Maximum Likelihood and Cross Validation for Kriging hyper-parameter estimation

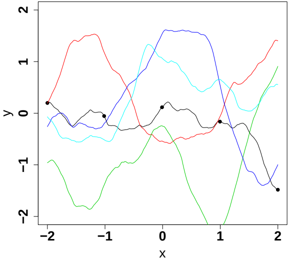
François Bachoc
Josselin Garnier
Jean-Marc Martinez

CEA-Saclay, DEN, DM2S, STMF, LGLS, F-91191 Gif-Sur-Yvette, France
LPMA, Université Paris 7

July 2013

Kriging model with Gaussian process

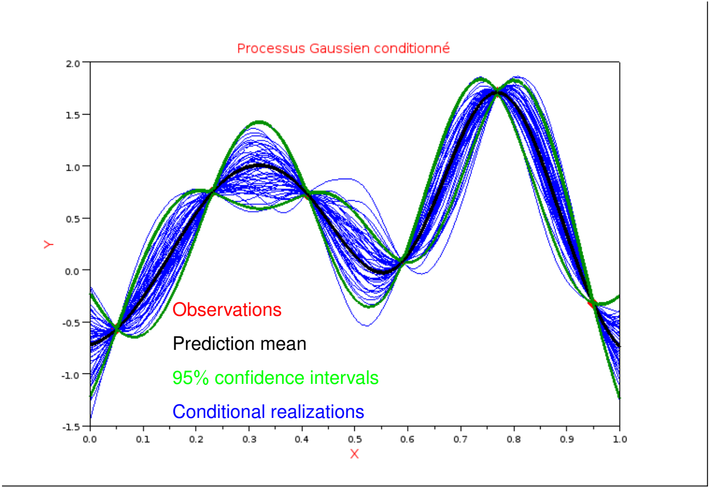
Basic idea : representing a deterministic and unknown function as the realization of a Gaussian process



Notation

Gaussian process Y defined on the set $\mathcal{X}$.

When the distribution of the Gaussian process is known



All this from explicit matrix vector formula

Covariance function estimation

Parameterization

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

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

Estimation

Y is observed at $x_1, ..., x_n \in \mathcal{X}$, yielding the Gaussian vector $y = (Y(x_1), ..., Y(x_n))$.
Estimators $\hat{\sigma}^2(y)$ and $\hat{\theta}(y)$

"Plug-in" Kriging prediction

1. Estimate the covariance function
2. Assume that the covariance function is fixed and carry out the explicit Kriging equations

Maximum Likelihood for estimation

Explicit Gaussian likelihood function for the observation vector y

Maximum Likelihood

Define $\mathbf{R}_\theta$ as the correlation matrix of $y = (Y(x_1), ..., Y(x_n))$ under correlation function K_θ.

The Maximum Likelihood estimator of (σ^2, θ) is

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

Cross Validation for estimation

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

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

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

Virtual Leave One Out formula

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

Virtual Leave-One-Out

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

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

Using the virtual Cross Validation formula :

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

and

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

Introduction to Kriging and covariance function estimation

Finite sample analysis of ML and CV under model misspecification

Asymptotic analysis of ML and CV in the well-specified case

Conclusion

Objectives

We want to study the cases of model misspecification, that is to say the cases when the true covariance function K_1 of Y is far from $\mathcal{K} = \left\{\sigma^2 K_\theta, \sigma^2 \geq 0, \theta \in \Theta\right\}$

In this context we want to compare Leave-One-Out and Maximum Likelihood estimators from the point of view of prediction mean square error and point-wise estimation of the prediction mean square error

We proceed in two steps

- When $\mathcal{K} = \left\{\sigma^2 K_2, \sigma^2 \geq 0\right\}$, with K_2 a correlation function, and K_1 the true unit-variance covariance function : theoretical formula and numerical tests
- In the general case : numerical studies

Case of variance hyper-parameter estimation

- $\hat{y}_0$: Kriging prediction of $y_0 := Y(x_0)$ with fixed misspecified correlation function K_2
- $\mathbb{E}\left[(\hat{y}_0 - y_0)^2 \middle| y \right]$: conditional mean square error of the non-optimal prediction
- One estimates σ^2 by $\hat{\sigma}^2$.
- Conditional mean square error of $\hat{y}_0$ estimated by $\hat{\sigma}^2 c_{x_0}^2$ with $c_{x_0}^2$ fixed by K_2

The Risk

We study the Risk criterion for an estimator $\hat{\sigma}^2$ of σ^2

$$\mathcal{R}_{\hat{\sigma}^2, x_0} = \mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - y_0)^2 \middle| y \right] - \hat{\sigma}^2 c_{x_0}^2 \right)^2 \right]$$

$\longrightarrow$ Explicit formula for estimators of σ^2 that are quadratic forms of the observation vector

