Gaussian processes indexed on the symmetric group: prediction and learning

François Bachoc[1], Baptiste Broto[1,2], Fabrice Gamboa[1], Jean-Michel Loubes[1]

[1] Institut de Mathématiques de Toulouse
[2] CEA Saclay

Saint-Etienne - 2018

Gaussian process regression (Kriging model)

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

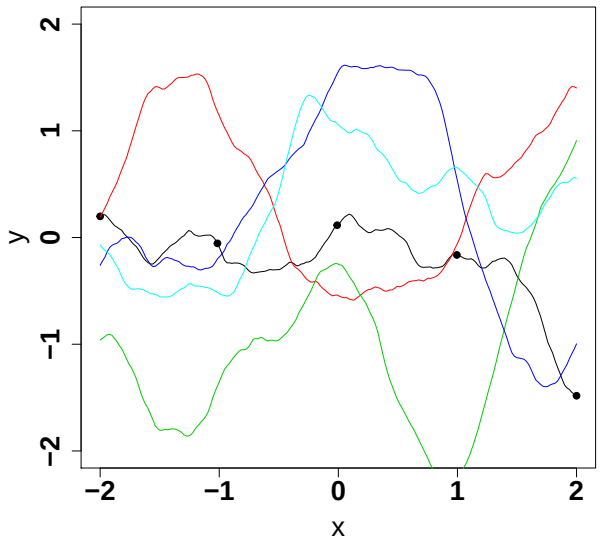


FIGURE: Illustration with $\mathcal{X} = [-2, 2]$

Goal

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

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

The Gaussian process

Definition

A stochastic process $Y : \mathcal{X} \to \mathbb{R}$ is a Gaussian process if and only if, for any $n \in \mathbb{N}$, for any $x_1, \ldots, x_n \in \mathcal{X}$, the random vector $(Y(x_1), \ldots, Y(x_n))$ is a Gaussian vector

Mean and covariance function

- The mean function of Y is the function $m : \mathcal{X} \to \mathbb{R}$ defined by $m(x) = \mathbb{E}(Y(x))$ for $x \in \mathcal{X}$
  - Can be any function from $\mathcal{X}$ to $\mathbb{R}$
- The covariance function of Y is the function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined by $K(x_1, x_2) = \mathrm{cov}(Y(x_1), Y(x_2))$ for $x_1, x_2 \in \mathcal{X}$
  - The covariance function is symmetric non-negative definite (SNND)

SNND functions

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is SNND if

- It is symmetric : $K(x_1, x_2) = K(x_2, x_1)$ for $x_1, x_2 \in \mathcal{X}$
- For any $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix $[K(x_i, x_j)]_{i,j=1,\ldots,n}$ is symmetric non-negative definite

Existence of Gaussian processes

Theorem

For any set $\mathcal{X}$ for any function $m : \mathcal{X} \to \mathbb{R}$, and any SNND function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there exists a Gaussian process Y on $\mathcal{X}$ with mean function m and covariance function K

(Kolmogorov's extension theorem)

- In this talk we let $m = 0$
- We focus on constructing SNND functions K

Properties of covariance functions on $\mathbb{R}^d$

The most studied case for Gaussian processes is when $\mathcal{X} = \mathbb{R}^d$

Classical properties

For a covariance function K on $\mathcal{X} = \mathbb{R}^d$

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

Classical example

$K(x_1, x_2)$ is a decreasing function of $||x_1 - x_2||_2$ or $||x_1 - x_2||_1$

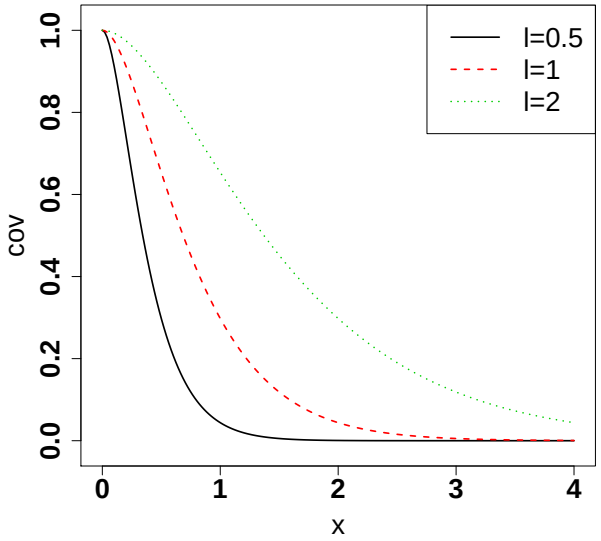
Example of the Matérn $\frac{3}{2}$ covariance function on $\mathbb{R}$

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 $\mathbb{R}^d$

Conditional distribution

Gaussian process Y with zero mean function and covariance function K observed at $x_1, ..., x_n \in \mathcal{X}$

