



HAL
open science

An analytic comparison of regularization methods for Gaussian Processes

Hossein Mohammadi, Rodolphe Le Riche, Nicolas Durrande, Eric Touboul,
Xavier Bay

► **To cite this version:**

Hossein Mohammadi, Rodolphe Le Riche, Nicolas Durrande, Eric Touboul, Xavier Bay. An analytic comparison of regularization methods for Gaussian Processes. [Research Report] Ecole Nationale Supérieure des Mines de Saint-Etienne; LIMOS. 2016. hal-01264192v1

HAL Id: hal-01264192

<https://hal.science/hal-01264192v1>

Submitted on 2 Feb 2016 (v1), last revised 5 May 2017 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An analytic comparison of regularization methods for Gaussian Processes

Hossein Mohammadi^{1, 2}, Rodolphe Le Riche^{*2, 1}, Nicolas Durrande^{1, 2},
Eric Touboul^{1, 2}, and Xavier Bay^{1, 2}

¹Ecole des Mines de Saint Etienne, H. Fayol Institute

²CNRS LIMOS, UMR 5168

Abstract

Gaussian Processes (GPs) are often used to predict the output of a parameterized deterministic experiment. They have many applications in the field of Computer Experiments, in particular to perform sensitivity analysis, adaptive design of experiments and global optimization. Nearly all of the applications of GPs to Computer Experiments require the inversion of a covariance matrix. Because this matrix is often ill-conditioned, regularization techniques are required. Today, there is still a need to better regularize GPs.

The two most classical regularization methods are *i*) pseudoinverse (PI) and *ii*) nugget (or jitter or observation noise). This article provides algebraic calculations which allow comparing PI and nugget regularizations. It is proven that pseudoinverse regularization averages the output values and makes the variance null at redundant points. On the opposite, nugget regularization lacks interpolation properties but preserves a non-zero variance at every point. However, these two regularization techniques become similar as the nugget value decreases. A distribution-wise GP is introduced which interpolates Gaussian distributions instead of data points and mitigates the drawbacks of pseudoinverse and nugget regularized GPs. Finally, data-model discrepancy is discussed and serves as a guide for choosing a regularization technique.

Keywords: Degeneracy of covariance matrices; Gaussian process regression; Kriging; Regularization.

*Corresponding author: Ecole Nationale Supérieure des Mines de Saint Etienne, Institut H. Fayol, 158, Cours Fauriel, 42023 Saint-Etienne cedex 2 - France
Tel : +33477420023
Email: leriche@emse.fr

Nomenclature

Abbreviations

CV, Cross-validation
GP, Gaussian Process
ML, Maximum Likelihood
PI, Pseudoinverse

Greek symbols

δ , nugget value.
 Δ , the difference between two likelihood functions.
 κ , condition number of a matrix.
 κ_{max} , maximum condition number after regularization.
 λ_i , the i th largest eigenvalue of the covariance matrix.
 $\mu(\cdot)$, Gaussian process mean.
 ω_i , i th weight of a linear combination.
 σ^2 , process variance.
 Σ , diagonal matrix made of covariance matrix eigenvalues.
 τ , tolerance of pseudoinverse.
 θ_i , characteristic length-scale in dimension i .

Latin symbols

\mathbf{c} , vector of covariances between a new point and the design points \mathbf{X} .
 \mathbf{C} , covariance matrix of process $Y(\cdot)$ evaluated at \mathbf{X} .
 \mathbf{C}^i , i th column of \mathbf{C} .
 $discr$, model-data discrepancy.
 \mathbf{e}_i , i th unit vector.
 $f: \mathbb{R}^d \rightarrow \mathbb{R}$, true function, to be predicted.
 K , kernel or covariance function.
 $\tilde{\mathbf{K}}$, correlation matrix.
 n , number of design points.
 N , number of redundant points.
 \mathbf{P}_{Im} , orthogonal projection matrix onto the image space of a matrix (typically \mathbf{C}).
 \mathbf{P}_{Nul} , orthogonal projection matrix onto the null space of a matrix (typically \mathbf{C}).
 r , rank of the matrix \mathbf{C} .
 s_i^2 , variance of response values at i -th repeated point.
 \mathbf{V} , column matrix of eigenvectors of \mathbf{C} associated to strictly positive eigenvalues.
 \mathbf{W} , column matrix of eigenvectors of \mathbf{C} associated to zero eigenvalues.
 \mathbf{X} , matrix of design points.
 $\mathbf{Y}(\mathbf{x})$, Gaussian process.
 \mathbf{y} , vector of response or output values.
 \bar{y}_i , mean of response values at i -th repeated point.

1 Introduction

Conditional Gaussian Processes, also known as kriging models, are commonly used for predicting from a set of spatial observations. Kriging performs a linear combination of the observed response values. The weights in the combination depend only, through a covariance function, on the locations of the points where one wants to predict and that of the points where samples are available [8, 23, 29, 20].

Although GPs can model stochastic or deterministic spatial phenomena, the focus of this paper is on experiments with deterministic outputs. Computer simulations provide examples of such noiseless experiments. Furthermore, we assume that the location of the observed points and the covariance function are given a priori. This occurs frequently within algorithms performing adaptive design of experiments [4], global sensitivity analysis [16] and global optimization [14].

Kriging models require the inversion of a covariance matrix which is made of the covariance function evaluated at every pair of observed locations. In practice, anyone who has used a kriging model has experienced one of the circumstances where the covariance matrix is not numerically invertible. This happens when observed points are repeated, or even are close to each other, or when the covariance function makes the information provided by observations redundant.

In the literature, various strategies have been employed to avoid degeneracy of the covariance matrix. A first set of approaches proceed by controlling the locations of design points (the Design of Experiments or DoE). The influence of the DoE on the condition number of the covariance matrix has been investigated in [24]. [21] proposes to build kriging models from a uniform subset of design points to improve the condition number. In [17], new points are taken suitably far from all existing data points to guarantee a good conditioning.

Other strategies select the covariance function so that the covariance matrix remains well-conditioned. In [9] for example, the influence of all kriging parameters on the condition number, including the covariance function, is discussed. Ill-conditioning also happens in the related field of linear regression with the Gauss-Markov matrix $\Phi^T \Phi$ that needs to be inverted, where Φ is the matrix of basis functions evaluated at the DoE. In regression, work has been done on diagnosing ill-conditioning and the solution typically involves working on the definition of the basis functions to recover invertibility [5]. The link between the choice of the basis functions and the choice of the covariance functions is given by Mercer’s theorem, [20].

Instead of directly inverting the covariance matrix, an iterative method has been proposed in [11] to solve the kriging equations and avoid numerical instabilities.

Two generic solutions to overcome the degeneracy of covariance matrix are the pseudoinverse (PI) and the “nugget” regularizations. They have a wide range of applications because, contrarily to the methods mentioned above, they

can be used a posteriori in computer experiments algorithms without major redesign of the methods. This is the reason why most kriging implementations contain PI or nugget regularization.

The singular value decomposition and the idea of pseudoinverse have already been suggested in [14] to overcome degeneracy. The Model-Assisted Pattern Search (MAPS) software [26] relies on an implementation of the pseudoinverse to invert the (covariance) matrices.

The most often used approach to deal with ill-conditioning in the covariance is to introduce a “nugget” [7, 25, 15, 1], that is to say add a small positive scalar on the covariance diagonal. The popularity of the nugget regularization may be either due to its simplicity or to its interpretation as the variance of a noise on the observations. The value of the nugget term can be estimated by maximum likelihood (ML). It is reported in [18] that the presence of a nugget term significantly changes the modes of the likelihood function of a GP. Similarly in [12], the authors have advocated a nonzero nugget term in the design and analysis of their computer experiments. They have also stated that estimating a nonzero nugget value may improve some statistical properties of the kriging models such as their predictive accuracy [13]. However, some references like [19] recommend that the magnitude of nugget remains as small as possible to preserve the interpolation property.

Because of the diversity of arguments regarding GP regularization, we feel that there is a need to provide analytical explanations on the effects of the main approaches. This paper provides new results regarding the analysis and comparison of pseudoinverse and nugget kriging regularizations in the context of deterministic outputs. Our analysis is made possible by approximating ill-conditioned covariance matrices with the neighboring truly non-invertible covariance matrices that stem from redundant points. Some properties of kriging regularized by PI and nugget are stated and proved. The paper finishes with the description of a new type of regularization associated to a distribution-wise GP.

2 Kriging models and degeneracy of the covariance matrix

2.1 Context: conditional Gaussian processes

This section contains a summary of conditional GP concepts and notations. Readers who are familiar with GP may want to proceed to the next section (2.2).

Let f be a real-valued function defined over $D \subseteq \mathbb{R}^d$. Assume that the values of f are known at a limited set of points called design points. One wants to infer the value of this function elsewhere. Conditional GP is one of the most important technique for this purpose [23, 29].

A GP defines a distribution over functions. Formally, a GP indexed by D is a collection of random variables $(Y(\mathbf{x}); \mathbf{x} \in D)$ such that for any $n \in \mathbb{N}$ and any

$\mathbf{x}^1, \dots, \mathbf{x}^n \in D$, $(Y(\mathbf{x}^1), \dots, Y(\mathbf{x}^n))$ follows a multivariate Gaussian distribution. The distribution of the GP is fully characterized by a mean function $\mu(\mathbf{x}) = \mathbb{E}(Y(\mathbf{x}))$ and a covariance function $K(\mathbf{x}, \mathbf{x}') = \text{Cov}(Y(\mathbf{x}), Y(\mathbf{x}'))$ [20].

The choice of kernel plays a key role in the obtained kriging model. In practice, a parametric family of kernels is selected (e.g., Matérn, polynomial, exponential) and then the unknown kernel parameters are estimated from the observed values. For example, a separable squared exponential kernel is expressed as

$$K(\mathbf{x}, \mathbf{x}') = \sigma^2 \prod_{i=1}^d \exp\left(-\frac{|x_i - x'_i|^2}{2\theta_i^2}\right). \quad (1)$$

In the above equation, σ^2 is a scaling parameter known as process variance and x_i is the i th component of \mathbf{x} . The parameter θ_i is called length-scale and determines the correlation length along coordinate i . It should be noted that $\text{Cov}(Y(\mathbf{x}), Y(\mathbf{x}'))$ in Equation (1) is only a function of the difference between \mathbf{x} and \mathbf{x}' . A GP with this property is said to be stationary, otherwise it is nonstationary. Interested readers are referred to [20] for further information about GPs and kernels.

Let $Y(\mathbf{x})_{\mathbf{x} \in D}$ be a GP with kernel $K(\cdot, \cdot)$ and zero mean ($\mu(\cdot) = 0$). $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^n)$ denotes the n data points where the samples are taken and the corresponding response values are $\mathbf{y} = (y_1, \dots, y_n)^\top = (f(\mathbf{x}^1), \dots, f(\mathbf{x}^n))^\top$. The posterior distribution of the GP ($Y(\mathbf{x})$) knowing it interpolates the data points is still Gaussian with mean and covariance [20]

$$m_K(\mathbf{x}) = \mathbb{E}(Y(\mathbf{x})|Y(\mathbf{X}) = \mathbf{y}) = \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1} \mathbf{y}, \quad (2)$$

$$\begin{aligned} c_K(\mathbf{x}, \mathbf{x}') &= \text{Cov}(Y(\mathbf{x}), Y(\mathbf{x}')|Y(\mathbf{X}) = \mathbf{y}) \\ &= K(\mathbf{x}, \mathbf{x}') - \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1} \mathbf{c}(\mathbf{x}'), \end{aligned} \quad (3)$$

where $\mathbf{c}(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}^1), \dots, K(\mathbf{x}, \mathbf{x}^n))^\top$ is the vector of covariances between a new point \mathbf{x} and the n already observed sample points. The $n \times n$ matrix \mathbf{C} is a covariance matrix between the data points and its elements are defined as $\mathbf{C}_{i,j} = K(\mathbf{x}^i, \mathbf{x}^j) = \sigma^2 \mathbf{K}$, where \mathbf{K} is the correlation matrix. Hereinafter, we call $m_K(\mathbf{x})$ and $v_K(\mathbf{x}) = c_K(\mathbf{x}, \mathbf{x})$ the kriging mean and variance, respectively.

