Geostatistics

Mathieu Ribatet—Full Professor of Statistics

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_{Craft} 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 $s_* \in \mathcal{X}$.

Different type of spatial data

- geostatistical data: data are defined continuously on $\mathcal{X}$, e.g., rainfall;
- punctual data: the data are points falling randomly over some space $\mathcal{X}$, e.g., tree locations.
- lattice data: data are aggregated over sub-regions, e.g., number of citizen in counties.

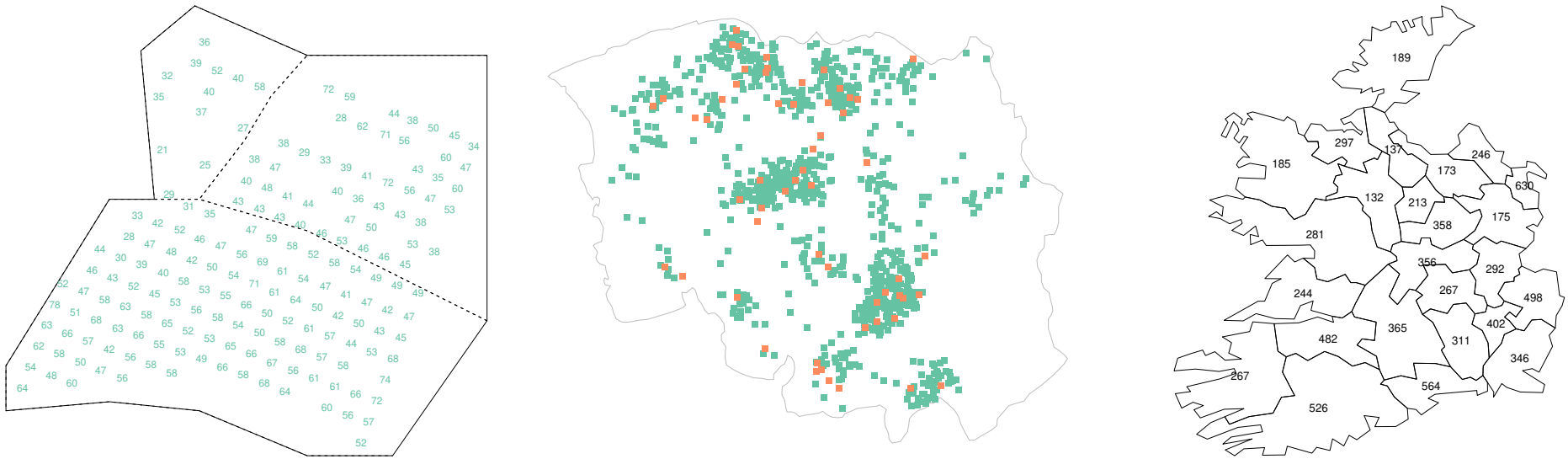
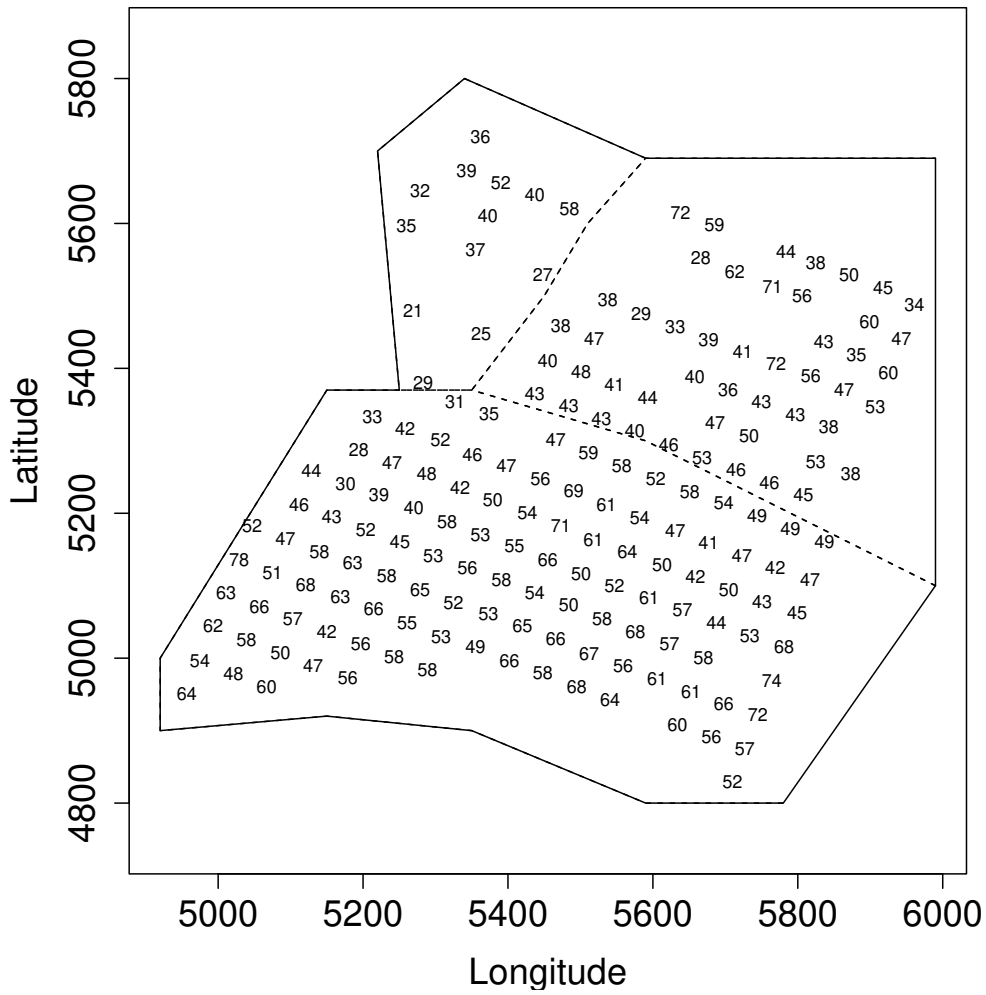


Figure 1: *The three different type of spatial data. From left to right: geostatistical (Calcium concentration), punctual (location of lung and larynx) and lattice data.*

The Ca_{20} data set



Focus is on geostatiscal data only!

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 $\{Y(s) \colon s \in \mathcal{X}\}$ is completely characterised from its finite dimensional distribution functions, i.e., for any $k \geq 1$ and $s_1, \ldots, s_k \in \mathcal{X}$*

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

(provided they satisfy the hypothesis of the Kolmogorov extension theorem, i.e., invariance to permutation and consistent marginalisation)

Strictly stationary processes

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

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

where A_j are Borel sets.

☞ In practice, strict stationarity 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., $\mathsf{Var}[Y(s)] < \infty$ for all $s \in \mathcal{X}$.

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

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

  - the covariance function

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

Weak stationarity and isotropy

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

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

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

$$\{Y(Rs) \colon s \in \mathcal{X}\} \stackrel{\mathsf{d}}{=} \{Y(s) \colon s \in \mathcal{X}\}. \qquad \texttt{(rotation invariance)}$$