Notation

- $y = (Y(x_1), ..., Y(x_n))^t$
- R is the $n \times n$ matrix $[K(x_i, x_j)]$
- $r(x) = (K(x, x_1), ..., K(x, x_n))^t$

Conditional mean

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

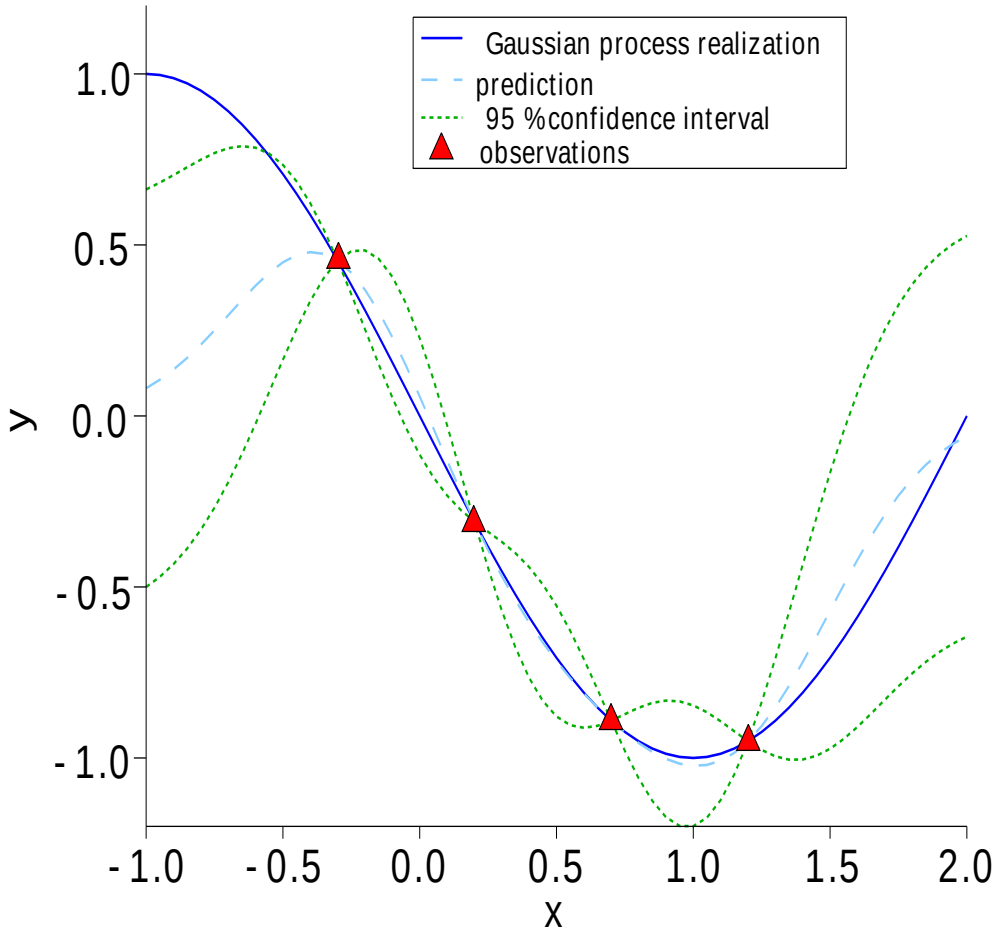
Conditional variance

The conditional variance is
$K_n(x, x) = var(Y(x)|Y(x_1), ..., Y(x_n)) = \mathbb{E}\left[(Y(x) - m_n(x))^2\right] = K(x, x) - r(x)^t R^{-1} r(x)$.

