

---

# Geostatistics

## École d'été Traitement des données massives et apprentissage

Mathieu Ribatet




# Motivation

- Many variables are spatial in extent, e.g., rainfall, petroleum, elevation
- The use of univariate or even multivariate statistical models may be too restrictive.
- An example would be to try to estimate the expected surface of a pollutant exceeding some critical level  $u_{\text{crit}}$  in a study region  $\mathcal{X} \subset \mathbb{R}^d$ , i.e.,

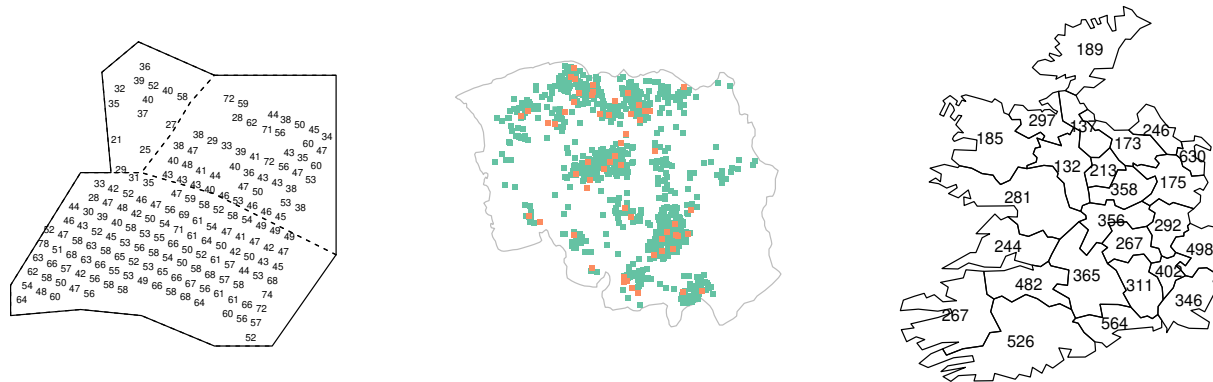
$$\text{Area}(u_{\text{crit}}) = \mathbb{E} \left[ \int_{\mathcal{X}} 1_{\{Y(s) > u_{\text{crit}}\}} \mathrm{d}s \right],$$

where  $Y(s)$  is the amount of pollutant at location  $s$ .

 The use of univariate models may still be useful provided the focus is on pointwise quantities, e.g., quantiles at  $s_* \in \mathcal{X}$ .

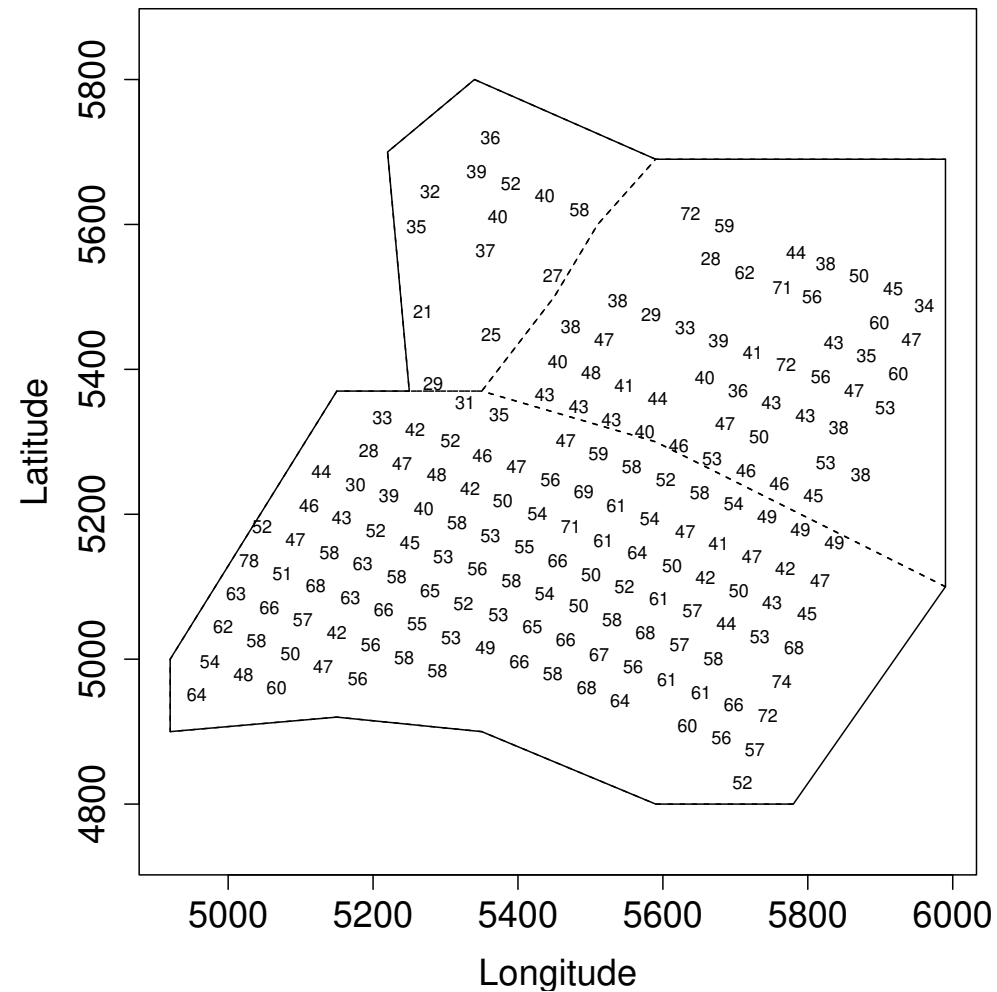
# Geostatistics vs Point-processes vs Lattice

- Spatial data may be of (at least) two types:
  - geostatistical data where the variable is defined **continuously** on  $\mathcal{X}$ , e.g., rainfall;
  - punctual data where data are **points falling randomly** over some (possibly functional) space  $\mathcal{X}$ , e.g., tree locations.
  - lattice data where data are **aggregated over sub-regions**, e.g., number of citizen in countries.



**Figure 1:** *The three different type of spatial data. From left to right: geostatistical, punctual and lattice data.*

# The Ca<sub>20</sub> data set



Focus is on geostatistical data only!

▷ 1. Framework

2. Inference

3. Model-based  
geostatistics

4. Simulation

5. Bayesian  
hierarchical models

6. Big data

# 1. Framework

# Stochastic processes

---

**Definition 1.** A **stochastic process** defined on  $\mathcal{X}$  is a collection of random variables indexed by  $\mathcal{X}$  on the **same probability space**  $(\Omega, \mathcal{F}, \Pr)$ .

**Proposition 1.** A stochastic process  $\{Y(s) : s \in \mathcal{X}\}$  is completely characterized from its finite dimensional distribution functions, i.e., for any  $k \geq 1$  and  $s_1, \dots, s_k \in \mathcal{X}$

$$\Pr \{Y(s_1) \leq A_1, \dots, Y(s_k) \leq A_k\}, \quad A_1, \dots, A_k \text{ Borel sets,}$$

*provided they satisfy the hypothesis of the Kolmogorov extension theorem, i.e., invariance to permutation and consistent marginalization.*

# Strictly stationnary processes

**Definition 2.** A stochastic process  $\{Y(s) : s \in \mathcal{X}\}$  is said (strictly) stationnary if its finite dimensional distribution functions are **invariant by translation**, i.e., for any  $k \geq 1$ ,  $s_1, \dots, s_k \in \mathcal{X}$  and  $h \in \mathcal{X}$  we have

$$\Pr \{Y(s_1 + h) \leq A_1, \dots, Y(s_k + h) \leq A_k\} = \Pr \{Y(s_1) \leq A_1, \dots, Y(s_k) \leq A_k\},$$

where  $A_j$  are Borel sets.

 In practice, strict stationnarity is too strong and cannot be checked. Need a weaker hypothesis.

# Second order processes

---

**Definition 3.** A **second order** stochastic process is a stochastic process whose second order moment exists, i.e.,  $\text{Var}[Y(s)] < \infty$  for all  $s \in \mathcal{X}$ .

- Working with second order processes allows to define
  - the **mean function / trend / drift**

$$\begin{aligned}\mu: \mathcal{X} &\longrightarrow \mathbb{R} \\ s &\longmapsto \mathbb{E}[Y(s)],\end{aligned}$$

- the **covariance function**

$$\begin{aligned}K: \mathcal{X} \times \mathcal{X} &\longrightarrow \mathbb{R} \\ (s, s') &\longmapsto \text{Cov}\{Y(s), Y(s')\}.\end{aligned}$$



## Weak stationnarity and isotropy

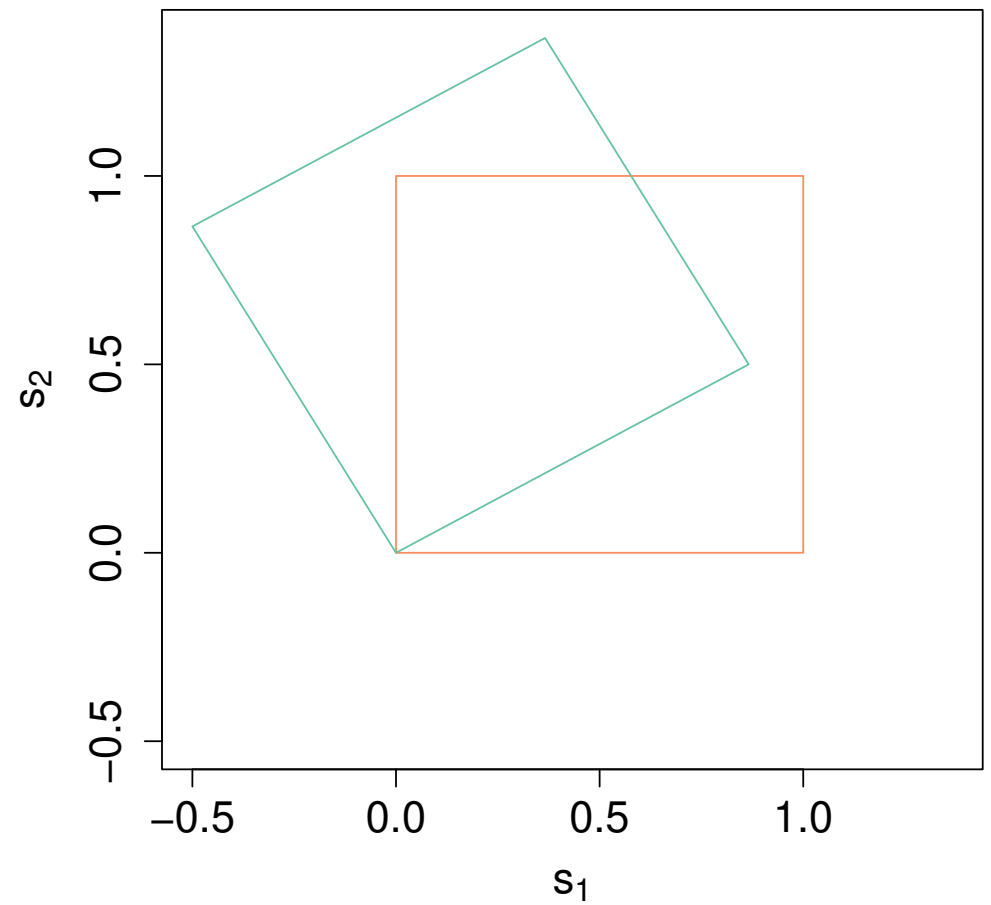
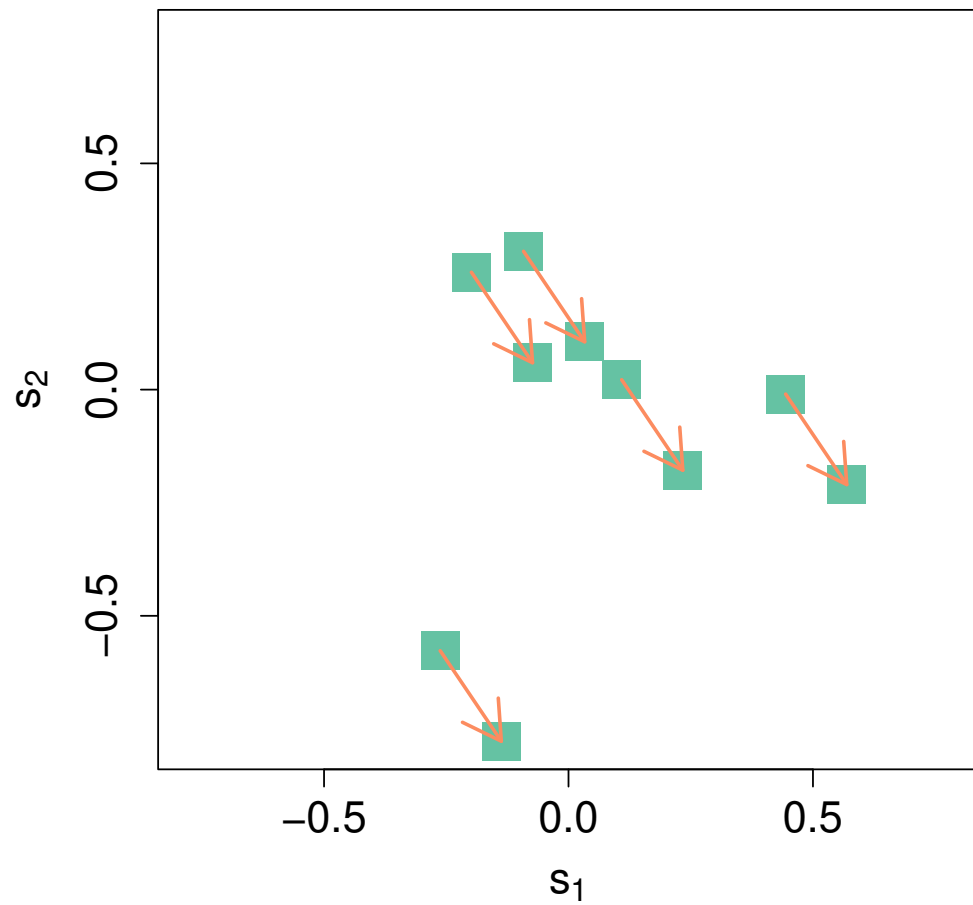
---

**Definition 4.** A second order process is said **weakly stationnary**, or just **stationnary**, if for any  $s, s' \in \mathcal{X}$  and  $h \in \mathcal{X}$  we have

$$\mu(s+h) = \mu(s), \quad K(s+h, s'+h) = K(s, s'). \quad (\text{translation invariance})$$

**Definition 5.** A stochastic process  $\{Y(s) : s \in \mathcal{X}\}$  is said **isotropic** if for any rotation matrix  $R$ , i.e.,  $|R| = 1$  and  $R^{-1} = R^T$ , we have

$$\{Y(Rs) : s \in \mathcal{X}\} \stackrel{d}{=} \{Y(s) : s \in \mathcal{X}\}. \quad (\text{rotation invariance})$$



