

Patched Gaussian Processes

Tom Perrin, Leo Roux & Thibault Lambert

Option Géostatistique, Mines Paris - PSL

`tom.perrin@etu.minesparis.psl.eu`

`leo.roux@etu.minesparis.psl.eu`

`thibault.lambert@etu.minesparis.psl.eu`



■ TABLE OF CONTENTS

I. Prerequisites

II. The "Big N " Problem and Beyond

III. Methods to Tackle the Issue

III.1 Patchwork Kriging (Park and Aley)

III.2 Nearest-Neighbors Gaussian Processes (Datta et al.)

IV. Performance Comparison

V. Conclusion

■ GAUSSIAN PROCESS

Gaussian Process : Random surface $W(s)$ over a domain \mathcal{D} such that :

$$\forall \{s_1, \dots, s_n\} \in \mathcal{D}^n, W := (W(s_1), \dots, W(s_n)) \sim \mathcal{N}(\mu, C).$$

- **Mean function** : $\mu(s) := \mathbb{E}[W(s)] \longrightarrow \mu := [\mu(s_1), \dots, \mu(s_N)]^T$
- **Covariance function** : $C(s, s') := \mathbb{Cov}[W(s), W(s')] \longrightarrow C := [C(s_i, s_j)]_{1 \leq i, j \leq N}$.

Law of W :

$$p(W) = \frac{1}{\sqrt{(2\pi)^d \det(C)}} \exp \left(-\frac{1}{2} (W - \mu)^T C^{-1} (W - \mu) \right).$$

■ KRIGING

Observations : $\mathcal{D} = \{(x_i, y_i) \mid i \in \{1, \dots, N\}\}$.

Locations $x := [x_1, \dots, x_N]^T \longrightarrow$ **Responses** $y := [y_1, \dots, y_N]^T$.

Kriging : Stochastic predictions based on these observations.

- **Reponses** y seen as realizations of a **gaussian process** f with noise :

$$y_i = f(x_i) + \varepsilon_i, \text{ where } \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

- **Prediction** $f_* := f(x^*)$ at new location $x^* \longrightarrow$ **Joint law of** (f_*, y) :

$$p(f_*, y) \sim \mathcal{N} \left(0, \begin{bmatrix} c_{**} & c_{*x} \\ c_{x*} & \sigma^2 I + C_{xx} \end{bmatrix} \right).$$

where $c_{**} = C(x^*, x^*)$, $c_{x*} = [C(x_1, x^*), \dots, C(x_N, x^*)]^T$, $C_{xx} = [C(x_i, x_j)]_{1 \leq i, j \leq N}$.

Predictive distribution :

$$p(f_* \mid y) \sim \mathcal{N}(\underbrace{c_{x*}^T (\sigma^2 I + C_{xx})^{-1} y}_{\text{predictive mean}}, \underbrace{c_{**} - c_{x*}^T (\sigma^2 I + C_{xx})^{-1} c_{x*}}_{\text{predictive variance}})$$

- **Predicted value :** $f_*(x^*) \approx c_{x*}^T (\sigma^2 I + C_{xx})^{-1} y.$
- **Uncertainty :** $c_{**} - c_{x*}^T (\sigma^2 I + C_{xx})^{-1} c_{x*}.$

■ BAYESIAN APPROACH

Hierarchical model :

$$y_i = X(x_i) \cdot \beta + f_\theta(x_i) + \varepsilon_i.$$

- **Regression term** $X(x_i) \cdot \beta$: Captures effects of known explanatory variables X .
- **Spatial effect** $f_\theta(x_i)$: Realization of a gaussian process $f_\theta \sim \mathcal{N}(0, C_\theta)$.
- **Noise** ε_i : $\varepsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, \tau^2)$.

■ BAYESIAN APPROACH

Bayesian Inference : Estimation of the **hidden state** f_θ and **parameters** $\phi := \{\beta, \theta, \tau^2\}$.