Results for the variance hyper-parameter estimation

Procedure

- Designs of experiments studied : SRS, LHS-Maximin and regular grid
- We make the distance between K_1 and K_2 vary, starting from 0.
  - For sample K_1 and K_2 are Matérn, with $\ell_1 = \ell_2 = 1.2$, $\nu_1 = 1.5$, and $\nu_2 \in [0.5, 2.5]$
- We calculate and study the Risk criterion

Results

- For not too regular design of experiments : CV is more robust than ML to misspecification
  - Larger variance but smaller bias for CV
  - The bias term becomes dominating when $K_1 \neq K_2$
- For regular design of experiments, CV is less robust to model misspecification

Case of variance and correlation hyper-parameter estimation

For variance and correlation hyper-parameter estimation

- Numerical study on analytical functions
  - Ishigami function ($d = 3$)
  - Morris function ($d = 10$)
- Confirmation of the results of the variance estimation case

For more details

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, http ://dx.doi.org/10.1016/j.csda.2013.03.016*.

Introduction to Kriging and covariance function estimation

Finite sample analysis of ML and CV under model misspecification

Asymptotic analysis of ML and CV in the well-specified case

Conclusion

Framework and objectives

Estimation

We do not make use of the distinction σ^2, θ. 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 Maximum Likelihood is asymptotically more efficient
- Study the influence of the spatial sampling on the estimation

Spatial sampling for hyper-parameter estimation

- Spatial sampling : Initial design of experiment for Kriging
- It has been shown that irregular spatial sampling is often an advantage for hyper-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*.

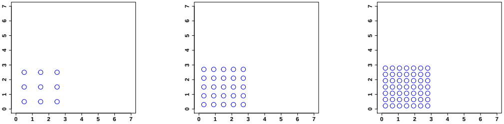
- Our question : Is irregular sampling always better than regular sampling for hyper-parameter estimation ?

Two asymptotic frameworks for hyper-parameter estimation

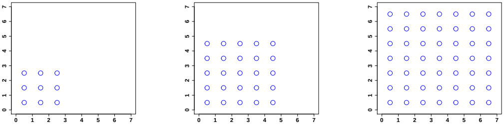
Asymptotics (number of observations $n \to +\infty$) is an area of active research (Maximum-Likelihood estimator)

Two main asymptotic frameworks

- fixed-domain asymptotics : The observations 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 general conditions

  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*.

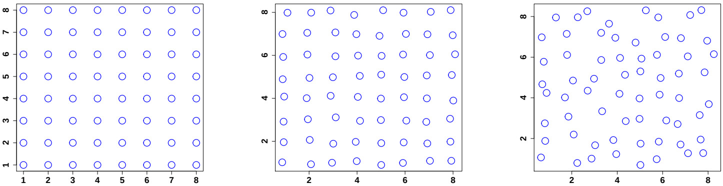
Randomly perturbed regular grid (1/2)

- Observation point i :

$$v_i + \epsilon X_i$$

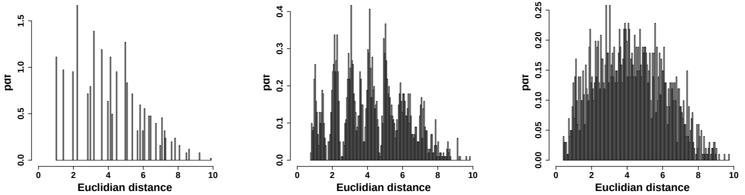
  - $(v_i)_{i \in \mathbb{N}^*}$: regular square grid of step one in dimension d
  - $(X_i)_{i \in \mathbb{N}^*}$: *iid* with uniform distribution on $[-1, 1]^d$
- $\epsilon \in]-\frac{1}{2}, \frac{1}{2}[$ is the regularity parameter.
  - $\epsilon = 0 \longrightarrow$ regular grid.
  - $|\epsilon|$ close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

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



Randomly perturbed regular grid (2/2)

Histograms of the interpoint distances with $\epsilon = 0, \frac{1}{8}, \frac{3}{8}$



Consistency and asymptotic normality

Under general conditions

For ML

- a.s convergence of the random Fisher information : The random trace

$$\frac{1}{n} Tr \left(\mathbf{R}_{\theta_0}^{-1} \frac{\partial \mathbf{R}_{\theta_0}}{\partial \theta_i} \mathbf{R}_{\theta_0}^{-1} \frac{\partial \mathbf{R}_{\theta_0}}{\partial \theta_j} \right)$$

  converges a.s to the element $(\mathbf{I}_{ML})_{i,j}$ of a $p \times p$ deterministic matrix $\mathbf{I}_{ML}$ as $n \to +\infty$
- asymptotic normality : With $\mathbf{\Sigma}_{ML} = 2\mathbf{I}_{ML}^{-1}$

$$\sqrt{n} \left(\hat{\theta}_{ML} - \theta_0 \right) \to \mathcal{N}\left(0, \mathbf{\Sigma}_{ML}\right)$$

For CV

Same result with more complex random traces for asymptotic covariance matrix $\mathbf{\Sigma}_{CV}$

$\longrightarrow$ consistency and same rate of convergence for CV

Objectives for the analysis of the spatial sampling impact

The asymptotic covariance matrices $\mathbf{\Sigma}_{ML,CV}$ depend only on the regularity parameter ϵ.