**Figure 2:** *Illustration of stationnarity and isotropy.*

# Consequences

- If a process is stationnary we have

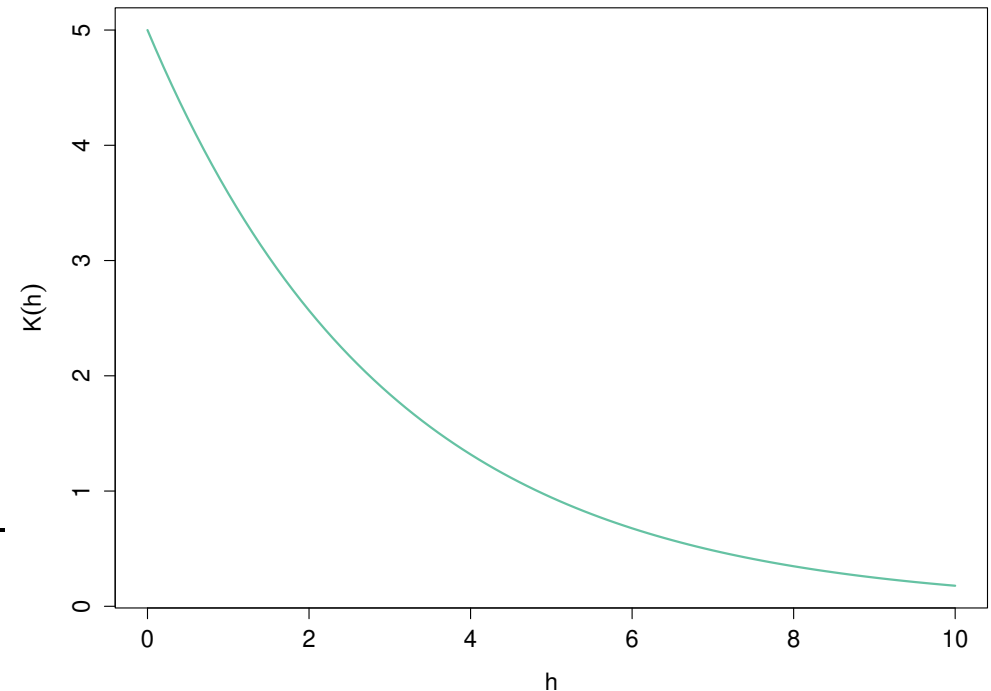
$$K(s, s') = K(o, s' - s) = K(h),$$

where  $h = s - s'$  and is **even** since

$$K(s, s') = K(s', s).$$

- If we further assume isotropy, the covariance function now satisfies

$$\begin{aligned} K(Rh) &= K(h) \\ &= K(\|h\|) \\ &= K(-\|h\|). \end{aligned}$$



**Figure 3:** Plot of a stationnary isotropic covariance function. *What is  $K(0)$ ?*

# Processes with stationary increments

---

**Definition 6.** A stochastic process  $\{Y(s) : s \in \mathcal{X}\}$  is said to have **stationary increments** if for all  $s \in \mathcal{X}$  and  $h \in \mathcal{X}$ , the distribution of

$$Y(s + h) - Y(s) \stackrel{\text{fidi}}{=} Y(h) - Y(o),$$

i.e., depends only on the lag  $h$  and where  $o \in \mathcal{X}$  is an arbitrary origin.

- The motivation for using stationary increments processes is that we are **no longer restricted to stationary processes**.
- We can even work with **non second order processes** and simply assume

$$\text{Var}[Y(h) - Y(o)] < \infty.$$

**Example 1.** Consider the following **random walk** defined on  $\mathcal{X} = \mathbb{Z}$

$$Y(s+1) = Y(s) + \varepsilon_{s+1}, \quad \varepsilon_j \stackrel{\text{iid}}{\sim} N(0, \sigma^2).$$

It has indeed stationary increments since

$$Y(s+h) - Y(s) = \sum_{j=0}^{h-1} \{Y(s+h-j) - Y(s+h-j-1)\} = \sum_{j=0}^s \varepsilon_{s+h-j} \sim N(0, h\sigma^2).$$

Note that  $\text{Var}\{Y(s)\} \rightarrow \infty$  as  $s \rightarrow \infty$  and thus is **not second order**.

**i** Extension of the above random walk to  $\mathcal{X} = \mathbb{R}^d$  leads to the so-called **Brownian random fields**. If we further assume dependence across innovation  $\varepsilon_j$  we have **fractional Brownian random fields**.

# Semi-variogram

- The **covariance function** is a summary statistic of the spatial dependence function for at most **second order processes**.
- To get an analogue for **stationnary increment processes** we rather consider the **semi-variogram**

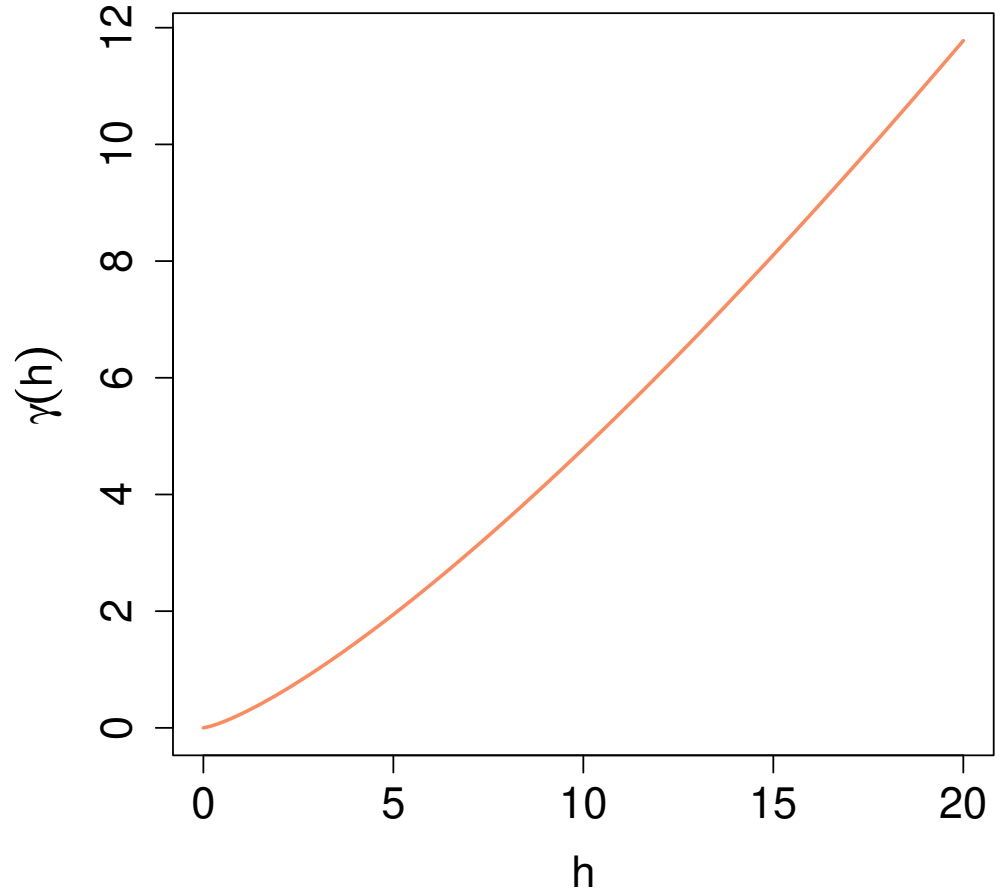
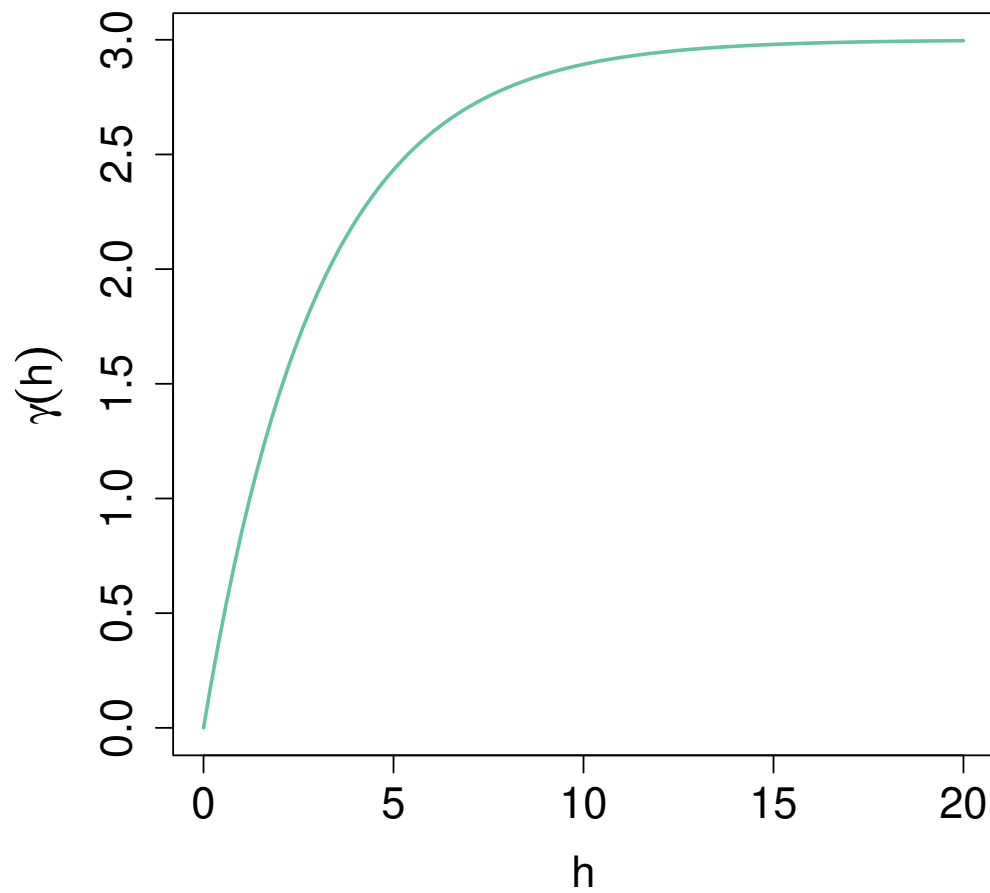
$$\gamma(h) = \frac{1}{2} \text{Var}[Y(h) - Y(o)] = \frac{1}{2} \mathbb{E} \left[ \{Y(h) - Y(o)\}^2 \right].$$

 If the process is indeed second order we have

$$\gamma(h) = \frac{1}{2} \{2K(o, o) - 2K(o, h)\} = K(o, o)\{1 - \rho(h)\},$$

where  $h \mapsto \rho(h)$  is the correlation function and

$$\gamma(h) \longrightarrow K(o, o), \quad \|h\| \rightarrow \infty, \quad (\text{as long as } \rho(h) \rightarrow 0)$$



**Figure 4:** *Bounded (left) and unbounded (right) semi-variograms. If it exists, what is  $\gamma(\infty)$ ?*

# Some isotropic stationary correlation functions and variograms

Family	$\rho(h)$	$\gamma(h)$	Support
Exponential	$\exp(-h/\lambda)$	$1 - \exp(-h/\lambda)$	$\lambda > 0$
Gaussian	$\exp\left\{-\left(h/\lambda\right)^2\right\}$	$1 - \exp\left\{-\left(h/\lambda\right)^2\right\}$	$\lambda > 0$
Stable / Powered exponential	$\exp\left\{-\left(h/\lambda\right)^\kappa\right\}$	$1 - \exp\left\{-\left(h/\lambda\right)^\kappa\right\}$	$\lambda > 0, 0 \leq \kappa \leq 2$
Whittle–Matérn	$\frac{2^{1-\kappa}}{\Gamma(\kappa)} \left(\frac{u}{\lambda}\right)^\kappa K_\kappa\left(\frac{u}{\lambda}\right)$	$1 - \frac{2^{1-\kappa}}{\Gamma(\kappa)} \left(\frac{u}{\lambda}\right)^\kappa K_\kappa\left(\frac{u}{\lambda}\right)$	$\lambda > 0, \kappa > 0$
Fractional	—	$(h/\lambda)^\kappa$	$0 \leq \kappa \leq 2$

- The parameters  $\lambda$  and  $\kappa$  are known as the **range** and **smooth** parameters.
- Associated covariance functions are derived using a **sill parameter**  $\tau$ , i.e.,

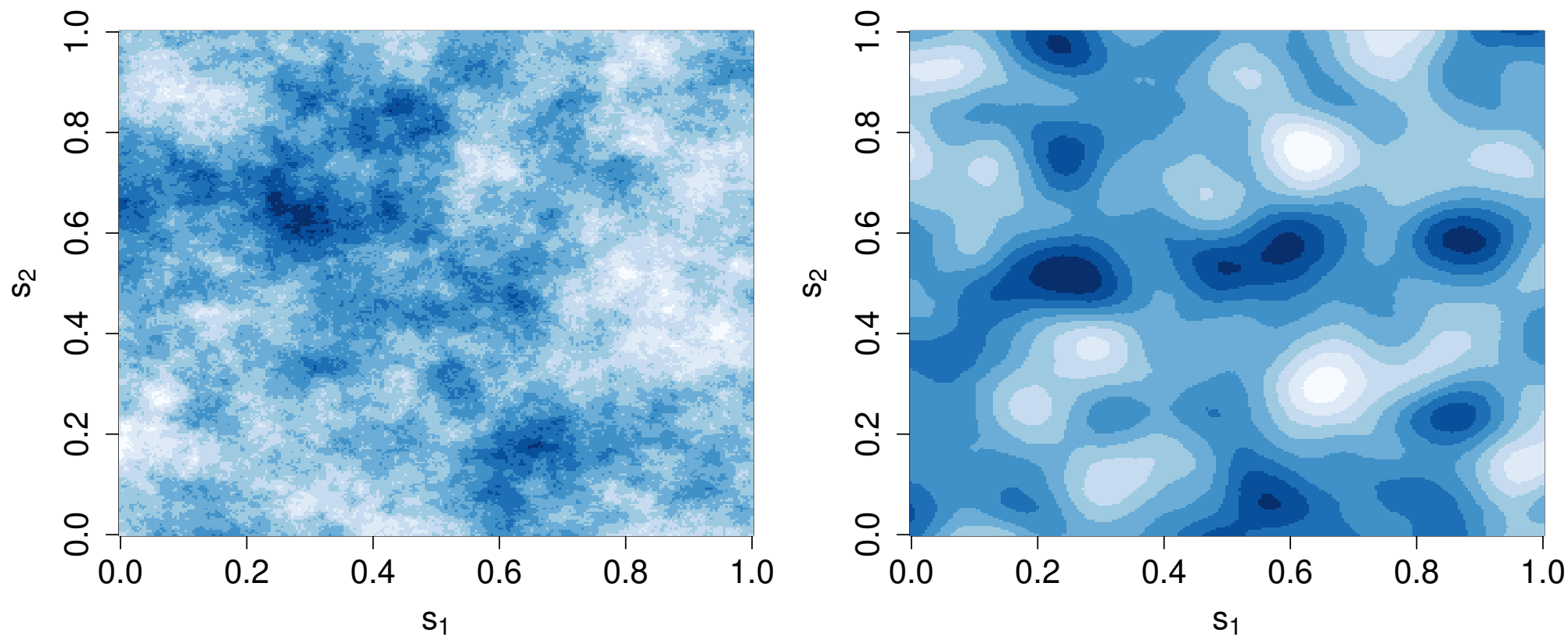
$$K(h) = \tau\rho(h), \quad \tau > 0. \quad (\tau = K(o))$$

- The **smooth** and **range** parameters drives respectively the **smoothness** of the random process and **the range of spatial dependence**.