One essential question is how to estimate the unknown parameters in the covariance function. Typically, the values of the model parameters (i.e., σ and the θ_i 's) are learned via maximization of the likelihood. The likelihood function of the unknown parameters given observations $\mathbf{y} = (y_1, \dots, y_n)^\top$ is defined as follows:

$$L(\mathbf{y}|\boldsymbol{\theta}, \sigma^2) = \frac{1}{(2\pi)^{n/2} |\mathbf{C}|^{1/2}} \exp\left(-\frac{\mathbf{y}^\top \mathbf{C}^{-1} \mathbf{y}}{2}\right). \quad (4)$$

In the above equation, $|\mathbf{C}|$ indicates the determinant of the covariance matrix \mathbf{C} and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^\top$ is a vector made of the length-scales in each dimension. It is usually more convenient to work with the natural logarithm of the likelihood function that is:

$$\ln L(\mathbf{y}|\boldsymbol{\theta}, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{C}| - \frac{1}{2} \mathbf{y}^\top \mathbf{C}^{-1} \mathbf{y}. \quad (5)$$

The ML estimator of the process variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n} \mathbf{y}^\top \tilde{\mathbf{K}}^{-1} \mathbf{y}, \quad (6)$$

and if it is inserted in (5), it yields (minus) the concentrated log-likelihood,

$$-2 \ln L(\mathbf{y}|\boldsymbol{\theta}, \sigma^2) = n \ln(2\pi) + n \ln \hat{\sigma}^2 + \ln |\tilde{\mathbf{K}}| + n. \quad (7)$$

Finally, $\boldsymbol{\theta}$ is estimated by numerically minimizing Equation (7).

2.2 Degeneracy of the covariance matrix

Computing the kriging mean (Equation (2)) or (co)variance (Equation (3)) or even samples of GP trajectories, requires inverting the covariance matrix \mathbf{C} . In practice, the covariance matrix should not only be invertible, but also well-conditioned. A matrix is said to be near singular or ill-conditioned or degenerated if its condition number is too large. For covariance matrices, which are symmetric and positive semidefinite, the condition number $\kappa(\mathbf{C})$ is the ratio of the largest to the smallest eigenvalue.

There are many situations where the covariance matrix is near singular. The most frequent and easy to understand case is when some data points are too close to each other, where closeness is measured with respect to the metric induced by the covariance function. This is a recurring issue in sequential DoEs like the EGO algorithm [14] where the search points tend to pile up around the points of interest such as the global optimum [19]. When this happens, the resulting covariance matrix is no longer numerically invertible because some columns are almost identical.

In this paper, to analyze PI and nugget regularizations, we are going to consider matrix degeneracy pushed to its limit, that is true non-invertibility (or rank deficiency) of \mathbf{C} . Non invertibility happens if a linear dependency exists between \mathbf{C} 's columns (or rows). Appendix A provides a collection of examples where the covariance matrix is not invertible with calculation details that will become clear later. Again, the easiest to understand and the most frequent occurrence of \mathbf{C} 's rank deficiency is when some of the data points tend towards each other until they are at the same \mathbf{x}^i position. They form *repeated* points, the simplest example of what we more generally call redundant points which will be formally defined shortly. Figure 9 in Appendix A.1 is an example of repeated points. Repeated points lead to strict non-invertibility of \mathbf{C} since the corresponding columns are identical. The special case of repeated points will be instrumental in understanding some aspects of kriging regularization in Sections 3.2 and 4.2 because the eigenvectors of the covariance matrix associated to null eigenvalues are known.

The covariance matrix of GPs may lose invertibility even though the data points are not close to each other in Euclidean distance. This occurs for example with additive GPs for which the kernel is the sum of kernels defined in each

dimension, $K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^d K_i(x_i, x'_i)$. The additivity of a kernel may lead to linear dependency in some columns of the covariance matrix. For example, in the DoE shown in Figure 5, only three of the first four points which form a rectangle provide independent information in the sense that the GP response at any of the four points is fully defined by the response at the three other points. This is explained by a linear dependency between the first four columns, $\mathbf{C}^4 = \mathbf{C}^3 + \mathbf{C}^2 - \mathbf{C}^1$, which comes from the additivity of the kernel and the rectangular design [10]:

$$\mathbf{C}_i^4 = \text{Cov}(x_1^i, x_1^4) + \text{Cov}(x_2^i, x_2^4) = \text{Cov}(x_1^i, x_1^2) + \text{Cov}(x_2^i, x_2^3),$$

and completing the covariances while accounting for $x_2^2 = x_2^1$, $x_1^3 = x_1^1$, yields

$$\mathbf{C}_i^4 = \text{Cov}(\mathbf{x}^i, \mathbf{x}^3) + \text{Cov}(\mathbf{x}^i, \mathbf{x}^2) - \text{Cov}(\mathbf{x}^i, \mathbf{x}^1) = \mathbf{C}_i^3 + \mathbf{C}_i^2 - \mathbf{C}_i^1.$$

Note that if the measured outputs y^1, \dots, y^4 are not additive ($y^4 \neq y^2 + y^3 - y^1$), none of the four measurements can be easily deleted without loss of information, hence the need for the general regularization methods that will be discussed later.

Periodic kernels may also yield non-invertible covariance matrices although data points are far from each other. This is illustrated in Figure 12 (and further detailed in Appendix A.4) where points 1 and 2, and points 3 and 4, provide the same information as they are one period away from each other. Thus, $\mathbf{C}^1 = \mathbf{C}^2$ and $\mathbf{C}^3 = \mathbf{C}^4$.

Our last example comes from the dot product (or linear) kernel (cf. Appendix A.5). Because the GP trajectories and mean are linear, no uncertainty is left in the model when the number of data points n reaches $d + 1$ and when $n > d + 1$ the covariance matrix is no longer invertible.

2.3 Eigen analysis and definition of redundant points

We start by introducing our notations for the eigendecomposition of the covariance matrix. Let the $n \times n$ covariance matrix \mathbf{C} have rank r , $r \leq n$. A covariance matrix is positive semidefinite, thus its eigenvalues are greater than or equal to zero. The eigenvectors associated to strictly positive eigenvalues are denoted \mathbf{V}^i , $i = 1, \dots, r$, and those associated to null eigenvalues are \mathbf{W}^i , $i = 1, \dots, (n - r)$, that is $\mathbf{C}\mathbf{V}^i = \lambda_i \mathbf{V}^i$ where $\lambda_i > 0$ and $\mathbf{C}\mathbf{W}^i = \mathbf{0}$. The eigenvectors are grouped columnwise into the matrices $\mathbf{V} = [\mathbf{V}^1, \dots, \mathbf{V}^r]$ and $\mathbf{W} = [\mathbf{W}^1, \dots, \mathbf{W}^{n-r}]$. In short, the eigenvalue decomposition of the covariance matrix \mathbf{C} obeys

$$\mathbf{C} = [\mathbf{V} \ \mathbf{W}] \boldsymbol{\Sigma} [\mathbf{V} \ \mathbf{W}]^\top, \quad (8)$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix containing the eigenvalues of \mathbf{C} , $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > 0$ and $\lambda_{r+1} = \dots = \lambda_n = 0$. \mathbf{V} spans the image space and \mathbf{W} spans the null space of \mathbf{C} , $\text{Im}(\mathbf{C})$ and $\text{Nul}(\mathbf{C})$, respectively. $[\mathbf{V} \ \mathbf{W}]$ is an orthogonal matrix,

$$[\mathbf{V} \ \mathbf{W}]^\top [\mathbf{V} \ \mathbf{W}] = [\mathbf{V} \ \mathbf{W}] [\mathbf{V} \ \mathbf{W}]^\top = \mathbf{V}\mathbf{V}^\top + \mathbf{W}\mathbf{W}^\top = \mathbf{I}. \quad (9)$$

$\mathbf{V}\mathbf{V}^\top$ is the orthogonal projection matrix onto $Im(\mathbf{C})$. Similarly, $\mathbf{W}\mathbf{W}^\top$ is the orthogonal projection matrix onto $Nul(\mathbf{C})$. For a given matrix \mathbf{C} , the eigenvectors \mathbf{W}^i are not uniquely defined because any linear combination of them is also an eigenvector associated to a null eigenvalue. However, the orthogonal projection matrices onto the image and null spaces of \mathbf{C} are unique and will be cornerstones in the definition of redundant points.

Before formally defining redundant points, we present the examples of singular covariance matrices of Appendix A. These examples are two dimensional to allow for a graphical representation. The kernels, designs of points, eigenvalues and eigenvectors and the $\mathbf{V}\mathbf{V}^\top$ projection matrix are given.

The first example detailed in Appendix A.1 has two groups of repeated data points (points 1, 2 and 6, on the one hand, points 3 and 4, on the other hand), in which there are 3 redundant, points. The covariance matrix has 3 null eigenvalues. It should be noted that the off-diagonal coefficients of the $\mathbf{V}\mathbf{V}^\top$ projection matrix associated to the indices of repeated points are not 0.

Appendix A.2 shows how additive kernels may generate singular covariance matrices: points 1, 2, 3 and 4 are arranged in a rectangular pattern which makes columns 1 to 4 linearly dependent (as already explained in Section 2.2). The additive property makes any one of the 4 points of a rectangular pattern redundant in that the value of the GP there is uniquely set by the knowledge of the GP at the 3 other points. The same stands for points 5 to 8. Two points are redundant (1 in each rectangle) and there are two null eigenvalues. Again, remark how the off-diagonal coefficients of $\mathbf{V}\mathbf{V}^\top$ associated to the points of the rectangles are not zero. Another example of additivity and singularity is depicted in Section A.3: although the design points are not set in a rectangular pattern, there is a shared missing vertex between two orthogonal triangles so that, because of additivity, the value at this missing vertex is defined twice. In this case, there is one redundant point, one null eigenvalue, and all the points of the design are coupled: all off-diagonal terms in $\mathbf{V}\mathbf{V}^\top$ are not zero.

Finally, Section A.4 is a case with a periodic kernel and a periodic pattern of points so that points 1 and 2 provide the same information, and similarly with points 3 and 4. There are 2 null eigenvalues, and the (1,2) and (3,4) off-diagonal terms in $\mathbf{V}\mathbf{V}^\top$ are not zero.

In general, we call *redundant* the set of data points that make the covariance matrix non-invertible by providing linearly dependent information.

Definition 1 (Redundant points set)

Let \mathbf{C} be the $n \times n$ non-invertible positive semidefinite covariance matrix of a Gaussian process. It has rank r , $r < n$. \mathbf{V} is the $n \times r$ matrix of the eigenvectors associated to strictly positive eigenvalues. R is a set of at least two redundant points indices if for any i and j in R , $(\mathbf{V}\mathbf{V}^\top)_{ij} \neq 0$.

Redundant points could be equivalently defined with the \mathbf{W} matrix since, from Equation (9), $\mathbf{V}\mathbf{V}^\top$ and $\mathbf{W}\mathbf{W}^\top$ have the same non-zero off-diagonal terms with opposite signs. Subsets of redundant points are also redundant. The *degree of redundancy* of a set of points R is the number of zero eigenvalues of the

covariance matrix restricted to the points in R , i.e., $[\mathbf{C}_{ij}]$ for all $(i, j) \in R^2$. The degree of redundancy is the number of points that should be removed from R to recover invertibility of the covariance restricted to the points in R . When $r = n$, \mathbf{C} is invertible and there is no redundant point. An interpretation of redundant points will be made in the next Section on pseudoinverse regularization.