Conditional distribution

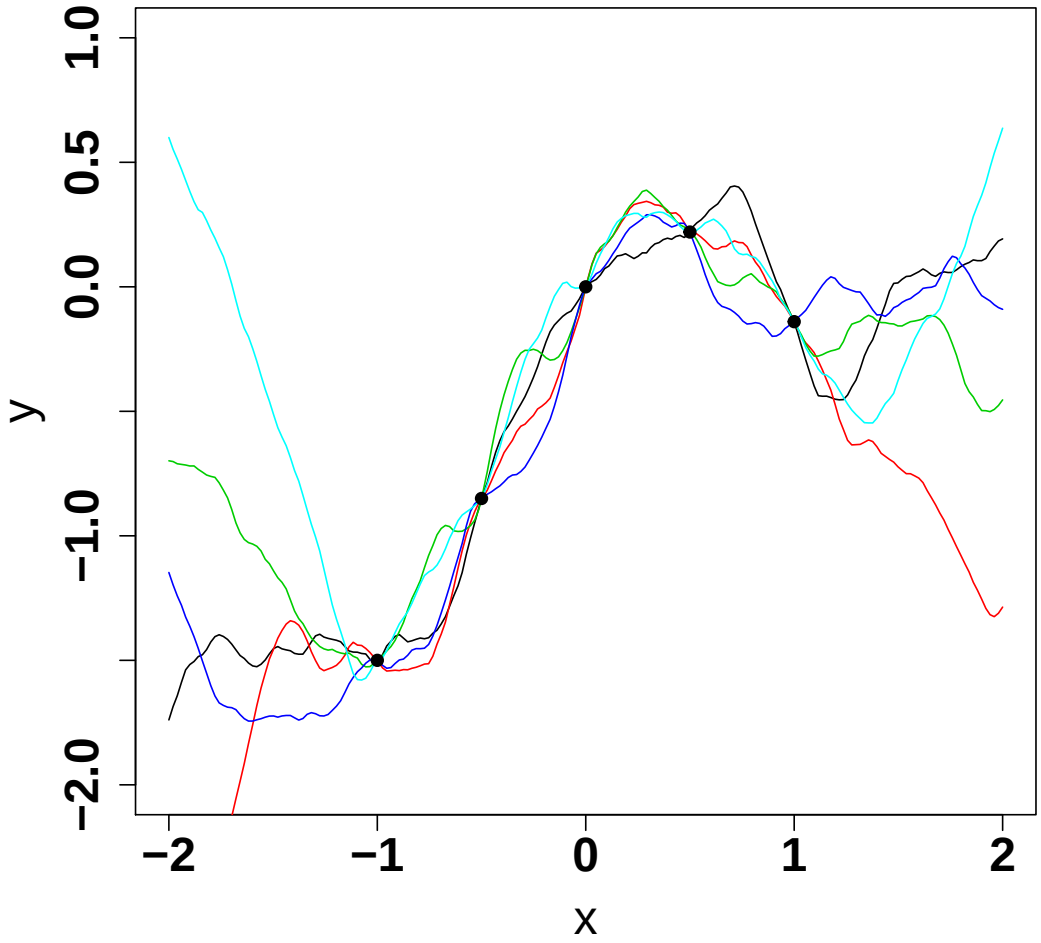
Conditionally to $Y(x_1), ..., Y(x_n)$, Y is a Gaussian process with (conditional) mean function m_n and (conditional) covariance function $(x, y) \to k_n(x, y) = k(x, y) - r(x)^t R^{-1} r(y)$

Illustration of conditional mean and variance



(Case $\mathcal{X} = [-1, 2]$)

Illustration of the conditional distribution



(Case $\mathcal{X} = [-2, 2]$)

Covariance function estimation

- 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 Y.

- θ 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 $y = (Y(x_1), ..., Y(x_n))$

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 correlation matrix of $y = (Y(x_1), ..., Y(x_n))^t$ with covariance function K_θ : $R_\theta = [K(x_i, x_j)]_{i,j=1,...,n}$.
The Maximum Likelihood estimator of θ is

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

$\Rightarrow$ Numerical optimization with $O(n^3)$ criterion
$\Rightarrow$ Most standard estimation method
$\Rightarrow$ Other estimation methods exits : empirical variogram (Book, Cressie), Cross validation (Zhang and Wang 10, Bachoc 13)

The symmetric group

Permutations

For $N \in \mathbb{N}$, a permutation of $\{1, \dots, N\}$ is a bijection $\sigma : \{1, \dots, N\} \to \{1, \dots, N\}$

Interpretation

- Let $I_1, \dots, I_N$ be a number of 'items' and let $\succ$ be a preference relation
- A permutation $\sigma : \{1, \dots, N\} \to \{1, \dots, N\}$ can be interpreted as a total ranking

$$I_{\sigma^{-1}(1)} \succ I_{\sigma^{-1}(2)} \cdots I_{\sigma^{-1}(N)}$$

in which case $\sigma(i)$ is the rank of item I_i

The symmetric group

We let S_N be the symmetric group : the set of all the permutations of $\{1, \dots, N\}$
$\Longrightarrow$ Finite set with $N!$ elements

Data set

Consider a data set of the form $(\sigma_1, y_1, \ldots, \sigma_n, y_n)$

- σ_i is a permutation in S_N for $i = 1, \ldots, n$
- $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$

Gaussian process model

We study the model

- $y_i = Y(\sigma_i)$

or

- $y_i = Y(\sigma_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \tau)$

where Y is a Gaussian process on S_N with zero mean function and covariance function K

Alternative point of view

Y is defined by a Gaussian vector of dimension $N!$ and K is defined by a $N! \times N!$ covariance matrix.

Motivations and potential applications

- Kernels (covariance functions) on S_N are studied in the statistics and machine learning literature


H. Mania, A. Ramdas, M. J. Wainwright, M. I. Jordan, and B. Recht "On kernel methods for covariates that are rankings", *Electronic Journal of Statistics, Volume 12, Number 2 (2018), 2537-2577*.


