Maximum Likelihood and Cross Validation for covariance function estimation in Gaussian process regression

François Bachoc

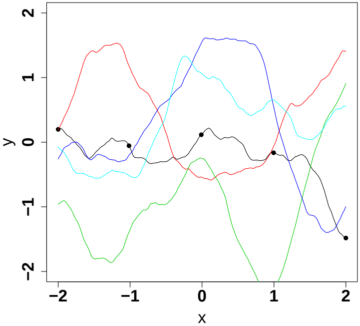
former PhD advisor: Josselin Garnier
former PhD co-advisor: Jean-Marc Martinez

Department of Statistics and Operations Research, University of Vienna

Zurich - November 2014

Gaussian process regression (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**

The Gaussian process

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

The covariance function

- The function $K_1 : \mathcal{X}^2 \to \mathbb{R}$, defined by $K_1(x_1, x_2) = cov(Y(x_1), Y(x_2))$

In most classical cases :

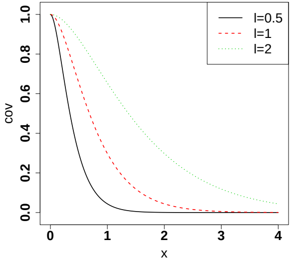
- Stationarity : $K_1(x_1, x_2) = K_1(x_1 - x_2)$
- Continuity : $K_1(x)$ is continuous $\Rightarrow$ Gaussian process realizations are continuous
- Decrease : $K_1(x)$ decreases with $||x||$ and $\lim_{||x|| \to +\infty} K_1(x) = 0$

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

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

It is defined as

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



Interpretation

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

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

Covariance function estimation

Parameterization

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

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

Observations

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

Estimation

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

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 parameters have been estimated and fixed to $\hat{\sigma}$ and $\hat{\theta}$

- $\mathbf{R}_{\hat{\theta}}$ is the correlation matrix of Y at $x_1, ..., x_n$ under correlation function $K_{\hat{\theta}}$
- $r_{\hat{\theta}}$ is the correlation vector of Y between $x_1, ..., x_n$ and x_{new} under correlation function $K_{\hat{\theta}}$

Prediction

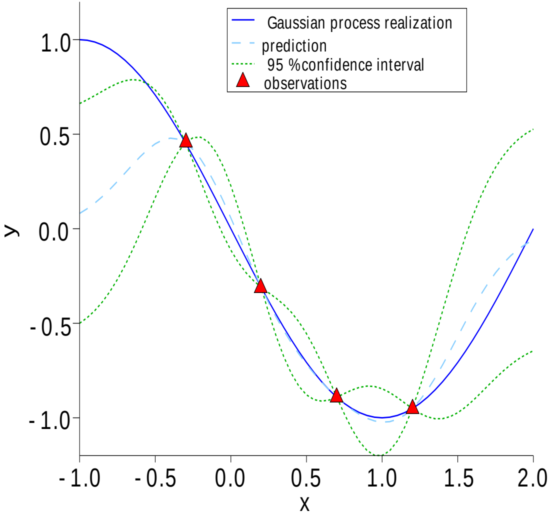
The prediction is $\hat{Y}_{\hat{\theta}}(x_{new}) := \mathbb{E}_{\hat{\theta}}(Y(x_{new})|Y(x_1), ..., Y(x_n)) = r_{\hat{\theta}}^t \mathbf{R}_{\hat{\theta}}^{-1} y$.

Predictive variance

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

("Plug-in" approach)

Illustration of prediction



Application to computer experiments

Computer model

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

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

Gaussian process model for computer experiments

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

- Bayesian framework on a fixed function

What we obtain

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

Conclusion

Gaussian process regression :

- The covariance function characterizes the Gaussian process
- It is estimated first (main topic of the talk, cf below)
- Then we can compute prediction and predictive variances with explicit matrix vector formulas
- Widely used for computer experiments

Maximum Likelihood for estimation

Explicit Gaussian likelihood function for the observation vector y

Maximum Likelihood

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

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

$\Rightarrow$ Numerical optimization with $O(n^3)$ criterion
$\Rightarrow$ Most standard estimation method