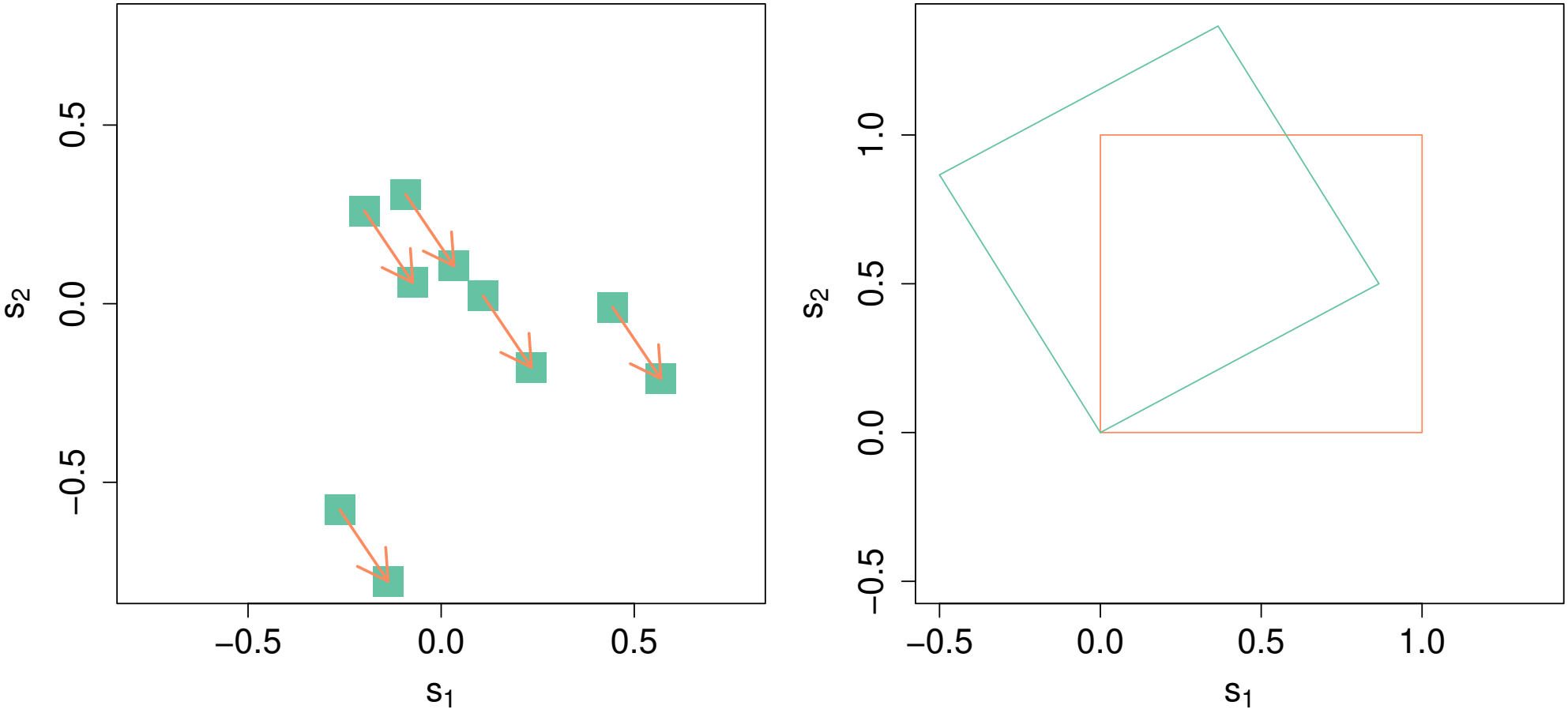


Figure 2: *Illustration of stationarity and isotropy.*

Consequences

- If a process is stationary we have

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

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

$$\text{Cov}\{Y(s), Y(s')\} = \text{Cov}\{Y(s'), Y(s)\}.$$

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

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

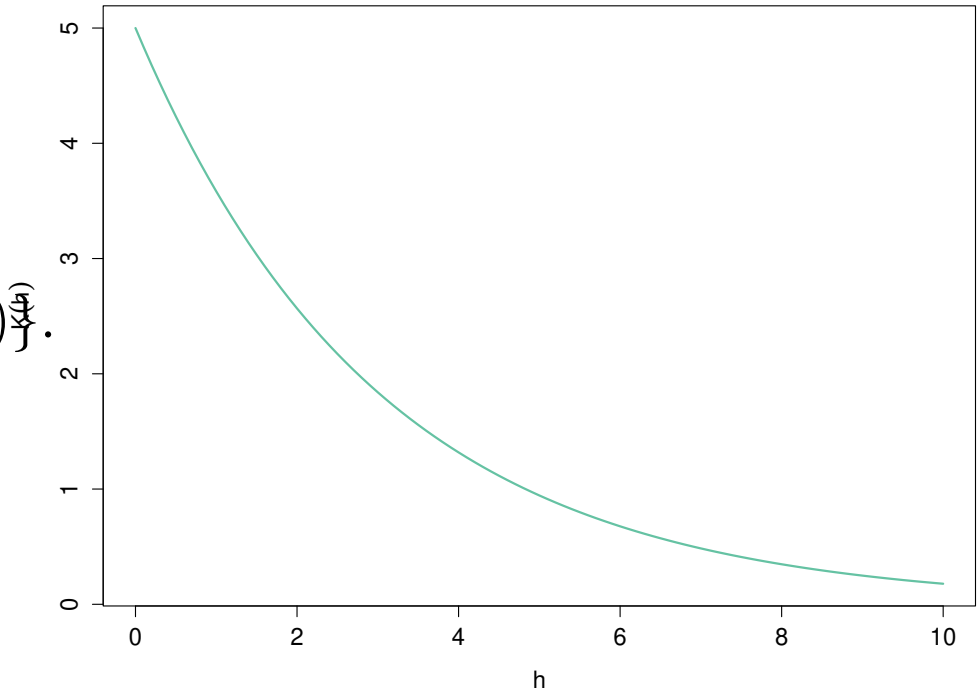


Figure 3: *Plot of a stationary isotropic covariance function. What is $K(0)$?*

Processes with stationary increments

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

$$Y(s+h) - Y(s) \overset{\text{fin}}{=} Y(h) - Y(o),$$

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 event 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 $\mathcal{X} = \mathbb{Z}$

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

It has indeed stationary increments since

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

but is not stationary. Even worse we have $\text{Var}\{Y(s)\} \to \infty$ as $s \to \infty$.

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 increments we get fractional Brownian processes.

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 stationary increment processes we rather consider the semi–variogram

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

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

$$\gamma(h) \longrightarrow K(o,o), \qquad \|h\| \to \infty, \quad \texttt{(as long as } \rho(h) \to 0\texttt{)}$$

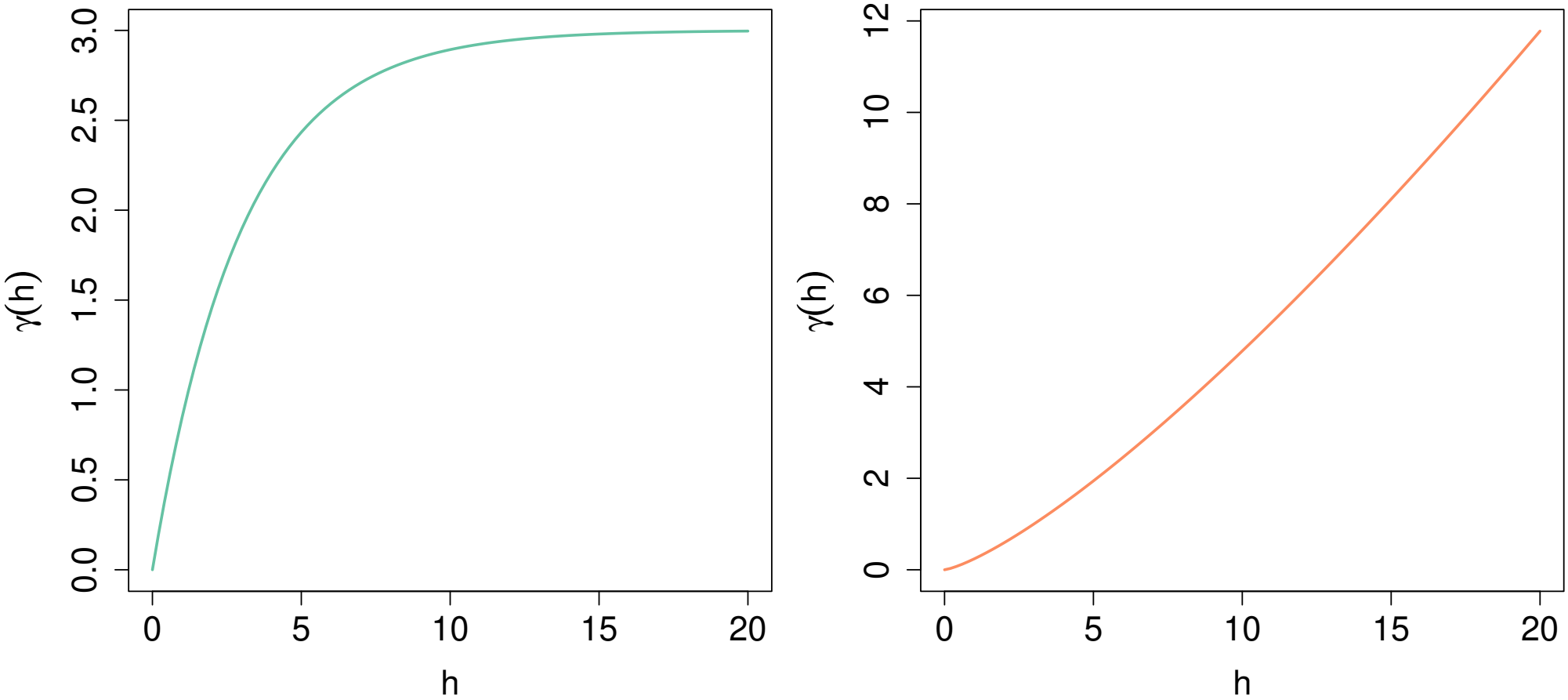


Figure 4: *Bounded (left) and unbounded (right) semi–variograms.* *If it exists, what is $\gamma(\infty)$?*

Some isotropic stationary correlation functions and variograms

Family	$\rho(h)$	$\gamma(h)$	Support
Exponential	$\exp(-h/\lambda)$	$1-\exp(-h/\lambda)$	$\lambda > 0$
Gaussian	$\exp\left\{-(h/\lambda)^2\right\}$	$1-\exp\left\{-(h/\lambda)^2\right\}$	$\lambda > 0$
Stable / Powered exponential	$\exp\left\{-(h/\lambda)^\kappa\right\}$	$1-\exp\left\{-(h/\lambda)^\kappa\right\}$	$\lambda > 0, 0 \leq \kappa \leq 2$
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 λ and κ are known as the range and smooth parameters.
- Associated covariance functions are derived using a sill parameter τ, i.e.,

$$K(h) = \tau\rho(h), \qquad \tau > 0. \qquad (\tau = \mathtt{K}(\mathtt{o}))$$

- 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 $\rho(h_p) = 0.05$.

