

# Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes

François Bachoc

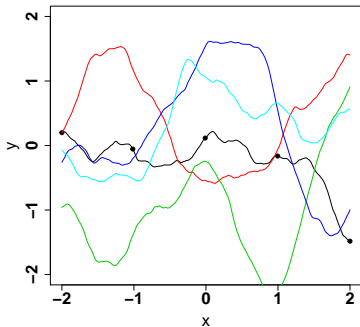
Department of Statistics and Operations Research, University of Vienna

(This work was performed while the author was a PhD student, supervised by **Josselin Garnier** (Paris Diderot University) and **Jean-Marc Martinez** (French Atomic Energy commission))

UCM 2014 - Sheffield - July 2014

- 1 Covariance function estimation for Gaussian processes
- 2 Objective : asymptotic analysis of estimation and of spatial sampling impact
- 3 Randomly perturbed regular grid and asymptotic normality
- 4 Impact of spatial sampling

- **Kriging model** : study of a **single realization** of a **Gaussian process**  $Y(x)$  on a domain  $\mathcal{X} \subset \mathbb{R}^d$
- **Goal** : **predicting** the continuous realization function, from a finite number of **observation points**



## Classical plug-in approach

Given an observation vector of  $Y$  at  $x_1, \dots, x_n \in \mathcal{X}$ ,  $y = (Y(x_1), \dots, Y(x_n))$  :

- 1 **Estimation** of the covariance function
- 2 Assume the covariance function is **known** and equal to its estimate. Then **prediction** of the Gaussian process realization is carried out with the **explicit** Kriging equations

⇒ This talk is mainly focused on covariance function estimation

## Covariance function

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

- We assume here for simplicity that the Gaussian process is **centered** ( $\mathbb{E}(Y(x)) = 0$ )  
⇒ the covariance function **characterizes** the Gaussian process

## Parameterization

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

- $\sigma^2$  is the **variance** parameter
- $\theta$  is a multidimensional **correlation parameter**.  $K_\theta$  is a **stationary** correlation function

## Observations

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

## Estimation

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

Explicit Gaussian likelihood function for the observation vector  $y$

## Maximum Likelihood

Define  $\mathbf{R}_\theta$  as the correlation matrix of  $y = (Y(x_1), \dots, Y(x_n))$  with correlation function  $K_\theta$  and  $\sigma^2 = 1$ .

The Maximum Likelihood estimator of  $(\sigma^2, \theta)$  is

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

⇒ Numerical optimization with  $O(n^3)$  criterion

⇒ Most **standard** estimation method. Expected to work best when the covariance function model is **well specified**

# Cross Validation for estimation

- $\hat{y}_{\theta,i,-i} = \mathbb{E}_{\sigma^2,\theta}(Y(x_i)|y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$
- $\sigma^2 c_{\theta,i,-i}^2 = \text{var}_{\sigma^2,\theta}(Y(x_i)|y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$

## Leave-One-Out criteria we study

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

and

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

## Robustness

We showed 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

Let  $\mathbf{R}_\theta$  be the covariance matrix of  $y = (y_1, \dots, y_n)$  with correlation function  $K_\theta$  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 :

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

and

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

⇒ Same computational cost as ML

- The covariance function characterizes the Gaussian process
- Standard Kriging approach : estimation and prediction with "fixed" estimated covariance function
  - ⇒ we focus on the estimation step
- We consider Maximum Likelihood and Cross Validation estimation
  - ⇒ numerical optimization with similar computational cost for both methods
  - ⇒ Maximum Likelihood : the standard method
  - ⇒ Cross Validation : can be a more appropriate alternative



- 1 Covariance function estimation for Gaussian processes
- 2 Objective : asymptotic analysis of estimation and of spatial sampling impact
- 3 Randomly perturbed regular grid and asymptotic normality
- 4 Impact of spatial sampling