I. R. Kondor, "Group Theoretical Methods in Machine Learning", *PhD Thesis, Columbia University, New York, NY, USA, 2008.*.


Y. Jiao and J.-P. Vert, "The Kendall and Mallows kernels for permutations", *IEEE transactions on pattern analysis and machine intelligence, 2017*.

- Applications :
  - σ_i is the response of an individual to a survey and y_i is a characteristic of the individual (marketing, social sciences,...)
  - σ_i an order of task processing and y_i is the resulting performances (computer science, logistics)

Distances between permutations

For any permutations π and σ of S_n let

- The Kendall's tau distance defined by

$$d_\tau(\pi,\sigma) := \sum_{\substack{i,j=1\dots,N \\ i<j}} \left(\mathbb{1}_{\sigma(i)>\sigma(j),\ \pi(i)<\pi(j)} + \mathbb{1}_{\sigma(i)<\sigma(j),\ \pi(i)>\pi(j)}\right),$$

  that is, it counts the number of pairs on which the permutations disagree in ranking

- The Hamming distance defined by

$$d_H(\pi,\sigma) := \sum_{i=1}^{N} \mathbb{1}_{\tau(i)\neq\sigma(i)}$$

- The Spearman's footrule distance defined by

$$d_S(\pi,\sigma) := \sum_{i=1}^{N} |\tau(i)-\sigma(i)|$$

Covariance functions

- Let d be one the the three distances above
- We consider functions of the form $K_\theta : S_N \times S_N \to \mathbb{R}$ defined by

$$K_\theta(\sigma_1, \sigma_2) = \theta_2 e^{-\theta_1 (d(\sigma_1, \sigma_2))},$$

with $\theta = (\theta_1, \theta_2) \in (0, \infty)^2$

Proposition : SNND and non-degenerate

The three functions defined above are SNND and non-degenerate :
For $\sigma_1, \ldots, \sigma_n \in S_N$, two-by-two distinct, the $n \times n$ matrix $[K_\theta(\sigma_i, \sigma_j)]_{i,j=1,\ldots,n}$ is invertible

Proof : see Hania et al 2018 for Kendall's distance and our paper for the other two
Two possible proofs

- A rather short one by embedding into $\mathbb{R}^N$ or $\mathbb{R}^{(N(N-1)/2)}$
- A more technical one based on Fourier analysis on S_N
  This second proof can be extended to partial rankings, see below

Partial ranking

- Recall that we consider N items $I_1, \dots, I_N$
- A partial ranking R is a statement of the form

$$X_1 \succ X_2 \succ \dots \succ X_m$$

where $X_1, \dots, X_m$ are disjoint subsets of $\{I_1, \dots, I_N\}$
  - Relevant when N becomes large (surveys)
- To a partial ranking R we associate

$$\begin{aligned} E_R \quad := \quad & \{\sigma \in S_n : \ \sigma(i_1) < \sigma(i_2) < \dots < \sigma(i_m) \\ & \text{for any choice of } (I_{i_1}, \dots, I_{i_m}) \in X_1 \times \dots \times X_m \ \} \end{aligned}$$

(set of permutations that are in agreement with the partial ranking)

Covariance functions on partial ranking

- One possibility is the convolution kernel (Jiao and Vert 2017)
- For any covariance function K on S_N, let

$$\mathcal{K}(R, R') := \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma' \in E_{R'}} K(\sigma, \sigma')$$

- It is SNND on the set of all partial rankings (Jiao and Vert 2017)
- In our paper, we study instead

$$d_{\text{avg}}(R, R') := \frac{1}{|E_R||E_{R'}|} \sum_{\sigma \in E_R} \sum_{\sigma \in E_{R'}} d(\sigma, \sigma')$$

and let K_θ defined by

$$K_\theta(R, R') = \theta_2 e^{-\theta_1 (d_{\text{avg}}(R,R'))},$$

Proposition

The three functions obtained from d_{avg}, with the Kendall's, Hamming's and Spearman's distances are SNND

Open question : are they non-degenerate ?

Kernel renormalization

- Now $K_\theta(R, R)$ depends on R
- $K_\theta(R, R)$ can be very small when E_R contains many permutations
- We recommend to use the normalized kernel

$$K_{\text{norm},\theta}(R, R') = \theta_2 \frac{1}{\sqrt{e^{-\theta_1(d_{\text{avg}}(R,R))} e^{-\theta_1(d_{\text{avg}}(R',R'))}}} e^{-\theta_1(d_{\text{avg}}(R,R'))}$$

Computational simplifications

- The expression of $d_{\text{avg}}(R, R')$ involves a number of terms that can grows exponentially with N