In the repeated points example of Appendix A.1, the two largest redundant points sets are $\{1, 2, 6\}$ and $\{3, 4\}$ with degrees of redundancy 2 and 1, respectively. The first additive example (Appendix A.2) has two sets of redundant points, $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8\}$ each with a degree of redundancy equal to 1. In the second additive example (Appendix A.3), all the points are redundant with a degree equal to 1. The periodic case (Appendix A.4) has two sets of redundant points of degree 1, $\{1, 2\}$ and $\{3, 4\}$. With the linear kernel (Appendix A.5) all data points are redundant and in the given example where $n = d + 2$ the degree of redundancy is 1.

3 Pseudoinverse regularization

3.1 Definition

In this Section, we state well-known properties of pseudoinverse matrices without proofs (which can be found, e.g., in [6]) and apply them to the kriging equations (2) and (3). Pseudoinverse matrices are generalizations of the inverse matrix. The most popular pseudoinverse is the *Moore–Penrose pseudoinverse* which is hereinafter referred to as pseudoinverse.

When \mathbf{C}^{-1} exists (i.e., \mathbf{C} has full rank, $r = n$), we denote as $\boldsymbol{\beta}$ the term $\mathbf{C}^{-1}\mathbf{y}$ of the kriging mean formula, Equation (2). More generally, when \mathbf{C} is not a full rank matrix, we are interested in the vector $\boldsymbol{\beta}$ that simultaneously minimizes $\|\mathbf{C}\boldsymbol{\beta} - \mathbf{y}\|_2$ and $\|\boldsymbol{\beta}\|_2$. This solution is unique and obtained by $\boldsymbol{\beta}^{PI} = \mathbf{C}^\dagger\mathbf{y}$ where \mathbf{C}^\dagger is the pseudoinverse of \mathbf{C} . Each vector $\boldsymbol{\beta}$ can be uniquely decomposed into

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{PI} + \boldsymbol{\beta}_{Nul(\mathbf{C})}, \quad (10)$$

where $\boldsymbol{\beta}^{PI}$ and $\boldsymbol{\beta}_{Nul(\mathbf{C})}$ belong to the image space and the null space of the covariance matrix, respectively. The decomposition is unique since, \mathbf{C} being symmetric, $Im(\mathbf{C})$ and $Nul(\mathbf{C})$ have no intersection.

The pseudoinverse of \mathbf{C} is expressed as

$$\mathbf{C}^\dagger = [\mathbf{V} \ \mathbf{W}] \begin{bmatrix} \text{diag}(\frac{1}{\lambda})_{r \times r} & \mathbf{0}_{r \times (n-r)} \\ \mathbf{0}_{(n-r) \times r} & \mathbf{0}_{(n-r) \times (n-r)} \end{bmatrix} [\mathbf{V} \ \mathbf{W}]^\top, \quad (11)$$

where $\text{diag}(\frac{1}{\lambda})$ is a diagonal matrix with $\frac{1}{\lambda_i}$, $i = 1, \dots, r$, as diagonal elements. So $\boldsymbol{\beta}^{PI}$ reads

$$\boldsymbol{\beta}^{PI} = \sum_{i=1}^r \frac{(\mathbf{V}^i)^\top \mathbf{y}}{\lambda_i} \mathbf{V}^i. \quad (12)$$

Equation (12) indicates that β^{PI} is in the image space of \mathbf{C} , because it is a linear combination of eigenvectors associated to positive eigenvalues. A geometrical interpretation of β^{PI} and pseudo-inverse is given in Figure 1. The kriging mean (Equation (2)) with PI regularization can be written as

$$m^{PI}(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top \sum_{i=1}^r \frac{(\mathbf{V}^i)^\top \mathbf{y}}{\lambda_i} \mathbf{V}^i. \quad (13)$$

Similarly, the kriging covariance (3) regularized by PI is,

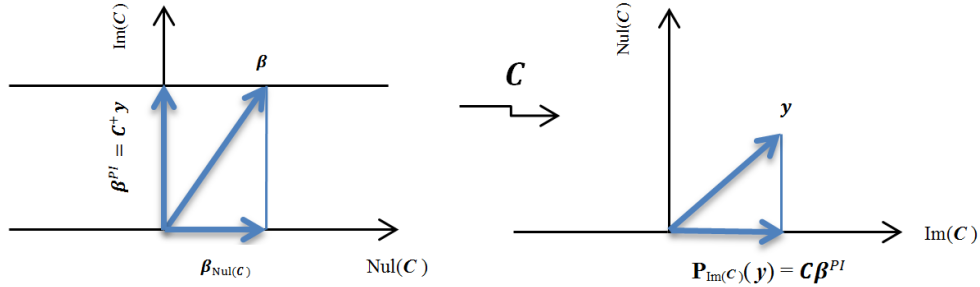


Figure 1: Geometrical interpretation of the Moore-Penrose pseudoinverse. In the leftmost picture, infinitely many vectors β are solutions to the system $\mathbf{C}\beta = \mathbf{y}$. But the minimum norm solution is $\mathbf{C}^\dagger \mathbf{y}$. The rightmost picture shows the orthogonal projection of \mathbf{y} onto the image space of \mathbf{C} , $\mathbf{P}_{Im(\mathbf{C})}(\mathbf{y})$, which is equal to $\mathbf{C}\mathbf{C}^\dagger \mathbf{y}$ (Property 1).

$$\begin{aligned} c^{PI}(\mathbf{x}, \mathbf{x}') &= K(\mathbf{x}, \mathbf{x}') - \mathbf{c}(\mathbf{x})^\top \sum_{i=1}^r \left(\frac{(\mathbf{V}^i)^\top \mathbf{c}(\mathbf{x}')}{\lambda_i} \right) \mathbf{V}^i \\ &= K(\mathbf{x}, \mathbf{x}') - \sum_{i=1}^r \frac{\left((\mathbf{V}^i)^\top \mathbf{c}(\mathbf{x}) \right) \left((\mathbf{V}^i)^\top \mathbf{c}(\mathbf{x}') \right)}{\lambda_i}. \end{aligned} \quad (14)$$

3.2 Properties of PI kriging

The PI kriging mean averages the outputs. Before proving this property, let us illustrate it with the simple example of Figure 2: there are redundant points at $\mathbf{x} = 1.5$, $\mathbf{x} = 2$ and $\mathbf{x} = 2.5$. We observe that the kriging mean with PI regularization is equal to the mean of the outputs, $m^{PI}(1.5) = -0.5 = (-1 + 0)/2$, $m^{PI}(2) = 5 = (1.5 + 4 + 7 + 7.5)/4$ and $m^{PI}(2.5) = 5.5 = (5 + 6)/2$. The PI averaging property is due to the more abstract fact that PI projects the observed \mathbf{y} onto the image space of \mathbf{C} .

Property 1 (PI as projection of outputs onto $Im(\mathbf{C})$)

The PI kriging prediction at \mathbf{X} is the projection of the observed outputs onto the image space of the covariance matrix, $Im(\mathbf{C})$.

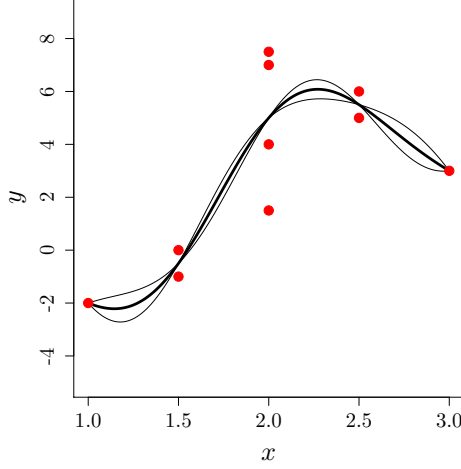


Figure 2: Kriging mean $m^{PI}(x)$ (thick line) and confidence intervals $m^{PI}(x) \pm 2\sqrt{v^{PI}(x)}$ (thin lines). Kriging mean using pseudoinverse goes exactly through the average of the outputs. The observed values are $\mathbf{y} = (-2, -1, 0, 1.5, 4, 7, 7.5, 6, 5, 3)^\top$. $m^{PI}(1.5) = -0.5$, $m^{PI}(2) = 5$, and $m^{PI}(2.5) = 5.5$. Note that v^{PI} is zero at redundant points.

Proof: The PI kriging means at all design points is given by

$$m^{PI}(\mathbf{X}) = \mathbf{C}\mathbf{C}^\dagger \mathbf{y} . \quad (15)$$

Performing the eigendecompositions of the matrices, one gets,

$$\begin{aligned} m^{PI}(\mathbf{X}) &= [\mathbf{V} \ \mathbf{W}] \begin{bmatrix} \text{diag}(\boldsymbol{\lambda}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}^\top \\ \mathbf{W}^\top \end{bmatrix} [\mathbf{V} \ \mathbf{W}] \begin{bmatrix} \text{diag}(\frac{1}{\boldsymbol{\lambda}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{V}^\top \\ \mathbf{W}^\top \end{bmatrix} \mathbf{y} \\ &= \mathbf{V}\mathbf{V}^\top \mathbf{y} \end{aligned} \quad (16)$$

The matrix

$$\mathbf{P}_{Im(\mathbf{C})} = \mathbf{V}\mathbf{V}^\top = (\mathbf{I} - \mathbf{W}\mathbf{W}^\top) \quad (17)$$

is the orthogonal projection onto the image space of \mathbf{C} because it holds that

$$\begin{aligned} \mathbf{P}_{Im(\mathbf{C})} &= \mathbf{P}_{Im(\mathbf{C})}^\top; \\ \mathbf{P}_{Im(\mathbf{C})}^2 &= \mathbf{P}_{Im(\mathbf{C})}; \\ \forall \mathbf{v} \in Im(\mathbf{C}) , \mathbf{P}_{Im(\mathbf{C})} \mathbf{v} &= \mathbf{v}; \\ \text{and } \forall \mathbf{u} \in Nul(\mathbf{C}) , \mathbf{P}_{Im(\mathbf{C})} \mathbf{u} &= \mathbf{0} \quad \square \end{aligned}$$

Redundant points can be further understood thanks to Property 1 and Equation (16): points redundant with \mathbf{x}^i are points \mathbf{x}^j where the observations influences $m^{PI}(\mathbf{x}^i)$. The kriging predictions at the redundant data points $m^{PI}(\mathbf{x}^i)$ and $m^{PI}(\mathbf{x}^j)$ are not \mathbf{y}_i and \mathbf{y}_j , as it happens at non-redundant points where the model is interpolating, but a linear combination of them. The averaging performed by PI becomes more clearly visible in the important case of repeated points.

Property 3 (Null variance of PI regularized models at data points)

The variance of Gaussian processes regularized by pseudoinverse is zero at data points.

Therefore $v^{PI}(\cdot)$ is null at redundant points.

Proof: From Equation (3), the PI kriging variances at all design points are

$$v^{PI}(\mathbf{X}) = c^{PI}(\mathbf{X}, \mathbf{X}) = K(\mathbf{X}, \mathbf{X}) - \mathbf{c}(\mathbf{X})^\top \mathbf{C}^\dagger \mathbf{c}(\mathbf{X}) = \mathbf{C} - \mathbf{C}^\top \mathbf{C}^\dagger \mathbf{C} = \mathbf{C} - \mathbf{C} = 0,$$

thanks to the pseudoinverse property [27], $\mathbf{C}\mathbf{C}^\dagger\mathbf{C} = \mathbf{C}$. \square

4 Nugget regularization

4.1 Definition and covariance orthogonality property

When regularizing a covariance matrix by nugget, a positive value, δ , is added to the main diagonal. This corresponds to a probabilistic model with an additive white noise of variance δ , $Y(\mathbf{x}) \mid Y(\mathbf{x}^i) + \varepsilon_i = \mathbf{y}_i$, $i = 1, \dots, n$, where the ε_i 's are i.i.d. $\mathcal{N}(0, \delta^2)$. Nugget regularization improves the condition number of the covariance matrix by increasing all the eigenvalues by δ : if λ_i is an eigenvalue of \mathbf{C} , then $\lambda_i + \delta$ is an eigenvalue of $\mathbf{C} + \delta\mathbf{I}$ and the eigenvectors remain the same (the proof is straightforward). The associated condition number is $\kappa(\mathbf{C} + \delta\mathbf{I}) = \frac{\lambda_{max} + \delta}{\lambda_{min} + \delta}$. The nugget parameter causes kriging to smoothen the data and become non-interpolating.

