - x'$ only. These processes are known as *second order*, *wide-sense (WSS)*, or *weakly stationary*. Evidently, strict stationarity implies weak stationarity, though the reverse need not be true. For a Gaussian process, however, the wide-sense stationarity conditions for the mean and covariance are necessary and sufficient for it to be strictly stationary. This follows from the fact that a Gaussian distribution is fully characterized by its first and second moments. If, moreover, the covariance function is a function of $x-x'$ only through the Euclidean distance $||x-x'||$, the process is said to be *isotropic*. The concept of isotropy arises when there is no special meaning attached to the axes being used.

For weakly stationary processes, there is a representation of the covariance funtion in the Fourier transform space:

**Theorem 2.3** (Bochner's Theorem, Theorem 1 of [Stein '99])**.** *A complex valued function $k(r)$ on $\mathbb{R}^D$ is the autocovariance function for a weakly stationary mean square continuous complex-valued random process on $\mathbb{R}^D$ if and only if it can be represented as*

$$k(r) = \int_{\mathbb{R}^D} e^{2\pi i s \cdot r} d\mu(s),$$

*where $\mu$ is a positive finite measure.*

If $\mu$ has a density $S(s)$, then

$$k(r) = \int_{\mathbb{R}^D} e^{2\pi i s \cdot r} S(s) ds,$$

and $S(s)$ is known as the spectral density (or power spectrum) of $k(r)$. The criterion to guarantee that the spectral density exists is to verify if $k(r)$ is an absolutely integrable function in $\mathbb{R}^D$. If, additionally, the covariance is isotropic and the spectral density exists, then $S(s)$ is a function of $||s||$ only. Refer to [Gihman & Skorohod '74] for the proof of Bochner's Theorem and further details.

If both $k(r)$ and $S(s)$ satisfy the conditions for the Fourier inversion to be valid, then by the Wiener-Khinchin theorem $k(r)$ and $S(s)$ are duals of each other and

$$S(s) = \int k(r) e^{-2\pi i s \cdot r} dr.$$

It is immediate that the power spectrum must be integrable, since $\int S(s) ds = k(0)$.

## 2.4 Continuity and differentiability

In many situations, when modeling a physical phenomenon, we may want the underlying stochastic process to be continuous, differentiable, or even smooth in time or space, for example. This required continuity or differentiability in a given sense translates the necessary physical realism. In some cases, we can relate the autocovariance function to these properties of the stochastic process.

Continuity and differentiability of a function $f(x)$, for $x \in \mathbb{R}^D$, at a point $x^*$ can be stated in terms of the convergence of sequences of the form $\{f(x_n)\}$, when $||x_n - x^*|| \to 0$ as $n \to \infty$. For stochastic processes, there are many forms of convergence. We will consider mean square (m.s.) and almost sure (a.s.) convergence and state properties that imply continuity and differentiability of a Gaussian process.

**Theorem 2.4** (Theorem 2.2.1 of [Adler '09]). *A random processes $Z(x)$ is continuous in mean square at the point $x^* \in \mathbb{R}^D$ if and only if its covariance function*

$$k(x, x') = \mathbb{E}[(Z(x) - \mathbb{E}[Z(x)])(Z(x') - \mathbb{E}[Z(x')])]$$

*is continuous at the point $x = x' = x^*$. Also, if $k(x, x')$ is continuous at every diagonal point $x = x'$, then the process is everywhere continuous in mean square.*

For a stationary process, this reduces to checking if $k(r)$ is continuous at $r = 0$. We stress that continuity in mean square does not imply sample path continuity, which is defined in the following.

**Definition 2.5.** Let $z(x)$ be an $\mathbb{R}^m$-valued function that is a realization of the random process $Z(x)$ for $x \in \mathbb{R}^n$. Then, the set in $\mathbb{R}^{n+m}$ determined by the points $\{(x, z(x)), x \in \mathbb{R}^n\}$ is called a *sample function*, or *sample path* of the process $Z(x)$.

**Theorem 2.6** (Theorem 2.2.2 of [Adler '09]). *If the derivative $\partial^2 k(x, x')/\partial x_i \partial x_i'$ exists and is finite at the point $(x^*, x^*) \in \mathbb{R}^{2D}$, then, if $e_i$ denotes the $i$-th canonical basis vector, the limit*

$$\frac{\partial Z(x^*)}{\partial x_i} = \lim_{h \to 0} \frac{Z(x^* + he_i) - Z(x^*)}{h}$$

*exists, and $\partial Z(x^*)/\partial x_i$ is called the mean square derivative of $Z(x)$ at the point $x^*$. If this exists for every $x \in \mathbb{R}^D$, then $Z(x)$ is said to posses a m.s. derivative. The covariance function of $Z_i(x)$ is then given by*

$$\frac{\partial^2 k(x, x')}{\partial x_i \partial x_i'}.$$

Similarly, the second order derivatives of $Z(x)$, $\partial^2 Z(x)/\partial x_i \partial x_j$, for $1 \leq i, j \leq D$, are defined as

$$\frac{\partial^2 Z(x)}{\partial x_i \partial x_j} = \lim_{h,l \to 0} \frac{Z(x + he_i + le_j) - Z(x + he_i) - Z(x + le_j) + Z(x)}{hl},$$

and are Gaussian Processes whose covariance function is the fourth order derivative of $k(x, x')$

$$\frac{\partial^4 k(x, x')}{\partial x_i \partial x_j \partial x'_i \partial x'_j}.$$

For a stationary process with covariance function $k(x, x')$, we can write $k(x, x') = k(r)$, where $r = x - x'$, and the m.s. continuity and differentiability properties of the process are determined by the smoothness of $k(r)$ at the point $r = 0$. In this case, if the $2m$-th order partial derivative of $k(r)$, $\partial^{2m} k(r)/\partial^2 r_{i_1} \dots \partial^2 r_{i_m}$, exists and is finite at $r = 0$, then the $m$-th order partial derivative of $Z(x)$, $\partial^m Z(x)/\partial x_{i_1} \dots \partial x_{i_m}$, exists for every $x$ as a mean square limit.

A stronger definition of continuity is given by means of almost sure convergence.

**Definition 2.7.** A stochastic process $Z(x)$ is said to be almost surely continuous at $x^*$ if for every sequence $\{x_n\}_{n=1,\dots}$ for which $||x_n - x^*|| \to 0$ as $n \to \infty$, and is denoted by $Z(x_n) \xrightarrow{a.s.} Z(x^*)$. We say that $Z(x)$ is almost surely continuous throughout a set $A \subseteq \mathbb{R}^D$ if it is almost surely continuous at each point $x \in A$. This type of continuity is referred as sample path continuity.

In particular, for Gaussian Processes, a.s. continuity is, again, a consequence of a certain condition on the covariance function, as we see in the following.

**Theorem 2.8** (Theorem 3.4.1 of [Adler '09])**.** *Let $Z(x)$, with $x \in \mathbb{R}^D$, be a real-valued, zero-mean, Gaussian process with a continuous covariance function. Then, if for some $0 < C < \infty$ and some $\varepsilon \geq 0$,*

$$\mathbb{E}[(Z(x) - Z(x'))^2] \leq \frac{C}{|\log(||x - x'||)|^{1+\varepsilon}},$$

*for all $x, x'$ in the unit cube $I_0$, $Z$ has, with probability one, continuous sample functions over $I_0$.*

If the Gaussian process $Z(x)$ is stationary, this translates as requiring that

$$k(0) - k(r) \leq \frac{C}{|\log(||r||)|^{1+\varepsilon}},$$

for some $0 < C < \infty$ and some $\varepsilon \geq 0$.

## 2.5 Length-scale

For a 1-dimensional Gaussian process, the length-scale of the process can be understood in terms of the number of upcrossings at a certain level $u$ as in [Adler '09].

**Definition 2.9.** Let $f(x)$, with $x \in \mathbb{R}$, be a continuous function on an interval $I = [a, b]$ such that $f(x)$ is not identically equal to $u$ in any subinterval, and neither $f(a)$ nor $f(b)$

is equal to $u$. Then $f$ is said to have an upcrossing of level $u$ at the point $x_0$ if there exists an $\varepsilon > 0$ such that $f(x) \leq u$ in $(x_0 - \varepsilon, x_0)$ and $f(x) \geq u$ in $(x_0, x_0 + \varepsilon)$.

The number of such points $x_0$ in $I$ is called the number of upcrossings of $u$ by $f$ in $I$, and it is denoted by $N_u$.

**Theorem 2.10** (Theorem 4.1.1 of [Adler '09]). *If $N_u$ is the number of upcrossings of the level $u$ by a zero-mean stationary almost surely continuous Gaussian process on $[0, 1]$, then*

$$\mathbb{E}[N_u] = \frac{1}{2\pi}\sqrt{-\frac{k''(0)}{k(0)}}\exp\left\{-\frac{u^2}{2k(0)}\right\}. \tag{2.13}$$

This theorem is valid regardless of the finiteness of $k''(0)$. Thus, only if the Gaussian process is mean square differentiable, there is a finite number of upcrossings in a given finite interval (refer to Theorem 2.6). For the squared exponential kernel in dimension 1, $k_{SE}(d) = \exp\{-d^2/(2l^2)\}$, with $d = ||x - x'||$, the expected number of upcrossing of the corresponding 1-dimensional Gaussian process in the interval $[0, 1]$ is $(2\pi l)^{-1}$, which confirms $l$ as a length-scale parameter, see Figure 2.7 for an illustrative example of the behavior of sample paths when different length-scales are fixed.

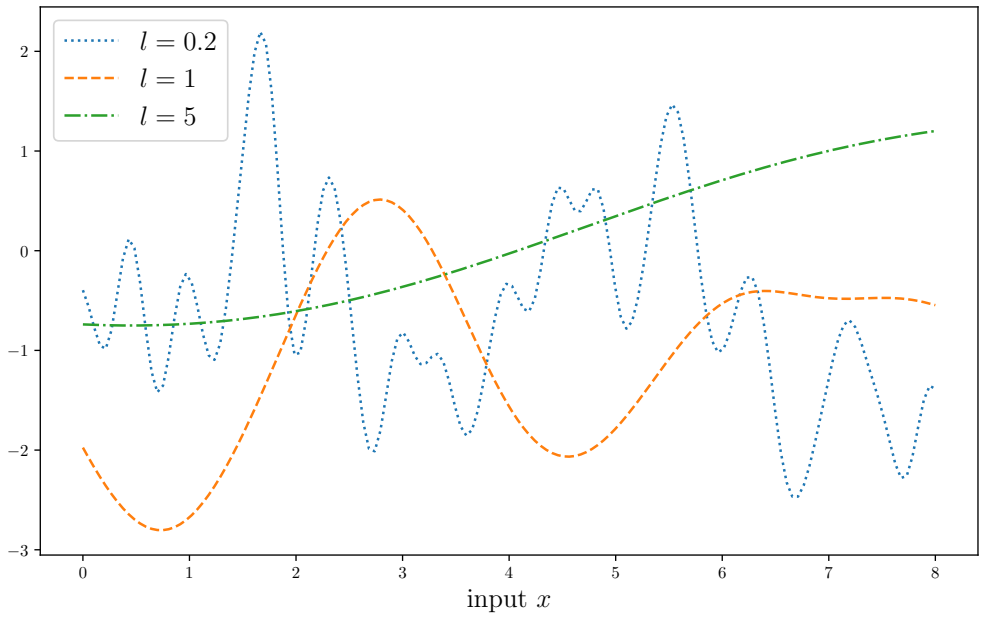



Figure 2.7: Sample functions of GP with squared exponential kernel $k_{SE}(d) = \exp\{-d^2/(2l^2)\}$, with $d = ||x - x'||$, and different length-scales $l$.

## 2.6 Examples of covariance functions

In the study of integral operators, any integral transform of a function $f$ can be written as

$$(Tf)(x) = \int_{\chi} f(x)k(x,x')d\mu(x'),$$

where $\mu$ denotes a measure and $k(x,x')$ is the *kernel* or *nucleus* of the transform, a function mapping a pair of inputs $x \in \chi$ and $x' \in \chi$ into $\mathbb{R}$. An arbitrary function will not necessarily be a covariance function, since the Gram matrix $K$ for a set $\{x_i\}_{i=1,\dots,n}$ with entries $K_{ij} = k(x_i, x_j)$ must be a valid covariance matrix for any number of arbitrary input points. A valid covariance matrix $K$ is symmetric and positive semidefinite, this translates to a kernel that is symmetric, $k(x,x') = k(x',x)$, and positive semidefinite, that is

$$\int_{\chi\times\chi} f(x)k(x,x')f(x')d\mu(x)d\mu(x') \geq 0,$$

for all functions $f \in L^2(\chi,\mu)$, which means that $f : \chi \to \mathbb{R}$ is such that $||f||_{L^2(\chi,\mu)} = \left(\int_{\chi} |f(x)|^2 d\mu(x)\right)^{1/2} < \infty$.

We present a selection of the most relevant covariance functions used for inputs in $\mathbb{R}^D$. For a broader discussion refer to [Rasmussen & Williams '05], [MacKay '98], and [Duvenaud '14].

1. The *squared exponential* covariance function is the most widely used covariance kernel in machine learning. It is given by the Gaussian function

$$k_{SE}(x,x') = \exp\left\{-\frac{||x-x'||^2}{2l^2}\right\},$$

and gives rise to an infinitely m.s. differentiable Gaussian process, given that it is a stationary kernel with smooth covariance function at the origin. Furthermore, the squared exponential is a function of $d = ||x-x'||$, and has an analytic Fourier transform, which is also a Gaussian function,

$$\mathcal{F}(k_{SE})(s) = S(s) = (2\pi l^2)^{D/2}\exp\{-2\pi^2 l^2 s^2\}.$$

2. A class of more realistic isotropic covariance functions, that unlike the squared exponential do not confer infinite derivatives to the GP, are the *Matérn* covariance functions named after the Swedish forestry statistician Bertil Matérn (1917-2007). They are given by

$$k_\nu(d) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}d}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}d}{l}\right),$$

with $\nu, l > 0$ and $K_\nu$ is the modified Bessel function of the second kind of order $\nu$. The parameter $\nu$ is a smoothness parameter which relates to $k_\nu(d)$ being $\lceil\nu\rceil - 1$ times differentiable, while $l$ has a role of length-scale parameter. As seen

in [Rasmussen & Williams '05], for $\nu = p + 1/2$, and $p$ a non-negative integer, the expression of the covariance simplifies to

$$k_{\nu=p+1/2}(d) = \exp\left\{-\frac{\sqrt{2\nu}d}{l}\right\}\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{\sqrt{8\nu}d}{l}\right)^{p-i}.$$

For $p = 1$ and $p = 2$, we obtain the most interesting cases, which are differentiable but yet distinguishable from a smooth process. Their covariance functions are

$$k_{\nu=3/2}(d) = \exp\left\{-\frac{\sqrt{3}d}{l}\right\}\left(1+\frac{\sqrt{3}d}{l}\right),$$

and

$$k_{\nu=5/2}(d) = \exp\left\{-\frac{\sqrt{5}d}{l}\right\}\left(1+\frac{\sqrt{5}d}{l}+\frac{5d^2}{3l^2}\right).$$

All covariance functions of the Matérn have analytic expressions for their respective spectral densities and for $\nu \to \infty$ they converge to the squared exponential kernel. See [Stein '99] and [Rasmussen & Williams '05] for further details.

In Figure (2.8), we have the functions for different value of the parameter $\nu$, and Figures (2.9) and (2.10) exemplify the behavior of sample paths corresponding to such kernels.

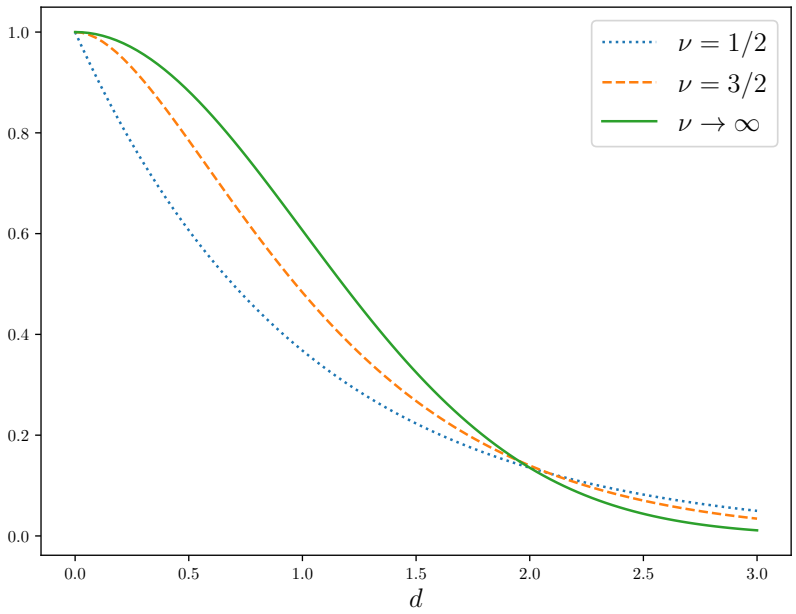



Figure 2.8: Matérn kernels for different values of $\nu$ and $l = 1$

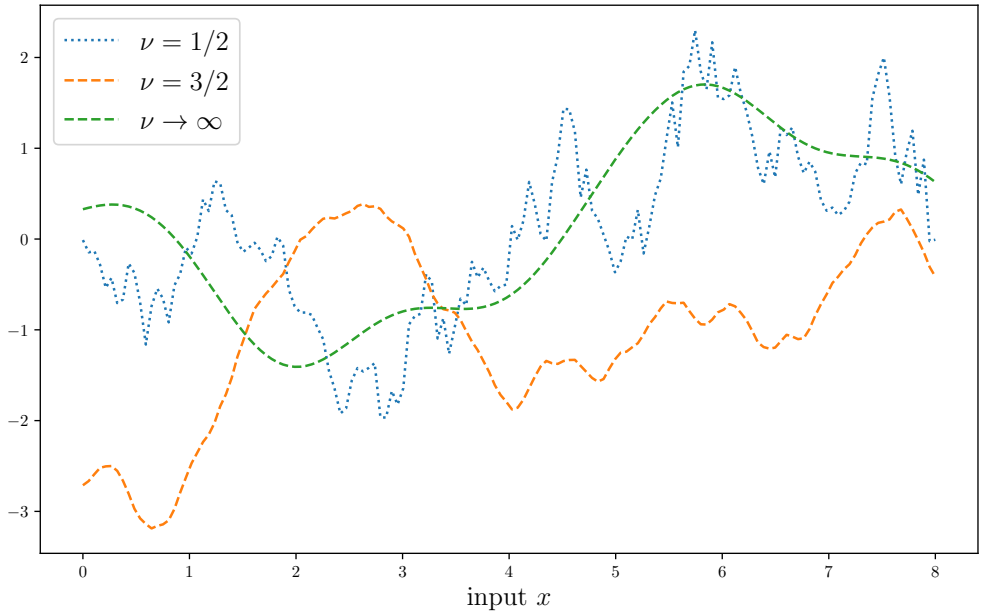



Figure 2.9: Prior Matérn samples for different values of $\nu$ and $l = 1$.

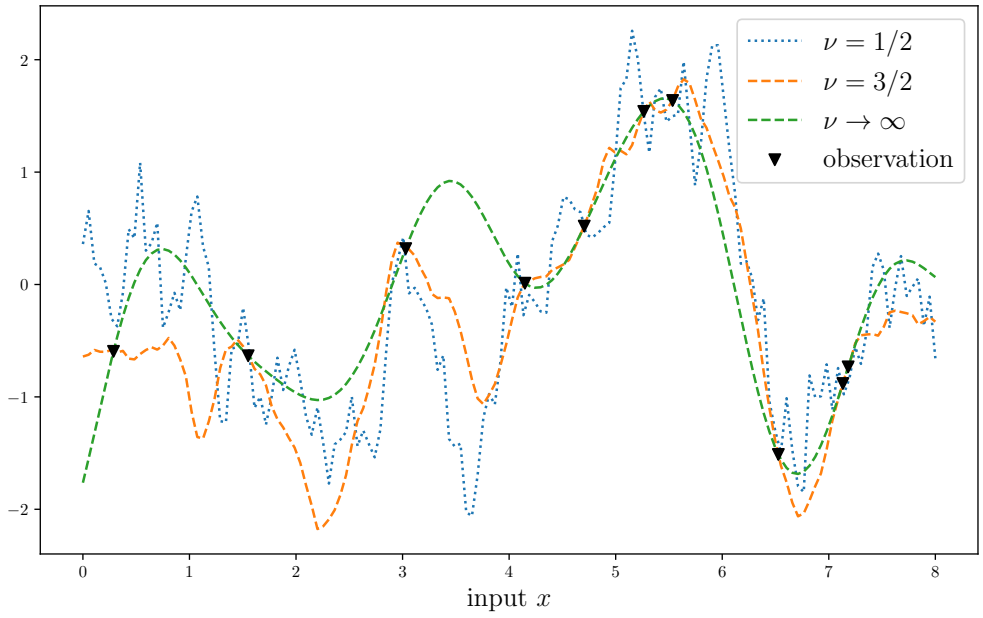



Figure 2.10: Posterior samples for different values of $\nu$ and optimized hyperparameter $l$.

3. When the parameter $\nu$ of the Matérn class is equal to $1/2$, we have a rough process which is known as the *Ornstein-Uhlenbeck (OU)* process with the exponential covariance function
$$k_{OU}(d) = \exp\left\{-\frac{d}{l}\right\}.$$

4. A similar covariance kernel to the Ornstein-Uhlenbeck is given by the $\gamma$-*exponential* class with covariance function
$$k_{\gamma-\exp}(d) = \exp\left\{-\left(\frac{d}{l}\right)^{\gamma}\right\},$$
for $0 < \gamma \leq 2$, which is an alternative but less flexible class than the Matérn as mentioned in [Stein '99], since it is not m.s. differentiable except for $\gamma = 2$. In Figure (2.11), we see the kernel for different values of the parameter $\gamma$.

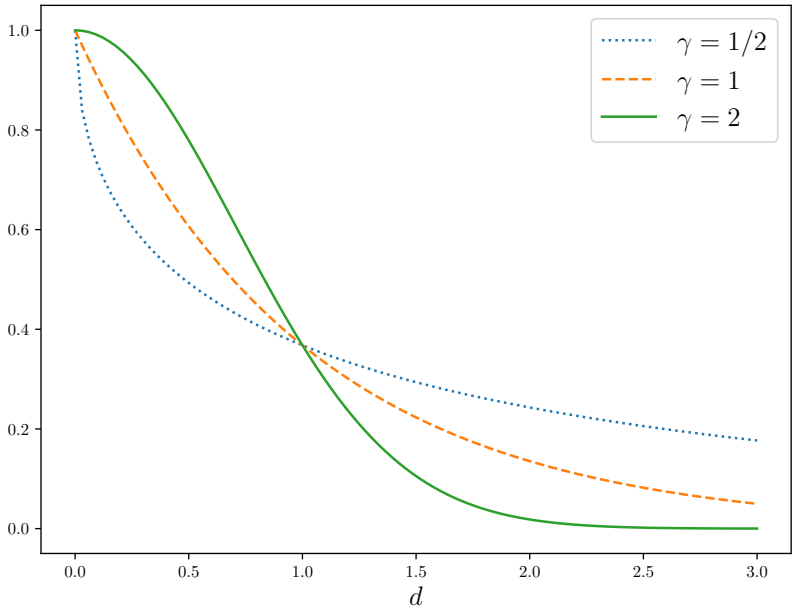



Figure 2.11: $\gamma$-exponential kernels for different values of $\gamma$ and $l = 1$.

5. The *rational quadratic* covariance function, with parameters $\alpha, l > 0$, is given by
$$k_{RQ}(d) = \left(1 + \frac{d^2}{2\alpha l^2}\right)^{-\alpha}.$$
A Gaussian process with this covariance kernel is m.s. differentiable for any value of $\alpha$, see [Rasmussen & Williams '05].

   In Figure (2.12), we observe the kernel for different values of the parameter $\alpha$, and Figures (2.13) and (2.14) exemplify the behavior of sample paths.

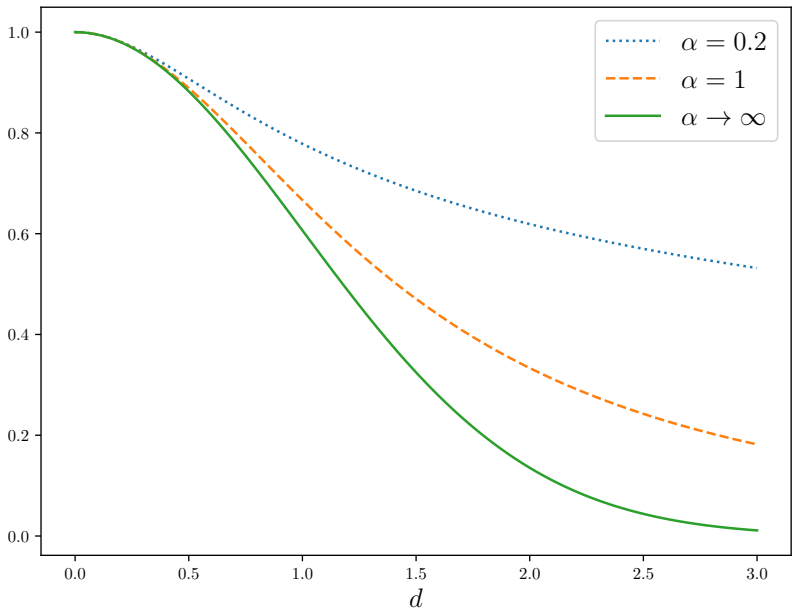



Figure 2.12: Rational quadratic kernels for different values of $\alpha$ and $l = 1$

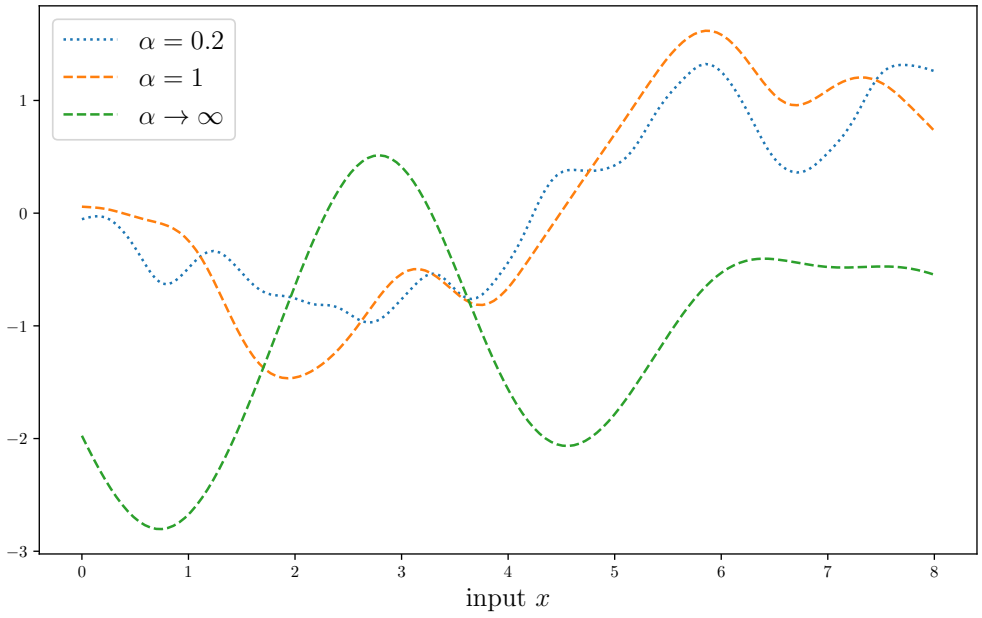



Figure 2.13: Prior samples with rational quadratic kernel for different values of $\alpha$ and $l = 1$.

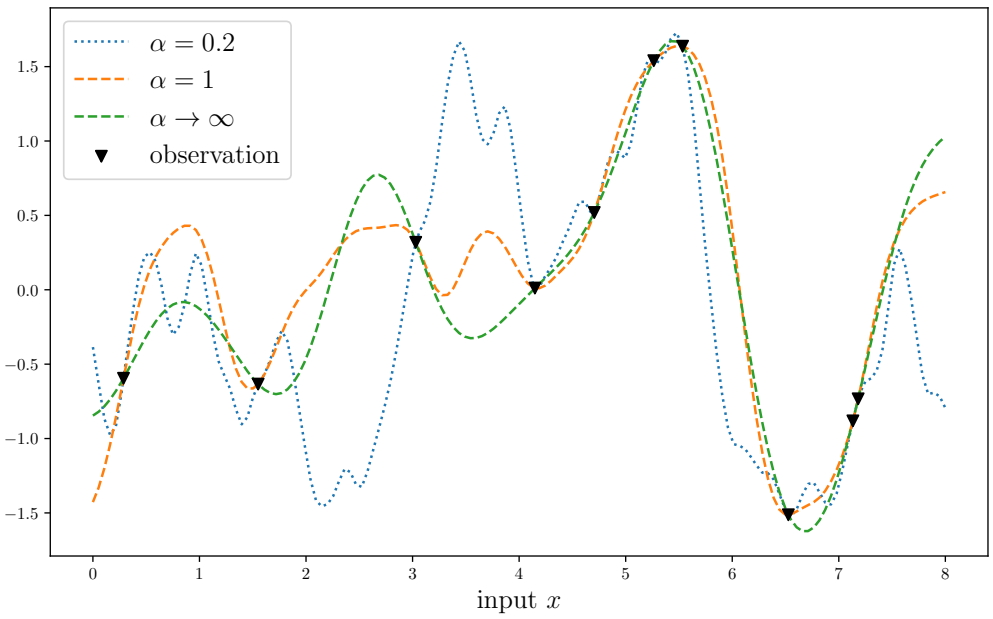



Figure 2.14: Posterior samples with rational quadratic kernel different values of $\alpha$ and hyperparameter $l$ optimized.

6. For non-stationary processes, a simple covariance function is given by using a general covariance matrix $\Sigma$ to create a *dot product* kernel:

$$k(x, x') = \sigma_0^2 + x^T \Sigma x'.$$

The special case $\Sigma = I$, yields $k(x, x') = \sigma_0^2 + x^T x'$, while $\Sigma = 0$, yields the *constant* covariance function $k(x, x') = \sigma_0^2$. Another possible choice is the polynomial one, $k(x, x') = (\sigma_0^2 + x^T \Sigma x')^p$, for a positive integer $p$.