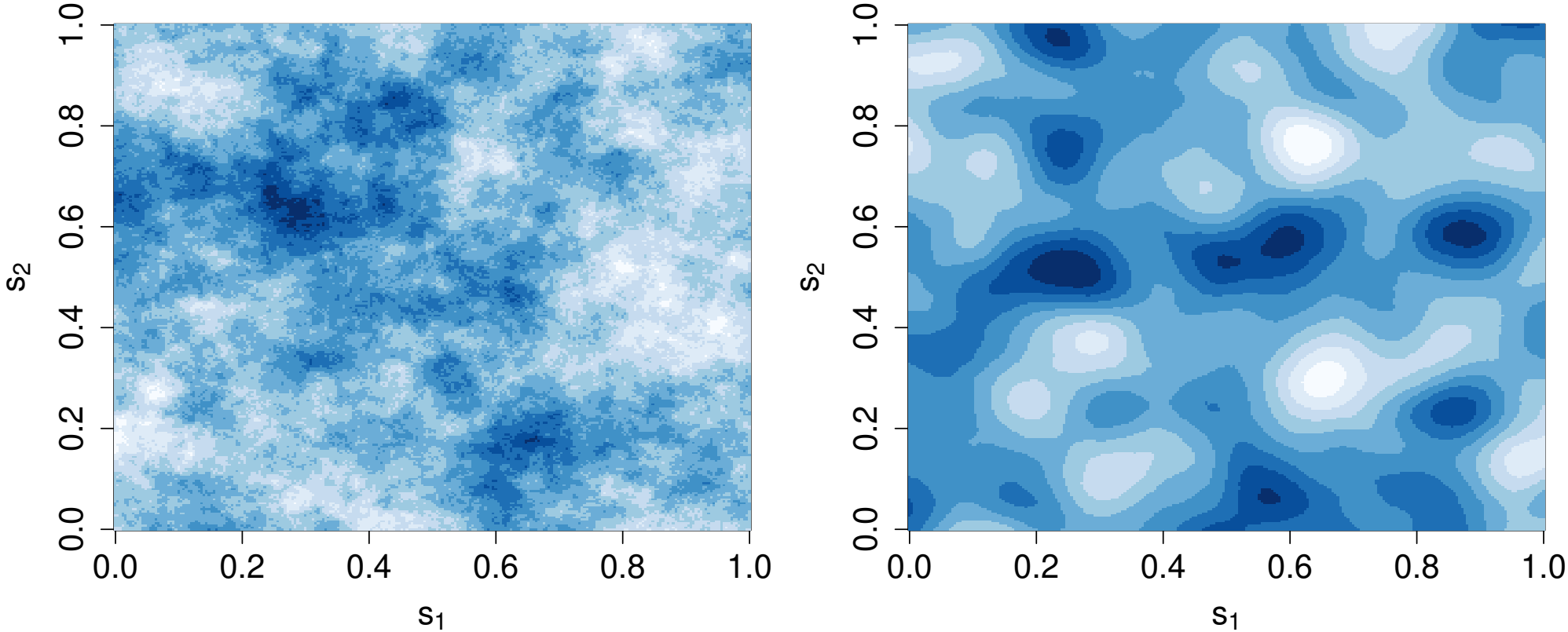
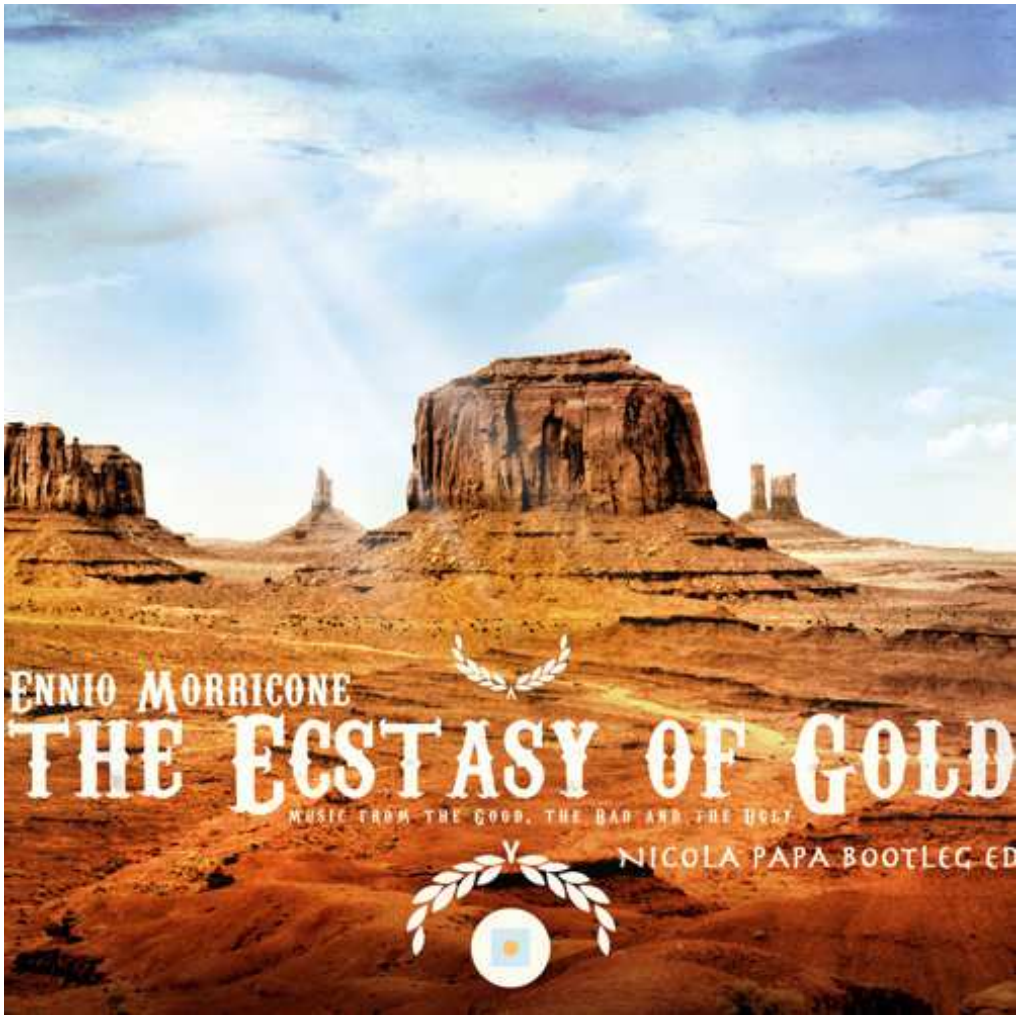


Figure 5: *Two realisations of a random fields with a powered exponential correlation function. Left:* $\kappa = 1$. *Right:* $\kappa = 2$.


ENNIO MORRICONE
THE ECSTASY OF GOLD
MUSIC FROM THE GOOD, THE BAD AND THE UGLY
NICOLA PAPA BOOTLEG ED

- 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 orders are possible!

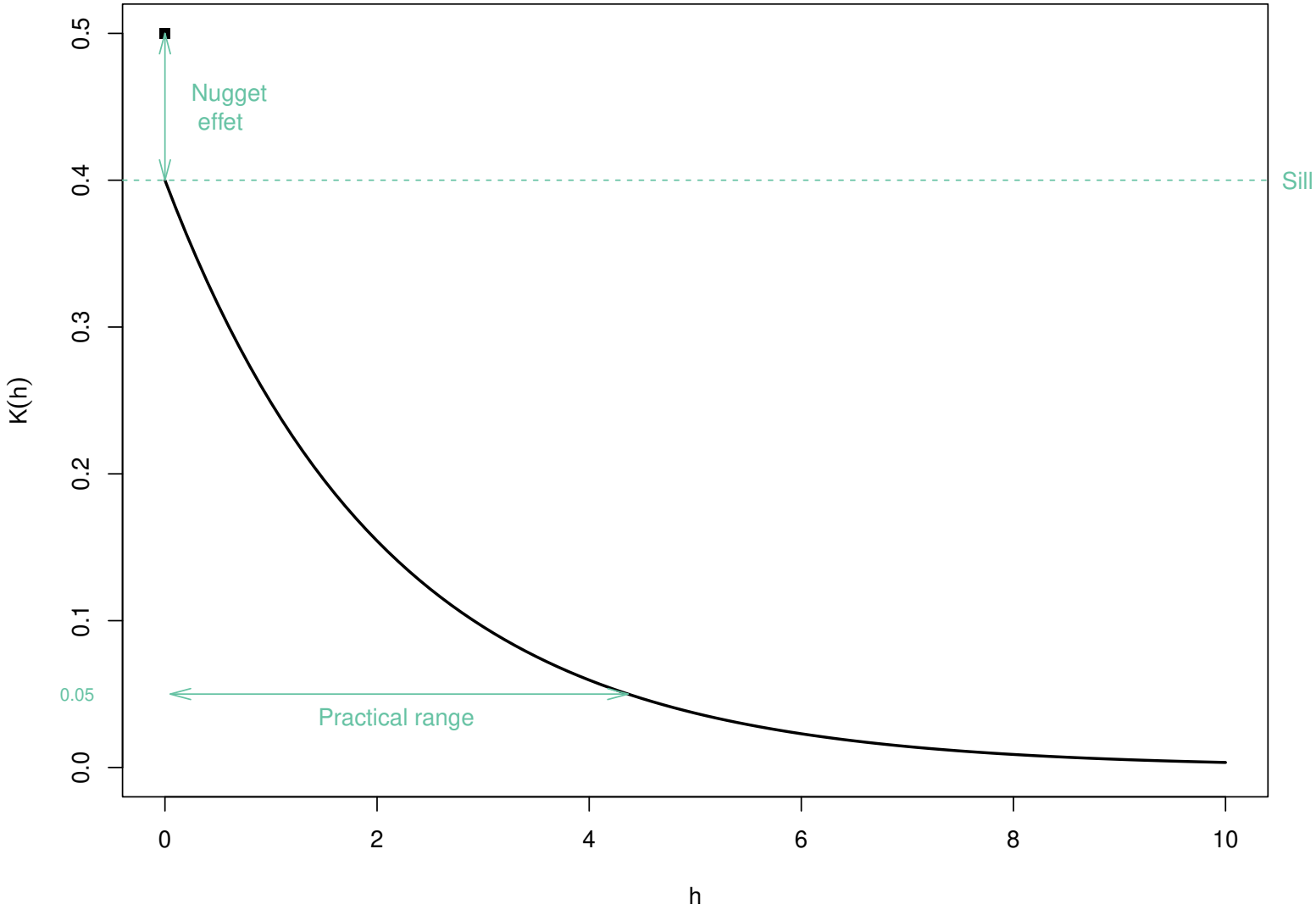


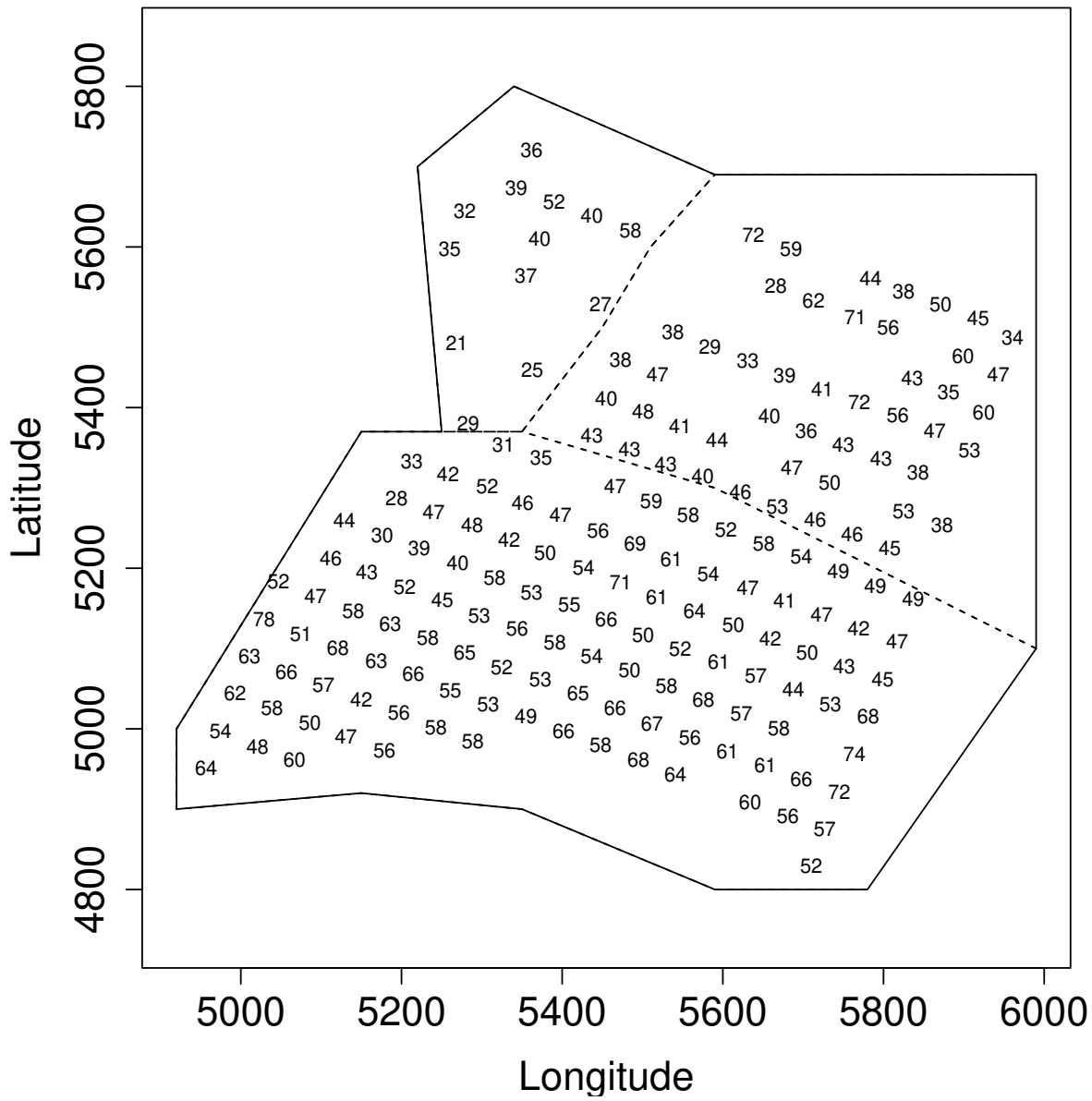
Figure 6: *Illustration of the nugget effect, the sill parameter and the practical range.*

