Kriging models with Gaussian processes - covariance function estimation and impact of spatial sampling

François Bachoc

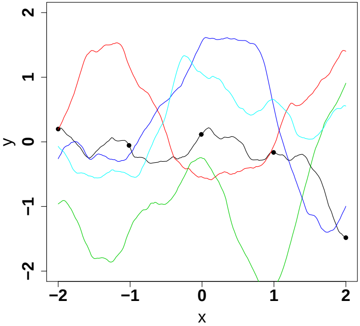
former PhD advisor: Josselin Garnier
former CEA advisor: Jean-Marc Martinez

Department of Statistics and Operations Research, University of Vienna
(Former PhD student at CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France
and LPMA, Université Paris Diderot)

JPS 2014 - Forges-les-Eaux - March 2014

Kriging model

Study of a **single realization** of a Gaussian process $Y(x)$ on a domain $\mathcal{X} \in \mathbb{R}^d$



Goal

Predicting the continuous realization function, from a finite number of **observation points**

The Gaussian process

The Gaussian process

- We consider that the Gaussian process is centered, $\forall x, \mathbb{E}(Y(x)) = 0$
- The Gaussian process is hence characterized by its covariance function

The covariance function

- The function $K : \mathcal{X}^2 \to \mathbb{R}$, defined by $K(x_1, x_2) = cov(Y(x_1), Y(x_2))$

In most classical cases :

- Stationarity : $K(x_1, x_2) = K(x_1 - x_2)$
- Continuity : $K(x)$ is continuous $\Rightarrow$ Gaussian process realizations are continuous
- Decrease : $K(x)$ is a decreasing function for $x \geq 0$ and $\lim_{x \to +\infty} K(x) = 0$

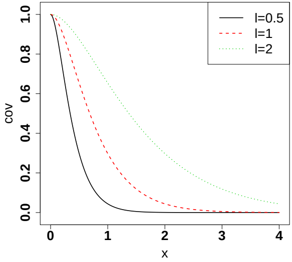
Example of the Matérn $\frac{3}{2}$ covariance function on $\mathbb{R}$

The Matérn $\frac{3}{2}$ covariance function, for a Gaussian process on $\mathbb{R}$ is parameterized by

- A variance parameter $\sigma^2 > 0$
- A correlation length parameter $\ell > 0$

It is defined as

$$C_{\sigma^2,\ell}(x_1, x_2) = \sigma^2 \left(1 + \sqrt{6}\frac{|x_1 - x_2|}{\ell} \right) e^{-\sqrt{6}\frac{|x_1 - x_2|}{\ell}}$$



Interpretation

- Stationarity, continuity, decrease
- σ^2 corresponds to the order of magnitude of the functions that are realizations of the Gaussian process
- ℓ corresponds to the speed of variation of the functions that are realizations of the Gaussian process

$\Rightarrow$ Natural generalization on $\mathbb{R}^d$

Covariance function estimation

Parameterization

Covariance function model $\{\sigma^2 K_\theta, \sigma^2 \geq 0, \theta \in \Theta\}$ for the Gaussian Process Y.

- σ^2 is the variance parameter
- θ is the multidimensional correlation parameter. K_θ is a stationary correlation function.

Observations

Y is observed at $x_1, ..., x_n \in \mathcal{X}$, yielding the Gaussian vector $y = (Y(x_1), ..., Y(x_n))$.

Estimation

Objective : build estimators $\hat{\sigma}^2(y)$ and $\hat{\theta}(y)$

Maximum Likelihood for estimation

Explicit Gaussian likelihood function for the observation vector y

Maximum Likelihood

Define $\mathbf{R}_\theta$ as the correlation matrix of $y = (Y(x_1), ..., Y(x_n))$ with correlation function K_θ and $\sigma^2 = 1$.
The Maximum Likelihood estimator of (σ^2, θ) is

$$(\hat{\sigma}^2_{ML}, \hat{\theta}_{ML}) \in \underset{\sigma^2 \geq 0, \theta \in \Theta}{\operatorname{argmin}} \frac{1}{n} \left(\ln\left(|\sigma^2 \mathbf{R}_\theta|\right) + \frac{1}{\sigma^2} y^t \mathbf{R}_\theta^{-1} y \right)$$

