Gaussian processes indexed on the symmetric group: prediction and learning

François Bachoc\textsuperscript{1}, Baptiste Broto\textsuperscript{1,2}, Fabrice Gamboa\textsuperscript{1}, Jean-Michel Loubes\textsuperscript{1}

\textsuperscript{1}Institut de Mathématiques de Toulouse
\textsuperscript{2}CEA Saclay

Saint-Etienne - 2018
1. Gaussian processes

2. Covariance function estimation and prediction

3. Construction of covariance functions on the symmetric group

4. Extension to partial rankings

5. Asymptotic results
Gaussian process regression (Kriging model)

Study of a single realization of a Gaussian process \( x \rightarrow Y(x) \) 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

Definition

A stochastic process $Y: \mathcal{X} \rightarrow \mathbb{R}$ 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 $m: \mathcal{X} \rightarrow \mathbb{R}$ 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 $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ defined by $K(x_1, x_2) = \text{cov}(Y(x_1), Y(x_2))$ for $x_1, x_2 \in \mathcal{X}$.
  - The covariance function is symmetric non-negative definite (SNND).

SNND functions

A function $K: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 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 $[K(x_i, x_j)]_{i,j=1}^{n}$ 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 $K$
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 $||x_1 - x_2||_1$
The Matérn $\frac{\alpha}{2}$ 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
- $\sigma^2$ corresponds to the order of magnitude of the functions that are realizations of the Gaussian process
- $\ell$ corresponds to the speed of variation of the functions that are realizations of the Gaussian process

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

2 Covariance function estimation and prediction

3 Construction of covariance functions on the symmetric group

4 Extension to partial rankings

5 Asymptotic results
Conditional distribution

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

**Notation**

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

**Conditional mean**

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

**Conditional variance**

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

**Conditional distribution**

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

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

**Estimation**
Objective: build estimator $\hat{\theta}(y)$
Explicit Gaussian likelihood function for the observation vector $y$

**Maximum Likelihood**

Define $R_\theta$ as the correlation matrix of $y = (Y(x_1), ..., Y(x_n))^t$ with covariance function $K_\theta$:

$$R_\theta = [K(x_i, x_j)]_{i,j=1,...,n}.$$ 

The Maximum Likelihood estimator of $\theta$ is

$$\hat{\theta}_{ML} \in \arg\min_{\theta \in \Theta} \frac{1}{n} \left( \ln (|R_\theta|) + y^tR_\theta^{-1}y \right)$$

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

2 Covariance function estimation and prediction

3 Construction of covariance functions on the symmetric group

4 Extension to partial rankings

5 Asymptotic results
**Permutations**

For \( N \in \mathbb{N} \), a permutation of \( \{1, \ldots, N\} \) is a bijection \( \sigma : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\} \).

**Interpretation**

- Let \( l_1, \ldots, l_N \) be a number of ‘items’ and let \( \succ \) be a preference relation.
- A permutation \( \sigma : \{1, \ldots, N\} \rightarrow \{1, \ldots, N\} \) can be interpreted as a total ranking
  \[
  l_{\sigma^{-1}(1)} \succ l_{\sigma^{-1}(2)} \cdots l_{\sigma^{-1}(N)}
  \]
  in which case \( \sigma(i) \) is the rank of item \( l_i \).

**The symmetric group**

We let \( S_N \) be the symmetric group: the set of all the permutations of \( \{1, \ldots, N\} \)

\( \implies \) Finite set with \( N! \) elements
Data set

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

- \(\sigma_i\) 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

- Applications:
  - $\sigma_i$ is the response of an individual to a survey and $y_i$ is a characteristic of the individual (marketing, social sciences,
  - $\sigma_i$ an order of task processing and $y_i$ is the resulting performances (computer science, logistics)
For any permutations $\pi$ and $\sigma$ of $S_n$ let

- The **Kendall's tau distance** defined by
  
  $$
  d_{\tau}(\pi, \sigma) := \sum_{\substack{i,j=1,\ldots,N \atop i<j}} \left( 1_{\sigma(i) > \sigma(j), \ \pi(i) < \pi(j)} + 1_{\sigma(i) < \sigma(j), \ \pi(i) > \pi(j)} \right),
  $$

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

- The **Hamming distance** defined by
  
  $$
  d_{H}(\pi, \sigma) := \sum_{i=1}^{N} 1_{\tau(i) \neq \sigma(i)}
  $$

- 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 of 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.
1. Gaussian processes
2. Covariance function estimation and prediction
3. Construction of covariance functions on the symmetric group
4. Extension to partial rankings
5. Asymptotic results
Recall that we consider \( N \) items \( I_1, \ldots, I_N \).

A partial ranking \( R \) is a statement of the form

\[
X_1 \succ X_2 \succ \ldots \succ X_m
\]

where \( X_1, \ldots, X_m \) are disjoint subsets of \( \{I_1, \ldots, I_N\} \).

