Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case

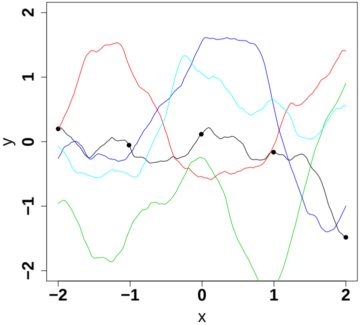
François Bachoc

University Paul Sabatier, Toulouse

CLAPEM 2016

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**
- Widely applied in machine learning, geostatistics, computer experiments...

Covariance function parametrization

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 $K_0(x_1, x_2) = \mathrm{Cov}(Y(x_1), Y(x_2))$

Covariance function parameterization

Covariance function model $\{K_\theta ; \theta \in \Theta\}$ for the Gaussian process Y.

- $\theta \in \Theta \subset \mathbb{R}^p$ is the multidimensional covariance parameter. K_θ is a covariance function

Observations

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

Estimation

Objective : build estimator $\hat{\theta}(y)$

Prediction with estimated covariance function

Gaussian process Y observed at $x_1, ..., x_n$ and predicted at x
$y = (Y(x_1), ..., Y(x_n))^t$

Once the covariance parameters have been estimated and fixed to $\hat{\theta}$

- $\mathbf{R}_{\hat{\theta}}$ is the covariance matrix of Y at $x_1, ..., x_n$ under covariance function $K_{\hat{\theta}}$
- $r_{\hat{\theta}}(x)$ is the covariance vector of Y between $x_1, ..., x_n$ and x under covariance function $K_{\hat{\theta}}$

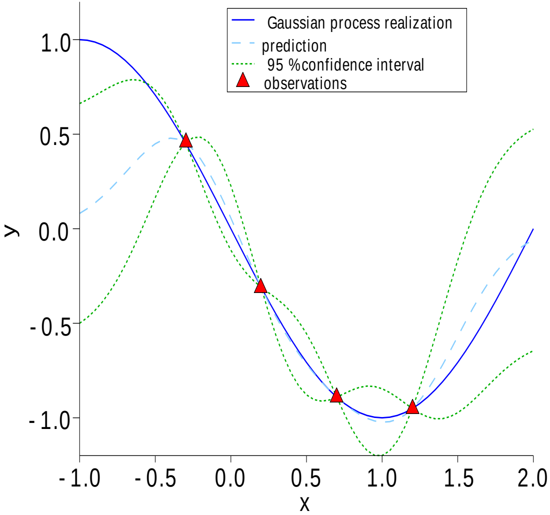
Prediction

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

Predictive variance

The predictive variance is $var_{\hat{\theta}}(Y(x)|Y(x_1), ..., Y(x_n)) = K_{\theta}(x, x) - r_{\hat{\theta}}^t(x)\mathbf{R}_{\hat{\theta}}^{-1}r_{\hat{\theta}}(x)$

Illustration of prediction



Maximum Likelihood (ML) for estimation

Explicit Gaussian likelihood function for the observation vector y

Maximum Likelihood

Define $\mathbf{R}_\theta$ as the covariance matrix of $y = (Y(x_1), ..., Y(x_n))$ with covariance function K_θ
The Maximum Likelihood estimator of θ is

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

$\Rightarrow$ Most standard estimation method

Cross Validation (CV) for estimation

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

Leave-One-Out criteria we study

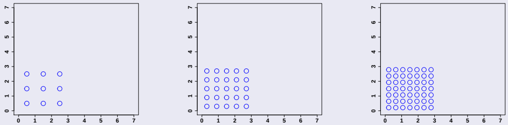
$$\hat{\theta}_{CV} \in \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_{\theta,i,-i})^2$$

$\Longrightarrow$ Alternative method used by some authors

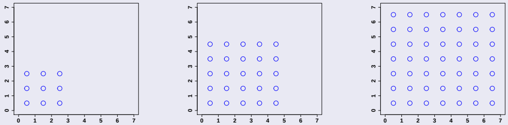
Asymptotic framework

Two main asymptotic frameworks

- fixed-domain asymptotics : As $n \to \infty$ the observation points are dense in a bounded domain (e.g. book Stein 99)



- increasing-domain asymptotics : As $n \to \infty$ the observation point density is constant and the observation domain is unbounded (e.g. Mardia and Marshall 83, Cressie and Lahiri 93, Bachoc 14)