- **Prior distributions** : We assume prior knowledge $p(\beta)$, $p(\theta)$ and $p(\tau^2)$.
- **Likelihood** : $p(y \mid f_\theta, \beta, \tau^2) = \mathcal{N}(X\beta + f_\theta, \tau^2 I)$.
- **Latent GP prior** : $p(f_\theta \mid \theta) = \mathcal{N}(0, C_\theta)$.

Joint posterior distribution :

$$p(f_\theta, \beta, \theta, \tau^2 \mid y) \propto \underbrace{p(y \mid f_\theta, \beta, \tau^2)}_{\text{data likelihood}} \times \underbrace{p(f_\theta \mid \theta)}_{\text{spatial link}} \times \underbrace{p(\beta)p(\theta)p(\tau^2)}_{\text{parameter priors}}.$$

■ TABLE OF CONTENTS

I. Prerequisites

II. The "Big N " Problem and Beyond

III. Methods to Tackle the Issue

III.1 Patchwork Kriging (Park and Aley)

III.2 Nearest-Neighbors Gaussian Processes (Datta et al.)

IV. Performance Comparison

V. Conclusion

■ COMPUTATIONAL COSTS

Both the predictive mean and variance require **solving linear systems** involving $\Sigma = \sigma^2 I + C_{xx}$.

- **Time complexity** : Matrix inversion $\longrightarrow \mathcal{O}(N^3)$.
- **Space complexity** : Storing covariance matrix $\longrightarrow \mathcal{O}(N^2)$.

Solutions :

- Acting on the **covariance matrix** Σ : compact support, covariance tapering, markovian models \longrightarrow **Sparse matrix**.
- Acting on the **amount of observations** N ...

■ INDEPENDANT LOCAL KRIGING

Idea : Split the observations \mathcal{D} into K subsets $\mathcal{D}_1, \dots, \mathcal{D}_K$, with $\mathcal{D}_k := \{(x_i, y_i) \mid x_i \in \Omega_k\}$.

Region $\Omega_k \longrightarrow$ Local gaussian process $f_k \longrightarrow$ Associated covariance function $C_k(\cdot, \cdot)$.

- **Stationary process** : $C_k(\cdot, \cdot) = C(\cdot, \cdot)$.
- Otherwise : Different covariance functions C_k .

$$y_{k,i} = f_k(x_i) + \varepsilon_{k,i}.$$

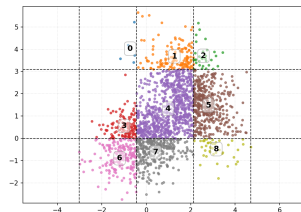
Two questions :

- How to split the observations efficiently ?
- How to deal with shared boundaries $\Gamma_{k,\ell} := \overline{\Omega}_k \cap \overline{\Omega}_\ell$?

■ SPLITTING THE OBSERVATIONS

Grid partitioning : Splitting the data using an uniform grid to cover the space.

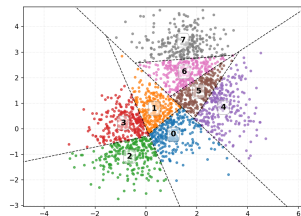
- ⊕ Very easy implementation.
- ⊖ High density variance between regions.



Grid partitioning

PCA partitioning : Splitting the data based on principal component projections values.

- ⊕ Balanced number of points ($N_k \approx \text{cst}$).
- ⊖ Complex boundary definition.



PCA partitioning

■ BORDER DISCONTINUITY

Issue :

$$f_k(x) \neq f_\ell(x) \text{ at frontier } \Gamma_{k,\ell}.$$

Solution :

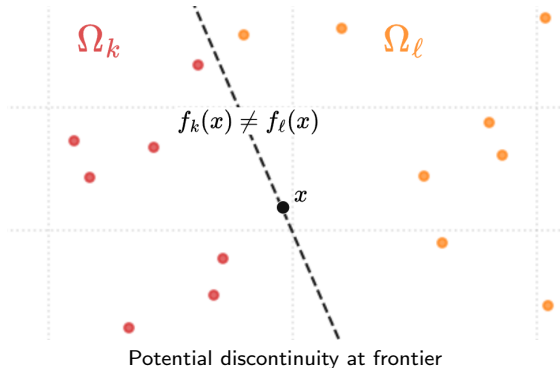
$$\text{Set } \delta_{k,\ell} := f_k - f_\ell = 0 \text{ at } \Gamma_{k,\ell}.$$