$\Rightarrow$ Numerical optimization with $O(n^3)$ criterion

Cross Validation for estimation

- $\hat{y}_{\theta,i,-i} = \mathbb{E}_{\sigma^2,\theta}(Y(x_i)|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$
- $\sigma^2 c^2_{\theta,i,-i} = var_{\sigma^2,\theta}(Y(x_i)|y_1, ..., y_{i-1}, y_{i+1}, ..., y_n)$

Leave-One-Out criteria we study

$$\hat{\theta}_{CV} \in \underset{\theta\in\Theta}{\mathrm{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_{\theta,i,-i})^2$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_{\hat{\theta}_{CV},i,-i})^2}{\hat{\sigma}^2_{CV} c^2_{\hat{\theta}_{CV},i,-i}} = 1 \Leftrightarrow \hat{\sigma}^2_{CV} = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{y}_{\hat{\theta}_{CV},i,-i})^2}{c^2_{\hat{\theta}_{CV},i,-i}}$$

Robustness

We show that Cross Validation can be preferable to Maximum Likelihood when the covariance function model is misspecified

Bachoc F, Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification, *Computational Statistics and Data Analysis 66 (2013) 55-69*

Virtual Leave One Out formula

Let $\mathbf{R}_\theta$ be the covariance matrix of $y = (y_1, ..., y_n)$ with correlation function K_θ and $\sigma^2 = 1$

Virtual Leave-One-Out

$$y_i - \hat{y}_{\theta,i,-i} = \frac{1}{(\mathbf{R}_\theta^{-1})_{i,i}} \left(\mathbf{R}_\theta^{-1} y\right)_i \quad \text{and} \quad c_{i,-i}^2 = \frac{1}{(\mathbf{R}_\theta^{-1})_{i,i}}$$

O. Dubrule, Cross Validation of Kriging in a Unique Neighborhood, *Mathematical Geology*, 1983.

Using the virtual Cross Validation formula :

$$\hat{\theta}_{CV} \in \underset{\theta \in \Theta}{\text{argmin}} \frac{1}{n} y^t \mathbf{R}_\theta^{-1} diag(\mathbf{R}_\theta^{-1})^{-2} \mathbf{R}_\theta^{-1} y$$

and

$$\hat{\sigma}_{CV}^2 = \frac{1}{n} y^t \mathbf{R}_{\hat{\theta}_{CV}}^{-1} diag(\mathbf{R}_{\hat{\theta}_{CV}}^{-1})^{-1} \mathbf{R}_{\hat{\theta}_{CV}}^{-1} y$$

$\Rightarrow$ Same computational cost as ML

Prediction with fixed covariance function

Gaussian process Y observed at $x_1, ..., x_n$ and predicted at x_{new}
$y = (Y(x_1), ..., Y(x_n))^t$

Once the covariance function has been estimated and fixed

- $\mathbf{R}$ is the covariance matrix of Y at $x_1, ..., x_n$
- r is the covariance vector of Y between $x_1, ..., x_n$ and x_{new}

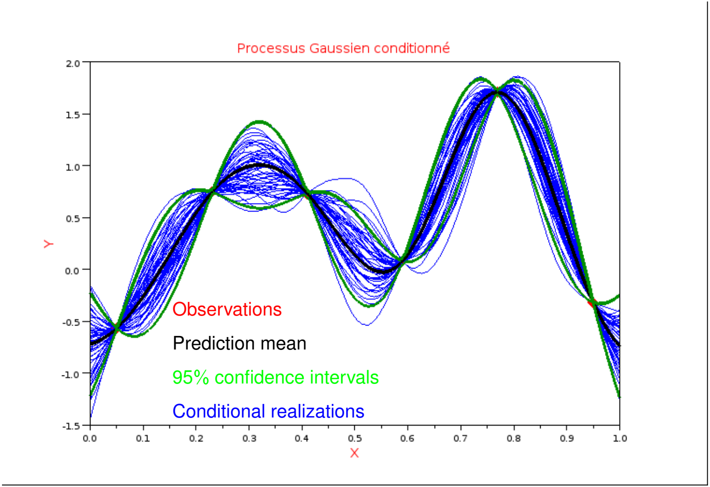
Prediction

The prediction is $\hat{Y}(x_{new}) := \mathbb{E}(Y(x_{new})|Y(x_1), ..., Y(x_n)) = r^t \mathbf{R}^{-1} y$.

