Kriging outside of the Gaussian framework

J. Muré
Université Paris Diderot

Supervisor(s): Prof. J. Garnier (Université Paris Diderot), and Dr. L. Le Gratiet (EDF R&D)

Ph.D. expected duration: 2015-2018

Adress:
EDF Recherche et Développement - Dpt. de Management des Risques Industriels
6 quai Watier
78401 Chatou
France

Email: joseph.mure@edf.fr

Abstract:

Kriging has become a popular metamodel when it comes to emulating computer code. In addition to being easy to implement, it has the advantage of yielding confidence intervals regarding its predictions. However, it operates under the assumption that the code it is trying to emulate is a realisation of a stationary Gaussian process. Of course, the impact of this assumption on the prediction declines when the number of observation points increases, because Kriging operates conditionaly to the observations and is therefore able to adapt to them. But in a context where observation points are scarce (for example, when the emulated code is time-consuming), this assumption needs to be checked if the confidence intervals yielded at the end of the process are to mean anything.

In some cases the assumption is patently false. When the emulated code is known to yield only nonnegative values, one cannot expect it to follow a Gaussian distribution, be it stationary or not. The goal of this thesis is therefore to review all existing approaches to moving Kriging beyond its classical stationary Gaussian framework and to refine an existing one or to invent another one entirely. The chosen approach will then be applied to a code developed by EDF (Code_Carmel3D) in order to estimate probabilities of defect detection for certain components of nuclear power plants. We are currently favouring trans-Gaussian Kriging, where it is assumed that one member of a family of transformations indexed by a parameter λ must be applied to the code in order to get a stationary Gaussian process.

Let us set up some notations regarding classical Kriging. The emulated code is assumed to follow a Gaussian distribution whose mean function is (in most cases) assumed to be a linear combination of a few known functions h_1, h_2, ... , h_p with linear coefficients grouped in the unknown p-dimensional vector $\boldsymbol{\beta}$ and whose covariance function is unknown but assumed to be the product of the unknown one-dimensional variance parameter σ^2 and an unknown member of a known family of stationary correlation functions indexed by a (likely multidimensional) parameter $\boldsymbol{\theta}$. The code is observed at $n > p$ observation points, and $\boldsymbol{y}$ is the n-dimensional vector containing the observations. $\boldsymbol{H}$ is the known $n \times p$ full-ranked matrix whose p columns contain the values of the functions h_1, h_2, ..., h_p at every observation point. Therefore, $\boldsymbol{H}\boldsymbol{\beta}$ is the mean vector of the Gaussian vector $\boldsymbol{y}$. In order to fully determine the Kriging model, one must thus estimate the mean parameter $\boldsymbol{\beta}$, the variance parameter σ^2 and the correlation parameter $\boldsymbol{\theta}$. Noting $\boldsymbol{\Sigma_\theta}$ the correlation matrix of the Gaussian vector $\boldsymbol{y}$, as it naturally depends on correlation parameter $\boldsymbol{\theta}$, the likelihood of $\boldsymbol{y}$ is :

$$L(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\boldsymbol{\Sigma_\theta}|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{H\beta})^\top \boldsymbol{\Sigma_\theta}^{-1}(\boldsymbol{y}-\boldsymbol{H\beta})\right\} . \tag{1}$$

Let us note g_λ a generic member of the transformation family we are using, J_λ the absolute value of the Jacobian determinant of g_λ at the observation points. If $\boldsymbol{z}$ is the vector containing the observations, its likelihood can be thus written :

$$L(\boldsymbol{z}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta},\lambda) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\boldsymbol{\Sigma_\theta}|^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(g_\lambda(\boldsymbol{z})-\boldsymbol{H\beta})^\top \boldsymbol{\Sigma_\theta}^{-1}(g_\lambda(\boldsymbol{z})-\boldsymbol{H\beta})\right\} J_\lambda(\boldsymbol{z}) . \tag{2}$$

How are the parameters $\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta},\lambda$ to be estimated ? Using a Maximum Likelihood Estimator (MLE) may seem straightforward, but is not a very robust option, especially with few observation points. This problem could be remedied by using some prior distribution on parameters $\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta},\lambda$ to "help the MLE out".

At this point, we believe we had best separate the "Gaussian parameters" $\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}$ from the "transformation parameter" λ, and that the prior distribution on the parameters should be written as $\pi(\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta} \mid \lambda)\, \pi(\lambda)$. This allows for a more generic definition of the prior distribution, as $\pi(\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta} \mid \lambda)$ does not actually depend on the chosen transformation family.

The prior distribution on transformation parameter λ will of course depend on the chosen transformation family. It is thus impossible to specify a generic prior $\pi(\lambda)$ that would fit all cases. However, as far as Gaussian parameters $\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta}$ are concerned, there has been some research by [BDOS01] and [Pau05]. They recommend the use of the Bernardo reference prior with the "mean parameter" β as a nuisance parameter to "covariance parameters" $\sigma^2,\boldsymbol{\theta}$. We would argue that the covariance parameters should be further split, because there is no reason the variance parameter σ^2 should play the same role as the correlation parameter $\boldsymbol{\theta}$. Indeed, we would consider σ^2 to be a nuisance parameter with respect to $\boldsymbol{\theta}$. Of course, the greatest difficulty lies in determining the prior distribution on correlation parameter $\boldsymbol{\theta}$. In order to do that, we will focus on the case where the covariance function belongs to the tensorized Matérn class with fixed regularity. In this context, $\boldsymbol{\theta}$ is the vector containing the correlation lengths in every direction. We will then derive the Jeffreys independence prior on the correlation lengths with respect to the integrated likelihood $L^I(\boldsymbol{z} \mid \boldsymbol{\theta},\lambda) = \int L(\boldsymbol{z}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta},\lambda)\, \pi(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{\theta},\lambda)\, d\boldsymbol{\beta}\, d\sigma^2$. The Jeffreys independence prior strikes us as a better choice than the Jeffreys-rule prior because of the known problems of the Jeffreys-rule prior when the dimension increases and because the assumption of independence on the different correlation lengths does not seem unreasonable. After deriving this prior, we will endeavour to establish the propriety of the posterior distribution.

References

[BDOS01] James O Berger, Victor De Oliveira, and Bruno Sansó. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001.

[Pau05] Rui Paulo. Default priors for Gaussian processes. *Annals of Statistics*, pages 556–582, 2005.

Short biography – Joseph Muré has a Master's degree of Probability and Random models of the Université Pierre et Marie Curie and has just begun a PhD thesis financed by EDF. Its results will be applied to improving current metamodels regarding Probability of Defect Detection (POD) in nuclear power plants.