Property 4 (Loss of interpolation in models regularized by nugget)

A conditional Gaussian process regularized by nugget has its mean no longer equal to the output at data points, $m^{Nug}(\mathbf{x}^i) \neq \mathbf{y}^i$, $i = 1, n$.

This property can be understood as follows. A conditional GP with invertible covariance matrix is interpolating because $c(\mathbf{x}^i)^\top \mathbf{C}^{-1} \mathbf{y} = \mathbf{C}^{i\top} \mathbf{C}^{-1} \mathbf{y} = \mathbf{e}_i^\top \mathbf{y} = y_i$. This does not stand when \mathbf{C}^{-1} is replaced by $(\mathbf{C} + \delta\mathbf{I})^{-1}$.

Recall that the term $\mathbf{C}^{-1} \mathbf{y}$ in the kriging mean of Equation (2) is denoted by $\boldsymbol{\beta}$. When nugget regularization is used, $\boldsymbol{\beta}$ is shown as $\boldsymbol{\beta}^{Nug}$ and, thanks to the eigenvalue decomposition of $(\mathbf{C} + \delta\mathbf{I})^{-1}$, it is written

$$\boldsymbol{\beta}^{Nug} = \sum_{i=1}^r \frac{(\mathbf{V}^i)^\top \mathbf{y}}{\lambda_i + \delta} \mathbf{V}^i + \sum_{i=r+1}^n \frac{(\mathbf{W}^i)^\top \mathbf{y}}{\delta} \mathbf{W}^i. \quad (20)$$

The main difference between $\boldsymbol{\beta}^{PI}$ (Equation (12)) and $\boldsymbol{\beta}^{Nug}$ lies in the second part of $\boldsymbol{\beta}^{Nug}$: the part that spans the null space of the covariance matrix. In the following, we show that this term cancels out when multiplied by $\mathbf{c}(\mathbf{x})^\top$, a product that intervenes in kriging.

Property 5 (Orthogonality Property of \mathbf{c} and $\text{Nul}(\mathbf{C})$)

The covariance vector $\mathbf{c}(\mathbf{x})$ is perpendicular to the null space of the covariance matrix \mathbf{C} .

Proof: The kernel $K(.,.)$ is a covariance function [2], hence the matrix

$$\mathbf{C}_x = \begin{bmatrix} K(\mathbf{x}, \mathbf{x}) & \mathbf{c}(\mathbf{x})^\top \\ \mathbf{c}(\mathbf{x}) & \mathbf{C} \end{bmatrix} \quad (21)$$

is positive semidefinite.

Let \mathbf{w} be a vector in the null space of \mathbf{C} . According to the definition of positive semidefinite matrices, we have

$$\begin{pmatrix} 1 \\ \mathbf{w} \end{pmatrix}^\top \mathbf{C}_x \begin{pmatrix} 1 \\ \mathbf{w} \end{pmatrix} = K(\mathbf{x}, \mathbf{x}) + 2 \sum_{i=1}^n K(\mathbf{x}, x_i) w_i + 0 \geq 0. \quad (22)$$

The above equation is valid for any vector $\gamma \mathbf{w}$ as well, in which γ is a real number. This happens only if $\sum_{i=1}^n K(\mathbf{x}, x_i) w_i$ is zero, that is to say, $\mathbf{c}(\mathbf{x})^\top$ is perpendicular to the null space of \mathbf{C} . \square

As a result of the Orthogonality Property of c and $\text{Nul}(\mathbf{C})$, the second term in Equation (20) disappears in the kriging mean with nugget regularization which becomes

$$m^{Nug}(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top \sum_{i=1}^r \frac{(\mathbf{V}^i)^\top \mathbf{y}}{\lambda_i + \delta} \mathbf{V}^i. \quad (23)$$

The Orthogonality Property applies similarly to the kriging covariance (Equation (3)), which yields

$$\begin{aligned} c^{Nug}(\mathbf{x}, \mathbf{x}') &= K(\mathbf{x}, \mathbf{x}') - \mathbf{c}(\mathbf{x})^\top \sum_{i=1}^r \frac{(\mathbf{V}^i)^\top \mathbf{c}(\mathbf{x}')}{\lambda_i + \delta} \mathbf{V}^i \\ &= K(\mathbf{x}, \mathbf{x}') - \sum_{i=1}^r \frac{\left((\mathbf{V}^i)^\top \mathbf{c}(\mathbf{x}) \right) \left((\mathbf{V}^i)^\top \mathbf{c}(\mathbf{x}') \right)}{\lambda_i + \delta}. \end{aligned} \quad (24)$$

Comparing equations (13) and (23) indicates that the behavior of m^{PI} and m^{Nug} will be similar to each other if δ is small. The same holds for kriging covariances (hence variances) c^{PI} and c^{Nug} in equations (14) and (24).

Property 6 (Equivalence of PI and nugget regularizations)

The mean and covariance of conditional GPs regularized by nugget tend toward the ones of GPs regularized by pseudoinverse as the nugget value δ tends to 0.

In addition, equations (14) and (24) show that c^{Nug} is always greater than c^{PI} . These results will be illustrated later in the Discussion Section.

4.2 Nugget and maximum likelihood

It is common to estimate the nugget parameter by maximum likelihood (ML, cf. Appendix B, Equation (34)). As will be detailed below, the amplitude of the nugget estimated by ML is increasing with the spread of observations

at redundant points. It matches the interpretation of nugget as the amount of noise put on data: an increasing discrepancy between responses at a given point is associated to more observations noise.

In Figure 3 two vectors of response values are shown, \mathbf{y} (bullets) and \mathbf{y}^+ (crosses), located at k different \mathbf{x} sites. The spread of response values \mathbf{y}^+ is larger than that of \mathbf{y} at some redundant points. Let s_i^2 and s_i^{+2} , $i = 1, \dots, k$, denote the variances of \mathbf{y} and \mathbf{y}^+ at the redundant points,

$$s_i^2 = \frac{\sum_{j=N_1+\dots+N_{i-1}+1}^{N_1+\dots+N_i} (y_j - \bar{y}_i)^2}{N_i - 1}, \quad (25)$$

and the same stands with \mathbf{y}^+ and its variance s_i^{+2} . The nugget that maximizes the likelihood, the other GP parameters being fixed (the length-scales θ_i and the process variance σ^2), is increasing when the variance of the outputs increases.

Theorem 1

Suppose that there are observations located at k different sites. If we are given two vectors of response values \mathbf{y} and \mathbf{y}^+ such that

1. $s_i^{+2} \geq s_i^2$ for all $i = 1, \dots, k$ and
2. $\bar{y}_i = \bar{y}_i^+$ for all $i = 1, \dots, k$,

then the nugget amplitudes $\hat{\delta}$ and $\hat{\delta}^+$ that maximize the likelihood with other GP parameters being fixed are such that $\hat{\delta}^+ \geq \hat{\delta}$.

The Theorem is proved in Appendix B.

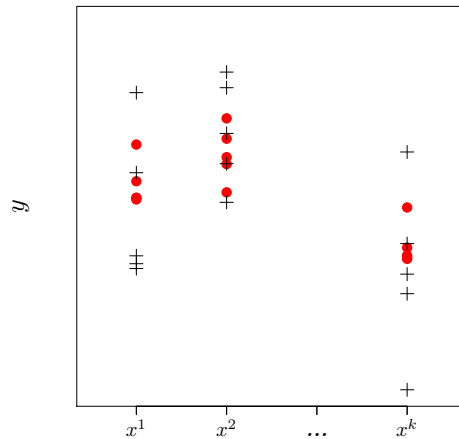


Figure 3: The response values \mathbf{y} and \mathbf{y}^+ are denoted by bullets and crosses, respectively. At each location, the mean of \mathbf{y} and \mathbf{y}^+ are identical, $\bar{y}_i = \bar{y}_i^+$, but the spread of observations in \mathbf{y}^+ is never less than that of \mathbf{y} at redundant points.

5 Discussion: choice and tuning of the classical regularization methods

This section carries out a practical comparison of PI and nugget regularization methods, which are readily available in most GP softwares. We start with a discussion of how data and model match, which further allows to decide whether nugget or PI should be used. Finally, we provide guidelines to tune the regularization parameters.

Note that nugget regularization should be used when the observed data is known to be noisy since it has a physical meaning [22]. The loss of the interpolating property at data points associated to nugget regularization is here a beneficial filtering effect. This discussion on non-deterministic outputs is out of the scope of the current article.

5.1 Model-data discrepancy

Model-data discrepancy can be measured as the distance between the observations \mathbf{y} and the GP model regularized by pseudoinverse.

Definition 2 (Model-data discrepancy) *Let \mathbf{X} be a set of design points with associated observations \mathbf{y} . Let \mathbf{V} and \mathbf{W} be the normalized eigenvectors spanning the image space and the null space of the covariance matrix \mathbf{C} , respectively. The model-data discrepancy is defined as*

$$discr = \frac{\|\mathbf{y} - m^{PI}(\mathbf{X})\|^2}{\|\mathbf{y}\|^2} = \frac{\|\mathbf{W}\mathbf{W}^\top\mathbf{y}\|^2}{\|\mathbf{y}\|^2} \quad (26)$$

where $m^{PI}(\cdot)$ is the pseudoinverse regularized GP model of Equation (15).

The last equality in the definition of $discr$ comes from Equations (16) and (17). The discrepancy is a normalized scalar, $0 \leq discr \leq 1$, where $discr = 0$ indicates that the model and the data are perfectly compatible, and vice versa when $discr = 1$. The definition of redundant points does not depend on the observations \mathbf{y} and the model-data discrepancy is a scalar globalizing the contributions of all observations. An intermediate object between redundant points and discrepancy is the gradient of the squared model-data error with respect to the observations,

$$\nabla_{\mathbf{y}}\|\mathbf{y} - m^{PI}(\mathbf{X})\|^2 = \mathbf{W}\mathbf{W}^\top\mathbf{y} . \quad (27)$$

It appears that the gradient of the error, $\|\mathbf{y} - m^{PI}(\mathbf{X})\|^2$, is the model-data distance, which comes from the quadratic form of the error. The magnitude of the components of the vector $\mathbf{W}\mathbf{W}^\top\mathbf{y}$ measures the sensitivity of the error to a particular observation. At repeated points, a gradient-based approach would advocate to make the observations closer to their mean proportionally to their distance to the mean.

$-\mathbf{W}\mathbf{W}^\top\mathbf{y}$ is a direction of reduction of the model-data distance in the space of observations. Because the distance considered is quadratic, this direction is colinear to the error, $(\mathbf{y} - m^{PI}(\mathbf{X}))$. The indices of the non-zero components of $\mathbf{W}\mathbf{W}^\top\mathbf{y}$ also designate the redundant points.

5.2 Two detailed examples

A common practice when the nugget value, δ , is not known beforehand is to estimate it by ML or cross-validation. In Appendix B, we show that the ML estimated nugget value, $\hat{\delta}$, is increasing with the spread of responses at redundant points. This is one situation (among others, e.g., the additive example hereafter) where the data and the model mismatch, and $\hat{\delta}$ is large. Figure 4 is an example where $\hat{\delta}$ is equal to 7.06. Some authors such as in [28, 3] recommend using cross-validation instead of ML for learning the kriging parameters. In the example of Figure 4, the estimated nugget value by leave-one-out cross-validation, denoted by $\hat{\delta}_{CV}$, is 1.75. The dash-dotted lines represent the kriging model regularized by nugget that is estimated by cross-validation. The model-data discrepancy is $discr = 0.36$ and $\mathbf{W}\mathbf{W}^\top \mathbf{y} = (0, 0, -3, 3, 0, 0)^\top$ which shows that points 3 and 4 are redundant and their outputs should be made closer to reduce the model-data error.