$\boxed{\longrightarrow}$ in the sequel, we study the functions $\epsilon \to \mathbf{\Sigma}_{ML,CV}$

Small random perturbations of the regular grid

We study $\left(\frac{\partial^2}{\partial \epsilon^2} \mathbf{\Sigma}_{ML,CV} \right)_{\epsilon=0}$

- Closed form expression for ML for $d = 1$ using Toeplitz matrix sequence theory
- Otherwise, it is calculated by exchanging limit in n and derivatives in ϵ

Large random perturbations of the regular grid

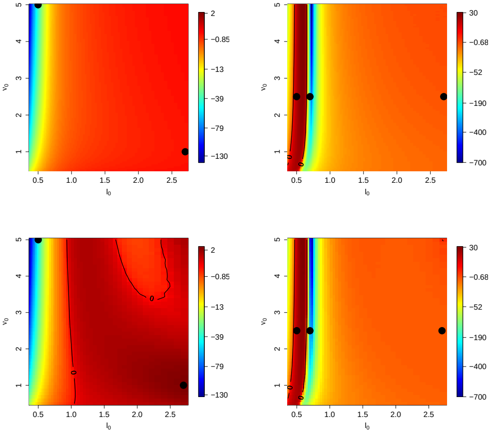
We study $\epsilon \to \mathbf{\Sigma}_{ML,CV}$

- Closed form expression for ML and CV for $d = 1$ and $\epsilon = 0$ using Toeplitz matrix sequence theory
- Otherwise, it is calculated by taking n large enough

Small random perturbations of the regular grid

Matèrn model. Dimension one. One estimated hyper-parameter.
Levels plot of $(\partial_\epsilon^2 \Sigma_{ML,CV})/\Sigma_{ML,CV}$ in $\ell_0 \times \nu_0$

Top : ML
Bot : CV
Left : $\hat{\ell}$ (ν_0 known)
Right : $\hat{\nu}$ (ℓ_0 known)

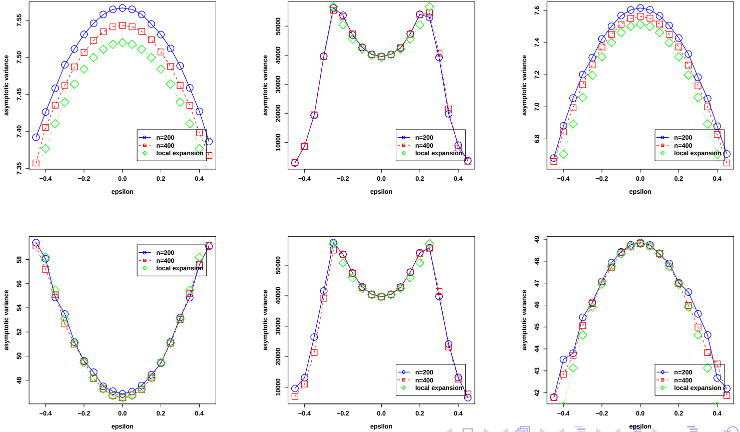


There exist cases of degradation of the estimation for small perturbation for ML and CV. Not easy to interpret

Large random perturbations of the regular grid

Plot of $\Sigma_{ML,CV}$. Top : ML. Bot : CV.

From left to right : $(\hat{\ell}, \ell_0 = 2.7, \nu_0 = 1)$, $(\hat{\nu}, \ell_0 = 0.5, \nu_0 = 2.5)$, $(\hat{\nu}, \ell_0 = 2.7, \nu_0 = 2.5)$



Conclusion on the well-specified case

- CV is consistent and has the same rate of convergence than ML
- We confirm that ML is more efficient
- Irregularity in the sampling is generally an advantage for the estimation, but not necessarily
  - With ML, irregular sampling is more often an advantage than with CV
  - Large perturbations of the regular grid are often better than small ones for estimation
  - Keep in mind that hyper-parameter estimation and Kriging prediction are strongly different criteria for a spatial sampling

For further details :

Bachoc F, Asymptotic analysis of the role of spatial sampling for hyper-parameter estimation of Gaussian processes, *Submitted, available at http ://arxiv.org/abs/1301.4321*.

Conclusion

General conclusion

- ML preferable to CV in the well-specified case
- In the misspecified-case, with not too regular design of experiments : CV preferable because of its smaller bias
- In both misspecified and well-specified cases : the estimation benefits from an irregular sampling
- The variance of CV is larger than that of ML in all the cases studied.

Perspectives

- Designing other CV procedures (LOO error ponderation, decorrelation and penalty term) to reduce the variance
- Expansion-domain asymptotic analysis of the misspecified case

Thank you for your attention !