Top k partial ranking

A top k partial ranking is a statement of the form

$$I_{i_1} \succ I_{i_2} \succ \cdots \succ I_{i_k} \succ X_{rest},$$

where $X_{rest} := \{I_1, \ldots, I_N\} \setminus \{I_{i_1}, \cdots, I_{i_k}\}$

- We write $I = (i_1, \cdots, i_k)$ for a top-k partial ranking

Some notation for two top k partial rankings $I := (i_1, \cdots, i_k)$ and $I' := (i'_1, \cdots, i'_k)$

- Let
$$\{j_1, \cdots, j_p\} := \{i_1, \cdots, i_k\} \cap \{i'_1, \cdots, i'_k\}$$
where $j_1 < j_2 < \cdots < j_p$
- Let, for $l = 1, \cdots p$, c_{j_l} (resp. c'_{j_l}) denotes the rank of j_l in I (resp. in I')
- Let $r := k - p$ and define $\tilde{I}$ (resp. $\tilde{I}'$) as the complementary set of $\{j_1, \cdots, j_p\}$ in $\{i_1, \cdots, i_k\}$ (resp. in $\{i'_1, \cdots, i'_k\}$)
- Writing these two sets in ascending order, we may finally define for $j = 1, \cdots, r$, u_j (resp. u'_j) as the rank in I (resp I') of the j-th element of $\tilde{I}$ (resp. $\tilde{I}'$)

Proposition

Let l and l' be two top k-partial rankings. Set $n' := n - k - 1$ and $m := n - |l \cup l'|$. Then,

$$\begin{aligned}
d_{\tau,\text{avg}}(l, l') &= \sum_{1 \le l < l' \le p} \mathbb{1}_{(c_{j_l} < c_{j_{l'}}, c'_{j_l} > c'_{j_{l'}}) \text{ or } (c_{j_l} > c_{j_{l'}}, c'_{j_l} < c'_{j_{l'}})} + r(2k + 1 - r) \\
&\quad - \sum_{j=1}^{r} (u_j + u'_j) + r^2 + \binom{n-k}{2} - \frac{1}{2}\binom{m}{2}, \\
d_{H,\text{avg}}(l, l') &= \sum_{l=1}^{p} \mathbb{1}_{c_{j_l} \neq c'_{j_l}} + m\frac{n-k-1}{n-k} + 2r, \\
d_{S,\text{avg}}(l, l') &= \sum_{l=1}^{p} |c_{j_l} - c'_{j_l}| + r(n + k + 1) - \sum_{j=1}^{r} (u_j + u'_j) \\
&\quad + mn' - \frac{mn'(2n' + 1)}{3(n' + 1)}
\end{aligned}$$

Computational simplifications

Let $\mathcal{K}^{\tau}_{\theta_1,\theta_2}$, $\mathcal{K}^{H}_{\theta_1,\theta_2}$ and $\mathcal{K}^{S}_{\theta_1,\theta_2}$ be the covariance functions on partial rankings based on the averaged Kendall, Hamming and Spearman's distances

Corollary

Let l be a k-top partial ranking. Then,

$$\begin{aligned}
\mathcal{K}^{\tau}_{\theta_1,\theta_2}(l,l) &= \theta_2 \exp\left(-\frac{\theta_1}{2}\binom{n-k}{2}\right)\\
\mathcal{K}^{H}_{\theta_1,\theta_2}(l,l) &= \theta_2 \exp\left(-\theta_1(n-k-1)\right)\\
\mathcal{K}^{S}_{\theta_1,\theta_2}(l,l) &= \theta_2 \exp\left(-\theta_1\left[(n-k)(n-k-1) - \frac{(n-k-1)(2n-2k-1)}{3}\right]\right)
\end{aligned}$$

(In the paper, we also provide simplifications for the Hamming distance, when the two partial rankings have same numbers of sets and same set cardinalities)

Asymptotic setting

- For each $n \in \mathbb{N}$ we consider $N_n \in \mathbb{N}$
- We consider a Gaussian process Y on S_{N_n} with covariance function K_{θ_0} defined by

$$K_{\theta_0}(\sigma_1, \sigma_2) = \theta_{0,2} e^{-\theta_{0,1}(d(\sigma_1,\sigma_2))} + \theta_{0,3} \mathbf{1}_{\sigma_1=\sigma_2},$$

