Maximum Likelihood and Cross Validation for covariance hyper-parameter estimation of Gaussian processes

François Bachoc

PhD advisors:

Josselin Garnier
&
Jean-Marc Martinez

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

March 2013

The PhD

PhD on the subject of Kriging models ($\approx$ Gaussian process regression)

Two components of the PhD

- Work on the problem of the covariance function estimation
  - Bachoc F, Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification, *Computational Statistics and Data Analysis,* .
  - Bachoc F, Asymptotic analysis of the role of spatial sampling for hyper-parameter estimation of Gaussian processes, *Submitted*.
- Use of Kriging models for numerical model validation
  - Bachoc F, Bois G, Garnier J, and Martinez J.M, Calibration and improved prediction of computer models by universal Kriging, *Accepted in Nuclear Science and engineering*.

Kriging models

A Kriging model

The study of a single realization of a Gaussian process Y on a domain $\mathcal{X} \subset \mathbb{R}^d$

Objectives

Here, given n observations $Y(x_1), ..., Y(x_n)$, estimating, for a new point x_{new},

- the predictive mean $\mathbb{E}(Y(x_{new})|Y(x_1), ..., Y(x_n))$
- the predictive variance $var(Y(x_{new})|Y(x_1), ..., Y(x_n))$

The Gaussian process

- $\forall x_1, ..., x_n \in \mathcal{X}$, the vector $(Y(x_1), ..., Y(x_n))$ is Gaussian
- 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

Covariance function

Covariance function

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

Stationarity

We consider the covariance function K stationary : $K(x_1, x_2) = K(x_1 - x_2)$
In this case, Bochner's theorem holds : the Fourier transform $\hat{K}$ of K is non-negative
This is well interpreted because, for all, $x_1, ..., x_n \in \mathcal{X}$, $\alpha_1, ..., \alpha_n \in \mathbb{R}$:

$$0 \leq \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) = \int_{\mathbb{R}^d} \hat{K}(f) \left| \sum_{i=1}^{n} \alpha_i e^{(Jf^t x_i)} \right|^2 df$$

Smoothness of the covariance function

Smoothness in $\mathbb{R}$

The following equivalence (in $\mathbb{R}$) is very important in both theory and practice :
The Gaussian process Y is k times mean square differentiable $\Leftrightarrow$ The covariance function K is $2k$ times differentiable at zero $\Leftrightarrow$ The Fourier transform $\hat{K}$ verifies $\int_{\mathbb{R}} f^{2k} \hat{K}(f) < +\infty$

$\longrightarrow$ Motivation for the following Matérn model

Matérn model

Covariance function parameterized by the hyper-parameters $\phi > 0$, $\nu > 0$ and $\alpha > 0$ and defined by

$$\hat{K}(f) = \phi \frac{1}{(\alpha^2 + f^2)^{\frac{1}{2}+\nu}}$$

ν : smoothness hyper-parameter. $\nu > k \Leftrightarrow Y$ is k times mean square differentiable.

Parameterization of the covariance function in $\mathbb{R}$

Parameterization of the Matérn model

Alternative parameterization by $\sigma^2 > 0$, $l_c > 0$, $\nu > 0$:

$$K(x) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}x}{l_c}\right)^{\nu} K_{\nu}\left(\frac{2\sqrt{\nu}x}{l_c}\right)$$

Interpretation of the hyper-parameters

- σ^2 : $\sigma^2 = K(0)$ is the variance hyper-parameter $\rightarrow$ scale of the Gaussian Process
- l_c is the correlation length $\rightarrow$ scale of variation of the Gaussian Process
- ν is the smoothness hyper-parameter $\rightarrow$ smoothness of the realizations of the Gaussian Process

Particular cases

- $\nu = \frac{1}{2}$: exponential model

$$K(x) = e^{-\sqrt{2}\frac{|x|}{l_c}}$$

- $\nu = +\infty$: Gaussian model

$$K(x) = e^{-\frac{x^2}{l_c^2}}$$

Parameterization of the covariance function in $\mathbb{R}^d$

Multidimensional Matérn model

Parameterized by $\sigma^2 > 0$, $l_{c,1} > 0$, ...,$l_{c,d} > 0$, $\nu > 0$
Defined by, with

$$|x|_{l_c} = \sqrt{\sum_{i=1}^{d} \frac{x_i^2}{l_{c,i^2}}},$$

and $K_{m,1}$ the Matérn covariance function in dimension one,

$$K(x) = K_{m,1}(|x|_{l_c})$$

$\rightarrow$ $l_{c,i}$ is the i-th correlation length and is the scale of variation corresponding to the i-th component