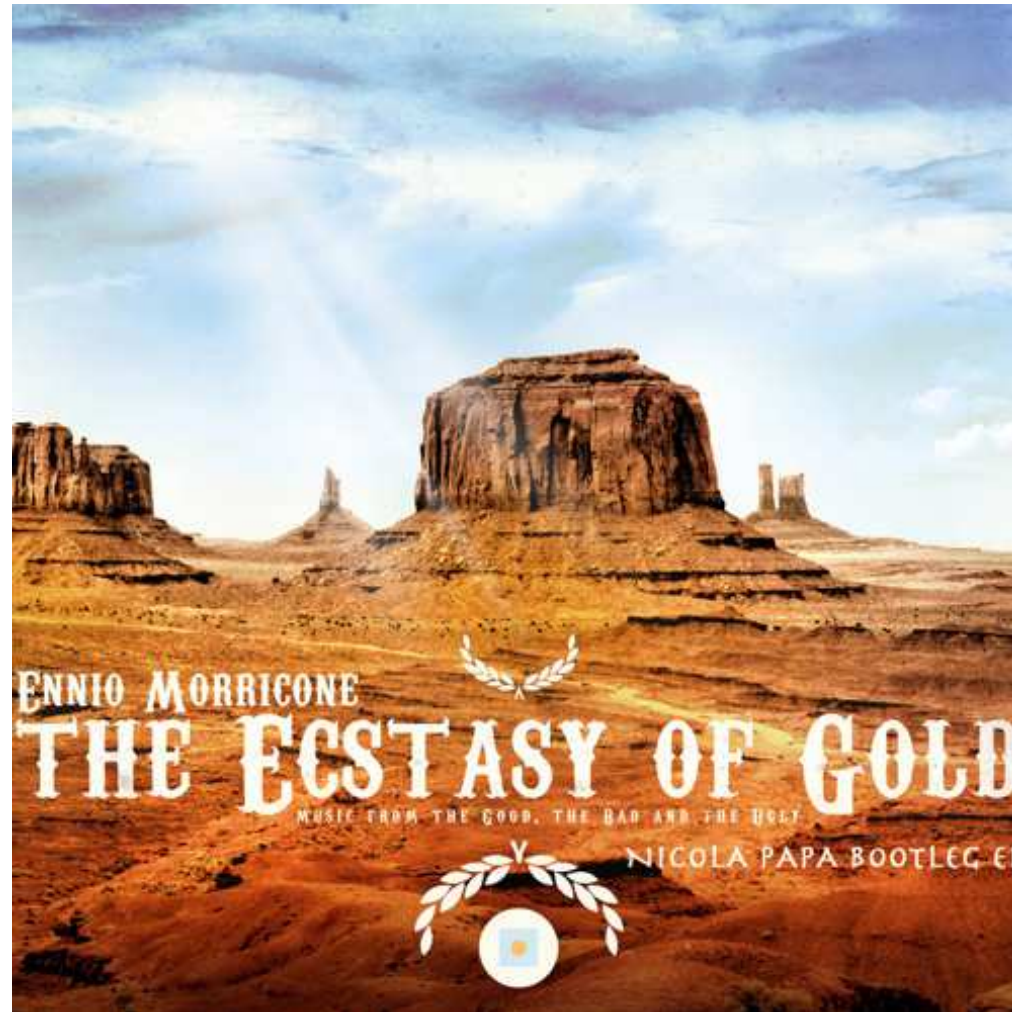


The practical range  $h_p$  is the distance such that  $\rho(h_p) = 0.05$ .





**Figure 5:** *Two realizations of a random fields with a powered exponential correlation function. Left:  $\kappa = 1$ . Right:  $\kappa = 2$ .*



- 
- The covariance function may have a **discontinuity** at the origin, called **nugget effect**, i.e.,

$$K(h) = \begin{cases} \eta + \tau, & h = 0, \\ \tau\rho(h), & h > 0. \end{cases}$$

- The nugget effect may have two possible interpretations:
  - error in measurements, i.e.,  $Y(s) = S(s) + \varepsilon(s)$
  - spatial variation on a scale smaller than the minimum distance between measurements (if no replicate)

## Some interesting properties and quantities

---

**Proposition 2.** *If a correlation of a stationary process is **discontinuous**, then discontinuity has to be at the **origin**.*

*If a stationary process has a correlation function which is **continuous (at the origin)** then it is continuous and if twice differentiable, the process is differentiable (both from a  $L^2$  sense).*

**i** Extension to higher order are possible!

1. Framework

▷ 2. Inference

3. Model-based  
geostatistics

4. Simulation

5. Bayesian  
hierarchical models

6. Big data

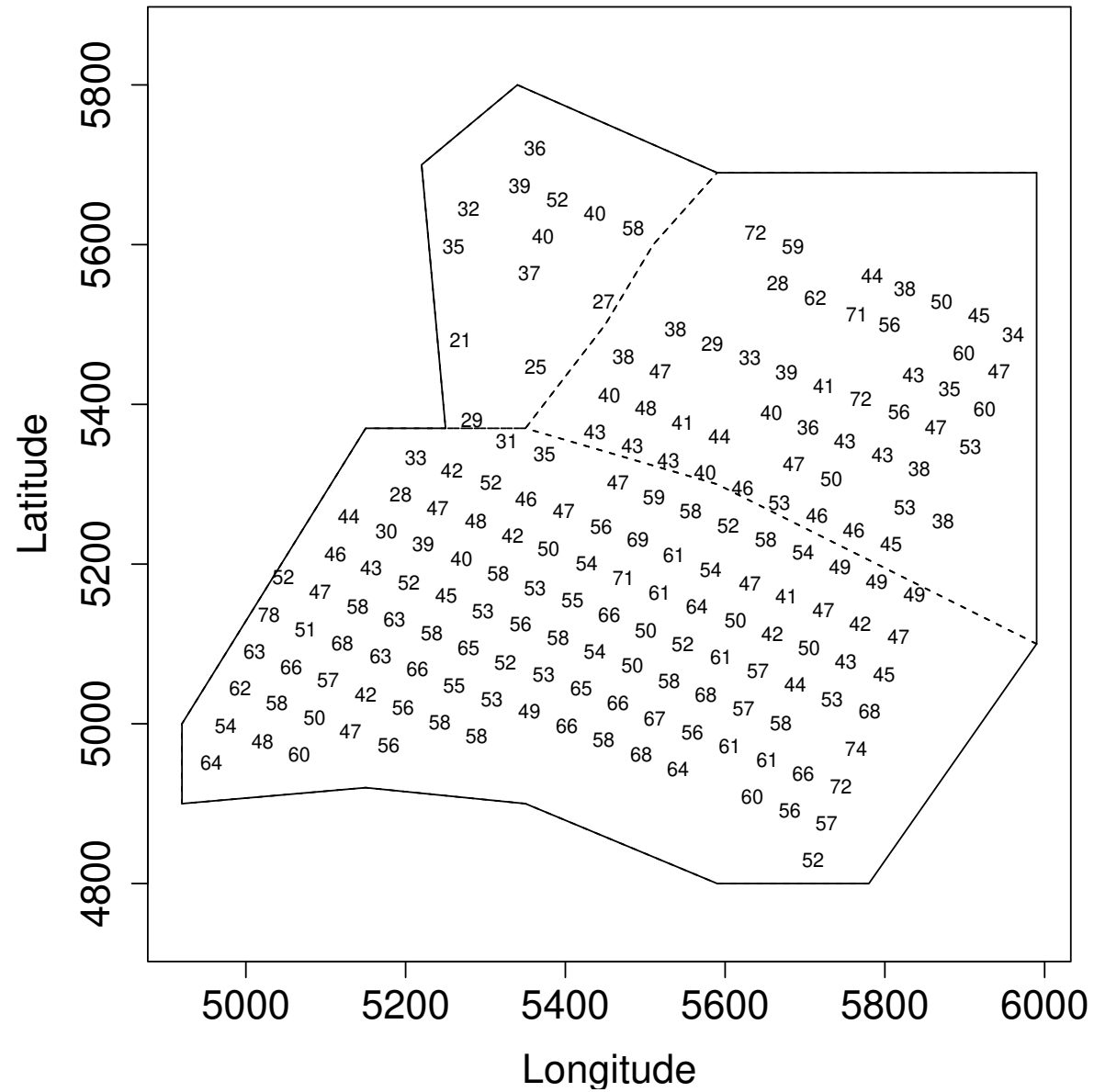
## 2. Inference

# Descriptive analysis

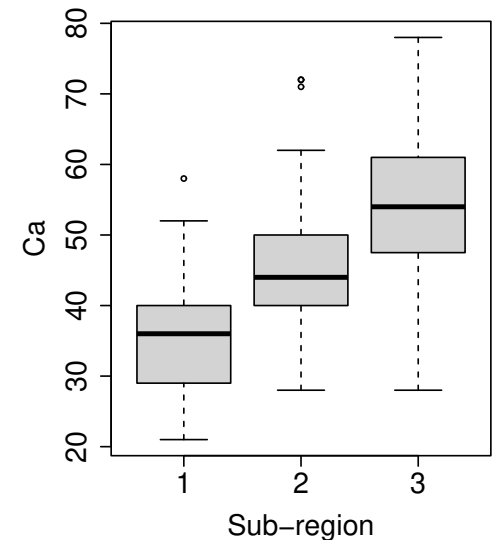
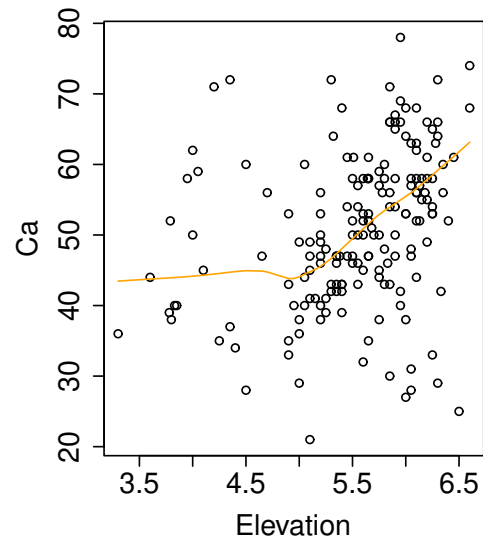
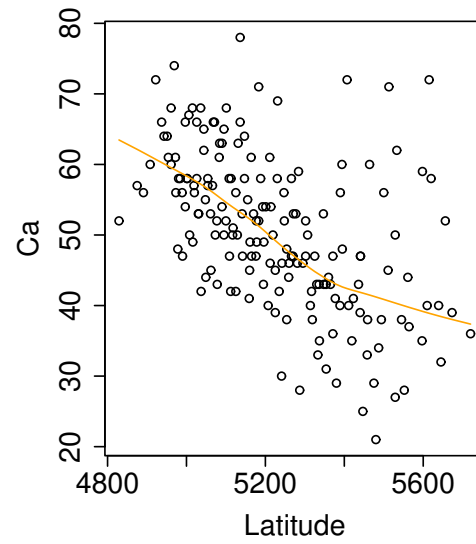
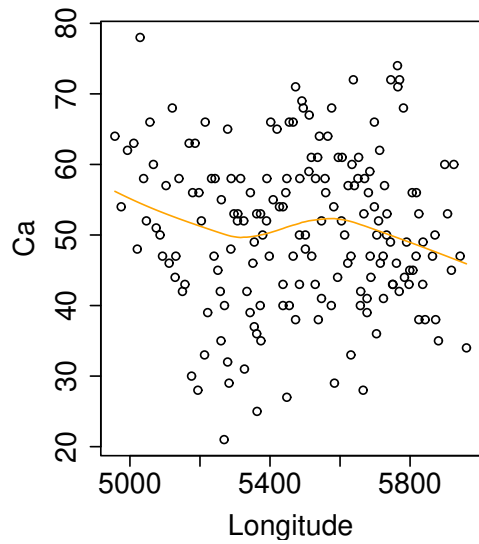
---

- Before trying to model the data we need to check whether the data can safely be assumed stationary / isotropic / ...
- Essential we start with a **descriptive analysis** which, for our context, consists in
  - checking for any trend in the mean function  $s \mapsto \mu(s)$
  - inspecting the semi-variogram.
- The first stage is very simple. Just plot data w.r.t. some covariates, e.g., longitude, latitude, ...

