Universidade Federal do Rio de Janeiro

Gaussian Processes and Multi-fidelity

Ivani Ivanova Ivanova

Rio de Janeiro

Agosto de 2019

Universidade Federal do Rio de Janeiro

Gaussian Process and Multifidelity

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
Agosto de 2019

Universidade Federal do Rio de Janeiro

Gaussian Process and Multifidelity

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 Ântonio Tavares Ramos

———————————————————

Prof. ?

———————————————————

Prof. ?

———————————————————

Prof. ?

Rio de Janeiro
Dezembro de 2016

Verdun, Claudio Mayrink
V487c Compressive Sensing / Claudio Mayrink Verdun. –
Rio de Janeiro: UFRJ/ IM, 2016.
202 f. 29,7 cm.
Orientador: César Javier Niche Mazzeo.
Coorientadores: Bernardo Freitas Paulo da Costa,
Eduardo Ântonio Barros da Silva.
Dissertação (Mestrado) — Universidade Federal do
Rio de Janeiro, Instituto de Matemática, Programa de
Pós-Graduação em Matemática, 2016.

1. Compressive Sensing. 2. Sparse Modelling.
3. Signal Processing. 4. Applied Harmonic Analysis.
5. Inverse Problems I. Mazzeo, César Javier Niche,
orient. II. Costa, Bernardo Freitas Paulo da, coorient.
Silva, Eduardo Ântonio Barros da, coorient. III. Título

Contents

Chapter 1

Gaussian Process Regression

1.1 Basics

Definition 1.1. 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. We define the mean function $m(x)$ and the covariance function $k(x, x')$ of a real process $Z(x)$ as

$$\begin{aligned} m(x) &= \mathbb{E}[Z(x)] \\ k(x, x') &= \mathbb{E}[(Z(x) - m(x))(Z(x') - m(x'))], \end{aligned} \tag{1.1}$$

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

$$Z(x) \sim \mathcal{GP}(m(x), k(x, x')). \tag{1.2}$$

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

It is common to take the mean function to be zero for simplicity, since it is usually unknown and would require a parametrization and subsequent estimation of a larger set of hiperparameters.

As a first example, consider the squared exponential covariance function, given by $k_{SE}(x, x') = \exp\big\{ - \frac{||x-x'||^2}{l^2} \big\}$, where l is a length-scale parameter.

1.2 Prediction

We have a training set $\mathcal{D}$ with n observations, $\mathcal{D} = \{(x_i, y_i)\}_{i=1,\ldots,n}$, where $x_i \in \mathbb{R}^D$ denotes an input vector of dimension D and y_i the associated scalar output called target. We aggregate the n inputs in a $n \times D$ matrix X, and the outputs in a vector y. We want to make inference about the output value for any input. We will obtain the necessary expressions for the zero-mean case, with which it is simple to generalize for an arbitrary $m(x)$.

1.2.1 Noise-free observations

For the noise-free case, we model our output in a location x as $Z(x)$ with

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

Let our observations be $\{(x_i, z_i)\}_{i=1,\dots,n}$ and the test inputs be $\{x_{*i}\}_{i=1,\dots,n_*}$. If we aggregate the inputs and test inputs in matrices X and X_*, respectively, and if $Z(X) = (Z(x_1), \dots, Z(x_n))^T$ and $Z(X_*) = (Z(x_{*1}), \dots, Z(x_{*n}))^T$. We call $K(X, X_*)$ the $n \times n_*$ covariance matrix of the process evaluated at all pairs of points in the training and test sets, and similarly for $K(X_*, X), K(X, X)$ and $K(X_*, X_*)$. Then, the joint distribution of the training and test outputs, according to the specified prior is

$$\begin{bmatrix} Z(X) \\ Z(X^*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right). \tag{1.3}$$

When observing $Z(X) = z$, we can condition the joint Gaussian distribution on the observations using (3.2) to obtain

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

with

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

and

$$\mathrm{Cov}[\bar{z}_*] = K(X_*, X_*) - K(X_*, X)K(X, X)^{-1}K(X, X_*).$$

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_* must necessarily decrease when compared to the prior covariance $K(x_*, x_*) = k(x_*, x_*)$.

1.2.2 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 ε denotes an additive independent and identically distributed Gaussian with mean 0 and variance σ_n^2. 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 δ_{ij} denotes the Kronecker delta. The joint distribution then becomes

$$\begin{bmatrix} Z(X) \\ Z(X_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right) \tag{1.4}$$

which, as in the noiseless case, gives rise to the predictive distribution

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

with

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

and

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

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

1.2.3 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 or noisy case, then

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

with

$$\bar{z}_* = m(X_*) + K(X_*, X) K_y^{-1}(z - m(X)) \tag{1.5}$$

while the predictive covariance remains

$$\mathrm{Cov}[\bar{z}_*] = K(X_*, X_*) - K(X_*, X) K_y^{-1} K(X, X_*). \tag{1.6}$$

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

To incorporate a mean function may be usefull for the interpretability of the model and the incorporation of prior knowledge, though 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 a kind of regression, expressing it as a combination of fixed basis functions. For this, let $h(x) = (h_1(x), \ldots, h_p(x))^T$ be the fixed basis functions, for example polynomials $(1, x, x^2, \ldots, x^p)^T$ and β a p-dimensional parameter vector which is to be inferred from the data. Then, the model consists of

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

with $Z(x) \sim \mathcal{GP}(0, k(x,x'))$.

It is common to put an independent Gaussian prior on the parameters β, such that $\beta \sim \mathcal{N}(b, B)$. Hence, using equations (1.5) and (1.6), the predictive distribution at the test inputs 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

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

where H and H_* are the matrices that collect the vectors $h(x)$ at the training and test locations, respectively, $H = (h(x_1), \ldots, h(x_n))$ and $H_* = (h(x_{*1}), \ldots, h(x_{*n}))$. After rearranging the term (see section (3.4)), we can rewrite the predictive mean and covariance as

$$\bar{w}_* = H_*^T \bar{\beta} + K_*^T K_y^{-1}(y - H^T \bar{\beta}) = \bar{z}_* + R^T \bar{\beta}, \tag{1.7}$$

$$\mathrm{Cov}[\bar{w}_*] = \mathrm{Cov}[\bar{z}_*] + R^T (B^{-1} + HK_y^{-1}H^T)^{-1} R, \tag{1.8}$$

with $\bar{\beta} = (B^{-1} + HK_y^{-1}H^T)^{-1}(HK_y^{-1}y + B^{-1}b)$ and $R = H_* - HK_y^{-1}H_*$. 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 (we add uncertainty when we included uncertainty on β).

1.2.4 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. 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 $Z(X)|X \sim \mathcal{N}(0, K)$. Using equation (3.3), we easily obtain

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

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 $B^{-1} \to O$ if the prior is vague, see [Rasmussen & Williams '05].

1.3 Properties

Definition 1.2. A stochastic process $Z(x)$ is said to be strictly stationary if its finite dimensional distributions are invariant under translations in the parameter 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 as 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 (WWS)*, or *weakly* stationary. Evidently, strict stationarity implies weak stationarity, though the reverse must 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 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 frequency space:

Theorem 1.3 (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 μ is a positive finite measure.

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

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

Theorem 1.4 (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 $x = x' = x^*$. If $k(x, x')$ is continuous at every diagonal point $x = x'$, 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 function continuity.

Theorem 1.5 (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 derivative $\partial^2 Z(x)/\partial x_i \partial x_j$, for $1 \leq i, j \leq D$ are defined as

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

which have as its covariance function the fourth order derivative

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

For a stationary process, 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 $\partial^{2m} k(r)/\partial^2 r_{i_1} \dots \partial^2 r_{i_m}$ exists and is finite at $r = 0$, then the m-th order partial derivative $\partial^m Z(x)/\partial x_{i_1} \dots \partial x_{i_m}$ exists for every x as a mean square limit.

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

Definition 1.6. A stochastic process $Z(x)$ is said to be almost surely continuous at x^* if for every sequence $\{x_n\}_{n=1,\ldots}$ for which $||x_n - x^*|| \to 0$ as $n \to \infty$, $Z(x_n) \xrightarrow{a.s.} Z(x^*)$. We say that Z is almost surely continuous throughout a set $A \subseteq \mathbb{R}^D$ if it is almost surely continuous at each $x \in A$. This type of continuity is refered as sample path continuity.

In particular, for Gaussian Processes, a.s. continuity is, again, a consequence if a certain condition on the covariance function is satisfied.

Theorem 1.7 (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 an interval I, Z has, with probability one, continuous sample functions over I.

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 $\varepsilon \geq 0$.

1.5 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 the level u as in [Adler '09].

Definition 1.8. Let $f(x)$ 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 is denoted N_u.

Theorem 1.9 (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*

$$\mathbb{E}[N_u] = \frac{1}{2\pi}\sqrt{-\frac{k''(0)}{k(0)}}\exp\left\{-\frac{u^2}{2k(0)}\right\}. \tag{1.9}$$

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 1.5).

blake & lindsey?

