Estimation de la fonction de covariance dans le modèle de Krigeage et applications: bilan de thèse et perspectives

François Bachoc
Josselin Garnier
Jean-Marc Martinez

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

Juin 2013

The PhD

PhD started in October 2010

Two components in the PhD

- Use of Kriging model for code validation and metamodeling
- Work on the problem of the covariance function estimation

Numerical code and physical system

A numerical code, or parametric numerical model, is represented by a function f :

$$\begin{array}{lll} f & : \mathbb{R}^d \times \mathbb{R}^m & \rightarrow \mathbb{R} \\ & (x, \beta) & \rightarrow f(x, \beta) \end{array}$$

Observations can be made of a physical system Y_{real}

$$x_i \quad \rightarrow \boxed{Y_{real}} \rightarrow y_i$$

- The inputs x are the experimental conditions
- The inputs β are the calibration parameters of the numerical code
- The outputs $f(x_i, \beta)$ and y_i are the variable of interest

A numerical code modelizes (gives an approximation of) a physical system

Statistical model (1/2)

Statistical model followed in the phd

- The physical system Y_{real} is a deterministic function
- There exist a correct parameter β and a model error function Z so that

$$Y_{real}(x) = f(x, \beta) + Z(x)$$

- The observations are noised $Y_i = Y_{real}(x_i) + \epsilon_i$, where ϵ_i are *i.i.d* centered Gaussian variables

The model is different from other classical models for inverse problems where

- $Y_i = f(x_i, \beta_i) + \epsilon_i$. The β_i are *i.i.d* realizations of the random vector β
- Hence the physical system is random

Fu S, An adaptive kriging method for characterizing uncertainty in inverse problems, *Journée du GdR MASCOT-NUM - 23 mars 2011*.

de Crécy A, A methodology to quantify the uncertainty of the physical models of a code, *Rapport CEA DEN/DANS/DM2S/STMF*.

Statistical model (2/2)

- Z is modeled as the realization of a centered Gaussian process
- A Bayesian modelling is possible for the correct parameter β
  - no prior information case. β is constant and unknown
  - prior information case. β is the realization of a Gaussian vector $\mathcal{N}(\beta_{prior}, Q_{prior})$
- Linear approximation of the code w.r.t β

$$f(x, \beta) = \sum_{j=1}^{m} h_j(x)\beta_j$$

Eventually, the statistical model is a Kriging model

$$\boxed{Y_i = \sum_{j=1}^{m} h_j(x_i)\beta_j + Z(x_i) + \epsilon_i}$$

Hence calibration and prediction are carried out within the Kriging framework

Results with the thermal-hydraulic code Flica IV (1/2)

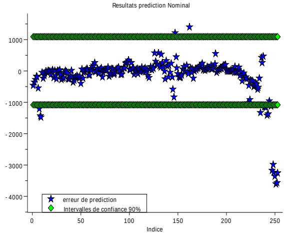
The experiment consists in pressurized and possibly heated water passing through a cylinder. We measure the pressure drop between the two ends of the cylinder.
Quantity of interest : The part of the pressure drop due to friction : ΔP_{fro}
Two kinds of experimental conditions :

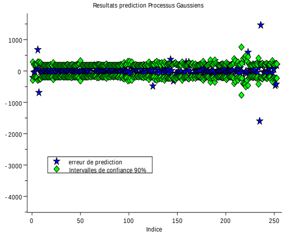
- System parameters : Hydraulic diameter D_h, Friction height H_f, Channel width e
- Environment variables : Output pressure P_s, Flowrate G_e, Parietal heat flux Φ_p, Liquid enthalpy h_e^l, Thermodynamic title X_{th}^e, Input temperature T_e

We dispose of 253 experimental results
Important : Among the 253 experimental results, only 8 different system parameters $\rightarrow$ Not enough to use the Gaussian processes model for prediction for new system parameters $\rightarrow$ We predict for new environment variables only

Results with the thermal-hydraulic code Flica IV (2/2)

	RMSE	90% Confidence Intervals
Nominal code	$661\,Pa$	$234/253 \approx 0.925$
Gaussian Processes	$189\,Pa$	$235/253 \approx 0.93$





Influence of the linear approximation

- The Gaussian process model of the model error is also tractable in the non-linear case

  M. J. Bayarri, J.O. Berger, R. Paulo, J. Sacks, J.A. Cafeo, J. Cavendish, C.H. Lin and J. Tu A framework for validation of computer models, *Technometrics, 49 (2), 138-154*.