Cross Validation for estimation

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

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

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

$\Longrightarrow$ Can be used as an alternative method

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

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

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

Using the virtual Cross Validation formula :

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

and

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

$\Rightarrow$ Same computational cost as ML

Well-specified case

Estimation of θ only

For simplicity, we do not distinguish the estimations of σ^2 and 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_1 of the Gaussian Process belongs to the set $\{K_\theta, \theta \in \Theta\}$. Hence

$$K_1 = K_{\theta_0}, \theta_0 \in \Theta$$

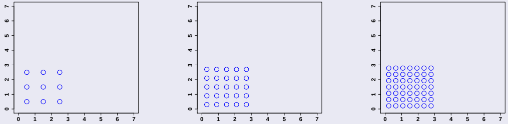
$\Longrightarrow$ Most standard theoretical framework for estimation
$\Longrightarrow$ ML and CV estimators can be analyzed and compared w.r.t. estimation error criteria (based on $|\hat{\theta} - \theta_0|$)

Two asymptotic frameworks for covariance parameter estimation

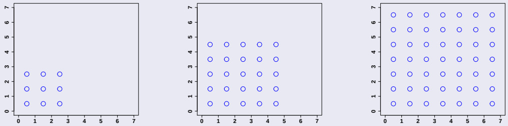
- Asymptotics (number of observations $n \to +\infty$) is an active area of research
- There are several asymptotic frameworks because they are several possible location patterns for the observation points

Two main asymptotic frameworks

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



- increasing-domain asymptotics : number of observation points is proportional to domain volume $\longrightarrow$ unbounded observation domain.



Existing fixed-domain asymptotic results

- From 80'-90' and onward. Fruitful theory for interaction estimation-prediction.
  - Stein M, Interpolation of Spatial Data : Some Theory for Kriging, *Springer, New York, 1999*.
- Consistent estimation is impossible for some covariance parameters (identifiable in finite-sample), see e.g.
  - Zhang, H., Inconsistent Estimation and Asymptotically Equivalent Interpolations in Model-Based Geostatistics, *Journal of the American Statistical Association (99), 250-261, 2004*.
- Proofs (consistency, asymptotic distribution) are challenging in several ways
  - They are done on a case-by-case basis for the covariance models
  - They may assume gridded observation points
- No impact of spatial sampling of observation points on asymptotic distribution
- (No results for CV)

Existing increasing-domain asymptotic results

- Consistent estimation is possible for all covariance parameters (that are identifiable in finite-sample). [More independence between observations]
- Asymptotic normality proved for Maximum-Likelihood

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

  N. Cressie and S.N Lahiri, The asymptotic distribution of REML estimators, *Journal of Multivariate Analysis 45 (1993) 217-233*.

  N. Cressie and S.N Lahiri, Asymptotics for REML estimation of spatial covariance parameters, *Journal of Statistical Planning and Inference 50 (1996) 327-341*.

  Under conditions that are
  - General for the covariance model
  - Not simple to check or specific for the spatial sampling
- (No results for CV)

$\Longrightarrow$ We study increasing-domain asymptotics for ML and CV

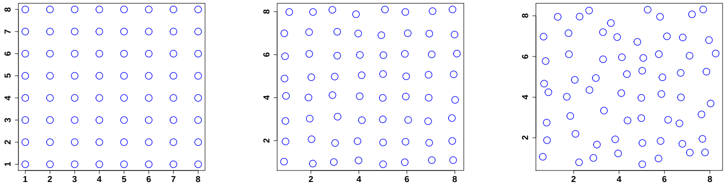
The randomly perturbed regular grid that we study

- Observation point X_i :

$$v_i + \epsilon U_i$$

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

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



Why a randomly perturbed regular grid ?

- Realizations car correspond to various sampling techniques for the observation points
- In the corresponding paper, one main objective is to study the impact of the irregularity (regularity parameter ϵ) :

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*.