1.6 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 μ 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 we wish for 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 a 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 funcions $f \in L_2(\chi,\mu)$.

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].

The *squared exponential* covariance function is the most widely-used covariance kernel in machine learning and it is given by the following gaussian

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

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. By equation (1.9), we know that the expected number of upcrossing of the 1-dimensional Gaussian process in the interval $[0,1]$ is $(2\pi l)^{-1}$ which confirms l as a length-scale parameter. Furthermore, the squared exponential as a function of $d = ||x-x'||$ has an analytic Fourier transform, which is also a gaussian function:

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

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 forestry statistician Bertil Matérn. They are given by

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

with $\nu, l > 0$ and K_ν being the modified Bessel function of the second kind of order ν. The parameter ν is a smoothness parameter which relates to k_ν being $\lceil \nu \rceil - 1$ times differentiable and l has a role of length-scale parameter. As in [Rasmussen & Williams '05], for $\nu = p + 1/2$ 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 iteresting 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 this class have analytic expressions for their respective spectral densities and for $\nu \to \infty$ they converge to the squared exponential. See [Stein '99] and [Rasmussen & Williams '05] for further details.

When the parameter ν of the Matérn class is equal to 1/2, we have a rough process which is known as the Ornstein-Uhlenbeck with *exponential* the covariance function

$$k_{OU}(d) = \exp\left\{ -\frac{d}{l}\right\}$$

A similar covariance kernel is given by the γ-*exponential* class with covariance

$$k_{\gamma-\exp}(d) = \exp\left\{ -\left(\frac{d}{l}\right)^{\gamma}\right\}$$

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 $\gamma = 2$.

The *rational quadratic* covariance function for $\alpha, l > 0$ is given by

$$k_{RQ}(d) = \left(1+\frac{d^2}{2\alpha l^2}\right)^{-\alpha}.$$

For non-stationary processes, a simple covariance function is given by using a general covariance matrix Σ to create a *dot product* kernel:

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

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

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].

We also may have a different length-scale behavior throughout the input space. However, just 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, it is known that 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 the *blurring* effect when performing a convolution with a fixed kernel $h(w,w')$, with which it is possible to construct the covariance function $\int h(x,z)k(z,z')h(z',x')dzdz'$.

1.7 Model selection

The families of covariance functions have free hyper-parameters such as length-scale which must be chosen in some way. While some hyper-parameters may be interpretable, selecting the best values as precisely as possible is extremely important in order to make 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 values for each parameter or shape 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 function, and after selecting the hyper-parameters, for example. We will briefly explore two ways of performing model selection: Bayesian and cross-validation.

1.7.1 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 $\mathcal{M}_i$, there are parameters w which depend on hyper-parameters θ, but there may be as many levels as needed. We first specify priors $p(\mathcal{M}_i)$, $p(\theta|\mathcal{M}_i)$ and $p(w|\theta,\mathcal{M}_i)$. We could choose broad or non-informative priors if our knowledge about each set of elements is vague. Then, one level at a time, we infer the free elements of the level. To begin, we use Bayes rule to infer on 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 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 we use Bayes rule again and the marginal likelihood as

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

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 evaluations of integrals. If these integrals are not analytically tractable, we must resort to analytical approximations of Markov Chain Monte Carlo (MCMC). 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).

For GP regression, the role of hyper-parameters and models is quite straightforward, with the various covariance functions determining the model and the free variables in each covariance function being the hyper-parameters. 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 1.2.

We, thus, 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 hyper-parameters θ through K_y. Maximizing on the hiperparameters includes obtaining the derivatives

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

Now observe that $y^T K^{-1} \frac{\partial K}{\partial \theta_i} K^{-1} y = \mathrm{tr}(y^T K^{-1} \frac{\partial K}{\partial \theta_i} K^{-1} y) = \mathrm{tr}(K^{-1} y (K^{-1} y)^T \frac{\partial K}{\partial \theta_i})$

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

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

occam? p 110 rasmussen, meio confuso

1.7.2 Cross-validation

Gaussian processes are a very powerfull 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. When using this hold-out method, if the validation set is small one obtains estimates with large variance and with one split you lose 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. Generally, this is an extremely expensive procedure, however some models including GP have computational shortcuts. We will briefly discuss them for GP.

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. How this is done will be clarified later. Let ξ 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 ξ, 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

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

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

For simplicity of notation, will use v_ζ for a vector containing the entries of v with indices in the set ζ, $v_{-\zeta}$ for the vector of entries of v with indices not in ζ, $[A]_{[\zeta,\gamma]}$ for the submatrix of A containing the row indices in ζ and column indices in γ and similarly as in the vector case for the submatrices $[A]_{[-\zeta,\gamma]}$, $[A]_{[\zeta,-\gamma]}$ and $[A]_{[-\zeta,-\gamma]}$. The sets ζ and γ 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 inverting the matrix $K(X_{-\xi}, X_{-\xi})$, which will be of size $(n - \#\xi) \times (n - \#\xi)$ and $\#\xi = n/k$ in the k-fold setting. If we rearrange the points so that the ones with indices in ξ 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, $y_\xi | X, y_{-\xi}, \theta \sim \mathcal{N}\big([K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]} y_{-\xi}, [K]_{\xi,\xi} - [K]_{[\xi,-\xi]}[K]^{-1}_{[-\xi,-\xi]}[K]_{[-\xi,\xi]}\big)$.

By the block matrix inversion identity (3.5), 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

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

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 the inverse of $[K]_{[-\xi,-\xi]}$ as

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

For the particular case of the LOO, the set of indices ξ consists only of one index i and $\mathcal{Q} = 1/[K^{-1}]_{ii}$, thus the predictive probability is

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

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

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

and

$$\frac{\partial \sigma_i^2}{\theta_j} = -\frac{\left[K^{-1}\frac{\partial K}{\partial \theta_j}K^{-1}\right]_{ii}}{[K^{-1}]_{ii}^2},$$

and the chain rule gives us the derivative of the log predictive probability

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

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

Then, 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. Thus, this method is more costly than the one based on the marginal likelihood.

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

1.8 Historical Notes

Gaussian processes are by no means a recent topic, having been studied in probability theory and statistics as they are probably the simplest representation of a random process.

In the field of geostatistics, Gaussian process regression is known as kriging.

Chapter 2

Multi-Fidelity

2.1 A gist of multi-fidelity

· how to build a surrogate model for different levels of fidelity
· what type of models exist and history

2.2 A first autoregressive model

Many 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 possible problem is the need to specify a large number of parameters, which can be difficult to identify or measure directly. However, since many experiments can be simulated at different orders of fidelity (for example, simulations of physical processes by finite elements may be made using different grids, a smaller grids being the higher fidelity code), a method to integrate the information of the simple and inexpensive simulations that still capture essential features to the data of the expensive and more reliable code would be a practical approach for understanding the phenomenon given the cost/time restrictions.

The work of Kennedy and O'Hagan in [Kennedy & O'Hagan '98] concerns the use of a autoregressive model based on Gaussian processes for combining data from deterministic simulations of different accuracy in order to infer about the most accurate and reliable code and perform a uncertainty analysis. 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, since individual runs of rough codes don't provide information outside of a very small neighbourhood.

(3) Prior beliefs of each level of the code can be modelled using Gaussian Process.

(4) The outputs of each level are scalars.

We suppose that we have s levels of code $\{z_t(x)\}_{t=1,\ldots,s}$ sorted by increasing order of fidelity and modeled by Gaussian Processes $\{Z_t(x)\}_{t=1,\ldots,s}$, with $x \in Q \subseteq \mathbb{R}^n$, thus considering the $z_s(x)$ being the most accurate and costly code. The object of inference will be $Z_s(x)$ conditioned on all outputs of all levels of code we have.

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

$$\mathrm{Cov}\{Z_t(x), Z_{t-1}(x')|Z_{t-1}(x)\} = 0 \;\text{ for all }\; x' \neq x. \tag{2.1}$$

This translates as a kind of Markov property: given the nearest point $Z_{t-1}(x)$, we can learn no 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 $t = 2, \ldots, s$:

$$\begin{cases} Z_t(x) = \rho_{t-1}(x) Z_{t-1}(x) + \delta_t(x) \\ Z_{t-1}(x) \perp \delta_t(x) \\ \rho_{t-1}(x) = g_{t-1}^T(x)\beta_{\rho_{t-1}} \end{cases} \tag{2.2}$$

where

$$\delta_t(x) \sim \mathcal{GP}(f_t^T(x)\beta_t, \sigma_t^2 r_t(x,x')) \tag{2.3}$$

and

$$Z_1(x) \sim \mathcal{GP}(f_1^T(x)\beta_1, \sigma_1^2 r_1(x,x')). \tag{2.4}$$

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, β_t is a p_t-dimensional vector, $\beta_{\rho_{t-1}}$ is a q_{t-1}-dimensional vector, and σ_t^2 is a positive real number. We denote $\sigma^2 = (\sigma_1^2, \ldots, \sigma_s^2)$, $\beta = (\beta_1^T, \ldots, \beta_s^T)^T$ and $\beta_\rho = (\beta_{\rho_1}^T, \ldots, \beta_{\rho_{s-1}})$. This way, we can obtain the expected value of $Z_t(x)$ as