# Ca20 dataset.

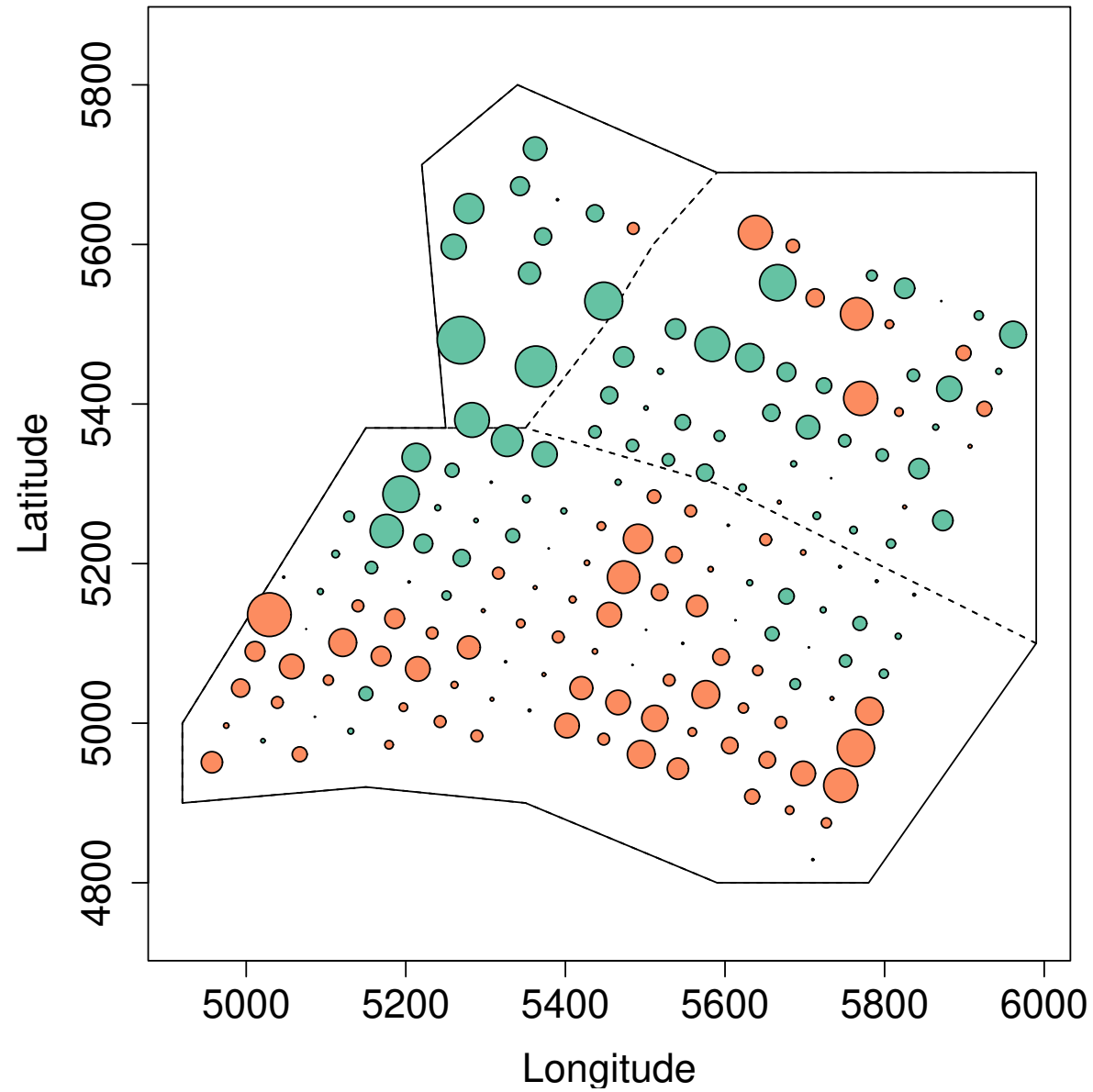


# Ca20 dataset.





# Ca20 dataset.



# Empirical variograms

- Given some data  $\mathcal{D}_n = \{Y_i(s_j) : i = 1, \dots, n, j = 1, \dots, k\}$ , we easily estimate the semi-variogram

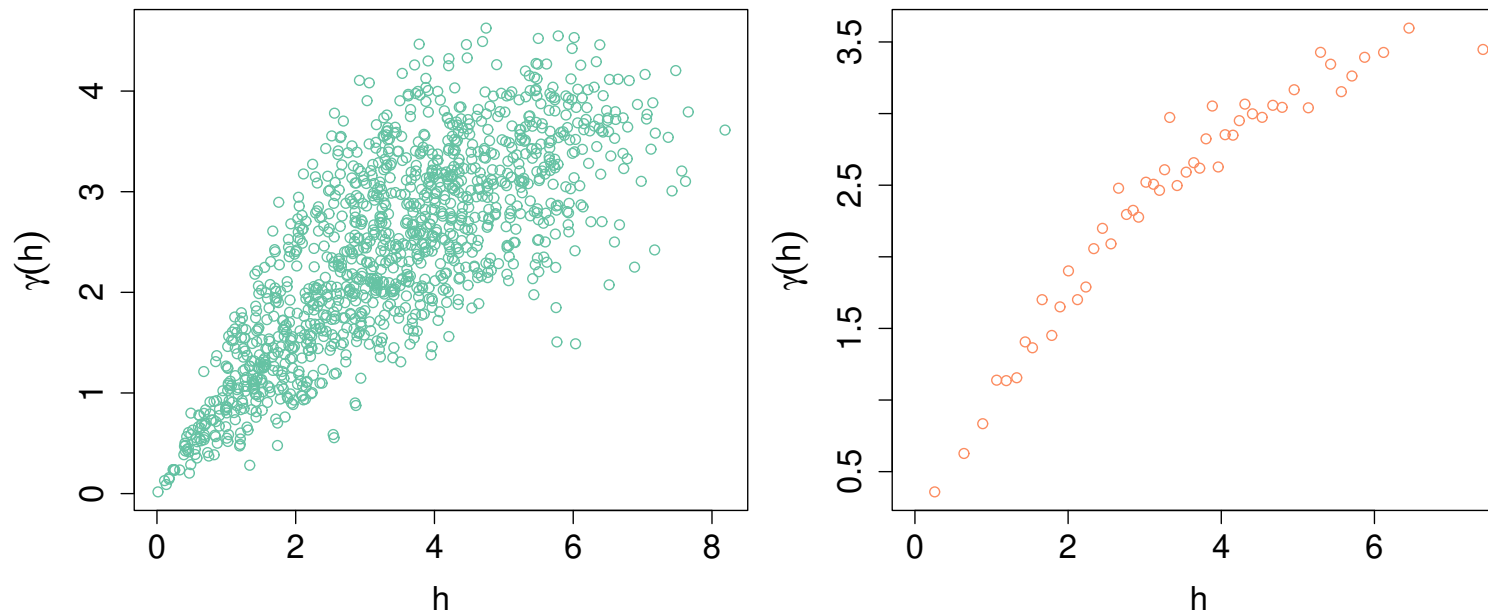
$$\hat{\gamma}(h_{j,\ell}) = \frac{1}{2n} \sum_{i=1}^n \{Y_i(s_j) - Y_i(s_\ell)\}^2, \quad h_{j,\ell} = \|s_j - s_\ell\|.$$

- We may have  $n = 1$  so that the above estimator has **huge variance** and we rather use a **binned version**, i.e.,

$$\tilde{\gamma}(h_b) = \frac{1}{2|B_b|} \sum_{i=1}^n \sum_{j,\ell=1}^k \{Y_i(s_j) - Y_i(s_\ell)\}^2 \mathbf{1}_{\{\|s_j - s_\ell\| \in B_b\}},$$


where  $\{B_b : b = 1, \dots, B\}$  is a partition of  $(0, \max h_{j,\ell})$  and  $h_b$  is the centroid of  $B_b$ .

 The binned estimator is however **biased** but has a (much) **lower variance**.



**Figure 6:** *Empirical variograms. Left: raw. Right: binned.*

The above estimator makes sense only if your data can be considered as **stationnary** or at least with **stationnary increments**.

 You may want to remove any possible trends (using a linear model for instance) and estimate the variogram on the residuals.

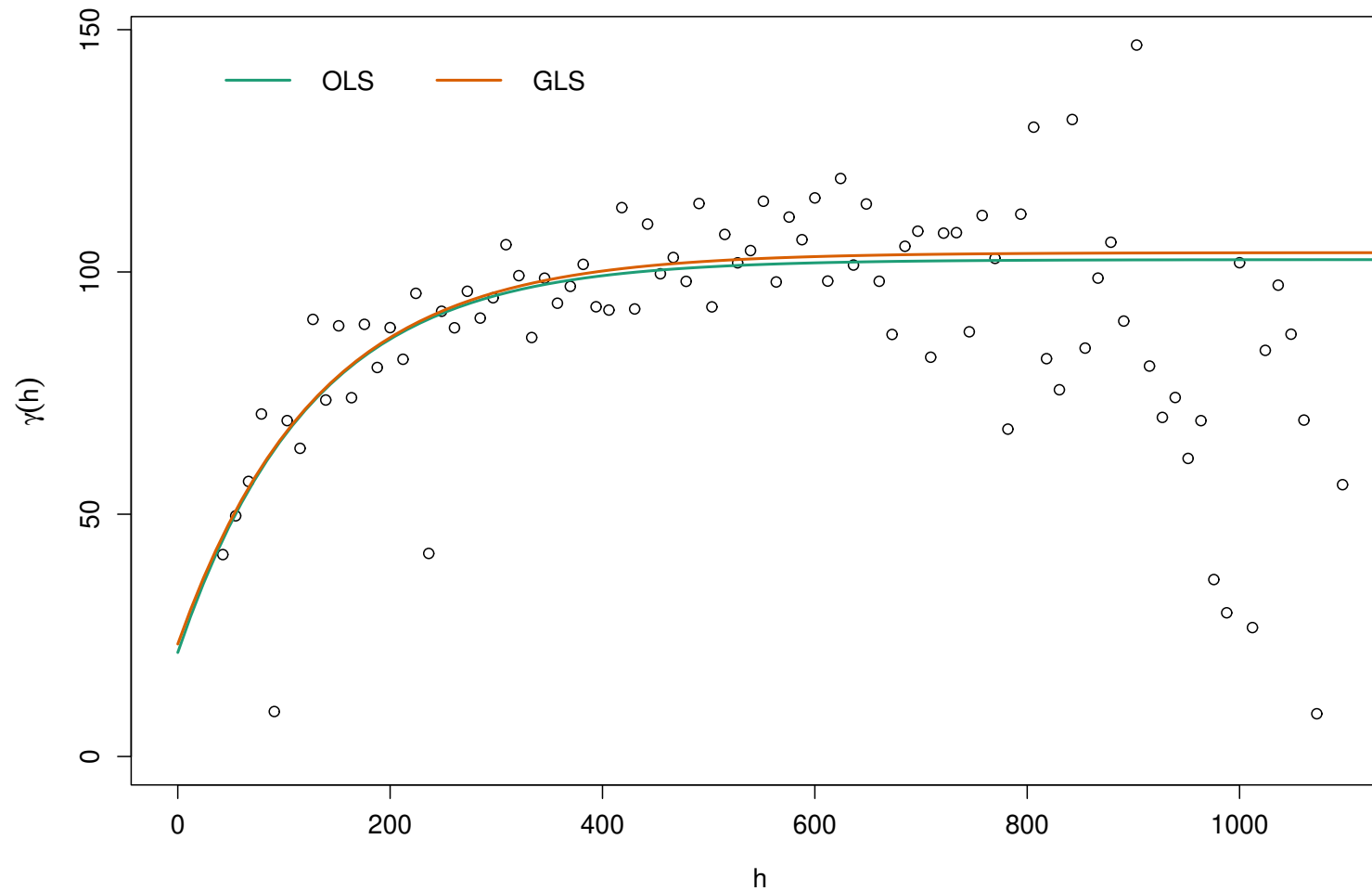
# Least square fitting of a variogram

- Suppose we have fitted a mean function, e.g., from linear models.
- We can fit any parametric variogram  $\gamma(\cdot; \psi)$  minimizing using the (weighted) least square estimator on the empirical variogram  $\hat{\gamma}$ , i.e.,

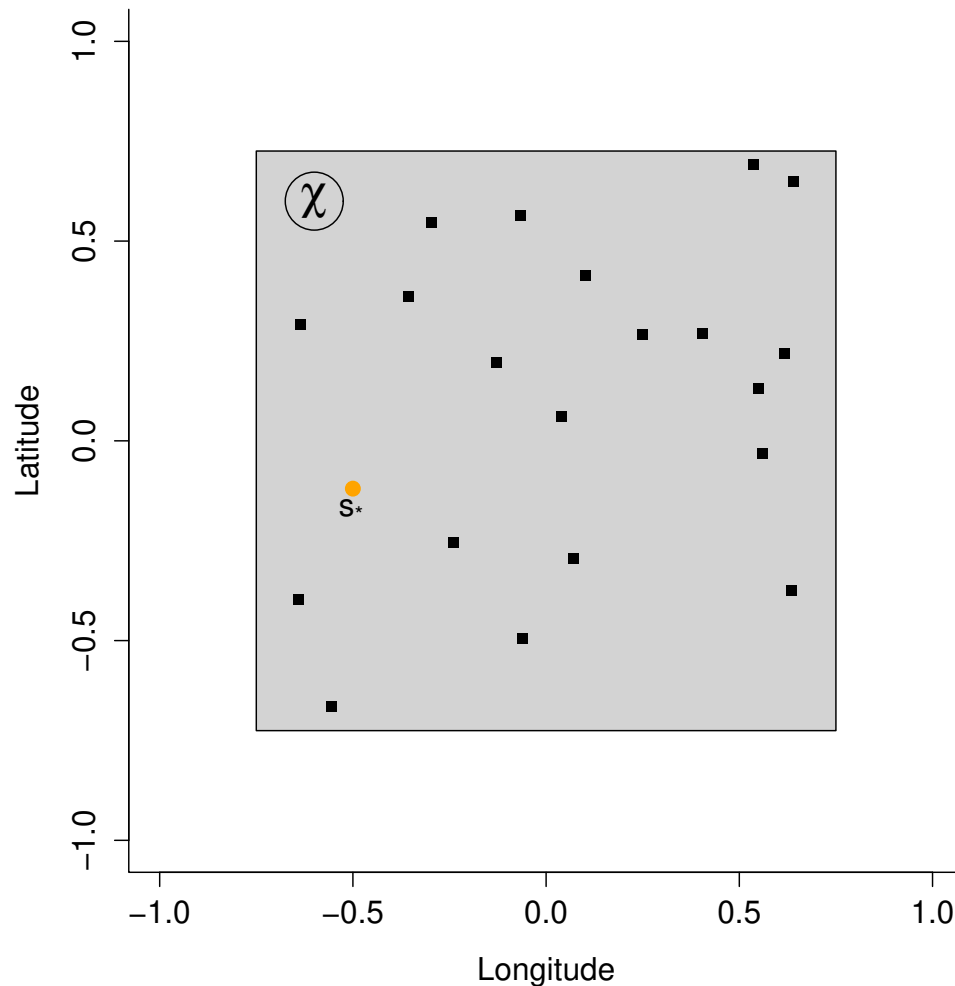
$$\hat{\psi} = \arg \min_{\psi \in \Psi} \sum_{j,l} \omega_{j,l} \{ \hat{\gamma}(h_{j,l}) - \gamma(h_{j,l}; \psi) \}^2 .$$

- The two fitted quantities are all we need to enable predictions!

💣 Nasty optimization problem: use several initial values!  
Often fix the smooth parameter to some values, e.g.,  $\kappa = 0.25, 0.5, \dots, 2$ .  
Always question yourself if a nugget effect makes sense.



**Figure 7:** *Least square fitting of a parametric variogram on the Calcium dataset.*



- Prediction of  $Y(s_*)$  based on observations  $Y(s_1), \dots, Y(s_k)$ .
- Restriction to **unbiased linear estimators**, i.e.,

$$\hat{Y}(s_*) = \sum_{j=1}^k \lambda_j Y(s_j),$$

with  $\mathbb{E}[\hat{Y}(s_*)] = \mu(s_*)$ .