- Note the condition $|\epsilon| < 1/2 \Longrightarrow$ minimum distance between observation points $\Longrightarrow$ technically convenient and appears in aforementioned publications

Consistency and asymptotic normality

Recall that $\mathbf{R}_\theta$ is defined by $(R_\theta)_{i,j} = K_\theta(X_i - X_j)$. (Family of random covariance matrices)
Under general summability, regularity and identifiability conditions, we show

Proposition : for ML

- a.s convergence of the random Fisher information : The random trace $\frac{1}{2n} Tr \left(\mathbf{R}_{\theta_0}^{-1} \frac{\partial \mathbf{R}_{\theta_0}}{\partial \theta_i} \mathbf{R}_{\theta_0}^{-1} \frac{\partial \mathbf{R}_{\theta_0}}{\partial \theta_j} \right)$ 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 $\Sigma_{ML} = \mathbf{I}_{ML}^{-1}$

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

Proposition : for CV

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

Main ideas for the proof

Based on applying classical M-estimator methods to the criteria functions

$$\frac{1}{n}\left(\ln\left(|\sigma^2\mathbf{R}_\theta|\right)+\frac{1}{\sigma^2}y^t\mathbf{R}_\theta^{-1}y\right) \quad \text{and} \quad \frac{1}{n}y^t\mathbf{R}_\theta^{-1}\mathit{diag}(\mathbf{R}_\theta^{-1})^{-2}\mathbf{R}_\theta^{-1}y$$

with

- observation vector y : $y_i = Y(X_i)$
- random covariance matrix $\mathbf{R}_\theta$: $(\mathbf{R}_\theta)_{i,j} = K_\theta(X_i - X_j)$

Then :

- 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 θ and θ_0 by the integrated square difference between K_θ and K_{θ_0}
- For almost-sure convergence of random traces : block-diagonal approximation of the random matrices involved and Cauchy criterion
- For asymptotic normality of criterion gradients : almost-sure (with respect to the random perturbations) Lindeberg-Feller Central Limit Theorem

Conclusion on well-specified case

In this expansion-domain asymptotic framework

- ML and CV are consistent and have the standard rate of convergence $\sqrt{n}$
- (not presented here) in the corresponding paper we show, numerically, than CV has a larger asymptotic variance $\Longrightarrow$ could be expected since we address the well-specified case
- (not presented here) in the paper we study numerically the impact of irregularity of spatial sampling on asymptotic variance $\Longrightarrow$ irregular sampling is beneficial to estimation

Misspecified case

The covariance function K_1 of Y does not belong to

$$\left\{ \sigma^2 K_\theta, \sigma^2 \geq 0, \boldsymbol{\theta} \in \Theta \right\}$$

$\Longrightarrow$ There is no true covariance parameter but there may be optimal covariance parameters for difference criteria :

- prediction mean square error
- confidence interval reliability
- multidimensional Kullback-Leibler distance
- ...

$\Longrightarrow$ Cross Validation can be more adapted than Maximum Likelihood for some of these criteria

Finite-sample analysis

We proceed in two steps

- When covariance function model is $\{\sigma^2 K_2, \sigma^2 \geq 0\}$, with K_2 a fixed correlation function, and K_1 is the true covariance function : explicit expressions and numerical tests
- In the general case : numerical studies

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*.

Case of variance parameter estimation

- $\hat{Y}_{new}$: prediction of $Y_{new} := Y(x_{new})$ with fixed misspecified correlation function K_2
- $\mathbb{E}\left[(\hat{Y}_{new} - Y_{new})^2 \middle| y \right]$: conditional mean square error of the prediction $\hat{Y}_{new}$
- One estimates σ^2 by $\hat{\sigma}^2$. $\hat{\sigma}^2$ may be $\hat{\sigma}^2_{ML}$ or $\hat{\sigma}^2_{CV}$
- Conditional mean square error of $\hat{Y}_{new}$ predicted by $\hat{\sigma}^2 c^2_{x_{new}}$ with $c^2_{x_{new}}$ fixed by K_2

Definition : the Risk

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

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

Explicit expression of the Risk

Let, for $i = 1, 2$:

- r_i be the covariance vector of Y between $x_1, ..., x_n$ and x_{new} with covariance function K_i
- $\mathbf{R}_i$ be the covariance matrix of Y at $x_1, ..., x_n$ with covariance function K_i

Proposition : 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

$$\begin{aligned} f(\mathbf{A}, \mathbf{B}) &= tr(\mathbf{A}) tr(\mathbf{B}) + 2tr(\mathbf{AB}) \\ \mathbf{M}_0 &= (\mathbf{R}_2^{-1} r_2 - \mathbf{R}_1^{-1} r_1)(r_2^t \mathbf{R}_2^{-1} - r_1^t \mathbf{R}_1^{-1}) \mathbf{R}_1 \\ \mathbf{M}_1 &= \mathbf{M} \mathbf{R}_1 \\ c_i &= 1 - r_i^t \mathbf{R}_i^{-1} r_i, \qquad i = 1, 2 \end{aligned}$$

Corollary : ML and CV are quadratic estimators $\Longrightarrow$ we can carry out an exhaustive numerical study of the Risk criterion

Two criteria for the numerical study

Definition : Risk on Target Ratio (RTR)

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

Definition : Bias on Target Ratio (BTR)

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

Integrated versions over the prediction domain $\mathcal{X}$

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

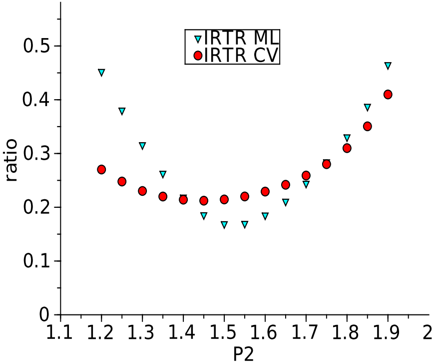
and

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

70 observation points on $[0,1]^5$. Mean over LHS-Maximin samplings.
K_1 and K_2 are power-exponential covariance functions,

$$K_i(x,y) = \exp\left(- \sum_{j=1}^{5} \left(\frac{|x_j - y_j|}{\ell_i} \right)^{p_i} \right),$$

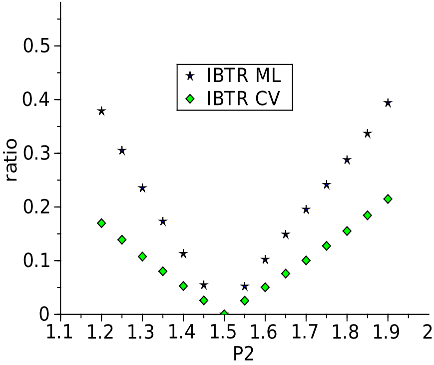
with $\ell_1 = \ell_2 = 1.2$, $p_1 = 1.5$, and p_2 varying.



70 observations on $[0,1]^5$. Mean over LHS-Maximin samplings.
K_1 and K_2 are power-exponential covariance functions,

$$K_i(x,y) = \exp\left(- \sum_{j=1}^{5} \left(\frac{|x_j - y_j|}{\ell_i} \right)^{p_i} \right),$$

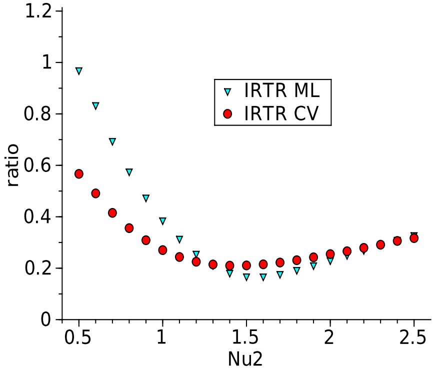
with $\ell_1 = \ell_2 = 1.2$, $p_1 = 1.5$, and p_2 varying.



70 observations on $[0,1]^5$. Mean over LHS-Maximin samplings.
K_1 and K_2 are Matérn covariance functions,

$$K_i(x,y) = \frac{1}{\Gamma(\nu_i)2^{\nu_i-1}} \left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right)^{\nu_i} K_{\nu_i} \left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right),$$

with Γ the Gamma function and K_{ν_i} the modified Bessel function of second order.
We use $\ell_1 = \ell_2 = 1.2$, $\nu_1 = 1.5$, and ν_2 varying.