7. Periodization may be obtained by mapping the inputs by a periodic function, as $u(x) = (sin(x), cos(x))$ for 1-dimensional inputs, and using this in a known kernel. For the squared exponential, this gives us

$$k(x, x') = \exp\left\{ - \frac{2 \sin^2(\frac{x - x'}{2})}{l^2} \right\}.$$

This kind of approach is known as *warping* or *embedding* as in [MacKay '98].

8. We may expect to have different length-scale behavior throughout the input space. Simply replacing the length-scale parameter $l$ with a function $l(x)$ in the covariance expression will not necessarily produce a positive semidefinite kernel. [Gibbs '97] constructs a covariance kernel based on the squared exponential for which the

characteristic length-scale is a function of the input points. This function is given by

$$k(x,x') = \prod_{d=1}^{D} \left( \frac{2l_d(x)l_d(x')}{l_d^2(x) + l_d^2(x')} \right)^{1/2} \exp\left\{ -\sum_{d=1}^{D} \frac{(x_d - x'_d)^2}{l_d^2(x) + l_d^2(x')} \right\},$$

where each $l_d(x)$ is a positive function.

Finally, it is worth mentioning that there are straightforward ways to construct new covariance functions from previously known ones. For this, if $k_1(x,x')$ and $k_2(x,x')$ are valid covariance functions, so is their sum $k_1(x,x') + k_2(x,x')$ and their product $k_1(x,x')k_2(x,x')$. If $k(x,x')$ is a covariance function, a deterministic function $a(x)$ produces the covariance kernel $a(x)k(x,x')a(x')$. An extension of this is called the *blurring* effect when performing a convolution with a fixed kernel $h(w,w')$, with which it is possible to construct the covariance function $\int h(x,z)k(z,z')h(z',x')dzdz'$.

## 2.7 Model selection

The families of covariance functions presented previously have free hyperparameters such as length-scale which must be chosen in some way. While some hyperparameters may be easily interpretable, this is not always the case. Nevertheless, efficiently selecting the best values is extremely important in order to make accurate predictions. Furthermore, while the context may give us some information about properties like stationarity, isotropicity or periodicity, for example, our knowledge about the exact form of the covariance function is vague. Therefore, we must compare different covariance functions, and values for their respective hyperparameter in order to determine these elements of the modeling. This may be made level-wise, first selecting the general model (GP vs. other types of regression), then the covariance kernel, and then the hyper-parameters, for example.

In the following, we will briefly explore two ways of performing model selection: Bayesian and cross-validation.

### 2.7.1 Bayesian model selection

In a general framework, we can construct a hierarchical approach with a finite number of models $\mathcal{M}_i$. For each model $\mathcal{M}_i$ (upper level), there are parameters $w$ (lower level) which depend on hyper-parameters $\theta$ (medium level), but there may be as many levels as needed. We intend to select the model, hyperparameters and parameters which maximize the posterior probability of each one of these elements.

First, we specify priors $p(\mathcal{M}_i)$, $p(\theta|\mathcal{M}_i)$ and $p(w|\theta,\mathcal{M}_i)$. Broad or non-informative priors can be chosen if the prior knowledge about each set of elements is vague. Then, one level at a time, we infer its free elements. To begin, we use Bayes rule to infer the parameters of the bottom level

$$p(w|y,X,\theta,\mathcal{M}_i) = \frac{p(y|X,w,\mathcal{M}_i)p(w|\theta,\mathcal{M}_i)}{p(y|X,\theta,\mathcal{M}_i)},$$

where $p(y|X, w, \mathcal{M}_i)$ is the likelihood with implicit dependence on $\theta$ through $w$, and

$$p(y|X, \theta, \mathcal{M}_i) = \int p(y|X, w, \mathcal{M}_i)p(w|\theta, \mathcal{M}_i)dw$$

is the marginal likelihood (also called evidence). For the next level, the hyperparamenter level, we again use Bayes rule and obtain the marginal likelihood

$$p(\theta|y, X, \mathcal{M}_i) = \frac{p(y|X, \theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)}{p(y|X, \mathcal{M}_i)},$$

with

$$p(y|X, \mathcal{M}_i) = \int p(y|X, \theta, \mathcal{M}_i)p(\theta|\mathcal{M}_i)d\theta.$$

Finally, at the top level we have

$$p(\mathcal{M}_i|y, X) = \frac{p(y|X, \mathcal{M}_i)p(\mathcal{M}_i)}{p(y|X)},$$

with normalizing constant

$$p(y|X) = \sum_i p(y|X, \mathcal{M}_i)p(\mathcal{M}_i).$$

This approach demands many integral evaluations. If these integrals are not analytically tractable, we must resort to analytical approximations such as Markov Chain Monte Carlo (MCMC), and if a step is particularly difficult, it may be substituted with the maximization of the likelihood instead of using the full knowledge (of the prior and marginal). When the expressions or approximation of the posteriors are available, the selection is straightforward.

For Gaussian process regression, the role of hyper-parameters and models is quite unambiguous, with the various covariance functions determining the model and the free variables in each covariance function being the hyperparameters. For other models, such as neural networks, the parameters are also identifiable and interpretable (see [Rasmussen & Williams '05] and [MacKay '03]), but since Gaussian Processes are non-parametric models this may not be so simple in this case. The parameters in the Gaussian process modeling are the noise-free values of the latent function $z(x)$ at the training locations, thus we have as many parameters as training points. The positive side is that the bottom level Bayesian inference concerning the parameters has already been performed in Section 2.2.

By Equation (2.12), thus, we have a log marginal likelihood given by

$$\log p(y|X, \theta) = \frac{1}{2}y^T K_y^{-1} y - \frac{1}{2}\log\det(K_y) - \frac{n}{2}\log(2\pi),$$

with dependence on the hyperparameters $\theta$ through $K_y$, see Figure 2.15 for an illustrative example of a negative log marginal likelihood.

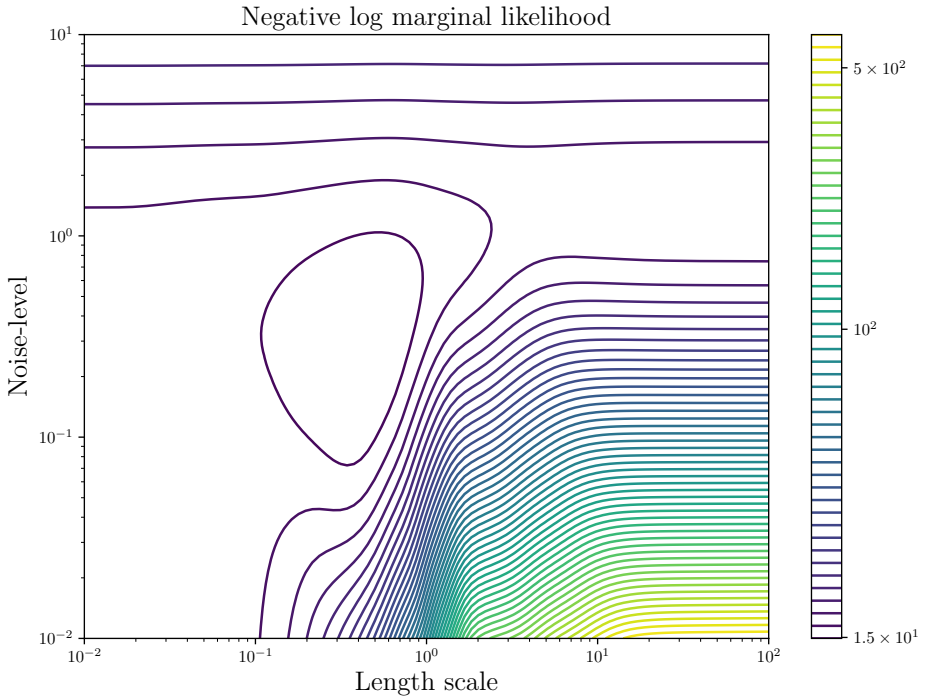



Figure 2.15: Negative log marginal likelihood for the data and model of Figure 2.6 with hyperparameter $\sigma^2 = 0.897^2$.

Maximizing on the hyperparameters passes through obtaining the derivatives

$$\frac{\partial}{\partial \theta_i} \log p(y|X,\theta) = \frac{1}{2} y^T K_y^{-1} \frac{\partial K_y}{\partial \theta_i} K_y^{-1} y - \frac{1}{2} \mathrm{tr}\Big( K_y^{-1} \frac{\partial K_y}{\partial \theta_i} \Big).$$

For this, observe that

$$y^T K_y^{-1} \frac{\partial K_y}{\partial \theta_i} K_y^{-1} y = \mathrm{tr}(y^T K_y^{-1} \frac{\partial K_y}{\partial \theta_i} K_y^{-1} y) = \mathrm{tr}(K_y^{-1} y (K_y^{-1} y)^T \frac{\partial K_y}{\partial \theta_i})$$

$$\implies \frac{\partial}{\partial \theta_i} \log p(y|X,\theta) = \frac{1}{2} \mathrm{tr}\Big( (K_y^{-1} y (K_y^{-1} y)^T - K_y^{-1}) \frac{\partial K_y}{\partial \theta_i} \Big).$$

This procedure has a cost of $\mathcal{O}(n^3)$ for the inversion of $K_y$ and $\mathcal{O}(n^2)$ for computing the derivative of $K_y$ with respect to each hyper-parameter $\theta_i$, therefore a gradient based optimizer may be used for maximizing the log marginal likelihood.

The use of the marginal likelihood (also refered as *evidence*), for example $p(y|X,\mathcal{M}_i)$ in $p(\mathcal{M}_i|y,X) = \frac{p(y|X,\mathcal{M}_i)p(\mathcal{M}_i)}{p(y|X)}$, automatically incorporates a trade-off between model complexity and model fit. This happens because frequently flat priors $p(M_i)$ are used for the models, such that we have approximately

$$p(\mathcal{M}_i|y,X) \propto p(y|X,\mathcal{M}_i).$$

In Figure 2.16, we have a schematic plot of the marginal likelihood $p(y|X, \mathcal{M}_i)$ in the vertical axis when the number of data points $n$ and input data $X$ are fixed for a particular model $\mathcal{M}_i$. The horizontal axis is a representation of all possible output sets $y$, with a particular set indicated by the vertical dotted line. Since the curves describe probability distributions, they must integrate 1 over all possible datasets.

A model which is too simple will describe well few possible data sets, for which the marginal likelihood will attain a large value, but it is unlikely that this model will generate a particular dataset $\{X, y\}$. A complex model has a broader range of possible targets, which translates to a flatter marginal likelihood, thus, it describes well a large collection of possible outputs, but since it must integrate 1, the value of the marginal likelihood at any particular data set is small. Therefore, there is a preference for a model with an intermediate level of complexity. This effect is known as *Occam's Razor*, a principle of "parsimony of explanations", which chooses the simplest model that still explains well the data. For more details, see [Rasmussen & Ghahramani '01].

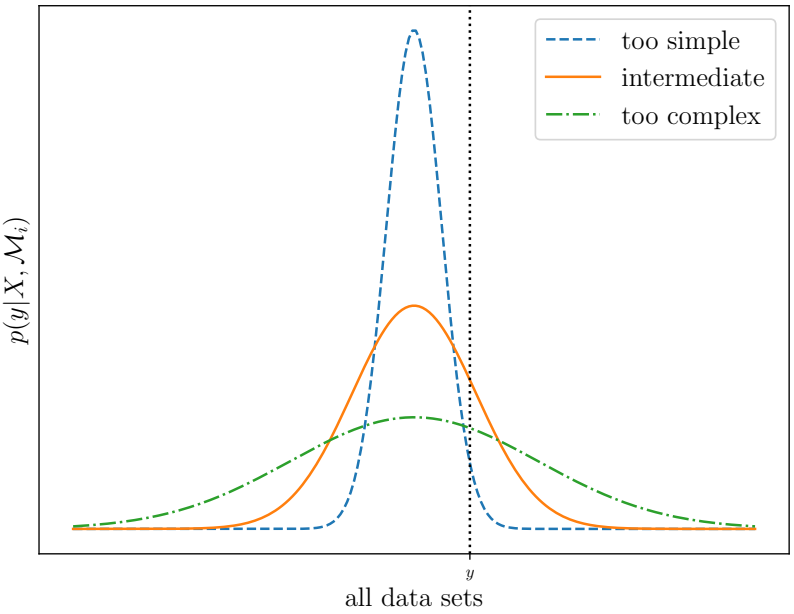



Figure 2.16: Schematic view of the marginal likelihood $p(y|X, \mathcal{M}_i)$ over all possible datasets, with fixed inputs $X$, for three levels of model complexity. A specific data set $\{X, y\}$ is indicated by the vertical dotted line.

### 2.7.2 Cross-validation

Gaussian processes are a very powerful and flexible tool, but, because of that, attention is needed when training the model. When using all the data for training, we may come across problems like overfitting, when the training error is very small, yet the model fails

to generalize for new points.

With the cross-validation (CV) procedure, we can lower the generalization error and prevent overfitting. It consists in splitting the data set in two: the *training set* which will be used for training the model and the *validation set* which is used to monitor the performance of the model. Using this hold-out method, if the validation set is small, one obtains estimates with large variance and one loses the information given by the points in the validation set. To avoid these problems, generally the $k$-fold setting is used: the original training set is split into $k$ equally sized disjoint sets and a cross-validation procedure is performed $k$ times, each of them using one of the sets as the validation set and the union of the remaining $k-1$ as the training set. The value of $k$ is usually set between 3 and 10. When $k = n$, this is called *leave-one-out cross-validation* (LOO-CV) and consists of training $n$ models. In general, this is an extremely expensive procedure, however some models including GP have computational shortcuts, which we will briefly discuss.

A possible objective function used for measuring the fit, which will be maximized w.r.t. the hyper-parameters, is the log predictive probability when leaving out of the training a validation set. In short, maximizing the log predictive probability means that we wish to select parameters that favor the dataset we have, giving higher probability on closer values of the outputs than to other ones.

Let $\xi$ be the set of indices of the points in the validation set. When training with the points in the set $X_{-\xi}$, which consists of all data points except for the ones with indices in $\xi$, the equations for the predictive mean and variance of a zero-mean GP (we drop the subscript $y$ of $K_y$ in a possible noisy case for simplicity of notation) give us

$$y_\xi | X, y_{-\xi}, \theta \sim \mathcal{N}\big(K(X_\xi, X_{-\xi})(K(X_{-\xi}, X_{-\xi})^{-1} y_{-\xi}\big),$$

$$K(X_\xi, X_\xi) - K(X_\xi, X_{-\xi})(K(X_{-\xi}, X_{-\xi}))^{-1} K(X_{-\xi}, X_\xi)).$$

The notation in this section and in the next chapter will be as follows. We will use $v_\zeta$ for a vector containing the entries of $v$ with indices in the set $\zeta$, $v_{-\zeta}$ for the vector of entries of $v$ with indices not in $\zeta$, $[A]_{[\zeta,\gamma]}$ for the submatrix of $A$ containing the rows with indices in the set $\zeta$ and columns with indices in the set $\gamma$, and similarly as in the vector case for the submatrices $[A]_{[-\zeta,\gamma]}$, $[A]_{[\zeta,-\gamma]}$ and $[A]_{[-\zeta,-\gamma]}$. The sets $\zeta$ and $\gamma$ may consist of only one index $i$, in this case we will simply use $i$ instead of a set.

Back to our Gaussian process prediction, observe that the costly part is the inversion of the matrix $K(X_{-\xi}, X_{-\xi})$, which will be of size $(n - \#\xi) \times (n - \#\xi)$, with $\#\xi = n/k$ in the $k$-fold setting. If we rearrange the points so that the ones with indices in $\xi$ come last, we can rewrite the matrix $K$ as

$$K = K(X, X) = \begin{bmatrix} [K]_{[-\xi,-\xi]} & [K]_{[-\xi,\xi]} \\ [K]_{[\xi,-\xi]} & [K]_{[\xi,\xi]} \end{bmatrix}.$$

Therefore, using Equations (2.5) and (2.6), we have that

$$y_\xi | X, y_{-\xi}, \theta \sim \mathcal{N}\big([K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]} y_{-\xi}, [K]_{\xi,\xi} - [K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]}[K]_{[-\xi,\xi]}\big).$$

By the block matrix inversion identity (A.7), we have that

$$K^{-1} = \begin{bmatrix} A & B \\ B^T & \mathcal{Q}^{-1} \end{bmatrix},$$

with

$$A = [K]^{-1}_{[-\xi,-\xi]} + [K]^{-1}_{[-\xi.-\xi]}[K]_{[-\xi,\xi]}\mathcal{Q}^{-1}[K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]},$$
$$B^T = -\mathcal{Q}^{-1}[K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]},$$

and

$$\mathcal{Q} = \Big[[K^{-1}]_{[\xi,\xi]}\Big]^{-1} = [K]_{[\xi,\xi]} - [K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]}[K]_{-\xi,\xi}.$$

And this implies that

$$\Big[[K^{-1}]_{[\xi,\xi]}\Big]^{-1}[K^{-1}y]_{[\xi]} = \Big[[K^{-1}]_{[\xi,\xi]}\Big]^{-1}\Big(B^T y_{-\xi} + \mathcal{Q}^{-1}y_\xi\Big) = -[K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]}y_{-\xi} + y_\xi.$$

Observe that now we are able to rewrite the expression of $p(y_\xi|X, y_{-\xi}, \theta)$ without needing to obtain the inverse of $[K]_{[-\xi,-\xi]}$ for each set $\xi$. This predictive probability is given by

$$y_\xi|X, y_{-\xi}, \theta \sim \mathcal{N}\big(y_\xi - \mathcal{Q}[K^{-1}y]_\xi, \mathcal{Q}\big).$$

For the particular case of the LOO-CV, the set of indices $\xi$ consists only of a index $i$. Then, $\mathcal{Q} = 1/[K^{-1}]_{ii}$, and the predictive probability is given by

$$y_i|X, y_{-i}, \theta \sim \mathcal{N}\big(\mu_i, \sigma_i^2\big).$$

with

$$\mu_i = y_i - [K^{-1}y]_i/[K^{-1}]_{ii},$$

and

$$\sigma_i^2 = 1/[K^{-1}]_{ii}.$$

Note that the computational cost is $\mathcal{O}(n^3)$ for inverting $K$ and $\mathcal{O}(n^2)$ for the LOO-CV when $K^{-1}$ is known.

Thus, the log predictive probability of the LOO-CV scheme, which will be the objective function for optimization, is given by

$$L_{\text{LOO}}(X, y, \theta) = \sum_{i=1}^{n} \log p(y_i|X, y_{-i}, \theta) = \sum_{i=1}^{n} -\frac{1}{2}\log(\sigma_i^2) - \frac{(y_i - \mu_i)^2}{2\sigma_i^2} - \frac{1}{2}\log(2\pi).$$

The derivatives for the means and variances w.r.t. the hyperparameters are

$$\frac{\partial \mu_i}{\partial \theta_j} = -\frac{\Big[K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}y\Big]_i[K^{-1}]_{ii} - [K^{-1}y]_i\Big[K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\Big]_{ii}}{[K^{-1}]^2_{ii}},$$

and

$$\frac{\partial \sigma_i^2}{\partial \theta_j} = -\frac{\Big[K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\Big]_{ii}}{[K^{-1}]^2_{ii}}.$$

Now, by the chain rule, the derivative of the log predictive probability is

$$\frac{\partial L_{\mathrm{LOO}}}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial \log p(y_i|X, y_{-i}, \theta)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \theta_j} + \frac{\partial \log p(y_i|X, y_{-i}, \theta)}{\partial \sigma_i^2} \frac{\partial \sigma_i^2}{\partial \theta_j} =$$

$$\sum_{i=1}^{n} [K^{-1}y]_i \left( \frac{\left[K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}y\right]_i}{[K^{-1}]_{ii}} - \frac{[K^{-1}y]_i \left[K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\right]_{ii}}{2[K^{-1}]_{ii}^2} \right) - \frac{\left[K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\right]_{ii}}{2[K^{-1}]_{ii}}.$$

The computational cost is $\mathcal{O}(n^3)$ for inverting $K$ and $\mathcal{O}(n^3)$ for the computation of the derivative for each hyper-parameter, since the matrix multiplication $K^{-1}\frac{\partial K}{\partial \theta_j}$ is unavoidable. Therefore, this method is more costly than the previously discussed Bayesian model selection which was based on the marginal likelihood.

A discussion about the CV procedure for a non-zero mean Gaussian process is present in [Le Gratiet '13].

# Chapter 3

# Multi-Fidelity Modeling

## 3.1 A gist of multi-fidelity

Many accurate computer simulations are too costly to be run a considerable amount of times for them to describe appropriately the underlying modeled phenomenon, with only a few data points that can be obtained in a reasonable amount of time. Another problem that can be encountered when modeling a phenomenon is the need to specify a large number of parameters, which can be difficult to identify or measure directly.

However, in some situations, there are multiple computational models available which describe the phenomenon of interest. These computational models can have varying fidelity and computational cost. High-fidelity models represent the behavior of the system accurately for the intended application, yet often are expensive and multiple realizations cannot be afforded. Low-fidelity models estimate the same phenomenon with a lower accuracy than the high-fidelity model, but are less expensive and many realizations are obtainable. The low-fidelity models are usually obtained through, for example, dimensionality reduction, linearization, simpler physics models, coarser domains, etc.

A method to integrate the information of the simpler and inexpensive simulations, which capture basic features of the phenomenon, and data of the expensive and more reliable simulations would be a practical approach for understanding the phenomenon given the cost restrictions. This is precisely what multi-fidelity models intend to accomplish by combining the information of both low and high-fidelity models.

The fidelity level of a model concerns how well the model approximates a physical phenomenon/system. It is commonly associated with:

(1) How close the mathematical model is to reality: many differential equations can model the same systems (for example, inclusion or not of turbulent effects when describing a flow, linearization of equations, simplification of boundary conditions, etc.).

(2) Changing the discretization model by using finer or coarser discretizations.

(3) Using experimental data, which constitute the high-fidelity data.

In general, multi-fidelity models require the construction of surrogate models to reduce the computational cost when a large number of expensive simulations are needed.

Surrogate models may be used for the low-fidelity model and for the high-fidelity one, but usually even with this the high-fidelity model alone would require too many evaluations. The idea behind multi-fidelity (surrogate) models is to correct the low-fidelity models using the high-fidelity models. Many correction methods are known, some of which we briefly describe below for a 2-level fidelity model (but which are easily extended for more levels). We denote by $\widehat{y}_{HF}(x)$ the estimator of the high-fidelity model at the point $x$ and $y_{LF}(x)$ the low-fidelity model at $x$. The following are some of the most common corrections:

1. *Additive correction:*
$$\widehat{y}_{HF}(x) = y_{LF}(x) + \delta(x),$$
with $\delta(x)$ an additive correction/discrepancy function based on the difference between the high and low fidelity models.
2. *Multiplicative correction:*
$$\widehat{y}_{HF}(x) = \rho(x) y_{LF}(x),$$
with $\rho(x)$ a multiplicative correction constructed using the ratio between the high and low fidelity models.
3. *Comprehensive corrections:* two examples are when both additive and multiplicative corrections can be used as in
$$\widehat{y}_{HF}(x) = \rho(x) y_{LF}(x) + \delta(x),$$
or a hybrid version of both of them
$$\widehat{y}_{HF}(x) = w(x)\rho(x) y_{LF}(x) + (1 - w(x))(y_{LF}(x) + \delta(x)),$$
where $w(x)$ is a weight function.

To illustrate one way of obtaining the corrections, we will use the additive and multiplicative cases. In these cases, the difference $y_{HF} - y_{LF}$ and ratio $y_{HF}/y_{LF}$ on sampling points are used to obtain the corrections $\delta(x)$ and $\rho(x)$, respectively. We suppose that we have $M$ data points for the low-fidelity model and $m$ for the high-fidelity, with $M \gg m$. First, we observe if the low-fidelity model is cheap enough to generate response output at other necessary locations. If it isn't, we need to build a surrogate, such as a Gaussian process model, to replace it. After, a surrogate is constructed for the difference or ratio of high and low fidelity models on the $m$ common points. Last, we can either use the surrogate for low-fidelity points and create the multi-fidelity model by adding (for the difference) or multiplying (for the ratio) the two surrogate models or use the surrogate for the difference or ratio to approximate the discrepancy at the $M - m$ points where only the low-fidelity model is available and fit a surrogate model for the $m$ high-fidelity data and $M - m$ approximated data.

To summarize, the basic idea behind multi-fidelity models is that high-fidelity data is used to establish accuracy and convergence, while the low-fidelity are used for speedup. More details and a survey on multi-fidelity models can be found in [Peherstorfer et al. '18] and [Fernández-Godino et al. '16].

## 3.2 A first autoregressive model

The work of Kennedy and O'Hagan in [Kennedy & O'Hagan '98] concerns the use of an autoregressive model based on Gaussian processes for combining data from deterministic simulations of different accuracies in order to infer about the most accurate and reliable code and perform uncertainty analysis, that is, it is a multi-fidelity (surrogate) model. The following assumptions are made:

(1) Different levels of code are correlated in some way.

(2) The codes have a degree of smoothness: the output values for similar inputs are close. Individual runs of rough codes do not provide information outside of a very small neighborhood.

(3) Prior beliefs of each level of the code can be modeled using Gaussian processes.

(4) The outputs of each level are scalars.

We suppose that we have $s$ levels of code $\{z_t(x)\}_{t=1,\dots,s}$ sorted by increasing order of fidelity and modeled by Gaussian processes $\{Z_t(x)\}_{t=1,\dots,s}$, with $x \in D \subseteq \mathbb{R}^n$, thus considering the $z_s(x)$ being the most accurate and costly code. This mean that we consider each level $z_t(x)$, with $x \in D$, as a realization of the random process $Z(x)$. The object of inference is $Z_s(x)$, the highest-fidelity level, conditioned on all outputs of all levels of code that we have.

Furthermore, consider the following assumption about two levels $Z_t(x)$ and $Z_{t-1}(x)$:

$$\mathrm{Cov}\{Z_t(x), Z_{t-1}(x')|Z_{t-1}(x)\} = 0 \quad \text{for all} \quad x' \neq x. \tag{3.1}$$

It translates as a kind of Markov property: given the nearest point to $Z_t(x)$ at the level $t-1$, which is $Z_{t-1}(x)$, we learn nothing more about $Z_t(x)$ from any other point $Z_{t-1}(x')$, for $x' \neq x$.

In [O'Hagan '98], it is proved that this Markov property implies the following model. Consider for $t = 2, \dots, s$:

$$\begin{cases} Z_t(x) = \rho_{t-1}(x)Z_{t-1}(x) + \delta_t(x) \\ Z_{t-1}(x) \perp \delta_t(x) \\ \rho_{t-1}(x) = g_{t-1}^T(x)\beta_{\rho_{t-1}} \end{cases}, \tag{3.2}$$

where

$$\delta_t(x) \sim \mathcal{GP}(f_t^T(x)\beta_t, \sigma_t^2 r_t(x, x')), \tag{3.3}$$

and

$$Z_1(x) \sim \mathcal{GP}(f_1^T(x)\beta_1, \sigma_1^2 r_1(x, x')). \tag{3.4}$$

Also, $g_{t-1}(x)$ is a vector of $q_{t-1}$ regression functions, $f_t(x)$ is a vector of $p_t$ regression functions, $r_t(x, x')$ is a correlation function ($r_t(x, x') \in [-1, 1]$ for all $x, x' \in D$ and $\sigma_t^2 r_t(x, x')$ is a valid covariance function), $\beta_t$ is a $p_t$-dimensional parameter vector, $\beta_{\rho_{t-1}}$ is a $q_{t-1}$-dimensional parameter vector, and $\sigma_t^2$ is a positive real number. We denote $\sigma^2 = (\sigma_1^2, \dots, \sigma_s^2)$, $\beta = (\beta_1^T, \dots, \beta_s^T)^T$ and $\beta_\rho = (\beta_{\rho_1}^T, \dots, \beta_{\rho_{s-1}}^T)^T$. In this way, we write the expected value of $Z_t(x)$ as

$$\mathbb{E}[Z_t(x)|\sigma^2, \beta, \beta_\rho] = \mathbb{E}[\rho_{t-1}(x)Z_{t-1}(x) + \delta_t(x)|\sigma^2, \beta, \beta_\rho] =$$

$$\rho_{t-1}(x)\mathbb{E}[Z_{t-1}(x)|\sigma^2,\beta,\beta_\rho] + f_t^T(x)\beta_t = \cdots = \sum_{i=1}^{t}\left(\prod_{j=1}^{t-1}\rho_j(x)\right)f_i^T(x)\beta_i = h_t(x)^T\beta,$$

where

$$h_t(x)^T = \left(\left(\prod_{i=1}^{t-1}\rho_i(x)\right)f_1^T(x), \left(\prod_{i=2}^{t-1}\rho_i(x)\right)f_2^T(x), \ldots, \rho_{t-1}(x)f_{t-1}^T(x), f_t^T(x), 0, \ldots, 0\right),$$

with $dim(h_t(x)) = dim(\beta) = \sum_{i=1}^{s} p_i$, thus $h_t(x)^T$ having $\sum_{i=t+1}^{s} p_i$ zeros at its right end.

The covariance of the process $Z_t(x)$ at two different points is given by

$$\begin{gathered}\mathrm{Cov}\{Z_t(x), Z_t(x')|\sigma^2,\beta,\beta_\rho\} = \\ \mathrm{Cov}\{\rho_{t-1}(x)Z_{t-1}(x) + \delta_t(x), \rho_{t-1}(x')Z_{t-1}(x') + \delta_t(x')|\sigma^2,\beta,\beta_\rho\} = \\ \rho_{t-1}(x)\rho_{t-1}(x')\mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(x')|\sigma^2,\beta,\beta_\rho\} + \sigma_t^2 r_t(x,x') = \cdots = \\ \sum_{j=1}^{t}\sigma_j^2\left(\prod_{i=j}^{t-1}\rho_i(x)\rho_i(x')\right)r_j(x,x').\end{gathered} \tag{3.5}$$

We use the convention that the empty product is equal to 1.