Predictive variance

The predictive variance is
$var(Y(x_{new})|Y(x_1), ..., Y(x_n)) = \mathbb{E}\left[(Y(x_{new}) - \hat{Y}(x_{new}))^2\right] = var(Y(x_{new})) - r^t \mathbf{R}^{-1} r$.

Illustration of prediction



Application to computer experiments

Computer model

A computer model, computing a given variable of interest, corresponds to a deterministic function $\mathbb{R}^d \to \mathbb{R}$. Evaluations of this function are time consuming

- Examples : Simulation of a nuclear fuel pin, of thermal-hydraulic systems, of components of a car, of a plane...

Kriging model for computer experiments

Basic idea : representing the code function by a realization of a Gaussian process

- Bayesian framework on a fixed function

What we obtain

- metamodel of the code : the Kriging prediction function approximates the code function, and its evaluation cost is negligible
- Error indicator with the predictive variance
- Full conditional Gaussian process $\Rightarrow$ possible goal-oriented iterative strategies for optimization, failure domain estimation, small probability problems, code calibration...

Conclusion

Kriging models :

- The covariance function characterizes the Gaussian process
- It is estimated first. Here we consider Maximum Likelihood and Cross Validation estimation
- Then we can compute prediction and predictive variances with explicit matrix vector formulas
- Widely used for computer experiments

1. Kriging models with Gaussian processes

2. Asymptotic analysis of covariance function estimation and of spatial sampling impact

Framework and objectives

Estimation

We do not make use of the distinction $\sigma^2, \boldsymbol{\theta}$. Hence we use the set $\{K_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ of stationary covariance functions for the estimation.

Well-specified model

The true covariance function K of the Gaussian Process belongs to the set $\{K_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$. Hence

$$K = K_{\boldsymbol{\theta}_0}, \boldsymbol{\theta}_0 \in \Theta$$

Objectives

- Study the consistency and asymptotic distribution of the Cross Validation estimator
- Confirm that, asymptotically, Maximum Likelihood is more efficient
- Study the influence of the spatial sampling on the estimation

- Spatial sampling : initial design of experiments for Kriging
- It has been shown that irregular spatial sampling is often an advantage for covariance parameter estimation

Stein M, Interpolation of Spatial Data : Some Theory for Kriging, *Springer, New York, 1999. Ch.6.9*.

Zhu Z, Zhang H, Spatial sampling design under the infill asymptotics framework, *Environmetrics 17 (2006) 323-337*.

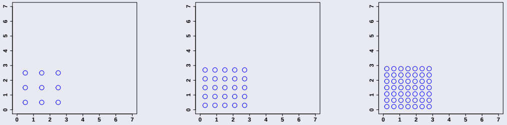
- Our question : can we confirm this finding in an asymptotic framework

Two asymptotic frameworks for covariance parameter estimation

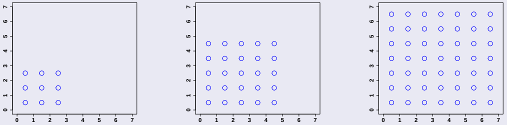
Asymptotics (number of observations $n \to +\infty$) is an active area of research (Maximum-Likelihood estimator)

Two main asymptotic frameworks

- fixed-domain asymptotics : The observation points are dense in a bounded domain



- increasing-domain asymptotics : A minimum spacing exists between the observation points $\longrightarrow$ infinite observation domain.



Choice of the asymptotic framework

Comments on the two asymptotic frameworks

- fixed-domain asymptotics
  From 80'-90' and onwards. Fruitful theory

  Stein, M., Interpolation of Spatial Data Some Theory for Kriging, *Springer, New York, 1999*.