## Estimation

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



## Well-specified model

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

$$K = K_{\theta_0}, \theta_0 \in \Theta$$

## 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

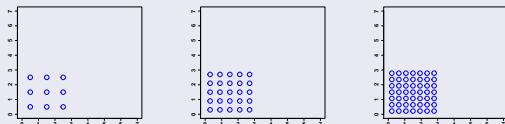
- **Spatial sampling** : initial design of experiments for Kriging
- It has been shown that irregular spatial sampling is often an advantage for covariance parameter estimation
  -  Stein M, Interpolation of Spatial Data : Some Theory for Kriging, *Springer, New York, 1999. Ch.6.9.*
  -  Zhu Z, Zhang 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

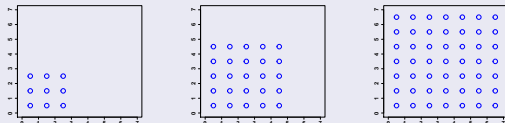
Asymptotics (number of observations  $n \rightarrow +\infty$ ) is an active area of research (Maximum Likelihood estimator)

## 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  
→ infinite observation domain.



## 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 → **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 with spatial sampling with tunable irregularity

- 1 Covariance function estimation for Gaussian processes
- 2 Objective : asymptotic analysis of estimation and of spatial sampling impact
- 3 Randomly perturbed regular grid and asymptotic normality**
- 4 Impact of spatial sampling

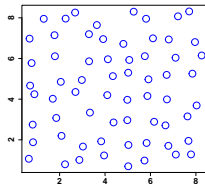
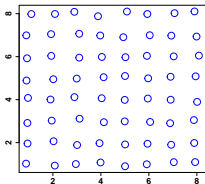
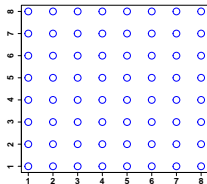
# The randomly perturbed regular grid that we study

- Observation point  $i$  :

$$v_i + \epsilon X_i$$

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

$\frac{1}{2n} \text{Tr} \left( \mathbf{R}_{\theta_0}^{-1} \frac{\partial \mathbf{R}_{\theta_0}}{\partial \theta_i} \mathbf{R}_{\theta_0}^{-1} \frac{\partial \mathbf{R}_{\theta_0}}{\partial \theta_j} \right)$  converges a.s to the element  $(\mathbf{I}_{ML})_{i,j}$  of a  $p \times p$  deterministic matrix  $\mathbf{I}_{ML}$  as  $n \rightarrow +\infty$

- **asymptotic normality** : With  $\Sigma_{ML} = \mathbf{I}_{ML}^{-1}$

$$\sqrt{n} (\hat{\theta}_{ML} - \theta_0) \rightarrow \mathcal{N}(0, \Sigma_{ML})$$

## Proposition : for CV

Same result with more complex expressions for asymptotic covariance matrix  $\Sigma_{CV}$

$\Rightarrow$  Same rate of convergence for ML and CV

$\Rightarrow$  The asymptotic covariance matrices  $\Sigma_{ML,CV}$  depend **only** on the regularity parameter  $\epsilon$

$\boxed{\rightarrow}$  we can study the functions  $\epsilon \rightarrow \Sigma_{ML,CV}$



- 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  $\theta$  and  $\theta_0$  by the integrated square difference between  $K_\theta$  and  $K_{\theta_0}$
- 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

- 1 Covariance function estimation for Gaussian processes
- 2 Objective : asymptotic analysis of estimation and of spatial sampling impact
- 3 Randomly perturbed regular grid and asymptotic normality
- 4 Impact of spatial sampling**

We study the functions  $\epsilon \rightarrow \Sigma_{ML,CV}$

## 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  $\Gamma$  the Gamma function and  $K_\nu$  the modified Bessel function of second order