2. Inference

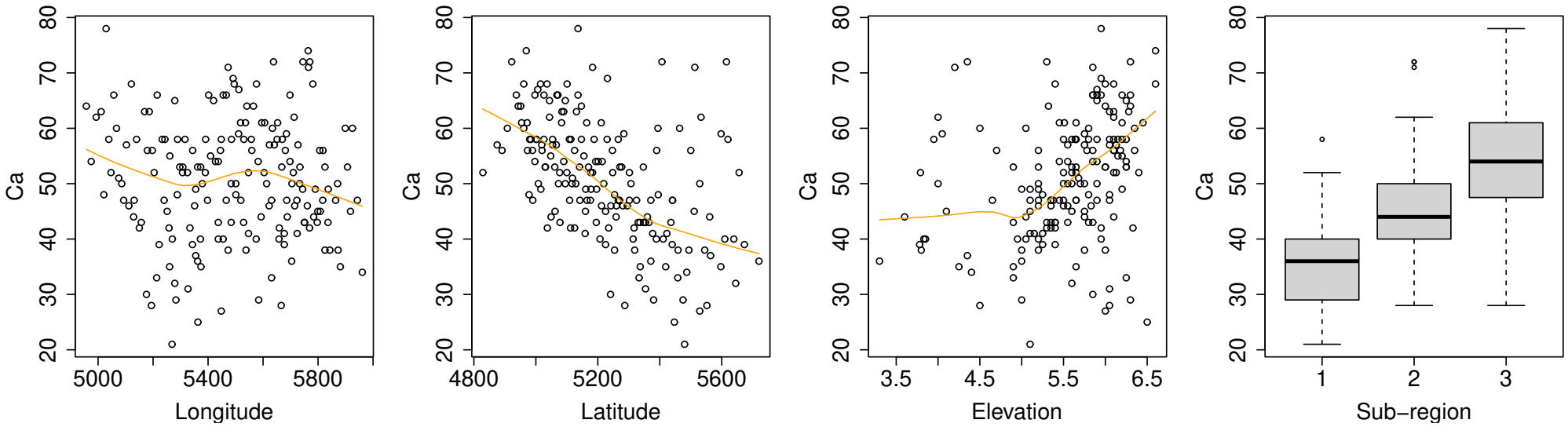
Descriptive analysis

- Before trying to model the data we need to check whether the data can safely be assumed stationary / isotropic / ...
- Essentially we start with a descriptive analysis which, four 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, ...

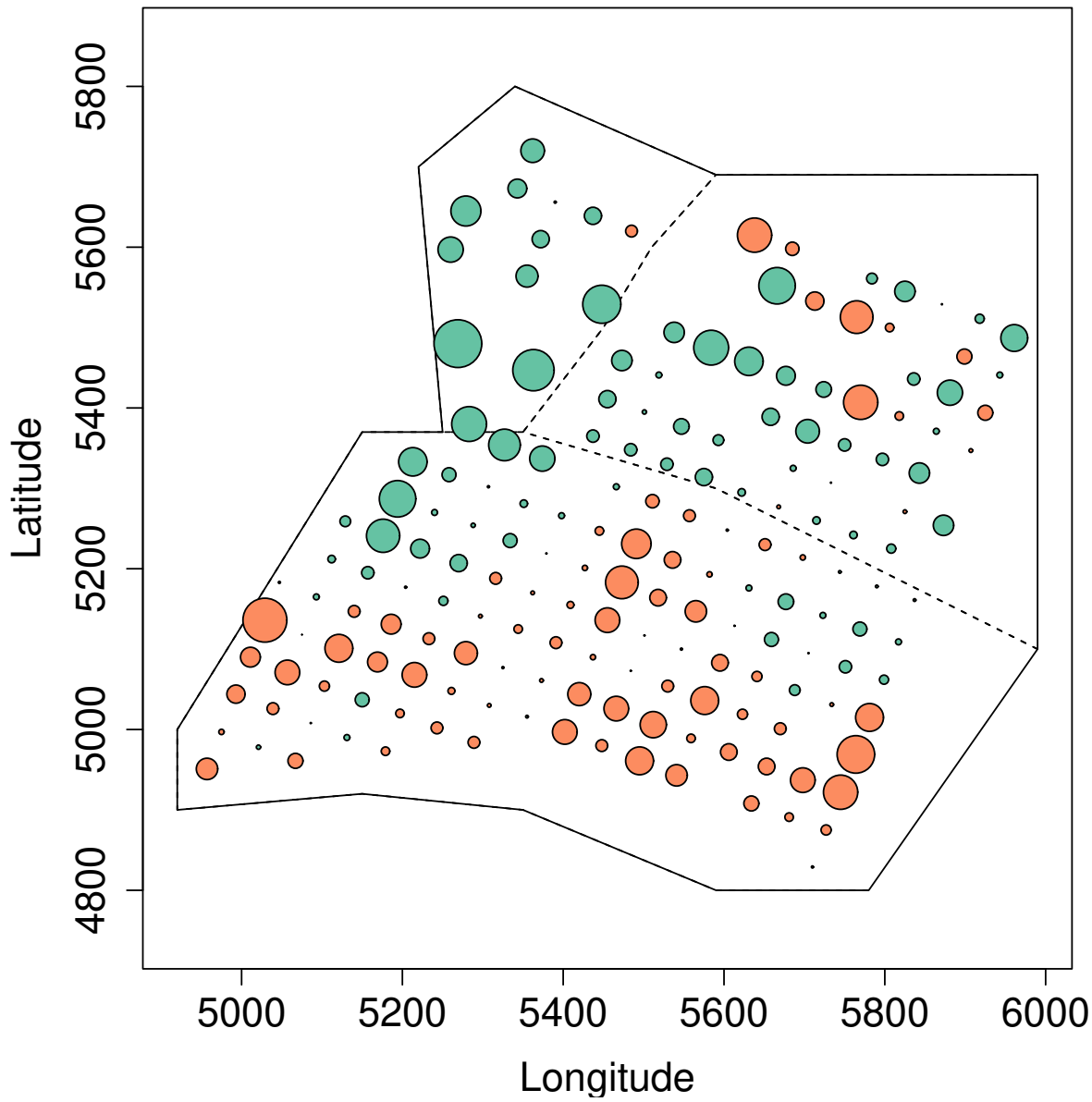
Ca20 data set.



Ca20 data set.



Ca20 data set.



Empirical variograms

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

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

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

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

where $\{B_b \colon b = 1, \ldots, B\}$ 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.

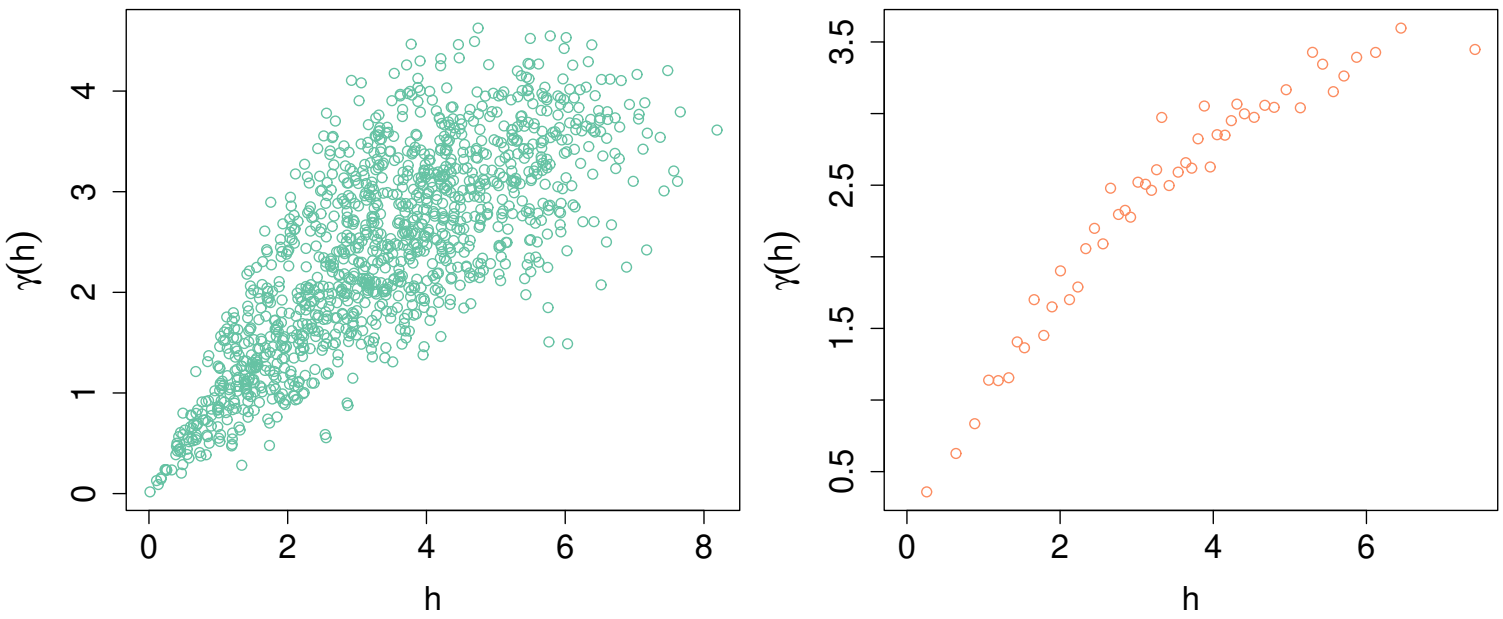


Figure 7: *Empirical variograms. Left: raw. Right: binned.*

The above estimator makes sense only if your data can be considered as stationary or at least with stationary 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.,

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

- 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., $\kappa = 0.25, 0.5, \ldots, 2$.
Always question yourself if a nugget effect makes sense.