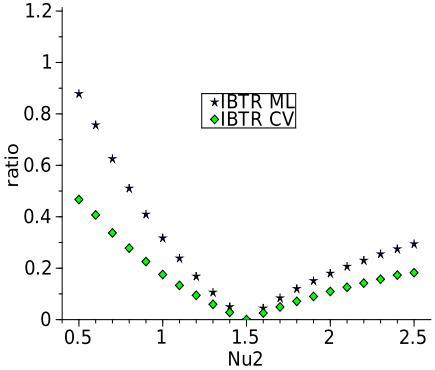


For designs of observation points that are not too regular (4/6)

70 observations on $[0,1]^5$. Mean over LHS-Maximin samplings.
K_1 and K_2 are Matérn covariance functions,

$$K_i(x,y) = \frac{1}{\Gamma(\nu_i)2^{\nu_i - 1}} \left(2\sqrt{\nu_i} \frac{||x-y||_2}{\ell_i} \right)^{\nu_i} K_{\nu_i} \left(2\sqrt{\nu_i} \frac{||x-y||_2}{\ell_i} \right),$$

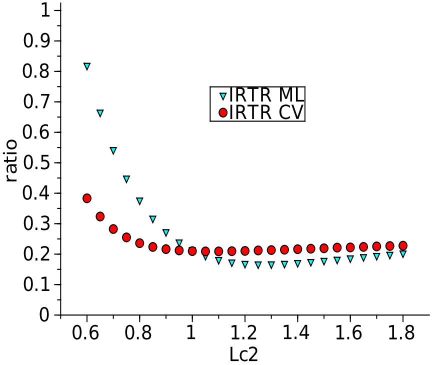
with Γ the Gamma function and K_{ν_i} the modified Bessel function of second order.
We use $\ell_1 = \ell_2 = 1.2$, $\nu_1 = 1.5$, and ν_2 varying.



70 observations on $[0,1]^5$. Mean over LHS-Maximin samplings.
K_1 and K_2 are Matérn covariance functions,

$$K_i(x,y) = \frac{1}{\Gamma(\nu_i)2^{\nu_i-1}} \left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right)^{\nu_i} K_{\nu_i}\left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right),$$

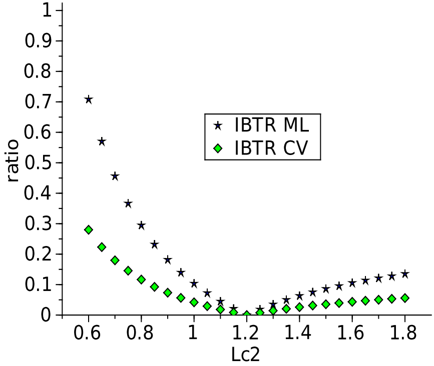
with Γ the Gamma function and K_{ν_i} the modified Bessel function of second order.
We use $\nu_1 = \nu_2 = \frac{3}{2}$, $\ell_1 = 1.2$ and ℓ_2 varying.



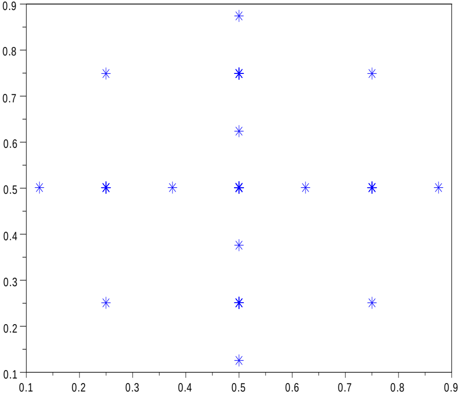
70 observations on $[0,1]^5$. Mean over LHS-Maximin samplings.
K_1 and K_2 are Matérn covariance functions,

$$K_i(x,y) = \frac{1}{\Gamma(\nu_i)2^{\nu_i-1}} \left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right)^{\nu_i} K_{\nu_i}\left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right),$$

with Γ the Gamma function and K_{ν_i} the modified Bessel function of second order.
We use $\nu_1 = \nu_2 = \frac{3}{2}$, $\ell_1 = 1.2$ and ℓ_2 varying.