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

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 having $\sum_{i=t+1}^{s} p_i$ zeros at its right, and

$$\begin{gathered} \mathrm{Cov}\{Z_t(x), Z_t(x')|\sigma^2, \beta, \beta_\rho\} = \\ \mathrm{Cov}\{\rho_{t-1}(x)Z_{t-1}(x) + \delta_t(x), \rho_{t-1}(x')Z_{t-1}(x')|\sigma^2, \beta, \beta_\rho\} + \delta_t(x')\} = \\ \rho_{t-1}(x)\rho_{t-1}(x')\mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(x')|\sigma^2, \beta, \beta_\rho\} + \sigma_t^2 r_t(x,x') = \\ = \cdots = \sum_{j=1}^{t} \sigma_j^2 \left(\prod_{i=j}^{t-1} \rho_i(x)\rho_i(x') \right) r_j(x,x'), \end{gathered} \tag{2.5}$$

$$\begin{aligned}\mathrm{Cov}\{Z_t(x), Z_{t'}(x')|\sigma^2,\beta,\beta_\rho\} &= \mathrm{Cov}\{\rho_{t-1}(x)Z_{t-1}(x)+\delta_t(x), Z_{t'}(x')|\sigma^2,\beta,\beta_\rho\} = \\ &\rho_{t-1}(x)\mathrm{Cov}\{Z_{t-1}(x), Z_{t'}(x')|\sigma^2,\beta,\beta_\rho\} = \cdots = \\ &\left(\prod_{i=t'}^{t-1}\rho_i(x)\right)\mathrm{Cov}\{Z_{t'}(x), Z_{t'}(x')\} \text{ for } t' < t.\end{aligned} \tag{2.6}$$

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

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

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 a 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.

$$H_s = \begin{bmatrix}[\text{—} & h_1(D_1) & \text{—}] \\ [\text{—} & h_2(D_2) & \text{—}] \\ & \vdots & \\ [\text{—} & h_{s-1}(D_{s-1}) & \text{—}] \\ [\text{—} & h_s(D_s) & \text{—}]\end{bmatrix} = \begin{bmatrix}\text{—} & h_1(x_1^1) & \text{—} \\ & \vdots & \\ \text{—} & h_1(x_{n_1}^1) & \text{—} \\ \text{—} & h_2(x_1^2) & \text{—} \\ & \vdots & \\ \text{—} & h_{s-1}(x_{n_{s-1}}^{s-1}) & \text{—} \\ \text{—} & h_s(x_1^s) & \text{—} \\ & \vdots & \\ \text{—} & h_s(x_{n_s}^s) & \text{—}\end{bmatrix}. \tag{2.7}$$

We now construct the vector $k_s(x)$ of covariances between $Z_s(x)$ and $\mathcal{Z}^{(s)}$:

$$k_s^T(x) = (c_1^T(x, D_1),\dots,c_s^T(x, D_s))^T \tag{2.8}$$

with $c_t^T(x, D_t) = \mathrm{Cov}\{Z_s(x), Z_t(D_t)\} = (\mathrm{Cov}\{Z_s(x), Z_t(x_1^t)\},\dots,\mathrm{Cov}\{Z_s(x), Z_t(x_{nt}^t)\})$, which, using the expressions (2.5) and (2.6), can be rewritten as

$$\begin{aligned} c_t^T(x, D_t) &= \prod_{i=t}^{s-1}\rho_i(x)\mathrm{Cov}\{Z_t(x), Z_t(D_t)\} = \\ \left(\prod_{i=t}^{s-1}\rho_i(x)\right)&(\rho_{t-1}(x)\rho_{t-1}(D_t)\odot\mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\} + \sigma_t^2 r_t(x, D_t)) = \\ &\rho_{t-1}(D_t)\odot c_{t-1}^T(x, D_t) + \left(\prod_{i=t}^{s-1}\rho_i(x)\right)\sigma_t^2 r_t(x, D_t),\end{aligned} \tag{2.9}$$

where $\odot$ represents the element by element matrix (or vector) product, $c_i^T(x, D_t) = \text{Cov}\{Z_s(x), Z_i(D_t)\}$ 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 $c_1^T(x, D_t) = \prod_{i=1}^{s-1} \rho_i(x)\text{Cov}\{Z_1(x), Z_1(D_t)\} = \Big(\prod_{i=1}^{s-1} \rho_i(x)\Big)\sigma_1^2 r_1(x, D_t)$.

The covariance matrix V_s of $\mathcal{Z}^{(s)}$, can also be contructed using (2.5) and (2.6):

$$V_s = \text{Cov}\{\mathcal{Z}^{(s)}, \mathcal{Z}^{(s)}\} = \begin{bmatrix} V_{1,1} & \ldots & V_{1,s} \\ \vdots & \ddots & \vdots \\ V_{s,1} & \ldots & V_{s,s} \end{bmatrix}, \tag{2.10}$$

with the diagonal elements

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

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

$$V_{t',t} = \text{Cov}\{\mathcal{Z}_{t'}, \mathcal{Z}_t\} = \text{Cov}\{Z_{t'}(D_{t'}), \rho_{t-1}(D_t) \odot Z_{t-1}(D_t) + \delta_t(D_t)\} =$$

$$(\mathbf{1}_{n_{t'}}\rho_{t-1}^T(D_t)) \odot \text{Cov}\{Z_{t'}(D_{t'}), Z_{t-1}(D_t)\} = \cdots = \tag{2.11}$$

$$= \Bigg(\bigodot_{i=t'}^{t-1} \mathbf{1}_{n_{t'}}\rho_i^T(D_t)\Bigg) \odot \text{Cov}\{Z_{t'}(D_{t'}), Z_{t'}(D_t)\}$$

when $1 \leq t' < t \leq s$ and $V_{t',t}^T$ otherwise. Here, $V_{t',t'}(D_t, D_{t'})$ is 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.

Lastly, let $v_{Z_s}^2(x)$ denote the variance of $Z_s(x)$. By (2.5), it is given by

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

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

$$\begin{bmatrix} Z_s(x) \\ \mathcal{Z}^{(s)} \end{bmatrix} \Bigg| \sigma^2, \beta, \beta_\rho \Bigg] \sim \mathcal{N}\Bigg(\begin{bmatrix} h_s(x)^T\beta \\ H_s\beta \end{bmatrix}, \begin{bmatrix} v_{Z_s}^2(x) & k_s^T(x) \\ k_s(x) & V_s \end{bmatrix}\Bigg). \tag{2.12}$$

By the predictive identities for Gaussian processes (1.5) and (1.6), it is straightforward that

$$Z_s(x)|\mathcal{Z}^{(s)} = z^{(s)}, \sigma^2, \beta, \beta_\rho \sim \mathcal{N}(m_{Z_s}(x), s_{Z_s}^2(x)), \tag{2.13}$$

with

$$m_{Z_s}(x) = h_s^T(x)\beta + k_s^T(x)V_s^{-1}(z^{(s)} - H_s\beta), \tag{2.14}$$

and

$$s_{Z_s}^2(x) = v_{Z_s}^2(x) - k_s^T(x)V_s^{-1}k_s(x). \tag{2.15}$$

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 σ_1, by Proposition 3.2, we have that $k_s^T(x)V_s^{-1}$ is independent of σ_t^2 for $t = 1, \dots, s$ and, therefore, the predictive mean $m_{Z_s}(x)$ does not depend on the variance parameter of any level.

2.3 The recursive autoregressive model

The work in [Le Gratiet '13] and [Le Gratiet & Garnier '14] is an extension and apprimoration of the autoregressive model of Kennedy and O'Hagan in [Kennedy & O'Hagan '98]. There, a new way of performing the co-kriging is present for reducing the computational complexity by breaking the s-level co-kriging into s independent krigings.