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 hyper-parameter
- θ is the multidimensional correlation hyper-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 $\hat{\theta}(y)$

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

Remark : Taking systematically the uncertainty on the covariance function into account in the predictive variance is a subject of research, but is not (yet) classical in Kriging

Conclusion

- The covariance function characterizes the Gaussian process
- It is estimated first
- Then we can compute prediction and predictive variances with closed form matrix vector formulas

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 (Leave-One-Out)

Gaussian Process Y observed at $x_1, \ldots, x_n$ with values $y = (y_1, \ldots, y_n)^t$

Cross Validation (Leave-One-Out) principle

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

Let $\mathbf{R}$ be the covariance matrix of $y = (y_1, \ldots, y_n)$

Virtual Leave-One-Out

$$y_i - \hat{y}_{i,-i} = (diag(\mathbf{R}^{-}1))^{-1}\mathbf{R}^{-}y \quad \text{and} \quad c^2_{i,-i} = \frac{1}{(\mathbf{R}^{-})_{i,i}}$$

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

Cross Validation for covariance function estimation (1/3)

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

Leave-One-Out criteria we study

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

and

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

Cross Validation for covariance function estimation (2/3)

Using the virtual Cross Validation formula :

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

and

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

Cross Validation for covariance function estimation (3/3)

- Leave-One-Out estimation is tractable
- Other Cross-Validation criteria exist

C.E. Rasmussen and C.K.I. Williams, Gaussian Processes for Machine Learning, *The MIT Press, Cambridge*, 2006.

- To the best of our knowledge : problems of the choice of the cross validation criterion and of the cross validation procedure are not fully solved for Kriging
- It is our intuition that when one is primarily interested in point-wise predictive mean and variance, the Leave-One-Out criteria presented are reasonable

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 stationary correlation function, and K_1 is the true stationary unit-variance covariance function : Theoretical formula and numerical tests
- In the general case : Numerical studies

Setting for variance hyper-parameter estimation

Let

- r_1 be the covariance vector of Y between $x_1, ..., x_n$ and x_{new} with covariance function K_1
- r_2 be the covariance vector of Y between $x_1, ..., x_n$ and x_{new} with covariance function K_2
- $\mathbf{R}_1$ be the covariance matrix of Y at $x_1, ..., x_n$ with covariance function K_1
- $\mathbf{R}_2$ be the covariance matrix of Y at $x_1, ..., x_n$ with covariance function K_2

Then

- $\hat{Y}(x_{new}) = r_2^t \mathbf{R}_2^{-1} y$ is the Kriging prediction
- $\mathbb{E}\left[(\hat{Y}(x_{new}) - Y(x_{new}))^2 | y\right] = (r_1^t \mathbf{R}_1^{-1} y - r_2^t \mathbf{R}_2^{-1} y)^2 + 1 - r_1^t \mathbf{R}_1^{-1} r_1$ is the conditional mean square error of the non-optimal prediction
- One estimates σ^2 with $\hat{\sigma}^2$ and estimates the conditional mean square error with $\hat{\sigma}^2 c_{x_{new}}^2$ with $c_{x_{new}}^2 := 1 - r_2^t \mathbf{R}_2^{-1} r_2$

The Risk

The Risk

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

$$\mathcal{R}_{\hat{\sigma}^2, x_{new}} = \mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c_{x_{new}}^2\right)^2\right]$$

Formula for quadratic estimators

When $\hat{\sigma}^2 = y^t \mathbf{M} y$, we have

$$\begin{aligned} \mathcal{R}_{\hat{\sigma}^2, x_{new}} &= f(\mathbf{M}_0, \mathbf{M}_0) + 2c_1 tr(\mathbf{M}_0) - 2c_2 f(\mathbf{M}_0, \mathbf{M}_1) \\ &\quad + c_1^2 - 2c_1 c_2 tr(\mathbf{M}_1) + c_2^2 f(\mathbf{M}_1, \mathbf{M}_1) \end{aligned}$$

with

$$f(\mathbf{A}, \mathbf{B}) = tr(\mathbf{A}) tr(\mathbf{B}) + 2tr(\mathbf{AB})$$

$$\mathbf{M}_0 = (\mathbf{R}_2^{-1} r_2 - \mathbf{R}_1^{-1} r_1)(r_2^t \mathbf{R}_2^{-1} - r_1^t \mathbf{R}_1^{-1}) \mathbf{R}_1$$