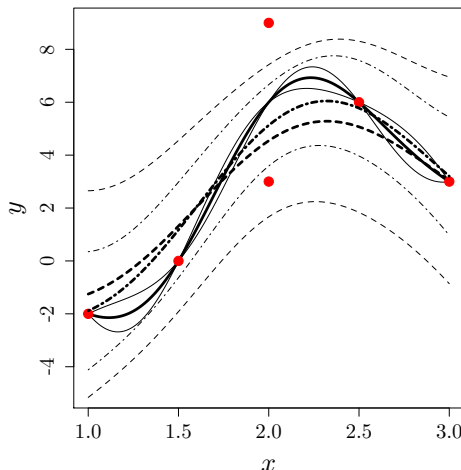


Figure 4: Comparison of kriging regularized by PI (solid lines), nugget estimated by ML (dashed lines) and nugget estimated by cross-validation (dash-dotted lines). $\mathbf{X} = [1; 1.5; 2; 2.00001; 2.5; 3]$ and $\mathbf{y} = (-2, 0, 3, 9, 6, 3)^\top$. The estimated nugget values are $\hat{\delta} = 7.06$ and $\hat{\delta}_{CV} = 1.75$.

We now give a two-dimensional example of a kriging model with additive kernel defined over $\mathbf{X} = [(1, 1), (2, 1), (1, 2), (2, 2), (1.5, 1.5), (1.25, 1.75), (1.75, 1.25)]$, cf. Figure 5. As explained in Section 2.2, the first four points of the DoE make the additive covariance matrix non-invertible eventhough the points are not near each other in Euclidean distance. Suppose that the design points have the response values $\mathbf{y} = (1, 4, -2, 1, 1, -0.5, 2.5)^\top$ which correspond to the additive true function $f(\mathbf{x}) = x_1^2 - x_2^2 + 1$. The covariance matrix is the sum of two parts

$$\mathbf{C}_{add} = \sigma_1^2 K_1 + \sigma_2^2 K_2 ,$$

where σ_i^2 are the process variances and $\sigma_i^2 K_i$ the kernel in dimension $i = 1, 2$.

To estimate the parameters of \mathbf{C}_{add} , the negative of the likelihood is minimized (see Equation (34)) which yields a nugget value $\hat{\delta} \approx 10^{-12}$ (the lower

bound on nugget used). A small nugget value is obtained because the associated output value follows an additive law compatible with the kernel: there is no discrepancy between the model and the data. Because of the small nugget value, the models regularized by PI and nugget are very close to each other (the left picture in Figure 5).

Let us now introduce model-data discrepancy in this example: the observations of the first four data points no longer follow an additive function after changing the third response from -2 to 2; additive kriging models cannot interpolate their outputs. The nugget value estimated by ML is equal to 1.91, so $m^{Nug}(\mathbf{x})$ does not interpolate any of the data points (\mathbf{x}^1 to \mathbf{x}^7). Regarding $m^{PI}(\mathbf{x})$, the projection onto $Im(\mathbf{C})$ make the GP predictions different from the observations at \mathbf{x}^1 , \mathbf{x}^2 , \mathbf{x}^3 and \mathbf{x}^4 . For example, $m^{PI}(\mathbf{x}^4) = 2$. The projection applied to points \mathbf{x}^5 to \mathbf{x}^7 where no linear dependency exists show that $m^{PI}(\mathbf{x})$ is interpolating there, which is observed on the right picture of Figure 5.

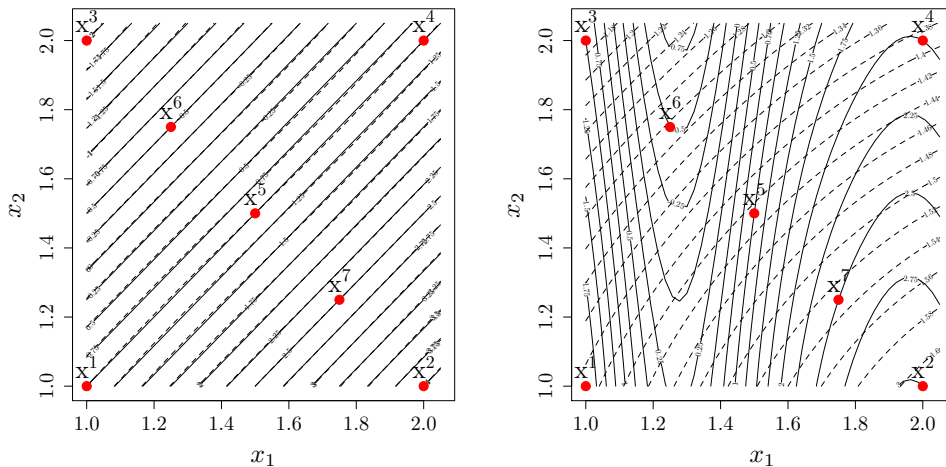


Figure 5: Contour plots of kriging mean regularized by pseudoinverse (solid line) vs. nugget (dashed line) for an additive GP. The bullets are data points. Left: the response values are additive, $\mathbf{y} = (1, 4, -2, 1, 1, -0.5, 2.5)^\top$ and $\hat{\delta} = 10^{-12}$. Right: the third observation is replaced by 2, creating non-additive observations and $\hat{\delta} \approx 1.91$; $m^{Nug}(\mathbf{x})$ is no longer interpolating, $m^{PI}(\mathbf{x})$ still interpolates \mathbf{x}^5 to \mathbf{x}^7 .

Our observations reflect that large estimated values of nugget (whether by ML or cross-validation) indicate model-data discrepancy. This agrees with the calculated discrepancies: in the last additive kernel example when all the outputs were additive, $discr = 0$ and $\mathbf{W}\mathbf{W}^\top \mathbf{y} = (0, 0, 0, 0, 0, 0, 0)^\top$ (no redundant point); when the value of the third output was increased to 2, $discr = 0.37$ and $\mathbf{W}\mathbf{W}^\top \mathbf{y} = (-1, 1, 1, -1, 0, 0, 0)^\top$ showing that points 1 to 4 are redundant and that, to reduce model error, points 1 and 4 should increase their outputs while points 2 and 3 should decrease theirs.

Of course, for the sole purpose of quantifying model-data discrepancy it is more efficient to use Formula (26) which involves one pseudo-inverse calculation and two matrix products against a nonlinear likelihood maximization with

repeated embedded \mathbf{C} eigenvalues analyses for the nugget estimation.

5.3 PI or nugget ?

On the one hand, models regularized by PI have predictions, $m^{PI}(\cdot)$, that interpolate uniquely defined points and go through the average output at redundant points (Property 2). The associated kriging variances, $v^{PI}(\cdot)$, are null at redundant points (Property 3). On the other hand, models regularized by nugget have predictions which are neither interpolating nor averaging (Property 4) while their variances are non-zero at data points. Note that kriging variance tends to σ^2 as the nugget value increases (see Equation (24)). These facts can be observed in Figure 6. Additionally, this Figure illustrates that nugget regularization tends to PI regularization as the nugget value decreases (Property 6). If there is a good agreement between the data and the GP model, the PI reg-

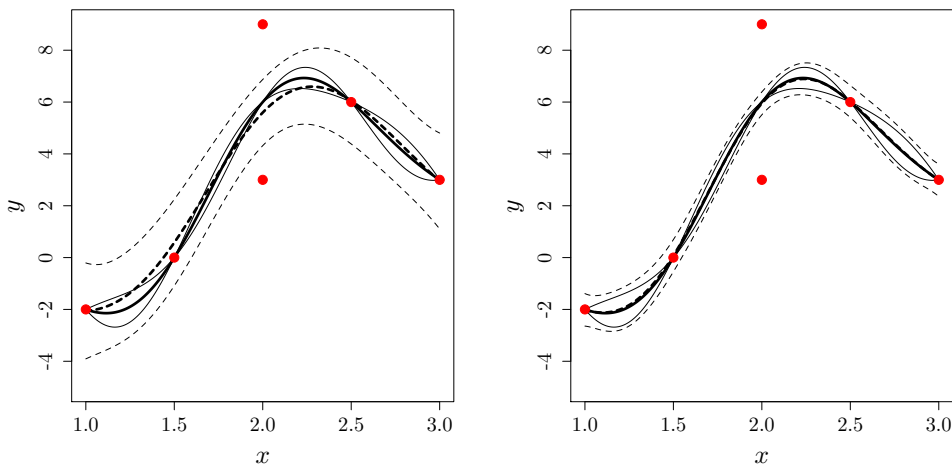


Figure 6: One dimensional kriging regularized by PI (solid lines) and nugget (dashed lines). The nugget amplitude is 1 on the left and 0.1 on the right. The cut-off eigenvalue for the pseudoinverse is $\tau = 10^{-3}$. $m^{Nug}(x)$ is not interpolating which is best seen at the second point on the left. On the right, the PI and nugget models are closer to each other. Same \mathbf{X} and \mathbf{y} as Figure 4.

ularization or equivalently, a small nugget, should be used. This can also be understood through the Definition of model-data discrepancy and Property 1: when $discr = 0$, the observations are perpendicular to $Nul(\mathbf{C})$ and, equivalently, $m^{PI}(\mathbf{X}) = \mathbf{y}$ since $m^{PI}(\cdot)$ performs a projection onto $Im(\mathbf{C})$. Vice versa, if the model-data discrepancy measure is large, choosing PI or nugget regularization is a matter of choice: either the prediction averaging property is regarded as most important and PI should be used, or a non-zero variance at redundant points is favored and nugget should be selected; If the discrepancy is concentrated on few redundant points, nugget regularized models will distribute the uncertainty (additional model variance) throughout the x domain while PI regularized models will ignore it.

5.4 Tuning regularization parameters

How small can a nugget value be? Adding nugget to the main diagonal of a covariance matrix increments all the eigenvalues by the nugget amplitude. The condition number of the covariance matrix with nugget is $\kappa(\mathbf{C} + \delta\mathbf{I}) = \frac{\lambda_{max} + \delta}{\lambda_{min} + \delta}$. Accordingly, a “small” nugget is the smallest value of δ such that $\kappa(\mathbf{C} + \delta\mathbf{I})$ is less than a reasonable condition number after regularization, κ_{max} (say, $\kappa_{max} = 10^8$). With such targeted condition number, the smallest nugget would be $\delta = \frac{\lambda_{max} - \kappa_{max}\lambda_{min}}{\kappa_{max} - 1}$ if $\lambda_{max} - \kappa_{max}\lambda_{min} \geq 0$, $\delta = 0$ otherwise.

Computing a pseudoinverse also involves a parameter, the positive threshold τ below which an eigenvalue is considered as null. The eigenvectors associated to eigenvalues smaller than τ are numerically regarded as null space basis vectors (even though they may not, strictly speaking, be part of the null space). A suitable threshold should filter out eigenvectors associated to points that are almost redundant. The heuristic we propose is to tune τ so that λ_1/τ , which is an upper bound of the PI condition number¹, is equal to κ_{max} , i.e., $\tau = \lambda_1/\kappa_{max}$.

In the example shown in Figure 7, the covariance matrix is not numerically invertible because the points 3 and 4 are near $x = 2$. The covariance matrix has six eigenvalues, $\lambda_1 = 34.89 \geq \dots \geq \lambda_5 = 0.86 \geq \lambda_6 = 8.42 \times 10^{-11} \approx 0$ and the eigenvector related to the smallest eigenvalue is $\mathbf{W}^1 = (\mathbf{e}^4 - \mathbf{e}^3)/\sqrt{2}$. In Figure 7, we have selected $\tau = 10^{-3}$, hence $\kappa_{PI}(\mathbf{C}) = 40.56$. Any value of τ in the interval $\lambda_6 < \tau < \lambda_5$ would have yielded the same result. But if the selected tolerance were e.g., $\tau = 1$, which is larger than λ_5 , the obtained PI kriging model no longer interpolates data points.