- Estimator is the one minimizing the mean squared error, i.e.,

$$\hat{Y}(s_*) = \arg \min_T \mathbb{E} \left[ \{T - Y(s_*)\}^2 \right].$$

# (Universal) Kriging

---

- There are several of kriging:

**Simple**  $\mu(s) \equiv 0$

**Ordinary**  $\mu(s) = m$ ,  $m$  unknown parameter

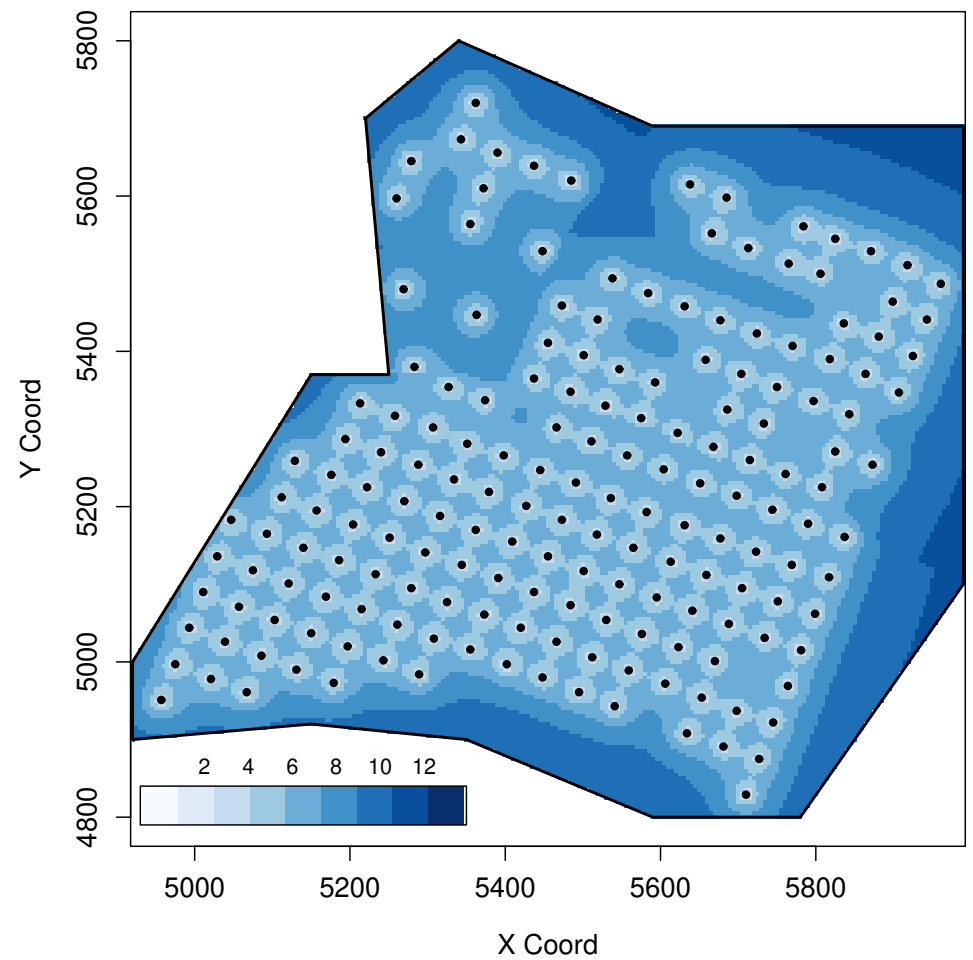
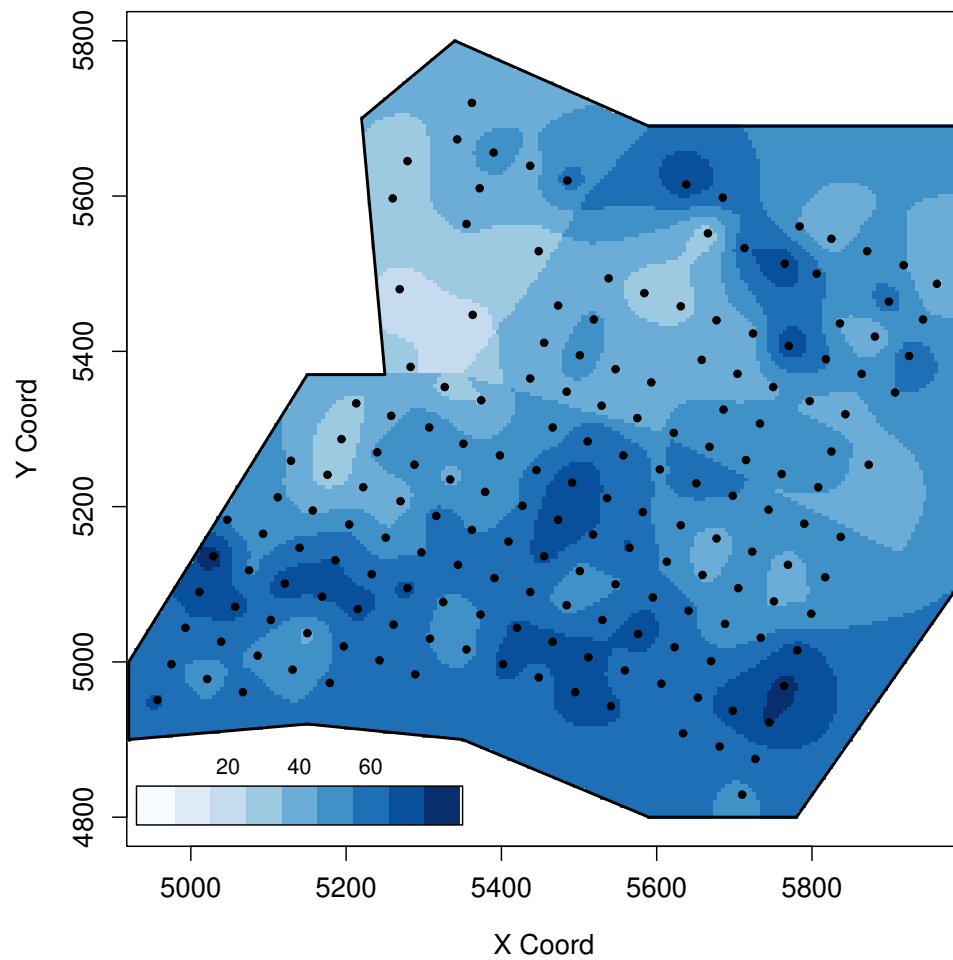
**Universal**  $\mu(s) = \mathbf{x}(s)^\top \boldsymbol{\beta}$ ,  $\boldsymbol{\beta}$  unknown parameter,  $\mathbf{x}(s)$  vector of covariates,

**Co-kriging**  $Y$  is multivariate

and their intrinsic counterpart.

- **Explicit expressions** for  $\hat{Y}(s_*)$  are available but not given here (nasty).
- We can also get expression for the **kriging variance**, i.e.,

$$\text{Var} \left\{ \hat{Y}(s_*) - Y(s_*) \right\}.$$



**Figure 8:** *Kriging estimator (left) and kriging standard error (right) for the Ca20 dataset.*



1. Framework

---

2. Inference

---

3. Model-based  
▷ geostatistics

---

4. Simulation

---

5. Bayesian  
hierarchical models

---

6. Big data

---

## 3. Model-based geostatistics

## Gaussian distributions (reminder)

**Definition 7.** The **multivariate Gaussian distribution** defined on  $\mathbb{R}^d$ ,  $d \geq 1$ , has probability density function

$$f(\mathbf{y}) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\}, \quad \mathbf{y} \in \mathbb{R}^d, \quad (1)$$

where  $\boldsymbol{\mu} \in \mathbb{R}^d$  is the **mean vector** and  $\Sigma \in M_d(\mathbb{R})$  is the **covariance matrix**.

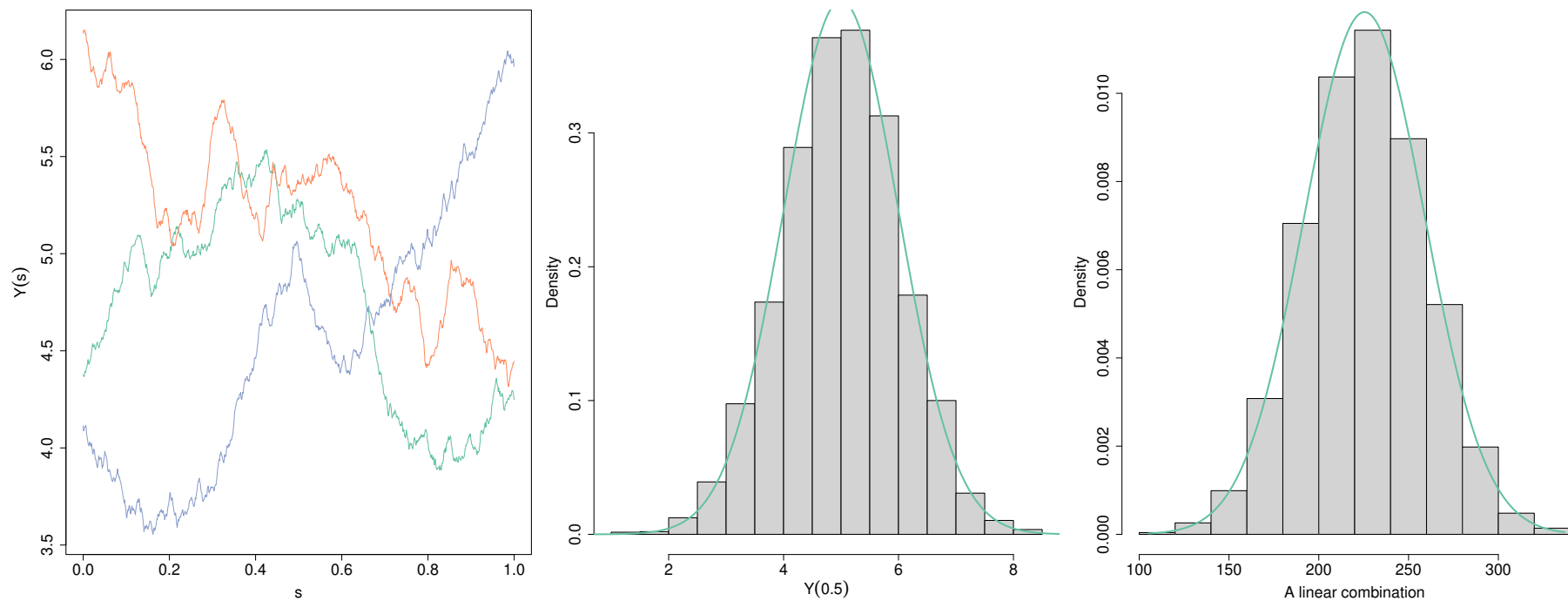
**i** The Mahalanobis distance is given by

$$a^2(\mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

# Gaussian processes

**Definition 8.** A Gaussian process  $\{Y(s) : s \in \mathcal{X}\}$  is a stochastic process whose finite dimensional distribution functions are multivariate Gaussian.

**Proposition 3.** A Gaussian process is completely characterized through its *mean function* and *covariance function*.




**Figure 9:** Numerical illustration of a Gaussian process.

## (semi) Definite positive functions

**Definition 9.** A function  $f: \mathbf{s} \in \mathbb{R}^d \mapsto f(\mathbf{s})$  is said to be (semi) definite positive if it is **symmetric** and

$$\boldsymbol{\lambda}^\top A \boldsymbol{\lambda} > 0, \quad A = (a_{i,j} = f(s_i - s_j) : i, j = 1, \dots, d), \quad x_1, \dots, x_p \in \mathbb{R}^d,$$

for any non-zero vector  $\boldsymbol{\lambda} \in \mathbb{R}^p$ . It is semi definite positive if the above inequality is not strict.

 The covariance function  $\gamma$  is (semi) definite positive to ensure that the Mahalanobis distance

$$a^2(\mathbf{s}, \mathbf{y}) = (\mathbf{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{s})(\mathbf{y} - \boldsymbol{\mu}), \quad \boldsymbol{\Sigma}(\mathbf{s}) = \{\sigma_{i,j} = \gamma(s_i, s_j)\}$$

is always **positive** and the multivariate Gaussian density is properly defined.

# Fitting a Gaussian process

- Having observed  $n$  independent observations at  $k$  spatial locations, i.e.,  $\mathcal{D}_n = \{y_i(s_j) : i = 1, \dots, n, j = 1, \dots, k\}$   $s_1, \dots, s_k$ , we define the **log-likelihood** as

$$\ell(\mu, \gamma; \mathcal{D}_n) = -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log |\Sigma(\mathbf{s})| - \frac{1}{2} \sum_{i=1}^n a^2(\mathbf{s}, \mathbf{y}_i).$$

♥ no likelihood theory here, but gaussian processes can (easily) be estimated by **maximizing the log-likelihood**.

# Parametric assumptions

---

- The above likelihood has some flaws:
  - it has  $d + d(d + 1)/2$  parameters to estimate which is typically too large;
  - cannot enable prediction at a new location  $s_*$  since both mean and covariance function cannot be computed at  $s_*$ .
- Hence we further place some parametric structures on
  - the mean function  $\mu(s)$ , e.g.,

$$\mu(s; \boldsymbol{\beta}) = \mathbf{x}(s)^\top \boldsymbol{\beta},$$

where  $\mathbf{x}(s)$  is a vector of additional covariates at  $s$  and  $\boldsymbol{\beta}$  a parameter vector to be estimated.

- the covariance function  $\gamma(s, s') = \gamma(s, s'; \boldsymbol{\psi})$  using some prescribed parametric expressions as the ones presented earlier.

# Non isotropic/stationnary covariance functions

- Defining non isotropic / stationnary covariance functions is a current research field and is far from being trivial.
- A quick and dirty way to get non isotropic covariance functions is to use any isotropic correlation function on a **transformed space**  $\mathcal{X}'$  given by

$$\begin{aligned}\phi: \mathcal{X} &\longrightarrow \mathcal{X}' \\ s &\longmapsto \phi(s; \kappa),\end{aligned}$$

for some prescribed parametric one–one mapping  $\phi(\cdot; \kappa)$ .

- A specific case, known as **geometric anisotropy**, is to set

$$\phi(s; \kappa) = C(\kappa)s, \quad C(\kappa) = \begin{bmatrix} \cos \kappa_1 & -\sin \kappa_1 \\ \sin \kappa_1 & \cos \kappa_1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \kappa_2^{-1} \end{bmatrix},$$

$\kappa_1, \kappa_2$  are respectively the anisotropy angle and ratio.

# Interpolation

- As usual the best estimator we can reach (in a  $L^2$  sense) is the conditional expectation, i.e.,

$$\hat{Y}(s_*) = \mathbb{E} \{Y(s_*) \mid Y(s_1), \dots, Y(s_k)\}.$$

- For the Gaussian case, the conditional expectation is **linear** in the  $Y(s_j)$ .
- Hence the above estimator is actually the **kriging estimator**!