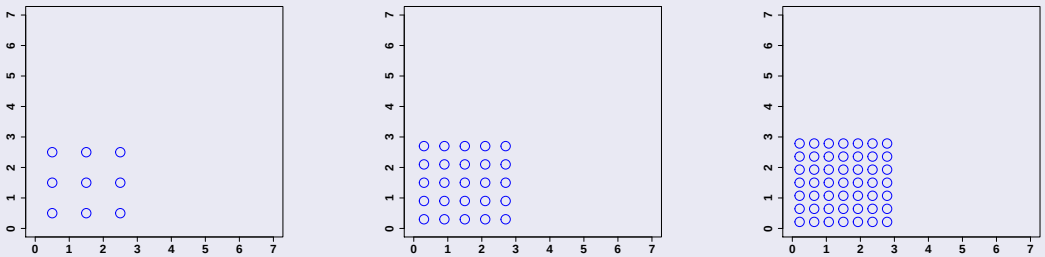
  with $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \theta_{0,3}) \in (0, \infty)^3$
- Kendall's, Hamming's and Spearman's distances
- We consider n permutations $\sigma_1, \ldots, \sigma_n \in S_{N_n}$
- We consider the parametric model of covariance functions $K_\theta; \theta \in \Theta$, for $\Theta = \prod_{i=1,2,3}[\theta_{\min,i}, \theta_{\max,i}] \subset (0, \infty)^3$ and with K_θ defined as K_{θ_0}
- We assume $\theta_0 \in \Theta$
- We aim at studying the consistency and asymptotic normality of the ML estimator $\hat{\theta}_{ML}$
- We need $N_n \to \infty$ as $n \to \infty$ for consistency to be possible

Two asymptotic frameworks for covariance parameter estimation on $\mathbb{R}^d$

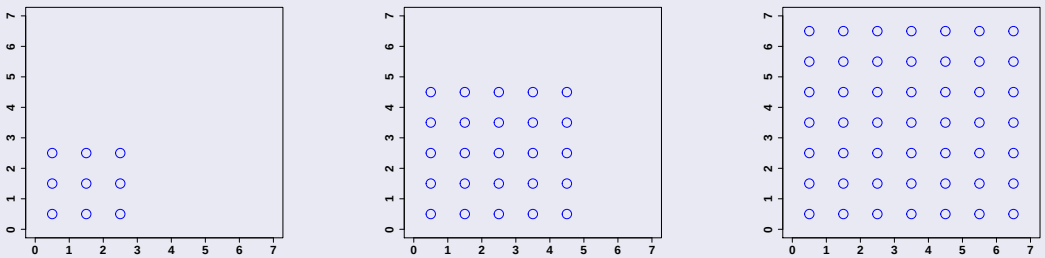
- Asymptotics for Gaussian processes on $\mathbb{R}^d$ is an active area of research
- There are several asymptotic frameworks because they are several possible location patterns for the observation points

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 asymptotic setting : mimicking expansion-domain asymptotics on $\mathbb{R}$

Observation assumption :

1. Condition 1 : There exists $\beta > 0$ such that $\forall i, j = 1, \ldots, n$, $d(\sigma_i, \sigma_j) \geq |i - j|^\beta$.
2. Condition 2 : There exists $c > 0$ such that $\forall i, j = 1, \ldots, n$ $d(\sigma_i, \sigma_{i+1}) \leq c$.

Asymptotic results for estimation

Theorem

Under Conditions 1 and 2, we get

$$\widehat{\theta}_{ML} \xrightarrow[n\to+\infty]{p} \theta_0$$

Theorem

Let M_{ML} be the 3×3 matrix defined by

$$(M_{ML})_{i,j} = \frac{1}{2n} \mathrm{Tr}\left(R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_i} R_{\theta_0}^{-1} \frac{\partial R_{\theta_0}}{\partial \theta_j} \right)$$

Then

$$\sqrt{n} M_{ML}^{\frac{1}{2}} \left(\widehat{\theta}_{ML} - \theta_0 \right) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, I_3)$$

Furthermore,

$$0 < \liminf_{n\to\infty} \lambda_{\min}(M_{ML}) \leq \limsup_{n\to\infty} \lambda_{\max}(M_{ML}) < +\infty$$

Extends existing increasing domain asymptotic results on $\mathbb{R}^d$ (Bachoc 14) and for distribution inputs (Bachoc 17) by showing specific local and global identifiability conditions

Asymptotic Impact on prediction

- Let $\hat{Y}_{\theta,n}(\sigma)$ be the conditional expectation of $Y(\sigma)$ given $Y(\sigma_1), \ldots, Y(\sigma_n)$ under covariance function K_θ

Theorem

$$\forall N \in \mathbb{N},\ \forall \sigma_N \in S_N,\quad \left| \widehat{Y}_{\widehat{\theta}_{ML}}(\sigma_{N_n}) - \widehat{Y}_{\theta_o}(\sigma_{N_n}) \right| = o_p(1)$$

