Asymptotic results for cross validation estimation of covariance parameters of Gaussian processes

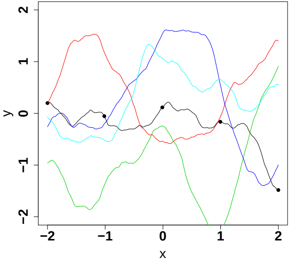
François Bachoc
Agnès Lagnoux, Thi Mong Ngoc Nguyen

Institut de Mathématiques de Toulouse

MathAmSud - 2020

Gaussian process regression (Kriging model)

Study of a **single realization** of a Gaussian process $x \to Y(x)$ on a domain $\mathcal{X} \subset \mathbb{R}^d$



Goal

Predicting the continuous realization function, from a finite number of **observation points**

Applications : Computer experiments, machine learning, geosciences,. . .

The Gaussian process

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 $\lim_{||x|| \to +\infty} K_1(x) = 0$

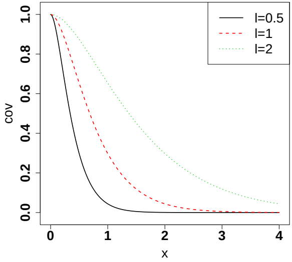
Example of the Matérn $\frac{3}{2}$ covariance function on $\mathbb{R}$

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 $\hat{\theta}(y)$

Maximum Likelihood (ML) for estimation

Explicit Gaussian likelihood function for the observation vector y

Maximum Likelihood

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

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

Leave-One-Out criteria

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

and

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

$\Longrightarrow$ Alternative method used by some authors. E.g. Sundararajan and Keerthi 2001, Zhang and Wang, 2010, Bachoc 2013

$\Longrightarrow$ Cost is $O(n^3)$ as well (Dubrule, 1983)

Well-specified case

Estimation of ψ

We let $\psi = (\sigma^2, \theta)$. Hence we consider the set $\left\{ K_\psi, \psi \in \Psi \right\}$ of covariance functions for the estimation

Well-specified model

The true covariance function K_1 of the Gaussian process belongs to the set $\left\{ K_\psi, \psi \in \Psi \right\}$. Hence

$$K_1 = K_{\psi_0}, \psi_0 \in \Psi$$

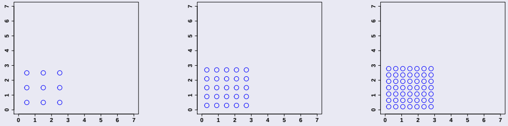
$\Longrightarrow$ Most standard theoretical framework for estimation
$\Longrightarrow$ ML and CV estimators can be analyzed and compared w.r.t. estimation error criteria (based on $||\hat{\psi} - \psi_0||$)

Two asymptotic frameworks for covariance parameter estimation

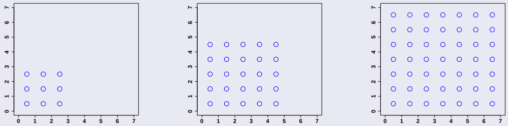
- Asymptotics (number of observations $n \to +\infty$) is an active area of research
- There are several asymptotic frameworks because they are several possible location patterns for the observation points

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

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 and Cross-Validation

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*.

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*.

Our fixed-domain asymptotic setting : exponential covariance function

Observation setting :

- Fixed one-dimensional domain $\mathcal{X} = [0, 1]$
- We consider a triangular array of observation points $\{x_i^{(n)}; 1 \leq i \leq n, n \in \mathbb{N}\}$
- We let $(x_1, ..., x_n) = (x_1^{(n)}, ..., x_n^{(n)})$
- We assume $0 = x_1 < x_2 < ... < x_n = 1$

Covariance function :

- $K_\psi(t) = K_{\sigma^2,\theta}(t) = \sigma^2 e^{-\theta|t|}$
- $(\sigma^2, \theta) \in [a, A] \times [b, B]$, with $0 < a < A < \infty$, $0 < b < B < \infty$
- Ornstein-Uhlenbeck process

Comments

- More amenable to theoretical analysis
  - Correlation matrix $\mathbf{R}_\theta = [e^{-\theta |x_i - x_j|}]_{1 \leq i,j \leq n}$ has an explicit inverse
  - Markovian process
- Studied by : Ying 1991, 1993, chen et al 2000, Antognini 2010, Chang et al 2017, Velandia et al 2017
- Covariance function not differentiable at 0 $\Longrightarrow$ realizations are not differentiable