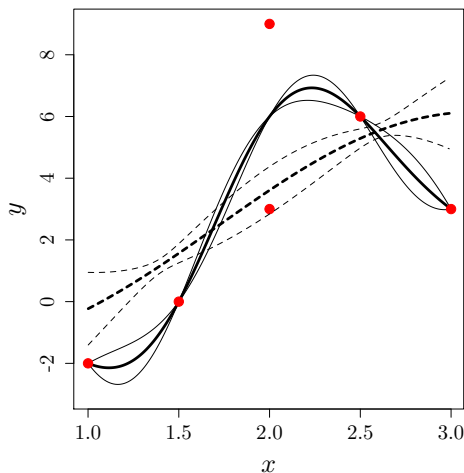


Figure 7: Effect of the tolerance τ on the kriging model regularized by PI. Dashed line, $\tau = 1$; continuous line, $\tau = 10^{-3}$. Except for τ , the setting is the same as that of Figure 6. When the tolerance is large ($\tau = 1$), the 5th eigenvector is deleted from the effective image space of \mathbf{C} in addition to the 6th eigenvector, and the PI regularized model is no longer interpolating. Same \mathbf{X} and \mathbf{y} as Figure 4.

¹By PI condition number we mean $\kappa_{PI}(\mathbf{C}) = \|\mathbf{C}\| \|\mathbf{C}^\dagger\| = \lambda_1/\lambda_r \leq \lambda_1/\tau$

6 Interpolating Gaussian distributions

6.1 Interpolation and repeated points

We have seen that redundant points reveal specific properties of regularization schemes. In our context of deterministic experiments, we are interested in interpolating data. The notion of interpolation should be clarified in the case of repeated points with different outputs (e.g., Figure 3) as a function cannot interpolate them. Here, we seek GPs that have the following interpolation properties.

Definition 3 (Interpolation properties at repeated points) *A GP exhibits interpolation properties when*

- *its trajectories pass through uniquely defined data points (therefore the GP has a null variance there),*
- *and at repeated points the GP's mean and variance are the empirical average and variance of the outputs, respectively.*

The following GP model has the above interpolation properties for deterministic outputs, even in the presence of repeated points. In this sense, it can be seen as a new regularization technique, although its potential use goes beyond regularization.

6.2 A GP model with interpolation properties

A *distribution-wise* model is obtained by interpolating Gaussian probability distributions instead of data points. Following the same notations as in Section 3.2, the model is built from observations at k different \mathbf{x} sites, which are considered as samples of Gaussian variables $Z(\mathbf{x}^1), \dots, Z(\mathbf{x}^k)$. Together, the k sets of observations make the random vector \mathbf{Z} . Redundant points are grouped by sites, e.g., y_1, \dots, y_{N_1} are the observations at \mathbf{x}^1 , etc. Recall that the output empirical mean and variance at \mathbf{x}^i are \bar{y}_i and \bar{s}_i^2 , that we gather in the vector $\bar{\mathbf{y}}$ and the $k \times k$ diagonal matrix Γ made of the \bar{s}_i^2 's. The conditional mean and variance of the distribution-wise GP come from the laws of total expectation and variance applied to the random variables Z and the GP outcomes $\omega \in \Omega$:

$$m^{Dist}(\mathbf{x}) = \mathbb{E}_Z(\mathbb{E}_\Omega(Y(\mathbf{x})|\mathbf{Z} = \mathbf{z})) = \mathbb{E}_Z(\mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1} \mathbf{z}) = \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1} \mathbb{E}_Z(\mathbf{z})$$

where the index Z is used to distinguish between the point-wise and the distribution-wise covariances. For example, \mathbf{C} is $n \times n$ and not necessarily invertible while \mathbf{C}_Z is $k \times k$ and invertible. Now, $\mathbb{E}_Z(\mathbf{z})$ is approximated by the empirical observations means which yields,

$$m^{Dist}(\mathbf{x}) \equiv \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1} \bar{\mathbf{y}}. \quad (28)$$

Equation (28) indicates that the model passes through the empirical mean of the outputs. The variance is calculated in a similar way,

$$\begin{aligned} v^{Dist}(\mathbf{x}) &= \mathbb{E}_Z(\text{Var}_\Omega(Y(\mathbf{x})|\mathbf{Z} = \mathbf{z})) + \text{Var}_Z(\mathbb{E}_\Omega(Y(\mathbf{x})|\mathbf{Z} = \mathbf{z})) \\ &= \mathbf{c}_Z(\mathbf{x}, \mathbf{x}) - \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1} \mathbf{c}_Z(\mathbf{x}) + \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1} (\text{Var}_Z \mathbf{z}) \mathbf{C}_Z^{-1} \mathbf{c}_Z(\mathbf{x}). \end{aligned}$$

Finally, the variance of the outputs is replaced by its empirical estimate

$$v^{Dist}(\mathbf{x}) \equiv \mathbf{c}_Z(\mathbf{x}, \mathbf{x}) - \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1} \mathbf{c}_Z(\mathbf{x}) + \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1} \Gamma \mathbf{C}_Z^{-1} \mathbf{c}_Z(\mathbf{x}). \quad (29)$$

$v^{Dist}(\mathbf{x})$ is no longer zero at redundant points, as illustrated in Figure 8: The conditional variance at \mathbf{x}^i is equal to the corresponding empirical variance s_i^2 because $\mathbf{C}_Z^{-1} \mathbf{c}_Z(\mathbf{x}^i) = \mathbf{e}_i$. Note the fundamental difference between the behaviours of a distribution-wise GP and a GP regularized by nugget: as the number of observations N_i at a redundant point \mathbf{x}^i increases, the variance $v^{Nug}(\mathbf{x}^i)$ tends to 0 while $v^{Dist}(\mathbf{x}^i)$ remains equal to s_i^2 .

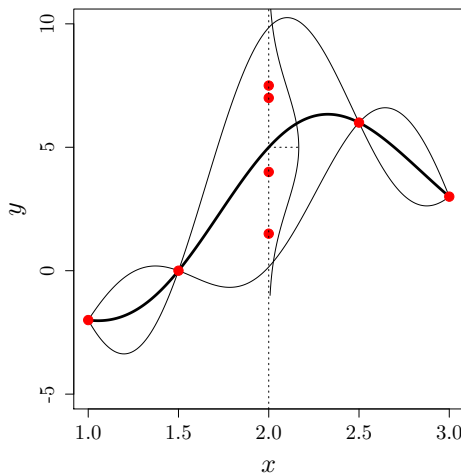


Figure 8: Distribution-wise GP, $m^{Dist}(x)$ (thick line) $\pm 2\sqrt{v^{Dist}(x)}$ (thin lines). At the redundant point $x = 2$, the outputs are 1.5, 4, 7 and 7.5. The mean of the distribution-wise GP passes through the average of outputs. Contrarily to PI (cf. Figure 2), distribution-wise GP preserves the empirical variance: the kriging variance at $x = 2$ is equal to $s_{x=2}^2 = 5.87$.

7 Conclusions

This paper provides a new algebraic comparison of pseudoinverse and nugget regularizations, two classical solutions to overcome the degeneracy of the covariance matrix in Gaussian processes (GPs). We propose a practical strategy when confronted with bad conditioning in GP regression. The analysis focuses on the interpolation properties of GPs when outputs are deterministic. Clear differences between pseudoinverse and nugget regularizations arise by looking at redundant points as a limit case of covariance matrix degeneracy. We have proved that, contrarily to GPs with nugget, GPs with pseudoinverse average the values of outputs and have null variance at redundant points. In GPs regularized by nugget, the discrepancy between model and data translates into a departure of the GP from observation points throughout the domain. In GPs regularized by pseudoinverse, this departure only occurs at redundant points, but the variance is null there.

We have proposed a distribution-wise GP model that interpolates normal distributions instead of data points. This model does not have the drawbacks from both nugget and pseudoinverse regularizations: it not only averages the outputs at redundant points but also preserves the redundant points variances.

Distribution-wise GPs shed a new light on regularization, which starts with the creation of redundant points by clustering. A potential benefit is the reduction in covariance matrix size. Further studying distribution-wise GPs is the main continuation of this work.

Acknowledgments

The authors would like to acknowledge support by the French national research agency (ANR) within the Modèles Numérique project “NumBBO- Analysis, Improvement and Evaluation of Numerical Blackbox Optimizers”.

A Examples of redundant points

This Appendix gives easily interpretable examples of DoEs with associated kernels that make the covariance matrix non-invertible. The eigenvalues, eigenvectors and orthogonal projection matrix onto the image space (cf. also Section 2.3) are described.

A.1 Repeated points

Repeated design points are the simplest example of redundancy in a DoE since columns of the covariance matrix \mathbf{c} are duplicated. An example is given in Figure 9 with a two-dimensional design, and a classical squared exponential kernel. The eigenvalues and eigenvectors of the covariance matrix associated to

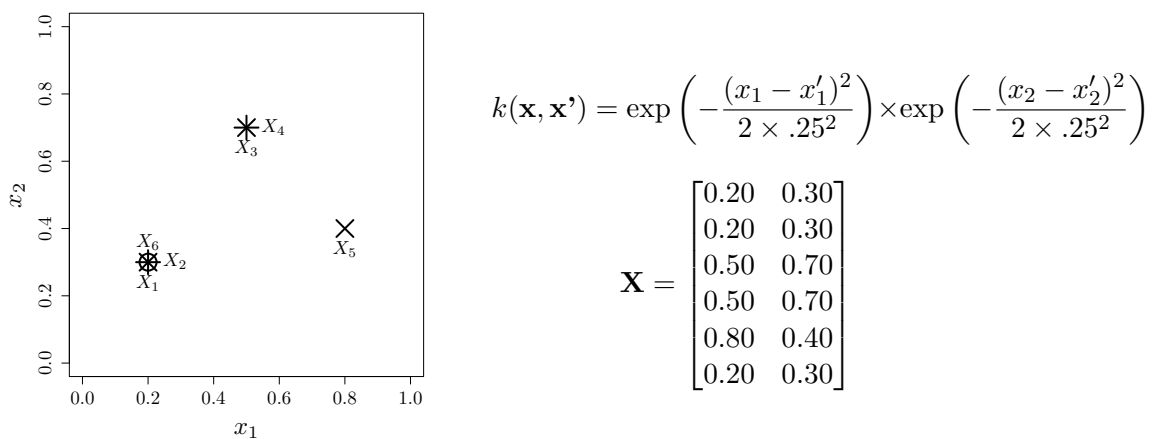


Figure 9: Kernel and DoE of the repeated points example

Figure 9 are

$$\boldsymbol{\lambda} = \begin{bmatrix} 3.12 \\ 1.99 \\ 0.90 \\ 0.00 \\ 0.00 \\ 0.00 \end{bmatrix}, \mathbf{V} = \begin{bmatrix} -0.55 & 0.19 & 0.00 \\ -0.55 & 0.19 & 0.00 \\ -0.22 & -0.64 & -0.21 \\ -0.22 & -0.64 & -0.21 \\ -0.09 & -0.28 & 0.96 \\ -0.55 & 0.19 & 0.00 \end{bmatrix} \text{ and } \mathbf{W} = \begin{bmatrix} 0.00 & -0.30 & 0.76 \\ -0.71 & 0.12 & -0.39 \\ -0.04 & 0.66 & 0.26 \\ 0.04 & -0.66 & -0.26 \\ 0.00 & 0.00 & 0.00 \\ 0.71 & 0.18 & -0.37 \end{bmatrix},$$

with the orthogonal projection matrix onto $Im(\mathbf{C})$

$$\mathbf{V}\mathbf{V}^\top = \begin{bmatrix} 0.33 & 0.33 & 0.00 & 0.00 & 0.00 & 0.33 \\ 0.33 & 0.33 & 0.00 & 0.00 & 0.00 & 0.33 \\ 0.00 & 0.00 & 0.50 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.50 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.33 & 0.33 & 0.00 & 0.00 & 0.00 & 0.33 \end{bmatrix}$$

Points $\{1, 2, 6\}$ and $\{3, 4\}$ are repeated and redundant.

