Habilitation defense
Contributions to Gaussian processes, uncertainty quantification and post-model-selection inference

François Bachoc

Institut de Mathématiques de Toulouse
Université Paul Sabatier

Toulouse
November 29 2018

Jury

Jury members :

M. Bernard BERCU	Université Bordeaux 1
Mme Béatrice LAURENT	INSA Toulouse
M. Jean-Michel LOUBES	Université Paul Sabatier
M. Emilio PORCU	Newcastle University
Mme Clémentine PRIEUR	Université Grenoble Alpes
M. Vincent RIVOIRARD	Université Paris Dauphine

Based on reports from :

Mme Gerda CLAESKENS	KU Leuven
M. Emilio PORCU	Newcastle University
Mme Clémentine PRIEUR	Université Grenoble Alpes

A short CV

Career

- **2013** PhD defense in October at University Paris Diderot
- **2013-2015** Post-doctoral fellow at the University of Vienna
- **2015-...** Maître de Conférences at Institut de Mathématiques de Toulouse

Teaching and service

- Gave various courses in Vienna and Toulouse on mathematics and statistics
- **2016-2018** Responsible of the "CMI" track for bachelor students in mathematics
- **2016-2018** Co-organizer of the statistics seminar
- Reviewer for statistics journals and machine learning conferences

Funding acknowledgement

- Chaire OQUAIDO
- ANR projects PEPITO, RISCOPE, SansSoucis, MESA
- IDEX, Université Fédérale Toulouse Midi-Pyrénées
- CNRS (PEPS)

Student mentoring

PhD theses co-advision

- **2016-...** Andrés Felipe López-Lopera,
  - Gaussian processes with inequality constraints
  - With École des Mines de Saint Etienne
  - Co-supervision with Nicolas Durrande and Olivier Roustant
- **2017-...** Baptiste Broto
  - Shapley effects in sensitivity analysis + Gaussian processes with permutations
  - With CEA Saclay (alternative energies and atomic energy commission)
  - Co-supervision with Marine Depecker and Jean-Marc Martinez
- **2017-...** José Daniel Betancourt
  - Gaussian processes with functional inputs for coastal flooding
  - Institut de Mathématiques de Toulouse
  - Co-supervision with Thierry Klein

Bachelor and master theses advision

- **2016** Antonin Lavigne (bachelor), with Sébastien Gerchinovitz
- **2017** Théo Barthe (master)

1. Covariance parameter estimation for Gaussian processes

- since PhD thesis beginning in 2010
- Includes funding from OQUAIDO, PEPITO, RISCOPE

2. Other contributions to Gaussian processes

- mostly since 2015 in Toulouse
- Includes Andrés', Baptiste's and José's theses
- Includes funding from OQUAIDO, PEPITO, RISCOPE

3. Valid confidence intervals post-model-selection

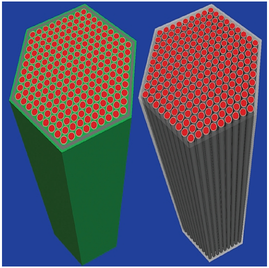
- since post-doc beginning in 2013
- Includes funding from SansSoucis

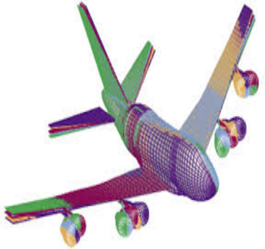
1. Covariance parameter estimation for Gaussian processes
   - Introduction to Gaussian processes
   - A focus on one paper
   - Short description of other papers

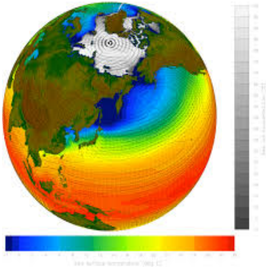
2. Other contributions to Gaussian processes
   - A focus on one paper
   - Short description of other papers

3. Valid confidence intervals post-model-selection
   - Introduction to post-model-selection inference
   - A focus on one paper
   - Short description of other papers

Computer models have become essential in science and industry !







For clear reasons : cost reduction, possibility to explore hazardous or extreme scenarios...