- Relevant when \( N \) becomes large (surveys).

To a partial ranking \( R \) we associate

\[
E_R := \{ \sigma \in S_n : \sigma(i_1) < \sigma(i_2) < \cdots < \sigma(i_m) \}
\]

for any choice of \( (I_{i_1}, \ldots, I_{i_m}) \in X_1 \times \cdots \times X_m \)

(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

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

- 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 $K_\theta$ defined by

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

**Proposition**

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

**Open question**: are they non-degenerate?
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

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

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

Top $k$ partial ranking

A top $k$ partial ranking is a statement of the form

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

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

- 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 $\bar{I}$ (resp. $\bar{I}'$) 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 $\bar{I}$ (resp. $\bar{I}'$)
Proposition

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

\[
d_{r, \text{avg}}(I, I') = \sum_{1 \leq l < l' \leq p} \mathbb{1}(c_{jl} < c_{jl}', c_{jl}' > c_{jl}) + r(2k + 1 - r)
\]

\[
- \sum_{j=1}^{r} (u_j + u'_j) + r^2 + \binom{n-k}{2} - \frac{1}{2} \binom{m}{2},
\]

\[
d_{H, \text{avg}}(I, I') = \sum_{l=1}^{p} \mathbb{1}_{c_{jl} \neq c_{jl}'} + m \frac{n - k - 1}{n - k} + 2r,
\]

\[
d_{S, \text{avg}}(I, I') = \sum_{l=1}^{p} |c_{lj} - c_{lj}'| + r(n + k + 1) - \sum_{j=1}^{r} (u_j + u'_j)
\]

\[
+ mn' - \frac{mn'(2n' + 1)}{3(n' + 1)},
\]
Let $K_{\tau_{\theta_1,\theta_2}}$, $K_{H_{\theta_1,\theta_2}}$ and $K_{S_{\theta_1,\theta_2}}$ be the covariance functions on partial rankings based on the averaged Kendall, Hamming and Spearman’s distances.

**Corollary**

Let $I$ be a $k$-top partial ranking. Then,

$$
K_{\tau_{\theta_1,\theta_2}}(I, I) = \theta_2 \exp \left( -\frac{\theta_1}{2} \binom{n-k}{2} \right)
$$

$$
K_{H_{\theta_1,\theta_2}}(I, I) = \theta_2 \exp (-\theta_1 (n-k-1))
$$

$$
K_{S_{\theta_1,\theta_2}}(I, I) = \theta_2 \exp \left( -\theta_1 \left[ (n-k)(n-k-1) - \frac{(n-k-1)(2n-2k-1)}{3} \right] \right)
$$

(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)
1 Gaussian processes

2 Covariance function estimation and prediction

3 Construction of covariance functions on the symmetric group

4 Extension to partial rankings

5 Asymptotic results
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 $K_{\theta_0}$ defined by

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

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 $K_\theta$ defined as $K_{\theta_0}$

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 $\rightarrow$ unbounded observation domain.
Existing fixed-domain asymptotic results

- From 80’-90’ and onward. Fruitful theory for interaction estimation-prediction.

- Consistent estimation is **impossible** for some covariance parameters (identifiable in finite-sample), see e.g.

- 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

  


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

\[ \hat{\theta}_{ML} \xrightarrow{p} \theta_0 \]

\( n \rightarrow +\infty \)

**Theorem**

Let \( M_{ML} \) be the \( 3 \times 3 \) matrix defined by

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

Then

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

Furthermore,

\[ 0 < \lim \inf_{n \rightarrow \infty} \lambda_{\min}(M_{ML}) \leq \lim \sup_{n \rightarrow \infty} \lambda_{\max}(M_{ML}) < +\infty \]

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 \( K_\theta \)

**Theorem**

\[
\forall N \in \mathbb{N}, \ \forall \sigma_N \in S_N, \quad \left| \hat{Y}_{\theta,ML}(\sigma_{N_n}) - \hat{Y}_{\theta,o}(\sigma_{N_n}) \right| = o_p(1)
\]

where \( \sigma_{N_n} \) is the extension of \( \sigma_N \) from \( S_N \) to \( S_{N_n} \) for \( N_n \geq N \)
\textbf{Figure:} Estimates of $P(\|\hat{\theta}_{ML} - \theta_0\| > 0.5)$ for different values of $n$, the number of observations, with $	heta_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.
**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.
Figure: Estimates of $P \left( \left| \hat{Y}_{\hat{\theta}_{ML}, n} (\sigma) - \hat{Y}_{\theta_0} (\sigma) \right| > 0.3 \right)$ 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 $\sigma$ 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 $\implies$ more natural expansion domain setting

The preprint:


Thank you for your attention !