where σ_{N_n} is the extension of σ_N from S_N to S_{N_n} for $N_n \geq N$

Somme numerical illustrations

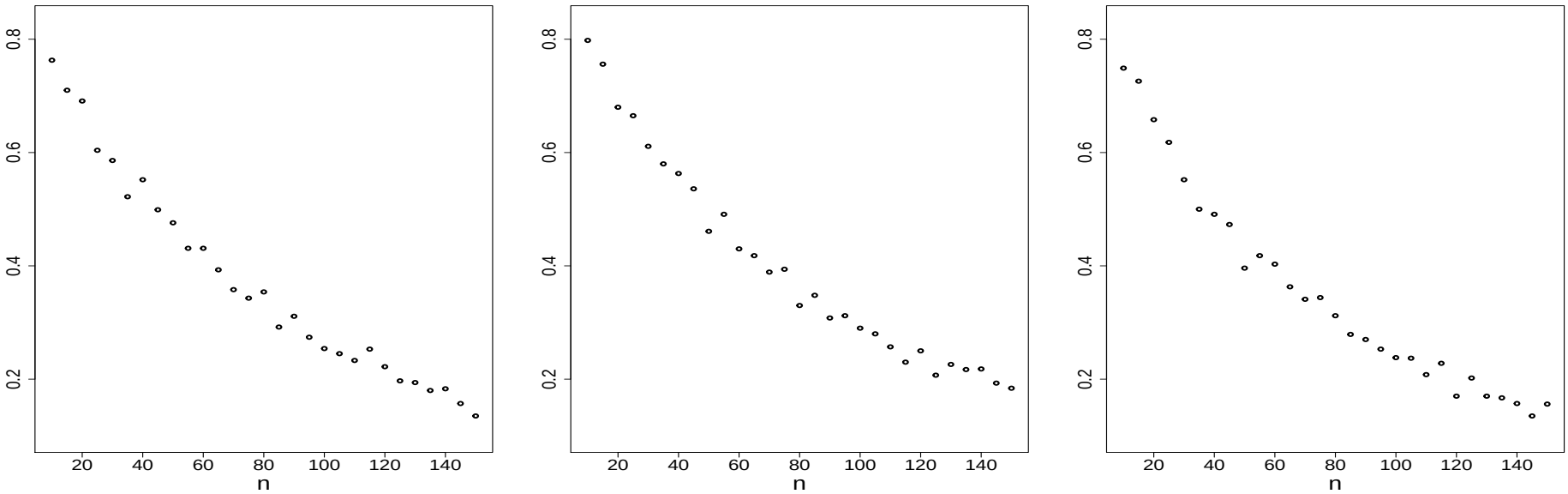


FIGURE: Estimates of $P(\|\hat{\theta}_{ML} - \theta_0\| > 0.5)$ for different values of n, the number of observations, with $\theta_0 = (0.1, 0.8, 0.3)$ and Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right