Next, for different levels $t$ and $t'$, with $t > t'$, and different input points $x$ and $x'$, the covariance is

$$\begin{gathered}\mathrm{Cov}\{Z_t(x), Z_{t'}(x')|\sigma^2,\beta,\beta_\rho\} = \\ \mathrm{Cov}\{\rho_{t-1}(x)Z_{t-1}(x) + \delta_t(x), Z_{t'}(x')|\sigma^2,\beta,\beta_\rho\} = \\ \rho_{t-1}(x)\mathrm{Cov}\{Z_{t-1}(x), Z_{t'}(x')|\sigma^2,\beta,\beta_\rho\} = \cdots = \\ \left(\prod_{i=t'}^{t-1}\rho_i(x)\right)\mathrm{Cov}\{Z_{t'}(x), Z_{t'}(x')\}.\end{gathered} \tag{3.6}$$

Let us now consider $\mathcal{Z}_t$, the Gaussian vector containing the values of $Z_t(x)$ evaluated at the points in $D_t = \{x_i^t\}_{i=1,\ldots,n_t}$ for $t = 1, \ldots, s$, and let $\mathcal{Z}^{(s)} = (\mathcal{Z}_1^T, \ldots, \mathcal{Z}_s^T)^T$ be the Gaussian vector containing the values of all processes $Z_t(x)$ at their respective points in $D_t$, and let us assume that $D_s \subseteq D_{s-1} \subseteq \cdots \subseteq D_1$. Namely, let

$$\mathcal{Z}^{(s)} = (Z_1(x_1^1), \ldots, Z_1(x_{n_1}^1), Z_2(x_1^2), \ldots, Z_{s-1}(x_{n_{s-1}}^{s-1}), Z_s(x_1^s), \ldots, Z_s(x_{n_s}^s)).$$

With the $h_t(\cdot)$ vectors, it is easy to construct the mean of $\mathcal{Z}^{(s)}$, which is given by $H_s\beta$ with $H_s$ being the matrix constructed by stacking $h_1(\cdot)^T$ evaluated at the points in $D_1$,

followed by the values of $h_2(\cdot)^T$ evaluated at the points in $D_2$, and so forth,

$$H_s = \begin{bmatrix} [— & h_1(D_1) & —] \\ [— & h_2(D_2) & —] \\ & \vdots & \\ [— & h_{s-1}(D_{s-1}) & —] \\ [— & h_s(D_s) & —] \end{bmatrix} = \begin{bmatrix} — & h_1(x_1^1) & — \\ & \vdots & \\ — & h_1(x_{n_1}^1) & — \\ — & h_2(x_1^2) & — \\ & \vdots & \\ — & h_{s-1}(x_{n_{s-1}}^{s-1}) & — \\ — & h_s(x_1^s) & — \\ & \vdots & \\ — & h_s(x_{n_s}^s) & — \end{bmatrix}. \tag{3.7}$$

The covariances are all conditioned by the values of the same hyperparameters, therefore, for simplicity, the dependencies on the vectors $\sigma^2, \beta$ and $\beta_\rho$ are left implicit. Now, we can construct the vector $k_s(x)$ of covariances between $Z_s(x)$ and $\mathcal{Z}^{(s)}$

$$k_s^T(x) = (c_1^T(x, D_1), \ldots, c_s^T(x, D_s))^T, \tag{3.8}$$

with $c_t^T(x, D_t) = \mathrm{Cov}\{Z_s(x), Z_t(D_t)\} = (\mathrm{Cov}\{Z_s(x), Z_t(x_1^t)\}, \ldots, \mathrm{Cov}\{Z_s(x), Z_t(x_{n_t}^t)\})$. Using (3.5) and (3.6), the expression of $c_t^T(x, D_t)$ can be rewritten as

$$c_t^T(x, D_t) = \left(\prod_{i=t}^{s-1} \rho_i(x)\right) \mathrm{Cov}\{Z_t(x), Z_t(D_t)\} =$$

$$\left(\prod_{i=t}^{s-1} \rho_i(x)\right) \left(\rho_{t-1}(x)\rho_{t-1}(D_t) \odot \mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\} + \sigma_t^2 r_t(x, D_t)\right) = \tag{3.9}$$

$$\rho_{t-1}(D_t) \odot c_{t-1}^T(x, D_t) + \left(\prod_{i=t}^{s-1} \rho_i(x)\right) \sigma_t^2 r_t(x, D_t),$$

where $\odot$ represents the element by element matrix (or vector) product,

$$c_i^T(x, D_t) = \mathrm{Cov}\{Z_s(x), Z_i(D_t)\} \text{ for } i \leq t,$$
$$r_t^T(x, D_t) = (r_t(x, x_1^t), \ldots, r_t(x, x_{n_t}^t)),$$

and

$$c_1^T(x, D_t) = \left(\prod_{i=1}^{s-1} \rho_i(x)\right) \mathrm{Cov}\{Z_1(x), Z_1(D_t)\} = \left(\prod_{i=1}^{s-1} \rho_i(x)\right) \sigma_1^2 r_1(x, D_t).$$

The covariance matrix $V_s$ of $\mathcal{Z}^{(s)}$, can also be contructed using (3.5) and (3.6):

$$V_s = \mathrm{Cov}\{\mathcal{Z}^{(s)}, \mathcal{Z}^{(s)}\} = \begin{bmatrix} V_{1,1} & \ldots & V_{1,s} \\ \vdots & \ddots & \vdots \\ V_{s,1} & \ldots & V_{s,s} \end{bmatrix}, \tag{3.10}$$

with diagonal elements

$$V_{t,t} = \mathrm{Cov}\{\mathcal{Z}_t, \mathcal{Z}_t\} = \sigma_t^2 R_t + \sum_{j=1}^{t-1} \sigma_j^2 \left(\prod_{i=j}^{t-1} \rho_i(D_t)\rho_i^T(D_t)\right) \odot R_j,$$

for $t = 1, \ldots, s$, where $R_j = [r_j(x, x')]_{x,x' \in D_j}$, and off-diagonal entries given by

$$\begin{aligned} V_{t',t} = \mathrm{Cov}\{\mathcal{Z}_{t'}, \mathcal{Z}_t\} = \mathrm{Cov}\{Z_{t'}(D_{t'}), \rho_{t-1}(D_t) \odot Z_{t-1}(D_t) + \delta_t(D_t)\} = \\ (\mathbf{1}_{n_{t'}} \rho_{t-1}^T(D_t)) \odot \mathrm{Cov}\{Z_{t'}(D_{t'}), Z_{t-1}(D_t)\} = \cdots = \\ \left( \bigodot_{i=t'}^{t-1} \mathbf{1}_{n_{t'}} \rho_i^T(D_t) \right) \odot \mathrm{Cov}\{Z_{t'}(D_{t'}), Z_{t'}(D_t)\}, \end{aligned} \tag{3.11}$$

for $1 \leq t' < t \leq s$, and $V_{t',t}^T$ otherwise. Here, for $t > t'$, we denote by $V_{t',t'}(D_t, D_{t'})$ the submatrix of $V_{t',t'}$ with entries corresponding to the points in $D_t \subseteq D_{t'}$ in the rows and points in $D_{t'}$ in the columns.

Last, let $v_{Z_s}^2(x)$ denote the variance of $Z_s(x)$. By Equation (3.5), this variance is

$$v_{Z_s}^2(x) = \mathrm{Var}[Z_s(x)|\sigma^2, \beta, \beta_\rho] = \sigma_s^2 + \sum_{i=1}^{s-1} \sigma_i^2 \left( \prod_{j=i}^{s-1} \rho_j(x)^2 \right) = \sum_{i=1}^{s} \sigma_i^2 \left( \prod_{j=i}^{s-1} \rho_j(x)^2 \right).$$

Thus, the joint distribution of $Z_s(x)$ and $\mathcal{Z}^{(s)}$, given $\sigma^2, \beta, \beta_\rho$, is the following multivariate normal:

$$\begin{bmatrix} Z_s(x) \\ \mathcal{Z}^{(s)} \end{bmatrix} \Big| \sigma^2, \beta, \beta_\rho \Bigg] \sim \mathcal{N}\left( \begin{bmatrix} h_s(x)^T \beta \\ H_s \beta \end{bmatrix}, \begin{bmatrix} v_{Z_s}^2(x) & k_s^T(x) \\ k_s(x) & V_s \end{bmatrix} \right). \tag{3.12}$$

By the predictive identities for Gaussian processes (2.8) and (2.9), it is straightforward that, when observing $\mathcal{Z}^{(s)} = z^{(s)}$

$$Z_s(x)|\mathcal{Z}^{(s)} = z^{(s)}, \sigma^2, \beta, \beta_\rho \sim \mathcal{N}(m_{Z_s}(x), s_{Z_s}^2(x)), \tag{3.13}$$

with

$$m_{Z_s}(x) = h_s^T(x)\beta + k_s^T(x) V_s^{-1}(z^{(s)} - H_s \beta), \tag{3.14}$$

and

$$s_{Z_s}^2(x) = v_{Z_s}^2(x) - k_s^T(x) V_s^{-1} k_s(x). \tag{3.15}$$

Note that, since

$$k_1(x)^T V_1^{-1} = \mathrm{Cov}\{Z_1(x), Z_1(D_1)\} \frac{R_1^{-1}}{\sigma_1^2} = \sigma_1^2 r_1(x, D_1) \frac{R_1^{-1}}{\sigma_1^2} = r_1(x, D_1) R_1^{-1}$$

does not depend on $\sigma_1^2$, by Proposition A.2, we have that $k_s^T(x) V_s^{-1}$ is independent of $\sigma_t^2$, for $t = 1, \ldots, s$, and, therefore, the predictive mean $m_{Z_s}(x)$ does not depend on the variance hyperparameters of any level.

## 3.3 The recursive autoregressive model

The work in [Le Gratiet '13] and [Le Gratiet & Garnier '14] is an extension and improvement of the autoregressive model of Kennedy and O'Hagan in [Kennedy & O'Hagan '98] previously presented. In his work, Le Gratiet discusses a new way of performing the co-kriging (this expression arises from the idea that multiple

correlated kriging procedures are performed) with the aim of reducing the computational complexity by breaking the $s$-level co-kriging into $s$ independent Gaussian processes.

In this new model, for $t = 2, \ldots, s$, let

$$\begin{cases} Z_t(x) = \rho_{t-1}(x)\widetilde{Z}_{t-1}(x) + \delta_t(x) \\ \widetilde{Z}_{t-1}(x) \perp \delta_t(x) \\ \rho_{t-1}(x) = g_{t-1}^T(x)\beta_{\rho_{t-1}} \end{cases}, \tag{3.16}$$

where $\widetilde{Z}_{t-1}(x)$ is a Gaussian process with the distribution of

$$Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_{t-1}^2, \beta_{t-1}, \beta_{\rho_{t-2}},$$

$\delta(x)$ is a Gaussian process with distribution (3.3), and the experimental design sets have the nested property

$$D_s \subseteq D_{s-1} \subseteq \ldots, \subseteq D_1.$$

The only difference from the classical autoregressive multi-fidelity model (3.2) is that, instead of expressing $Z_t$ as a function of $Z_{t-1}$, we first condition $Z_{t-1}$ by the realization of $Z_i$ at the points in $D_i$, for $i = 1, \ldots, t-1$, that is, $\mathcal{Z}^{(t-1)}$ is equal to the values $z^{(t-1)} = (z_1, \ldots, z_{t-1})$.

Since the joint distribution of $Z_{t-1}(x)$ and $\mathcal{Z}^{(t-1)}$ conditioned on $\sigma_{t-1}^2, \beta_{t-1}, \beta_{\rho_{t-2}}$ is Gaussian, for $t = 2, \ldots, s$, so will be the distribution of

$$[\widetilde{Z}_{t-1}(x) = Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_{t-1}^2, \beta_{t-1}, \beta_{\rho_{t-2}}],$$

whose mean and variance we will denote by $\mu_{Z_{t-1}}(x)$ and $\sigma_{Z_{t-1}}^2(x)$. By Equation (3.16), we have that

$$\begin{aligned}[Z_t(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}}] &= \rho_{t-1}(x)\widetilde{Z}_{t-1}(x) + \delta_t(x) \\ &\sim \mathcal{N}(\rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T\beta_t, \rho_{t-1}^2(x)\sigma_{Z_{t-1}}^2(x) + \sigma_t^2(x)),\end{aligned}$$

since $r_t(x, x) = 1 \; \forall x$, given that it is a correlation function. This way, the joint distribution of $Z_t(x)$ and $\mathcal{Z}_t$ conditioned by $\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_t^2, \beta_t$ and $\beta_{\rho_{t-1}}$ is

$$\begin{aligned}&\begin{bmatrix} Z_t(x) \\ \mathcal{Z}_t \end{bmatrix} \Big| \mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}} \Bigg] \sim \\ &\mathcal{N}\left( \begin{bmatrix} \rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t \\ \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) + F_t\beta_t \end{bmatrix}, \begin{bmatrix} \rho_{t-1}^2(x)\sigma_{Z_{t-1}}^2(x) + \sigma_t^2(x) & r_t^T(x) \\ r_t(x) & R_t \end{bmatrix} \right).\end{aligned} \tag{3.17}$$

For simplicity, we use $R_t = [r_t(x, x')]_{x,x' \in D_t}$ for the correlation matrix of the Gaussian process $\delta(x)$ at the points in $D_t$, $r_t^T(x)$ for the correlation vector $r_t^T(x) = (r_t(x, x'))_{x' \in D_t}$, $\rho_{t-1}(D_t)$ for the vector containing the values $\rho_{t-1}(x)$ for $x \in D_t$, and $F_t$ the experience matrix containing the values of $f_t^T(x)$ on $D_t$ as rows. In other words,

$$R_t = [r_t(x, x')]_{x,x' \in D_t} = \begin{bmatrix} r_t(x_1^t, x_1^t) & \ldots & r_t(x_1^t, x_{n_t}^t) \\ \vdots & \ddots & \vdots \\ r_t(x_{n_t}^t, x_1^t) & \ldots & r_t(x_{n_t}^t, x_{n_t}^t) \end{bmatrix},$$

$$r_t^T(x) = (r_t(x, x_1^t), \ldots, r_t(x, x_{n_t}^t)),$$
$$\rho_{t-1}^T(D_t) = (\rho_{t-1}(x_1), \ldots, \rho_{t-1}(x_{n_t}^t)),$$

and

$$F_t = \begin{bmatrix} — & f_t^T(x_1^t) & — \\ & \vdots & \\ — & f_t^T(x_{n_t}^t) & — \end{bmatrix}.$$

Observe that Equation (3.17) shows that the process $Z_t$ conditioned by $\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_t^2, \beta_t$, and $\beta_{\rho_{t-1}}$ is Gaussian. Therefore, using again the Gaussian process predictive equations, Equations (2.8) and (2.9), for further conditioning $Z_t(x)$ by $\mathcal{Z}_t = z_t$, we obtain the expressions for $\mu_{Z_t}(x)$ and $\sigma_{Z_t}^2(x)$ of the distribution of

$$\widetilde{Z}_t(x) = [Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}}] \sim \mathcal{N}(\mu_{Z_t}(x), \sigma_{Z_t}^2(x)). \tag{3.18}$$

Thus, these functions are given by

$$\mu_{Z_t}(x) = \rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) - F_t\beta_t), \tag{3.19}$$

and

$$\sigma_{Z_t}^2(x) = \rho_{t-1}^2(x)\sigma_{Z_{t-1}}^2(x) + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t(x)), \tag{3.20}$$

These last predictive equations are referent to the *simple co-kriging model (SK)*, when we consider fixed values for the hyperparameters. Both the the predictive mean and variance at the level $t$ are expressed as functions of the predictive mean and variance at the level $t-1$, respectively. Furthermore, as in basic Gaussian process regression when using covariance kernels of the form $k(x, x') = \sigma^2 r(x, x')$, the predictive mean does not depend on the variance parameters $\{\sigma_t^2\}_{t=1,\ldots,s}$, and the variance does not depend on any of the observed values $z^{(t)}$.

Note that, similarly, for $t = 1$,

$$\begin{bmatrix} Z_1(x) \\ \mathcal{Z}_1 \end{bmatrix} \Big| \sigma_1^2, \beta_1 \sim \mathcal{N}\left( \begin{bmatrix} f_1^T\beta_1 \\ F_1\beta_1 \end{bmatrix}, \begin{bmatrix} \sigma_1^2(x) & r_1^T(x) \\ r_1(x) & R_1 \end{bmatrix} \right)$$
$$\implies Z_1(x)|\mathcal{Z}^{(1)} = z^{(1)}, \sigma_1^2, \beta_1 \sim \mathcal{N}(\mu_{Z_1}(x), \sigma_{Z_1}^2(x)),$$

with

$$\begin{cases} \mu_{Z_1}(x) = f_1^T(x)\beta_1 + r_1^T(x)R_1^{-1}(z_1 - F_1\beta_1) \\ \sigma_{Z_1}^2(x) = \sigma_1^2(1 - r_1^T(x)R_1^{-1}r_1(x)). \end{cases}$$

**Remark 1.** *For the recursive model above, it is true that for, $t = 1, \ldots, s$,*

$$\mu_{Z_t}(D_t) = z_t,$$

*where $z_t = z_t(D_t)$ is the vector containing the observed values of $Z_t(x)$ at the points in $D_t$.*

*Proof.* For $t = 1$,

$$\mu_{Z_1}(x) = f_1^T(x)\beta_1 + r_1^T(x)R_1^{-1}(z_1 - F_1\beta_1)$$

$$\implies \mu_{Z_1}(D_1) = F_1\beta_1 + R_1R_1^{-1}(z_1 - F_1\beta_1) = z_1.$$

Equivalently, for $t \geq 1$, we know that

$$\mu_{Z_t}(x) = \rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) - F_t\beta_t)$$
$$\implies \mu_{Z_t}(D_t) = \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) + F_t\beta_t + R_tR_t^{-1}(z_t - \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) - F_t\beta_t) =$$
$$= z_t.$$

□

This, in addition to the nested property of the sets $D_t$, gives

$$\mu_{Z_{t-1}}(D_t) = z_{t-1}(D_t),$$

for $t = 2, \ldots, s$, which can be replaced in equation (3.19) to obtain

$$\mu_{Z_t}(x) = \rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t) \odot z_{t-1}(D_t) - F_t\beta_t). \quad (3.21)$$

**Remark 2.** *In the same conditions of the previous remark,*

$$\sigma^2_{Z_t}(x_i^t) = 0 \quad \forall x_i^t \in D_t.$$

*Proof.* Observe that the $i$-th column of $R_t$ is equal to $r_t(x_i^t)$. Therefore, using the identity given by Equation (A.10), we obtain

$$R_t^{-1}r_t(x_i^t) = \begin{bmatrix} \mathbf{0}_{(i-1)\times 1} \\ 1 \\ \mathbf{0}_{(n_t-i)\times 1} \end{bmatrix}$$
$$\implies r_t^T(x_i^t)R_t^{-1}r_t(x_i^t) = r_t(x_i^t, x_i^t) = 1.$$

Substituting this in the expression for $\sigma^2_{Z_t}(x_i^t)$ and using the recursion of this expression combined to the nested property of the sets gives us the desired relation. □

**Remark 3.** *The use of the nested property, $D_s \subseteq D_{s-1} \subseteq \cdots \subseteq D_1$, can be relaxed. This extension is found in Appendix B of the thesis of Loic Le Gratiet, [Le Gratiet '13].*

Despite the different formulation of the the classical autoregressive model (3.2) and the recursive autoregressive model (3.16), both of them have, in fact, the same predictive equations. This result is stated in the following proposition:

**Proposition 3.1** (Proposition 1 of [Le Gratiet & Garnier '14])**.** *Let us consider $s$ Gaussian processes $\{Z_t(x)\}_{t=1,\ldots,s}$, and $\mathcal{Z}^{(s)} = (\mathcal{Z}_t)_{t=1,\ldots,s}$ the Gaussian vector containing the values of $\{Z_t(x)\}_{t=1,\ldots,s}$ at points in $\{D_t\}_{t=1,\ldots,s}$, with $D_s \subseteq D_{s-1} \subseteq \cdots \subseteq D_1$. If we consider the mean (3.14), and the variance (3.15) induced by the model (3.2) when we condition the Gaussian process $Z_s(x)$ by the observed values $z^{(s)}$ of $\mathcal{Z}^{(s)}$, and parameters $\beta, \beta_\rho$ and $\sigma^2$, and the mean (3.19) and variance (3.20) induced by the model (3.16) when we condition $Z_s(s)$ by $z^{(s)}$ and parameters $\beta, \beta_\rho$ and $\sigma^2$, then, we have:*

$$\mu_{Z_s}(x) = m_{Z_s}(x),$$

*and*

$$\sigma^2_{Z_s}(x) = s^2_{Z_s}(x).$$

*Proof.* Throughout this proof, we will use the nested property of the sets, specifically that $D_t \subseteq D_{t-1}$, and a particular ordering of the points in each of these sets, that is $D_t = (D_{t-1} \backslash D_t, D_t)$.

*For the mean:* By Equation (3.14), we know that for the classical model,

$$m_{Z_s}(x) = h_s^T(x)\beta + k_s^T(x)V_s^{-1}(z^{(s)} - H_s\beta).$$

Then, for a $t$-level model with $t = 2, \ldots, s$, we have

$$m_{Z_t}(x) = h_t^T(x)\beta^{(t)} + k_t^T(x)V_t^{-1}(z^{(t)} - H_t\beta^{(t)}), \tag{3.22}$$

where $\beta^{(t)} = (\beta_1^T, \ldots, \beta_t^T)^T$, $z^{(t)} = (z_1^T, \ldots, z_t^T)^T$, and

$$h_t^T(x) = \left(\left(\prod_{i=1}^{t-1}\rho_i(x)\right)f_1^T(x), \left(\prod_{i=2}^{t-1}\rho_i(x)\right)f_2^T(x), \ldots, \rho_{t-1}(x)f_{t-1}^T(x), f_t^T(x)\right)$$

$$\implies h_t^T(x) = (\rho_{t-1}(x)h_{t-1}^T(x), \quad f_t^T(x)).$$

This way,

$$h_t^T(x)\beta^{(t)} = \sum_{i=1}^{t}\left(\prod_{j=i}^{t-1}\rho_j(x)\right)f_i^T(x)\beta_i.$$

For $t = 1, \ldots, s$, $H_t$ can be constructed similarly to the $H_s$ matrix of equation (3.7), but it is simpler to observe that $H_t$ is a submatrix of $H_s$ containing its first $\sum_{i=1}^{t} n_i$ rows and its first $\sum_{i=1}^{t} p_i$ columns. If $t > 1$, we can use the same idea to write $H_t$ as

$$H_t = \begin{bmatrix} H_{t-1} & 0 \\ A & F_t(D_t) \end{bmatrix},$$

where $A$ is the submatrix of $H_t$ containing its last $n_t$ rows and its first $\sum_{i=1}^{t-1} p_i$ columns:

$$A = [\rho_{t-1}(D_t)\mathbf{1}^T_{\sum_{i=1}^{t-1} p_i}] \odot h_{t-1}(D_t),$$

where

$$h_{t-1}(D_t) = \begin{bmatrix} h_{t-1}(x_1^t) \\ \vdots \\ h_{t-1}(x_{n_t}^t) \end{bmatrix}.$$

By Proposition $A.2$ found in the Appendix,

$$k_t^T(x)V_t^{-1} = \big(\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1} - (0, \quad [\rho_{t-1}^T(D_t) \odot r_t^T(x)]R_t^{-1}), \quad r_t^T(x)R_t^{-1}\big).$$

This equality implies that

$$k_t^T(x)V_t^{-1}z^{(t)} = k_t^T(x)V_t^{-1}\begin{bmatrix} z^{(t-1)} \\ z_t \end{bmatrix} =$$

$$\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1}z^{(t-1)} - [\rho_{t-1}^T(D_t) \odot r_t^T(x)]R_t^{-1}z_{t-1}(D_t) + r_t^T(x)R_t^{-1}z_t.$$

Using again Proposition $A.2$, and the expression we obtained for $H_t$, we get

$$k_t^T(x)V_t^{-1}H_t\beta^{(t)} = \rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1}H_{t-1}\beta^{(t-1)} -$$

$$[\rho_{t-1}^T(D_t) \odot r^T(x)]R_t^{-1}h_{t-1}(D_t)\beta^{(t-1)}+$$
$$r^T(x)R_t^{-1}\big([\rho_{t-1}(D_t)\mathbf{1}^T_{\sum_{i=1}^{t-1}p_i}] \odot h_{t-1}(D_t)\big)\beta^{(t-1)} + r^T(x)R_t^{-1}F_t\beta_t =$$
$$\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1}H_{t-1}\beta^{(t-1)} + r^T(x)R_t^{-1}F_t\beta_t,$$

since the two middle terms cancel each other. Therefore, using the obtained expressions for $h_t^T(x)$, $k_t^T(x)V_t^{-1}z^{(s)}$ and $k_t^T(x)V_t^{-1}H_t\beta^{(t)}$ in Equation (3.22),

$$m_{Z_t}(x) = \rho_{t-1}(x)h_{t-1}(x)\beta^{(t-1)} + f_t^T(x)\beta_t+$$
$$\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1}z^{(t-1)} - [\rho_{t-1}^T(D_t) \odot r_t^T(x)]R_t^{-1}z_{t-1}(D_t) + r_t^T(x)R_t^{-1}z_t-$$
$$\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1}H_{t-1}\beta^{(t-1)} - r^T(x)R_t^{-1}F_t\beta_t =$$
$$\rho_{t-1}(x)m_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}\Big(z_t - \rho_{t-1}(D_t) \odot z_{t-1}(D_t) - F_t\beta_t\Big).$$

From the last line of the previous equation, we notice that both $m_{Zt}(x)$ and $\mu_{Zt}(x)$ follow the exact same recursive relations. This, added to the fact that $\mu_{Z_1}(x) = m_{Z_1}(x) = f_1^T(x)\beta^1$, gives us the desired equality

$$\mu_{Zs}(x) = m_{Zs}(x).$$

*For the variance:* We follow similar steps as before.

For the $t$-level classical co-kriging model, equation (3.15) states that

$$s^2_{Z_t}(x) = v^2_{Z_t}(x) - k_t^T(x)V_t^{-1}k_t(x). \tag{3.23}$$

For the variance term $v^2_{Z_t}(x)$, we use equation (3.5), to obtain

$$v^2_{Z_t}(x) = \mathrm{Var}[Z_t(x)] = \rho^2_{t-1}(x)\mathrm{Var}[Z_{t-1}(x)] + \sigma_t^2 = \rho^2_{t-1}(x)v^2_{Z_{t-1}}(x) + \sigma_t^2. \tag{3.24}$$

For the $k_t^T(x)V_t^{-1}k_t(x)$ term, we know from Proposition A.2 of the Appendix chapter that

$$k_t^T(x)V_t^{-1} = \big[\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1} - [0, \quad [\rho_{t-1}^T(D_t) \odot r_t^T(x)]R_t^{-1}], \quad r_t^T(x)R_t^{-1}\big],$$

and, by Equations (A.22) and (3.5), it is clear that

$$k_t^T(x) = (\rho_{t-1}(x)k_{t-1}^T(x), \quad \mathrm{Cov}\{Z_t(x), \mathcal{Z}_t\}) =$$
$$(\rho_{t-1}(x)k_{t-1}^T(x), \quad \rho_{t-1}(x)\rho_{t-1}^T(D_t) \odot \mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\} + \sigma_t^2 r_t^T(x)).$$

These last two equalities, in turn, imply that

$$k_t^T(x)V_t^{-1}k_t(x) =$$
$$\big[\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1} - [0, \quad [\rho_{t-1}^T(D_t) \odot r_t^T(x)]R_t^{-1}], \quad r_t^T(x)R_t^{-1}\big] \times$$
$$\begin{bmatrix} \rho_{t-1}(x)k_{t-1}(x) \\ \rho_{t-1}(x)\rho_{t-1}(D_t) \odot \mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\}^T + \sigma_t^2 r_t(x) \end{bmatrix}.$$

Note that, because of the ordering of the points in $D_{t-1}$, the last $n_t$ terms of $k_{t-1}^T(x)$ are exactly $\mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\}$. For this reason,

$$k_t^T(x)V_t^{-1}k_t(x) = \rho^2_{t-1}(x)k_{t-1}^T V_{t-1}^{-1}k_{t-1}(x)-$$

$$[\rho_{t-1}^T(D_t) \odot r_t^T(x)]R_t^{-1}\rho_{t-1}(x)\mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\}^T +$$
$$r_t^T(x)R_t^{-1}\big(\rho_{t-1}(x)\rho_{t-1}(D_s) \odot \mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\}^T + \sigma_t^2 r_t^T(x)\big) =$$
$$\rho_{t-1}^2(x)k_{t-1}^T V_{t-1}^{-1}k_{t-1}(x) + \sigma_t^2 r_t^T(x)R_t^{-1}r_t(x)$$

Using this result, together with Equation (3.24), in Equation (3.23), gives us

$$s_{Z_t}^2(x) = \rho_{t-1}^2(x)(v_{Z_{t-1}}^2(x) - k_{t-1}^T V_{t-1}^{-1}k_{t-1}(x)) + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t(x)) =$$
$$\rho_{t-1}^2(x)s_{Z_{t-1}}^2(x) + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t(x)).$$

This is the same recursive relation that $\sigma_{Z_t}^2(x)$ satisfies. Noting that $\sigma_{Z_1}^2 = s_{Z_1}^2(x)$, we obtain

$$\sigma_{Z_s}^2(x) = s_{Z_s}^2(x).$$

An analogous argument proves the equivalence for predictive covariances, see [Le Gratiet '13]. □

Therefore, we proved that both the classical autoregressive model (3.2) and the recursive autoregressive model have the same predictive Gaussian distribution for $Z_s(x)$, and, while the computational cost of the model (3.2) proposed in [Kennedy & O'Hagan '98] is dominated by the inversion of the matrix $V_s$ of size $\sum_{i=1}^s n_i \times \sum_{i=1}^s n_i$, the recursive model proposed in [Le Gratiet & Garnier '14] is built on $s$ independent krigings, each having its computational cost dominated by the inversion of the $R_t$ matrix of size $n_t \times n_t$ for $t = 1, \ldots, s$, which results in a lower computational cost. Besides that, the memory cost is also lower for this model, since it requires storing the $s$ matrices $\{R_t\}_{t=1,\ldots,s}$ instead of the matrix $V_s$ for the classical approach.