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

$$\begin{cases} Z_t(x) = \rho_{t-1}(x)\widetilde{Z}_{t-1}(x) + \delta_t(x) \\ \widetilde{Z}_{t-1}(x) \perp \delta_t(x) \\ \rho_{t-1}(x) = g_{t-1}^T(x)\beta_{\rho_{t-1}} \end{cases} \tag{2.16}$$

where $\widetilde{Z}_{t-1}(x)$ is a Gaussian process with distribution $[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_{t-1}^2, \beta_{t-1}, \beta_{\rho_{t-2}}]$, $\delta(x)$ is a Gaussian process with distribution (2.3), the experimental design sets have the nested property $D_s \subseteq D_{s-1} \subseteq \dots, \subseteq D_1$. The only difference from the classical autorregressive multi-fidelity model (2.2) is that, instead of expressing Z_t as a function of Z_{t-1}, we condition Z_{t-1} by the values $z^{(t-1)} = (z_1, \dots, z_{t-1})$ in the sets $\{D_i\}_{i=1,\dots,t-1}$. 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 $t = 2, \dots, s$, 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 (2.16), 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 $r_t(x, x) = 1\ \forall x$ as it is a correlation. This way, the joint distribution of $Z_t(x)$ and $\mathcal{Z}_t$ conditioned on $\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_t^2, \beta_t$ and $\beta_{\rho_{t-1}}$ is

$$\left[\begin{matrix} Z_t(x) \\ \mathcal{Z}_t \end{matrix}\middle| \mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}}\right] \sim$$

$$\sim \mathcal{N}\left(\begin{bmatrix} \rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T\beta_t \\ \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) + F_t\beta_t \end{bmatrix}, \begin{bmatrix} \rho_{t-1}^2(x)\sigma_{Z_{t-1}}^2(x) + \sigma_t^2(x)) & r_t^T(x) \\ r_t(x) & R_t \end{bmatrix} \right).$$

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

Using again (1.5) and (1.6) for conditioning $Z_t(x)$ by $\mathcal{Z}_t = z_t$, we obtain the expressions for $\mu_{Z_t}(x)$ and $\sigma_{Z_t}^2(x)$ in the distribution of

$$\widetilde{Z}_t(x) = [Z_t(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}, \sigma_t^2, \beta_t, \beta_{\rho_{t-1}}] \sim \mathcal{N}(\mu_{Z_t}(x), \sigma_{Z_t}^2(x)) \tag{2.17}$$

which are

$$\mu_{Z_t}(x) = \rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t) \odot \mu_{Z_{t-1}}(D_t) - F_t\beta_t), \quad (2.18)$$

and

$$\sigma^2_{Z_t}(x) = \rho^2_{t-1}(x)\sigma^2_{Z_{t-1}}(x) + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t(x)), \quad (2.19)$$

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, similarly as in the basic Gaussian process regression, the predictive mean does not depend on the variance parameters $\{\sigma_t^2\}_{t=1,\ldots,s}$.

Note that, for $t = 1$,

$$\begin{bmatrix} Z_1(x) \\ \mathcal{Z}_1 \end{bmatrix} \Big| \sigma_1^2, \beta_1 \Bigg] \sim \mathcal{N}\left(\begin{bmatrix} f_1^T\beta_1 \\ F_1\beta_1 \end{bmatrix}, \begin{bmatrix} \sigma_1^2(x) & r_1^T(x) \\ r_1(x) & R_t \end{bmatrix} \right),$$

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

with

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

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 known values of $Z_t(x)$ at the points in D_t.

Proof. For $t = 1$,

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

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

Similarly, 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 D_t sets, gives us

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

which can be replaced in equation (2.18) to obtain

$$\mu_{Z_t}(x) = \rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t) \odot z_{t-1}(D_t) - F_t\beta_t). \quad (2.20)$$

Remark 2. *In the same conditions,*

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

Proof. Observe that, the i-th column of R_t is equal to $r_t(x_i^t)$, therefore, by (3.8),

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

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. □

Despite the different formulation of the classical autoregressive model (2.2) and the recursive autoregressive model (2.16), both of them have, in fact, the same predictive equations. This result is stated in the following proposition:

Proposition 2.1 (Proposition 1 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 points in $\{D_t\}_{t=1,\dots,s}$ with $D_s \subseteq D_{s-1} \subseteq \cdots \subseteq D_1$. If we consider the mean (2.14) and the variance (2.15) induced by the model (2.2) when we condition the Gaussian process $Z_s(x)$ by the known values $z^{(s)}$ of $\mathcal{Z}^{(s)}$ and parameters β, β_ρ and σ^2 and the mean (2.18) and variance (2.19) induced by the model (2.16) when we condition $Z_s(s)$ by $z^{(s)}$ and parameters β, β_ρ and σ^2, then, we have:*

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

Proof. Throughout this proof, we will use the nested property of the sets, specifically that $D_t \subseteq D_{t-1}$, and the special 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(2.14), we know that for the classical model, we have

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

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

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

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

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

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

This way,

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

For $t = 1, \dots, s$, H_t can be constructed similarly to H_s of equation (2.7), 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 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 or, more precisely,

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

where

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

From Proposition 3.2, we have that

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

$$\implies k_t^T(x)V_t^{-1}z^{(s)} = k_t^T(x)V_t^{-1}\begin{bmatrix} z^{(s-1)} \\ z_s \end{bmatrix} =$$

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

Again by Proposition 3.2 and the expression we obtained for H_t, we have

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

$$[\rho_{t-1}^T(D_t) \odot r^T(x)]R_t^{-1}h_{t-1}(D_t)\beta^{(t-1)} +$$

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

$$\rho_{t-1}(x)k_{t-1}^T(x)V_{t-1}^{-1}H_{t-1}\beta^{(t-1)} + r^T(x)R_t^{-1}F_t\beta_t$$

since the two middle terms cancel each other out. Therefore,

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

$$\rho_{t-1}(x)m_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}\Big(z_t - \rho_{t-1}(D_t) \odot z_{t-1}(D_t) - F_t\beta_t\Big).$$

Thus, both $m_{Zt}(x)$ and $\mu_{Zt}(x)$ follow the 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 relation

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

For the variance: We follow similar steps as above to prove the result.

For the t-level classical co-kriging model, equation (2.15) states that

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

For the variance term, we use equation (2.5), to get

$$v^2_{Z_t}(x) = \mathrm{Var}[Z_t(x)] = \rho^2_{t-1}(x)\mathrm{Var}[Z_{t-1}(x)] + \sigma_t^2 = \rho^2_{t-1}(x)v^2_{Z_{t-1}}(x) + \sigma_t^2. \qquad (2.21)$$

For the $k_t^T(x)V_t^{-1}k_t(x)$ term, we know from Proposition 3.2 that

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

and by equations (3.12) and (2.5), it is clear that

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

These equalities, in turn, imply that

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

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 that reason,

$$k_t^T(x)V_t^{-1}k_t(x) = \rho^2_{t-1}(x)k_{t-1}^T V_{t-1}^{-1}k_{t-1}(x) - [\rho_{t-1}^T(D_t) \odot r_t^T(x)]R_t^{-1}\rho_{t-1}(x)\mathrm{Cov}\{Z_{t-1}(x), Z_{t-1}(D_t)\}^T$$

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

$$\rho^2_{t-1}(x)k_{t-1}^T V_{t-1}^{-1}k_{t-1}(x) + \sigma_t^2 r_t^T(x)R_t^{-1}r_t(x)$$

This result, together with equation (2.21), gives us

$$s^2_{Z_t}(x) = \rho^2_{t-1}(x)(v^2_{Z_{t-1}}(x) - k_{t-1}^T V_{t-1}^{-1}k_{t-1}(x)) + \sigma_t^2(1 - r_t^T(x)R_t^{-1}r_t(x)) =$$

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

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].

□

With this result, we proved that both (2.2) and (2.16) models have the same predictive Gaussian distribution for $Z_s(x)$, and, while the computational cost of the model (2.2) 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 $t = 1, \dots, s$. This results in a lower computational cost for the recursive model. Besides that, the memory cost is also lower in this model, since it requires storing the s matrices $\{R_t\}_{t=1,\dots,s}$ instead of the matrix V_s for the classical approach.

2.3.1 Bayesian parameter estimation

The parameter vector β, β_ρ and σ_t^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 each t and (β_1, σ_1^2) can be estimated separately. For the bayesian approach, a smart choice of prior distributions give us closed form expressions for the posteriors. We will consider two such choices:

(i) all priors are informative

(ii) all priors are non-informative.

Case (i): we consider the Jeffreys priors

$$p(\beta_1|\sigma_1^2) \propto 1, \;\; p(\sigma_1^2) \propto \frac{1}{\sigma_1^2}, \;\; p(\beta_{\rho_{t-1}}, \beta_t|z^{(t-1)}, \sigma_t^2) \propto 1, \;\; p(\sigma_t^2|z^{(t-1)}) \propto \frac{1}{\sigma_t^2}. \tag{2.22}$$