We address increasing-domain asymptotic here

Well-specified and misspecified case

- Well-specified case : The covariance function $K_0 = K_{\theta_0}$ of Y belongs to

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

  - Estimators are evaluated w.r.t. the estimation error $|\hat{\theta} - \theta_0|$
  - Maximum Likelihood is preferable over Cross Validation (e.g. Bachoc 14)

- Misspecified case : The covariance function K_0 of Y does not belong to

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

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

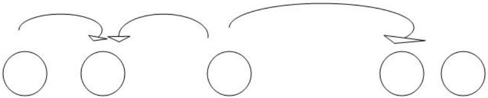
  - prediction mean square error
  - confidence interval reliability
  - multidimensional Kullback-Leibler distance
  - ...

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

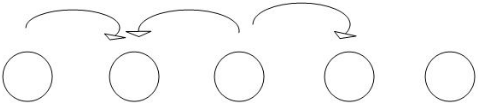
$\Longrightarrow$ We aim at providing asymptotic results supporting this last point

Impact of the spatial sampling

- For irregularly spaced observations points, prediction for new points can be similar to Leave-One-Out prediction $\Longrightarrow$ the Cross Validation criterion can be unbiased



- For regularly spaced observations points, prediction for new points is different from Leave-One-Out prediction $\Longrightarrow$ the Cross Validation criterion is biased



$\Longrightarrow$ we aim at supporting this interpretation in an asymptotic framework

Assumptions

- 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

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

with stationary K_θ of the form

$$K_\theta(x_1 - x_2) = \underbrace{K_{c,\theta}(x_1 - x_2)}_{\text{continuous part}} + \underbrace{\delta_\theta \mathbf{1}_{x_1 = x_2}}_{\text{noise part}}$$

where $K_{c,\theta}(x)$ is continuous in x 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 true covariance function is also of the form

$$K_0(x_1 - x_2) = K_{c,0}(x_1 - x_2) + \delta_0 \mathbf{1}_{x_1 = x_2}$$

- The model satisfies regularity and summability conditions
- The true covariance function K_0 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 covariance 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 point 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)

We show

Theorem

With

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

we have

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

Comments :

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

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

We show

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

A numerical illustration

- Dimension $d = 2$
- The true covariance function is isotropic Matérn with $\sigma_0^2 = 1$, $\ell_0 = 4$ and $\nu_0 = 10$
- The true noise variance is $\delta_0 = 0.25^2$
- The model covariance function is isotropic Matérn with known $\nu = 10$ and with $\theta = (\sigma^2, \ell)$ estimated by $\hat{\theta} = (\hat{\sigma}^2, \hat{\ell})$
- The noise variance δ_θ is enforced
  - to 0.25^2 in the well-specified case
  - to 0.1^2 in the misspecified case

n	Specification	Estimation	Average of $\hat{\ell}$	Standard deviation of $\hat{\ell}$	Average of $E_{n,\hat{\sigma}^2,\hat{\ell}}$	Average of $KL_{n,\hat{\sigma}^2,\hat{\ell}}$
100	Well-specified	ML	4.014	0.600	0.021	0.026
	Well-specified	CV	4.525	1.564	0.024	0.123
	Misspecified	ML	1.279	0.385	0.112	1.120
	Misspecified	CV	4.637	1.754	0.024	3.725
500	Well-specified	ML	3.990	0.244	0.016	0.004
	Well-specified	CV	4.158	0.698	0.016	0.031
	Misspecified	ML	1.216	0.122	0.104	1.076
	Misspecified	CV	4.167	0.727	0.016	3.477

TABLE: Monte Carlo simulations with 2000 samples. For each sample, we generate the data, compute $\hat{\sigma}^2$ and $\hat{\ell}$ by ML and CV, and compute the corresponding $E_{n,\hat{\sigma}^2,\hat{\ell}}$ and $KL_{n,\hat{\sigma}^2,\hat{\ell}}$.

Conclusion

- For well-specified models, ML generally appears to be optimal
- In the misspecified case with random observation points, CV is optimal for the integrated square prediction error
- In the misspecified case, a comparison of ML and CV would be criterion-dependent
- In practice, significantly different estimates between ML and CV can be a sign of model misspecification

Some potential perspectives

- Extension to other CV estimators
- Obtaining Central Limit Theorems
- Non-Gaussian case

The manuscript :

F. Bachoc, "Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case", *Bernoulli, in press*.

Thank you for your attention !