🔊 You will sometimes hear: “kriging is the optimal estimator” in a  $L^2$  sense. It is **wrong** unless if we assume Gaussianity. However it is indeed optimal if we restrict to linear unbiased estimators.



## By the way...

---

- What if my data are **not Gaussian**, e.g., rainfall amount.
- A quick and dirty way is to work on a **transformation of your data**, e.g.,  $\log Y(s)$ , so that Gaussianity is a sensible choice.
- One widely used choice for positive variable is the **Box–Cox transformation**

$$y \mapsto \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0. \end{cases}$$

- However it implicitly assumes that the data are stationary so you need to remove any trend first to estimate the shape parameter  $\lambda$  in the Box–Cox transformation.



1. Framework

2. Inference

3. Model-based  
geostatistics

▷ 4. Simulation

5. Bayesian  
hierarchical models

6. Big data

## 4. Simulation

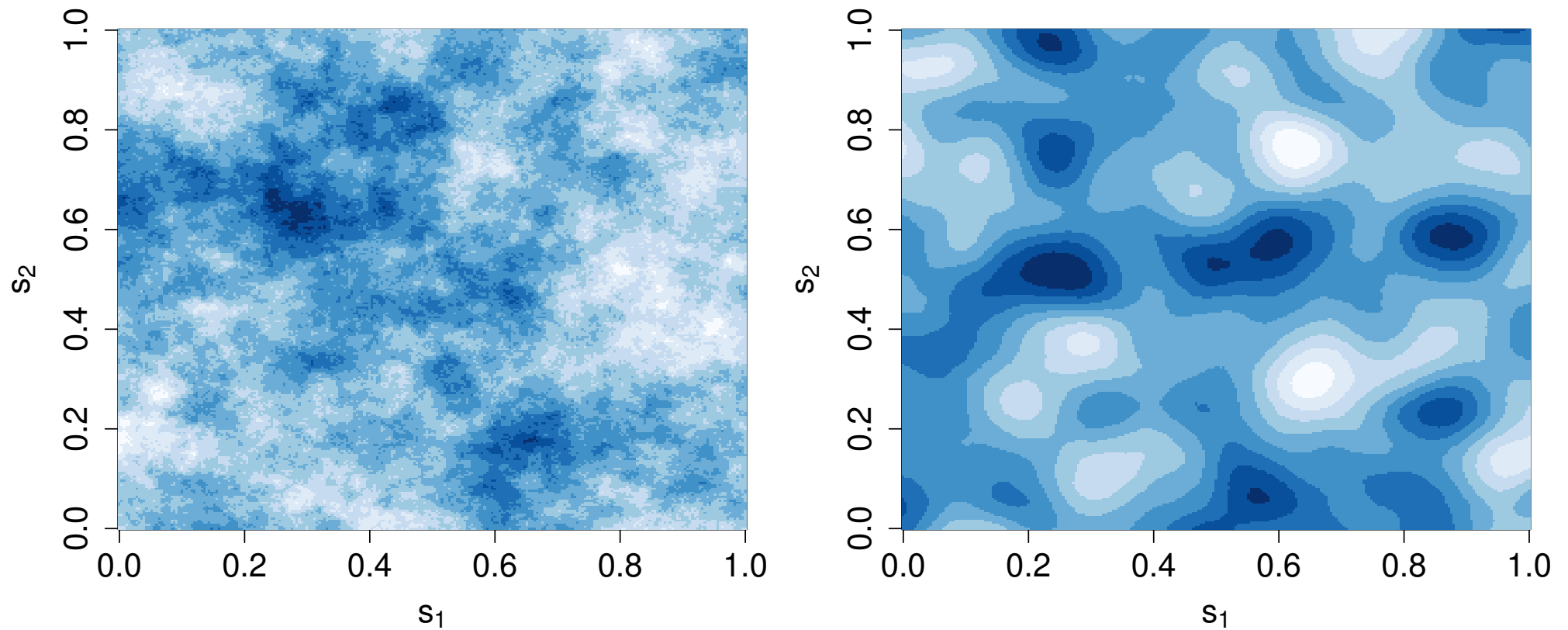
## (Unconditional) Simulations

- It is rather straightforward to simulate a Gaussian process at a moderate number of locations, e.g.,  $k \leq 3000$ , from the **Cholesky decomposition** of the covariance matrix.
- More precisely for any  $\mathbf{s} = (s_1, \dots, s_k) \in \mathcal{X}$ , we have

$$Y(\mathbf{s}) \stackrel{d}{=} \mu(\mathbf{s}) + C(\mathbf{s})^\top \boldsymbol{\varepsilon}, \quad \Sigma(\mathbf{s}) = C(\mathbf{s})^\top C(\mathbf{s}),$$

where  $\boldsymbol{\varepsilon}$  is a vector of  $k$  independent standard normal random variables.

**i** More sophisticated techniques, e.g., turning bands, circulant embedding methods, exist to get faster simulations on large (gridded) number of locations.



**Figure 10:** *Two realizations of a random fields with a powered exponential correlation function. Left:  $\kappa = 1$ . Right:  $\kappa = 2$ .*

# Conditional simulations

- Estimating **areal quantities** from kriging may be too smooth.
- **Conditional simulations** can be used to get Monte Carlo estimate (and thus the entire distribution) of it.
- Conditional simulations are random simulations that **honors some constraints**, e.g., simulating from

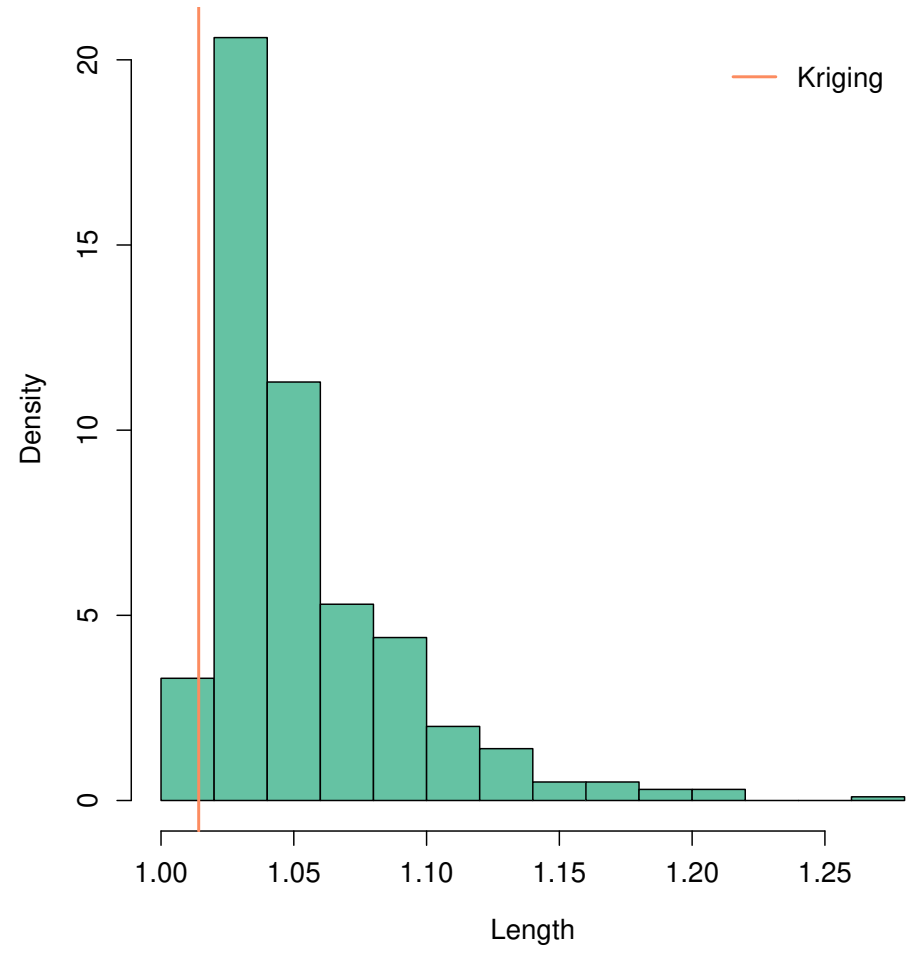
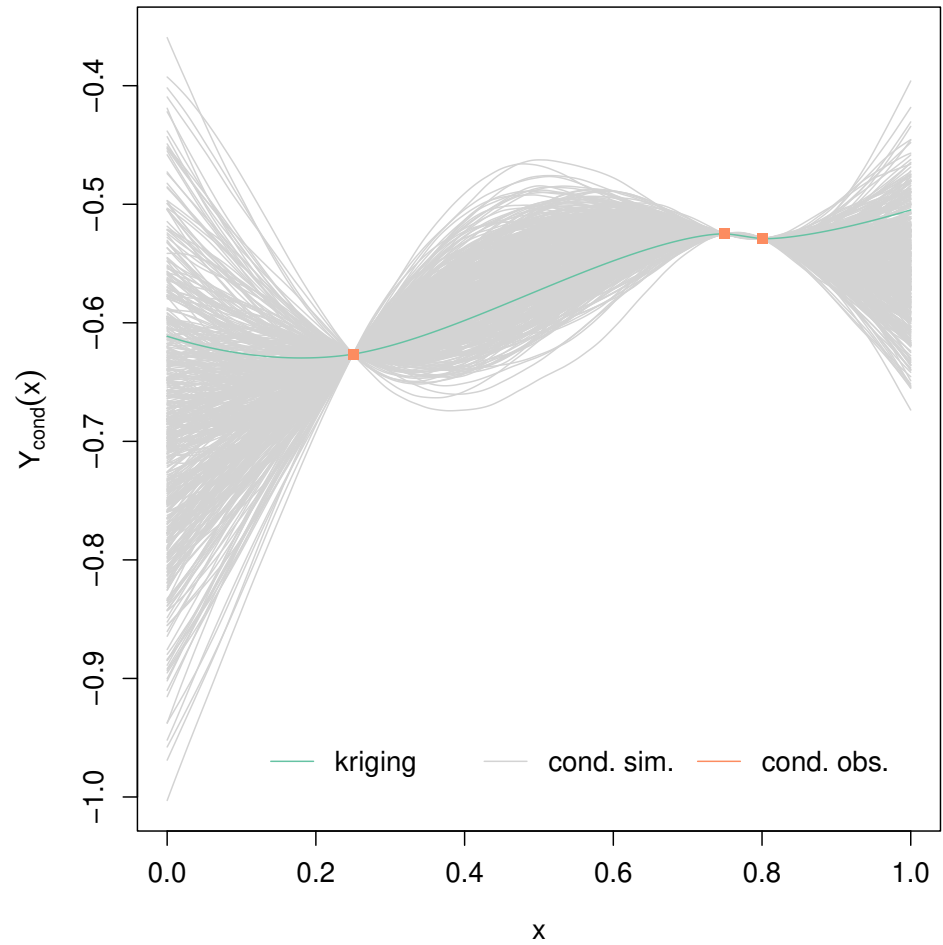
$$Y(s_*) \mid Y(\mathbf{s}) = \mathbf{y},$$

where  $\mathbf{y}$  is the vector of held fixed values at prescribed location  $\mathbf{s}$ .

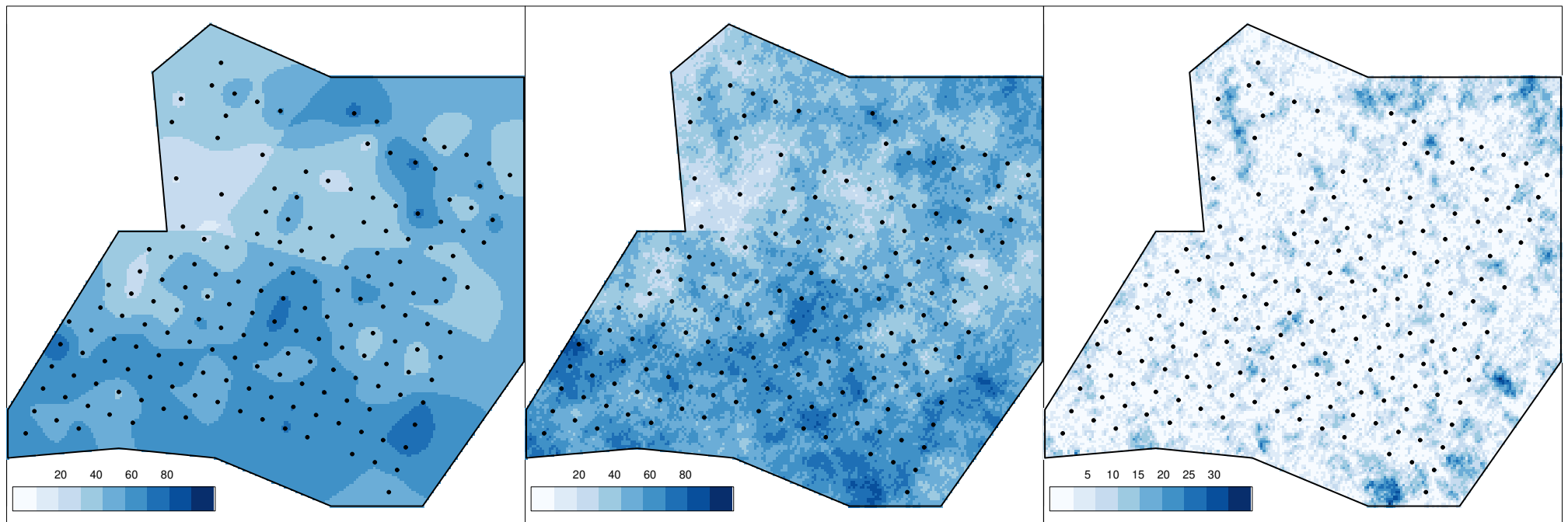
- Under the Gaussian setting, one can use the decomposition

$$Y(s_*) \mid \{Y(\mathbf{s}) = \mathbf{y}\} \stackrel{d}{=} \underbrace{Y_{\text{krig}}(s_*)}_{\text{kriging of } Y} + \tilde{Y}(s_*) - \underbrace{\tilde{Y}_{\text{krig}}(s_*)}_{\text{kriging of } \tilde{Y}},$$

where  $\tilde{Y}$  is an independent copy of  $Y$ .