$$\mathbf{M}_1 = \mathbf{M}\mathbf{R}_1$$

$$c_1 = 1 - r_1^t \mathbf{R}_1^{-1} r_1$$

$$c_2 = 1 - r_2^t \mathbf{R}_2^{-1} r_2$$

CV and ML estimation

CV and ML estimation

$$\hat{\sigma}^2_{ML} = \frac{1}{n} y^t \mathbf{R}_2^{-1} y$$

$$\hat{\sigma}^2_{CV} = \frac{1}{n} y^t \mathbf{R}_2^{-1} \left[diag(\mathbf{R}_2^{-1}) \right]^{-1} \mathbf{R}_2^{-1} y$$

Well-specified case : Risk $\sim$ estimation Mean Square Error for σ^2

- ML estimation : $var(\hat{\sigma}^2_{ML})$ is the Cramer-Rao bound $\frac{2}{n}$
- CV estimation : $var(\hat{\sigma}^2_{CV})$ can reach 2

$\longrightarrow$ When $K_2 = K_1$, ML is best. Numerical study when $K_2 \neq K_1$

Criteria for numerical studies (1/2)

Risk on Target Ratio (RTR),

$$RTR(x_{new}) = \frac{\sqrt{\mathcal{R}_{\hat{\sigma}^2, x_{new}}}}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]} = \frac{\sqrt{\mathbb{E}\left[\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c^2_{x_{new}}\right)^2\right]}}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]}$$

Bias-variance decomposition

$$\mathcal{R}_{\hat{\sigma}^2, x_{new}} = \Bigg(\underbrace{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right] - \mathbb{E}\left(\hat{\sigma}^2 c^2_{x_{new}}\right)}_{\text{bias}}\Bigg)^2 + \underbrace{var\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2 | y\right] - \hat{\sigma}^2 c^2_{x_{new}}\right)}_{\text{variance}}$$

Bias on Target Ratio (BTR) criterion

$$BTR(x_{new}) = \frac{\left|\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right] - \mathbb{E}\left(\hat{\sigma}^2 c^2_{x_{new}}\right)\right|}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]}$$

Criteria for numerical studies (2/2)

$$\Bigg(\underbrace{RTR}_{\text{relative error}}\Bigg)^2 = \Bigg(\underbrace{BTR}_{\text{relative bias}}\Bigg)^2 + \underbrace{\frac{var\left(\mathbb{E}\left[(\hat{y}_0 - Y_0)^2|y\right] - \hat{\sigma}^2 c^2_{x_{new}}\right)}{\mathbb{E}\left[(\hat{Y}_0 - Y_0)^2\right]^2}}_{\text{relative variance}}$$

Integrated criteria on the prediction domain $\mathcal{X}$

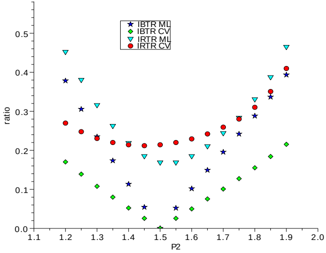
$$IRTR = \sqrt{\int_{\mathcal{X}} RTR^2(x_{new})d\mu(x_{new})}$$

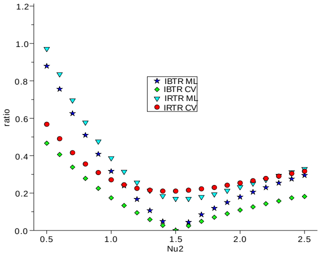
and

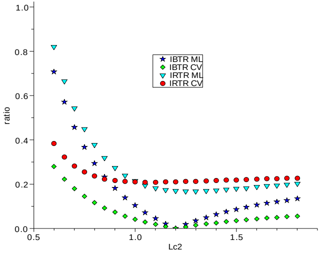
$$IBTR = \sqrt{\int_{\mathcal{X}} BTR^2(x_{new})d\mu(x_{new})}$$

Numerical results

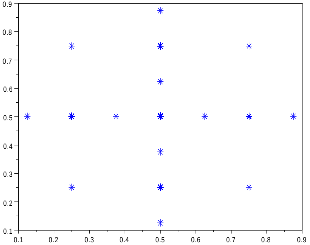
70 observations on $[0,1]^5$. Mean over LHS-Maximin DoE's.
Top : K_1 and K_2 are power-exponential, with $l_{c,1} = l_{c,2} = 1.2$, $p_1 = 1.5$, and p_2 varying.
Bot left : K_1 and K_2 are Matérn, with $l_{c,1} = l_{c,2} = 1.2$, $\nu_1 = 1.5$, and ν_2 varying.
Bot right : K_1 and K_2 are Matérn ($\nu = \frac{3}{2}$), with $l_{c,1} = 1.2$, and $l_{c,2}$ varying.