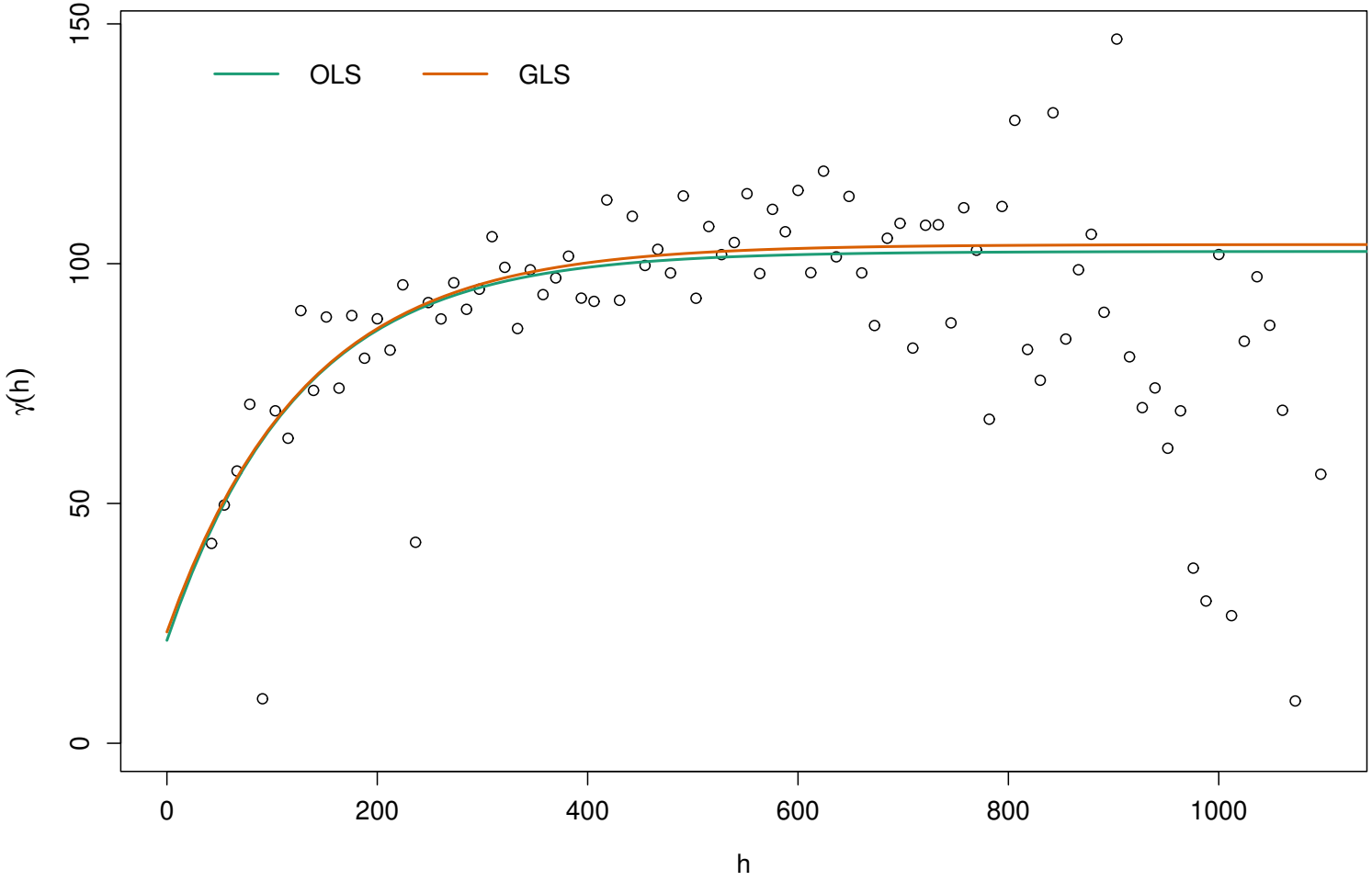
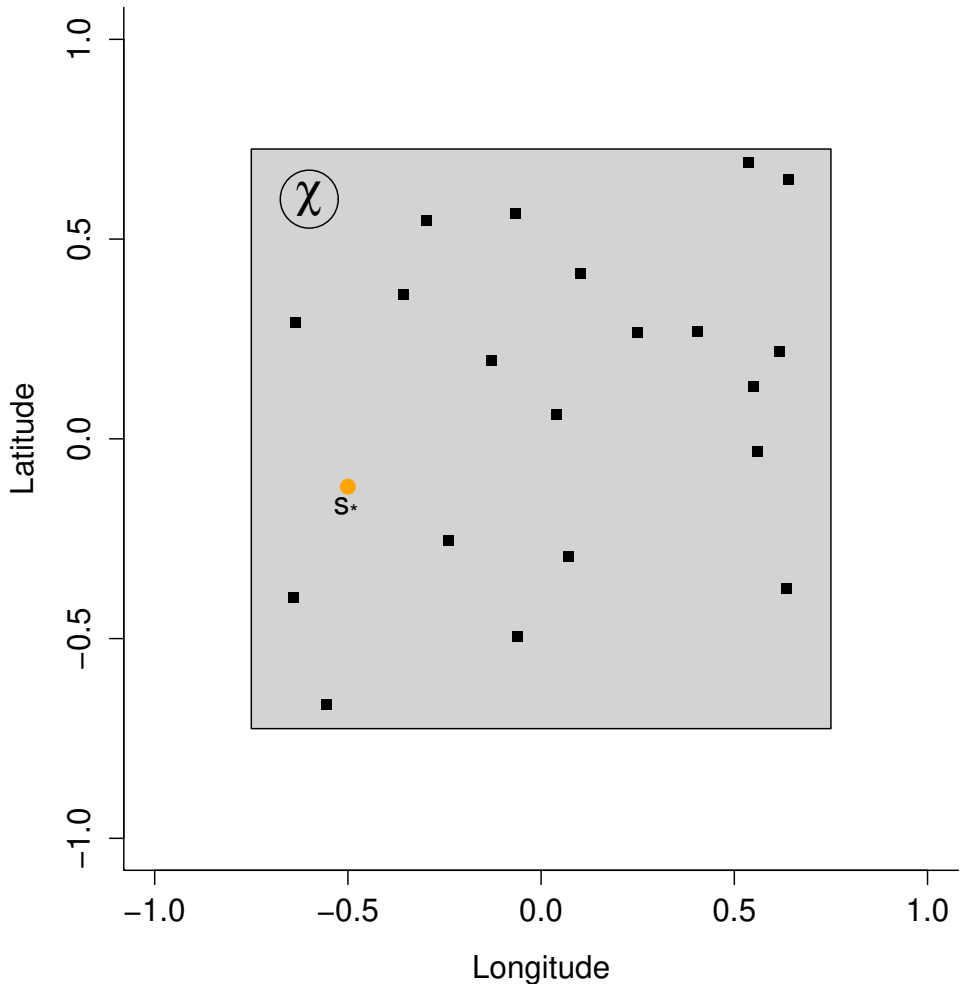


Figure 8: *Least square fitting of a parametric variogram on the Calcium data set.*



- Prediction of $Y(s_*)$ based on observations $Y(s_1), \ldots, Y(s_k)$.
- 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.,