Computer models as expensive functions

A computer model can be seen as a deterministic function

$$\begin{aligned} f : \mathbb{X} \subset \mathbb{R}^d &\to \mathbb{R} \\ x &\mapsto f(x) \end{aligned}$$

- x : tunable simulation parameter (e.g. geometry)
- $f(x)$: scalar quantity of interest (e.g. energetic efficiency)

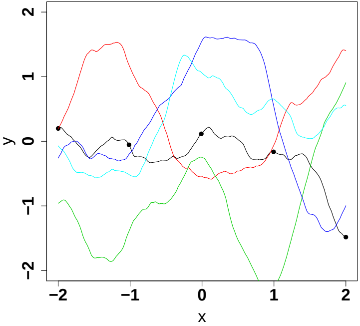
The function f is usually

- continuous (at least)
- non-linear
- only available through evaluations $x \mapsto f(x)$

$\Longrightarrow$ black box model

Gaussian processes

Modeling the **black box function** as a **single realization** of a Gaussian process $x \to \xi(x)$ on the domain $\mathbb{X} \subset \mathbb{R}^d$



Usefulness

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

Gaussian processes

Definition

A stochastic process $\xi : \mathbb{X} \to \mathbb{R}$ is Gaussian if for any $x_1, ..., x_n \in \mathbb{X}$, the vector $(\xi(x_1), ..., \xi(x_n))$ is a Gaussian vector

Mean and covariance functions

The distribution of a Gaussian process is characterized by

- Its mean function : $x \mapsto m(x) = \mathbb{E}(\xi(x))$. Can be any function $\mathbb{X} \to \mathbb{R}$
- Its covariance function $(x_1, x_2) \mapsto k(x_1, x_2) = Cov(\xi(x_1), \xi(x_2))$. Must yield valid covariance matrices

The covariance function

In most classical cases :

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

Example $k(x_1, x_2) = \sigma^2 e^{-||x_1 - x_2||/\ell}$

Conditional distribution

Gaussian process ξ observed at $x_1, ..., x_n$

Notation

- $y = (\xi(x_1), ..., \xi(x_n))'$
- R is the $n \times n$ matrix $[k(x_i, x_j)]$
- $r(x) = (k(x, x_1), ..., k(x, x_n))'$
- $m = (m(x_1), ..., m(x_n))'$

Conditional mean

The conditional mean is $m_n(x) := \mathbb{E}(\xi(x)|\xi(x_1), ..., \xi(x_n)) = m(x) + r(x)'R^{-1}(y - m)$

Conditional variance

The conditional variance is $k_n(x, x) = var(\xi(x)|\xi(x_1), ..., \xi(x_n)) = k(x, x) - r(x)'R^{-1}r(x)$

Conditional distribution

Conditionally to $\xi(x_1), ..., \xi(x_n)$, ξ is a Gaussian process with (conditional) mean function m_n and (conditional) covariance function $(x, y) \rightarrow k_n(x, y) = k(x, y) - r(x)'R^{-1}r(y)$

Illustration of conditional mean and variance

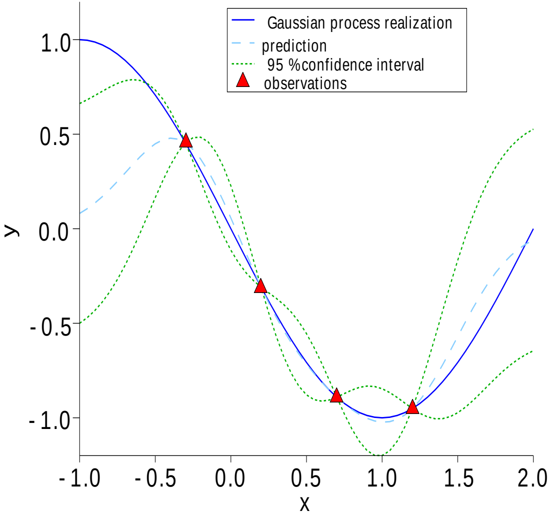
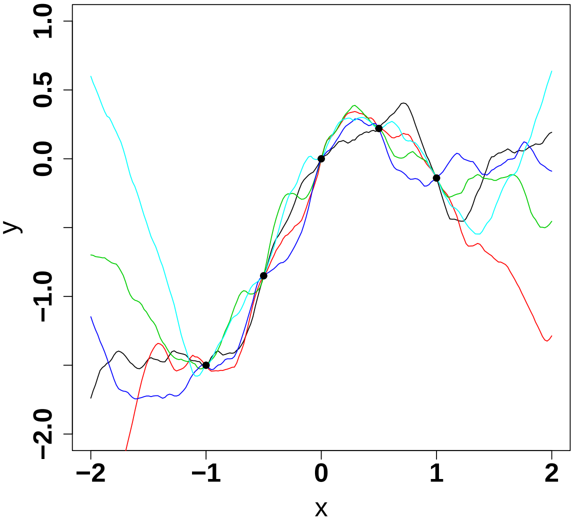