In practice : pseudo-observations

$$\delta_{k,\ell}(x) = 0 \text{ at the frontiers.}$$

$$\mathbb{E} \left[f_*^{(k)} \mid y \right] \longrightarrow \mathbb{E} \left[f_*^{(k)} \mid y, \delta = 0 \right].$$

$$\mathbb{V} \left[f_*^{(k)} \mid y \right] \longrightarrow \mathbb{V} \left[f_*^{(k)} \mid y, \delta = 0 \right].$$



■ TABLE OF CONTENTS

I. Prerequisites

II. The "Big N " Problem and Beyond

III. Methods to Tackle the Issue

III.1 Patchwork Kriging (Park and Aley)

III.2 Nearest-Neighbors Gaussian Processes (Datta et al.)

IV. Performance Comparison

V. Conclusion

■ TABLE OF CONTENTS

I. Prerequisites

II. The "Big N " Problem and Beyond

III. Methods to Tackle the Issue

III.1 Patchwork Kriging (Park and Aley)

III.2 Nearest-Neighbors Gaussian Processes (Datta et al.)

IV. Performance Comparison

V. Conclusion

■ PSEUDO-OBSERVATIONS

For two neighbor regions Ω_k, Ω_ℓ , place B observations on the border $\Gamma_{k,\ell}$:

- **Pseudo-locations** : $x^{(k,\ell)} := \left(x_1^{(k,\ell)}, \dots, x_B^{(k,\ell)} \right)^T$.
- **Pseudo-values** : $\delta_{k,\ell} := \left(\delta_{k,\ell}(x_1^{(k,\ell)}), \dots, \delta_{k,\ell}(x_B^{(k,\ell)}) \right)^T$.
- **Observations** : $y := \left(y_1^T, \dots, y_K^T \right)^T$.
- **Pseudo-observations** : $\delta := \left(\delta_{1,1}^T, \dots, \delta_{1,K}^T, \dots, \delta_{K,K}^T \right)^T$.

■ PATCHWORK KRIGING

Predict $f_*^{(k)} := f_k(x^*)$ at $x^* \in \Omega_k \longrightarrow$ **Joint law of $(f_*^{(k)}, y, \delta)$** :

$$\begin{bmatrix} f_*^{(k)} \\ y \\ \delta \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} c_{**} & c_{*\mathcal{D}}^{(k)} & c_{*\delta}^{(k)} \\ c_{\mathcal{D}*}^{(k)} & C_{\mathcal{D}\mathcal{D}} & C_{\mathcal{D}\delta} \\ c_{\delta*}^{(k)} & C_{\delta\mathcal{D}} & C_{\delta\delta} \end{bmatrix} \right).$$

■ PATCHWORK KRIGING

Definitions of the covariance blocks :

$$c_{**} := \mathbb{Cov}(f_*^{(k)}, f_*^{(k)}),$$

$$c_{*\mathcal{D}}^{(k)} := \left(\mathbb{Cov}(f_*^{(k)}, y_1), \dots, \mathbb{Cov}(f_*^{(k)}, y_K) \right),$$

$$c_{*\delta}^{(k)} := \left(\mathbb{Cov}(f_*^{(k)}, \delta_{1,1}), \dots, \mathbb{Cov}(f_*^{(k)}, \delta_{K,K}) \right),$$

$$C_{\mathcal{D}\mathcal{D}} := \begin{bmatrix} \mathbb{Cov}(y_1, y_1) & \dots & \mathbb{Cov}(y_1, y_K) \\ \dots & \ddots & \dots \\ \mathbb{Cov}(y_K, y_1) & \dots & \mathbb{Cov}(y_K, y_K) \end{bmatrix}, \quad C_{\mathcal{D}\delta} := \begin{bmatrix} \mathbb{Cov}(y_1, \delta_{1,1}) & \dots & \mathbb{Cov}(y_1, \delta_{K,K}) \\ \dots & \ddots & \dots \\ \mathbb{Cov}(y_K, \delta_{1,1}) & \dots & \mathbb{Cov}(y_K, \delta_{K,K}) \end{bmatrix} \quad \text{and}$$