A.2 First additive example

The first example of GP with additive kernel is described in Figure 10. As explained in Section 2.2, the rectangular patterns of points $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8\}$ create linear dependencies between the columns of \mathbf{C} . The eigen-

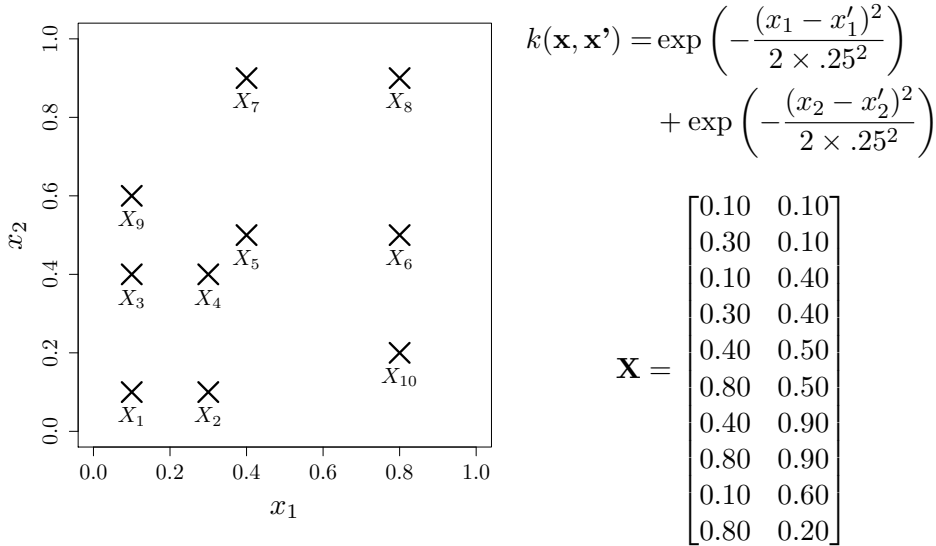


Figure 10: Kernel and DoE of the first additive GP example

values and eigenvectors of the covariance matrix are,

$$\boldsymbol{\lambda} = \begin{bmatrix} 9.52 \\ 3.58 \\ 2.60 \\ 2.31 \\ 1.46 \\ 0.39 \\ 0.09 \\ 0.06 \\ 0.00 \\ 0.00 \end{bmatrix}, \mathbf{V} = \begin{bmatrix} -0.30 & -0.32 & 0.45 & -0.15 & 0.34 & -0.10 & 0.22 & 0.40 \\ -0.33 & -0.24 & 0.29 & -0.43 & -0.22 & -0.30 & -0.43 & 0.04 \\ -0.38 & -0.22 & -0.01 & 0.31 & 0.22 & 0.59 & 0.17 & 0.17 \\ -0.41 & -0.14 & -0.17 & 0.04 & -0.34 & 0.40 & -0.47 & -0.19 \\ -0.38 & 0.01 & -0.37 & 0.03 & -0.40 & -0.29 & 0.43 & 0.18 \\ -0.28 & 0.45 & 0.03 & 0.44 & -0.13 & -0.27 & -0.15 & 0.40 \\ -0.25 & 0.19 & -0.38 & -0.62 & 0.11 & 0.13 & 0.30 & -0.07 \\ -0.15 & 0.64 & 0.02 & -0.22 & 0.38 & 0.15 & -0.29 & 0.15 \\ -0.34 & -0.13 & -0.24 & 0.26 & 0.54 & -0.43 & -0.10 & -0.51 \\ -0.25 & 0.34 & 0.59 & 0.05 & -0.22 & 0.08 & 0.35 & -0.54 \end{bmatrix}$$

$$\text{and } \mathbf{W} = \begin{bmatrix} 0.00 & 0.50 \\ 0.00 & -0.50 \\ 0.00 & -0.50 \\ 0.00 & 0.50 \\ 0.50 & 0.00 \\ -0.50 & 0.00 \\ -0.50 & 0.00 \\ 0.50 & 0.00 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \end{bmatrix}.$$

The projection matrix onto the image space is

$$\mathbf{V}\mathbf{V}^\top = \begin{bmatrix} 0.75 & 0.25 & 0.25 & -0.25 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.25 & 0.75 & -0.25 & 0.25 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.25 & -0.25 & 0.75 & 0.25 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ -0.25 & 0.25 & 0.25 & 0.75 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.75 & 0.25 & 0.25 & -0.25 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.25 & 0.75 & -0.25 & 0.25 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.25 & -0.25 & 0.75 & 0.25 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & -0.25 & 0.25 & 0.25 & 0.75 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}.$$

The redundancy between points 1 to 4 on the one hand, and 5 to 8 on the other hand, is readily seen on the matrix.

A.3 Second additive example

This example shows how an incomplete rectangular pattern with additive kernels can also make covariance matrices singular. In Figure 11, the point at coordinates (0.3, 0.4), which is not in the design, has a GP response defined twice, once by the points {1, 2, 3} and once by the points {4, 5, 6}. This redun-

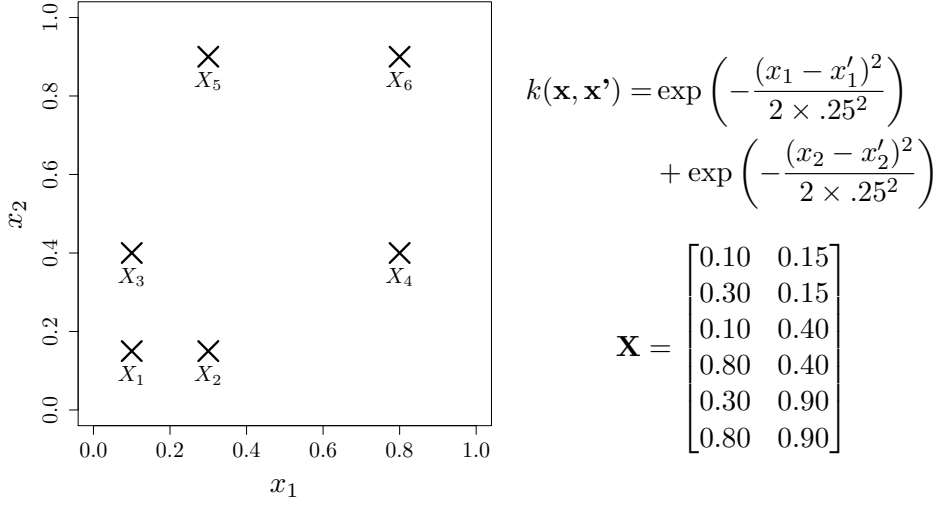


Figure 11: Kernel and DoE of the second additive GP example

dancy in the DoE explains why \mathbf{C} has one null eigenvalue:

$$\boldsymbol{\lambda} = \begin{bmatrix} 5.75 \\ 2.90 \\ 2.07 \\ 0.80 \\ 0.49 \\ 0.00 \end{bmatrix}, \mathbf{V} = \begin{bmatrix} -0.50 & 0.34 & -0.01 & 0.18 & 0.66 \\ -0.49 & 0.25 & 0.20 & 0.57 & -0.40 \\ -0.48 & 0.17 & -0.29 & -0.69 & -0.01 \\ -0.32 & -0.39 & -0.65 & 0.17 & -0.35 \\ -0.36 & -0.28 & 0.66 & -0.33 & -0.28 \\ -0.20 & -0.75 & 0.09 & 0.15 & 0.45 \end{bmatrix}, \mathbf{W} = \begin{bmatrix} -0.41 \\ 0.41 \\ 0.41 \\ -0.41 \\ -0.41 \\ 0.41 \end{bmatrix}.$$

The orthogonal projection matrix onto the image space of \mathbf{C} tells us that all the points in the design are redundant,

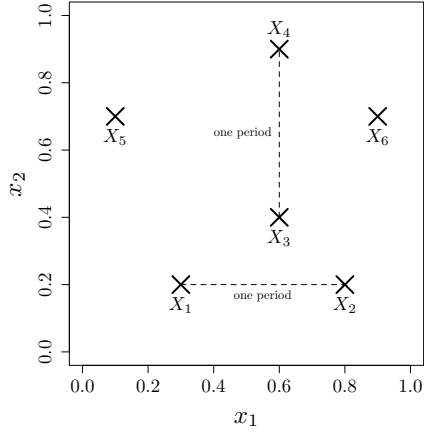
$$\mathbf{V}\mathbf{V}^\top = \begin{bmatrix} 0.83 & 0.17 & 0.17 & -0.17 & -0.17 & 0.17 \\ 0.17 & 0.83 & -0.17 & 0.17 & 0.17 & -0.17 \\ 0.17 & -0.17 & 0.83 & 0.17 & 0.17 & -0.17 \\ -0.17 & 0.17 & 0.17 & 0.83 & -0.17 & 0.17 \\ -0.17 & 0.17 & 0.17 & -0.17 & 0.83 & 0.17 \\ 0.17 & -0.17 & -0.17 & 0.17 & 0.17 & 0.83 \end{bmatrix}.$$

A.4 Periodic example

The kernel and DoE of the periodic example are given in Figure 12.

The eigenvalues and eigenvectors of the associated covariance matrix \mathbf{C} are,

$$\boldsymbol{\lambda} = \begin{pmatrix} 2.00 \\ 2.00 \\ 1.01 \\ 0.99 \\ 0.00 \\ 0.00 \end{pmatrix}, \mathbf{V} = \begin{bmatrix} -0.50 & 0.50 & 0.01 & -0.01 \\ -0.50 & 0.50 & 0.01 & -0.01 \\ -0.50 & -0.50 & 0.01 & -0.01 \\ -0.50 & -0.50 & 0.01 & -0.01 \\ -0.03 & 0.00 & -0.70 & 0.72 \\ 0.00 & 0.00 & -0.72 & -0.70 \end{bmatrix} \text{ and } \mathbf{W} = \begin{bmatrix} 0.00 & 0.71 \\ 0.00 & -0.71 \\ 0.71 & 0.00 \\ -0.71 & 0.00 \\ 0.00 & 0.00 \\ 0.00 & 0.00 \end{bmatrix}.$$



$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\sin(4\pi(x_1 - x'_1))^2}{2 \times .25^2}\right) \times \exp\left(-\frac{\sin(4\pi(x_2 - x'_2))^2}{2 \times .25^2}\right)$$

$$\mathbf{X} = \begin{bmatrix} 0.30 & 0.20 \\ 0.80 & 0.20 \\ 0.60 & 0.40 \\ 0.60 & 0.90 \\ 0.10 & 0.70 \\ 0.90 & 0.70 \end{bmatrix}$$

Figure 12: Kernel and DoE of the periodic example

There are two null eigenvalues. The projector onto the image space is

$$\mathbf{V}\mathbf{V}^\top = \begin{bmatrix} 0.50 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.50 & 0.50 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.50 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.50 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix}$$

which shows that points 1 and 2, on the one hand, and points 3 and 4, on the other hand, are redundant.

A.5 Dot product kernel example

The non-stationary dot product or linear kernel is $k(\mathbf{x}, \mathbf{x}') = 1 + \mathbf{x}^\top \mathbf{x}'$.

We consider a set of three one dimensional, non-overlapping, observation points:

$$\mathbf{X} = \begin{bmatrix} 0.20 \\ 0.60 \\ 0.80 \end{bmatrix}. \quad \text{The associated eigenvalues and eigenvectors are,}$$

$$\boldsymbol{\lambda} = \begin{bmatrix} 3.90 \\ 0.14 \\ 0.00 \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} -0.49 & 0.83 \\ -0.59 & -0.09 \\ -0.64 & -0.55 \end{bmatrix} \quad \text{and} \quad \mathbf{W} = \begin{bmatrix} 0.27 \\ -0.80 \\ 0.53 \end{bmatrix}$$

The projection matrix onto the image space of \mathbf{C} is

$$\mathbf{V}\mathbf{V}^\top = \begin{bmatrix} 0.93 & 0.21 & -0.14 \\ 0.21 & 0.36 & 0.43 \\ -0.14 & 0.43 & 0.71 \end{bmatrix}$$

Because there are 3 data points which is larger than $d + 1 = 2$, all points are redundant. With less than 3 data points, the null space of \mathbf{C} is empty.