- On the FLICA 4 data, we compare the linear approximation we use with the Bayes formula in the non-linear case for calibration and prediction.
  - Integrals are evaluated on a 5×5 grid in the calibration parameter space
  - The same grid is used for the linear-case
- We obtain
  - a 10% difference for calibration
  - a 1% difference for prediction
- In this case, the model error takes the linearization error into account

Conclusion on code Validation

- We can improve the prediction capability of the code by completing the physical representation with a statistical model
- Number of experimental results needs to be sufficient. No extrapolation
- More influence of the non-linearity of the code for calibration than for prediction

For more details

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, http ://arxiv.org/abs/1301.4114v2*

GERMINAL code meta-modeling (1/3)

Context

- GERMINAL code : simulation of the thermal-mechanical impact of the irradiation on a nuclear fuel pin
- Its utilization is part of a multi-physics and multi-objective optimization problem from reactor core design
- PhD thesis of Karim Ammar (CEA, DEN)

Characteristics for Kriging

- 12 inputs
- 2 outputs : a simple and a complicated relationship
- Numerical instabilities $\longrightarrow$ nugget effect to estimate
- Large data bases (thousands)

GERMINAL code meta-modeling (2/3)

Simple Kriging model, Matérn ($\frac{3}{2}$) covariance function with nugget effect, estimation by Maximum Likelihood

Summary of the results

- For estimation : the correlation lengths and the nugget effect make sense
- For prediction : good prediction results compared to neural network methods (+ predictive variances).
- Virtual Leave-One-Out standardized errors for detection of outlier GERMINAL calculations : confirmed by a new version of GERMINAL
- The size of the data base is a computational problem for Kriging estimation and prediction

GERMINAL code meta-modeling (3/3)

From the point of view of GERMINAL users

- Kriging gives good prediction (RMSE) compared to neural networks
- However, the cost of a call to the metamodel function is significantly more costly ($0(n)$)
  - We have not investigated Kriging methods dedicated to large data bases
- The predictive variances are considered a "plus" of Kriging and the Leave-One-Out for outlier detection has been appreciated

For more details

- In the PhD manuscript (planned for September)
- Karim AMMAR, Edouard HOURCADE, Cyril PATRICOT, François BACHOC and Jean-Marc MARTINEZ Improvement of supercomputing based core design process with parallel estimations and statistical analysis, *Submitted proceeding to the Joint International Conference on Supercomputing in Nuclear Applications and Monte Carlo 2013, Paris, 27-31 October, 2013*

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

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

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

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

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

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

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

Using the virtual Cross Validation formula :

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

and

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

Kriging for code validation and metamodeling

Statistical model and method for code validation

Application to the FLICA 4 thermal-hydraulic code

GERMINAL code meta-modeling

Maximum Likelihood and Cross Validation for hyper-parameter estimation

Finite sample analysis of ML and CV under model misspecification

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

Asymptotic framework

Consistency and asymptotic normality

Sketch of proof

Analysis of the asymptotic variance

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

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

The Risk

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

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

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

Summary of numerical results

For variance hyper-parameter estimation

- We make the distance between K_1 and K_2 vary, starting from 0
- 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

For variance and correlation hyper-parameter estimation

- Numerical study on analytical functions
- Confirmation of the results of the variance estimation case