Asymptotic results for maximum likelihood

- The parameters σ^2 and θ can not be estimated consistently
- The product $\sigma^2\theta$ can
- Ying, 1991 showed

$$\hat{\sigma}^2_{ML}\hat{\theta}_{ML} \to^{a.s.}_{n\to\infty} \sigma_0^2\theta_0 \quad \text{and} \quad \frac{\sqrt{n}}{\sqrt{2}\sigma_0^2\theta_0}(\hat{\sigma}^2_{ML}\hat{\theta}_{ML} - \sigma_0^2\theta_0) \to^{\mathcal{D}}_{n\to\infty} \mathcal{N}(0,1)$$

- Asymptotic variance is $(\sqrt{2}\sigma_0^2\theta_0)^2$ independently of the triangular array of observation points

Our result for cross validation

- Joint work with Agnès Lagnoux and Jade Nguyen (University of Toulouse)

  F. Bachoc, A. Lagnoux and T.M.N. Nguyen Cross-validation estimation of covariance parameters under fixed-domain asymptotics, *Journal of Multivariate Analysis 160 (2017) 42-67*.

- We study the cross validation estimator

$$(\hat{\sigma}_{CV}^2, \hat{\theta}_{CV}) \in \underset{a \leq \sigma^2 \leq A, b \leq \theta \leq B}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\log(\sigma^2 c_{\theta,i,-i}^2) + \frac{(y_i - \hat{y}_{\theta,i,-i})^2}{\sigma^2 c_{\theta,i,-i}^2} \right]$$

- We show

$$\hat{\sigma}_{CV}^2 \hat{\theta}_{CV} \to_{n \to \infty}^{a.s.} \sigma_0^2 \theta_0 \quad \text{and} \quad \frac{\sqrt{n}}{\tau_n \sigma_0^2 \theta_0} (\hat{\sigma}_{CV}^2 \hat{\theta}_{CV} - \sigma_0^2 \theta_0) \to_{n \to \infty}^{\mathcal{D}} \mathcal{N}(0,1)$$

- $(\tau_n \sigma_0^2 \theta_0)^2$ is the asymptotic variance

Analysis of the asymptotic variance

- Let $\Delta_i = x_{i+1} - x_i$ for $i = 2, ..., n$
- We have

$$\tau_n^2 = \frac{2}{n} \sum_{i=3}^{n-1} \left[\left(\frac{\Delta_{i+1}}{\Delta_i + \Delta_{i+1}} + \frac{\Delta_{i-1}}{\Delta_i + \Delta_{i-1}} \right)^2 + 2 \frac{\Delta_i \Delta_{i+1}}{(\Delta_i + \Delta_{i+1})^2} \right]$$

- We show, for any triangular array $\{x_1, ..., x_n\}$ satisfying $\max_{i=2,\dots,n} \Delta_i \to_{n \to \infty} 0$

$$2 \leq \liminf_{n \to \infty} \tau_n^2 \leq \limsup_{n \to \infty} \tau_n^2 \leq 4$$

- Asymptotic variance larger than for Maximum Likelihood
- We provide examples of triangular arrays reaching the lower and upper bound
- We extend the results to unknown non-zero mean functions

Misspecified case

The covariance function K_1 of Y does not belong to

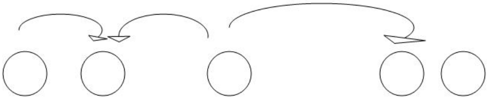
$$\{K_\psi, \psi \in \Psi\}$$

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

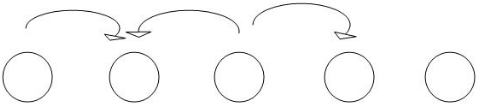
- prediction mean square error
- confidence interval reliability
- multidimensional Kullback-Leibler distance
- ...

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

- 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

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

$$\{K_\psi, \psi \in \Psi\}$$

with stationary K_ψ of the form

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

where $K_{c,\psi}(t)$ is continuous in t and $\delta_\psi > 0$
$\Longrightarrow \delta_\psi$ 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 $\hat{Y}_\psi(t)$ be the prediction of the Gaussian process Y at t, under correlation function K_ψ, from observations $Y(x_1), ..., Y(x_n)$

Integrated prediction error :

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

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

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

and so when n is large

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

$\Longrightarrow$ 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 in

F. Bachoc, "Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case", *Bernoulli, 2018*.

Theorem

With

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

we have

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

Comments :

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

Conclusion

The results shown support the following general picture

- For well-specified models, ML would be optimal
- CV can be preferable in the misspecified case for specific prediction-purposes (e.g. integrated prediction error).
  - beware of regularly spaced observation points
  - CV can yield large variances

Thank you for your attention !