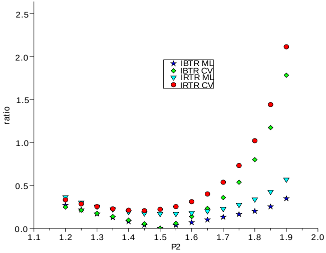


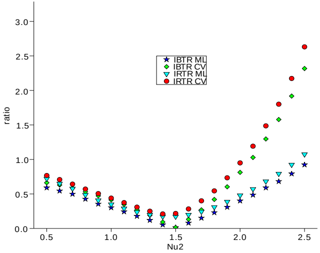


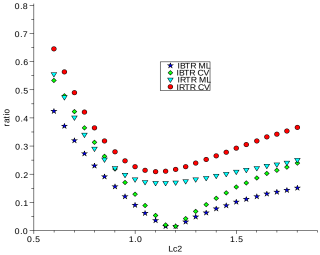


Case of a regular grid (Smolyak construction)









Work on analytical functions

Consider a deterministic function f on $[0,1]^d$

- Ishigami function :

  $$f(x_1,x_2,x_3) = \sin(-\pi+2\pi x_1)+7\sin((-\pi+2\pi x_2))^2+0.1\sin(-\pi+2\pi x_1).(-\pi+2\pi x_3)^4$$

- Morris function :

$$f(x) = \sum_{i=1}^{10} w_i(x) + \sum_{1\le i<j\le 6} w_i(x)w_j(x) + \sum_{1\le i<j<k\le 5} w_i(x)w_j(x)w_k(x) + \sum_{1\le i<j<k<l\le 4} w_i(x)w_j(x)w_k(x)w_l(x),$$

with $$w_i(x) = \begin{cases} 2\left(\frac{1.1x_i}{x_i+0.1}-0.5\right), & \text{if } i=3,5,7 \\ 2(x_i-0.5) & \text{otherwise} \end{cases}$$

Comparison criteria

Learning sample $y_{a,1}, \ldots, y_{a,n}$. Test sample $y_{t,1}, \ldots, y_{t,n_t}$

Mean Square Error (MSE) criterion :

$$MSE = \frac{1}{n_t} \sum_{i=1}^{n_t} (y_{t,i} - \hat{y}_{t,i}(y_a))^2$$

Predictive Variance Adequation (PVA) criterion :

$$PVA = \left| \log \left(\frac{1}{n_t} \sum_{i=1}^{n_t} \frac{(y_{t,i} - \hat{y}_{t,i}(y_a))^2}{\hat{\sigma}^2 c_{t,i}^2(y_a)} \right) \right|$$

We average MSE and PVA over $n_p = 100$ LHS Maximin DoE's. For each DoE : covariance estimation and Kriging prediction

Results with enforced correlation

We use (tensorized) Exponential and Gaussian correlation functions for the Ishigami function

Correlation model	Enforced hyper-parameters	MSE	PVA
Exponential	$[1, 1, 1]$	2.01	ML : 0.50 CV : 0.20
Exponential	$[1.3, 1.3, 1.3]$	1.94	ML : 0.46 CV : 0.23
Exponential	$[1.20, 5.03, 2.60]$	1.70	ML : 0.54 CV : 0.19
Gaussian	$[0.5, 0.5, 0.5]$	4.19	ML : 0.98 CV : 0.35
Gaussian	$[0.31, 0.31, 0.31]$	2.03	ML : 0.16 CV : 0.23
Gaussian	$[0.38, 0.32, 0.42]$	1.32	ML : 0.28 CV : 0.29

- Misspecified cases : Exponential and Gaussian isotropic
- ML have the highest PVA in the worst misspecification cases

Setting for estimated correlation

- Work on three correlation families
  - Exponential (tensorized)
  - Gaussian
  - Matérn with estimated regularity hyper-parameter
- Work in the isotropic and anisotropic case
  - Case2.i : A common correlation length is estimated
  - Case2.a : d different correlation lengths are estimated