$\Rightarrow \ell \geq 0$  : correlation length

$\Rightarrow \nu \geq 0$  : smoothness parameter

We consider

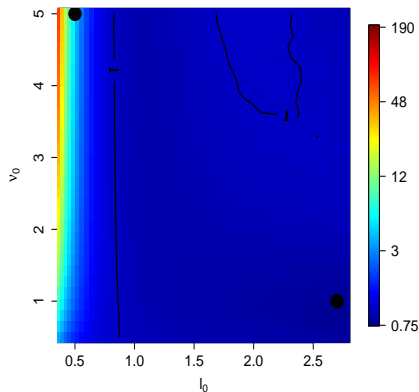
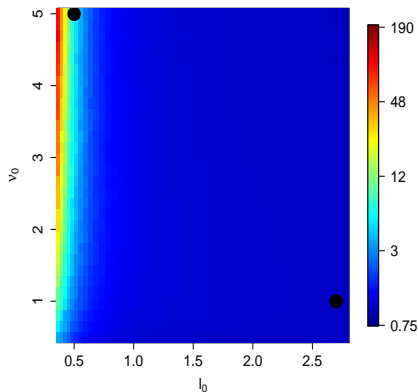
- The estimation of  $\ell$  when  $\nu_0$  is known
- The estimation of  $\nu$  when  $\ell_0$  is known

$\Rightarrow$  We study scalar asymptotic variances

# Results for the Matérn model (1/2)

Estimation of  $\ell$  when  $\nu_0$  is known.

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

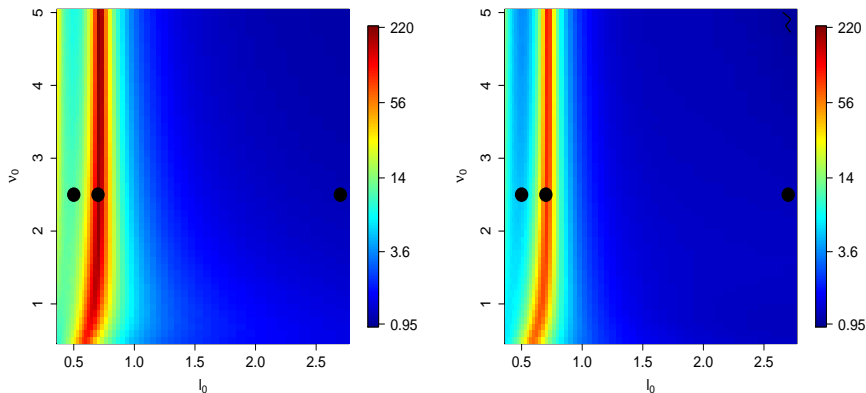


Perturbations of the regular grid are always beneficial for ML

## Results for the Matérn model (2/2)

Estimation of  $\nu$  when  $\ell_0$  is known.

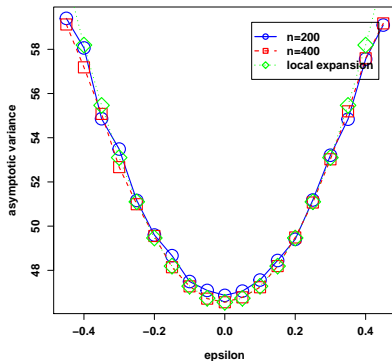
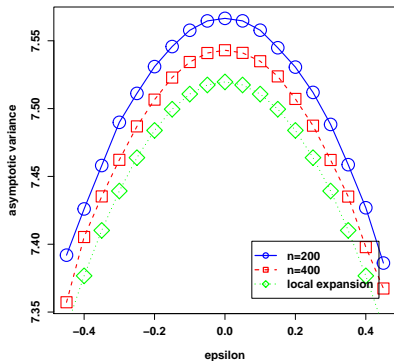
Level plot of  $[\Sigma_{ML,CV}(\epsilon = 0)] / [\Sigma_{ML,CV}(\epsilon = 0.45)]$  in  $\ell_0 \times \nu_0$  for ML (left) and CV (right)



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

# Some particular functions $\epsilon \rightarrow \Sigma_{ML,CV}$ (1/2)

Estimation of  $\ell$  when  $\nu_0$  is known, for  $\ell_0 = 2.7$ ,  $\nu_0 = 1$ .  
Plot of  $\epsilon \rightarrow \Sigma_{ML,CV}$  for ML (left) and CV (right)