### 3.3.1 Bayesian parameter estimation

The parameter vectors $\beta$, $\beta_\rho$, and $\sigma^2$ of the recursive autoregressive model may be determined using methods such as maximum likelihood or Bayesian estimation. Given the recursive formulation, $(\beta_t, \beta_{\rho_t}, \sigma_t^2)$, for $t = 2 \ldots, s$, and $(\beta_1, \sigma_1^2)$ can be estimated separately. For the Bayesian approach, a smart choice of prior distributions gives us closed form expressions for the posterior distributions. We consider two such choices:

(i) all priors are informative

(ii) all priors are non-informative.

Case (ii): we consider the Jeffreys priors

$$p(\beta_1|\sigma_1^2) \propto 1, \quad p(\sigma_1^2) \propto \frac{1}{\sigma_1^2}, \quad p(\beta_{\rho_{t-1}}, \beta_t|z^{(t-1)}, \sigma_t^2) \propto 1, \quad p(\sigma_t^2|z^{(t-1)}) \propto \frac{1}{\sigma_t^2}. \tag{3.25}$$

Case (i): all prior means and variances can be prescribed by using the following priors

$$[\beta_1|\sigma_1^2] \sim \mathcal{N}_{p_1}(b_1, \sigma_1^2 W_1)$$
$$[\beta_{\rho_{t-1}}, \beta_t|z^{(t-1)}, \sigma_t^2] \sim \mathcal{N}_{q_{t-1}+p_t}\left(b_t = \begin{bmatrix} b_{t-1}^\rho \\ b_t^\beta \end{bmatrix}, \sigma_t^2 W_t = \sigma_t^2 \begin{bmatrix} W_{t-1}^\rho & 0 \\ 0 & W_t^\beta \end{bmatrix}\right) \tag{3.26}$$

$$[\sigma_1^2] \sim \mathcal{IG}(\alpha_1, \gamma_1), \qquad [\sigma_t^2 | z^{(t-1)}] \sim \mathcal{IG}(\alpha_t, \gamma_t),$$

where

$b_1$ is a vector of size $p_1$,

$b_{t-1}^{\rho}$ a vector of size $q_t - 1$,

$b_t^{\beta}$ is a vector of size $p_t$,

$W_1$ is a $p_1 \times p_1$ matrix,

$W_{t-1}^{\rho}$ a $q_{t-1} \times q_{t-1}$ matrix,

$W_t^{\beta}$ a $p_t \times p_t$ matrix,

and $\alpha_1, \gamma_1, \alpha_t, \gamma_t > 0$ parameters of inverse Gamma distributions.

The posterior distributions are obtained in Section $A.7$, and are given by

$$[\beta_1 | z_1, \sigma_1^2] \sim \mathcal{N}_{p_1}(\Sigma_1 \nu_1, \Sigma_1), \qquad [\beta_{\rho_{t-1}}, \beta_t | z^{(t)}, \sigma_t^2] \sim \mathcal{N}_{q_{t-1}+p_t}(\Sigma_t \nu_t, \Sigma_t), \tag{3.27}$$

where

$$\Sigma_t = \begin{cases} \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} \mathcal{H}_t + \frac{W_t^{-1}}{\sigma_t^2}\right]^{-1} & \text{(i)} \\ \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} \mathcal{H}_t\right]^{-1} & \text{(ii)} \end{cases}, \tag{3.28}$$

$$\nu_t = \begin{cases} \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} z_t + \frac{W_t^{-1}}{\sigma_t^2} b_t\right] & \text{(i)} \\ \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} z_t\right] & \text{(ii)} \end{cases}, \tag{3.29}$$

with $\mathcal{H}_1 = F_1$, and $\mathcal{H}_t = [G_{t-1} \odot \left(z_{t-1}(D_t)\mathbf{1}_{q_{t-1}}^T\right) \quad F_t]$, with $G_{t-1}$ being the experience matrix containing the values of $g_{t-1}(x)^T$ at the points in $D_t$ as rows:

$$G_{t-1} = g_{t-1}^T(D_t) = \begin{bmatrix} — & g_{t-1}^T(x_1^t) & — \\ & \vdots & \\ — & g_{t-1}^T(x_{n_t}^t) & — \end{bmatrix}.$$

Also, for $t \geq 1$,

$$[\sigma_t^2 | z^{(t)}] \sim \mathcal{IG}\Big(a_t, \frac{Q_t}{2}\Big), \tag{3.30}$$

with

$$Q_t = \begin{cases} \gamma_t + (b_t + \widehat{\lambda}_t)^T (W_t + [\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t]^{-1})^{-1} (b_t - \widehat{\lambda}_t) + \widehat{Q}_t & \text{(i)} \\ \widehat{Q}_t & \text{(ii)} \end{cases},$$

$$\widehat{Q}_t = (z_t + \mathcal{H}\widehat{\lambda}_t)^T R_t^{-1} (z_t - \mathcal{H}_t \widehat{\lambda}_t),$$

$$\widehat{\lambda}_t = [\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t]^{-1} \mathcal{H}_t^T R_t^{-1} z_t,$$

$$a_t = \begin{cases} \frac{n_t}{2} + \alpha_t & \text{(i)} \\ \frac{n_t - p_t - q_{t-1}}{2} & \text{(ii)} \end{cases},$$

and $q_0 = 0$.

Interestingly, there are some equivalences when using the non-informative case (ii) to maximum likelihood estimates. It is straightforward that the posterior mean of $(\beta_t, \beta_{\rho_t})$,

for $t = 2, \ldots, s$ and $\beta_1$, is the maximum likelihood estimator of these parameters given that the prior distribution is constant.

For the variance, in [Patterson & Thompson '71] the concept of restricted likelihood was introduced in order to reduce bias in estimates for variance components via maximum likelihood. We follow [Santner et al. '03] and [Harville '74] to obtain the restricted maximum likelihood estimate for $\sigma_t^2$.

First, we need to transform our vector $z_t$ by a matrix $C^T$ of size $n_t \times (n_t - p_t - q_{t-1})$ with rank $n_t - p_t - q_{t-1}$, such that the transformed vector $C^T z_t$ has mean equal to 0 (the particular choice of $C$ is not important, see [Harville '74]). The idea behind this is that the transformed vector will not depend on parameters other than $\sigma_t^2$, and this implies that there will not be an increase of bias due to the estimation of the parameters $\beta_t$ and $\beta_{\rho_{t-1}}$, therefore prior information of these parameters is ignored. For simplicity, let $\tilde{\beta}_t = \begin{bmatrix} \beta_{\rho_{t-1}} \\ \beta_t \end{bmatrix}$, for $t > 1$, and $\tilde{\beta}_1 = \beta_1$. Observe that $\mathcal{Z}_t$ conditioned by $\mathcal{Z}^{(t-1)} = z^{(t-1)}$, $\beta_t$, $\beta_{\rho_{t-1}}$, and $\sigma_t^2$ has distribution

$$\mathcal{Z}_t | z^{(t-1)}, \tilde{\beta}_t, \sigma_t^2 \sim \mathcal{N}(\mathcal{H}_t \tilde{\beta}_t, \sigma_t^2 R_t).$$

To clarify where this distribution comes from, return to the expression of $Z_t(x)$ in Equation (3.16), and the check Remarks 1 and 2.

Thus, a possible choice of $C$ is one such that $CC^T = I - \mathcal{H}_t(\mathcal{H}_t^T \mathcal{H}_t)^{-1}\mathcal{H}_t^T$ and $C^T C = I$. Then, we have that

$$C^T \mathcal{H}_t = (C^T C) C^T \mathcal{H}_t = C^T (I - \mathcal{H}_t (\mathcal{H}_t^T \mathcal{H}_t)^{-1} \mathcal{H}_t^T) \mathcal{H}_t = 0$$

$$\implies C^T \mathcal{H}_t \tilde{\beta}_t = 0 \quad \forall \tilde{\beta}_t.$$

Then, the likelihood of $\zeta_t = C^T z_t$ (we let the dependencies on $z^{(t-1)}$ implicit) is given by

$$\ell_{rest}(\zeta_t; \sigma_t^2) = \frac{1}{(2\pi)^{(n_t - p_t - q_{t-1})/2}} \frac{1}{\sqrt{\det(C^T(\sigma_t^2 R_t) C)}} \exp\left\{ -\frac{1}{2\sigma_t^2} \zeta_t^T (C^T R_t C)^{-1} \zeta_t \right\},$$

which can be rewritten as

$$\frac{1}{(2\pi)^{(n_t - p_t - q_{t-1})/2}} \frac{\sqrt{\det(\mathcal{H}_t^T \mathcal{H}_t)}}{\sqrt{\det(\sigma_t^2 R_t) \det(\mathcal{H}_t^T (\sigma_t^2 R_t)^{-1} \mathcal{H}_t)}} \exp\left\{ -\frac{1}{2\sigma_t^2} (z_t - \mathcal{H}_t \widehat{\lambda}_t)^T R_t^{-1} (z_t - \mathcal{H}_t \widehat{\lambda}_t) \right\},$$

with $\widehat{\lambda}_t = [\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t]^{-1} \mathcal{H}_t^T R_t^{-1} z_t$ being the maximum likelihood estimate of $\tilde{\beta}_t$ using the data $z^t$. This implies that the log-likelihood is

$$\begin{aligned} \log(\ell_{rest}(\zeta_t; \sigma_t^2)) = & -\frac{n_t - p_t - q_{t-1}}{2} \log(2\pi) + \frac{1}{2} \log(\det(\mathcal{H}_t^T \mathcal{H}_t)) - \frac{n_t - p_t - q_{t-1}}{2} \log(\sigma_t^2) - \\ & \frac{1}{2} \log(\det(R_t)) - \frac{1}{2} \log(\det(\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t)) - \frac{1}{2\sigma_t^2} (z_t - \mathcal{H}_t \widehat{\lambda}_t)^T R_t^{-1} (z_t - \mathcal{H}_t \widehat{\lambda}_t) \\ \implies & \frac{\partial \log(\ell_{rest}(\zeta_t; \sigma_t^2))}{\partial \sigma_t^2} = -\frac{n_t - p_t - q_{t-1}}{2} \frac{1}{\sigma_t^2} + \frac{1}{2(\sigma_t^2)^2} (z_t - \mathcal{H}_t \widehat{\lambda}_t)^T R_t^{-1} (z_t - \mathcal{H}_t \widehat{\lambda}_t). \end{aligned} \tag{3.31}$$

Maximizing the log-likelihood by taking its derivative equal to zero, gives us the maximum likelihood estimate of $\sigma_t^2$,

$$\widehat{\sigma}^2_{t,\mathrm{EML}} = \frac{(z_t - \mathcal{H}_t\widehat{\lambda}_t)^T R_t^{-1}(z_t - \mathcal{H}_t\widehat{\lambda}_t)}{n_t - p_t - q_{t-1}} = \frac{Q_t}{2a_t},$$

which is closely related to the posterior distribution of $\sigma_t^2$.

Note that $\{r_t(x,x')\}_{x,x'\in D_t}$ is considered as known, but in a practical application the correlation function $r_t(x,x')$ would have to be chosen from a family of correlation functions $r_t(x,x';\varphi_t)$. Thus, the matrix $R_t$ is, in fact, a function $R_t(\varphi_t)$. The hyperparameter $\varphi_t$ has to be estimated in some way. One possible approach is to maximize the concentrated restricted log-likelihood, which is obtained by plugging the value $\widehat{\sigma}^2_{t,\mathrm{EML}}(\varphi_t)$ (it depends on $\varphi$ through $R_t(\varphi_t)$) for $\sigma_t^2$ in the expression of the log-likelihood (3.31). Therefore, we would need to minimize

$$\log(det(R_t(\varphi_t))) + \log(\det(\mathcal{H}_t^T R_t^{-1}(\varphi_t)\mathcal{H}_t)) + (n_t - p_t - q_{t-1})\log(\widehat{\sigma}^2_{t,\mathrm{EML}}(\varphi_t)),$$

and this has to be performed numerically.

### 3.3.2 Universal co-kriging model

The predictive distribution of $Z_s(x)$ given the observations $\mathcal{Z}^{(s)} = z^{(s)}$, and parameters $\beta$, $\beta_\rho$ and $\sigma_t^2$ of the recursive co-kriging model is given in (3.18). This corresponds to the *universal co-kriging model (UK)*, when the hyperparameters are not treated as known constants. In a Bayesian approach, we need to integrate the uncertainty of the parameters to obtain the distribution of $Z_s(x)$ conditioned by $\mathcal{Z}^{(s)} = z^{(s)}$ only. We already obtained the posterior distributions of $\sigma_t^2|z^{(t)}$ and $\beta_t, \beta_{\rho_{t-1}}|z^{(t)}, \sigma_t^2$, for $t = 2,\dots,s$, and $\sigma_1^2|z^{(1)}$ and $\beta_1|z^{(1)},\sigma_1^2$. Thus, the desired marginal distribution for $t = 1,\dots,s$ is obtained by performing the following integration:

$$p(Z_t(x)|z^{(t)}) = \int p(Z_t(x)|z^{(t)},\sigma_t^2,\beta_t,\beta_{\rho_{t-1}})p(\beta_t,\beta_{\rho_{t-1}}|z^{(t)},\sigma_t^2)p(\sigma_t^2|z^{(t)})d\beta_t d\beta_{\rho_{t-1}}d\sigma_t^2.$$

Nevertheless, the distribution of $Z_t(x)|z^{(t)}$ is not Gaussian and it does not have a closed form expression, requiring approximations or Monte Carlo integration. However, both mean and variance, $\mathbb{E}[Z_t(x)|z^{(t)}]$ and $\mathrm{Var}[Z_t(x)|z^{(t)}]$, respectively, have closed form expressions. This is summarized in the following proposition.

**Proposition 3.2** (Proposition 2 of [Le Gratiet & Garnier '14]). *Let us consider $s$ Gaussian processes $\{Z_t(x)\}_{t=1,\dots,s}$, and $\mathcal{Z}^{(s)} = (\mathcal{Z}_t)_{t=1,\dots,s}$ the Gaussian vector containing the values of $\{Z_t(x)\}_{t=1,\dots,s}$ at the points in $\{D_t\}_{t=1,\dots,s}$, with $D_s \subseteq D_{s-1} \subseteq \cdots \subseteq D_1$. If we consider the conditional predictive distribution in equation (3.18), and the posterior distribution of the parameters given in equations (3.27) and (3.30), then we have, for $t = 1,\dots,s$,*

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = u_t^T(x)\Sigma_t\nu_t + r_t^T(x)R_t^{-1}(z_t - \mathcal{H}_t\Sigma_t\nu_t) \tag{3.32}$$

*with*

$$u_1^T = f_1^T,$$

$$\mathcal{H}_1 = F_1,$$
$$u_t^T(x) = (g_{t-1}^T(x)\mathbb{E}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}], f_t(x)^T),$$

*and*

$$\mathcal{H}_t = [G_{t-1} \odot z_{t-1}(D_t)\mathbf{1}_{q_{t-1}}^T, F_t],$$

*for $t > 1$. Furthermore, we have*

$$Var[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = \widehat{\sigma}^2_{\rho_{t-1}}(x)\mathrm{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}]$$
$$+\frac{Q_t}{2(a_t-1)}(1 - r_t^T(x)R_t^{-1}r_t^T(x)) + (u_t^T(x) - r_t^T(x)R_t^{-1}\mathcal{H}_t)\widehat{\Sigma}_t(u_t^T(x) - r_t^T(x)R_t^{-1}\mathcal{H}_t)^T \tag{3.33}$$

*with*

$$\widehat{\sigma}^2_{\rho_{t-1}}(x) = \widehat{\rho}^2_{t-1}(x) + g_{t-1}^T(x)\widehat{\Sigma}_{\rho,t}g_{t-1}(x),$$

*and*

$$\widehat{\rho}_{t-1}(x) = g_{t-1}^T(x)[\widehat{\Sigma}_t, \widehat{\nu}_t]_{1,\dots,q_{t-1}},$$

*where $\widehat{\Sigma}_{\rho,t}$ is the submatrix with the first $q_{t-1}$ rows and columns of $\widehat{\Sigma}_t$ (relative to the hyperparameter vector $\beta_{\rho_{t-1}}$), which has the same expression of $\Sigma_t$, but with $\sigma_t^2$ replaced by its posterior mean, and similarly for $\widehat{\nu}_t$.*

*Proof.*
*Mean for $t > 1$:* By the law of total expectation (see Section A.3.1), we have that

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = \mathbb{E}[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}}]|\mathcal{Z}^{(t)} = z^{(t)}].$$

Using equations (3.18) and (3.19), we have that, for $t > 1$,

$$[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}}] \sim \mathcal{N}(\mu_{Z_t}(x), \sigma^2_{Z_t}(x))$$
$$\implies \mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}}] = \mu_{Z_t}(x) =$$
$$\rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t) \odot z_{t-1}(D_t) - F_t\beta_t),$$

thus, using the fact that $\rho_{t-1}(x) = g_{t-1}^T(x)\beta_{\rho_{t-1}}$, and given that $\mu_{Z_{t-1}}(x)$ is independent of both $\rho_{t-1}(x)$ and $z_t$ (which is a constant vector of observed values), we obtain

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = \mathbb{E}[\mu_{Z_t}(x)|\mathcal{Z}^{(t)} = z^{(t)}] =$$
$$g_{t-1}^T(x)\mathbb{E}[\beta_{\rho_{t-1}}|\mathcal{Z}^{(t)} = z^{(t)}]\mathbb{E}[\mu_{Z_{t-1}}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}] + f_t^T(x)\mathbb{E}[\beta_t(x)|\mathcal{Z}^{(t)} = z^{(t)}]$$
$$+r_t^T(x)R_t^{-1}\big(z_t - G_{t-1}\mathbb{E}[\beta_{\rho_{t-1}}|\mathcal{Z}^{(t)} = z^{(t)}] \odot z_{t-1}(D_t) - F_t\mathbb{E}[\beta_t|\mathcal{Z}^{(t)} = z^{(t)}]\big).$$

Note that, when we take the expectation of $(\beta_{\rho_{t-1}}, \beta_t)$ conditioned by $z^{(t)}$, we need to use the law of total expectation again, since we only have their posterior distribution conditioned by $\sigma_t^2$, which is greatly facilitated by the fact that the posterior mean, $\Sigma_t\nu_t$, does not depend on $\sigma_t^2$:

$$\mathbb{E}[\beta_{\rho_{t-1}}, \beta_t|Z^{(t)} = z^{(t)}] = \mathbb{E}[\mathbb{E}[\beta_{\rho_{t-1}}, \beta_t|\sigma_t^2, Z^{(t)} = z^{(t)}]|Z^{(t)} = z^{(t)}] =$$
$$\mathbb{E}[\Sigma_t\nu_t|Z^{(t)} = z^{(t)}] = \Sigma_t\nu_t = \widehat{\Sigma}_t\widehat{\nu}_t$$

Thus,

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = u_t(x)\widehat{\Sigma}_t\widehat{\nu}_t + r_t^T(x)R_t^{-1}(z_t - \mathcal{H}_t\widehat{\Sigma}_t\widehat{\nu}_t),$$

*Mean for $t = 1$:* Again, from the law of total expectation, we have that

$$\mathbb{E}[Z_1(x)|\mathcal{Z}^{(1)} = z^{(1)}] = \mathbb{E}[\mathbb{E}[Z_1(x)|\mathcal{Z}^{(1)} = z^{(1)}, \sigma_1^2, \beta_1]|\mathcal{Z}^{(1)} = z^{(1)}].$$

We know that

$$[Z_1(x)|\mathcal{Z}^{(1)} = z^{(1)}, \sigma_1^2, \beta_1] \sim \mathcal{N}(\mu_{Z_1}(x), \sigma_{Z_1}^2(x)),$$

with

$$\begin{cases} \mu_{Z_1}(x) = f_1^T(x)\beta_1 + r_1(x)R_1^{-1}(z^{(1)} - F_1\beta_1) \\ \sigma_{Z_1}^2 = \sigma_1^2(1 - r_1^T(x)R_1^{-1}r_1(x)) \end{cases}.$$

Therefore,

$$\mathbb{E}[Z_1(x)|\mathcal{Z}^{(1)} = z^{(1)}] = \mathbb{E}[\mu_{Z_1}(x)|\mathcal{Z}^{(1)} = z^{(1)}] =$$
$$f_1^T(x)\mathbb{E}[\beta_1|\mathcal{Z}^{(1)} = z^{(1)}] + r_1^T(x)R_1^{-1}(z^{(1)} - F_1\mathbb{E}[\beta_1|\mathcal{Z}^{(1)} = z^{(1)}]) =$$
$$f_1^T(x)\widehat{\Sigma}_1\widehat{\nu}_1 + r_1^T(x)R_1^{-1}(z^{(1)} - F_1\widehat{\Sigma}_1\widehat{\nu}_1) =$$
$$u_1^T(x)\widehat{\Sigma}_1\widehat{\nu}_1 + r_1^T(x)R_1^{-1}(z^{(1)} - \mathcal{H}_1\widehat{\Sigma}_1\widehat{\nu}_1).$$

*Variance for $t > 1$:* For this step of the proof, we will use the law of total variance (see Section A.3.2) twice to obtain the desired variance identity. We know that

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2] = \mathbb{E}[\widetilde{Z}_t(x)] = \mu_{Z_t}(x)$$
$$\implies \text{Var}[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] =$$
$$\text{Var}[\rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) - F_t\beta_t)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] =$$
$$= (u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)\Sigma_t(u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)^T, \tag{3.34}$$

when we observe that, here, $\mu_{Z_{t-1}}(x)$ and $r_t^T(x)R_t^{-1}z_t$ act as constants.

We also know that

$$\text{Var}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2] = \text{Var}[\widetilde{Z}_t(x) = \sigma_{Z_t}^2(x)$$
$$\implies \mathbb{E}[\text{Var}[Z_t(x)]|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] = \mathbb{E}[\sigma_{Z_t}^2(x)||\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] =$$
$$\mathbb{E}[\rho_{t-1}^2(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]\mathbb{E}[\sigma_{Z_{t-1}}^2(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t(x)),$$

by observing that $\rho_{t-1}(x)$ and $\sigma_{Z_{t-1}}^2(x)$ are independent. Furthermore, $\sigma_{Z_{t-1}}^2(x)$ depends on $\mathcal{Z}^{(t)} = z^{(t)}$ through $\mathcal{Z}^{(t-1)} = z^{(t-1)}$ only, and it is independent of $\sigma_t^2$. With this result, we obtain

$$\mathbb{E}[\sigma_{Z_{t-1}}^2(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] = \text{Var}[\widetilde{Z}_{t-1}(x)] = \text{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}, \beta_{t-1}, \beta_{\rho_{t-2}}, \sigma_{t-1}^2].$$

Note that

$$\mathbb{E}[\rho_{t-1}^2(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] = g_{t-1}^T(x)\mathbb{E}[\beta_{\rho_{t-1}}\beta_{\rho_{t-1}}^T|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]g_{t-1}(x) =$$

$$g_{t-1}^T(x)(\Sigma_{\rho,t} + [\Sigma_t, \nu_t]_{1,\dots,q_{t-1}})g_{t-1}(x),$$

therefore, we obtain

$$\begin{aligned}\mathrm{Var}[Z_t(x)]|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2] = \widehat{\sigma}^2_{\rho_{t-1}}(x)\mathrm{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}] \\ +\sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t^T(x)).\end{aligned} \tag{3.35}$$

By the law of total variance and equations (3.34) and (3.35), we get:

$$\begin{aligned}\mathrm{Var}[Z_t|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] = \mathrm{Var}\big[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2\big] + \\ +\mathbb{E}\big[\mathrm{Var}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2\big] = \\ (u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)\Sigma_t(u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)^T + \\ +\widehat{\sigma}^2_{\rho_{t-1}}(x)\mathrm{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}] + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t^T(x))\end{aligned}$$

To drop the dependence on $\sigma_t^2$ in $\mathrm{Var}[Z_t|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]$, we use the law of total variance again, and obtain

$$\begin{aligned}\mathrm{Var}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = \\ = \mathrm{Var}[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}] + \mathbb{E}[\mathrm{Var}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}].\end{aligned}$$

Note that, as previously stated, $\mathbb{E}[\mu_{Z_t}(x)|\mathcal{Z}^{(t)} = z^{(t)}]$ is independent of $\sigma_t^2$, thus

$$\begin{aligned}\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2] = \mathbb{E}[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}]|\mathcal{Z}^{(t)} = z^{(t)}] = \\ \mathbb{E}[\mu_{Z_t}(x)|\mathcal{Z}^{(t)} = z^{(t)}] \\ \implies \mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}] \perp \sigma_t^2.\end{aligned}$$

For this reason, the term $\mathrm{Var}[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}]$ is equal to 0, since

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] = \mathbb{E}[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2]|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2],$$

and this is independent of $\sigma_t^2$.

Therefore, now, we only need to take the expectation in $\sigma_t^2$ of the term $\mathrm{Var}[Z_t|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]$ which is equal to

$$\begin{aligned}(u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)\Sigma_t(u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)^T + \\ +\widehat{\sigma}^2_{\rho_{t-1}}(x)\mathrm{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}] + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t^T(x)).\end{aligned}$$

Observe that $[\sigma_t^2|z^{(t)}] \sim \mathcal{IG}\Big(a_t, \frac{Q_t}{2}\Big)$ implies that the posterior mean of $\sigma_t^2$ is $\frac{Q_t}{2(a_t-1)}$. We also note that, since $\Sigma_t$ is linear in $\sigma_t^2$, the expectation of $\Sigma_t$ is the expression for $\Sigma_t$ with $\sigma_t^2$ replaced by its posterior mean. Thus,

$$\begin{aligned}\mathrm{Var}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = \widehat{\sigma}^2_{\rho_{t-1}}(x)\mathrm{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}] + \\ +\frac{Q_t}{2(a_t-1)}(1 - r_t^T(x)R_t^{-1}r_t^T(x)) + (u_t^T - r_t^T(x)R_t^{-1}\mathcal{H}_t)\widehat{\Sigma}_t(u_t^T - r_t^T(x)R_t^{-1}\mathcal{H}_t)^T\end{aligned}$$

*Variance for $t = 1$:* Follows from easier but similar steps as for the ones performed above for $t > 1$, noting that every term $\rho_{t-1}(x)$ must be equal to 0. □

**Remark 4.** *Where we need to take the expectation of $\Sigma_t\nu_t$ with respect to $\sigma_t^2$, we, in fact, have an expression that does not depend on $\sigma_t^2$ anymore. We have replaced that with $\widehat{\Sigma}_t\widehat{\nu}_t$ only to simplify the notation and understanding.*

### 3.3.3 Cross-validation procedure

Though a cross-validation procedure might be extremely time-consuming for a general model, the recursive formulation of the co-kriging model allows shortcuts for it.

We let $\xi_s$ be the set of indices of $n_{test}$ test points in $D_s$, which constitute the test set $D_{test}$, and $\xi_t$, for $1 \leq t < s$, be the corresponding set of indices in $D_t$. Notice that this is possible because of the nested property $D_s \subseteq D_{s-1} \subseteq \cdots \subseteq D_1$, meaning that if we remove a subset of the data from the highest level of code, we can remove it from all of the other levels as well.

The following proposition gives the predictive error and variance vectors for the cross-validation procedure for the non-informative case. This is an extension of the cross-validation for Gaussian processes (subsection 2.7.2) and it provides the predictive error and variance for known or unknown parameters $(\beta_t, \beta_{\rho_{t-1}}, \sigma_t^2)$, for $1 < t \leq s$, and $(\beta_1, \sigma_1^2)$, the simple co-kriging and universal co-kriging models, respectively.

As in subsection 2.7.2, we use $v_\zeta$ for a vector containing the entries of $v$ with indices in the set $\zeta$, $v_{-\zeta}$ for the vector of entries of $v$ with indices not in $\zeta$, $[A]_{[\zeta,\gamma]}$ for the submatrix of $A$ containing the rows with indices in $\zeta$ and columns with indices in $\gamma$, and similarly to the vector case for the submatrices $[A]_{[-\zeta,\gamma]}$, $[A]_{[\zeta,-\gamma]}$, and $[A]_{[-\zeta,-\gamma]}$.