Illustration of the conditional distribution



Covariance function estimation

- Assume in the rest of the section that the mean function of ξ is zero
- One needs to select (estimate) a covariance function in order to apply the prediction formulas
- Classically, it is assumed that the covariance function k belongs to a parametric set

Parameterization

Covariance function model $\{k_\theta, \theta \in \Theta\}$ for the Gaussian process ξ

- θ is the multidimensional covariance parameter. k_θ is a covariance function

Observations

ξ is observed at $x_1, ..., x_n \in \mathbb{X}$, yielding the Gaussian vector $y = (\xi(x_1), ..., \xi(x_n))'$

Estimation

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

Maximum likelihood (ML) for estimation

Explicit Gaussian likelihood function for the observation vector y

Maximum likelihood

Define R_θ as the covariance matrix of $y = (\xi(x_1), ..., \xi(x_n))'$ with covariance function k_θ :
$R_\theta = [k_\theta(x_i, x_j)]_{i,j=1,\ldots,n}$
The maximum likelihood estimator of θ is

$$\hat{\theta}_{ML} \in \underset{\theta \in \Theta}{\operatorname{argmax}} \left(\frac{1}{(2\pi)^{n/2} |R_\theta|} e^{-\frac{1}{2} y' R_\theta^{-1} y} \right)$$

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

Cross Validation (CV) for estimation

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

Cross Validation

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

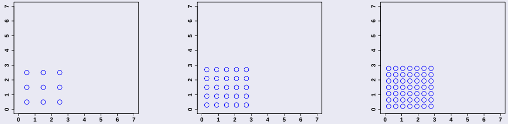
$\Longrightarrow$ Alternative method used by some authors. E.g. Sundararajan and Keerthi 2001, Zhang and Wang, 2010
$\Longrightarrow$ Cost is $O(n^3)$ as well (Dubrule, 1983)

Two asymptotic frameworks for Gaussian processes

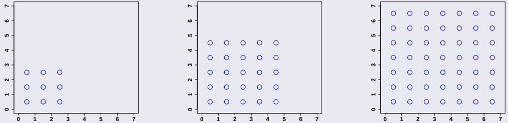
- Asymptotics (number of observations $n \to +\infty$) is an active area of research
- There are several asymptotic frameworks because there are several possible location patterns for the observation points

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.



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

Misspecified case

The covariance function k of ξ 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

Expansion-domain asymptotics with random sampling

- The observation points $(x_1, \ldots, x_n) = (X_1, \ldots, X_n)$ are *iid* and uniformly distributed on $[0, n^{1/d}]^d$
- We consider a covariance model $\{k_\theta; \theta \in \Theta\}$
- Regularity and summability conditions

CV minimizes the integrated prediction error

Let $\hat{\xi}_\theta(t)$ be the prediction of $\xi(t)$, under covariance function k_θ, from observations $\xi(x_1), ..., \xi(x_n)$

Integrated prediction error :

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

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

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

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

CV minimizes the integrated prediction error

We show

Theorem

With

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

we have

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

Comment :

- The optimal (unreachable) prediction error $\inf_{\theta \in \Theta} E_{n,\theta}$ is lower-bounded $\Longrightarrow$ CV is indeed asymptotically optimal

Short description of other papers

- In Furrer, Bachoc, Du 2016, we show the increasing-domain asymptotic consistency of covariance tapering
  - Motivation : approximation to circumvent the $O(n^3)$ cost
- In Bachoc, Furrer 2017, we lower bound the smallest eigenvalues of covariance matrices from multivariate processes
  - Motivation : appears as a necessary condition for increasing-domain asymptotic results
- In Velandia, Bachoc, Bevilacqua, Gendre, Loubes 2017 and Bachoc, Lagnoux, Nguyen 2017, we study consistency and asymptotic normality under fixed-domain asymptotics
  - For exponential covariance function in dimension one
  - Bivariate maximum likelihood and cross validation

J. Bect, F. Bachoc and D. Ginsbourger, A supermartingale approach to Gaussian process based sequential design of experiments, *Bernoulli, forthcoming*

We consider a Gaussian process ξ on a fixed compact $\mathbb{X} \subset \mathbb{R}^d$