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

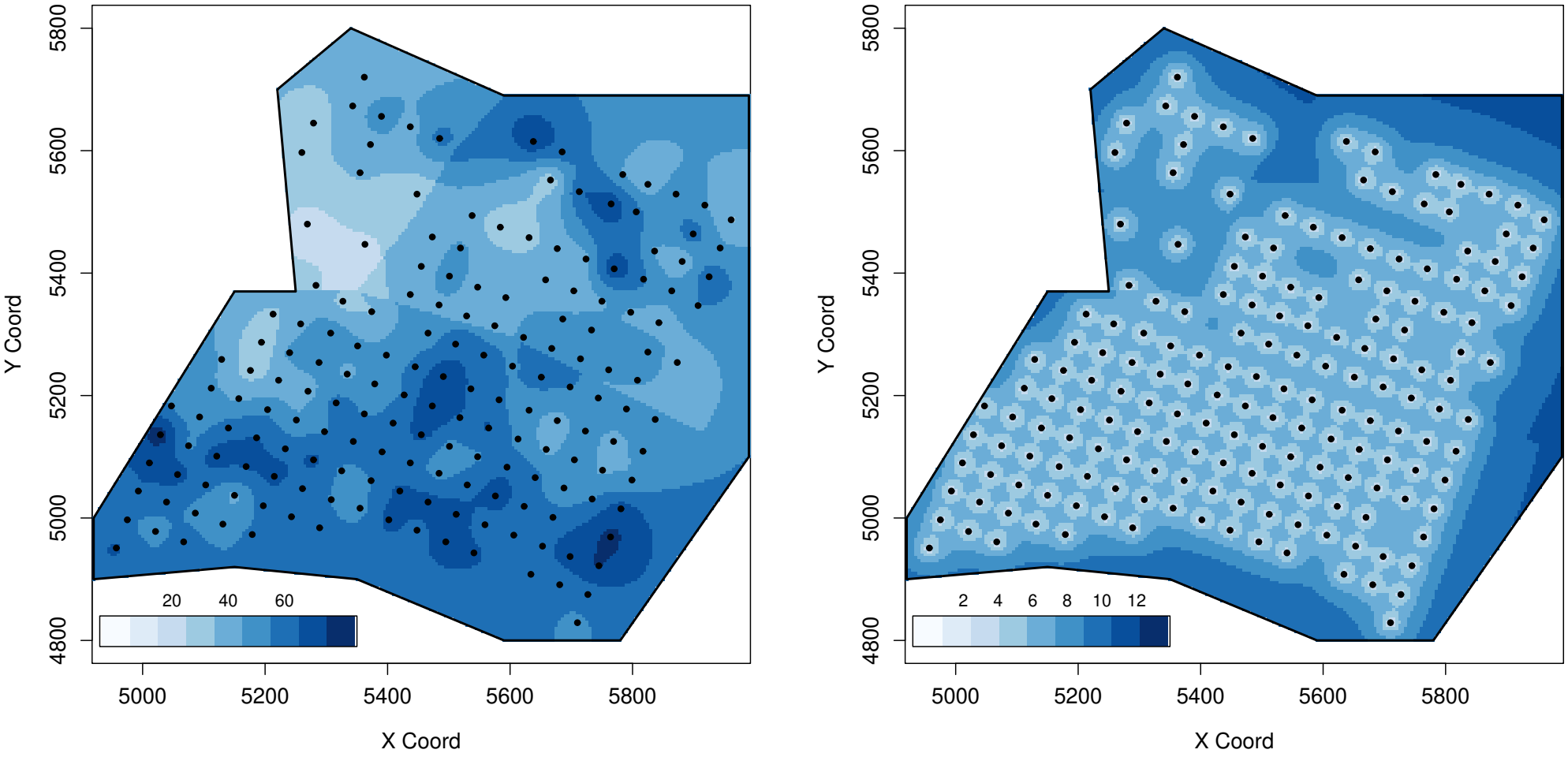


Figure 9: *Kriging estimator (left) and kriging standard error (right) for the Ca20 data set.*

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

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

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

The Mahalanobis distance is given by

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

Gaussian processes

Definition 8. A Gaussian process $\{Y(s) \colon s \in \mathcal{X}\}$ 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.*

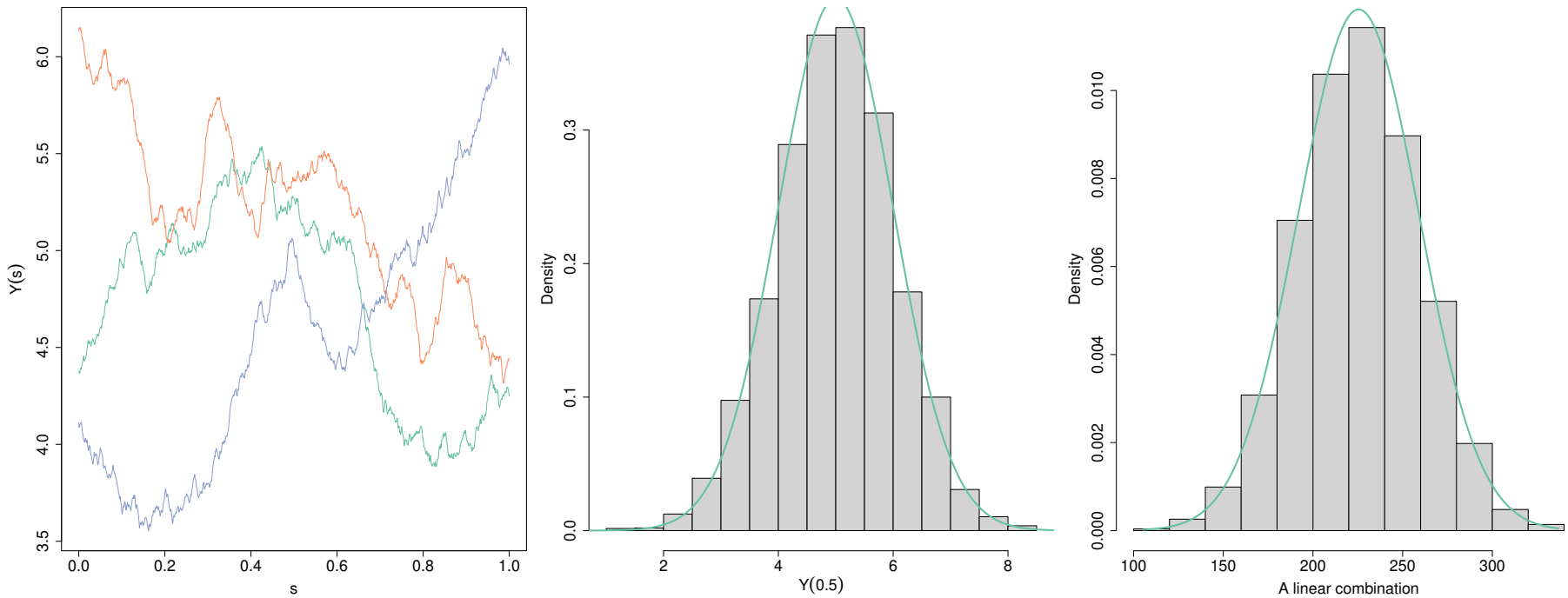


Figure 10: *Numerical illustration of a Gaussian process.*

(semi) Definite positive functions

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

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

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 γ is (semi) definite positive to ensure that the Mahalanobis distance

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

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., $\mathcal{D}_n = \{y_i(s_j) \colon i = 1, \ldots, n, j = 1, \ldots, k\}$ $s_1, \ldots, s_k$, 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 $\gamma(s, s') = \gamma(s, s'; \psi)$ using some prescribed parametric expressions as the ones presented earlier.

Non isotropic/stationary covariance functions

- Defining non isotropic / stationary 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

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

  for some prescribed parametric one–one mapping $\phi(\cdot;\kappa)$.
- A specific case, known as geometric anisotropy, is to set

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

  κ_1, κ_2 are respectively the anisotropy angle and ratio.

Interpolation

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

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

- 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 Gaussian. 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 Gaussian is a sensible choice.
- One widely used choice for positive variable is the Box–Cox transformation

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

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


Tu bluffes Martoni, il bluffe

4. Simulation

(Unconditional) Simulations

- It is rather straightforward to simulated 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 $\mathbf{s} = (s_1, \ldots, s_k) \in \mathcal{X}$, we have

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

  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.

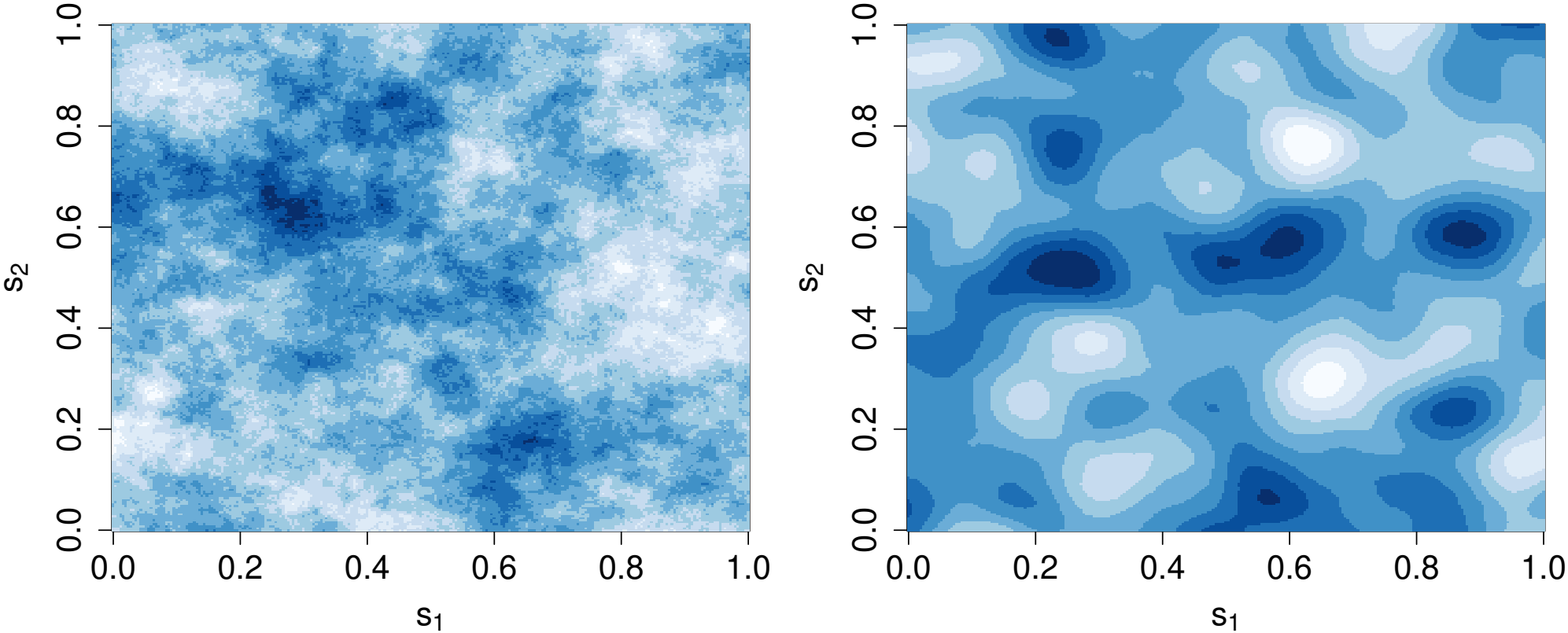


Figure 11: *Two realizations of a random fields with a powered exponential correlation function. Left:* $\kappa = 1$. *Right:* $\kappa = 2$.