  However, when convergence in distribution is proved, the asymptotic distribution does not depend on the spatial sampling $\longrightarrow$ Impossible to compare sampling techniques for estimation in this context
- increasing-domain asymptotics :
  Asymptotic normality proved for Maximum-Likelihood (under conditions that are not simple to check)

  Sweeting, T., Uniform asymptotic normality of the maximum likelihood estimator, *Annals of Statistics 8 (1980) 1375-1381*.

  Mardia K, Marshall R, Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika 71 (1984) 135-146*.

  (no results for CV)

We study increasing-domain asymptotics for ML and CV under irregular sampling

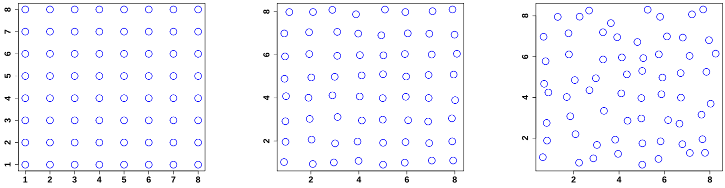
The randomly perturbed regular grid that we study

- Observation point i :

$$\boldsymbol{v}_i + \epsilon X_i$$

  - $(\boldsymbol{v}_i)_{i \in \mathbb{N}^*}$: regular square grid of step one in dimension d
  - $(X_i)_{i \in \mathbb{N}^*}$: *iid* with symmetric distribution on $[-1, 1]^d$
- $\epsilon \in (-\frac{1}{2}, \frac{1}{2})$ is the regularity parameter of the grid.
  - $\epsilon = 0 \longrightarrow$ regular grid.
  - $|\epsilon|$ close to $\frac{1}{2} \longrightarrow$ irregularity is maximal

Illustration with $\epsilon = 0, \frac{1}{8}, \frac{3}{8}$



Consistency and asymptotic normality

Under general summability, regularity and identifiability conditions, we show

Proposition : for ML

- a.s convergence of the random Fisher information : The random trace $\frac{1}{2n} Tr \left(\mathbf{R}_{\boldsymbol{\theta}_0}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}_0}}{\partial \boldsymbol{\theta}_i} \mathbf{R}_{\boldsymbol{\theta}_0}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}_0}}{\partial \boldsymbol{\theta}_j} \right)$ converges a.s to the element $(\mathbf{I}_{ML})_{i,j}$ of a $p \times p$ deterministic matrix $\mathbf{I}_{ML}$ as $n \to +\infty$
- asymptotic normality : With $\Sigma_{ML} = \mathbf{I}_{ML}^{-1}$

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0 \right) \to \mathcal{N} \left(0, \Sigma_{ML} \right)$$

Proposition : for CV

Same result with more complex expressions for asymptotic covariance matrix Σ_{CV}

Main ideas for the proof

- A central tool : because of the minimum distance between observation points : the eigenvalues of the random matrices involved are uniformly lower and upper bounded
- For consistency : bounding from below the difference of M-estimator criteria between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ by the integrated square difference between $K_{\boldsymbol{\theta}}$ and $K_{\boldsymbol{\theta}_0}$
- For almost-sure convergence of random traces : block-diagonal approximation of the random matrices involved and Cauchy criterion
- For asymptotic normality of criterion gradient : almost-sure (with respect to the random perturbations) Lindeberg-Feller Central Limit Theorem
- Conclude with classical M-estimator method

The asymptotic covariance matrices $\Sigma_{ML,CV}$ depend only on the regularity parameter ϵ.
$\boxed{\longrightarrow}$ in the sequel, we study the functions $\epsilon \to \Sigma_{ML,CV}$

Matérn model in dimension one

$$K_{\ell,\nu}(x_1, x_2) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu} \frac{|x_1 - x_2|}{\ell} \right)^{\nu} K_{\nu} \left(2\sqrt{\nu} \frac{|x_1 - x_2|}{\ell} \right),$$