Results for estimated correlation : Ishigami

Function	Correlation model	MSE	PVA
Ishigami	exponential case 2.i	ML : 1.99 CV : 1.97	ML : 0.35 CV : 0.23
Ishigami	exponential case 2.a	ML : 2.01 CV : 1.77	ML : 0.36 CV : 0.24
Ishigami	Gaussian case 2.i	ML : 2.06 CV : 2.11	ML : 0.18 CV : 0.22
Ishigami	Gaussian case 2.a	ML : 1.50 CV : 1.53	ML : 0.53 CV : 0.50
Ishigami	Matérn case 2.i	ML : 2.19 CV : 2.29	ML : 0.18 CV : 0.23
Ishigami	Matérn case 2.a	ML : 1.69 CV : 1.67	ML : 0.38 CV : 0.41

- Gaussian and Matérn are more adapted than exponential because of smoothness ($\rightarrow$ smaller MSE)
- Estimating several correlation lengths is more adapted
- In the exponential case, CV has smaller PVA and smaller or equal MSE
- In the Gaussian and Matérn cases, ML has MSE and PVA slightly smaller

Results for estimated correlation : Morris

Function	Correlation model	MSE	PVA
Morris	exponential case 2.i	ML : 3.07 CV : 2.99	ML : 0.31 CV : 0.24
Morris	exponential case 2.a	ML : 2.03 CV : 1.99	ML : 0.29 CV : 0.21
Morris	Gaussian case 2.i	ML : 1.33 CV : 1.36	ML : 0.26 CV : 0.26
Morris	Gaussian case 2.a	ML : 0.86 CV : 1.21	ML : 0.79 CV : 1.56
Morris	Matérn case 2.i	ML : 1.26 CV : 1.28	ML : 0.24 CV : 0.25
Morris	Matérn case 2.a	ML : 0.75 CV : 1.06	ML : 0.65 CV : 1.43

- Gaussian and Matérn are more adapted than exponential because of smoothness ($\rightarrow$ smaller MSE)
- Estimating several correlation lengths is more adapted
- In the Exponential case, CV has slightly smaller MSE and smaller PVA
- For Gaussian and Matérn 2.a, ML has smaller MSE and PVA
- For Gaussian and Matérn, going from 2.a to 2.i causes much more harm to ML than CV

Conclusion

Conclusion

- We study robustness relatively to prediction mean square errors and point-wise mean square error estimation
- For the variance estimation, CV is more robust than ML to correlation function misspecification
- This is not true for the Smolyak construction we tested
- In the general case of correlation function estimation $\rightarrow$ this is globally confirmed in a case study on analytical functions

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 function for the estimation.

Well-specified model

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

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

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 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 covariance 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 ?

Asymptotics for hyper-parameters 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
  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 : A minimum spacing exists between the observations $\longrightarrow$ infinite observation domain.
  Asymptotic normality proved for Maximum-Likelihood under general conditions

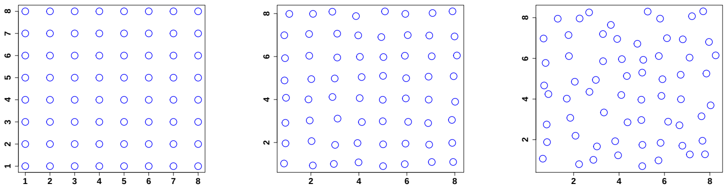
  Sweeting, T., Uniform asymptotic normality of the maximum likelihood estimator, *Annals of Statistics 8 (1980) 1375-1381*.

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

Randomly perturbed regular grid

- Our sampling model : regular square grid of step one in dimension d, $(v_i)_{i\in\mathbb{N}^*}$. The observation points are the $v_i + \epsilon X_i$. The $(X_i)_{i\in\mathbb{N}^*}$ are *iid* and uniform 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 $\epsilon = 0, \frac{1}{8}, \frac{3}{8}$



Main assumptions (1/2)

Control of the derivatives

$$K_\theta(t) \leq \frac{C}{1+|t|^{d+1}},$$

$$\forall i, \frac{\partial}{\partial \theta_i} K_\theta(t) \leq \frac{C}{1+|t|^{d+1}},$$

$$\forall i,j, \frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j} K_\theta(t) \leq \frac{C}{1+|t|^{d+1}},$$

$$\forall i,j,k, \frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j}\frac{\partial}{\partial \theta_k} K_\theta(t) \leq \frac{C}{1+|t|^{d+1}},$$

Positive continuous Fourier transform

- K_θ has a Fourier transform $\hat{K}_\theta$
- $(\theta, f) \to \hat{K}_\theta(f)$ is strictly-positive on $\Theta \times \mathbb{R}^d$.