$$C_{\delta\delta} := \begin{bmatrix} \mathbb{Cov}(\delta_{1,1}, \delta_{1,1}) & \dots & \mathbb{Cov}(\delta_{1,1}, \delta_{K,K}) \\ \dots & \ddots & \dots \\ \mathbb{Cov}(\delta_{K,K}, \delta_{1,1}) & \dots & \mathbb{Cov}(\delta_{K,K}, \delta_{K,K}) \end{bmatrix}.$$

■ PATCHWORK KRIGING

Predictive mean :

$$\mathbb{E} \left[f_*^{(k)} \mid y, \delta = 0 \right] = \left(c_{*\mathcal{D}}^{(k)} - c_{*\delta}^{(k)} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right) \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} y.$$

Predictive variance :

$$\begin{aligned} \mathbb{V} \left[f_*^{(k)} \mid y, \delta = 0 \right] = & c_{**} - c_{*\delta}^{(k)} C_{\delta\delta}^{-1} c_{\delta*}^{(k)} \\ & - \left(c_{*\mathcal{D}}^{(k)} - c_{*\delta}^{(k)} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right) \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} \left(c_{\mathcal{D}*}^{(k)} - c_{\delta*}^{(k)} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right) \end{aligned}$$

Noting $Q := \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1}$, $v := L^{-1} C_{\delta\mathcal{D}}$ and $w_* := L^{-1} c_{\delta*}^{(k)}$ (where $C_{\delta\delta} = LL^T$),

$$\begin{aligned} \mathbb{E} \left[f_*^{(k)} \mid y, \delta = 0 \right] &= (c_{*\mathcal{D}}^{(k)} - w_*^T v) Q y. \\ \mathbb{V} \left[f_*^{(k)} \mid y, \delta = 0 \right] &= c_{**} - w_*^T w_* - (c_{*\mathcal{D}}^{(k)} - w_*^T v) Q (c_{*\mathcal{D}}^{(k)} - w_*^T v)^T. \end{aligned}$$

■ PROOF

For two gaussian vectors $\mathbf{x}_1 \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $\mathbf{x}_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$, the conditional distribution $p(\mathbf{x}_1 | \mathbf{x}_2)$ is gaussian and verifies :

$$\mathbb{E}[\mathbf{x}_1 | \mathbf{x}_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2).$$

$$\mathbb{V}[\mathbf{x}_1 | \mathbf{x}_2] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Applying this to $\mathbf{x}_1 = f_*^{(k)}$ and $\mathbf{x}_2 = (y, \delta)$ gives :

$$\mathbb{E}[f_*^{(k)} | y, \delta] = [c_{*\mathcal{D}}^{(k)}, c_{*\delta}^{(k)}] \begin{bmatrix} C_{\mathcal{D}\mathcal{D}} & C_{\mathcal{D}\delta} \\ C_{\delta\mathcal{D}} & C_{\delta\delta} \end{bmatrix}^{-1} \begin{bmatrix} y \\ \delta \end{bmatrix}$$

and

$$\mathbb{V}[f_*^{(k)} | y, \delta] = c_{**} - [c_{*\mathcal{D}}^{(k)}, c_{*\delta}^{(k)}] \begin{bmatrix} C_{\mathcal{D}\mathcal{D}} & C_{\mathcal{D}\delta} \\ C_{\delta\mathcal{D}} & C_{\delta\delta} \end{bmatrix}^{-1} \begin{bmatrix} c_{\mathcal{D}*}^{(k)} \\ c_{\delta*}^{(k)} \end{bmatrix}.$$