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

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

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

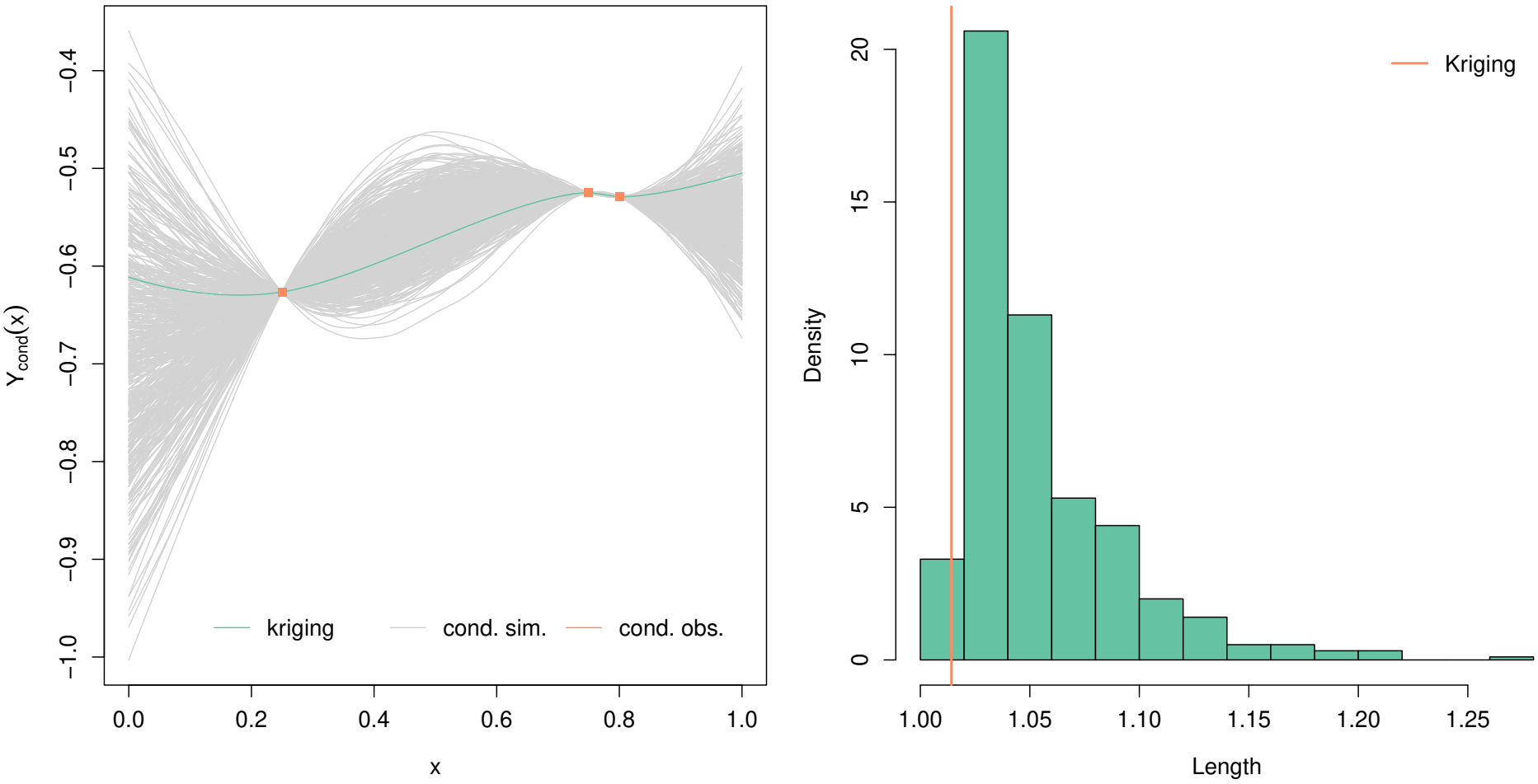


Figure 12: *Comparison between conditional simulations and kriging. Right: length of the curve estimated from kriging and conditional simulations.*

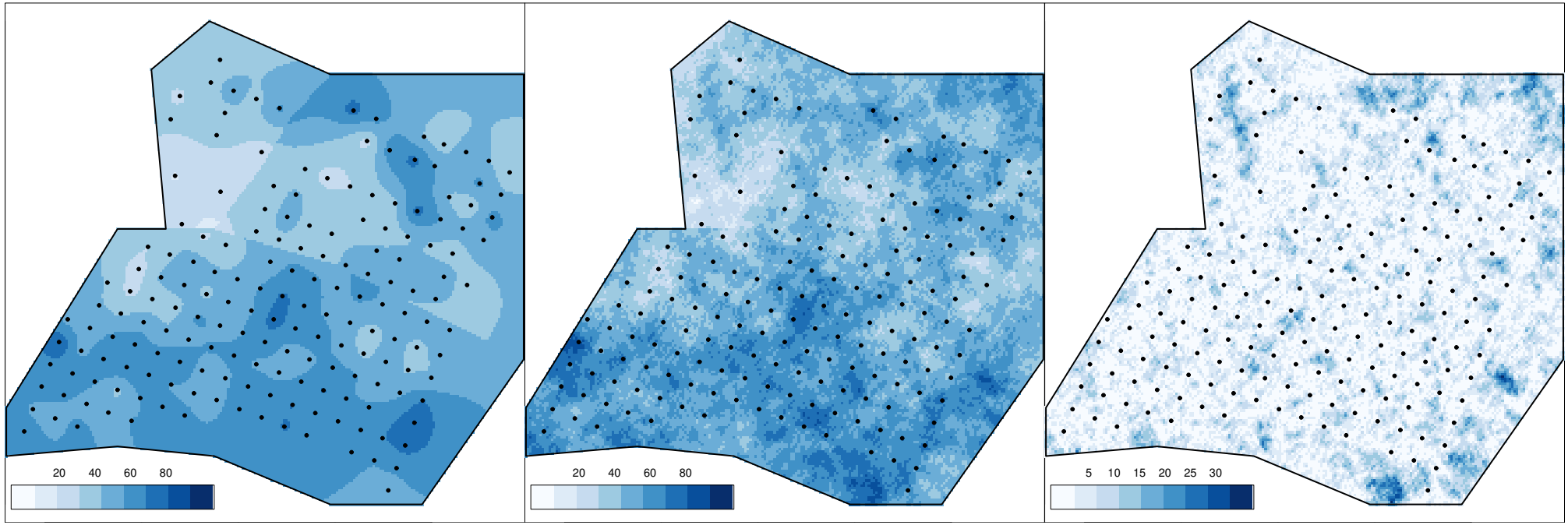


Figure 13: *Comparison between kriging (left) and a conditional simulation (middle). Right: absolute difference of the two.*

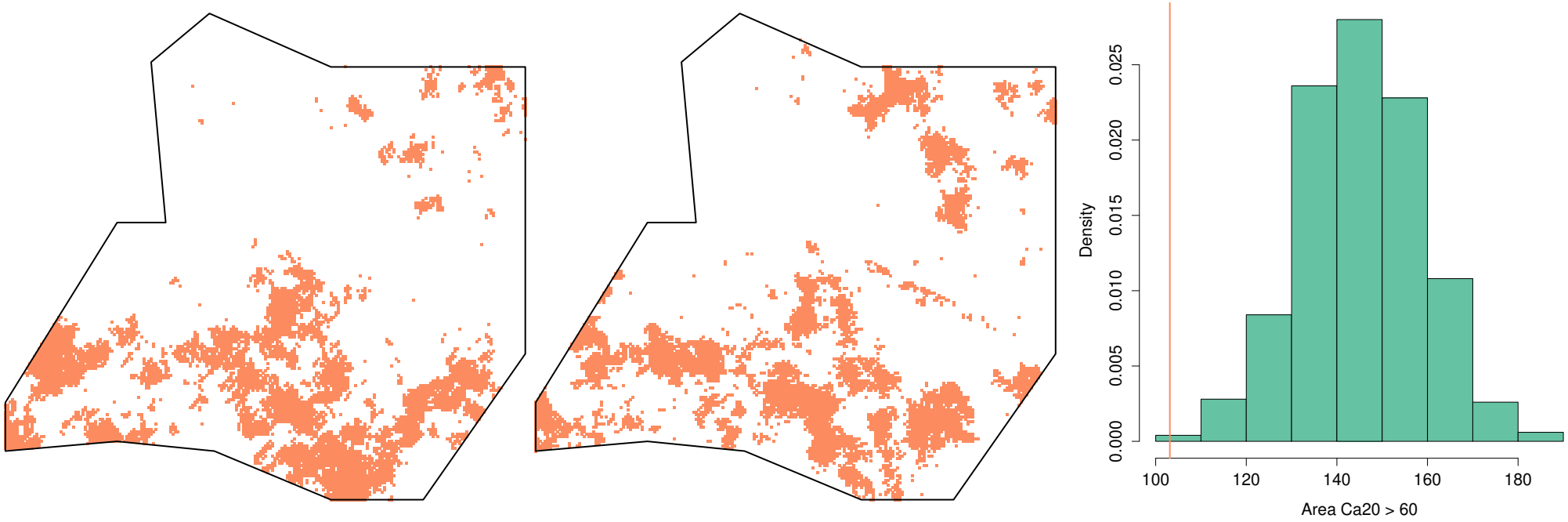


Figure 14: *Left and middle: Two sampled level sets with $u_{crit} = 60$. Right: Distribution of the expected level set area from conditional simulations (histogram) and kriging (vertical line).*

☞ As expected, the kriging–based estimator underestimates.

5. Bayesian hierarchical models

Hierarchical models: Motivations

- Data often depict different layers of variation, that one has to modeled:
  - 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 $\{f(y;\psi)\colon x \in \mathbb{R}^p, \psi \in \psi\}$ is a hierarchical model if we have

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

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.