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

$$\begin{gathered}
[\beta_1|\sigma_1^2] \sim \mathcal{N}_{p_1}(b_1, \sigma_1^2 W_1) \\
[\beta_{\rho_{t-1}}, \beta_t|z^{(t-1)}, \sigma_t^2] \sim \mathcal{N}_{q_{t-1}+p_t}\left(b_t = \begin{bmatrix} b_{t-1}^\rho \\ b_t^\beta \end{bmatrix}, \sigma_t^2 W_t = \sigma_t^2 \begin{bmatrix} W_{t-1}^\rho & 0 \\ 0 & W_t^\beta \end{bmatrix} \right) \\
[\sigma_1^2] \sim \mathcal{IG}(\alpha_1, \gamma_1), \qquad [\sigma_t^2|z^{(t-1)}] \sim \mathcal{IG}(\alpha_t, \gamma_t),
\end{gathered} \tag{2.23}$$

where b_1 is a vector of size p_1, b_{t-1}^ρ a vector of size $q_t - 1$, b_t^β is a vector of size p_t, W_1 is a $p_1 \times p_1$ matrix, W_{t-1}^ρ a $q_{t-1} \times q_{t-1}$ matrix, W_t^β 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 3.6 and are given by

$$[\beta_1|z_1, \sigma_1^2] \sim \mathcal{N}_{p_1}(\Sigma_1\nu_1, \Sigma_1), \qquad [\beta_{\rho_{t-1}}, \beta_t|z^{(t)}, \sigma_t^2] \sim \mathcal{N}_{q_{t-1}+p_t}(\Sigma_t\nu_t, \Sigma_t) \tag{2.24}$$

where

$$\Sigma_t = \begin{cases} \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} \mathcal{H}_t + \frac{W_t^{-1}}{\sigma_t^2}\right]^{-1} & \text{(i)} \\ \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} \mathcal{H}_t\right]^{-1} & \text{(ii)} \end{cases} \tag{2.25}$$

$$\nu_t = \begin{cases} \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} z_t + \frac{W_t^{-1}}{\sigma_t^2} b_t\right] & \text{(i)} \\ \left[\mathcal{H}_t^T \frac{R_t^{-1}}{\sigma_t} z_t\right] & \text{(ii)} \end{cases} \tag{2.26}$$

with $\mathcal{H}_1 = F_1$ and for $H_t = [G_{t-1} \odot z_{t-1}(D_t)\mathbf{1}_{q_{t-1}}^T) \quad F_t]$ with G_{t-1} being the experience matrix containing the values of $g_{t-1}(x)^T$ at the points in D_t. Also, for $t \geq 1$,

$$[\sigma_t^2 | z^{(t)}] \sim \mathcal{IG}\Big(a_t, \frac{Q_t}{2}\Big) \tag{2.27}$$

with

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

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

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

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 β_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 σ_t^2.

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 0 (the particular choice of C is not important, see [Harville '74]). The idea behind this is that the transformed vector won't depend on parameters other than σ_t^2, and this implies that there won't be an increase of bias due to the estimation of the parameters β_t and $\beta_{\rho_{t-1}}$, this means ignoring prior information of these parameters. For simplicity, if $\tilde{\beta}_t = \begin{bmatrix} \beta_{\rho_{t-1}} \\ \beta_t \end{bmatrix}$ for $t > 1$ and $\tilde{\beta}_1 = \beta_1$, z_t (given $z^{(t-1)}$, β_t, $\beta_{\rho_{t-1}}$ and σ_t^2) follows the distribution

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

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 $\mathcal{H}_t^T \mathcal{H}_t = I$. Then, as we need

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

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

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

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

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

$$\begin{aligned}\log(\ell_{rest}(\zeta_t;\sigma_t^2)) &= -\frac{n_t-p_t-q_{t-1}}{2}\log(2\pi) + \frac{1}{2}\log(\det(\mathcal{H}_t^T\mathcal{H}_t)) - \frac{n_t-p_t-q_{t-1}}{2}\log(\sigma_t^2)\\ &-\frac{1}{2}\log(\det(R_t)) - \frac{1}{2}\log(\det(\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t)) - \frac{1}{2\sigma_t^2}(z_t-\mathcal{H}_t\hat{\lambda}_t)^T R_t^{-1}(z_t-\mathcal{H}_t\hat{\lambda}_t) \qquad (2.28)\\ \implies \frac{\partial\log(\ell_{rest}(\zeta_t;\sigma_t^2))}{\partial\sigma_t^2} &= -\frac{n_t-p_t-q_{t-1}}{2}\frac{1}{\sigma_t^2} + \frac{1}{2(\sigma_t^2)^2}(z_t-\mathcal{H}_t\hat{\lambda}_t)^T R_t^{-1}(z_t-\mathcal{H}_t\hat{\lambda}_t).\end{aligned}$$

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

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

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 hyper-parameter φ_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 $\widehat{\sigma}^2_{t,\text{EML}}(\varphi_t)$ (it depends on φ through $R_t(\varphi_t)$) for σ_t^2 in the expression of the log-likelihood (2.28). Therefore, we would need to minimize

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

which has to be performed numerically.

2.3.2 Universal co-kriging model

mudar titulo e explicar aqui

Proposition 2.2. *Let us consider s Gaussian process $\{Z_t(x)\}_{t=1,\dots,s}$ 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 (2.17) and the posterior distribution of the parameters given in equations (2.24) and (2.27), then we have for $t = 1,\dots,s$:*

$$\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = u_t^T(x)\Sigma_t\nu_t + r_t^T(x)R_t^{-1}(z_t - \mathcal{H}_t\Sigma_t\nu_t) \tag{2.29}$$

with $u_1^T = f_1^T$, $\mathcal{H}_1 = F_1$ *and for* $t > 1$, $u_t^T(x) = (g_{t-1}^T(x)\mathbb{E}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}] \quad f_t(x)^T)$, *and* $\mathcal{H}_t = [G_{t-1} \odot z_{t-1}(D_t)\mathbf{1}_{q_{t-1}}^T, F_t]$. *Furthermore, we have*

$$Var[Z_t(x)|\mathcal{Z}^{(t)} = z^{(t)}] = \hat{\sigma}^2_{\rho_{t-1}}(x)\mathrm{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)} = z^{(t-1)}]$$
$$+\frac{Q_t}{2(a_t-1)}(1 - r_t^T(x)R_t^{-1}r_t^T(x)) + (u_t^T(x) - r_t^T(x)R_t^{-1}\mathcal{H}_t)\widehat{\Sigma}_t(u_t^T(x) - r_t^T(x)R_t^{-1}\mathcal{H}_t)^T \tag{2.30}$$

with $\hat{\sigma}^2_{\rho_{t-1}}(x) = \hat{\rho}^2_{t-1}(x) + g_{t-1}^T(x)\widehat{\Sigma}_{\rho,t}g_{t-1}(x)$, $\hat{\rho}_{t-1}(x) = g_{t-1}^T(x)[\widehat{\Sigma}_t, \widehat{\nu}_t]_{1,\dots,q_{t-1}}$, $\widehat{\Sigma}_{\rho,t}$ *is the submatrix of elements* $(1,\dots,q_{t-1}) \times (1,\dots,q_{t-1})$ *of* $\widehat{\Sigma}_t$, *which has the same expression of* Σ_t *but with* σ_t^2 *replaced by its posterior mean, and similarly for* $\widehat{\nu}_t$.

Proof. Mean for $t > 1$*:*

From the law of total expectation, we have that

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

By equations (2.17) and (2.18), we know that for $t > 1$,

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

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

and, given that $\mu_{Z_{t-1}}$ is independent of both $\rho_{t-1}(x)$ and z_t, we obtain

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

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

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

Variance for $t > 1$:

We will use the law of total variance twice to obtain the desired variance identity. We know that

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

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

$$= \operatorname{Var}[\rho_{t-1}(x)\mu_{Z_{t-1}}(x) + f_t^T(x)\beta_t + r_t^T(x)R_t^{-1}(z_t - \rho_{t-1}(D_t)\odot\mu_{Z_{t-1}}(D_t) - F_t\beta_t)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] =$$

$$= (u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)\Sigma_t(u_t^T(x) - r_t^T(x)R^{-1}\mathcal{H}_t)^T, \tag{2.31}$$

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

We also know that

$$\operatorname{Var}[Z_t(x)]|\mathcal{Z}^{(t)} = z^{(t)}, \beta_t, \beta_{\rho_{t-1}}, \sigma_t^2] = \operatorname{Var}[\widetilde{Z}_t(x)] = \sigma_{Z_t}(x)$$

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

$$= \mathbb{E}[\rho_{t-1}^2(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2]\mathbb{E}[\sigma_{Z_{t-1}}^2(x)|\mathcal{Z}^{(t)} = z^{(t)}, \sigma_t^2] + \sigma_t^2(1 - r_t^t(x)R_t^{-1}r_t(x)),$$

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

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

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

$$\mathrm{Var}[Z_t(x)]|\mathcal{Z}^{(t)}=z^{(t)},\beta_t,\beta_{\rho_{t-1}},\sigma_t^2]=\hat{\sigma}_{\rho_{t-1}}^2(x)\mathrm{Var}[Z_{t-1}(x)|\mathcal{Z}^{(t-1)}=z^{(t-1)}]+\sigma_t^2(1-r_t^T(x)R_t^{-1}r_t^T(x)) \tag{2.32}$$

By the law of total variance and equations (2.31) and (2.32), we obtain:

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

Again by the law of total variance again,

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

Note that, as stated previously, $\mathbb{E}[Z_t(x)]|\mathcal{Z}^{(t)}=z^{(t)},\beta_t,\beta_{\rho_{t-1}}]$ is independent of σ_t^2 and for this reason the term $\mathrm{Var}[\mathbb{E}[Z_t(x)|\mathcal{Z}^{(t)}=z^{(t)},\sigma_t^2]|\mathcal{Z}^{(t)}=z^{(t)}]$ in the previous equality is equal to 0. Now we only need to perform the expectation in σ_t^2 of

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

Since $[\sigma_t^2|z^{(t)}]\sim\mathcal{IG}\Big(a_t,\frac{Q_t}{2}\Big)$, the posterior mean of σ_t^2 is $\frac{Q_t}{2(a_t+1)}$. We also observe that as Σ_t is linear in σ_t, the expectation of Σ_t is the expression for Σ_t with σ_t replaced by its posterior mean. Thus,

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

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

Remark 3. *Where we need to take the expectation of $\Sigma_t\nu_t$, we in fact have an expression that doesn't depend of σ_t anymore. We have replaced that with $\widehat{\Sigma_t}\widehat{\nu_t}$ to simplify the notation and understanding.*

□

2.3.3 Cross-validation

bla bla + notação + dizer que é no caso nao informativo, extender pro informativo num remark?