B Proof of Theorem 1

Before starting the proof, we need equations resulting from the positive definiteness of the covariance matrix \mathbf{C} :

$$\mathbf{y} = \mathbf{P}_{Nul(\mathbf{C})}\mathbf{y} + \mathbf{P}_{Im(\mathbf{C})}\mathbf{y} \quad (30)$$

$$\mathbf{P}_{Im(\mathbf{C})}\mathbf{y} = \sum_{i=1}^{n-N+k} \langle \mathbf{y}, \mathbf{U}^i \rangle \mathbf{U}^i \quad (31)$$

$$\mathbf{P}_{Nul(\mathbf{C})}\mathbf{y} = \sum_{i=n-N+k+1}^n \langle \mathbf{y}, \mathbf{U}^i \rangle \mathbf{U}^i \quad (32)$$

$$\|\mathbf{P}_{Nul(\mathbf{C})}\mathbf{y}\|^2 = \|\mathbf{y} - \mathbf{P}_{Im(\mathbf{C})}\mathbf{y}\|^2, \quad (33)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

The ML formula given in Equation (5), after removing fixed terms and incorporating nugget effect, is:

$$-2 \ln L(\mathbf{y}|\delta) \approx \ln(|\mathbf{C} + \delta\mathbf{I}|) + \mathbf{y}^\top (\mathbf{C} + \delta\mathbf{I})^{-1} \mathbf{y}. \quad (34)$$

The eigenvalue decomposition of matrix $\mathbf{C} + \delta\mathbf{I}$ in (34) consists of

$$\mathbf{U} = (\mathbf{U}^1, \dots, \mathbf{U}^{N-k}, \mathbf{U}^{N-k+1}, \dots, \mathbf{U}^n) \quad (35)$$

$$\mathbf{\Sigma} = \text{diag}(\underbrace{\delta, \dots, \delta}_{N-k}, \delta + \lambda_{N-k+1}, \dots, \delta + \lambda_n), \quad (36)$$

where the first $N - k$ eigenvectors span the null space of \mathbf{C} . If Equation (34) is written based on the eigenvalue decomposition, we have

$$-2 \ln L(\mathbf{y}|\delta) \approx \sum_{i=1}^n \ln(\delta + \lambda_i) + \frac{1}{\delta} \sum_{i=1}^{N-k} \langle \mathbf{y}, \mathbf{U}^i \rangle^2 + \sum_{i=N-k+1}^n \frac{\langle \mathbf{y}, \mathbf{U}^i \rangle^2}{\delta + \lambda_i}, \quad (37)$$

or equivalently

$$-2 \ln L(\mathbf{y}|\delta) \approx \sum_{i=1}^n \ln(\delta + \lambda_i) + \frac{1}{\delta} \|\mathbf{y} - \mathbf{P}_{Im(\mathbf{C})}\mathbf{y}\|^2 + \sum_{i=N-k+1}^n \frac{\langle \mathbf{P}_{Im(\mathbf{C})}\mathbf{y}, \mathbf{U}^i \rangle^2}{\delta + \lambda_i}, \quad (38)$$

with the convention $\lambda_1 = \lambda_2 = \dots = \lambda_{N-k} = 0$. Based on (19), the term $\mathbf{y} - \mathbf{P}_{Im(\mathbf{C})}\mathbf{y}$ in the above formula is

$$\mathbf{y} - \mathbf{P}_{Im(\mathbf{C})}\mathbf{y} = \begin{bmatrix} y_1 - \bar{y}_1 \\ \vdots \\ y_{N_1} - \bar{y}_1 \\ \vdots \\ y_{N_1+\dots+N_{k-1}+1} - \bar{y}_k \\ \vdots \\ y_{N_1+\dots+N_k} - \bar{y}_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (39)$$

where \bar{y}^i , $i = 1, \dots, k$, designates the mean of response values at location i .

According to equations (39) and (25), $\|\mathbf{y} - \mathbf{P}_{Im(\mathbf{C})}\mathbf{y}\|^2 = \sum_{i=1}^k N_i s_i^2$. Hence,

Equation (38) using s_i^2 is updated as

$$-2 \ln L(\mathbf{y}|\delta) \approx \sum_{i=1}^n \ln(\delta + \lambda_i) + \frac{1}{\delta} \sum_{i=1}^k N_i s_i^2 + \sum_{i=N-k+1}^n \frac{\langle \mathbf{P}_{Im(\mathbf{C})}\mathbf{y}, \mathbf{U}^i \rangle^2}{\delta + \lambda_i}. \quad (40)$$

Let function $\Delta(\delta)$ express the difference between $-2 \ln L(\mathbf{y}|\delta)$ and $-2 \ln L(\mathbf{y}^+|\delta)$. Remark that $\mathbf{P}_{Im(\mathbf{C})}\mathbf{y} = \mathbf{P}_{Im(\mathbf{C})}\mathbf{y}^+$ because of our hypothesis $\bar{y}^i = \bar{y}^{+i}$, $i = 1, \dots, k$. The function $\Delta(\delta)$ is defined as

$$\Delta(\delta) \equiv -2 \ln L(\mathbf{y}^+|\delta) + 2 \ln L(\mathbf{y}|\delta) = \frac{1}{\delta} \sum_{i=1}^k N_i (s_i^{+2} - s_i^2), \quad (41)$$

and is monotonically decreasing.

Now we show that $\hat{\delta}^+$, the ML estimation of nugget from \mathbf{y}^+ , is never smaller than $\hat{\delta}$, the ML estimation of nugget from \mathbf{y} . Firstly, $\hat{\delta}^+$ cannot be smaller than $\hat{\delta}$. Indeed, if $\delta \leq \hat{\delta}$, then

$$\begin{aligned} -2 \ln L(\mathbf{y}^+|\delta) &= -2 \ln L(\mathbf{y}|\delta) + \Delta(\delta) \\ &\geq -2 \ln L(\mathbf{y}|\hat{\delta}) + \Delta(\delta) \\ &\geq -2 \ln L(\mathbf{y}|\hat{\delta}) + \Delta(\hat{\delta}) \\ &= -2 \ln L(\mathbf{y}^+|\hat{\delta}), \end{aligned} \quad (42)$$

which shows that $\hat{\delta}^+ \geq \hat{\delta}$. Secondly, if s_i^{+2} is strictly larger than s_i^2 , then $\hat{\delta}^+ > \hat{\delta}$ because the slope of $-2 \ln L(\mathbf{y}^+|\delta)$ is strictly negative at $\delta = \hat{\delta}$: The derivative of $-2 \ln L(\mathbf{y}^+|\delta)$ with respect to δ can be written as

$$\frac{d}{d\delta} (-2 \ln L(\mathbf{y}^+|\delta)) = \frac{d}{d\delta} (-2 \ln L(\mathbf{y}|\delta)) + \frac{d\Delta(\delta)}{d\delta}. \quad (43)$$

Since $\hat{\delta} = \arg \min -2 \ln L(\mathbf{y}|\delta)$, the second term in the right hand side of the above equation is equal to zero. Therefore, the derivative of $-2 \ln L(\mathbf{y}^+|\delta)$ with respect to δ reduces to

$$\frac{d}{d\delta} (-2 \ln L(\mathbf{y}^+|\delta)) = \frac{d}{d\delta} \left(\frac{1}{\delta} \sum_{i=1}^k N_i (s_i^{+2} - s_i^2) \right) = \frac{-1}{\delta^2} \sum_{i=1}^k N_i (s_i^{+2} - s_i^2). \quad (44)$$

The above derivative is strictly negative because $s_i^{+2} - s_i^2$ is positive and the proof is complete. \square

References

- [1] Ioannis Andrianakis and Peter G. Challenor. The effect of the nugget on Gaussian process emulators of computer models. *Computational Statistics & Data Analysis*, 56(12):4215–4228, 2012.

- [2] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.
- [3] François Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013.
- [4] Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vázquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793, 2012.
- [5] David A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley series in probability and mathematical statistics. Wiley, New York, 1991.
- [6] Adi Ben-Israel and Dan Cohen. On iterative computation of generalized inverses and associated projections. *SIAM Journal on Numerical Analysis*, 3(3):410–419, 1966.
- [7] Andrew J. Booker, J. E. Dennis, Jr., Paul D. Frank, David B. Serafini, Virginia Torczon, and Michael Trosset. A rigorous framework for optimization of expensive functions by surrogates. *Structural Optimization*, 17(17):1–13, 1998.
- [8] Noel A. C. Cressie. *Statistics for Spatial Data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1993.
- [9] George Davis and Max Morris. Six factors which affect the condition number of matrices associated with kriging. *Mathematical Geology*, 29(5):669–683, 1997.
- [10] Nicolas Durrande, David Ginsbourger, and Olivier Roustant. Additive covariance kernels for high-dimensional Gaussian process modeling. *Annales de la faculté des sciences de Toulouse Mathématiques*, 21(3):481–499, 2012.
- [11] Mark Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, 1997.
- [12] Robert B. Gramacy and Herbert K. Lee. Adaptive design and analysis of supercomputer experiments. *Technometrics*, 51(2):130–144, 2009.
- [13] Robert B. Gramacy and Herbert K. Lee. Cases for the nugget in modeling computer experiments. *Statistics and Computing*, 22(3):713–722, 2012.
- [14] Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.
- [15] Radford M. Neal. *Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification*. Technical report (University of Toronto .Dept. of Statistics). University of Toronto, 1997.

- [16] Jeremy E. Oakley and Anthony O’Hagan. Probabilistic sensitivity analysis of complex models: A bayesian approach. *Journal of the Royal Statistical Society, Series B*, 66:751–769, 2002.
- [17] Michael A. Osborne, Roman Garnett, and Stephen J. Roberts. Gaussian processes for global optimization. In *Proceedings of the 3rd Learning and Intelligent OptimizatioN Conference (LION 3)*, 2009.
- [18] Andrey Pepelyshev. The role of the nugget term in the Gaussian process method. In *mODa 9 – Advances in Model-Oriented Design and Analysis*, pages 149–156. Physica-Verlag HD, 2010.
- [19] Pritam Ranjan, Ronald Haynes, and Richard Karsten. A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. *Technometrics*, 53(4):366–378, 2011.
- [20] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press, 2005.
- [21] Gijs Rennen. Subset selection from large datasets for kriging modeling. *Structural and Multidisciplinary Optimization*, 38(6):545–569, 2009.
- [22] Olivier Roustant, David Ginsbourger, and Yves Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1):1–55, 2012.
- [23] Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):433–435, 1989.
- [24] Raviprakash Salagame and Russell Barton. Factorial hypercube designs for spatial correlation regression. *Journal of Applied Statistics*, 24(4):453–474, 1997.
- [25] Thomas J. Santner, Brian J. Williams, and William Notz. *The Design and Analysis of Computer Experiments*. Springer series in statistics. Springer-Verlag, New York, 2003.
- [26] Christopher Siefert, Virginia Torczon, and Michael W. Trosset. MAPS: Model-assisted pattern search, 1997–2000. www.cs.wm.edu/~va/software/maps/.
- [27] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 2009.
- [28] Tobias Wagner. A subjective review of the state of the art in model-based parameter tuning. In *the Workshop on Experimental Methods for the Assessment of Computational Systems (WEMACS)*, 2010.

- [29] William J. Welch, Robert J. Buck, Jerome Sacks, Henry P. Wynn, Toby J. Mitchell, and Max D. Morris. Screening, predicting, and computer experiments. *Technometrics*, 34(1):15–25, 1992.