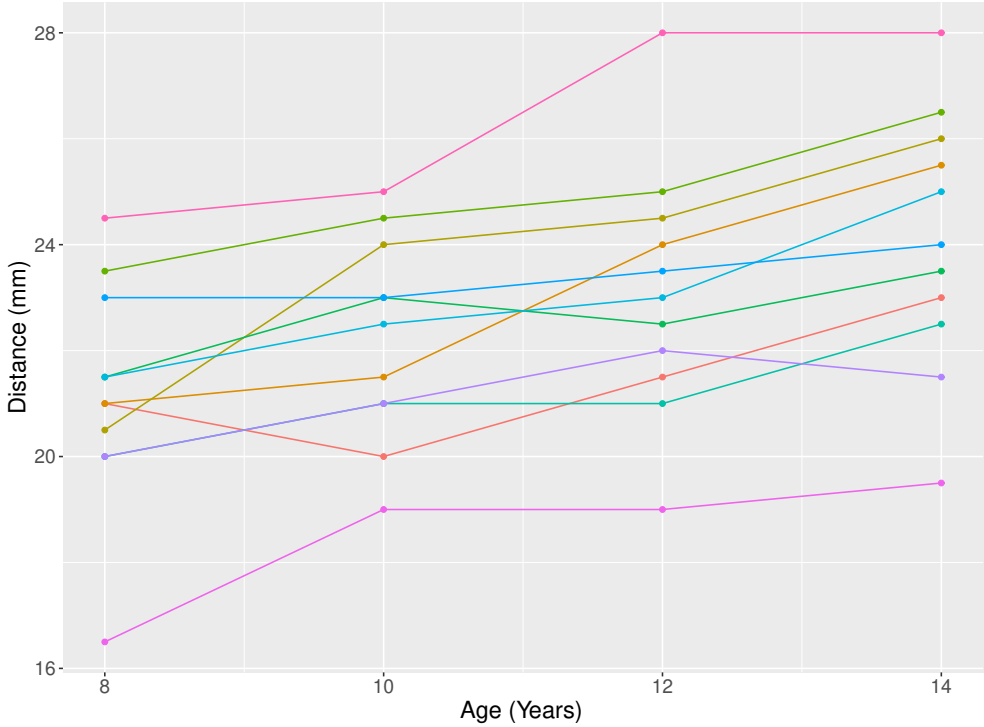


Figure 15: *The data collected by the orthodontists.*

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

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

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

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

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

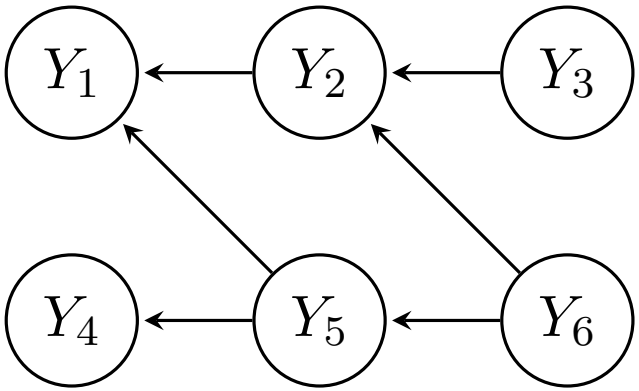
Directed acyclic graph (DAG)

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

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

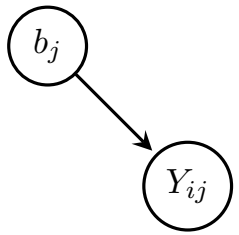
It is directed because it is a directed graph and acyclic because it is impossible to start from a node and get back to it using a path of arrows.

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:

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



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

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

where ... stands for all the other variables.

Bayesian statistical models

Definition 12. A parametric family of functions $\{f(x;\psi)\colon x \in E, \psi \in \Psi\}$ is a statistical model if, for any $\psi \in \Psi$, $x \mapsto f(x;\psi)$ is a probability density function on E. The sets Ψ 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.

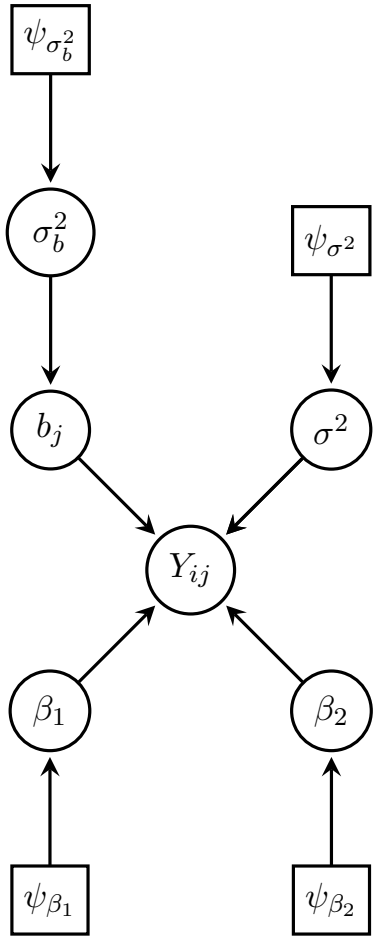
Definition 13. If we further place a prior distribution π on the parameter ψ we are dealing with a Bayesian statistical model (f, π) and the parameters of the prior distribution π are called the hyper–parameters.

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

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

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 $\mathcal{D}_n = (y_1, \ldots, y_n)$ and a Bayesian model (f, π). The main focus in Bayesian inference is on the posterior distribution

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

provided that the marginal distribution (normalizing constant)

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

☞ It is often very convenient to work up to a multiplicative factor independent of ψ 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.,

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

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

$$\pi(y_* \mid \mathcal{D}_n) = \int f(y_* \mid \mathcal{D}_n, \psi)\pi(\psi \mid \mathcal{D}_n)\mathrm{d}\psi, \qquad \mathcal{D}_n \text{ data set.}$$

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

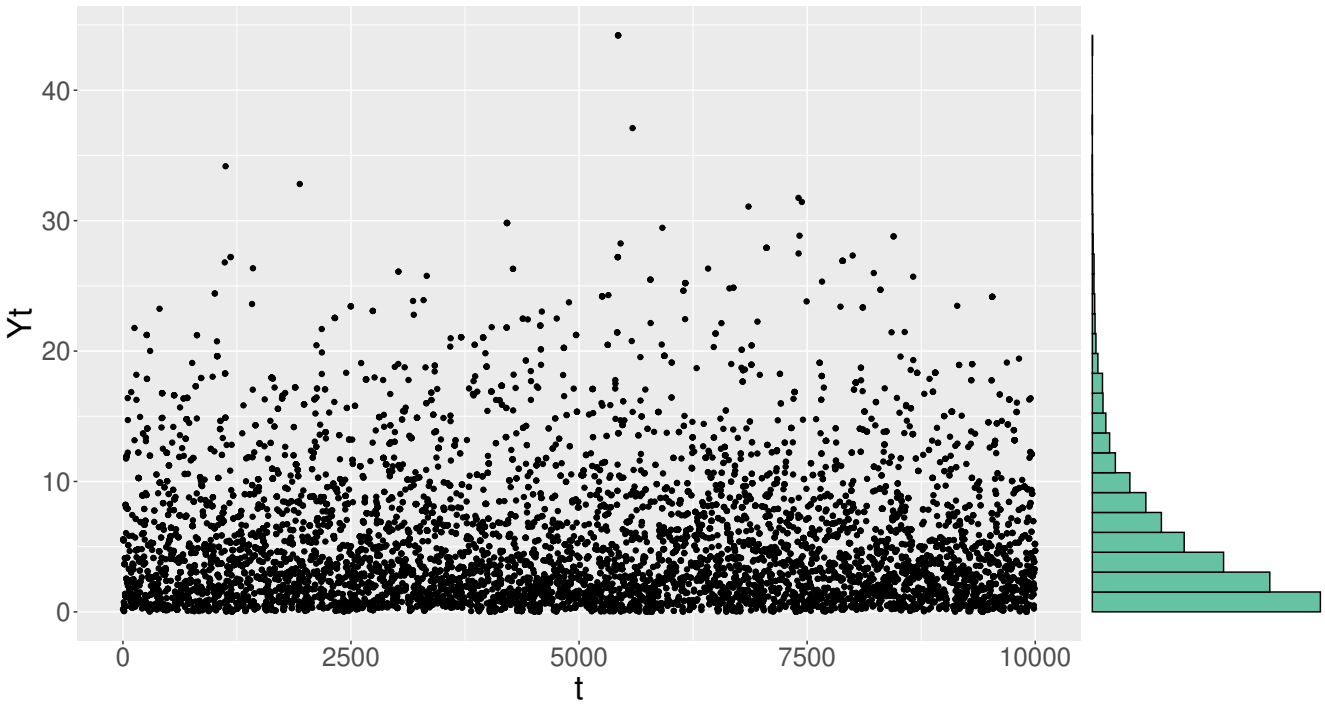


Figure 16: *A Markov chain whose stationary distribution is the Exponential(5).*

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

$$\pi(\psi_j \mid \ldots), \qquad \text{where "}\ldots\text{" means all the rest}$$

☞ In our orthodontist example, we sequentially sample from

$$\pi(\beta_1 \mid \ldots), \quad \pi(\beta_2 \mid \ldots), \quad \pi(\sigma^2 \mid \ldots), \quad \pi(\sigma_b^2 \mid \ldots), \quad \pi(b_j \mid \ldots)$$

Conditional independence model

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

$$
\begin{aligned}
Y(s) \mid \left\{\mu(\cdot), \sigma^2(\cdot)\right\} \overset{\text{ind}}{\sim} N\left\{\mu(s), \sigma^2(s)\right\}, \qquad s \in \mathcal{X} \\
\mu(\cdot) \sim \text{Gaussian Process} \\
\log \sigma^2(\cdot) \sim \text{Gaussian Process}
\end{aligned}
$$

- 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.,

$$Y(s) \mid \{\mu(s), \sigma(s), \xi(s)\} \overset{\text{ind}}{\sim} \text{GEV}\{\mu(s), \sigma(s), \xi(s)\}, \qquad s \in \mathcal{X}$$
$$\mu(\cdot), \log \sigma(\cdot), \xi(\cdot) \sim \text{Gaussian processes}$$

with prior distributions on the Gaussian processes parameters.

The full conditional distributions are

$$\begin{aligned}
&\pi(\mu(s_j) \mid \ldots), \quad \pi(\sigma(s_j) \mid \ldots), \quad \pi(\xi(s_j) \mid \ldots), \quad j = 1, \ldots, k \\
&\pi(\mu_\mu \mid \ldots), \quad \pi(\sigma_\mu \mid \ldots), \quad \pi(\xi_\mu \mid \ldots) \\
&\pi(\gamma_\mu \mid \ldots), \quad \pi(\gamma_\sigma \mid \ldots), \quad \pi(\gamma_\xi \mid \ldots),
\end{aligned}$$

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

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 ω_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

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

  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

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

- 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 ψ_0 be an initial state. The sequence*

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

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

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

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

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

Mini–batch gradient ascent

- Consider the following optimization problem

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

- 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

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

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

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

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

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 data set $\mathcal{D}_n$.
- The main difference is that these subsets are now random.
- The basic stochastic gradient ascent scheme is

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

- 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!