**Proposition 3.3** (Proposition 3 of [Le Gratiet & Garnier '14]). *Let us consider $s$ Gaussian processes $\{Z_t(x)\}_{t=1,\ldots,s}$, as in the recursive model presented in (3.16), and $\mathcal{Z}^{(s)} = (\mathcal{Z}_1, \ldots, \mathcal{Z}_s)$, with $\mathcal{Z}_t$ containing the values of $\{Z_t(x)\}_{x \in D_t}$, for $t = 1 \ldots, s$, and $D_s \subseteq D_{s-1} \subseteq \ldots D_1$. We denote by $D_{test}$ a set consisting of points of index $\xi_s$ of $D_s$ and $\xi_t$ the corresponding indices of the points in $D_t$ for $1 \leq t < s$. Let $\lambda_{t,-\xi_t}$ denote the posterior mean of the regression, and adjustment parameters $(\beta_{\rho_{t-1}}^T \quad \beta_t^T)^T$. Then, if $\varepsilon_{Z_t,\xi_t}$ are the errors (i.e. observed values minus predicted values) of the cross-validation procedure at the level $t$ when we remove the points of $D_{test}$ from levels $u$ to $t$, we have*

$$(\varepsilon_{Z_t,\xi_t} - \widehat{\rho}_{t-1}(D_{test}) \odot \varepsilon_{Z_{t-1},\xi_{t-1}})[R_t^{-1}]_{[\xi_t,\xi_t]} = [R_t^{-1}(z_t - \mathcal{H}_t \lambda_{t,-\xi_t})]_{[\xi_t]}, \tag{3.36}$$

*with $\varepsilon_{Z_i,\xi_i} = 0$, for $i < u$,*

$$\lambda_{t,-\xi_t} = \left([\mathcal{H}_t]_{[-\xi_t]}^T K_t [\mathcal{H}_t]_{[-\xi_t]}\right)^{-1} [\mathcal{H}_t]_{-\xi_t}^T K_t z_t(D_t \backslash D_{test}), \tag{3.37}$$

$$\widehat{\rho}_{t-1} = g_{t-1}^T(D_{test})[\lambda_{t,-\xi_t}]_{1,\ldots,q_{t-1}},$$

*and*

$$K_t = [R_t^{-1}]_{[-\xi_t,-\xi_t]} - [R_t^{-1}]_{[-\xi_t,\xi_t]} \left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1} [R_t^{-1}]_{[\xi_t,-\xi_t]}.$$

*Furthermore, if we denote by $\sigma^2_{Z_t,\xi_t}$ the variances of the corresponding cross-validation procedure, we have*

$$\sigma^2_{Z_t,\xi_t} = \widehat{\sigma}^2_{\rho_{t-1},-\xi_t}(D_{test}) \odot \sigma^2_{Z_{t-1},\xi_{t-1}} + \sigma^2_{t,-\xi_t} \mathrm{diag}\left(\left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1}\right) + \mathcal{V}_t, \tag{3.38}$$

*with*

$$\widehat{\sigma}^2_{\rho_{t-1},-\xi_t}(D_{test}) = g_{t-1}^T(D_{test})\left(\Sigma_{\rho_{t-1},\xi_t} + [\lambda_{t,-\xi_t}]_{1,\ldots,q_{t-1}} [\lambda_{t,-\xi_t}]_{1,\ldots,q_{t-1}}^T\right) g_{t-1}(D_{test}),$$

$$\Sigma_{\rho_{t-1},\xi_t} = \left[\left([\mathcal{H}_t^T]_{[-\xi_t]} K_t [\mathcal{H}_t]_{[-\xi_t]}\right)^{-1}\right]_{[1,\ldots,q_{t-1},1,\ldots,q_{t-1}]},$$

*and*

$$\sigma^2_{t,-\xi_t} = \frac{\big(z_t(D_t \backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}\lambda_{t,-\xi_t}\big)^T K_t\big(z_t(D_t \backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}\lambda_{t,-\xi_t}\big)}{n_t - p_t - q_{t-1} - n_{test}}, \tag{3.39}$$

*where* $\sigma^2_{i,-\xi_i} = 0$ *for* $i < u$, $n_{test}$ *is the length of the index vector* $\xi_s$, $\mathcal{H}_t = [G_{t-1} \odot (z_{t-1}(D_t)\mathbf{1}^T_{q_{t-1}}) \quad F_t]$, *and*

$$\mathcal{V}_t = \mathcal{U}_t\big([\mathcal{H}_t^T]_{[-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]}\big)^{-1}\mathcal{U}_t^T,$$

$$\mathcal{U}_t = v_t + \big([R_t^{-1}]_{[\xi_t,\xi_t]}\big)^{-1}[R_t^{-1}\mathcal{H}_t]_{[\xi_t]},$$

*and* $v_t = -[g^T_{t-1}(D_{test}) \odot (\varepsilon_{Z_{t-1},\xi_{t-1}}\mathbf{1}^T_{q_{t-1}}) \quad 0]$.

*Proof.* We begin by ordering the points in $D_t$ so that the points with index $\xi_t$ are the $n_{test}$ last points of $D_t$ for every $t$. Then,

$$R_t = \begin{bmatrix} [R_t]_{[-\xi_t,-\xi_t]} & [R_t]_{[-\xi_t,\xi_t]} \\ [R_t]_{[\xi_t,-\xi_t]} & [R_t]_{[\xi_t,\xi_t]} \end{bmatrix}.$$

Using the blockwise inversion formula (A.7), we have that

$$R_t^{-1} = \begin{bmatrix} A & B \\ B^T & \mathcal{Q}^{-1} \end{bmatrix},$$

with

$$A = \big([R_t]_{[-\xi_t,-\xi_t]}\big)^{-1} + \big([R_t]_{[-\xi_t,-\xi_t]}\big)^{-1}[R_t]_{[-\xi_t,\xi_t]}\mathcal{Q}^{-1}[R_t]_{[\xi_t,-\xi_t]}\big([R_t]_{[-\xi_t,-\xi_t]}\big)^{-1},$$

$$B^T = -\mathcal{Q}^{-1}[R_t]_{[\xi_t,-\xi_t]}\big([R_t]_{[-\xi_t,-\xi_t]}\big)^{-1}$$

and

$$\mathcal{Q} = [R_t]_{[\xi_t,\xi_t]} - [R_t]_{[\xi_t,-\xi_t]}\big([R_t]_{[-\xi_t,-\xi_t]}\big)^{-1}[R_t]_{[-\xi_t,\xi_t]}.$$

Now, we must compute the prediction for the points in $D_{test}$ at level the $t$. This will be done for two cases: the simple co-kriging, when the parameters are fixed, and the universal co-kriging, when they must be estimated.

*Simple co-kriging:* In this case, we have the variance and trend parameters fixed: $\sigma^2_{t,-\xi_t} = \frac{Q_t}{2(a_t-1)}$, $\lambda_{t,-\xi_t} = \Sigma_t\nu_t$ and $\mathcal{V}_t = 0$ (refer to Equation (3.20) and compare to (3.33)); $\mathcal{V}_t$ is an additive term related to the parameter estimations in the universal co-kriging case). Note that $\mathcal{Q} = \big([R_t^{-1}]_{[\xi_t,\xi_t]}\big)^{-1}$, thus $\frac{Q_t}{2(a_t-1)}\mathcal{Q}$ represents the covariance matrix of a Gaussian process with kernel $\frac{Q_t}{2(a_t-1)}r_t(x,x')$ on the points in $D_{test}$, and conditioned by the value of the process on the points in $D_t \backslash D_{test}$. Also,

$$\mathcal{Q}_{i,i} = 1 - r_t(x_t^i, D_t\backslash D_{test})^T\big([R_t]_{[-\xi_t,-\xi_t]}\big)^{-1}r_t(x_t^i, D_t\backslash D_{test}).$$

Therefore, by Equation (3.20), achieving (3.38) is straightforward.

For the predictive mean, by Equation (3.19), we have that the predicted output values for inputs in $D_{test}$ are

$$\mu_{Z_t}(D_{test}) = h_t^T(D_{test})\Sigma_t\nu_t + [R_t]_{[\xi_t,-\xi_t]}[R_t]^{-1}_{[-\xi_t,-\xi_t]}(z_t(D_t\backslash D_{test}) - [\mathcal{H}]^T_{[-\xi_t]}\Sigma_t\nu_t), \tag{3.40}$$

with $h_t^T(x) = [\mu_{Z_{t-1}}(x) g_{t-1}^T(x) \quad f_t^T(x)]$ and $\Sigma_t \nu_t = \lambda_{t,-\xi_t}$.

Now, note that

$$[R_t^{-1}(z_t - \mathcal{H}_t^T \lambda_{t,-\xi_t})]_{[\xi_t]} = [R_t^{-1}]_{[\xi_t,\xi_t]}(z_t(D_{test}) - [\mathcal{H}_t]_{[\xi_t]}^T \lambda_{t,-\xi_t}) +$$

$$[R_t^{-1}]_{[\xi_t,-\xi_t]}(z_t(D_t \backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}^T \lambda_{t,-\xi_t}).$$

Since $\left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1}[R_t^{-1}]_{[\xi_t,-\xi_t]} = \mathcal{Q}B^T = -[R_t]_{[\xi_t,-\xi_t]}[R_t]_{[-\xi_t,-\xi_t]}^{-1}$, we have that

$$\left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1}[R_t^{-1}(z_t - \mathcal{H}_t^T \lambda_{t,-\xi_t})]_{[\xi_t]} = z_t(D_{test}) - [\mathcal{H}_t]_{[\xi_t]}^T \lambda_{t,-\xi_t} -$$

$$[R_t]_{[\xi_t,-\xi_t]}[R_t]_{[-\xi_t,-\xi_t]}^{-1}(z_t(D_t \backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}^T \lambda_{t,-\xi_t}).$$

We then add and subtract $h_t^T(D_{test}) = [g_{t-1}^T(D_{test}) \odot (\mu_{Z_{t-1}}(D_{test}) \mathbf{1}_{q_{t-1}}^T) \quad f_t^T(D_{test})]$ multiplied by $\lambda_{t,-\xi_t}$ in the previous equation. This, Equation (3.40), the equality $\lambda_{t,-\xi_t} = \Sigma_t \nu_t$, and the fact that $\varepsilon_{Z_t,\xi_t} = z_t(D_{test}) - \mu_{Z_t}(D_{test})$ imply that

$$\left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1}[R_t^{-1}(z_t - \mathcal{H}_t^T \lambda_{t,-\xi_t})]_{[\xi_t]} = \varepsilon_{Z_t,\xi_t} - ([\mathcal{H}_t]_{[\xi_t]}^T - h_t^T(D_{test}))\lambda_{t,-\xi_t}.$$

Finally, note that

$$[\mathcal{H}_t]_{[\xi_t]}^T - h_t^T(D_{test}) = [g_{t-1}^T(D_{test}) \odot ((z_{t-1}(D_{test}) - \mu_{Z_{t-1}}(D_{test}))\mathbf{1}_{q_{t-1}}^T) \quad 0] =$$

$$[g_{t-1}^T(D_{test}) \odot (\varepsilon_{Z_{t-1},\xi_{t-1}} \mathbf{1}_{q_{t-1}}^T) \quad 0]$$

$$\implies ([\mathcal{H}_t]_{[\xi_t]}^T - h_t^T(D_{test}))\lambda_{t,-\xi_t} = \widehat{\rho}_{t-1}(D_{test}) \odot \varepsilon_{Z_{t-1},\xi_{t-1}}$$

$$\implies \left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1}[R_t^{-1}(z_t - \mathcal{H}_t^T \lambda_{t,-\xi_t})]_{[\xi_t]} = \varepsilon_{Z_t,\xi_t} - \widehat{\rho}_{t-1}(D_{test}) \odot \varepsilon_{Z_{t-1},\xi_{t-1}}.$$

*Universal co-kriging:* When the trend and variance parameters are unknown, they must be re-estimated with the data set with observations on the points in $D_t \backslash D_{test}$. We must refer to Subsection 3.3.1, where we have expressions for the estimates of the parameters when trained on $D_t$ and obtain similar ones training only on $D_t \backslash D_{test}$.

Notice that all expressions in Subsection 3.3.1 involve $R_t^{-1}$. In our case, this must be replaced by $[R_t]_{[-\xi_t,-\xi_t]}^{-1}$. Since we do not want to invert new (and big) matrices for each different set $D_{test}$, we must write an expression for $[R_t]_{[-\xi_t,-\xi_t]}^{-1}$ including only the previously obtained inverse matrix $R_t^{-1}$ and multiplications. For this, we will use the block matrix inversion formula (A.7) again. We write the inverse of $R_t$ as

$$R_t^{-1} = \begin{bmatrix} [R_t^{-1}]_{[-\xi_t,-\xi_t]} & [R_t^{-1}]_{[-\xi_t,\xi_t]} \\ [R_t^{-1}]_{[\xi_t,-\xi_t]} & [R_t^{-1}]_{[\xi_t,\xi_t]} \end{bmatrix}$$

$$\implies [R_t]_{[-\xi_t,-\xi_t]} = \left([R_t^{-1}]_{[-\xi_t,-\xi_t]} - [R_t^{-1}]_{[-\xi_t,\xi_t]}[R_t^{-1}]_{[\xi_t,\xi_t]}[R_t^{-1}]_{[\xi_t,-\xi_t]}\right)^{-1}.$$

Hence, $[R_t]_{[-\xi_t,-\xi_t]}^{-1} = K_t$. With this, it is easier to obtain the estimates for the trend and variance parameters. By (3.30), we promptly obtain the estimate for the variance parameter written in Equation (3.39). Similarly, the trend parameters estimate in 3.37 is obtained using (3.28) and (3.29).

For the mean, we recall Equation (3.32), with which we get the predictive mean when training on the set $D_t \backslash D_{test}$:

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = u_t^T(x)\lambda_{t,-\xi_t} + [r_t^T(x)]_{[-\xi_t]}[R_t]^{-1}_{[-\xi_t,-\xi_t]}(z_t(D_t\backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}\lambda_{t,-\xi_t}),$$

with $u_t^T(x) = (g_{t-1}^T(x)\mathbb{E}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}] \quad f_t(x)^T)$. We highlight that the conditioning term in the expectations is on the known values of the Gaussian processes $Z_i(x)$ on the points in $D_i\backslash D_{test}$, for $u \leq i \leq t$ and $D_i$ for $i \leq u$, but we keep the previous notation for simplicity. Then,

$$\mathbb{E}[Z_t(D_{test})|\mathcal{Z}^{(t)} = z^{(t)}] = u_t^T(D_{test})\lambda_{t,-\xi_t} + [R_t]_{[\xi_t,-\xi_t]}[R_t]^{-1}_{[-\xi_t,-\xi_t]}(z_t(D_t\backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}\lambda_{t,-\xi_t}).$$

We obtained an equivalent expression to the one in the simple co-kriging case. This way, the same algebraic manipulations performed in the previous case, only replacing $h_t$ with $u_t$, which also have similar expressions, yield equation (3.36).

The variance of the universal co-kriging is given by Equation (3.33), therefore when training on $D_i\backslash D_{test}$, for $u \leq i \leq t$ and $D_i$ for $i < u$, we obtain an equivalent expression to the simple co-kriging case except for the last term which then becomes

$$(u_t^T(D_{test}) - [R_t]_{[\xi_t,-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]})([\mathcal{H}_t^T]_{[-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]})^{-1}(u_t^T(D_{test}) - [R_t]_{[\xi_t,-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]})^T.$$

We have that

$$u_t^T(D_{test}) - [R_t]_{[\xi_t,-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]} = (u_t^T(D_{test}) - [\mathcal{H}_t]_{[\xi_t]}) + ([\mathcal{H}_t]_{[\xi_t]} - [R_t]_{[\xi_t,-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]}),$$

with

$$u_t^T(D_{test}) - [\mathcal{H}_t]^T_{[\xi_t]} = -[g_{t-1}^T(D_{test}) \odot (\varepsilon_{Z_{t-1},\xi_{t-1}}\mathbf{1}^T_{q_{t-1}}) \quad 0] = v_t,$$

and, since from the block matrix inversion of $R_t$ we know that

$$[R_t^{-1}\mathcal{H}_t]_{\xi_t} = B^T[\mathcal{H}_t]_{[-\xi_t]} + \mathcal{Q}^{-1}[\mathcal{H}_t]_{[\xi_t]}$$

$$\implies \mathcal{Q}[R_t^{-1}\mathcal{H}_t]_{\xi_t} = \mathcal{Q}B^T[\mathcal{H}_t]_{[-\xi_t]} + [\mathcal{H}_t]_{[\xi_t]} = \mathcal{Q}(-\mathcal{Q}^{-1}[R_t]_{[\xi_t,-\xi_t]}K_t)[\mathcal{H}_t]_{[-\xi_t]} + [\mathcal{H}_t]_{[\xi_t]}$$
$$= [\mathcal{H}_t]_{[\xi_t]} - [R_t]_{[\xi_t,-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]}.$$

Therefore, we obtained the expression for $\mathcal{V}_t$:

$$\mathcal{V}_t = \mathcal{U}_t([\mathcal{H}_t^T]_{[-\xi_t]}K_t[\mathcal{H}_t]_{[-\xi_t]})^{-1}\mathcal{U}_t^T$$

with $\mathcal{U}_t = v_t + ([R_t^{-1}]_{[\xi_t,\xi_t]})^{-1}[R_t^{-1}\mathcal{H}_t]_{[\xi_t]}$. □

**Remark 5.** *The expressions for $t = 1$ are obtained by ignoring all terms that are function of $\rho_{t-1}(x)$, $g_{t-1}(x)$ and $\beta_{\rho_{t-1}}$. For example, $h_t^T(x) = f_t^T(x)$ instead of $h_t^T(x) = [\mu_{Z_{t-1}}(x)g_{t-1}^T(x) \quad f_t^T(x)]$.*

The equations presented in the previous propositions are closed form expressions for the $k$-fold cross-validation procedure. They are still valid for $s = 1$, which defines a simple Gaussian process model, and allow re-estimation of the parameters of the model for each test set $D_{test}$. The cost of the procedure is determined by the inversion of the matrices $[R_t^{-1}]_{[\xi_t,\xi_t]}$ of size $n_{test} \times n_{test}$ for $u \leq t \leq s$ and $1 \leq u \leq s$ fixed, when intending to remove the points of $D_{test}$ from all sets $D_u$. This is far less when compared to the

standard estimation, which would require the inversion of matrices $[R_t]_{[-\xi_t,-\xi_t]}$ of size $(n_t - n_{test}) \times (n_t - n_{test})$. Notice that, for small $t$, we expect a big number $n_t$ of data points at level $t$.

Nevertheless, this proposition does not allow the re-estimation of the hyperparameters of the correlation functions $r_t(x, x')$, this would have to be done separately. Even in this case, however, the computational cost is reduced by using the aforementioned results.

## 3.4 Examples

In the next few sections, we present some toy data that exemplify the behavior of the recursive co-kriging model for various settings.

### 3.4.1 1-dimensional input data and 2 levels of fidelity

We consider the following functions for the high and low-fidelity levels:

$$f_{high}(x) = (6x - 2)^2 \sin(12x - 4),$$

$$f_{low}(x) = \frac{1}{2}(6x - 2)^2 \sin(12x - 4) + 10x - 10 = \frac{1}{2} f_{high}(x) + 10x - 10;$$

For the low-fidelity level, we use 11 equally spaced points in the interval $[0, 1]$, and for the high-fidelity level we use the first, fourth, seventh and last of them, see Figure 3.1. The most basic Gaussian process with zero-mean in the noiseless case is fitted to the data and the prediction is shown in Figure 3.2.

Different universal co-kriging models are fitted to the data and their predictions are shown in Figures 3.3 - 3.8. The root mean square error (RMSE) and mean absolute error are computed on 200 equispaced points in $[0, 1]$. Respective summaries are exhibited in Tables 3.1 - 3.6.

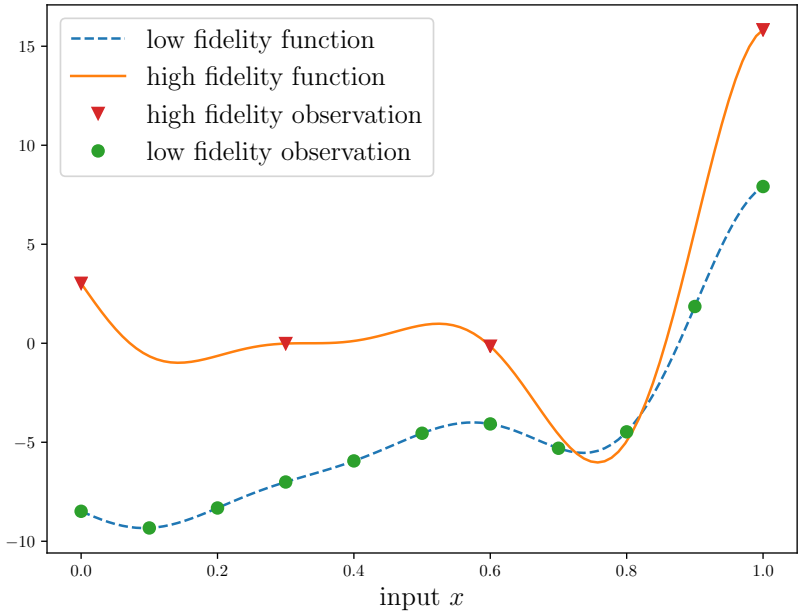



Figure 3.1: Low and high fidelity functions with 11 low fidelity observations and 4 high fidelity observations.

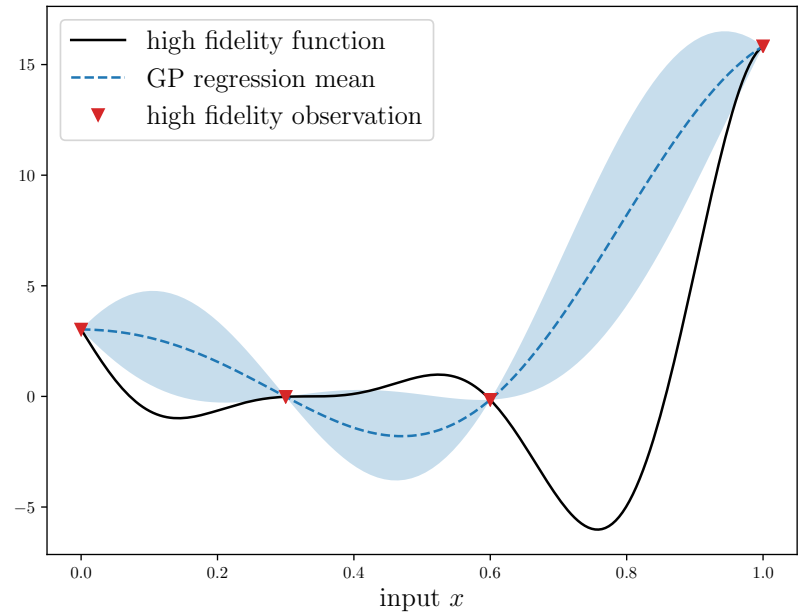



Figure 3.2: Noiseless Gaussian process regression with 95% confidence intervals using only the high fidelity observations for training. The kernel used was $\sigma^2 \exp\{-(x - x')^2/2l^2\}$ and the hyperparameters were optimized. The model presented a RMSE of 5.575 and a MAE of 3.947.

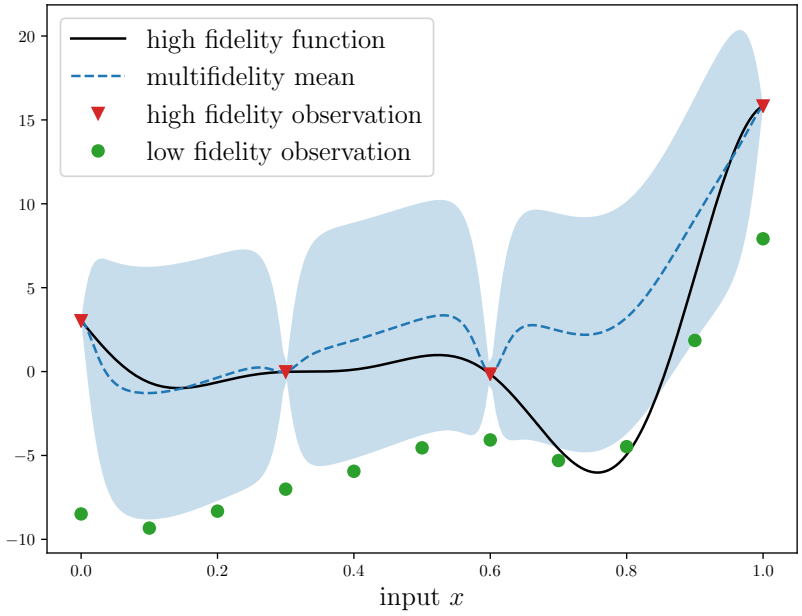



Figure 3.3: UK model 1 fitted to the data with 95% confidence intervals. The model presented a RMSE of 3.861 and a MAE of 2.793.

Table 3.1: Model 1 summary.

| | Level | Model specifications | Estimates |
|---|---|---|---|
| Kernel | 1 | Matérn 5/2 | $l = 0.346$ |
| | 2 | Matérn 5/2 | $l = 0.01$ |
| Trend | 1 | $f_1^T(x) = 1$ | $\beta_1^T = -1.028$ |
| | 2 | $f_2^T(x) = 1$ | $\beta_2^T = 7.383$ |
| Variance | 1 | - | $\sigma_1^2 = 59.888$ |
| | 2 | - | $\sigma_2^2 = 7.118$ |
| Adjustment | 2 | $g_1^T(x) = 1$ | $\beta_{\rho_1}^T = 0.93$ |

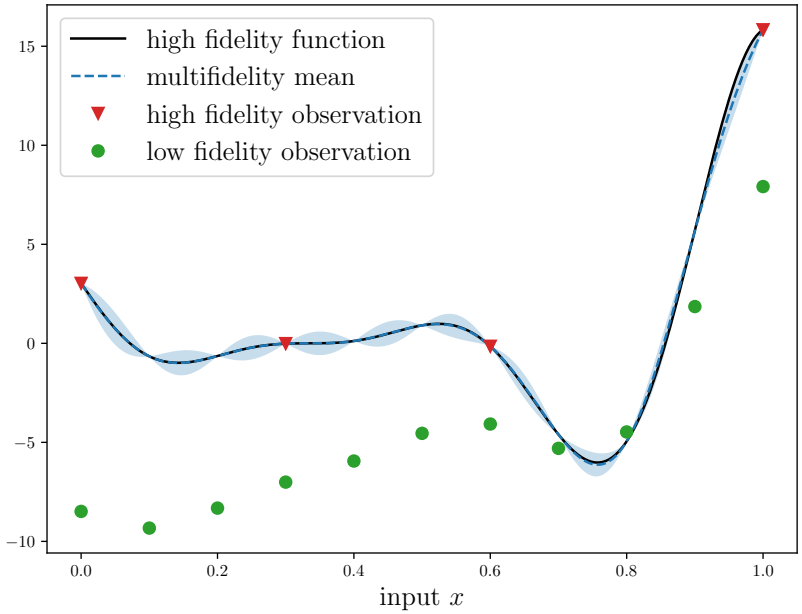



Figure 3.4: UK model 2 fitted to the data with 95% confidence intervals. The model presented a RMSE of 0.184 and a MAE of 0.079.

Table 3.2: Model 2 summary.

| | Level | Model specifications | Estimates |
|---|---|---|---|
| Kernel | 1 | Matérn 5/2 | $l = 0.346$ |
| | 2 | Matérn 5/2 | $l = 0.103$ |
| Trend | 1 | $f_1^T(x) = 1$ | $\beta_1^T = -1.028$ |
| | 2 | $f_2^T(x) = (1, x)$ | $\beta_2^T = (20, -20)$ |
| Variance | 1 | - | $\sigma_1^2 = 59.888$ |
| | 2 | - | $\sigma_2^2 = 1.3 \times 10^{-28}$ |
| Adjustment | 2 | $g_1^T(x) = 1$ | $\beta_{\rho_1}^T = 2$ |

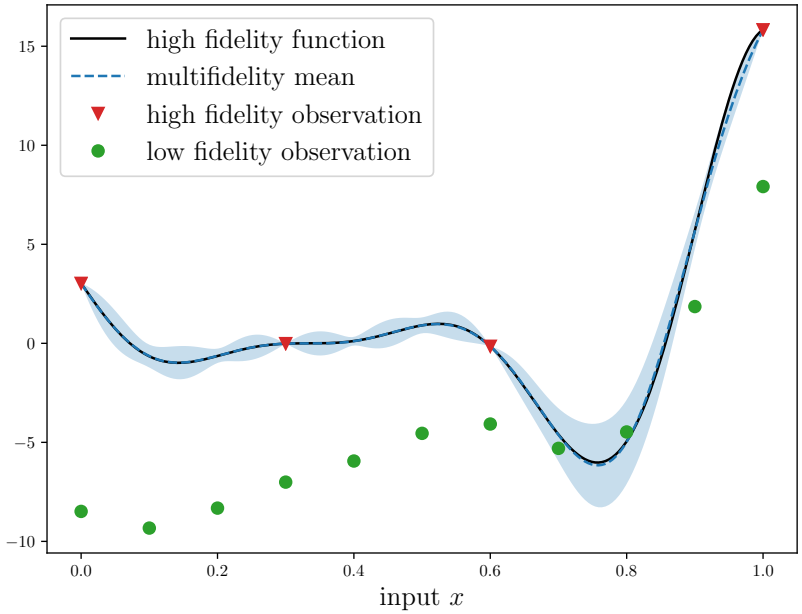



Figure 3.5: UK model 3 fitted to the data with 95% confidence intervals. The model presented a RMSE of 0.182 and a MAE of 0.08.

Table 3.3: Model 3 summary.

| | Level | Model specifications | Estimates |
|---|---|---|---|
| Kernel | 1 | Matérn 5/2 | $l = 0.346$ |
| | 2 | Matérn 5/2 | $l = 0.01$ |
| Trend | 1 | $f_1^T(x) = 1$ | $\beta_1^T = -1.028$ |
| | 2 | $f_2^T(x) = x$ | $\beta_2^T = -17.067$ |
| Variance | 1 | - | $\sigma_1^2 = 59.888$ |
| | 2 | - | $\sigma_2^2 = 114.965$ |
| Adjustment | 2 | $g_1^T(x) = 1$ | $\beta_{\rho_1}^T = 2.01$ |

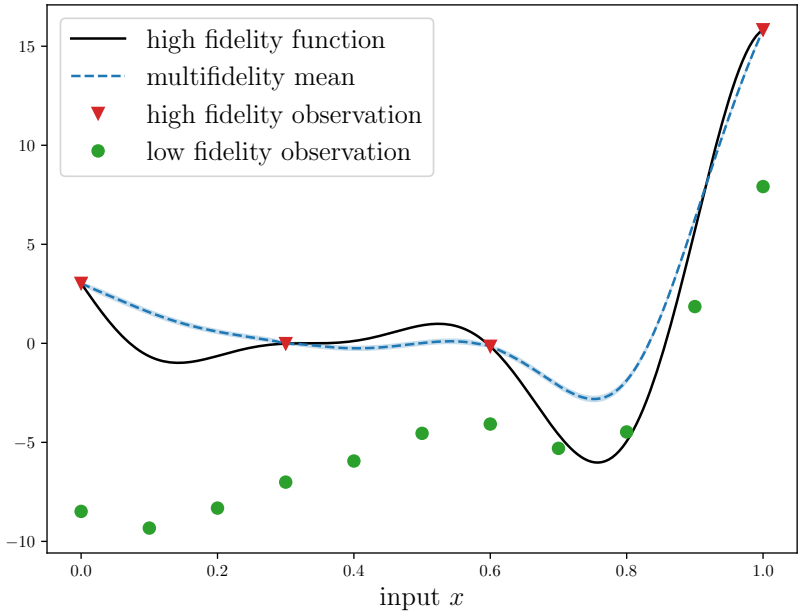



Figure 3.6: UK model 4 fitted to the data with 95% confidence intervals. The model presented a RMSE of 1.544 and a MAE of 1.211.