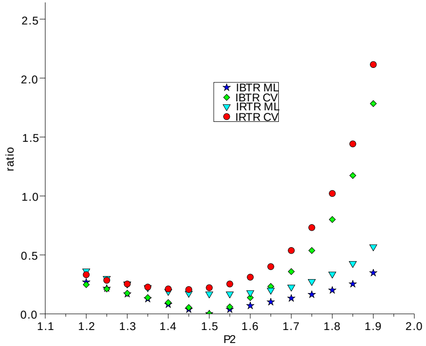
Projections of the observations points on the first two base vectors :



71 observations on $[0,1]^5$. Regular grid.
K_1 and K_2 are power-exponential covariance functions,

$$K_i(x,y) = \exp\left(- \sum_{j=1}^{5} \left(\frac{|x_j - y_j|}{\ell_i} \right)^{p_i} \right),$$

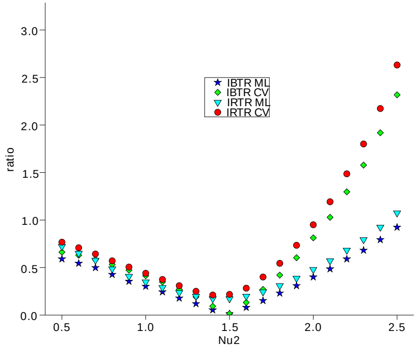
with $\ell_1 = \ell_2 = 1.2$, $p_1 = 1.5$, and p_2 varying.



71 observations on $[0,1]^5$. Regular grid.
K_1 and K_2 are Matérn covariance functions,

$$K_i(x,y) = \frac{1}{\Gamma(\nu_i)2^{\nu_i-1}} \left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right)^{\nu_i} K_{\nu_i}\left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i} \right),$$

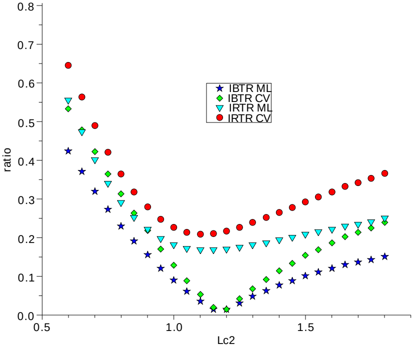
with Γ the Gamma function and K_{ν_i} the modified Bessel function of second order.
We use $\ell_1 = \ell_2 = 1.2$, $\nu_1 = 1.5$, and ν_2 varying.



71 observations on $[0,1]^5$. Regular grid.
K_1 and K_2 are Matérn covariance functions,

$$K_i(x,y) = \frac{1}{\Gamma(\nu_i)2^{\nu_i-1}} \left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i}\right)^{\nu_i} K_{\nu_i}\left(2\sqrt{\nu_i}\frac{||x-y||_2}{\ell_i}\right),$$

with Γ the Gamma function and K_{ν_i} the modified Bessel function of second order.
We use $\nu_1 = \nu_2 = \frac{3}{2}$, $\ell_1 = 1.2$ and ℓ_2 varying.



Summary of finite-sample analysis

For variance parameter estimation

- For not-too-regular designs of observation points : CV is more robust than ML to misspecification
  - Larger variance but smaller bias for CV
  - The bias term becomes dominant in the model misspecification case
- For regular design of experiments, CV is more biased than ML

$\Longrightarrow$ (not presented here) in the paper, a numerical study based on analytical functions confirms these findings for not-too-regular designs of observation points for the estimation of correlation parameters as well

Interpretation

- For irregular samplings of observations points, prediction for new points is similar to Leave-One-Out prediction $\Longrightarrow$ the Cross Validation criterion can be unbiased
- For regular samplings of observations points, prediction for new points is different from Leave-One-Out prediction $\Longrightarrow$ the Cross Validation criterion is biased