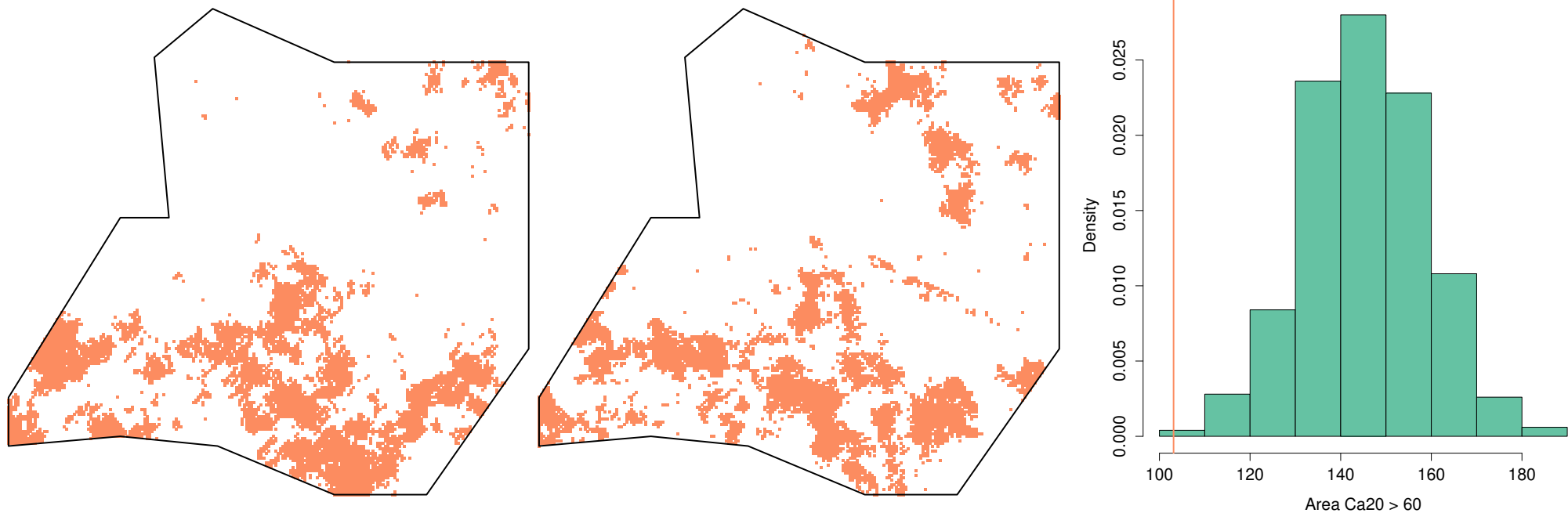


**Figure 11:** Comparison between conditional simulations and kriging. Right: length of the curve estimated from kriging and conditional simulations.



**Figure 12:** Comparison between kriging (left) and a conditional simulation (middle). Right: absolute difference of the two.





**Figure 13:** *Left and middle: Two sampled level sets with  $u_{crit} = 60$ . Right: Distribution of the expected level set area from conditional simulations (histogram) and kriging (vertical line).*

👉 As expected, the kriging-based estimator underestimates.

1. Framework

---

2. Inference

---

3. Model-based  
geostatistics

---

4. Simulation

---

5. Bayesian  
▷ hierarchical models

---

6. Big data

---

## 5. Bayesian hierarchical models

# Hierarchical models: Motivations

---

- Data often depict different layers of variation, that one has to modelled:
  - success of surgical interventions may depend on:
    - ▷ patients (age/state of health) within
    - ▷ surgeons (different experience/skill) within
    - ▷ hospitals (different environments/skill of nursing staff)
  - student's marks may depend on:
    - ▷ the classroom, which depend on
    - ▷ school, which depend on
    - ▷ school districts. . .
- For each layer we observed draws from their respective population, e.g., patients/doctors drawn from a given hospital.
- It suggests having different layer of randomness.

# Hierarchical model

**Definition 10.** A statistical model  $\{f(y; \psi) : x \in \mathbb{R}^p, \psi \in \psi\}$  is a **hierarchical model** if we have

$$f(y; \psi) = \int f_1(y | z_1) f_2(z_1 | z_2) \cdots f_d(z_{d-1} | z_d) f(z_d) dz_1 \cdots dz_d.$$

In the above expression, the  $z_j$ 's are called **latent variables**.

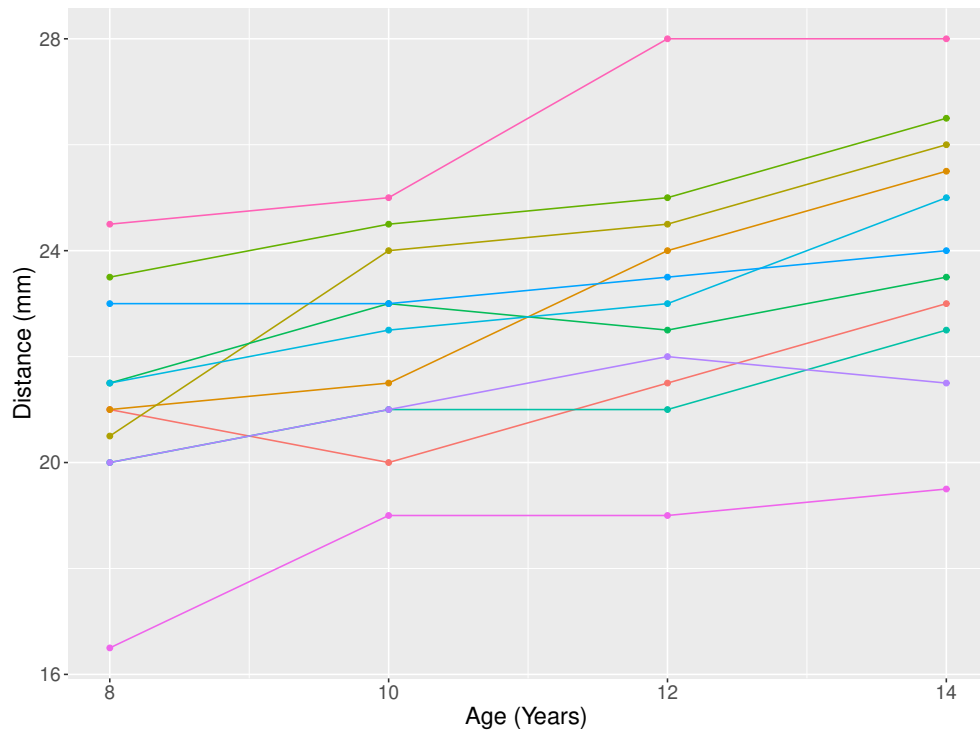
- The above integral representation often has **no closed form** and **dedicated strategies** for model fitting are needed, e.g.,

**Frequentist** EM-type algorithms

**Bayesian** Monte Carlo Markov Chain algorithms

 We will give a short focus on Bayesian statistics and MCMC algorithms in a moment.

**Example 2.** X-rays of the children's skulls were shot by orthodontists to measure the distance from the hypophysis to the pterygomaxillary fissure. Shots were taken every two years from 8 years of age until 14 years of age.



$$Y_{ij} \mid b_j \stackrel{\text{ind}}{\sim} N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$

$$b_j \sim N(0, \sigma_b^2),$$

- $Y_{ij}$ : distance
- $x_{ij}$ : age of subject  $j$  at index  $i$

**Figure 14:** *The data collected by the orthodontists.*

---

In our orthodontist example, the random variables  $b_j$  are latent variables and the integral representation is

$$f(y_{ij}; \psi) = \int \varphi(y_{ij}; \beta_1 + b_j + \beta_2 x_{ij}, \sigma^2) \varphi(b_j; 0, \sigma_b^2) db_j,$$

where  $\varphi(\cdot; \mu, \sigma^2)$  denotes the Gaussian density with mean  $\mu$  and variance  $\sigma^2$ .

# A spatial hierarchical model

- Recall that we typically assume a linear structure on the mean function of the Gaussian process, i.e.,

$$\mu(s) = f(s; \boldsymbol{\beta}) = \mathbf{x}(s)^\top \boldsymbol{\beta},$$

but in many situations it is **unrealistic** and we need to relax it.

- To bypass this hurdle, one way is to use hierarchical models where we now have

$$\begin{aligned} \mu(\cdot) \mid \varepsilon(\cdot) &\sim \text{Gaussian Process}(f(\cdot; \boldsymbol{\beta}) + \varepsilon(\cdot), \gamma) \\ \varepsilon(\cdot) &\sim \text{Gaussian Process}(0, \gamma_\varepsilon). \end{aligned}$$

 It enables departures from the inflexible linear structure by adding some **noise**.

# Directed acyclic graph (DAG)

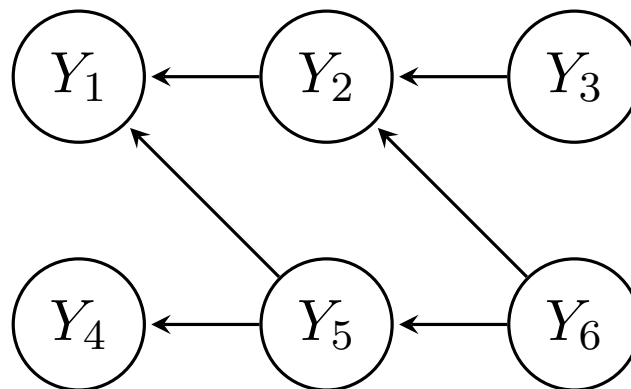
**Definition 11.** A **directed acyclic graph (DAG)** is a graphical model that represents a hierarchical dependence structure, i.e., for all  $i \in V$

$$Y_i \perp \text{non descendants of } Y_i \mid \text{parents of } Y_i.$$

It is **directed** because it is a directed graph and **acyclic** because it is impossible to start from a node and get back to it using a path of arrows.

**Example 3.** The hierarchical dependence structure

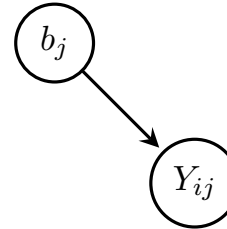
$f(y) = f(y_1 \mid y_2, y_5)f(y_2 \mid y_3, y_6)f(y_3)f(y_4 \mid y_5)f(y_5 \mid y_6)f(y_6)$  gives:





**Example 4.** Recall our model for the distance from the hypophysis to the pterygomaxillary fissure:

$$Y_{ij} \mid b_j \stackrel{\text{ind}}{\sim} N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$
$$b_j \sim N(0, \sigma_b^2).$$



# Factorization of a DAG and full conditional distributions

- Since, by definition, for any DAG  $G = (V, E)$  we have

$$f(y) = \prod_{j \in V} f(y_j \mid \text{parents of } y_j).$$

- Hence the full conditional distributions write

$$\begin{aligned} f(y_j \mid \dots) &\propto f(y_j), && (\propto \text{ stands for up to a multiplicative constant}) \\ &\propto \prod_{i \in V} f(y_i \mid \text{parents of } y_i) \\ &\propto f(y_j \mid \text{parents of } y_j) \prod_{\substack{i \in V: \\ y_i \text{ child of } y_j}} f(y_i \mid \text{parents of } y_i), \end{aligned}$$

where ... stands for all the other variables.

# Bayesian statistical models

**Definition 12.** A parametric family of functions  $\{f(x; \psi) : x \in E, \psi \in \Psi\}$  is a **statistical model** if, for any  $\psi \in \Psi$ ,  $x \mapsto f(x; \psi)$  is a probability density function on  $E$ . The sets  $\Psi$  and  $E$  are respectively called **parameter space** and **observational space**.

The above model is said to be **parametric** if  $\dim(\Psi) < \infty$ .

 Treat parameters as random variables.

**Definition 13.** If we further place a **prior distribution**  $\pi$  on the parameter  $\psi$  we are dealing with a **Bayesian statistical model**  $(f, \pi)$  and the parameters of the prior distribution  $\pi$  are called the **hyper-parameters**.

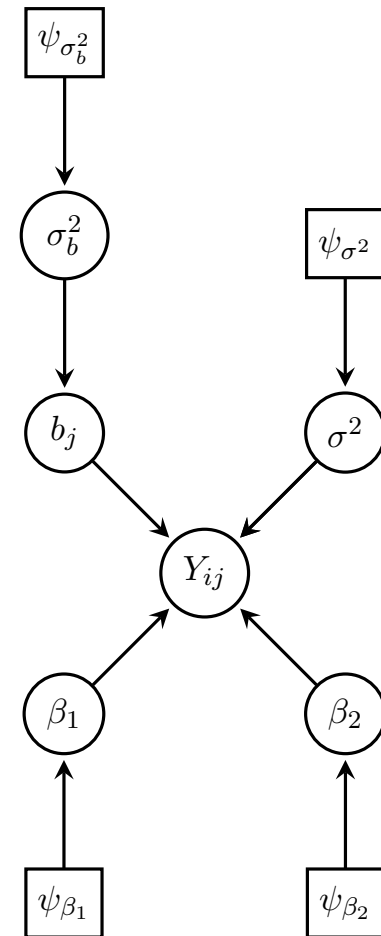
Recall our model for the distance from the hypophysis to the pterygomaxillary fissure:

$$Y_{ij} \mid b_j, \beta_1, \beta_2, \sigma^2 \stackrel{\text{ind}}{\sim} N(\beta_1 + b_j + \beta_2 x_{ij}, \sigma^2),$$

$$b_j \mid \sigma_b^2 \sim N(0, \sigma_b^2),$$

now with prior distribution

$$\pi(\theta) = \pi(\beta_1)\pi(\beta_2)\pi(\sigma_b^2)\pi(\sigma^2).$$



# Posterior distributions

**Definition 14.** Given a sample  $\mathcal{D}_n = (y_1, \dots, y_n)$  and a Bayesian model  $(f, \pi)$ . The main focus in Bayesian inference is on the **posterior distribution**

$$\pi(\psi \mid \mathcal{D}_n) = \frac{f(\mathcal{D}_n \mid \psi)\pi(\psi)}{\int f(\mathcal{D}_n \mid \psi)\pi(\psi)d\psi},$$

provided that the **marginal distribution (normalizing constant)**

$$m(\mathcal{D}_n) = \int f(\mathcal{D}_n \mid \psi)\pi(\psi)d\psi < \infty.$$

 It is often very convenient to work up to a multiplicative factor independent of  $\psi$  since it will cancel out in the above expression. In such situations we will write

$$\pi(\psi \mid \mathcal{D}_n) \propto f(\mathcal{D}_n \mid \psi)\pi(\psi).$$

# Pointwise estimation and prediction

---

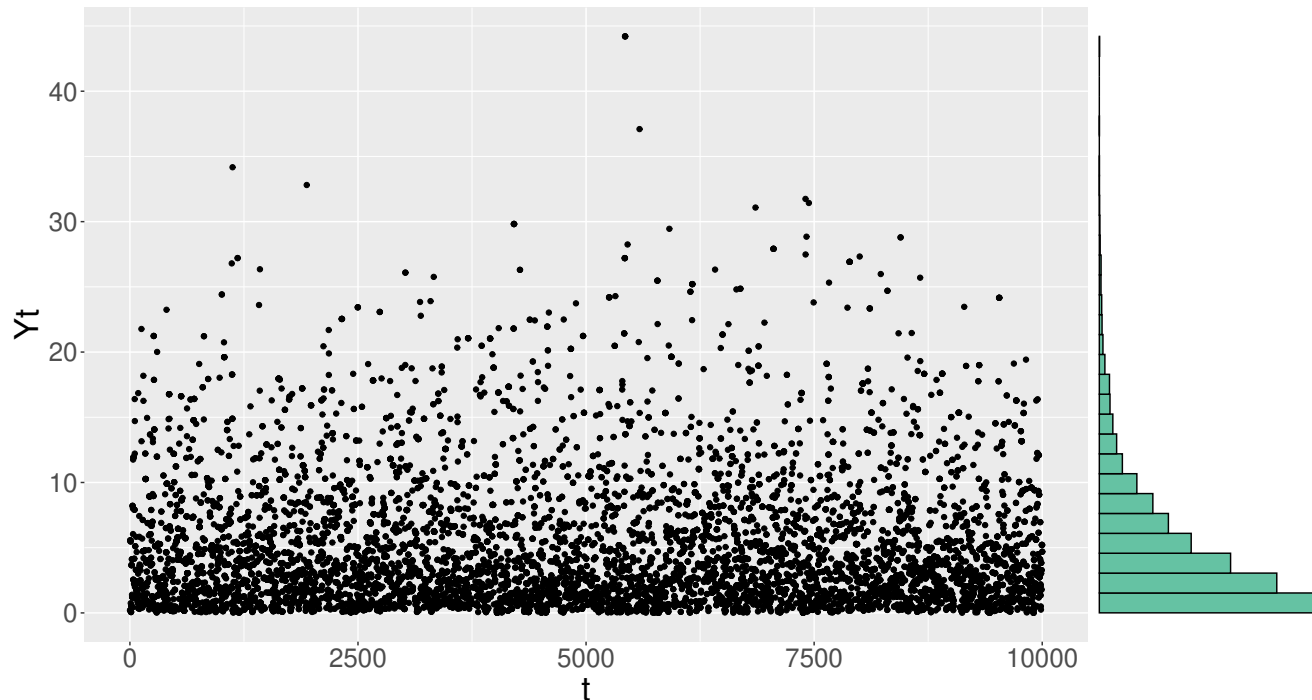
- To mimic **point estimates** in the frequentist world, model parameters may be estimated from the **posterior mean, median or mode**.
- The analogue of confidence intervals are **credible intervals**, i.e.,

$$\Pr_{\pi}(\psi \in I_{\alpha} \mid \mathcal{D}_n), \quad I_{\alpha} \text{ credible interval, } \alpha \text{ level.}$$