Table 3.4: Model 4 summary.

| | Level | Model specifications | Estimates |
|---|---|---|---|
| Kernel | 1 | Matérn 5/2 | $l = 0.346$ |
| | 2 | Matérn 5/2 | $l = 0.01$ |
| Trend | 1 | $f_1^T(x) = 1$ | $\beta_1^T = -1.028$ |
| | 2 | $f_2^T(x) = 1$ | $\beta_2^T = 3.69$ |
| Variance | 1 | - | $\sigma_1^2 = 59.888$ |
| | 2 | - | $\sigma_2^2 = 0.002$ |
| Adjustment | 2 | $g_1^T(x) = (1, x)$ | $\beta_{\rho_1}^T = (0.082, 1.452)$ |

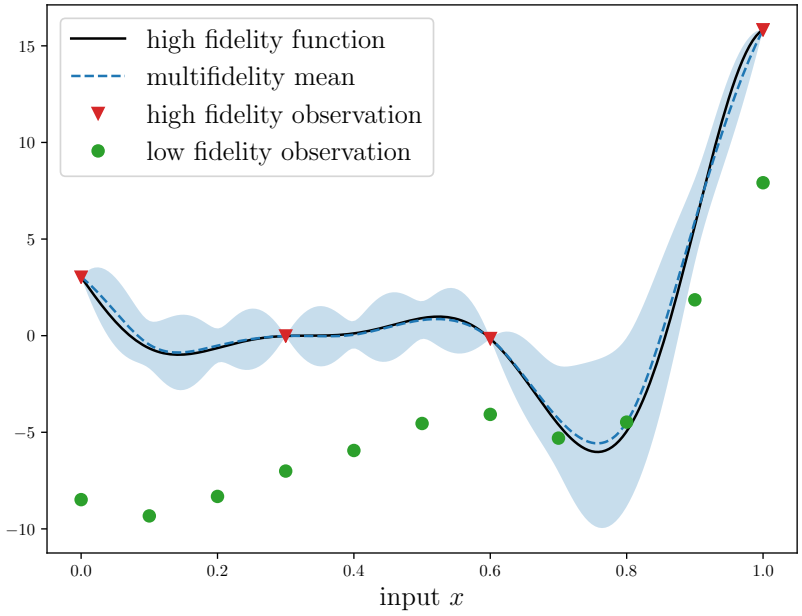



Figure 3.7: UK model 5 fitted to the data with 95% confidence intervals. The model presented a RMSE of 0.327 and a MAE of 0.233.

Table 3.5: Model 5 summary.

| | Level | Model specifications | Estimates |
|---|---|---|---|
| Kernel | 1 | Matérn 3/2 | $l = 0.279$ |
| | 2 | Matérn 3/2 | $l = 2$ |
| Trend | 1 | $f_1^T(x) = (1, x)$ | $\beta_1^T = (-10.159, 15.817)$ |
| | 2 | $f_2^T(x) = x$ | $\beta_2^T = -14.447$ |
| Variance | 1 | - | $\sigma_1^2 = 15.455$ |
| | 2 | - | $\sigma_2^2 = 104.317$ |
| Adjustment | 2 | $g_1^T(x) = 1$ | $\beta_{\rho_1}^T = 1.923$ |

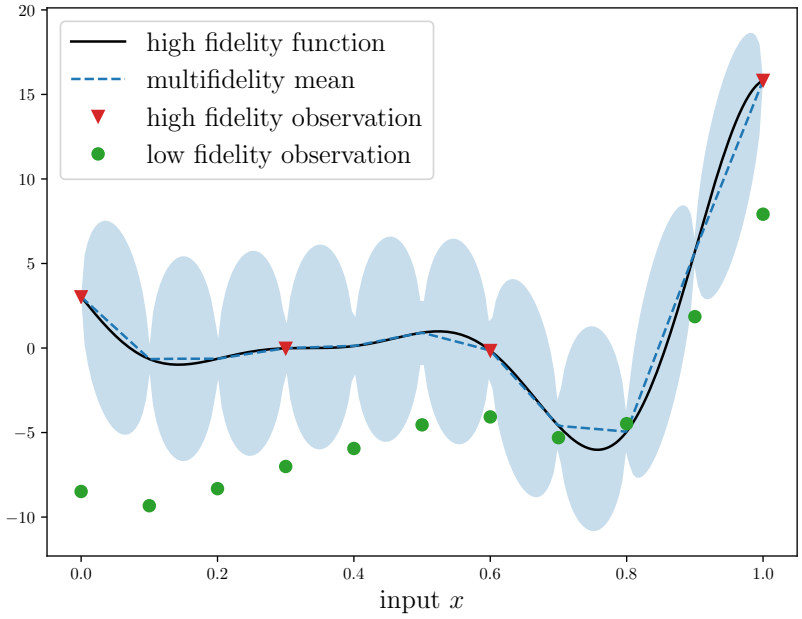



Figure 3.8: UK model 6 fitted to the data with 95% confidence intervals.The model presented a RMSE of 0.56 and a MAE of 0.369.

Table 3.6: Model 6 summary.

| | Level | Model specifications | Estimates |
|---|---|---|---|
| Kernel | 1 | OU | $l = 1.053$ |
| | 2 | OU | $l = 0.01$ |
| Trend | 1 | $f_1^T(x) = 1$ | $\beta_1^T = -1.72$ |
| | 2 | $f_2^T(x) = (1, x)$ | $\beta_2^T = (20, -20)$ |
| Variance | 1 | - | $\sigma_1^2 = 48.402$ |
| | 2 | - | $\sigma_2^2 = 1.2 \times 10^{-30}$ |
| Adjustment | 2 | $g_1^T(x) = 1$ | $\beta_{\rho_1}^T = 2$ |

In Figure 3.9, we compare the errors for different UK model setting. We use the same high and low-fidelity functions, but choose to use more observations, 20 equispaced points in $[0, 1]$ for $D_1$ and $D_2$, in order to better observe the behavior of the errors of the cross-validation procedure.

The errors were obtained by sequentially removing observations in a randomized order from the high fidelity level only and computing the mean absolute error and the root mean squared error on the removed observations. For the UK multi-fidelity models, we used a simple trend and adjustment setting, $f_1^T(x) = f_2^T(x) = g_{\rho_1}^T(x) = 1$, and different correlation kernels, using always the same for both levels. For comparison, the same was done for a 1-level Gaussian process with Matérn $5/2$ kernel and optimized constant mean $\mu$ fit to the high-fidelity data. The resulting errors were averaged over 100 runs of observation removal.

Note that when we have almost all of the observations of the high-level of fidelity, a Gaussian process is as good or even better compared to the multi-fidelity models. This happens because we have almost equal input sets $D_1$ and $D_2$ and, we gain not much more information from the low-fidelity observations. However, when there are few high-fidelity observations, all multi-fidelity models show smaller errors compared to the 1-level Gaussian process. This is exactly the context in which multi-fidelity is needed, when there is only a small amount of high-fidelity data.

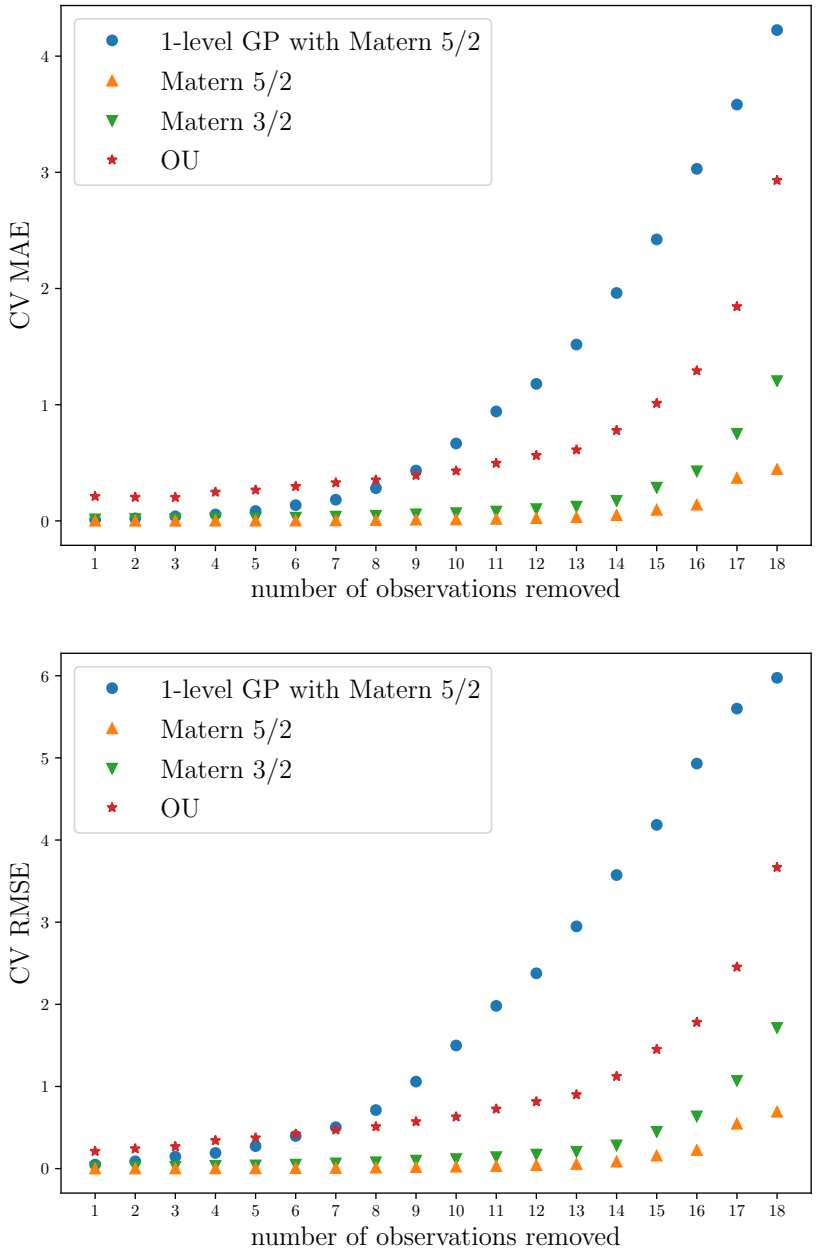



Figure 3.9: Comparison of the average cross-validation when 20 equally spaced observations in [0,1] are used and they are removed from the highest level of fidelity only.

### 3.4.2 1-dimensional input data and 3 levels of fidelity

For this brief example, we use the following functions for the three levels of fidelity:

$$f_{high}(x) = 6x^2 + \cos(12x) + \frac{1}{2}\sin(24x),$$

$$f_{medium}(x) = \frac{1}{2}f_{high}(x) + x^2 + 1,$$

and

$$f_{low}(x) = \left(x + \frac{1}{4}\right) f_{medium}(x) - x - 1.$$

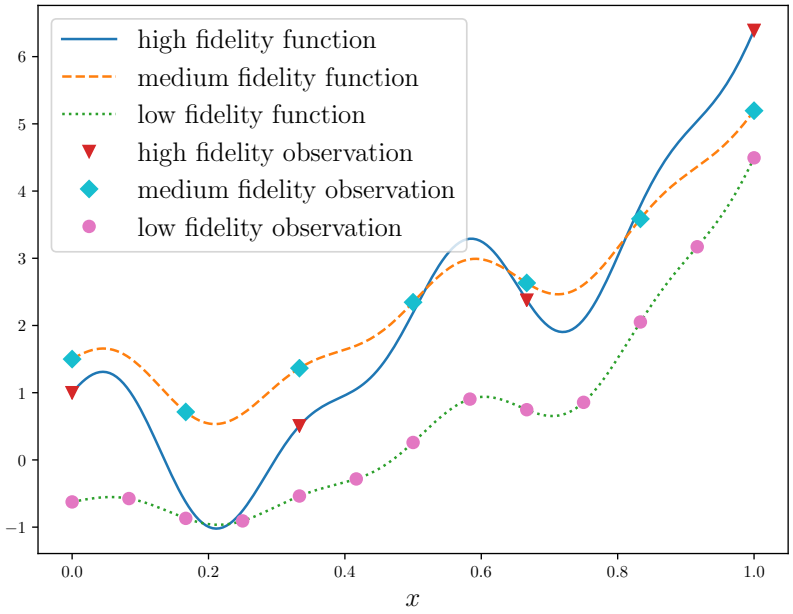



Figure 3.10: 3 levels of fidelity functions with 4 high fidelity, 7 medium fidelity and 13 low fidelity observations.

13 equispaced points in $[0, 1]$ were used for the low fidelity points, 7 for the medium level, and 4 for the high fidelity observations. A simple UK model, with $f_1^T(x) = f_2^T(x) = f_3^T(x) = g_{\rho_1}^T(x) = g_{\rho_2}^T(x) = 1$ is fitted to the data in Figure 3.11, and a slightly more complex one with $f_1^T(x) = f_2^T(x) = f_3^T(x) = g_{\rho_1}^T(x) = g_{\rho_2}^T(x) = (1, x)$, in Figure 3.12. Matérn 5/2 kernels were used for both. The average errors are more than halved in the second model compared to the simpler one.

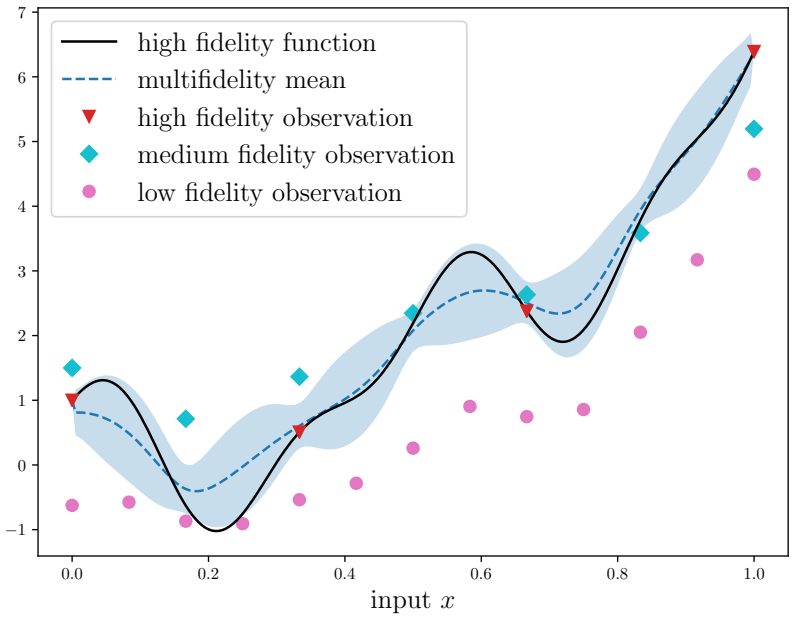



Figure 3.11: UK model 1 fitted to the data with 95% confidence intervals. Matérn 5/2 kernels were used for all levels, $f_1^T(x) = f_2^T(x) = f_3^T(x) = g_{\rho_1}^T(x) = g_{\rho_2}^T(x) = 1$. The RMSE on 200 equispaced points in $[0, 1]$ is 0.366 and the MAE is 0.286.

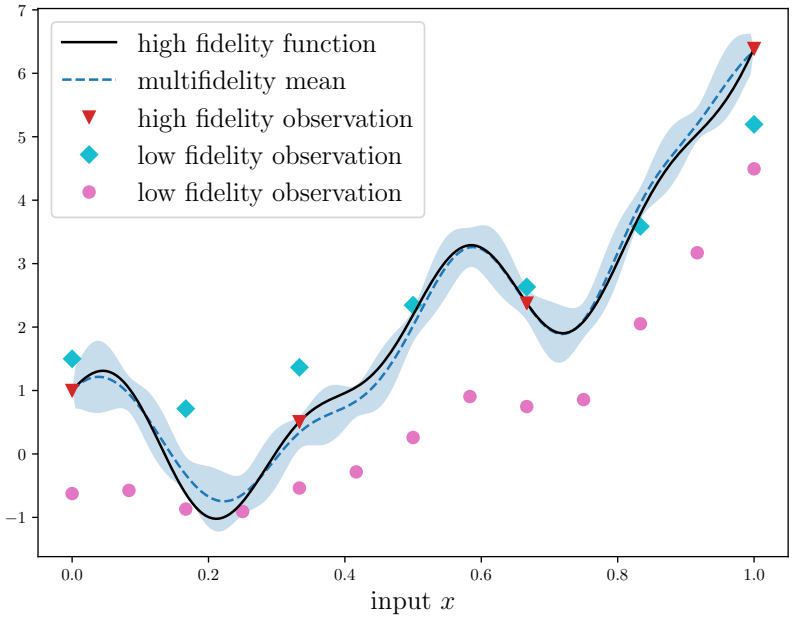



Figure 3.12: UK model 2 fitted to the data with 95% confidence intervals. Matérn 5/2 kernels were used for all levels, $f_1^T(x) = f_2^T(x) = f_3^T(x) = g_{\rho_1}^T(x) = g_{\rho_2}^T(x) = (1, x)$. The RMSE on 200 equispaced points in $[0, 1]$ is 0.155 and the MAE is 0.127.

### 3.4.3 2-dimensional input data and 2 levels of fidelity

For this two-dimensional example, we used the functions

$$f_{high}(x, y) = \cos(x^2 + y^2) + 0.5(\sin(xy) - x^2),$$

and

$$f_{low}(x, y) = (0.9 + 0.1x - 0.2y)f_{high}(x, y) + 0.5(xy + 0.5y + 1)$$

for low and high-fidelity. Observe the behavior of these two functions in the surface plots of Figures 3.14 and 3.15, and contour plots in Figure 3.16.

100 input points were uniformly selected in $[-2, 2] \times [2, 2]$, 30 of which were randomly selected for the high-fidelity data. The distribution of the input data points is shown in Figure 3.13.

We fit two models to the data, a simpler UK model, with $f_1^T(x, y) = f_2^T(x, y) = g_{\rho_1}^T(x, y) = 1$ and compare it to a Gaussian process with optimized mean in Figure 3.17, and a more complex UK model, with $f_1^T(x, y) = f_2^T(x, y) = g_{\rho_1}^T(x, y) = (1, x, y, xy)$, and compare it to the corresponding Gaussian process with prior mean $(1, x, y, xy)\eta$ with $\eta \in \mathbb{R}^4$ optimized via maximum likelihood (ML) in Figure 3.18. All kernels used are Matérn 5/2 and the Gaussian processes are fitted only to the high-fidelity data.

In Figures 3.19 and 3.20, we compare CV errors when using equispaced observations on a $10 \times 10$ grid in $[-2, 2] \times [-2, 2]$ for different models. Matérn 5/2 kernels are used for all models. For the UK models, observations are removed from the high-fidelity level only, and the Gaussian process has optimized constant mean and uses only the high-fidelity data.

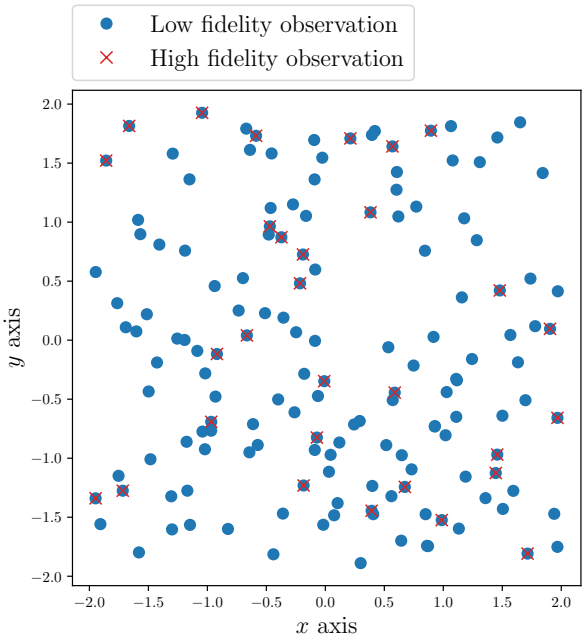



Figure 3.13: Location of the observations.

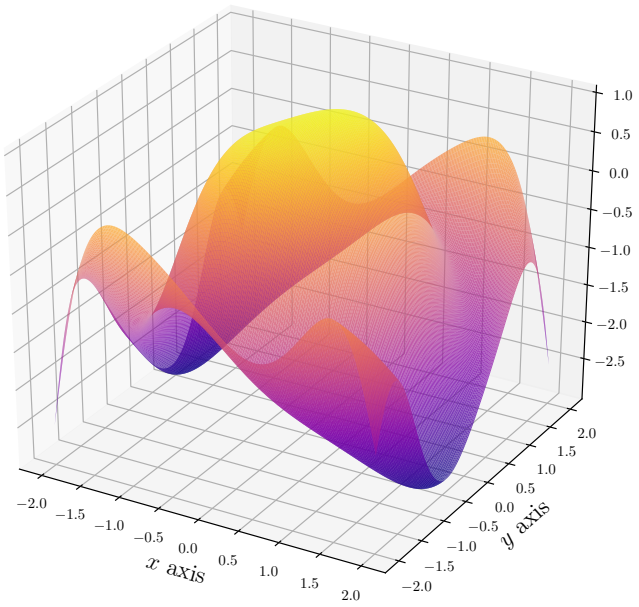



Figure 3.14: High fidelity function.

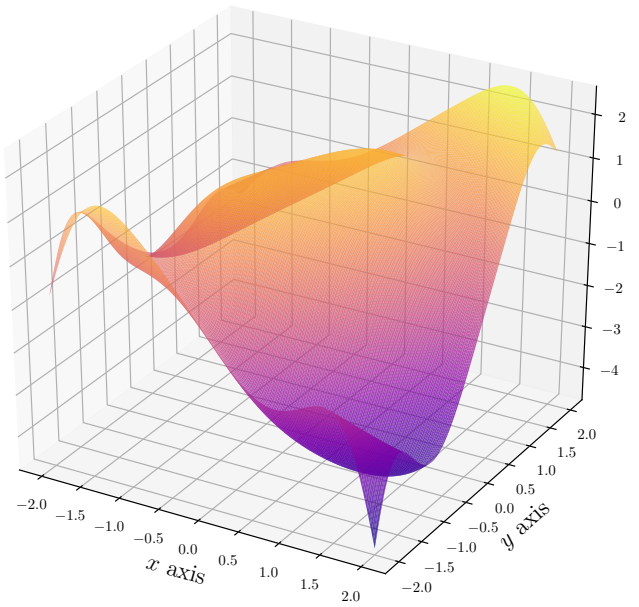



Figure 3.15: Low fidelity function.

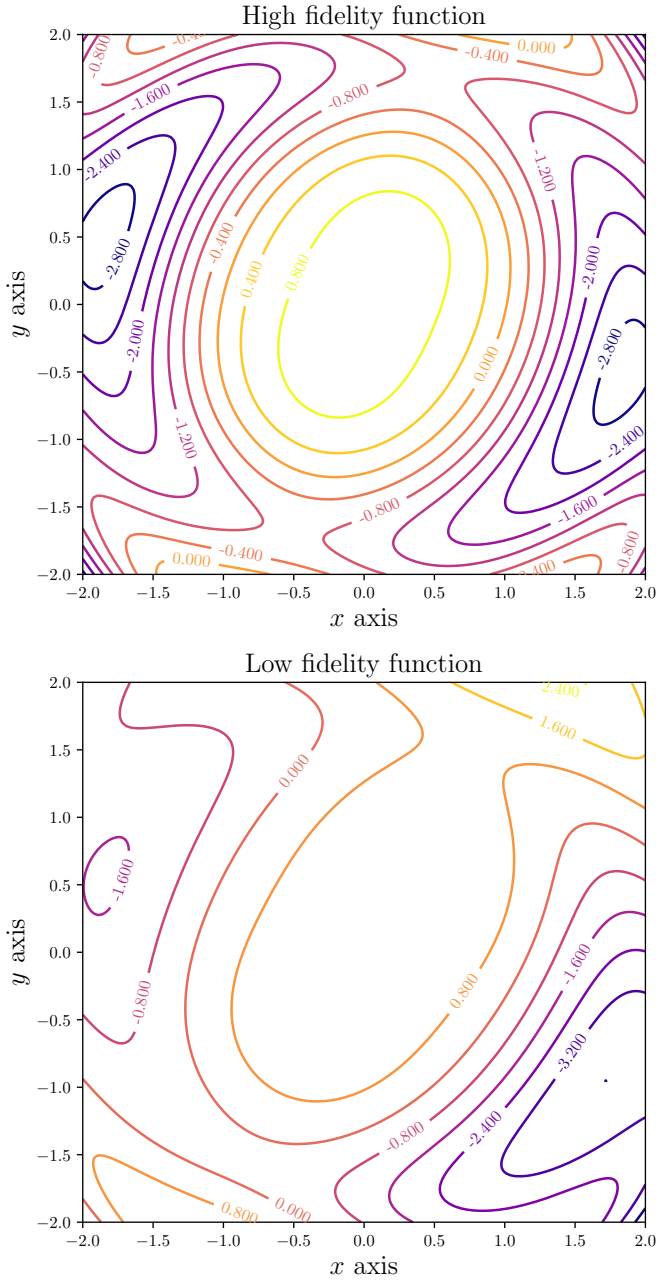



Figure 3.16: High and low fidelity functions with inputs in $[-2, 2] \times [-2, 2] \subset \mathbb{R}^2$ used for the multi-fidelity procedure.

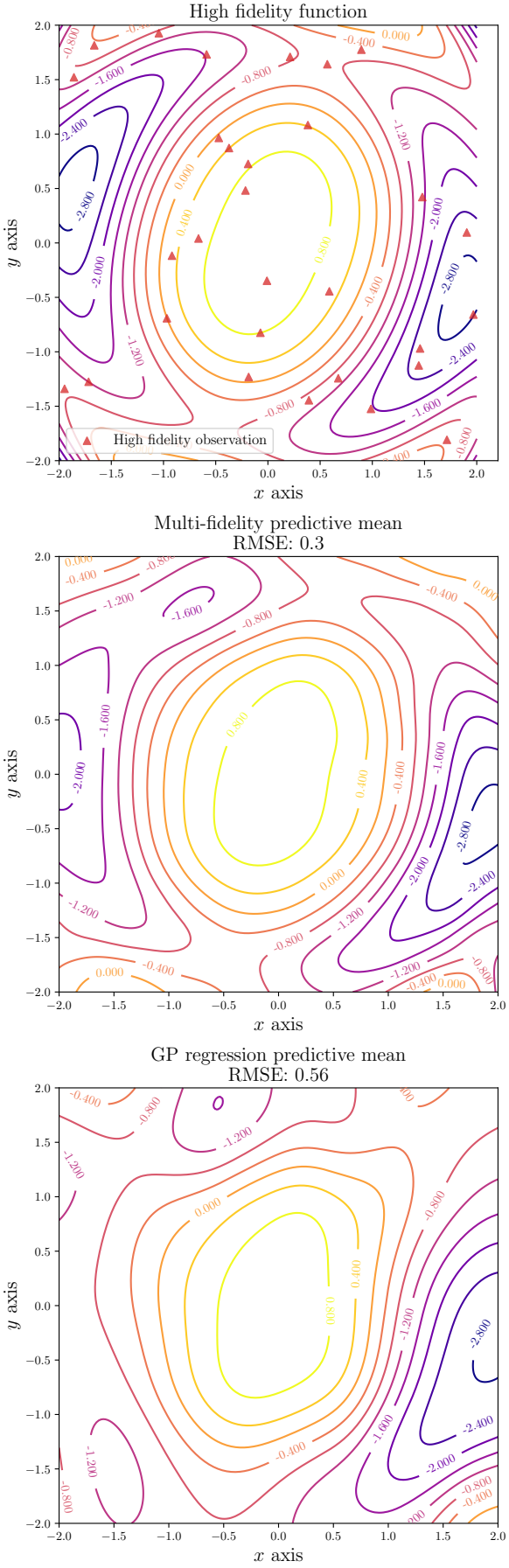



Figure 3.17: Comparative plots between the high fidelity function, multi-fidelity and Gaussian process regression. A UK multi-fidelity model with $f_1^T(x,y) = f_2^T(x,y) = f_3^T(x,y) = g_{\rho_1}^T(x,y) = g_{\rho_2}^T(x,y) = 1$ was fit to the data. The corresponding Gaussian process with prior mean $\eta$ with $\eta \in \mathbb{R}$ coefficient estimated via ML is fit in the third panel.

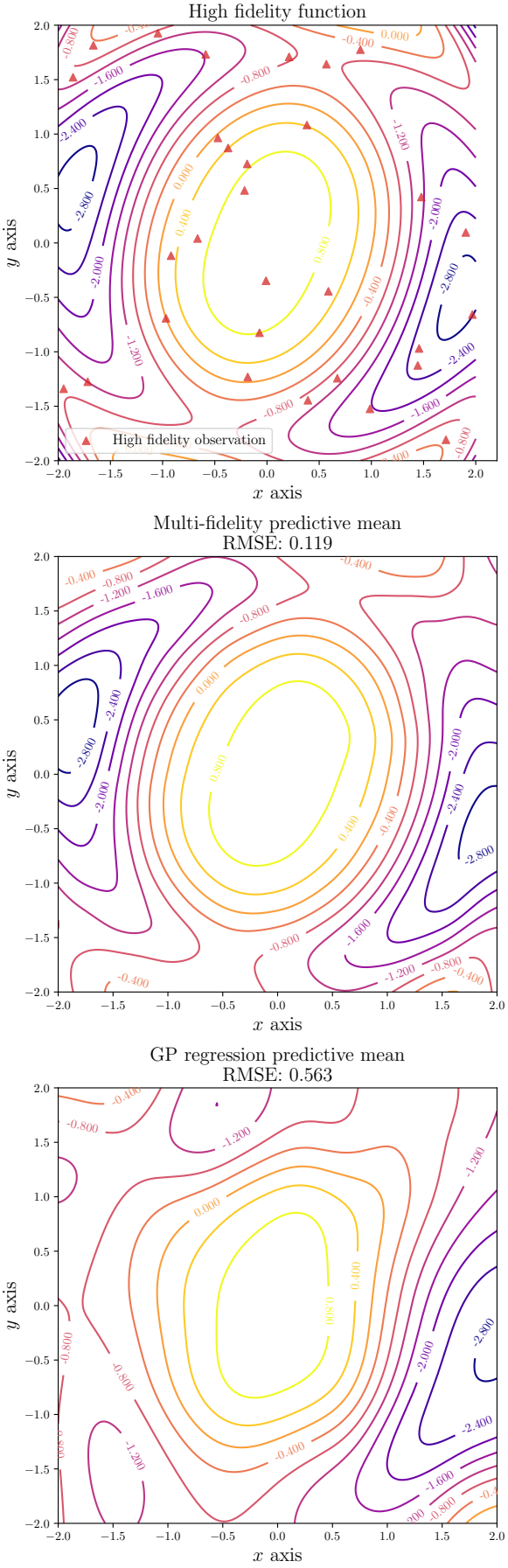



Figure 3.18: Comparative plots between the high fidelity function, multi-fidelity and Gaussian process regression. A UK multi-fidelity model with $f_1^T(x,y) = f_2^T(x,y) = g_{\rho_1}^T(x,y) = (1, x, y, xy)$ was fit to the data. The corresponding Gaussian process with prior mean $(1, x, y, xy)\eta$ with $\eta \in \mathbb{R}^4$ coefficients estimated via ML is fit in the third panel.