■ PROOF

For invertible matrix A, B, D , the following inversion formula holds true :

$$\begin{bmatrix} A & B \\ B^T & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}B^T)^{-1} & -(A - BD^{-1}B^T)^{-1}BD^{-1} \\ -D^{-1}B^T(A - BD^{-1}B^T)^{-1} & (D - B^TA^{-1}B)^{-1} \end{bmatrix}.$$

Applying this to $A = C_{\mathcal{D}\mathcal{D}}, B = C_{\mathcal{D}\delta}, D = C_{\delta\delta}$ gives, for the predictive mean :

$$\begin{aligned} \mathbb{E} \left[f_*^{(k)} \mid y, \delta \right] &= \begin{bmatrix} c_{*\mathcal{D}}^{(k)} \\ c_{*\delta}^{(k)} \end{bmatrix}^T \begin{bmatrix} (C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta}C_{\delta\delta}^{-1}C_{\delta\mathcal{D}})^{-1} & -(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta}C_{\delta\delta}^{-1}C_{\delta\mathcal{D}})^{-1}C_{\mathcal{D}\delta}C_{\delta\delta}^{-1} \\ -C_{\delta\delta}^{-1}C_{\delta\mathcal{D}}(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta}C_{\delta\delta}^{-1}C_{\delta\mathcal{D}})^{-1} & (C_{\delta\delta} - C_{\delta\mathcal{D}}C_{\mathcal{D}\mathcal{D}}^{-1}C_{\mathcal{D}\delta})^{-1} \end{bmatrix} \begin{bmatrix} y \\ \delta \end{bmatrix} \\ &= c_{*\mathcal{D}}^{(k)} (C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta}C_{\delta\delta}^{-1}C_{\delta\mathcal{D}})^{-1} y - c_{*\delta}^{(k)} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}}^T (C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta}C_{\delta\delta}^{-1}C_{\delta\mathcal{D}})^{-1} y \\ &\quad - c_{*\mathcal{D}}^{(k)} (C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta}C_{\delta\delta}^{-1}C_{\delta\mathcal{D}})^{-1} C_{\mathcal{D}\delta}C_{\delta\delta}^{-1} \delta + c_{*\delta}^{(k)} (C_{\delta\delta} - C_{\delta\mathcal{D}}C_{\mathcal{D}\mathcal{D}}^{-1}C_{\mathcal{D}\delta})^{-1} \delta. \end{aligned}$$

Taking $\delta = 0$ finally leads to :

$$\mathbb{E} \left[f_*^{(k)} \mid y, \delta = 0 \right] = (c_{*\mathcal{D}}^{(k)} - c_{*\delta}^{(k)} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}}) (C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta}C_{\delta\delta}^{-1}C_{\delta\mathcal{D}})^{-1} y.$$

PROOF

Concerning the predictive variance, the inversion formula also gives :

$$\begin{aligned}
 & \mathbb{V} \left[f_*^{(k)} \mid y, \delta \right] \\
 &= c_{**} - \begin{bmatrix} c_{*\mathcal{D}}^{(k)} \\ c_{*\delta}^{(k)} \end{bmatrix}^T \begin{bmatrix} \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} & - \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} \\ - C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} & \left(C_{\delta\delta} - C_{\delta\mathcal{D}} C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta} \right)^{-1} \end{bmatrix} \begin{bmatrix} c_{\mathcal{D}*}^{(k)} \\ c_{\delta*}^{(k)} \end{bmatrix} \\
 &= c_{**} - c_{*\mathcal{D}}^{(k)} \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} c_{\mathcal{D}*}^{(k)} + c_{*\mathcal{D}}^{(k)} \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} c_{\delta*}^{(k)} \\
 &\quad + c_{*\delta}^{(k)} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} c_{\mathcal{D}*}^{(k)} - c_{*\delta}^{(k)} \left(C_{\delta\delta} - C_{\delta\mathcal{D}} C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta} \right)^{-1} c_{\delta*}^{(k)}.
 \end{aligned}$$