Proposition 2.3. *Let us consider s Gaussian Process $\{Z_t(x)\}_{t=1,\dots,s}$ as in the recursive model presented in (2.16) and $\mathcal{Z}^{(s)} = (\mathcal{Z}_1, \dots, \mathcal{Z}_s)$ with $\mathcal{Z}_t$ containing the values of $\{Z_t(x)\}_{x \in D_t}$ for $t = 1 \dots, s$ and $D_s \subseteq D_{s-1} \subseteq \dots D_1$. We denote by D_{test} a set consisting of points of index ξ_s of D_s and ξ_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 ε_{Z_t,ξ_t} are the errors (i.e. real 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*

$$(\varepsilon_{Z_t,\xi_t} - \widehat{\rho}_{t-1}(D_{test}) \odot \varepsilon_{Z_{t-1},\xi_{t-1}})[R_t^{-1}]_{[\xi_t,\xi_t]} = [R_t^{-1}(z_t - \mathcal{H}_t\lambda_{t,-\xi_t})]_{[\xi_t]}, \tag{2.33}$$

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

$$\lambda_{t,-\xi_t} = \left([\mathcal{H}_t]_{[-\xi_t]}^T K_t [\mathcal{H}_t]_{[-\xi_t]}\right)^{-1} [\mathcal{H}_t]_{-\xi_t}^T K_t z_t(D_t \backslash D_{test}), \tag{2.34}$$

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

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 σ_{Z_t,ξ_t}^2 the variances of the corresponding cross-validation procedure, we have

$$\sigma_{Z_t,\xi_t}^2 = \widehat{\sigma}_{\rho_{t-1},-\xi_t}^2(D_{test}) \odot \sigma_{Z_{t-1},\xi_{t-1}}^2 + \sigma_{t,-\xi_t}^2 \operatorname{diag}\left(\left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1}\right) + \mathcal{V}_t \tag{2.35}$$

with

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

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

and

$$\sigma_{t,-\xi_t}^2 = \frac{\left(z_t(D_t \backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}\lambda_{t,-\xi_t}\right)^T K_t \left(z_t(D_t \backslash D_{test}) - [\mathcal{H}_t]_{[-\xi_t]}\lambda_{t,-\xi_t}\right)}{n_t - p_t - q_{t-1} - n_{test}}, \tag{2.36}$$

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

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

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

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

Proof. We begin by ordering the points in D_t so that the points with index ξ_t are the n_{test} last points of D_t for every t.

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

Using the blockwise inversion formula (3.5), we have that

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

with $A = \left([R_t]_{[-\xi_t,-\xi_t]}\right)^{-1} + \left([R_t]_{[-\xi_t,-\xi_t]}\right)^{-1}[R_t]_{[-\xi_t,\xi_t]}\mathcal{Q}^{-1}[R_t]_{[\xi_t,-\xi_t]}\left([R_t]_{[-\xi_t,-\xi_t]}\right)^{-1}$, $B^T = -\mathcal{Q}^{-1}[R_t]_{[\xi_t,-\xi_t]}\left([R_t]_{[-\xi_t,-\xi_t]}\right)^{-1}$ and $\mathcal{Q} = [R_t]_{[\xi_t,\xi_t]} - [R_t]_{[\xi_t,-\xi_t]}\left([R_t]_{[-\xi_t,-\xi_t]}\right)^{-1}[R_t]_{[-\xi_t,\xi_t]}$.

Now, we must compute the prediction for the points in D_{test} at level 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: $\sigma_{t,-\xi_t}^2 = \frac{Q_t}{2(a_t-1)}$, $\lambda_{t,-\xi_t} = \Sigma_t \nu_t$ and $\mathcal{V}_t = 0$ (refer to equation (2.19 and compare to (2.30)), $\mathcal{V}_t$ is an additive term related to parameter estimations in the universal co-kriging case). Note that $\mathcal{Q} = \left([R_t^{-1}]_{[\xi_t,\xi_t]}\right)^{-1}$, 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} conditioned on the value of the process on the points in $D_t \backslash D_{test}$. Note that

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

Therefore, by equation (2.19), achieving (2.35) is straightforward. alguma inconsistencia com o rho

For the mean, by equation (2.18), we have that the predicted values in D_{test} are

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

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

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

We then add and subtract $h_t^T(D_{test}) = [g_{t-1}^T(D_{test}) \odot (\mu_{Z_{t-1}}(D_{test})\mathbf{1}_{q_{t-1}}^T(D_{test})) \quad f_t^T(D_{test})]$ to the previous equation. This 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 2.3.1, 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 2.3.1 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 previously known inverses or small matrices. For this, we will use block matrix inversion again. We write

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

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

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 (2.27), we promptly obtain the estimate for the variance parameter written in equation (2.36). Similarly, the trend parameters estimate in 2.34 is obtained using (2.25) and (2.26).

For the mean, we recall equation (2.29) with which we get the predictive mean when training on the set $D_t \backslash D_{test}$

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

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 highligh 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 (2.33).

The variance of the universal co-kriging is given by equation (2.30), 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

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

where

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

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

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

We, then, obtained the expression for $\mathcal{V}_t$ with

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

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

□

Chapter 3

Appendix

3.1 Gaussian Identities

Let x and y be jointly Gaussian random vectors

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} A & C \\ C^T & B \end{bmatrix} \right), \tag{3.1}$$

then the conditional distribution of x given y is

$$x|y \sim \mathcal{N}(\mu_x + CB^{-1}(y - \mu_y), A - CB^{-1}C^T). \tag{3.2}$$

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

$$\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) = Z^{-1}\mathcal{N}(x|c, C) \tag{3.3}$$

with $c = C(A^{-1}a+B^{-1}b)$, $C = (A^{-1}+B^{-1})^{-1}$ and $Z^{-1} = (2\pi)^{-D/2}\det(A+B)^{-1/2}\exp\{-\frac{1}{2}(a-b)^T(A+B)^{-1}(a-b)\}$.

3.2 Matrix identities

3.2.1 Matrix inversion lemma

This lemma, also known as Woodbury matrix identity states that

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{3.4}$$

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

3.2.2 Block matrix inversion

If a matrix M is partitioned into four blocks, it can be partitioned blockwise as

$$M^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1}, \end{bmatrix} \tag{3.5}$$

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

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \tag{3.6}$$

3.2.3 bla

Suppose that we have M, an $m \times m$ 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,

$$M^{-1}X = \begin{bmatrix} 0 \\ \mathbf{I}_n \end{bmatrix}. \tag{3.7}$$

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 $m_1 \times n$, M_2 of size $m_2 \times n$ and Z of size $n \times n$, then

$$M^{-1}X = Y \iff X = MY = M_1Y_1 + XZ + M_2Y_2 \iff Y_1 = 0,\ Z = \mathbf{I}_n \text{ and } Y_2 = 0. \tag{3.8}$$

3.3 Probability

3.3.1 Law of total expectation

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

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]. \tag{3.9}$$

3.3.2 Law of total variance

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

$$\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y|X]] + \mathrm{Var}[\mathbb{E}[Y|X]]. \tag{3.10}$$

3.4 Equations (1.7) and (1.8)

Using the Matrix inversion lemma (3.4), it is possible to rewrite the expressions for the mean and covariance to obtain more interpretable ones. Indeed, note that

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

For the expression of the predictive covariance, we have that

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

$$K_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1})(K_* + H^T BH_*) =$$

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

$$(K_*^T + H_*^T BH)K_y^{-1}(K_* + H^T BH_*) = K_*^T K_y^{-1} K_* + K_*^T K_y^{-1} H^T BH_* + H_*^T BHK_y^{-1} K_* +$$

$$H_*^T BHK_y^{-1} H^T BH_* =$$

$$K_*^T K_y^{-1} K_* + K_*^T K_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1}(B^{-1} + HK_y^{-1}H^T)BH_* +$$

$$H_*^T B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1} K_* + H_*^T BHK_y^{-1} H^T BH_*$$

The last term in the previous expression may be rewritten as:

$$H_*^T BHK_y^{-1} H^T BH_* = H_*^T B(B^{-1} + HK_y^{-1}H^T)BH_* - H_*^T BH_* =$$

$$H_*^T B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1}(B^{-1} + HK_y^{-1}H^T)BH_* - H_*^T BH_*.$$

In addition to this, we have that

$$(K_*^T + H_*^T BH)(K_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1})(K_* + H^T BH_*) =$$

$$K_*^T K_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1} K_* + K_*^T K_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1} H^T BH_* +$$

$$H_* BHK_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1} K_* + H_*^T BHK_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1} H^T BH_*.$$

Lastly, observe that

$$-H_*^T B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1}(B^{-1} + HK_y^{-1}H^T)BH_* +$$

$$H_*^T BHK_y^{-1} H^T (B^{-1} + HK_y^{-1}H^T)^{-1} HK_y^{-1} H^T BH_* = -H_*^T (B^{-1} + HK_y^{-1}H^T)^{-1} H_* - 2H_*^T BH_*.$$

Substituting all these terms back in the expression for $\mathrm{Cov}[\bar{w}_*]$, we readily obtain $\mathrm{Cov}[\bar{w}_*] = \mathrm{Cov}[\bar{z}_*] + R^T(B^{-1} + HK_y^{-1}H^T)^{-1}R$ with $\mathrm{Cov}[\bar{z}_*] = K(X_*, X_*) - K_*^T K_y^{-1} K_*$.