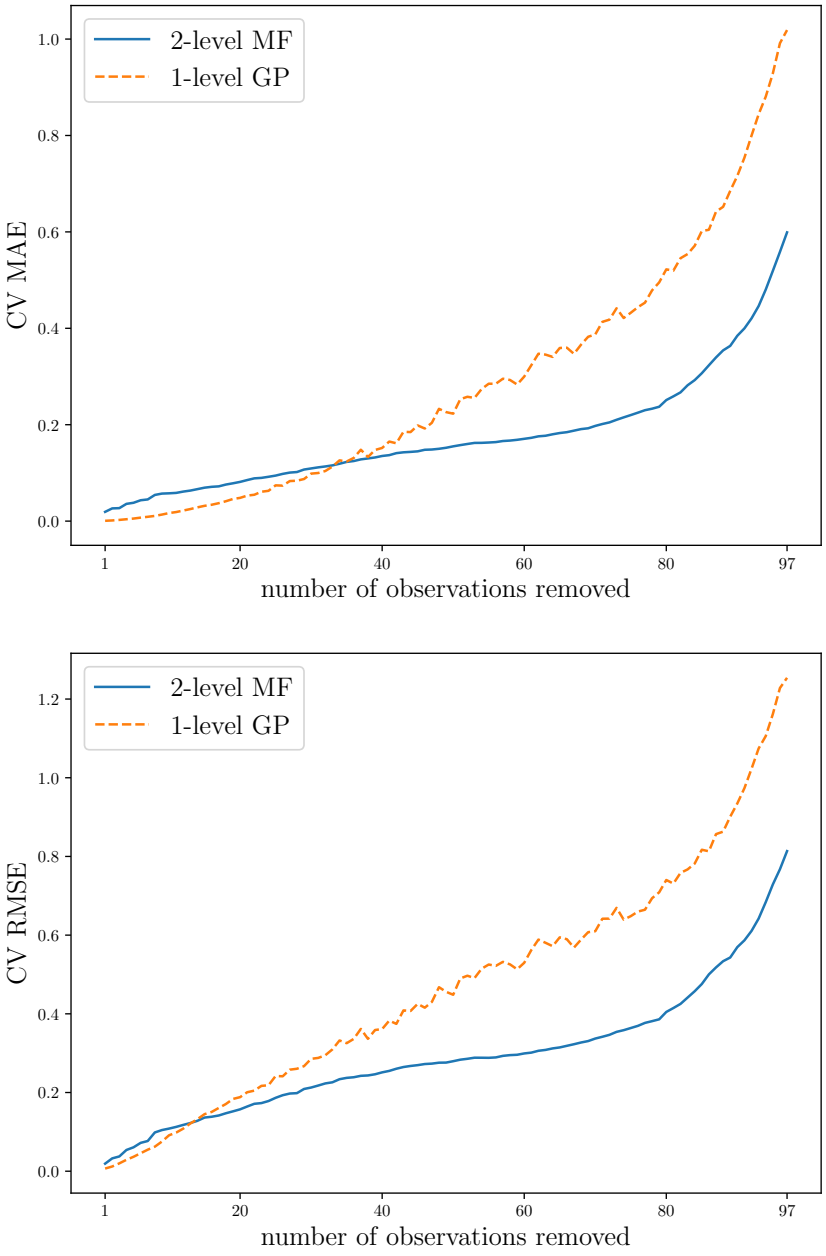



Figure 3.19: Comparison of average CV errors for a multi-fidelity model with $f_1^T(x,y) = f_2^T(x,y) = g_{\rho_1}^T(x,y) = 1$ and kernel Matérn $5/2$ and a Gaussian process with optimized constant mean and kernel Matérn $5/2$. In the multi-fidelity model, the observations are removed from the high-fidelity level only.

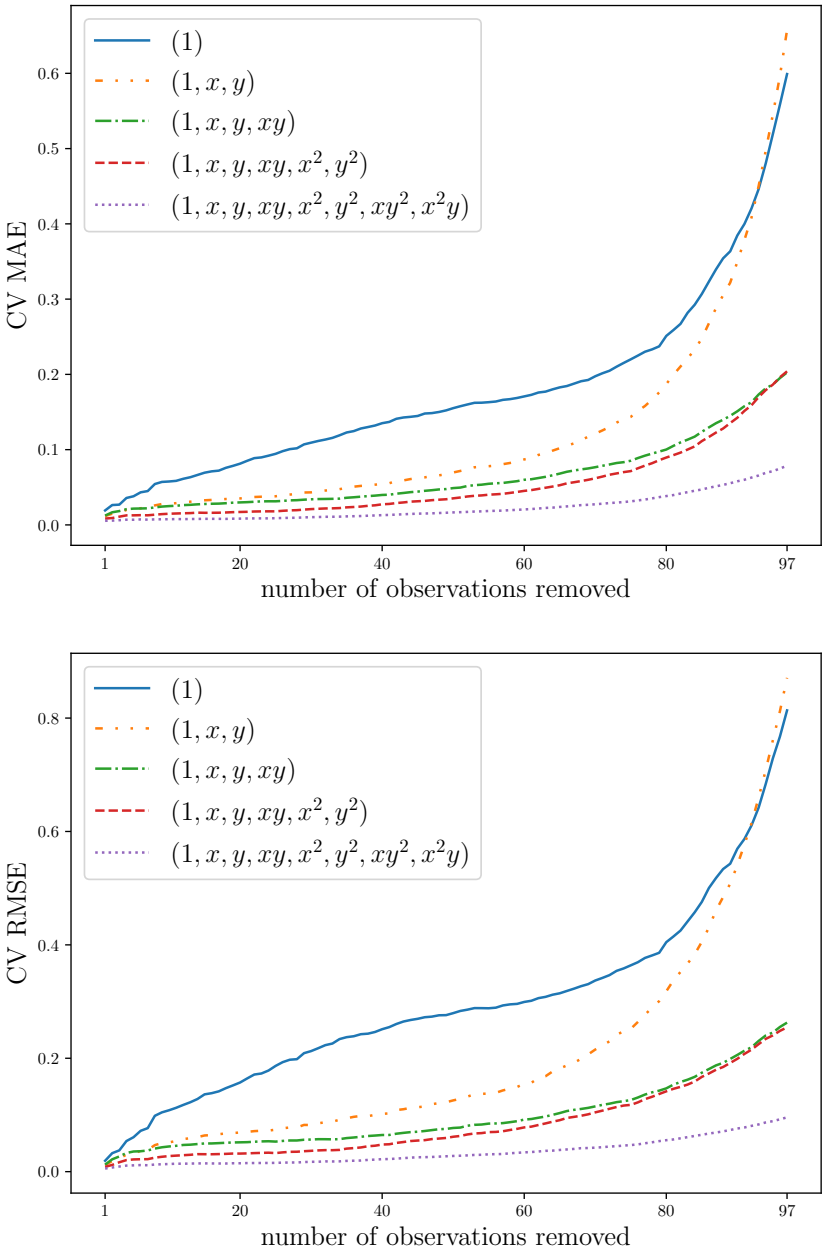



Figure 3.20: Comparison of average CV errors when removing observations from the high-level of fidelity for a multi-fidelity model with kernel Matérn 5/2, and different basis functions used for $f_1^T(x,y) = f_2^T(x,y) = g_{\rho_1}^T(x,y)$, which given in the legend.

# Chapter 4

# Conclusion and Perspectives

We present, in Chapter 3, the work of Le Gratiet in [Le Gratiet & Garnier '14] and [Le Gratiet '13] in establishing the equivalence of the recursive formulation of the co-kriging procedure and the classical co-kriging multi-fidelity model of Kennedy and O'Hagan proposed in [Kennedy & O'Hagan '98]. This new formulation provides a lower computational cost for the inference steps, since, for the Gaussian process prediction, the $s$-level model requires the inversion of $s$ matrices of sizes $n_t \times n_t$, with $t = 1, \ldots, s$, instead of the previous classical model, which required the inversion of a matrix of size $\sum_{t=1}^{s} n_t \times \sum_{t=1}^{s} n_t$. Furthermore, closed form expressions for cross-validation error and variance are available, which allow us to perform model selection via cross-validation at a lower computational cost than carrying out several necessary fittings.

In the examples of Section 3.4, we observed the fitting nature of multi-fidelity models. Models 1-6 of Section 3.4.1 show us how different kernels for the correlation function, and different trend and adjustment basis functions influence the predictions. Furthermore, in Figure 3.9 we note how the errors decrease for different kernels when we have more high-fidelity observations, and we compare them to a 1 level Gaussian process. It is evident that in the important case of few high-fidelity observations available, the multi-fidelity co-kriging models greatly surpass the power of Gaussian processes. Also, in Figure 3.20, we observe how increasing the number of basis functions for the trend and adjustment improves the fitting results for a more interesting case in 2 dimensions.

Thus, we see that the recursive co-kriging multi-fidelity model provides us a powerful tool for situations where the high-fidelity observations are scarse, yet we still have plenty of low-fidelity data.

In the future, we wish to further examine studies that are based on the recursive model presented in [Le Gratiet & Garnier '14]. One important extension of this model is the framework presented in [Perdikaris et al. '16] for high-dimensional input spaces and big data sets. In this paper, the authors use a graph-theoretic approach to encode the structure of the covariance matrix of the Gaussian processes priors of each level, and frequency domain machine learning algorithms to reduce the overall cost of the inference process.

Another work, presented in [Perdikaris et al. '17], generalized the recursive model with the classical autoregressive formulation to include nonlinearity by rewriting the expression

of $Z_t(x)$ as

$$Z_t(x) = f_{t-1}(Z_{t-1}(x) + \delta_t(x)),$$

where $f_{t-1}(x)$ is a function to be inferred from the data.

# Chapter 5

# Appendix

## A.1 Probability distributions

### A.1.1 Gaussian distribution

The multivariate normal (Gaussian) distribution of an $n$-dimensional random vector $X = (X_1, \ldots, X_n)^T$ is written following the notation

$$X \sim \mathcal{N}(\mu, \Sigma),$$

where $\mu$ denotes the mean vector and $\Sigma$ the covariance matrix of $X$. This distribution has density equal to

$$f(x) = \frac{\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}}{\sqrt{(2\pi)^n \det(\Sigma)}}.$$

### A.1.2 Inverse-gamma distribution

The inverse-gamma distribution of a positive random variable $X$ is written as

$$X \sim \mathcal{IG}(\alpha, \beta),$$

with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. Its density is given by

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left\{-\frac{\beta}{x}\right\}, \quad \text{for } x > 0.$$

$X \sim \mathcal{IG}(\alpha, \beta)$ has mean equal to $\frac{\beta}{\alpha-1}$, for $\alpha > 1$, and variance equal to $\frac{\beta^2}{(\alpha-1)(\alpha-2)}$, for $\alpha > 2$.

## A.2 Gaussian Identities

### A.2.1 Conditional probability

Let $x$ and $y$ be jointly Gaussian random vectors

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right). \tag{A.1}$$

Then, the conditional distribution of $x$ given $y$ is

$$x|y \sim \mathcal{N}(\mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}C^T). \tag{A.2}$$

### A.2.2 Product of Gaussian functions

A product of two Gaussian distributions is another (un-normalized) Gaussian (we write here $\mathcal{N}(x|m, \Sigma)$ for the density of the Gaussian distribution in $\mathbb{R}^D$ with mean $m$ and covariance $\Sigma$ at the point $x$):

$$\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) = Z^{-1}\mathcal{N}(x|c, C) \tag{A.3}$$

with

$$c = C(A^{-1}a + B^{-1}b)$$
$$C = (A^{-1} + B^{-1})^{-1}$$

and

$$Z^{-1} = (2\pi)^{-D/2} \det(A + B)^{-1/2} \exp\left\{ -\frac{1}{2}(a - b)^T(A + B)^{-1}(a - b) \right\}$$

## A.3 Probability identities

### A.3.1 Law of total expectation

Let $X$ and $Y$ be random variables such that $Y$ has finite mean. Then

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]. \tag{A.4}$$

### A.3.2 Law of total variance

If $X$ and $Y$ are arbitrary random variables for which the necessary expectations and variances are finite, then

$$\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathbb{E}[Y|X]]. \tag{A.5}$$

## A.4 Matrix identities

### A.4.1 Woodbury matrix identity

This identity, also known as *matrix inversion lemma*, states that

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{A.6}$$

when $A$ is an $n \times n$ invertible matrix, $C$ is an $m \times m$ invertible matrix, $U$ is $n \times m$, and $V$ is $m \times n$.

### A.4.2 Block matrix inversion

If a matrix $M$ is partitioned into four blocks, it's inverse can be partitioned blockwise as

$$M^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}, \tag{A.7}$$

where we assume that both $A$ and $D$ are invertible. Alternatively, we can write

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \tag{A.8}$$

### A.4.3 A particular block multiplication

Suppose that we have $M$, an invertible $m \times m$ matrix, partitioned as

$$M = \begin{bmatrix} M_1 & X \end{bmatrix},$$

where $M_1$ consists of the $m - n$ first columns of $M$ and $X$ of its last $n$ columns. Then

$$M^{-1}X = \begin{bmatrix} 0 \\ \mathbf{I}_n \end{bmatrix}. \tag{A.9}$$

This is derived simply by observing that if $Y = \begin{bmatrix} Y_1 \\ Z \end{bmatrix}$, with $Y_1$ of size $(m - n) \times n$ and $Z$ of size $n \times n$,

$$M^{-1}X = Y \iff X = MY = M_1Y_1 + XZ,$$

which is true if $Y_1 = 0$ and $Z = \mathbf{I}_n$. The result follows from the fact that $Y$ is unique.

Similarly, let

$$M = \begin{bmatrix} M_1 & X & M_2 \end{bmatrix},$$

with $M$ of size $m \times m$, $M_1$ of size $m \times m_1$, $M_2$ of size $m \times m_2$ and $X$ of size $m \times n$. If

$$Y = \begin{bmatrix} Y_1 \\ Z \\ Y_2 \end{bmatrix},$$

with $Y$ of size $m \times n$, $M_1$ of size $m_1 \times n$, $Y_2$ of size $m_2 \times n$ and $Z$ of size $n \times n$, then

$$M^{-1}X = Y \iff X = MY = M_1Y_1 + XZ + M_2Y_2 \iff Y_1 = 0,\ Z = \mathbf{I}_n \text{ and } Y_2 = 0. \tag{A.10}$$

## A.5 Proof of equations (2.10) and (2.11)

Using the Matrix inversion lemma (A.6), it is possible to rewrite the expressions for the mean and covariance

$$\bar{w}_* = H_*^T b + (K_*^T + H_*^T BH)(K_y + H^T BH)^{-1}(y - H^T b),$$

and

$$\text{Cov}[w_*] = K(X_*, X_*) + H_*^T BH_* - (K_*^T + H_*^T BH)^T (K_y + H^T BH)^{-1}(K_* + H^T BH_*),$$

to obtain more interpretable ones. Indeed, note that

$$(K_y + H^T BH)^{-1} = K_y^{-1} - K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}. \tag{A.11}$$

For the expression of the predictive covariance, we have that

$$\begin{gathered}\text{Cov}[\bar{w}_*] = K(X_*, X_*) + H_*^T BH_* - \\ (K_*^T + H_*^T BH)(K_y^{-1} - K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1})(K_* + H^T BH_*).\end{gathered} \tag{A.12}$$

Observe that

$$(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1} = (B^{-1} + HK_y^{-1}H^T)^{-1}(B^{-1} + HK_y^{-1}H^T) = I$$

therefore

$$\begin{gathered}(K_*^T + H_*^T BH)K_y^{-1}(K_* + H^T BH_*) = K_*^T K_y^{-1}K_* + K_*^T K_y^{-1}H^T BH_* + \\ H_*^T BHK_y^{-1}K_* + H_*^T BHK_y^{-1}H^T BH_* = \\ K_*^T K_y^{-1}K_* + K_*^T K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}(B^{-1} + HK_y^{-1}H^T)BH_* + \\ H_*^T B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}K_* + H_*^T BHK_y^{-1}H^T BH_*\end{gathered} \tag{A.13}$$

The last term in the previous expression can be rewritten as

$$\begin{gathered}H_*^T BHK_y^{-1}H^T BH_* = H_*^T B(B^{-1} + HK_y^{-1}H^T)BH_* - H_*^T BH_* = \\ H_*^T B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1}(B^{-1} + HK_y^{-1}H^T)BH_* - H_*^T BH_*.\end{gathered} \tag{A.14}$$

In addition to this, we have that

$$\begin{gathered}(K_*^T + H_*^T BH)(K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1})(K_* + H^T BH_*) = \\ K_*^T K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}K_* + K_*^T K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}H^T BH_* + \\ H_*^T BHK_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}K_* + \\ H_*^T BHK_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}H^T BH_*.\end{gathered} \tag{A.15}$$

Lastly, observe that

$$\begin{gathered}-H_*^T B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1}(B^{-1} + HK_y^{-1}H^T)BH_* + \\ H_*^T BHK_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}H^T BH_* = \\ -H_*^T(B^{-1} + HK_y^{-1}H^T)^{-1}H_* - 2H_*^T BH_*.\end{gathered} \tag{A.16}$$

Substituting all the terms of Equations (A.13)-(A.16) back in Equation (A.12) for $\text{Cov}[\bar{w}_*]$, we obtain

$$\text{Cov}[\bar{w}_*] = \text{Cov}[\bar{z}_*] + R^T(B^{-1} + HK_y^{-1}H^T)^{-1}R,$$

with

$$\text{Cov}[\bar{z}_*] = K(X_*, X_*) - K_*^T K_y^{-1}K_*.$$

For the predictive mean, using Equation (A.11), we have that

$$\begin{gathered}\bar{w}_* = H_*^T b + (K_*^T + H_*^T BH)(K_y^{-1} - K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1})(y - H^T b) = \\ H_*^T b + (K_*^T + H_*^T BH)K_y^{-1}(y - H^T b) - \\ (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}(y - H^T b) = \\ K_*^T K_y^{-1} y + \left(H_*^T B - (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}\right)HK_y^{-1}y + \\ \left(H_*^T - (K_*^T + H_*^T BH)K_y^{-1}H^T + (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}H^T\right)b\end{gathered} \tag{A.17}$$

We can now identify the term $K_*^T K_y^{-1} y = \bar{z}_*$. For the second term containing $y$, the second term of the fourth row of the last equation, note that

$$\begin{gathered}H_*^T B - (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1} = \\ \left(H_*^T B(B^{-1} + HK_y^{-1}H^T) - (K_*^T + H_*^T BH)K_y^{-1}H^T\right)(B^{-1} + HK_y^{-1}H^T)^{-1} = \\ R^T(B^{-1} + HK_y^{-1}H^T)^{-1},\end{gathered} \tag{A.18}$$

and for the term with $b$,

$$\begin{gathered}\left(H_*^T - (K_*^T + H_*^T BH)K_y^{-1}H^T + (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}H^T\right) = \\ \left(H_* B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1}B^{-1}\right) - \left((K_*^T + H_*^T BH)K_y^{-1}H^T\right) + \\ \left((K_*^T + H_*^T BH)K_y^{-1}H^T - (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}B^{-1}\right) = \\ R^T(B^{-1} + HK_y^{-1}H^T)B^{-1}.\end{gathered} \tag{A.19}$$

Using all these rewritten Equations (A.18) and (A.19) in Equation (A.17), we obtain

$$\bar{w}_* = \bar{z}_* + R^T\bar{\beta}.$$

## A.6 Requisites for the proof of Proposition 3.1 of Section 3.3

We leave all dependences on hyperparameters implicit for the sake of notation.

**Proposition A.1** (Proposition 3.1 of [Le Gratiet '13]). *If we consider the covariance matrix $V_s$ in (3.10) and sort the experimental design arranged so that, for $t = 2, \ldots, s$, first come the points that are in $D_{t-1}$ but not in $D_t$, and then the points in $D_t$, $(D_{t-1} \backslash D_t, D_t)$, then the inverse of $V_s$ has the form*

$$V_s^{-1} = \begin{bmatrix} V_{s-1}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{(\rho_{s-1}(D_s)\rho_{s-1}^T(D_s))\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} & -\begin{bmatrix} 0 \\ \frac{(\rho_{s-1}(D_s)\mathbf{1}_{n_s}^T)\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} \\ -\begin{bmatrix} 0 & \frac{(\mathbf{1}_{n_s}\rho_{s-1}^T(D_s))\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} & \frac{R_s^{-1}}{\sigma_s^2} \end{bmatrix}, \tag{A.20}$$

*and*

$$V_1^{-1} = \frac{R_1^{-1}}{\sigma_1^2},$$

*where* $V_{s-1}$ *is a* $(\sum_{i=1}^{s-1} n_i \times \sum_{i=1}^{s-1} n_i)$ *matrix, and* $R_s = [r_s(x,x')]_{x,x' \in D_s}$ *an* $(n_s \times n_s)$ *matrix.*

*Proof.* Let

$$V_s = \begin{bmatrix} V_{s-1} & U_{s-1} \\ U_{s-1}^T & V_{s,s} \end{bmatrix}, \quad \text{with} \quad U_{s-1} = \begin{bmatrix} V_{1,s} \\ \vdots \\ V_{s-1,s} \end{bmatrix} = \text{Cov}\{\mathcal{Z}^{(s-1)}, \mathcal{Z}_s\},$$

where $V_{s-1} = \text{Cov}\{\mathcal{Z}_{s-1}, \mathcal{Z}_{s-1}\}$ and $V_{t,s} = \text{Cov}\{\mathcal{Z}_t, \mathcal{Z}_s\}$.

Using (A.7), we can write the inverse of $V_s$ as

$$\begin{bmatrix} V_{s-1} & U_{s-1} \\ U_{s-1}^T & V_{s,s} \end{bmatrix}^{-1} = \begin{bmatrix} V_{s-1}^{-1} + V_{s-1}^{-1} U_{s-1} Q_s^{-1} U_{s-1}^T V_{s-1}^{-1} & -V_{s-1}^{-1} U_{s-1} Q_s^{-1} \\ -Q_s^{-1} U_{s-1}^T V_{s-1}^T & Q_s^{-1} \end{bmatrix}$$

where $Q_s = V_{s,s} - U_{s-1}^T V_{s-1}^{-1} U_{s-1}$. From (3.11), we know that for $t < s$,

$$V_{t,s} = [\mathbf{1}_{n_t} \rho_{s-1}^T(D_s)] \odot V_{t,s-1}(D_t, D_s)$$

$$\implies U_{s-1} = \begin{bmatrix} V_{1,s} \\ \vdots \\ V_{s-1,s} \end{bmatrix} = [\mathbf{1}_{\sum_{i=1}^{s-1} n_i} \rho_{s-1}^T(D_s)] \odot \begin{bmatrix} V_{1,s-1}(D_1, D_s) \\ \vdots \\ V_{s-1,s-1}(D_{s-1}, D_s) \end{bmatrix}.$$

Note that the $n_s$ last columns of $V_{s-1}$ are precisely

$$\begin{bmatrix} V_{1,s-1}(D_1, D_s) \\ \vdots \\ V_{s-1,s-1}(D_{s-1}, D_s) \end{bmatrix}.$$

By (A.9) and the fact that the Hadamard product is between a matrix with all identical rows and the one made of the $n_s$ last columns of $V_{s-1}$, we obtain

$$V_{s-1}^{-1} U_{s-1} = V_{s-1}^{-1} [\mathbf{1}_{\sum_{i=1}^{s-1} n_i} \rho_{s-1}^T(D_s)] \odot \begin{bmatrix} V_{1,s-1}(D_1, D_s) \\ \vdots \\ V_{s-1,s-1}(D_{s-1}, D_s) \end{bmatrix} =$$

$$[\mathbf{1}_{\sum_{i=1}^{s-1} n_i} \rho_{s-1}^T(D_s)] \odot \begin{bmatrix} 0 \\ \mathbf{I}_{n_s} \end{bmatrix},$$

with the 0 in the last equality being a $(\sum_{i=1}^{s-1} n_i - n_s) \times n_s$ matrix with all entries equal to 0.

Now we can rewrite $Q_s$ as something more familiar:

$$Q_s = V_{s,s} + U_{s-1}^T V_{s-1}^{-1} U_{s-1} =$$

$$= \text{Cov}\{\mathcal{Z}_s, \mathcal{Z}_s\} - \text{Cov}\{\mathcal{Z}^{s-1}, \mathcal{Z}_s\}^T \text{Var}[\mathcal{Z}^{(s-1)}] \text{Cov}\{\mathcal{Z}^{(s-1)}, \mathcal{Z}_s\}.$$

And this is exactly the predictive variance of $\mathcal{Z}_s$ conditioned by $\mathcal{Z}^{(s-1)}$. In addition to this,

$$\mathcal{Z}_s = Z_s(D_s) = \rho_{s-1}(D_s) \odot Z_{s-1}(D_s) + \delta_s(D_s)$$
$$\implies \text{Var}[\mathcal{Z}_s | \mathcal{Z}^{(s-1)}] = \text{Var}[\rho_{s-1}(D_s) \odot Z_{s-1}(D_s) + \delta_s(D_s) | \mathcal{Z}^{(s-1)}] =$$
$$= \text{Var}[\delta_s(D_s) | \mathcal{Z}^{(s-1)}] = \text{Var}[\delta_s(D_s)] = \sigma_s^2 R_s,$$

since $Z_{s-1}(D_s)$ is a constant when conditioned by $\mathcal{Z}^{(s-1)}$ and $\delta_s(x)$ is independent of $\mathcal{Z}^{(s-1)}$.

Having expressions for $V_{s-1}^{-1} U_{s-1}$ and $Q_s$, it becomes easier to construct the matrix $V_s^{-1}$. Note that

$$V_{s-1}^{-1} U_{s-1} Q_s^{-1} = \left( [\mathbf{1}_{\sum_{i=1}^{s-1} n_i} \rho_{s-1}^T(D_s)] \odot \begin{bmatrix} \mathbf{0} \\ \mathbf{I}_{n_s}, \end{bmatrix} \right) \frac{R_s^{-1}}{\sigma_s^2} =$$

$$\begin{bmatrix} \mathbf{0}_{(\sum_{i=1}^{s-1} n_i - n_s) \times n_s} \\ \left( [\mathbf{1}_{n_s} \rho_{s-1}^T(D_s)] \odot \mathbf{I}_{n_s} \right) \frac{R_s^{-1}}{\sigma_s^2} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_{(\sum_{i=1}^{s-1} n_i - n_s) \times n_s} \\ [\rho_{s-1}(D_s) \mathbf{1}_{n_s}^T] \odot \frac{R_s^{-1}}{\sigma_s^2} \end{bmatrix},$$

and this implies that

$$V_{s-1}^{-1} U_{s-1} Q_s^{-1} U_{s-1}^T V_{s-1}^{-1} = \begin{bmatrix} \mathbf{0}_{(\sum_{i=1}^{s-1} n_i - n_s) \times n_s} \\ [\rho_{s-1}(D_s) \mathbf{1}_{n_s}^T] \odot \frac{R_s^{-1}}{\sigma_s^2} \end{bmatrix} \begin{bmatrix} \mathbf{0}_{n_s \times (\sum_{i=1}^{s-1} n_i - n_s)} & [\rho_{s-1}(D_s) \mathbf{1}_{n_s}^T] \odot \mathbf{I}_{n_s} \end{bmatrix} =$$

$$= \begin{bmatrix} \mathbf{0}_{(\sum_{i=1}^{s-1} n_i - n_s) \times (\sum_{i=1}^{s-1} n_i - n_s)} & \mathbf{0}_{(\sum_{i=1}^{s-1} n_i - n_s) \times n_s} \\ \mathbf{0}_{n_s \times (\sum_{i=1}^{s-1} n_i - n_s)} & \left[ \frac{(\rho_{s-1}(D_s) \rho_{s-1}^T(D_s)) \odot R_s^{-1}}{\sigma_s^2} \right] \end{bmatrix}.$$

Therefore, using everything we constructed, we obtain a recursive form for $V_s^{-1}$:

$$V_s^{-1} = \begin{bmatrix} W_{1,1} & W_{1,2} \\ W_{1,2}^T & W_{2,2} \end{bmatrix}$$

with

$$W_{1,1} = \left[ V_{s-1}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \left[ \frac{(\rho_{s-1}(D_s) \rho_{s-1}^T(D_s)) \odot R_s^{-1}}{\sigma_s^2} \right] \end{bmatrix} \right],$$

$$W_{1,2} = - \begin{bmatrix} 0 \\ \frac{[\rho_{s-1}(D_s) \mathbf{1}_{n_s}^T] \odot R_s^{-1}}{\sigma_s^2} \end{bmatrix},$$

$$W_{1,2}^T = - \begin{bmatrix} 0 & \frac{[\mathbf{1}_{n_s} \rho_{s-1}^T(D_s)] \odot R_s^{-1}}{\sigma_s^2} \end{bmatrix},$$

$$W_{2,2} = \frac{R_s^{-1}}{\sigma_s^2},$$

and $V_1^{-1} = \frac{R_1^{-1}}{\sigma_1^2}$.