This finally yields to :

$$\mathbb{V} \left[f_*^{(k)} \mid y, \delta = 0 \right] = c_{**} - c_{*\delta}^{(k)} C_{\delta\delta}^{-1} c_{\delta*}^{(k)} - \left(c_{*\mathcal{D}}^{(k)} - c_{*\delta}^{(k)} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right) \left(C_{\mathcal{D}\mathcal{D}} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} C_{\delta\mathcal{D}} \right)^{-1} \left(c_{\mathcal{D}*}^{(k)} - C_{\mathcal{D}\delta} C_{\delta\delta}^{-1} c_{\delta*}^{(k)} \right).$$

■ CHOICE OF THE PARAMETERS

Covariance function :

- Different for each region $\longrightarrow c_k(\cdot, \cdot)$ for $k \in \{1, \dots, K\}$.
- Obtained by minimizing the negative log-likelihood :

$$NL(\theta) := -\log p(y, \delta = 0 \mid \theta)$$

$$= \frac{N}{2} \log(2\pi) + \frac{1}{2} \log \begin{vmatrix} C_{\mathcal{D}\mathcal{D}} & C_{\mathcal{D}\delta} \\ C_{\delta\mathcal{D}} & C_{\delta\delta} \end{vmatrix} + \frac{1}{2} \begin{bmatrix} y \\ 0 \end{bmatrix}^T \begin{bmatrix} C_{\mathcal{D}\mathcal{D}} & C_{\mathcal{D}\delta} \\ C_{\delta\mathcal{D}} & C_{\delta\delta} \end{bmatrix}^{-1} \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

Hyperparameters :

- Number of regions $K = 2^D$ (where D is the depth of the PCA partitioning).
 - $K \nearrow \implies$ Less accurate predictions.
- Number of pseudo-observations B per frontier.
 - $B \nearrow \implies$ More accurate predictions.

■ TABLE OF CONTENTS

I. Prerequisites

II. The "Big N " Problem and Beyond

III. Methods to Tackle the Issue

III.1 Patchwork Kriging (Park and Aley)

III.2 Nearest-Neighbors Gaussian Processes (Datta et al.)

IV. Performance Comparison

V. Conclusion

■ INTUITIVE IDEA

- It seems logical to assume that far away observations have little impact on the Kriging.
- Instead Kriging the whole dataset, we are Kriging each x^* only with the nearest neighbors.
- Restraining the Kriging to the m nearest neighbors of each prediction location x^* leads to a sub-optimal prediction.

■ INTUITIVE IDEA

- It seems logical to assume that far away observations have little impact on the Kriging.
- Instead Kriging the whole dataset, we are Kriging each x^* only with the nearest neighbors.
- Restraining the Kriging to the m nearest neighbors of each prediction location x^* leads to a sub-optimal prediction.

■ NEAREST-NEIGHBORS

Nearest-Neighbors are the points that have the highest spatial correlation with the prediction location x^* :

- In the case of stationary covariance functions, these are simply the closest points to x^* .
- For non-stationary covariance functions, the nearest neighbors can be found by evaluating the covariance between x^* and all observation locations and selecting the m locations with the highest covariance values.

■ CHOOSING M

- First idea : running cross-validation for different values of m and choosing the one that minimizes the prediction error.
- In Bayesian Inference, m is treated as a parameter and updated in the MCMC algorithm.

■ TABLE OF CONTENTS

I. Prerequisites

II. The "Big N " Problem and Beyond

III. Methods to Tackle the Issue

III.1 Patchwork Kriging (Park and Aley)

III.2 Nearest-Neighbors Gaussian Processes (Datta et al.)