- continuous mean function m
- continuous covariance function k
- continuous sample paths

Motivation

- When we observe $\xi(x_1), ..., \xi(x_n)$, the mean and covariance functions become m_n and k_n
- $\Longrightarrow$ We want to choose $x_1, ..., x_n$ so that m_n and k_n become maximally informative

e.g. $k_n(x,x)$ small, or $k_n(x,x)$ small when $m_n(x)$ is large

Sequential design

It is more efficient to select x_{i+1} after $\xi(x_1), ..., \xi(x_i)$ are observed

The observation points $x_1, ..., x_n$ become random observation points $X_1, ..., X_n$

Uncertainty functionals

Gaussian measures

- A Gaussian measure ν is a measure on $\mathcal{C}(\mathbb{X})$ corresponding to a Gaussian process with continuous sample paths (see e.g. Bogachev 98).

Uncertainty functional

It is a function $\mathcal{H} : \nu \mapsto \mathcal{H}(\nu) \in [0, \infty)$

Uncertainty functional : example

Expected improvement (EI) (Mockus 78, Jones et al. 98)

$$\mathcal{H}(\nu) = \mathbb{E}(\max_{u \in \mathbb{X}} \xi_\nu(u)) - \max_{u \in \mathbb{X}; k_\nu(u,u)=0} \mathbb{E}(\xi_\nu(u))$$

where

- ν has covariance function k_ν
- ξ_ν is a Gaussian process with distribution ν

$\Longrightarrow$ global optimization

- Let
$$\text{Cond}_{\xi(X_1),\dots,\xi(X_i),\xi(x)}$$
be the conditional distribution of ξ given $\xi(X_1),\dots,\xi(X_i),\xi(x)$

Stepwise Uncertainty Reduction (SUR)

The choice of observation points $(X_i)_{i \geq 1}$ follows a SUR strategy when

$$X_{i+1} \in \underset{x \in \mathbb{X}}{\operatorname{argmin}} \, \mathbb{E}_{|\xi(X_1),\dots,\xi(X_i)} \left(\mathcal{H} \left[\text{Cond}_{\xi(X_1),\dots,\xi(X_i),\xi(x)} \right] \right)$$

$\Longrightarrow$ minimizing the expected uncertainty after one additional evaluation of ξ

Expected improvement example

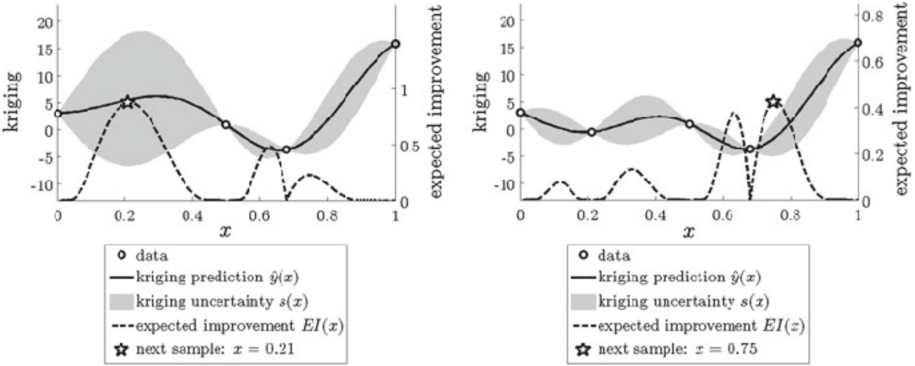
Let $\mathbb{E}_n$ be the conditional mean given $\xi(X_1), \ldots, \xi(X_n)$

- Expected improvement

$$X_{n+1} \in \operatorname*{argmax}_{x \in \mathbb{X}} \mathbb{E}_n \left(\left(\xi(x) - \max_{u \in \mathbb{X}; k_n(u,u)=0} \xi(u) \right)^+ \right)$$

Illustration of Expected Improvement

(for minimization)



(Figure borrowed from Viana et al. 13, Journal of global optimization)

Consistency

We want to provide general conditions ensuring that

$$\mathcal{H}\left(\mathrm{Cond}_{\xi(X_1),\ldots,\xi(X_n)}\right) \xrightarrow[n\to\infty]{a.s.} 0$$