$\Longrightarrow$ we now aim at supporting this interpretation in an asymptotic framework (ongoing work)

Expansion-domain asymptotics with purely random sampling

Context :

- The observation points $X_1, ..., X_n$ are *iid* and uniformly distributed on $[0, n^{1/d}]^d$
- We use a parametric noisy Gaussian process model with stationary covariance function model

$$\{K_\theta, \theta \in \Theta\}$$

  with stationary K_θ of the form

$$K_\theta(t_1 - t_2) = \underbrace{K_{c,\theta}(t_1 - t_2)}_{\text{continuous part}} + \underbrace{\delta_\theta \mathbf{1}_{t_1 = t_2}}_{\text{noise part}}$$

  where $K_{c,\theta}(t)$ is continuous in t and $\delta_\theta > 0$
  $\Longrightarrow \delta_\theta$ corresponds to a measure error for the observations or a small-scale variability of the Gaussian process
- The model satisfies regularity and summability conditions
- The true covariance function K_1 is also stationary and summable

Cross Validation asymptotically minimizes the integrated prediction error (1/2)

Let $\hat{Y}_\theta(t)$ be the prediction of the Gaussian process Y at t, under correlation function K_θ, from observations $Y(x_1), ..., Y(x_n)$

Integrated prediction error :

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

Intuition :
The variable t above plays the same role as a new observation points X_{n+1}, uniform on $[0, n^{1/d}]^d$ and independent of $X_1, ..., X_n$

So we have

$$\mathbb{E}\left(E_{n,\theta}\right) = \mathbb{E}\left(\left[Y(X_{n+1}) - \mathbb{E}_{\theta|X}(Y(X_{n+1})|Y(X_1), ..., Y(X_n)) \right]^2 \right)$$

and so when n is large

$$\mathbb{E}\left(E_{n,\theta}\right) \approx \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{\theta,i,-i})^2 \right)$$

$\Longrightarrow$ This is an indication that the Cross Validation estimator can be optimal for integrated prediction error

Cross Validation asymptotically minimizes the integrated prediction error (2/2)

Based on ongoing work, we have

Theorem

$$E_{n,\hat{\theta}_{CV}} = \inf_{\theta \in \Theta} E_{n,\theta} + o_p(1).$$

Comments :

- Same Gaussian process realization for both covariance parameter estimation and prediction error
- The optimal (unreachable) prediction error $\inf_{\theta \in \Theta} E_{n,\theta}$ is lower-bounded $\Longrightarrow$ CV is indeed asymptotically optimal

Challenges

Purely random sampling $\Longrightarrow$ potential clusters of observation points $\Longrightarrow$

- This situation has not been studied in the literature
- If we do not consider noisy Gaussian processes, the eigenvalues of the random covariance matrices are not lower-bounded
- These eigenvalues are not upper-bounded

As a consequence, with the proof techniques we have used, it is not clear how to

- Treat Gaussian process models without noise
- Study asymptotic distribution of estimators

Maximum Likelihood asymptotically minimizes the multidimensional Kullback-Leibler divergence

Let $KL_{n,\theta}$ be $1/n$ times the Kullback-Leibler divergence $d_{KL}(K_0||K_\theta)$, between the multidimensional Gaussian distributions of y, given observation points $X_1, ..., X_n$, under covariance functions K_θ and K_0.
Based on ongoing work, we have

Theorem

$$KL_{n,\hat{\theta}_{ML}} = \inf_{\theta \in \Theta} KL_{n,\theta} + o_p(1).$$

Comments :

- In increasing-domain asymptotics, when $K_\theta \neq K_0$, $KL_{n,\theta}$ is usually lower-bounded $\Longrightarrow$ ML is indeed asymptotically optimal
- Maximum Likelihood is optimal for a criterion that is not prediction oriented

Conclusion

The results shown support the following general picture

- For well-specified models, ML would be optimal
- For regular designs of observation points, the principle of CV does not really have ground
- For more irregular designs of observation points, CV can be preferable for specific prediction-purposes (e.g. integrated prediction error). (But its variance can be problematic)

Some 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 !