Main assumptions (2/2)

Set of interpoint spacings explored by the sampling

$D_\epsilon := \cup_{v \in \mathcal{Z}^d \backslash 0} \left(v + [-2\epsilon, 2\epsilon]^d \right)$

Identifiability

- Global
  - For $\epsilon = 0$, there does not exist $\theta \neq \theta_0$ so that $K_\theta(v) = K_{\theta_0}(v)$ for all $v \in \mathcal{Z}^d$
  - For $\epsilon \neq 0$ there does not exist $\theta \neq \theta_0$ so that $K_\theta = K_{\theta_0}$ a.s. on D_ϵ, and $K_\theta(0) = K_{\theta_0}(0)$
- Local
  - For $\epsilon = 0$, there does not exist $v_\lambda = (\lambda_1, \ldots, \lambda_p) \neq 0$ so that $\sum_{k=1}^p \lambda_k \frac{\partial}{\partial \theta_k} K_{\theta_0}(v) = 0$ for all $v \in \mathcal{Z}^d$
  - For $\epsilon \neq 0$ there does not exist $v_\lambda = (\lambda_1, \ldots, \lambda_p) \neq 0$ so that $\sum_{k=1}^p \lambda_k \frac{\partial}{\partial \theta_k} K_{\theta_0} = 0$ a.s. on D_ϵ, and $\sum_{k=1}^p \lambda_k \frac{\partial}{\partial \theta_k} K_{\theta_0}(0) = 0$

Correlation function family (only for Cross Validation)

$\forall \theta \in \Theta, K_\theta(0) = 1$

Assumptions verified by all classical stationary covariance function families

Consistency and asymptotic normality

For ML

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

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

For CV

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

Notations

- X : the random vector $(X_1, ..., X_n)$ of the perturbations
- x : A vector of $([-1, 1]^d)^n$, as a realization of X
- y : The random vector $(Y(X_1), ..., Y(X_n))$
- $\mathbf{R}_\theta := cov_\theta(y|X)$: The random covariance matrix
- $L_\theta := \frac{1}{n}\left(\ln\left(|\mathbf{R}_\theta|\right) + y^t\mathbf{R}_\theta^{-1}y\right)$: the ML criterion (to minimize)
- $CV_\theta := \frac{1}{n}y^t\mathbf{R}_\theta^{-1}diag(\mathbf{R}_\theta^{-1})^{-2}\mathbf{R}_\theta^{-1}y$: the CV criterion (to minimize)

Results on random matrices (1/)

Control of the eigenvalues

- The eigenvalues of $\mathbf{R}_\theta$, $\frac{\partial}{\partial \theta_i}\mathbf{R}_\theta$, $\frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j}\mathbf{R}_\theta$ and $\frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j}\frac{\partial}{\partial \theta_k}\mathbf{R}_\theta$ are upper-bounded uniformly in n, x, θ.
  - Because, e.g, $\sum_{j \in \mathbb{N}^*, j \neq i} K_\theta \left\{ v_i - v_j + \epsilon \left(x_i - x_j \right) \right\}$ is bounded uniformly in x, θ
- The eigenvalues of $\mathbf{R}_\theta$ are lower-bounded uniformly in n, x, θ.
  - Comes from

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j K_\theta (v_i + x_i, v_j + x_j) = \int_{\mathbb{R}^d} \hat{K}_\theta (f) \left| \sum_{i=1}^{n} \alpha_i e^{\left(J f^t (v_i + x_i) \right)} \right|^2 df$$

Results on random matrices (2/)

Class of matrix involved in the ML and CV criteria

Let a matrix sequence $\mathbf{M}$, whose expression uses $\mathbf{R}_\theta$, $\mathbf{R}_\theta^{-1}$, $\frac{\partial}{\partial \theta_i}\mathbf{R}_\theta$, $\frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j}\mathbf{R}_\theta$, $\frac{\partial}{\partial \theta_i}\frac{\partial}{\partial \theta_j}\frac{\partial}{\partial \theta_k}\mathbf{R}_\theta$, the matrix product, the diag operator and the matrix $diag(\mathbf{R}_\theta^{-1})^{-1}$.

e.g, $\mathbf{M} = \mathbf{R}_\theta^{-1}\frac{\partial}{\partial \theta_i}\mathbf{R}_\theta\mathbf{R}_\theta^{-1}\frac{\partial}{\partial \theta_j}\mathbf{R}_\theta$.

e.g, $\mathbf{M} = \mathbf{R}_\theta^{-1} diag(\mathbf{R}_\theta^{-1})^{-2}\mathbf{R}_\theta^{-1}$

Control of eigenvalues

The matrices $\mathbf{M}$ above have their eigenvalues upper-bounded uniformly in n, x, θ.