For the mean,

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

We can now identify the term $K_*^T K_y^{-1} y = \bar{z}_*$. For the term with y, note that

$$H_*^T B - (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1} =$$
$$\Big(H_*^T B(B^{-1} + HK_y^{-1}H^T) - (K_*^T + H_*^T BH)K_y^{-1}H^T\Big)(B^{-1} + HK_y^{-1}H^T)^{-1} = R^T(B^{-1} + HK_y^{-1}H^T)^{-1},$$

and for the term with b

$$\Big(H_*^T - (K_*^T + H_*^T BH)K_y^{-1}H^T + (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}HK_y^{-1}H^T\Big) =$$
$$\Big(H_* B(B^{-1} + HK_y^{-1}H^T)(B^{-1} + HK_y^{-1}H^T)^{-1}B^{-1}\Big) - \Big((K_*^T + H_*^T BH)K_y^{-1}H^T\Big) +$$
$$\Big((K_*^T + H_* BH)K_y^{-1}H^T - (K_*^T + H_*^T BH)K_y^{-1}H^T(B^{-1} + HK_y^{-1}H^T)^{-1}B^{-1}\Big) =$$
$$R^T(B^{-1} + HK_y^{-1}H^T)B^{-1}.$$

Using all these rewritten equations, we obtain $\bar{w}_* = \bar{z}_* + R^T\bar{\beta}$.

3.5 outras coisas ou encaixar no texto principal

em algum ponto nao tirei a transposta do r, renomear por R como em legratiet aqui e no texto do modelo classico, dizer na proposição 1 a equivalencia

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

Proposition 3.1. *If we consider the covariance matrix V_s in (2.10) 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*

$$V_s^{-1} = \begin{bmatrix} V_{s-1}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \frac{(\rho_{s-1}(D_s)\rho_{s-1}^T(D_s)) \odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} & -\begin{bmatrix} 0 \\ \frac{(\rho_{s-1}(D_s)\mathbf{1}_{n_s}^T) \odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} \\ -\begin{bmatrix} 0 & \frac{(\mathbf{1}_{n_s}\rho_{s-1}^T(D_s)) \odot R_s^{-1}}{\sigma_s^2} \end{bmatrix} & \frac{R_s^{-1}}{\sigma_s^2} \end{bmatrix}, \tag{3.11}$$

and

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

where V_{s-1} is an $(\sum_{i=1}^{s-1} n_i \times \sum_{i=1}^{s-1} n_i)$ 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 (3.5), we can write the inverse of V_s as

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

where $Q_s = V_{s,s} - U_{s-1}^TV_{s-1}^{-1}U_{s-1}$. From (2.11), 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 (3.7) 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 rewrite Q_s as something more familiar:

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

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

$$\mathcal{Z}_s = Z_s(D_s) = \rho_{s-1}(D_s) \odot Z_{s-1}(D_s) + \delta_s(D_s)$$

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

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

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

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

and this implies that

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

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 3.2. *If V_s is the covariance matrix in equation (2.10) and $k_s^T(x)$ the covariance vector in equation (2.8), the following equality is valid:*

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

Proof. In (2.8) and (2.9), we obtained a recursive expression for $k_s^T(x)$:

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

with

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

By Proposition 3.1,

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

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:

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

See that (2.6) implies that for $1 \leq t \leq s-1$,

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

$$\implies (c_1^T(x, D_1), \dots, c_{s-1}^T(x, D_{s-1}))^T = \rho_{s-1}(x)k_{s-1}^T(x) = \rho_{s-1}(x)\text{Cov}\{Z_{s-1}(x), \mathcal{Z}^{(s-1)}\}. \tag{3.12}$$

As in Proposition 3.1, 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 the 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

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

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

$$c_s^T(x, D_s) = \rho_{s-1}(D_s) \odot c_{s-1}^T(x, D_s) + \sigma_s^2 r_s(x, D_s) \tag{3.13}$$

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

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

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

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

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

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

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

And all together...

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

□

3.6 Parameter estimation of subsection 2.3.1

We'll use the approach presented in [Hoff '09] for the problem of finding the posterior of two parameters θ and γ when the priors 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 $p(\theta|X, \gamma)$ 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 of γ 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: We know that

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

and

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

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

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

(see to remarks (1)) and (2)).
Let $\mathcal{H}_1 = F_1$ and $\mathcal{H}_t = \big[G_{t-1} \odot [z_{t-1}(D_t)\mathbf{1}^T_{q_{t-1}} \quad F_t\big]$ 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 that $z^{(0)} = \emptyset$.

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,

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

$$\frac{1}{(2\pi)^{n_t/2}}\frac{1}{\sqrt{\det(\sigma_t^2 R_t)}} \exp\left\{-\frac{1}{2}\left((\tilde{\beta}_t - \Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t - \Sigma_t\nu_t) + z_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t - \nu_t^T\Sigma_t\nu_t\right)\right\}. \tag{3.14}$$

$$p(\tilde{\beta}_t|z^{(t)}, \sigma_t^2) \propto p(z_t|z^{(t-1),}\tilde{\beta}_t, \sigma_t^2)p(\tilde{\beta}_t|z^{(t-1)}, \sigma_t^2) \propto \exp\left\{ -\frac{1}{2}(\tilde{\beta}_t - \Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t - \Sigma_t\nu_t)\right\}.$$

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 $\tilde{\beta}_t$,

$$p(\tilde{\beta}_t|z^{(t)}, \sigma_t^2) \propto p(\tilde{\beta}_t|z^{(t-1)}, \sigma_t^2)p(z_t|\tilde{\beta}_t, \sigma_t^2) \propto \exp\bigg\{ -\frac{1}{2}(\tilde{\beta}_t - \Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t - \Sigma_t\nu_t)\bigg\}$$

$$\implies [\tilde{\beta}_t|z^{(t)}, \sigma_t^2] \sim \mathcal{N}(\Sigma_t\nu_t, \Sigma_t).$$

For the posterior of σ_t^2, we know that

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

$$\frac{1}{\sigma_t^2}\int \frac{1}{(2\pi)^{n_t/2}}\frac{1}{\sqrt{\det(\sigma_t^2 R_t)}}\exp\bigg\{-\frac{1}{2}\bigg((\tilde{\beta}_t - \Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t - \Sigma_t\nu_t) + z_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t - \nu_t^T\Sigma_t\nu_t\bigg)\bigg\}1d\tilde{\beta}_t$$

$$\propto \frac{1}{\sigma_t^2}\frac{1}{(\sigma_t^2)^{n_t/2}}\exp\bigg\{-\frac{1}{2}\bigg(z_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t - \nu_t^T\Sigma_t\nu_t\bigg)\bigg\}\int \exp\bigg\{-\frac{1}{2}\bigg((\tilde{\beta}_t - \Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t - \Sigma_t\nu_t)\bigg)\bigg\}d\tilde{\beta}_t$$

$$\propto \frac{1}{\sigma_t^2}\frac{1}{(\sigma_t^2)^{n_t/2}}\exp\bigg\{ -\frac{1}{2}\bigg(z_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t - \nu_t^T\Sigma_t\nu_t\bigg)\bigg\}\det(\Sigma_t)$$

$$\propto \frac{1}{\sigma_t^2}\frac{1}{(\sigma_t^2)^{n_t/2}}\exp\bigg\{ -\frac{1}{2}\bigg(z_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t - \nu_t^T\Sigma_t\nu_t\bigg)\bigg\}(\sigma_t^2)^{(p_t+q_{t-1})/2}.$$

Note that

$$z_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t - \nu_t^T\Sigma_t\nu_t = \frac{1}{\sigma_t^2}\bigg(z_t^T R_t^{-1}z_t - (\mathcal{H}_t^T R_t^{-1}z_t)^T\Big[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t\Big]^{-1}\mathcal{H}_t^T R_t^{-1}z_t\bigg),$$