$\Longrightarrow$ Uncertainty going to zero

Consistency

Let

$$\mathcal{G}_n = \sup_{x \in \mathbb{X}} \left(\mathcal{H}\left[\text{Cond}_{\xi(X_1),\ldots,\xi(X_n)}\right] - \mathrm{E}_{|\xi(X_1),\ldots,\xi(X_n)} \left\{ \mathcal{H}\left[\text{Cond}_{\xi(X_1),\ldots,\xi(X_n),\xi(x)}\right] \right\} \right)$$

(maximum expected uncertainty reduction)

Theorem

Let $\mathcal{H}$ denote an uncertainty functional with the supermartingale property

- uncertainty always decreases on average when adding an observation

Let (X_n) follow a SUR strategy

Then $\mathcal{G}_n \to 0$ almost surely

If, moreover, continuity conditions hold and if $\mathcal{H}$ is such that

no possible uncertainty reduction with one more observation $\Longrightarrow$ no uncertainty

then

$$\mathcal{H}\left(\text{Cond}_{\xi(X_1),\ldots,\xi(X_n)}\right) \xrightarrow[n \to \infty]{a.s.} 0$$

Summary of additional results

- We prove that the general results apply to four examples
- We introduce the notion of regular loss function, where $\mathcal{H}$ is an average loss when estimating a quantity of interest (e.g. maximum and maximizer of ξ).
- We provide a specific convergence result for regular loss functions, with easier to check assumptions

Short description of other papers

- In Bachoc, Ammar, Martinez 2016, we apply Gaussian processes to nuclear engineering
  - Comparison with neural networks and kernel regression
  - Outlier and numerical instability detection

- In Rullière, Durrande, Bachoc, Chevalier 2017 and Bachoc, Durrande, Rullière, Chevalier 2018+, we study the aggregation of Gaussian process models from data subsets
  - Motivation : approximation to circumvent the $O(n^3)$ cost

- In Bachoc, Gamboa, Loubes, Venet 2017, we study Gaussian processes indexed by one-dimensional probability distributions
  - Transport-based distances for covariance functions
  - Increasing-domain consistency and asymptotic normality for maximum likelihood

- In López-Lopera, Bachoc, Durrande, Roustant 2018 and Bachoc, Lagnoux, López-Lopera 2018+ we study Gaussian processes with inequality constraints
  - Boundedness and/or monotonicity and/or convexity
  - More intensive MCMC procedures
  - Fixed-domain consistency and asymptotic normality for constrained maximum likelihood

Data generating process

Location model

$$Y = \mu + U$$

- Y of size $n \times 1$: observation vector
- μ of size $n \times 1$: unknown mean vector
- $U \sim \mathcal{N}(0, \sigma^2 I_n)$
- σ^2 unknown

$\Longrightarrow$ Working distribution $P_{n,\mu,\sigma}$

Linear submodels

Consider a design matrix X of size $n \times p$

- $p < n$ or $p \geq n$

Linear submodels

Subsets $M \subset \{1, ..., p\}$ of the columns of X. Approximating μ by

$$X[M]v$$

- $X[M]$ of size $n \times |M|$: only the columns of X that are in M
- $X[M]$ full rank
- v of size $|M| \times 1$: needs to be selected/estimated

Restricted least square estimator

$$\hat{\beta}_M = \left(X'[M]X[M]\right)^{-1} X'[M]Y$$

The projection-based target

Let for $|M| \leq n$

$$\beta_M^{(n)} = \underset{v}{\operatorname{argmin}} ||\mu - X[M]v||$$

$$\beta_M^{(n)} = \left(X'[M]X[M]\right)^{-1} X'[M]\mu$$

Then $\beta_M^{(n)}$ is a target of inference here

Post-model-selection inference

Model selection procedure

Data-driven selection of the model with $\hat{M}(Y) = \hat{M}$
Ex : sequential testing, AIC, BIC, LASSO

- In Berk et al. 2013, annals of statistics, the target for inference is $\beta^{(n)}_{\hat{M}}$ and $\hat{M}$ can be any model selection procedure
  - Model selector $\hat{M}$ is "imposed"
  - Objective : best coefficients in this imposed model

This is what we call a post-model-selection inference problem

F. Bachoc, H. Leeb, B.M. Pötscher, “Valid confidence intervals for post-model-selection predictors”, *Annals of Statistics (forthcoming)*.