Results on random matrices (3/)

Almost sure convergence of random traces

$\frac{1}{n}Tr(\mathbf{M})$ converges a.s. to a deterministic limit S.

Sketch of proof

- Make the approximation that the Gaussian process Y is composed of a partition of independent Gaussian processes
- This boils down to approximating $\mathbf{M}$ of size $n \approx n_1 n_2$ by

$$\mathbf{M} \approx \mathbf{M}_{n_1,n_2} := \begin{pmatrix} \mathbf{M}_{n_1}^{(1)} & & & \\ & \mathbf{M}_{n_1}^{(2)} & & \\ & & \ddots & \\ & & & \mathbf{M}_{n_1}^{(n_2)} \end{pmatrix}$$

- $|\frac{1}{n}Tr(\mathbf{M}) - \frac{1}{n}Tr(\mathbf{M}_{n_1,n_2})| \to_{n_1,n_2 \to +\infty} 0$
- The $\mathbf{M}_{n_1}^{(i)}$ are *iid* so that $\frac{1}{n}Tr(\mathbf{M}_{n_1,n_2}) - \mathbb{E}\left(\frac{1}{n}Tr(\mathbf{M}_{n_1}^{(1)})\right) \to_{n_2 \to +\infty} 0$
- Conclude by letting $n_1, n_2 \to +\infty$ and by using the Cauchy criterion

Results on random matrices (4/)

Convergence of random quadratic forms

$\frac{1}{n} y^t \mathbf{M} y$ converges in mean square to $\frac{1}{n} Tr(\mathbf{M}\mathbf{R}_{\theta_0})$

Asymptotic normality of random quadratic forms

When $Tr(\mathbf{M}) = 0$, let S be the almost sure limit of $\frac{1}{n} Tr(\mathbf{M}\mathbf{R}_{\theta_0}\mathbf{M}\mathbf{R}_{\theta_0})$
Then $\frac{1}{\sqrt{n}} y^t \mathbf{M} y$ converges in law to a $\mathcal{N}(0, 2S)$

Sketch of proof

Let $\mathcal{L}(z_i|X) =_{iid} \mathcal{N}(0,1)$. Then

$$\frac{1}{\sqrt{n}} y^t \mathbf{M} y = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_i(\mathbf{M}\mathbf{R}_{\theta_0}) z_i^2$$

We then use an almost sure (with respect to X) Lindeberg-Feller criterion.

Asymptotic normality

- After having proved consistency and that the almost sure limit of $\frac{\partial^2}{\partial^2\theta}L_{\theta_0}$ is a strictly-positive matrix
- Using the results of random matrices above, we directly show that

$$\frac{\partial}{\partial\theta}L_{\theta_0} \to_{\mathcal{L}} \mathcal{N}(0, 2\mathbf{I}_{ML})$$

  and

$$\frac{\partial^2}{\partial^2\theta}L_{\theta_0} \to_p \mathbf{I}_{ML}$$

- We conclude using standard M-estimator techniques.
- Same method for CV

Consistency

Consistency

There exists $A > 0$ so that, uniformly in n, X, θ

$$\mathbb{E}(L_\theta - L_{\theta_0}|X) \geq A \sum_{i \in \mathbb{N}*} |K_\theta(v_i + X_i) - K_{\theta_0}(v_i + X_i)|^2$$

and $\sum_{i \in \mathbb{N}*} |K_\theta(v_i + X_i) - K_{\theta_0}(v_i + X_i)|^2$ converges in probability to

- $\sum_{v \in \mathcal{Z}^d} |K_\theta(v) - K_{\theta_0}(v)|^2$ if $\epsilon = 0$
- $\int_{D_\epsilon} f_T(t)|K_\theta(t) - K_{\theta_0}(t)|^2 + |K_\theta(0) - K_{\theta_0}(0)|^2$ if $\epsilon \neq 0$