IV. Performance Comparison

V. Conclusion

■ COMPLEXITY ANALYSIS I

For Method I (**Patchwork Kriging**), the expense is located in the computation of:

$$(C_{\delta\delta} - C_{\mathcal{D}\delta}^T C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta})^{-1}$$

Reminder

- N observations in a space of dimension d
- split into K regions of $M = \frac{N}{K}$ points, with B pseudo-observations at each boundary

■ COMPLEXITY ANALYSIS I

For Method I (**Patchwork Kriging**), the expense is located in the computation of:

$$(C_{\delta\delta} - C_{\mathcal{D}\delta}^T C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta})^{-1}$$

Reminder

- N observations in a space of dimension d
- split into K regions of $M = \frac{N}{K}$ points, with B pseudo-observations at each boundary

Complexity

- $C_{\mathcal{D}\mathcal{D}}^{-1}$ block diagonal $\rightarrow \mathcal{O}(KM^3)$
- $C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta} \rightarrow \mathcal{O}(KBM^2)$
- Inverting $C_{\delta\delta} - C_{\mathcal{D}\delta}^T C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta}$. As it is very sparse, we can use efficient methods to invert sparse matrix (Chan and George (1980)) $\rightarrow \mathcal{O}(d^3 B^3 K)$

■ COMPLEXITY ANALYSIS I

For Method I (**Patchwork Kriging**), the expense is located in the computation of:

$$(C_{\delta\delta} - C_{\mathcal{D}\delta}^T C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta})^{-1}$$

Reminder

- N observations in a space of dimension d
- split into K regions of $M = \frac{N}{K}$ points, with B pseudo-observations at each boundary

Complexity

- $C_{\mathcal{D}\mathcal{D}}^{-1}$ block diagonal $\rightarrow \mathcal{O}(KM^3)$
- $C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta} \rightarrow \mathcal{O}(KBM^2)$
- Inverting $C_{\delta\delta} - C_{\mathcal{D}\delta}^T C_{\mathcal{D}\mathcal{D}}^{-1} C_{\mathcal{D}\delta}$. As it is very sparse, we can use efficient methods to invert sparse matrix (Chan and George (1980)) $\rightarrow \mathcal{O}(d^3 B^3 K)$

$$\rightarrow \mathcal{O}(KM^3 + KBM^2 + d^3 B^3 K) \approx \mathcal{O}\left(\frac{N^3}{K^2} + d^3 B^3 K\right)$$

■ COMPLEXITY ANALYSIS II

For Method I (**NN-Kriging**),

Reminder

- N observations
- for each new location, a neighbourhood of size m is computed

■ COMPLEXITY ANALYSIS II

For Method I (**NN-Kriging**),

Reminder

- N observations
- for each new location, a neighbourhood of size m is computed

Complexity

- detail ?

■ COMPLEXITY ANALYSIS II

For Method I (**NN-Kriging**),

Reminder

- N observations
- for each new location, a neighbourhood of size m is computed

Complexity

- detail ?

$$\longrightarrow \mathcal{O}(Nm^3)$$

■ ACCURACY

For Method I (**NN-Kriging**),
Reminder

- N observations
- for each new location, a neighbourhood of size m is computed

■ ACCURACY

For Method I (**NN-Kriging**),

Reminder

- N observations
- for each new location, a neighbourhood of size m is computed

Complexity

- detail ?

■ ACCURACY

For Method I (**NN-Kriging**),

Reminder

- N observations
- for each new location, a neighbourhood of size m is computed

Complexity

- detail ?

$$\longrightarrow \mathcal{O}(Nm^3)$$

■ TABLE OF CONTENTS

I. Prerequisites

II. The "Big N " Problem and Beyond

III. Methods to Tackle the Issue

III.1 Patchwork Kriging (Park and Aley)

III.2 Nearest-Neighbors Gaussian Processes (Datta et al.)

IV. Performance Comparison

V. Conclusion

■ CONCLUSION

"On ne peut pas inverser une matrice de taille 20000 une fois, mais on peut inverser 20000 matrice de taille 20 une fois. . . enfin je crois"

— T. Lambert

■ REFERENCES I

- Chan, W. M. and George, A. (1980). A linear time implementation of the reverse cuthill-mckee algorithm. *BIT Numerical Mathematics*, 20:8–14.
- Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). On nearest-neighbor Gaussian process models for massive spatial data. *Wiley Interdisciplinary Reviews Computational Statistics*, 8(5):162–171.
- Park, C. and Apley, D. (2018). Patchwork Kriging for large-scale Gaussian process regression. *Journal of Machine Learning Research*.