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.*

- Kriging for code validation and metamodeling
  - Statistical model and method for code validation
  - Application to the FLICA 4 thermal-hydraulic code
  - GERMINAL code meta-modeling
- Maximum Likelihood and Cross Validation for hyper-parameter estimation
- Finite sample analysis of ML and CV under model misspecification
- **Asymptotic analysis of ML and CV in the well-specified case**
  - Asymptotic framework
  - Consistency and asymptotic normality
  - Sketch of proof
  - Analysis of the asymptotic variance
- 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, 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 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 observation points $\longrightarrow$ infinite observation domain.
  Asymptotic normality proved for Maximum-Likelihood under general conditions

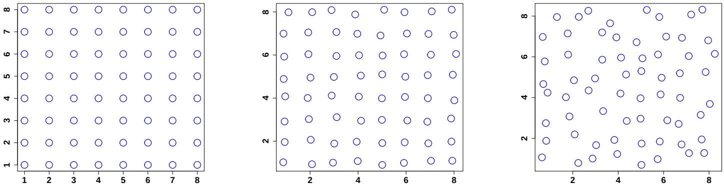
  Sweeting, T., Uniform asymptotic normality of the maximum likelihood estimator, *Annals of Statistics 8 (1980) 1375-1381*.

  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

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

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

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

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

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 $v_\lambda = (\lambda_1, ..., \lambda_p) \neq 0$ 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 $v_\lambda = (\lambda_1, ..., \lambda_p) \neq 0$ 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

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

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

- $\sum_{v\in\mathbb{Z}^d} |K_\theta(v) - K_{\theta_0}(v)|^2$ 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

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

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

- $\sum_{v \in \mathbb{Z}^d} |\frac{\partial}{\partial\theta} K_{\theta_0}(v)|^2$ 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 $\epsilon \rightarrow \mathbf{V}_{ML,CV}$

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 $\epsilon \rightarrow \mathbf{V}_{ML,CV}$

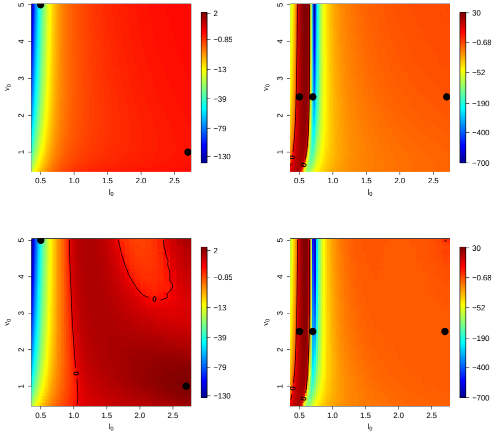
Closed form expression for ML and CV for $d = 1$ and $\epsilon = 0$ using Toeplitz matrix sequence theory

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 $(\partial_\epsilon^2 \Sigma_{ML,CV})/\Sigma_{ML,CV}$ in $\ell_0 \times \nu_0$

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

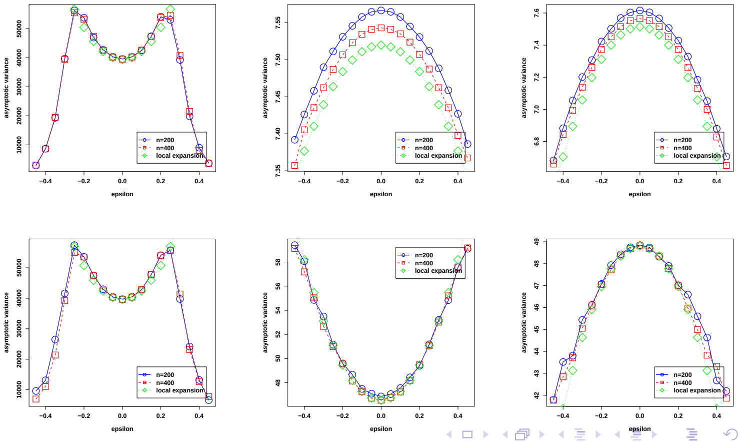


There exist cases of degradation of the estimation for small perturbation for ML and CV. Not easy to interpret

Large random perturbations of the regular grid

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

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



Conclusion 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 on covariance function estimation

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.

Thank you for your attention !