(with f_T the triangular pdf on $[-2\epsilon, 2\epsilon]^d$)
We conclude with the identifiability assumption.
Same method for CV

Strictly-positive second derivative

Strictly-positive second derivative (with $d = 1$)

There exist $A > 0$ so that, uniformly in n, X, θ

$$\mathbb{E}(\frac{\partial^2}{\partial \theta^2} L_{\theta_0} | X) \geq A \sum_{i \in \mathbb{N}*} |\frac{\partial}{\partial \theta} K_{\theta_0}(v_i + X_i)|^2$$

and $\sum_{i \in \mathbb{N}*} |\frac{\partial}{\partial \theta} K_{\theta_0}(v_i + X_i)|^2$ converges in probability to

- $\sum_{v \in \mathcal{Z}^d} |\frac{\partial}{\partial \theta} K_{\theta_0}(v)|^2$ if $\epsilon = 0$
- $\int_{D_\epsilon} f_T(t) |\frac{\partial}{\partial \theta} K_{\theta_0}(t)|^2 + |\frac{\partial}{\partial \theta} K_{\theta_0}(0)|^2$ if $\epsilon \neq 0$

(with f_T the triangular pdf on $[-2\epsilon, 2\epsilon]^d$)
We conclude with the identifiability assumption.
Same method for CV

Generalization to $d > 1$

Consider the covariance function family $\left\{ K_{(\theta_0)_1 + \delta\lambda_1, \ldots, (\theta_0)_p + \delta\lambda_p}, \delta_{inf} \leq \delta \leq \delta_{sup} \right\}$

Objectives

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

Small random perturbations of the regular grid

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

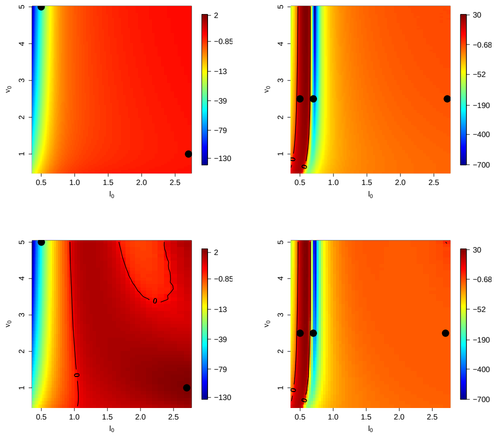
Large random perturbations of the regular grid

We study $\epsilon \longrightarrow \mathbf{V}_{ML,CV}$
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 $(\partial_{\epsilon}^{2} \mathbf{V}_{ML,CV}) / \mathbf{V}_{ML,CV}$ in $\ell_0 \times \nu_0$

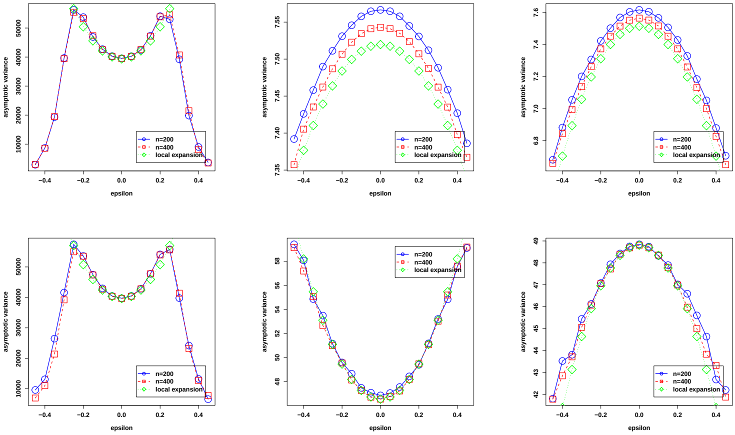
Top : ML
Bot : CV
Left : $\hat{\ell}$ (ν_0 known)
Right : $\hat{\nu}$ (ℓ_0 known)



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 $\mathbf{V}_{ML,CV}$. Top : ML. Bot : CV.
From left to right : $(\hat{\nu}, \ell_0 = 0.5, \nu_0 = 2.5)$, $(\hat{\ell}, \ell_0 = 2.7, \nu_0 = 1)$, $(\hat{\nu}, \ell_0 = 2.7, \nu_0 = 2.5)$



Conclusion

Conclusion

- Consistency and asymptotic normality for ML and CV. Same rate of convergence
- CV has the largest asymptotic variance
- 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