Predictors

Let

$$y_0 = \mu_0 + u_0$$

- $u_0 \sim \mathcal{N}(0, \sigma^2)$

Let x_0 be a $p \times 1$ vector
We consider the predictor target

$$x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$$

The confidence interval construction of Berk et al. 2013

Let a nominal level $1 - \alpha \in (0, 1)$ be fixed
The method of Berk et al. (2013) directly yields confidence intervals for $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ of the form

$$CI = x_0'[\hat{M}]\hat{\beta}_{\hat{M}} \pm K_1 ||s_{\hat{M}}||\hat{\sigma},$$

with

- $s_M' = x_0'[M] \left(X'[M]X[M]\right)^{-1} X'[M]$
- $\hat{\sigma}^2$ a variance estimator with appropriate properties
- "POSI Constant" K_1 does not depend on Y (but on X, x_0) (main novelty)

Interpretation

- Except from K_1 : standard confidence intervals for fixed M
- K_1 adresses the randomness of $\hat{M}$

The CIs satisfy

$$\inf_{\mu \in \mathbb{R}^n, \sigma > 0} P_{n,\mu,\sigma} \left(x_0'[\hat{M}]\beta_{\hat{M}}^{(n)} \in CI \right) \geq 1 - \alpha$$

$\Longrightarrow$ **Uniformly valid** confidence interval

Issues when x_0 is partially observed

The constant K_1 depends on all the components of x_0

It can happen that only $x_0[\hat{M}]$ is observed

- model selection for cost reason

We hence construct other constants so that

$$K_1 \leq K_2 \leq K_3 \leq K_4$$

(The CIs given by K_2, K_3, K_4 are hence universally valid)
K_2, K_3, K_4 depend only on $x_0[\hat{M}]$

Remarks :

- K_4 is introduced in a version of Berk et al. 2013
- The cost of computing K_1 can be exponential in p (in practice : $p \leq 30$ if all submodels considered)
- K_4 is cheap to compute

- K_1 depends on x_0 and X, and it does not seem easy to provide a systematic large p analysis, for any X, x_0
- When $x_0 = e_i$ (base vector), Berk et al. 2013 show that (for $p \leq n$)
  - When X has orthogonal columns, K_1 has rate $\sqrt{\log(p)}$
  - There exists sequences of X so that K_1 has rate $\sqrt{p}$

We show

Proposition

$\Longrightarrow$ When all submodels are allowed for

(a) Let X have orthogonal columns. There exists a sequence of vectors x_0 such that

$$\liminf_{p \to \infty} K_1(x_0)/\sqrt{p} > 0$$

(b) K_2, K_3, K_4 have rate $\sqrt{p}$ for any sequence of matrices X

$\Longrightarrow$ When submodels are restricted

K_4 has a smaller rate that is explicit

- **Issue :** The target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ depends on X but is a predictor of y_0 from x_0
- Issue is solved when lines of X and x_0' are realizations from the same distribution $\mathcal{L}$
- We define the design-independent target $x_0[\hat{M}]'\beta_{\hat{M}}^{(\star)}$
- It depends on $\mathcal{L}$ but not on X

Theorem : asymptotic coverage for fixed p

Under conditions on X and $\hat{M}$:

For *CI* obtained by K_1, K_2, K_3, K_4,

$$\inf_{\mu,\sigma} P_{n,\mu,\sigma}\left(x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)} \in CI \middle| X \right) \geq (1-\alpha) + o_p(1)$$

Short description of other papers

- In Bachoc, Ehler, Gräf 2017, we use optimal configurations of lines for computation of post-model-selection inference constants K
  - Link with potential minimization in applied mathematics

- In Bachoc, Blanchard, Neuvial 2018, we provide an upper bound on K_1 under restricted isometry properties (RIP)
  - Asymptotically tight
  - Extends results on orthogonal X

- In Bachoc, Preinerstorfer, Steinberger 2018, we extend the previous confidence intervals
  - General data generating processes
  - Non-linear models (e.g. binary regression)
  - Conservative intervals for unknown variances
  - Uniform asymptotic guarantees for fixed dimension

General conclusion

Summary

- Gaussian processes (Section 1 + Section 2)
  - Bayesian framework over functions
  - Asymptotic results for covariance estimation and sample path inference
  - Applications to computer models
- Post-model-selection inference (Section 3)
  - Selected model is imposed, inference over projection-based target
  - Asymptotic guarantees
  - Many numerical comparisons between procedures
- Other work and ongoing work $\rightarrow$ manuscript

Some open perspectives

- More general fixed-domain asymptotic results for Gaussian processes
- Tailored Gaussian processes for specific data
- Post-model-selection inference : algorithms for approximating/bounding K_1

Thank you for your attention !