- Prediction for a future observation  $y_*$  is usually done from the **predictive posterior distribution**

$$\pi(y_* \mid \mathcal{D}_n) = \int f(y_* \mid \mathcal{D}_n, \psi) \pi(\psi \mid \mathcal{D}_n) d\psi, \quad \mathcal{D}_n \text{ dataset.}$$

MCMC algorithms output a (dependent) sample from a **prespecified target distribution**.



**Figure 15:** A Markov chain whose stationary distribution is the  $\text{Exponential}(5)$ .

It is **not** specific to Bayesian statistics—widely used in this setting though.

- In a Bayesian setting, the target distribution is the posterior distribution.
- For Bayesian hierarchical models a sensible choice is the Gibbs sampler
- It consists in sampling successively from the full posterior distributions

$$\pi(\psi_j | \dots), \quad \text{where “}\dots\text{” means all the rest}$$

 In our orthodontist example, we sequentially sample from

$$\pi(\beta_1 | \dots), \quad \pi(\beta_2 | \dots), \quad \pi(\sigma^2 | \dots), \quad \pi(\sigma_b^2 | \dots), \quad \pi(b_j | \dots)$$



# Conditional independence model

- The conditional independence assumption states that the data are independent given the parameter model, e.g.,

$$Y(s) \mid \{\mu(\cdot), \sigma^2(\cdot)\} \stackrel{\text{ind}}{\sim} N \{\mu(s), \sigma^2(s)\}, \quad s \in \mathcal{X}$$
$$\mu(\cdot) \sim \text{Gaussian Process}$$
$$\log \sigma^2(\cdot) \sim \text{Gaussian Process}$$

- On the **data layer** we substitute a multivariate distribution for a product of univariate ones.

 The conditional assumption is appealing because one can easily switch the distribution in the data layer.

---

**Example 5.** If your data are pointwise block maxima you may want to use for the data layer the Generalized Extreme Value distribution (GEV), i.e.,

$$Y(s) \mid \{\mu(s), \sigma(s), \xi(s)\} \stackrel{\text{ind}}{\sim} \text{GEV}\{\mu(s), \sigma(s), \xi(s)\}, \quad s \in \mathcal{X}$$

$\mu(\cdot), \log \sigma(\cdot), \xi(\cdot) \sim$  Gaussian processes

with prior distributions on the Gaussian processes parameters.

The full conditional distributions are

$$\pi(\mu(s_j) \mid \dots), \quad \pi(\sigma(s_j) \mid \dots), \quad \pi(\xi(s_j) \mid \dots), \quad j = 1, \dots, k$$

$$\pi(\mu_\mu \mid \dots), \quad \pi(\sigma_\mu \mid \dots), \quad \pi(\xi_\mu \mid \dots)$$

$$\pi(\gamma_\mu \mid \dots), \quad \pi(\gamma_\sigma \mid \dots), \quad \pi(\gamma_\xi \mid \dots),$$

where  $\mu_\cdot$  and  $\gamma_\cdot$  are the mean function and variogram of the Gaussian processes whose parameters are updated in turn.

1. Framework

2. Inference

3. Model-based  
geostatistics

4. Simulation

5. Bayesian  
hierarchical models

▷ 6. Big data

## 6. Big data

- Broadly speaking, there are two different type of “big data”:

Type I when the number of covariates  $p$  is large

Type II when the sample size  $n$  is large

- From a statistical standpoint, Type I is the most challenging as parameter estimation is tricky or even impossible.
- Type II induces computational burden and we need numerical/optimization tricks.

💣 You can have the two of us!



# High-dimensional setting, a.k.a., big data I

---

- Fitting a Gaussian process when the **number of location is large**, i.e.,  $k \gg 1$ , is challenging.
- As stated previously, the most CPU demanding parts of the likelihood is the evaluation of  $|\Sigma(\mathbf{s})|$  and the Mahalanobis distance  $a^2(\mathbf{s})$ .
- To bypass this hurdle one can (at least) use one of the following options:
  - **composite likelihoods**
  - **covariance tapering**

# Composite likelihood

**Definition 15.** A **composite log-likelihood** is a linear combination of log-likelihoods of “**smaller dimensions**”.

**Example 6.** The **independent composite likelihood** uses only univariate densities, i.e.,

$$\ell_{\text{ind}}(\psi; \mathcal{D}_n) = \sum_{j=1}^k \omega_j \underbrace{\sum_{i=1}^n \log f\{y_i(s_j); \psi\}}_{\text{univariate log-likelihood}},$$

and the **pairwise composite likelihood** makes use of bivariate densities, i.e.,

$$\ell_{\text{pair}}(\psi; \mathcal{D}_n) = \sum_{j=1}^{k-1} \sum_{\ell=j+1}^k \omega_{j,\ell} \underbrace{\sum_{i=1}^n \log f\{y_i(s_j), y_i(s_\ell); \psi\}}_{\text{bivariate log-likelihood}},$$

where  $\omega_j$  and  $\omega_{j,\ell}$  are (positive) weights.

# Covariance Tapering

---

- Computational burden heavily relies on the inversion of the covariance matrix  $\Sigma(\mathbf{s})$
- Tapering consists in modify  $\Sigma(\mathbf{s})$  to get a sparse structure, i.e., many zeros.

**pros** efficient computation using sparse matrix algebra

**cons** approximate inference

---

**Proposition 4.** *Let  $f_1$  and  $f_2$  be two definite positive functions. Then the function*

$$f: s \mapsto f_1(s)f_2(s)$$

*is definite positive.*

- We can get a sparse version of  $\Sigma(\mathbf{s})$  from the above property. More precisely

$$\Sigma(\mathbf{s})_{\text{tap}} = \Sigma(\mathbf{s}) \odot \Sigma_c(\mathbf{s}),$$

where  $\odot$  stands for the direct product, i.e., componentwise, and  $\Sigma_c(\mathbf{s})$  is a covariance matrix obtained from a covariance function with compact support.

- The associated Cholesky decomposition will be sparse as well (up to a sensible permutation)



## Two taper approximation

- The tapering introduced above induce a bias in the parameter estimation.
- The bias can be severe if the tapering range is small compared to the practical range—prediction are slightly affected though.
- One may rather use a two-taper version where the Mahalanobis distance is now substituted with

$$\tilde{a}^2(\mathbf{s}, \mathbf{y}) = \mathbf{y}^\top \left[ \{\Sigma(\mathbf{s}) \odot \Sigma_c(\mathbf{s})\}^{-1} \Sigma_c(\mathbf{s}) \right] \mathbf{y}.$$

- The two-taper strategy yields unbiased parameters estimation
- The price to pay is that the computational cost is larger than the one-taper version

 Another approach consists in using a truncated SVD, as for PCA.

# High-dimensional setting, a.k.a., big data II

---

- Fitting a Gaussian process when the number of replicates is large, i.e.,  $n \gg 1$ , is challenging.
- In such situations evaluation of the likelihood is demanding due to the sum in  $n$ , i.e.,

$$\ell(\psi; \mathcal{D}_n) = \sum_{i=1}^n \log \varphi(\mathbf{y}_i; \boldsymbol{\mu}, \Sigma).$$

- Two (related) possible approaches are:
  - mini-batch gradient ascent
  - stochastic gradient ascent

## Gradient ascent (reminder)

**Proposition 5.** Let  $\psi_0$  be an initial state. The sequence

$$\psi_{n+1} = \psi_n + \eta \nabla J(\psi_n), \quad n \geq 0,$$

will converge to a local maxima (if it does), where  $\eta$  is known as the *step size* (*learning rate* if you're a noob!).

- The step size can be *adaptive*, i.e.,  $\eta$  now depends on  $t$  and / or  $\psi_n$ .
- Current popular choices are *Nesterov adaptive schemes*, i.e., so called *momentum*, where

$$\psi_{n+1} = \psi_n + \mu_n v_n + \eta_n \nabla J(\psi_n), \quad v_n \text{ some "measure of velocity"}.$$

 If minimizing, use gradient *descent*  $\psi_{n+1} = \psi_n - \eta \nabla J(\psi_n)$ .

# Mini-batch gradient ascent

- Consider the following optimization problem

$$\arg \max_{\psi \in \Psi} J(\psi), \quad J(\psi) = \sum_{i=1}^n J_i(\psi).$$

- If  $n \gg 1$ , evaluation of  $J$  is **prohibitive** and prevent the use of gradient ascent.
- One can minimize the CPU cost using **mini-batch gradient ascent**

$$\psi_{n+1,b+1} = \psi_{n,b} + \eta \sum_{i \in B_b} \nabla J_i(\psi_{n,b}), \quad b = 1, \dots, B,$$

and where by convention  $\psi_{n+1,1} = \psi_{n,B+1}$  and

$$\cup_{b=1,\dots,B} B_b = \{1, \dots, n\}, \quad B_b \cap B_{b'} = \emptyset,$$

i.e., a partition of  $\{1, \dots, n\}$ .

## If you want to show off (a bit)

---

- The gradient update step is done after browsing each batch  $B_b$
- The computational cost is thus reduced.
- One loop over the entire data set is called an **epoch**.

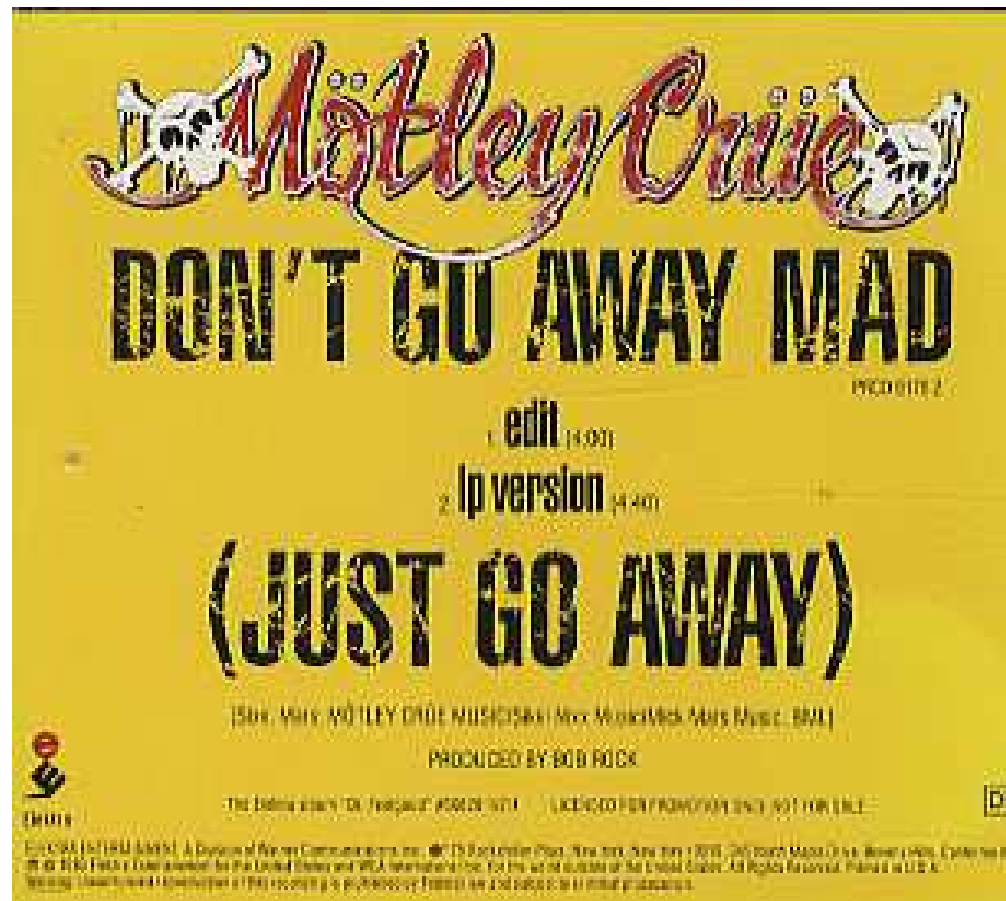
# Stochastic gradient ascent

---

- **Stochastic gradient ascent** is somehow similar to mini-batch gradient ascent as it compute the gradient on subset of the dataset  $\mathcal{D}_n$ .
- The main difference is that these subsets are now **random**.
- The basic stochastic gradient ascent scheme is

$$\psi_{n+1} = \psi_n + \eta \nabla J_I(\psi_n), \quad I \sim \text{Unif}\{1, \dots, n\}.$$

- Some generalization are possible:
  - random mini-batches where we drawn random batches
  - use a other distribution than the discrete uniform.
- Stochastic gradient ascent will converge to a local maxima as long as the learning rate goes to 0.
- Its randomness may helps escaping from local maxima.



Not enough details. . . you have plenty of keywords so that you can go deeper into the theory!