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

$$\widehat{Q}_t = (z_t - \mathcal{H}_t\hat{\lambda}_t)^T R_t^{-1}(z_t - \mathcal{H}_t\hat{\lambda}_t) = z_t^T R_t^{-1}z_t - z_t^T R_t^{-1}\mathcal{H}_t[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1}\mathcal{H}_t^T R_t^{-1}z_t -$$

$$(\mathcal{H}_t[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1}\mathcal{H}_t^T R_t^{-1}z_t)^T R_t^{-1}z_t + (\mathcal{H}_t[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1}\mathcal{H}_t^T R_t^{-1}z_t)^T R_t^{-1}\mathcal{H}_t[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t]^{-1}\mathcal{H}_t^T R_t^{-1}z_t$$

$$= z_t^T R_t^{-1}z_t - (\mathcal{H}_t^T R_t^{-1}z_t)^T\Big[\mathcal{H}_t^T R_t^{-1}\mathcal{H}_t\Big]^{-1}\mathcal{H}_t^T R_t^{-1}z_t.$$

$$\therefore p(\sigma_t^2|z^{(t)}) \propto \frac{1}{(\sigma_t^2)^{(n_t-p_t-q_{t-1})/2+1}}\exp\bigg\{ -\frac{\widehat{Q}_t}{2}\bigg\}$$

$$\implies \sigma_t^2|z^{(t)} \sim \mathcal{IG}(a_t, \widehat{Q}_t/2)$$

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 (i), recycling many expressions we found there. First recall the likelihood given in equation (3.14). Then, observe that as a function of $\tilde{\beta}_t$ and σ_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's complete squares without all indexes and parameters, with C and D generic self-adjoint matrices and x, c and d vectors with appropriate dimensions:

$$(x-d)^TD(x-d)+(x-C^{-1}c)^TC(x-C^{-1}c)=$$

$$=x^T(D+C)x-x^TDd-(Dd)^Tx+d^TDd-x^Tc-c^Tx+c^TC^{-1}c=$$

$$=x^T(D+C)x-x^T(Dd+c)-(Dd+c)^Tx+d^TDd+c^TC^{-1}c=$$

$$=(x-(D+C)^{-1}(Dd+c))^T(D+C)(x-(D+C)^{-1}(Dd+c))+$$

$$d^TDd+c^TC^{-1}c-(Dd+c)^T(D+C)^{-1}(Dd+c)$$

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 $C=\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]$. If we call $D+C=\frac{W_t^{-1}}{\sigma_t^2}+\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t=\Sigma_t^{-1}$ 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 Σ_t and ν_t compared to the non-informative case), we have

$$p(\tilde{\beta}_t|z^{(t-1)},\sigma_t^2)p(z_t|\tilde{\beta}_t,\sigma_t^2)\propto\frac{1}{(\sigma_t^2)^{(n_t+p_t+q_{t-1})/2}}\exp\Big\{-\frac{1}{2}(\tilde{\beta}_t-\Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t-\Sigma_t\nu_t)\Big\}\times$$

$$\exp\Big\{-\frac{1}{2}\Big(b_t^T\frac{W^{-1}}{\sigma_t^2}b_t+\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]^T\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]^{-1}\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]-\nu_t^T\Sigma_t\nu_t\Big)\Big\}\exp\Big\{-\frac{\widehat{Q}_t}{2\sigma_t^2}\Big\}.$$

Now, for simplicity, let's call $\mathcal{H}^TR_t^{-1}z_t=v$ and $\mathcal{H}_t^TR_t^{-1}\mathcal{H}_t=S$ and let's drop the index t. We will use the matrix inversion lemma (3.4) 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

$$b^TW^{-1}b+v^TS^{-1}v-(W^{-1}b+v)^T(W^{-1}+S)^{-1}(W^{-1}b+v)=b^TW^{-1}b+v^TS^{-1}v-$$

$$\big(b^TW^{-1}(W^{-1}+S)^{-1}W^{-1}b+v^T(W^{-1}+S)^{-1}W^{-1}b+b^TW^{-1}(W^{-1}+S)^{-1}v+v^T(W^{-1}+S)^{-1}v\big),$$

and that for the last 4 terms we have

$$b^TW^{-1}(W^{-1}+S)^{-1}W^{-1}b=b^TW^{-1}(W-W(W+S^{-1})^{-1}W)W^{-1}b=b^TW^{-1}b-b^T(W+S^{-1})^{-1}b,$$

$$v^T(W^{-1}+S)^{-1}W^{-1}b=v^TS^{-1}(W+S^{-1})^{-1}b,$$

$$b^TW^{-1}(W^{-1}+S)^{-1}v=b^T(W+S^{-1})^{-1}S^{-1}v,$$

and

$$v^T(W^{-1}+S)^{-1}v=v^T(S^{-1}-S^{-1}(W+S^{-1})^{-1}S^{-1})v=v^TS^{-1}v-v^TS^{-1}(W+S^{-1})^{-1}S^{-1}v.$$

Therefore, it is clear that

$$b^TW^{-1}b+v^TS^{-1}v-(W^{-1}b+v)^T(W^{-1}+S)^{-1}(W^{-1}b+v)=(b-S^{-1}v)^T(W+S^{-1})^{-1}(b-S^{-1}v)$$

and

$$\exp\Big\{-\frac{1}{2}\Big(b_t^T\frac{W^{-1}}{\sigma_t^2}b_t+\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]^T\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}\mathcal{H}_t\Big]^{-1}\Big[\mathcal{H}_t^T\frac{R_t^{-1}}{\sigma_t^2}z_t\Big]-\nu_t^T\Sigma_t\nu_t\Big)\Big\}=$$

$$\exp\Big\{-\frac{1}{2\sigma_t^2}\Big((b_t-S^{-1}v)^T(W_t+S^{-1})^{-1}(b_t-S^{-1}v)\Big)\Big\}=$$

$$\exp\Big\{-\frac{1}{2\sigma_t^2}\Big((b_t-\hat{\lambda}_t)^T(W_t+[\mathcal{H}_t^TR_t^{-1}\mathcal{H}_t]^{-1})^{-1}(b_t-\hat{\lambda}_t)\Big)\Big\}$$

By now, we already have all necessary expressions. Now, paying attention to what goes into the multiplicative constant, it is easy to obtain the posterior of $\tilde{\beta}_t$:

$$p(\tilde{\beta}_t|z^{(t)},\sigma_t^2)\propto p(\tilde{\beta}_t|z^{(t-1)},\sigma_t^2)p(z_t|\tilde{\beta}_t,\sigma_t^2)\propto\exp\Big\{-\frac{1}{2}(\tilde{\beta}_t-\Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t-\Sigma_t\nu_t)\Big\}$$

$$\implies\tilde{\beta}_t|z^{(t)},\sigma_t^2\sim\mathcal{N}(\Sigma_t\nu_t,\Sigma_t).$$

For the posterior of σ_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

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

$$\propto\frac{1}{(\sigma_t^2)^{\alpha_t+1}}\exp\Big\{-\frac{\gamma_t}{\sigma_t^2}\Big\}\int\frac{1}{(\sigma_t^2)^{(n_t+p_t+q_{t-1})/2}}\exp\Big\{-\frac{1}{2}(\tilde{\beta}_t-\Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t-\Sigma_t\nu_t)\Big\}$$

$$\exp\Big\{-\frac{1}{2\sigma_t^2}\Big((b_t-\hat{\lambda}_t)^T(W_t+[\mathcal{H}_t^TR_t^{-1}\mathcal{H}_t]^{-1})^{-1}(b_t-\hat{\lambda}_t)\Big)\Big\}\exp\Big\{-\frac{\widehat{Q}_t}{2\sigma_t^2}\Big\}d\tilde{\beta}_t=$$

$$\frac{1}{(\sigma_t^2)^{\alpha_t+(n_t+p_t+q_{t-1})/2+1}}\exp\left\{-\frac{1}{2\sigma_t^2}\left(2\gamma_t+(b_t-\hat{\lambda}_t)^T(W_t+[\mathcal{H}_t^TR_t^{-1}\mathcal{H}_t]^{-1})^{-1}(b_t-\hat{\lambda}_t)+\widehat{Q}_t\right)\right\}$$

$$\int\exp\left\{-\frac{1}{2}(\tilde{\beta}_t-\Sigma_t\nu_t)^T\Sigma_t^{-1}(\tilde{\beta}_t-\Sigma_t\nu_t)\right\}d\tilde{\beta}_t=\frac{1}{(\sigma_t^2)^{\alpha_t+(n_t+p_t+q_{t-1})/2+1}}\exp\left\{-\frac{1}{2\sigma_t^2}Q_t\right\}\sqrt{\det(\Sigma_t)}$$

$$\propto\frac{1}{(\sigma_t^2)^{\alpha_t+n_t/2+1}}\exp\Big\{-\frac{1}{2\sigma_t^2}Q_t\Big\}.$$

This way, we have $\sigma_t^2|z^{(t)}\sim\mathcal{IG}(n_t/2+\alpha_t,Q_t/2)$.

Bibliography

[] Books:

[Adler '09] Robert J. Adler *The Geometry of Random Fields*, Siam, 2009.

[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:

[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.

[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.

[Patterson & Thompson '71] H. D. Patterson & Robin Thompson *Recovery of interblock information when block sizes are unequal*, Biometrika, 58, 1971.

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.