Somme numerical illustrations

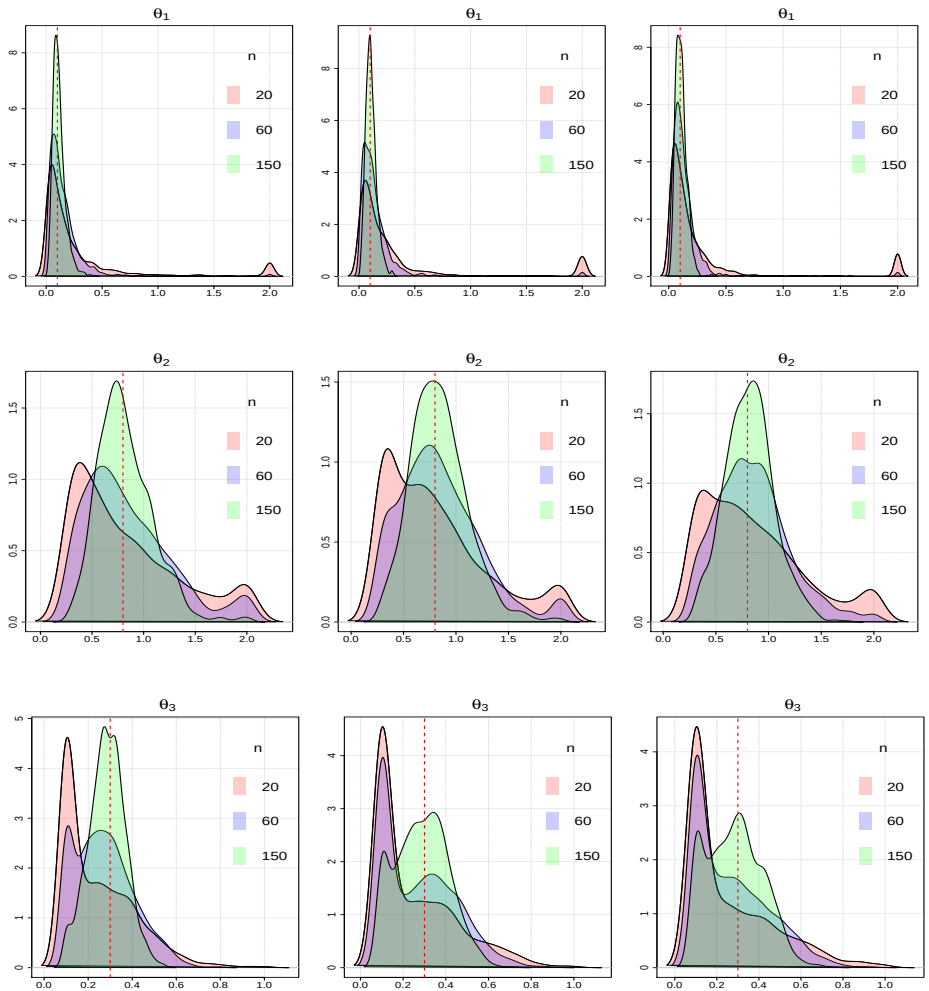


FIGURE: Density of the coordinates of $\hat{\theta}_{ML}$ for the number of observations $n = 20$ (in red), $n = 60$ (in blue), $n = 150$ (in green) with $\theta_0 = (0.1, 0.8, 0.3)$ (represented by the red vertical line). We used the Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right

Somme numerical illustrations

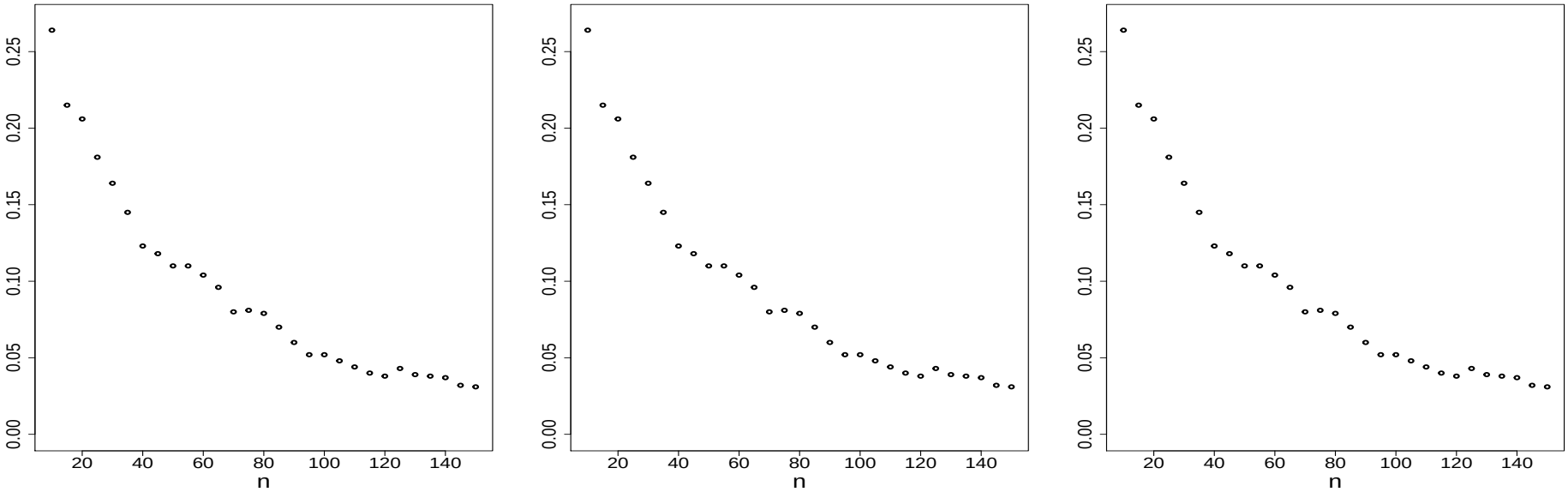


FIGURE: Estimates of $P\left(\left|\widehat{Y}_{\widehat{\theta}_{ML},n}(\sigma) - \widehat{Y}_{\theta_0}(\sigma)\right| > 0.3\right)$ for different values of n, the number of observations, with $\theta_0 = (0.1, 0.8, 0.3)$ and the Kendall's tau distance, the Hamming distance and the Spearman's footrule distance from left to right.

Conclusion and open questions

Conclusion

- Covariance functions on permutations are provided
- Extension to partial rankings and computational simplifications
- The asymptotic results of the Euclidean case can be extended

Open questions

- Application where input σ can be selected and sequential designs (e.g. optimization)
- Asymptotic settings where $\sigma_1, \ldots, \sigma_n$ are independent and uniformly distributed on S_{N_n} with sequence N_n carefully selected so that $d(\sigma_1, \{\sigma_2, \ldots, \sigma_n\})$ does not go to infinity or zero $\Longrightarrow$ more natural expansion domain setting

The preprint :

F. Bachoc, B. Broto, F. Gamboa and J.M. Loubes Gaussian Processes indexed on the symmetric group : prediction and learning, *arxiv.org/abs/1803.06118v3*.

Thank you for your attention !