with Γ the Gamma function and K_ν the modified Bessel function of second order

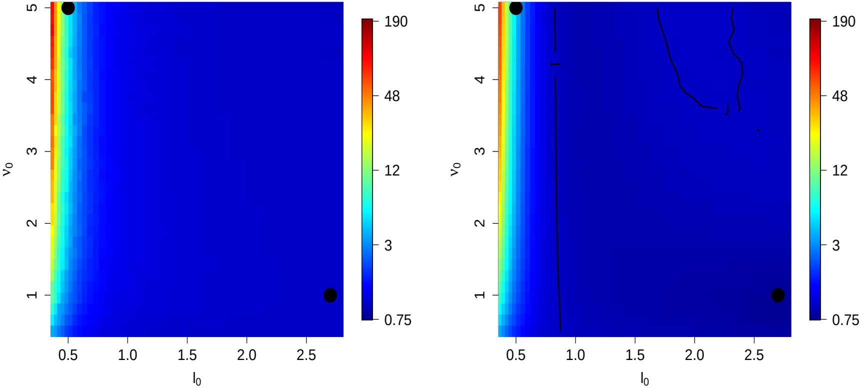
We consider

- The estimation of ℓ when ν_0 is known
- The estimation of ν when ℓ_0 is known

$\Longrightarrow$ We study scalar asymptotic variances

Results for the Matérn model (1/2)

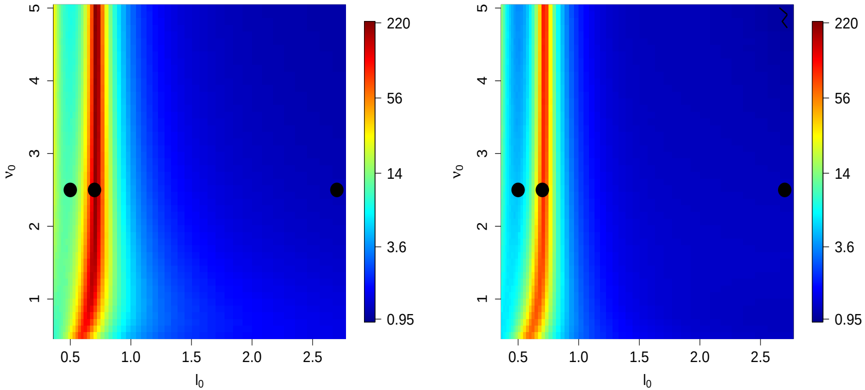
Estimation of ℓ when ν_0 is known.
Level plot of $\left[\Sigma_{ML,CV}(\epsilon = 0)\right] / \left[\Sigma_{ML,CV}(\epsilon = 0.45)\right]$ in $\ell_0 \times \nu_0$ for ML (left) and CV (right)



Perturbations of the regular grid are always beneficial for ML

Results for the Matérn model (2/2)

Estimation of ν when ℓ_0 is known.
Level plot of $\left[\Sigma_{ML,CV}(\epsilon = 0)\right] / \left[\Sigma_{ML,CV}(\epsilon = 0.45)\right]$ in $\ell_0 \times \nu_0$ for ML (left) and CV (right)



Perturbations of the regular grid are always beneficial for ML and CV

Conclusion on covariance function estimation and spatial sampling

- CV is consistent and has the same rate of convergence as ML
- We confirm (not presented here) that ML is more efficient
- Strong irregularity in the sampling is an advantage for covariance function estimation
  - With ML, irregular sampling is more often an advantage than with CV
  - We show that, however, regular sampling is better for prediction with known covariance function $\Longrightarrow$ motivation for using space-filling samplings augmented with some clustered observation points

Z. Zhu and H. Zhang, Spatial Sampling Design Under the Infill Asymptotics Framework, *Environmetrics 17 (2006) 323-337*.

L. Pronzato and W. G. Müller, Design of computer experiments : space filling and beyond, *Statistics and Computing 22 (2012) 681-701*.

For further details :

F. Bachoc, Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes, *Journal of Multivariate Analysis 125 (2014) 1-35*.

Some perspectives

Ongoing work

- Asymptotic analysis of the case of a misspecified covariance function model with purely random sampling

Other potential perspectives

- Designing other CV procedures (LOO error weighting, decorrelation and penalty term) to reduce the variance
- Start studying the fixed-domain asymptotics of CV, in the particular cases where it is done for ML

French community

French community

- GDR MASCOT-NUM *www.gdr-mascotnum.fr*
- consortium ReDICE *www.redice-project.org*

Thank you for your attention !