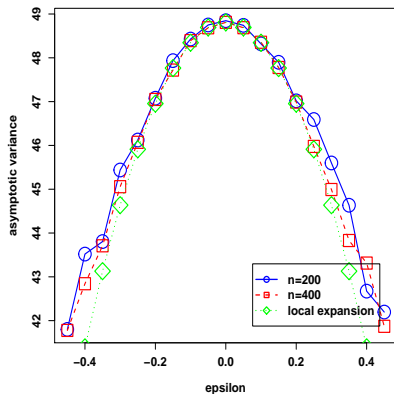
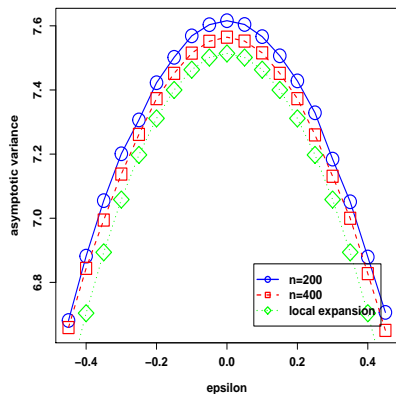


The asymptotic variance of CV is significantly larger than that of ML (but ML uses the known variance value, contrary to CV)

## Some particular functions $\epsilon \rightarrow \Sigma_{ML,CV}$ (2/2)

Estimation of  $\nu$  when  $\ell_0$  is known, for  $\ell_0 = 2.7$ ,  $\nu_0 = 2.5$ .

Plot of  $\epsilon \rightarrow \Sigma_{ML,CV}$  for ML (left) and CV (right)



The asymptotic variance of CV is significantly larger than that of ML (but ML uses the known variance value, contrary to CV)

# Prediction error with estimated covariance parameters

Let  $\hat{Y}_\theta(t)$  be the Kriging prediction of the Gaussian process  $Y$  at  $t$ , under correlation function  $K_\theta$   
Let  $N_{1,n}$  so that  $N_{1,n}^d \leq n < (N_{1,n} + 1)^d$  ( $\approx$  edge length of the spatial sampling)

## Integrated prediction error

$$E_{\epsilon,\theta} := \frac{1}{N_{1,n}^d} \int_{[0, N_{1,n}]^d} \left( \hat{Y}_\theta(t) - Y(t) \right)^2 dt$$

We show

## Proposition

Consider a consistent estimator  $\hat{\theta}$  of  $\theta_0$ . Then

$$|E_{\epsilon,\theta_0} - E_{\epsilon,\hat{\theta}}| = o_p(1)$$

Furthermore, there exists a constant  $A > 0$  so that for all  $n$ ,

$$\mathbb{E} (E_{\epsilon,\theta_0}) \geq A$$

$\Rightarrow$  No first-order difference of prediction error with estimated covariance between ML and CV (in the well-specified case)

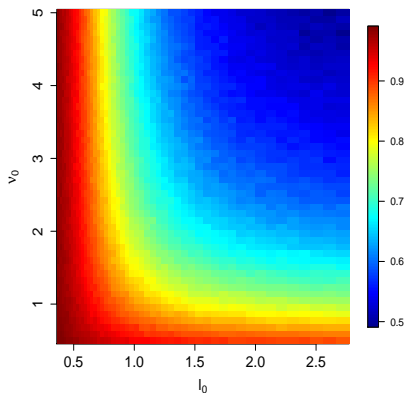
$\Rightarrow$  Other possible asymptotic framework showing a difference in the well-specified case (?)





# Impact of spatial sampling on prediction error

Matérn model in dimension one. Plot in  $\ell_0 \times \nu_0$  of an estimate (for  $n = 100$ ) of


$$\frac{\mathbb{E} [E_{\epsilon, \ell_0, \nu_0}(\epsilon = 0)]}{\mathbb{E} [E_{\epsilon, \ell_0, \nu_0}(\epsilon = 0.45)]}$$



The regular grid is always better for prediction mean square error

- CV is consistent and has the same rate of convergence as ML
  - We confirm that ML is more efficient
  - In our numerical study : strong irregularity in the sampling is an advantage for covariance function estimation
    - With ML, irregular sampling is more often an advantage than with CV
    - However, regular sampling is better for prediction with known covariance function
      - ⇒ 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.

## 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

Thank you for your attention !