□

**Proposition A.2** (Proposition 3.2 of [Le Gratiet '13]). *If $V_s$ is the covariance matrix in equation (3.10) and $k_s^T(x)$ the covariance vector in equation (3.8), the following equality is valid:*

$$k_s^T(x)V_s^{-1} = \big(\rho_{s-1}(x)k_{s-1}^T(x)V_{s-1}^{-1} - (0, \quad [\rho_{s-1}^T(D_s) \odot r_s^T(x, D_s)]R_s^{-1}), \quad r_s^T(x, D_s)R_s^{-1}\big).$$

*Proof.* In (3.8) and (3.9), we obtained a recursive expression for $k_s^T(x)$:

$$k_s^T(x) = \mathrm{Cov}\{Z_s(x), \mathcal{Z}^{(s)}\} = (c_1^T(x, D_1), \ldots, c_s^T(x, D_s))^T,$$

with

$$c_t^T(x, D_t) = \mathrm{Cov}\{Z_s(x), Z_t(D_t)\}$$

$$\implies c_t^T(x, D_t) = \rho_{t-1}(D_t) \odot c_{t-1}^T(x, D_t) + \left(\prod_{i=t}^{s-1} \rho_i(x)\right)\sigma_t^2 r_t^T(x, D_t). \tag{A.21}$$

By Proposition A.1,

$$V_s^{-1} = \begin{bmatrix} V_{s-1}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{(\rho_{s-1}(D_s)\rho_{s-1}^T(D_s))\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} & - \begin{bmatrix} 0 \\ \frac{(\rho_{s-1}(D_s)\mathbf{1}_{n_s}^T)\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} \\ - \begin{bmatrix} 0 & \frac{[\mathbf{1}_{n_s}\rho_{s-1}^T(D_s)]\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} & \frac{R_s^{-1}}{\sigma_s^2} \end{bmatrix},$$

which we will split as $V_s^{-1} = \begin{bmatrix} A & B \end{bmatrix}$ with

$$A = \begin{bmatrix} V_{s-1}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{(\rho_{s-1}(D_s)\rho_{s-1}^T(D_s))\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} \\ - \begin{bmatrix} 0 & \frac{[\mathbf{1}_{n_s}\rho_{s-1}^T(D_s)]\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} - \begin{bmatrix} 0 \\ \frac{(\rho_{s-1}(D_s)\mathbf{1}_{n_s}^T)\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} \\ \frac{R_s^{-1}}{\sigma_s^2} \end{bmatrix}.$$

This implies that

$$k_s^T(x)V_s^{-1} = \begin{bmatrix} k_s^T(x)A & k_s^T(x)B \end{bmatrix}.$$

*For A:*

$$k_s^T(x)A = (c_1^T(x, D_1), \ldots, c_{s-1}^T(x, D_{s-1})) \left[V_{s-1}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{(\rho_{s-1}(D_s)\rho_{s-1}^T(D_s))\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix}\right] -$$

$$c_s^T(x, D_s) \begin{bmatrix} 0 & \frac{[\mathbf{1}_{n_s}\rho_{s-1}^T(D_s)]\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix}$$

Note that Equation (3.6) implies that, for $1 \leq t \leq s-1$,

$$c_t(x, D_t) = \mathrm{Cov}\{Z_s(x), Z_t(D_t)\} = \rho_{s-1}(x)\mathrm{Cov}\{Z_{s-1}(x), Z_t(D_t)\}$$

$$\implies (c_1^T(x, D_1), \ldots, c_{s-1}^T(x, D_{s-1}))^T = \rho_{s-1}(x)k_{s-1}^T(x) = \rho_{s-1}(x)\mathrm{Cov}\{Z_{s-1}(x), \mathcal{Z}^{(s-1)}\}. \tag{A.22}$$

As in Proposition A.1, the points in the sets $D_{t-1}$ are ordered such that first come the points in $D_{t-1}\backslash D_t$ and after the ones in $D_t$. This ordering helps us manage many expressions we come across. Therefore,

$$c_{s-1}^T(x, D_{s-1}) = (c_{s-1}^T(x, D_{s-1}\backslash D_s), \quad c_{s-1}^T(x, D_s)).$$

and with these last expressions we obtain

$$k_s^T(x)A = \rho_{s-1}(x)k_{s-1}^T(x)V_{s-1}^{-1} + \left(0, \quad c_{s-1}^T(x, D_s)\frac{(\rho_{s-1}(D_s)\rho_{s-1}^T(D_s)) \odot R_s^{-1}}{\sigma_s^2}\right)$$

$$-c_s^T(x, D_s)\left[0 \quad \frac{[\mathbf{1}_{n_s}\rho_{s-1}^T(D_s)]\odot R_s^{-1}}{\sigma_s^2}\right].$$

Also, by Equation (A.21), we know that

$$c_s^T(x, D_s) = \rho_{s-1}(D_s) \odot c_{s-1}^T(x, D_s) + \sigma_s^2 r_s^T(x, D_s) \tag{A.23}$$

$$\implies c_s^T(x, D_s)\left[0 \quad \frac{[\mathbf{1}_{n_s}\rho_{s-1}^T(D_s)]\odot R_s^{-1}}{\sigma_s^2}\right] =$$

$$\left(0, \quad c_{s-1}^T(x, D_s)\frac{[\rho_{s-1}(D_s)\rho_{s-1}^T(D_s)] \odot R_s^{-1}}{\sigma_s^2} + [\rho_{s-1}^T(D_s) \odot r_s^T(x, D_s)]R_s^{-1}\right)$$

$$\implies k_s^T(x)A = \rho_{s-1}(x)k_{s-1}^T(x)V_{s-1}^{-1} - (\mathbf{0}_{1\times(\sum_{i=1}^{s-1} n_i - n_s)}, \quad [\rho_{s-1}^T(D_s) \odot r_s^T(x, D_s)]R_s^{-1}).$$

*For B:* We'll use the identities already obtained in the previous part of the proof.

$$k_s^T(x)B = -(c_1^T(x, D_1), \ldots, c_{s-1}^T(x, D_{s-1}))\begin{bmatrix} 0 \\ \frac{(\rho_{s-1}(D_s)\mathbf{1}_{n_s}^T)\odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} + c_s^T(x, D_s)\frac{R_s^{-1}}{\sigma_s^2} =$$

$$-c_{s-1}^T(x, D_s)\frac{(\rho_{s-1}(D_s)\mathbf{1}_{n_s}^T) \odot R_s^{-1}}{\sigma_s^2} + (\rho_{s-1}(D_s) \odot c_{s-1}^T(x, D_s) + \sigma_s^2 r_s^T(x, D_s))\frac{R_s^{-1}}{\sigma_s^2} =$$

$$= r_s^T(x, D_s)R_s^{-1}.$$

*And all together...*

$$k_s^T(x)V_s^{-1} = \big(\rho_{s-1}(x)k_{s-1}^T(x)V_{s-1}^{-1} - (0, \quad [\rho_{s-1}^T(D_s) \odot r_s^T(x, D_s)]R_s^{-1}), \quad r_s^T(x, D_s)R_s^{-1}\big).$$

□

## A.7 Parameter estimation of subsection 3.3.1

We'll use the approach presented in [Hoff '09] for the problem of finding the posterior of two parameters $\theta$ and $\gamma$ when their prior distributions are of the form $p(\theta|\gamma)$ and $p(\gamma)$. If we call the data $X$, we observe that the joint posterior distribution can be decomposed as

$$p(\theta, \gamma|X) = p(\theta|\gamma, X)p(\gamma|X).$$

Then, the posterior distribution $p(\theta|X, \gamma)$ is obtained by noting that

$$p(\theta|X, \gamma) = \frac{p(\theta|\gamma)p(X|\theta, \gamma)}{p(X|\gamma)} \propto p(\theta|\gamma)p(X|\theta, \gamma).$$

Next, the posterior distribution of $\gamma$ is given by a marginalization:

$$p(\gamma|X) = \frac{p(\gamma)p(X|\gamma)}{p(X)} \propto p(\gamma)p(X|\gamma) = p(\gamma)\int p(X|\theta, \gamma)p(\theta|\gamma)d\theta.$$

*First considerations:* Again, the obvious dependencies are left implicit.

We know that

$$Z_1(D_1) = \delta_1(D_1) \sim \mathcal{N}(F_1\beta_1, \sigma_1^2 R_1),$$

and that

$$Z_t(D_t) = \rho_{t-1}(D_t) \odot \widetilde{Z}_{t-1}(D_t) + \delta_t(D_t) = [G_{t-1}\beta_{\rho_{t-1}}] \odot \widetilde{Z}_{t-1}(D_t) + \delta_t(D_t) =$$

$$[G_{t-1} \odot [\widetilde{Z}_{t-1}(D_t)\mathbf{1}_{q_{t-1}}^T]\beta_{\rho_{t-1}} + \delta_t(D_t)$$

$$\sim \mathcal{N}(G_{t-1} \odot [z_{t-1}(D_t)\mathbf{1}_{q_{t-1}}^T]\beta_{\rho_{t-1}} + F_t\beta_t, \sigma_t^2 R_t)$$

(see Remarks 1 and 2 for clarification).

Let $\mathcal{H}_1 = F_1$ and $\mathcal{H}_t = \begin{bmatrix} G_{t-1} \odot [z_{t-1}(D_t)\mathbf{1}_{q_{t-1}}^T & F_t \end{bmatrix}$, for $t > 1$. Also, for simplicity, $\tilde{\beta}_t = \begin{bmatrix} \beta_{\rho_{t-1}} \\ \beta_t \end{bmatrix}$, for $t > 1$, and $\tilde{\beta}_1 = \beta_1$. Then, we can rewrite the previous expressions simply as

$$Z_t(D_t)|z^{(t-1)}, \tilde{\beta}_t, \sigma_t^2 \sim \mathcal{N}(\mathcal{H}_t\tilde{\beta}_t, \sigma_t^2 R_t),$$

for $t = 1, \ldots, s$, with the convention $z^{(0)} = \emptyset$.

Now we can construct the likelihood equations, for our observations $z_t$ for $t = 1, \ldots, s$:

$$p(z_t|z^{(t-1)}, \tilde{\beta}_t, \sigma_t^2) = \frac{1}{(2\pi)^{n_t/2}} \frac{1}{\sqrt{\det(\sigma_t^2 R_t)}} \exp\left\{ -\frac{1}{2}(z_t - \mathcal{H}_t\tilde{\beta}_t)^T \frac{R_t^{-1}}{\sigma_t^2}(z_t - \mathcal{H}_t\tilde{\beta}_t) \right\}$$

*All priors non-informative (ii):*

Note that

$$(z_t - \mathcal{H}_t\tilde{\beta}_t)^T \frac{R_t^{-1}}{\sigma_t^2}(z_t - \mathcal{H}_t\tilde{\beta}_t) =$$

$$z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t + \tilde{\beta}_t^T (\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t^2} \mathcal{H}_t) \tilde{\beta}_t - \tilde{\beta}_t^T \mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - z_t^T \frac{R_t^{-1}}{\sigma_t^2} \mathcal{H}_t \tilde{\beta}_t =$$

$$(\tilde{\beta}_t - \Sigma_t \nu_t)^T \Sigma_t^{-1} (\tilde{\beta}_t - \Sigma_t \nu_t) + z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t,$$

where $\Sigma_t = \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t^2} \mathcal{H}_t\right]^{-1}$ and $\nu_t = \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t\right]$. Therefore,

$$p(z_t | z^{(t-1)}, \tilde{\beta}_t, \sigma_t^2) =$$

$$\frac{1}{(2\pi)^{n_t/2}} \frac{1}{\sqrt{\det(\sigma_t^2 R_t)}} \exp\left\{ -\frac{1}{2} \left( (\tilde{\beta}_t - \Sigma_t \nu_t)^T \Sigma_t^{-1} (\tilde{\beta}_t - \Sigma_t \nu_t) + z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t \right) \right\}. \quad \text{(A.24)}$$

Since $z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t$ is constant with respect to $\tilde{\beta}_t$,

$$p(\tilde{\beta}_t | z^{(t)}, \sigma_t^2) \propto p(\tilde{\beta}_t | z^{(t-1)}, \sigma_t^2) p(z_t | z^{(t-1)} \tilde{\beta}_t, \sigma_t^2) \propto \exp\left\{ -\frac{1}{2} (\tilde{\beta}_t - \Sigma_t \nu_t)^T \Sigma_t^{-1} (\tilde{\beta}_t - \Sigma_t \nu_t) \right\}$$

$$\implies [\tilde{\beta}_t | z^{(t)}, \sigma_t^2] \sim \mathcal{N}(\Sigma_t \nu_t, \Sigma_t).$$

For the posterior of $\sigma_t^2$, we know that

$$p(\sigma_t^2 | z^{(t)}) \propto p(\sigma_t^2 | z^{(t-1)}) \int p(z_t | z^{(t-1)}, \tilde{\beta}_t, \sigma_t^2) p(\tilde{\beta}_t | z^{(t-1)}, \sigma_t^2) d\tilde{\beta}_t \propto$$

$$\frac{1}{\sigma_t^2} \int \frac{1}{(2\pi)^{n_t/2}} \frac{1}{\sqrt{\det(\sigma_t^2 R_t)}} \exp\left\{ -\frac{1}{2} \left( (\tilde{\beta}_t - \Sigma_t \nu_t)^T \Sigma_t^{-1} (\tilde{\beta}_t - \Sigma_t \nu_t) + z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t \right) \right\} 1 d\tilde{\beta}_t \propto$$

$$\frac{1}{\sigma_t^2} \frac{1}{(\sigma_t^2)^{n_t/2}} \exp\left\{ -\frac{1}{2} \left( z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t \right) \right\} \int \exp\left\{ -\frac{1}{2} \left( (\tilde{\beta}_t - \Sigma_t \nu_t)^T \Sigma_t^{-1} (\tilde{\beta}_t - \Sigma_t \nu_t) \right) \right\} d\tilde{\beta}_t \propto$$

$$\frac{1}{\sigma_t^2} \frac{1}{(\sigma_t^2)^{n_t/2}} \exp\left\{ -\frac{1}{2} \left( z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t \right) \right\} \det(\Sigma_t) \propto$$

$$\frac{1}{\sigma_t^2} \frac{1}{(\sigma_t^2)^{n_t/2}} \exp\left\{ -\frac{1}{2} \left( z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t \right) \right\} (\sigma_t^2)^{(p_t + q_{t-1})/2}.$$

Note that

$$z_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t - \nu_t^T \Sigma_t \nu_t = \frac{1}{\sigma_t^2} \left( z_t^T R_t^{-1} z_t - (\mathcal{H}_t^T R_t^{-1} z_t)^T \left[\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t\right]^{-1} \mathcal{H}_t^T R_t^{-1} z_t \right),$$

and that if $\widehat{Q}_t = (z_t - \mathcal{H}_t \hat{\lambda}_t)^T R_t^{-1} (z_t - \mathcal{H}_t \hat{\lambda}_t)$, and $\hat{\lambda}_t = [\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t]^{-1} \mathcal{H}_t^T R_t^{-1} z_t$, then

$$\widehat{Q}_t = (z_t - \mathcal{H}_t \hat{\lambda}_t)^T R_t^{-1} (z_t - \mathcal{H}_t \hat{\lambda}_t) = z_t^T R_t^{-1} z_t - z_t^T R_t^{-1} \mathcal{H}_t [\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t]^{-1} \mathcal{H}_t^T R_t^{-1} z_t -$$

$$(\mathcal{H}_t [\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t]^{-1} \mathcal{H}_t^T R_t^{-1} z_t)^T R_t^{-1} z_t +$$

$$(\mathcal{H}_t[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1}\mathcal{H}_t^T R_t^{-1} z_t)^T R_t^{-1}\mathcal{H}_t[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1}\mathcal{H}_t^T R_t^{-1} z_t$$
$$= z_t^T R_t^{-1} z_t - (\mathcal{H}_t^T R_t^{-1} z_t)^T \Big[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t\Big]^{-1}\mathcal{H}_t^T R_t^{-1} z_t$$
$$\therefore p(\sigma_t^2|z^{(t)}) \propto \frac{1}{(\sigma_t^2)^{(n_t-p_t-q_{t-1})/2+1}} \exp\Big\{ -\frac{\widehat{Q}_t}{2}\Big\}$$
$$\implies \sigma_t^2|z^{(t)} \sim \mathcal{IG}(a_t, \widehat{Q}_t/2),$$

with $a_t = (n_t - p_t - q_{t-1})/2 + 1$ and the convention $q_0 = 0$.

*All priors are informative (i):* We will follow the same steps as in the non-informative case (ii), recycling many expressions we found there. First, recall the likelihood function given in Equation (A.24). Now, observe that as a function of $\tilde{\beta}_t$ and $\sigma_t^2$,

$$p(\tilde{\beta}_t|z^{(t-1)}, \sigma_t^2)p(z_t|\tilde{\beta}_t, \sigma_t^2)$$
$$\propto \frac{1}{(\sigma_t^2)^{(p_t+q_{t-1})/2}} \exp\Big\{ -\frac{1}{2}(\tilde{\beta}_t - b)^T \frac{W_t^{-1}}{\sigma_t^2}(\tilde{\beta}_t - b)\Big\}\frac{1}{(\sigma_t^2)^{n_t/2}}\times$$
$$\exp\Big\{-\frac{1}{2}\Big(\tilde{\beta}_t-\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]^{-1}\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]\Big)^T\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]\Big(\tilde{\beta}_t-\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]^{-1}\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]\Big)\Big\}\times$$
$$\exp\Big\{ -\frac{\widehat{Q}_t}{2\sigma_t^2}\Big\}$$

For the sake of notation, let us complete squares without all indexes and parameters, with $C$ and $D$ generic self-adjoint matrices and $x$, $c$ and $d$ vectors with appropriate dimensions:

$$(x-d)^T D(x-d) + (x - C^{-1}c)^T C(x - C^{-1}c) =$$

$$= x^T(D+C)x - x^T Dd - (Dd)^T x + d^T Dd - x^T c - c^T x + c^T C^{-1} c =$$
$$= x^T(D+C)x - x^T(Dd+c) - (Dd+c)^T x + d^T Dd + c^T C^{-1} c =$$
$$= (x - (D+C)^{-1}(Dd+c))^T(D+C)(x-(D+C)^{-1}(Dd+c)) +$$
$$d^T Dd + c^T C^{-1} c - (Dd+c)^T(D+C)^{-1}(Dd+c)$$

In our case, $x = \tilde{\beta}_t$, $d = b$, $D = \frac{W_t^{-1}}{\sigma_t^2}$, $c = \Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]$, and $C = \Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]$. Then, $D + C = \frac{W_t^{-1}}{\sigma_t^2} + \mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t = \Sigma_t^{-1}$ and $Dd + c = \frac{W_t^{-1}b}{\sigma_t^2} + \mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t = \nu_t$ (observe the change in the expressions for $\Sigma_t$ and $\nu_t$ compared to the non-informative case), and we have

$$p(\tilde{\beta}_t|z^{(t-1)}, \sigma_t^2)p(z_t|\tilde{\beta}_t, \sigma_t^2) \propto \frac{1}{(\sigma_t^2)^{(n_t+p_t+q_{t-1})/2}} \exp\Big\{ -\frac{1}{2}(\tilde{\beta}_t - \Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t - \Sigma_t\nu_t)\Big\}\times$$
$$\exp\Big\{ -\frac{1}{2}\Big(b_t^T\frac{W^{-1}}{\sigma_t^2}b_t + \Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]^T\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]^{-1}\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big] - \nu_t^T\Sigma_t\nu_t\Big)\Big\}\exp\Big\{ -\frac{\widehat{Q}_t}{2\sigma_t^2}\Big\}.$$

To simplify the next algebraic manipulations, let us call $\mathcal{H}^T R_t^{-1} z_t = v$ and $\mathcal{H}_t^T R_t^{-1} \mathcal{H}_t = S$, and let's drop the index $t$. We will use the matrix inversion lemma (A.6) in the form

$$(W^{-1} + S)^{-1} = W - W(W + S^{-1})^{-1} W,$$

and

$$(W^{-1} + S)^{-1} = S^{-1} - S^{-1}(W + S^{-1})^{-1} S^{-1}.$$

Note that

$$b^T W^{-1} b + v^T S^{-1} v - (W^{-1} b + v)^T (W^{-1} + S)^{-1} (W^{-1} b + v) = b^T W^{-1} b + v^T S^{-1} v -$$

$$\big(b^T W^{-1} (W^{-1}+S)^{-1} W^{-1} b + v^T (W^{-1}+S)^{-1} W^{-1} b + b^T W^{-1} (W^{-1}+S)^{-1} v + v^T (W^{-1}+S)^{-1} v\big),$$

and that, for the last 4 terms, which we will call

$$\Theta = b^T W^{-1} (W^{-1} + S)^{-1} W^{-1} b,$$

$$\Omega = v^T (W^{-1} + S)^{-1} W^{-1} b,$$

$$\Xi = b^T W^{-1} (W^{-1} + S)^{-1} v,$$

and

$$\Lambda = v^T (W^{-1} + S)^{-1} v,$$

we have

$$\Theta = b^T W^{-1} (W - W(W + S^{-1})^{-1} W) W^{-1} b = b^T W^{-1} b - b^T (W + S^{-1})^{-1} b,$$

$$\Omega = v^T S^{-1} (W + S^{-1})^{-1} b,$$

$$\Xi = b^T (W + S^{-1})^{-1} S^{-1} v,$$

and

$$\Lambda = v^T (S^{-1} - S^{-1}(W + S^{-1})^{-1} S^{-1}) v = v^T S^{-1} v - v^T S^{-1} (W + S^{-1})^{-1} S^{-1} v.$$

Therefore, it becomes clear that

$$b^T W^{-1} b + v^T S^{-1} v - (W^{-1} b + v)^T (W^{-1} + S)^{-1} (W^{-1} b + v) =$$

$$b^T W^{-1} b + v^T S^{-1} v - (\Theta + \Omega + \Xi + \Lambda) =$$

$$(b - S^{-1} v)^T (W + S^{-1})^{-1} (b - S^{-1} v),$$

and, therefore,

$$\exp\Big\{ -\frac{1}{2}\Big( b_t^T \frac{W^{-1}}{\sigma_t^2} b_t + \Big[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t\Big]^T \Big[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t^2} \mathcal{H}_t\Big]^{-1} \Big[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t^2} z_t\Big] - \nu_t^T \Sigma_t \nu_t \Big)\Big\} =$$

$$\exp\Big\{ -\frac{1}{2\sigma_t^2}\Big( (b_t - S^{-1} v)^T (W_t + S^{-1})^{-1} (b_t - S^{-1} v) \Big)\Big\} =$$

$$\exp\Big\{-\frac{1}{2\sigma_t^2}\Big((b_t-\hat{\lambda}_t)^T(W_t+[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1})^{-1}(b_t-\hat{\lambda}_t)\Big)\Big\}.$$

We have already obtained all necessary expressions. Now, paying attention to what goes into the multiplicative constant, it is easy to obtain the posterior of $\tilde{\beta}_t$:

$$p(\tilde{\beta}_t|z^{(t)},\sigma_t^2)\propto p(\tilde{\beta}_t|z^{(t-1)},\sigma_t^2)p(z_t|\tilde{\beta}_t,\sigma_t^2)\propto\exp\Big\{-\frac{1}{2}(\tilde{\beta}_t-\Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t-\Sigma_t\nu_t)\Big\}$$

$$\Longrightarrow\ \tilde{\beta}_t|z^{(t)},\sigma_t^2\sim\mathcal{N}(\Sigma_t\nu_t,\Sigma_t).$$

For the posterior of $\sigma_t^2$, we have to integrate $p(\tilde{\beta}_t|z^{(t-1)},\sigma_t^2)p(z_t|\tilde{\beta}_t,\sigma_t^2)$, and using the tediously obtained expressions above, we get

$$p(\sigma_t^2|z^{(t)})\propto p(\sigma_t^2|z^{(t-1)})\int p(z_t|z^{(t-1)},\tilde{\beta}_t,\sigma_t^2)p(\tilde{\beta}_t|z^{(t-1)},\sigma_t^2)d\tilde{\beta}_t\propto$$

$$\frac{1}{(\sigma_t^2)^{\alpha_t+1}}\exp\Big\{-\frac{\gamma_t}{\sigma_t^2}\Big\}\int\frac{1}{(\sigma_t^2)^{(n_t+p_t+q_{t-1})/2}}\exp\Big\{-\frac{1}{2}(\tilde{\beta}_t-\Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t-\Sigma_t\nu_t)\Big\}\times$$

$$\exp\Big\{-\frac{1}{2\sigma_t^2}\Big((b_t-\hat{\lambda}_t)^T(W_t+[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1})^{-1}(b_t-\hat{\lambda}_t)\Big)\Big\}\exp\Big\{-\frac{\widehat{Q}_t}{2\sigma_t^2}\Big\}d\tilde{\beta}_t=$$

$$\frac{1}{(\sigma_t^2)^{\alpha_t+(n_t+p_t+q_{t-1})/2+1}}\exp\Big\{-\frac{1}{2\sigma_t^2}\Big(2\gamma_t+(b_t-\hat{\lambda}_t)^T(W_t+[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1})^{-1}(b_t-\hat{\lambda}_t)+\widehat{Q}_t\Big)\Big\}\times$$

$$\int\exp\Big\{-\frac{1}{2}(\tilde{\beta}_t-\Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t-\Sigma_t\nu_t)\Big\}d\tilde{\beta}_t=$$

$$\frac{1}{(\sigma_t^2)^{\alpha_t+(n_t+p_t+q_{t-1})/2+1}}\exp\Big\{-\frac{1}{2\sigma_t^2}Q_t\Big\}\sqrt{\det(\Sigma_t)}\propto$$

$$\frac{1}{(\sigma_t^2)^{\alpha_t+n_t/2+1}}\exp\Big\{-\frac{1}{2\sigma_t^2}Q_t\Big\}.$$

This way, we obtain

$$\sigma_t^2|z^{(t)}\sim\mathcal{IG}\Big(\frac{n_t}{2}+\alpha_t,\frac{Q_t}{2}\Big).$$

# Bibliography

## [] Books:


[Adler '09] Robert J. Adler *The Geometry of Random Fields*, Siam, 2009.

[Chilès & Desassis '18] Chilès Jean Paul & Nicolas Desassis *Fifty years of kriging* in Daya Sagar B., Cheng Q. Agterberg F., editors, *Handbook of Mathematical Geosciences*, Springer, 2018.

[Cressie '93] Noel A. C. Cressie *Statistics for Spatial Data, Revised Edition*, Wiley-Interscience, 1993.

[DeGroot & Schervish '11] Morris H. DeGroot & Mark J. Schervish *Probability and Statistics (4th Edition)*, Pearson, 2011.

[Gihman & Skorohod '74] Iosif I. Gihman & Anatoliy V. Skorohod *The Theory of Stochastic Processes, Vol. 1*, Springer-Verlag, 1974.

[Hoff '09] Peter D. Hoff *A First Course in Bayesian Statistical Methods*, Springer, 2009.

[MacKay '03] David J. C. MacKay *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[MacKay '98] David J. C. MacKay *Introduction to Gaussian processes* in Bishop C. M., editor, *Neural networks and machine learning*, Springer, 1998.

[Rasmussen & Ghahramani '01] Carl Edward Rasmussen & Zoubin Ghahramani *Occam's Razor* in Leen, T. Dietterich, T. G. & Tresp V., editors, *Advances in Neural Information Processing Systems 13*, MIT Press, 2001.

[Rasmussen & Williams '05] Carl Edward Rasmussen & Christopher K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2005.

[Santner et al. '03] Thomas J. Santner, Brian J. Williams & William I. Notz *The design and analysis of computer experiments*, Springer, 2003.

[Stein '99] Michael L. Stein, *Interpolation of Spatial Data: Some Theorey for Kriging*, Springer, 1999.

Papers:

[Cressie '90] Noel A. C. Cressie *The origins of kriging*, Mathematical Geology, Volume 22, 239-252, 1990.

[Fernández-Godino et al. '16] M. Giselle Fernández-Godino, Chanyoung Park, Nam H. Kim & Raphael T. Haftka *Review of multi-fidelity models*, preprint arXiv:1609.07196 [stat.AP]

[Kennedy & O'Hagan '98] Marc C. Kennedy & Anthony O'Hagan *Predicting the output from a complex computer code when fast approximations are available*, Biometrika, Volume 87, Issue 1, 1-13, 2000.

[Krige '51] Danie G. Krige *A statistical approach to some basic mine valuation problems on the Witwatersrand*, Journal of the Southern African Institute of Mining and Metallurgy, Volume 52, Issue 6, 119-139, 1951.

[Harville '74] David A. Harville *Bayesian inference for variance components using only error contrasts*, Biometrika, 61, 1974.

[Le Gratiet & Garnier '14] Loic Le Gratiet & Josselin Garnier *Recursive co-kriging model for Design of Computer experiments with multiple levels of fidelity with an application to hydrodynamic*, International Journal for Uncertainty Quantification, 2014.

[Matheron '63] Georges Matheron *Principles of Geostatistics*, Economic Geology, Volume 58, 1246-1266, 1963.

[O'Hagan '78] Anthony O'Hagan *Curve Fitting and Optimal Design for Prediction*, Journal of the Royal Statistical Society: Series B (Methodological), Volume 40, Issue 1, 1-24, 1978.

[Patterson & Thompson '71] H. D. Patterson & Robin Thompson *Recovery of interblock information when block sizes are unequal*, Biometrika, 58, 1971.

[Peherstorfer et al. '18] Benjamin Peherstorfer, Karen Willcox & Max Gunzburger *Survey of multifidelity methods in uncertainty propagation, inference, and optimization*, SIAM Review 60(3), 2018.

[Perdikaris et al. '16] Paris Perdikaris, George Em Karniadakis & Daniele Venturi *Multifidelity information fusion algorithms for high-dimensional systems and massive data sets*, SIAM Journal on Scientific Computing 38, B521-B538l, 2016.

[Perdikaris et al. '17] Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D. Lawrence & George Em Karniadakis *Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling*, Proceedings of the Royal Society A 473, 2017.

## Theses:

[Duvenaud '14] David Kristjanson Duvenaud *Automatic model construction with Gaussian processes*, University of Cambridge, 2014.

[Gibbs '97] Mark N. Gibbs *Bayesian Gaussian process for regression and classification*, University of Cambridge, 1997.

[Le Gratiet '13] Loic Le Gratiet *Multi-fidelity Gaussian process regression for computer experiments* Université Paris-Diderot - Paris VII, 2013.

## Notes:

[O'Hagan '98] Anthony O'Hagan *A Markov property for covariance structures* Report 98-13, University of Nottingham statistics section, 1998.

## R Packages:

[DiceKriging] Olivier Roustant, David Ginsbourger, Yves Deville. Contributors: Clement Chevalier & Yann Richet *DiceKriging: Kriging Methods for Computer Experiments 1.5.6*, `https://CRAN.R-project.org/package=DiceKriging`, 2018.

[MuFiCokriging] Loic Le Gratiet *MuFiCokriging: Multi-Fidelity Cokriging models 1.2*, `https://CRAN.R-project.org/package=MuFiCokriging`, 2012.

## Python Packages:

[Matplotlib] J. D. Hunter *Matplotlib: A 2D Graphics Environment*, Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

[Scikit-learn] Pedregosa et al. *Scikit-learn: Machine Learning in Python*, JMLR 12, pp. 2825-2830, 2011.