Nonparametric Spatial Statistics

Gérard Biau

Université Paris VI

Summary

1. Introduction
2. Nonparametric spatial density estimation
3. Nonparametric spatial prediction
4. Perspectives

Random fields

Let $Z_{\mathbf{x}}$ be a random field taking values in $\mathbb{R}^d$.

- $\mathbf{x}$ is **discrete**
  - $\rightarrow$ Examples in forestry.
  - $\rightarrow$ Discretization of a continuously indexed field.

- $\mathbf{x}$ is **continuous**
  - $\rightarrow$ Meteorology, soil science, oil engineering...
  - $\rightarrow$ Geostatistics.

Two nonparametric problems

- **Spatial density estimation**

  → For a discrete-space field.

  Tran (1990), Tran and Yakowitz (1993), Carbon, Hallin and Tran (1996), Carbon, Tran and Wu (1997) and Biau (2002).

  → For a continuous-space field.

  Biau (2002).

- **Spatial prediction**

  → For a discrete-space field.

  Hallin, Lu and Tran (2002)

  Biau and Cadre (2002).

  → For a continuous-space field.

  ??

Discrete-space density estimation

- **Basic framework**
  - $\rightarrow$ The field $(Z_{\mathbf{i}})_{\mathbf{i}\in\mathbb{N}^N}$ is observed on the set

$$\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{N}^N \;:\; 1 \le i_k \le n_k,\, k = 1, \ldots, N\}.$$

  - $\rightarrow$ Each $Z_{\mathbf{i}}$ has a density f on $\mathbb{R}^d$.
  - $\rightarrow$ The field is **mixing**.

$$\alpha_k = \sup_{\substack{\mathbf{i},\mathbf{j}\,\in\mathbb{N}^N \\ \|\mathbf{i}-\mathbf{j}\|\ge k}} \alpha\big(\sigma(Z_{\mathbf{i}}), \sigma(Z_{\mathbf{j}})\big), \quad k \in \mathbb{N},$$

and

$$\beta_k = \sup_{\substack{\mathbf{i},\mathbf{j}\,\in\mathbb{N}^N \\ \|\mathbf{i}-\mathbf{j}\|\ge k}} \beta\big(\sigma(Z_{\mathbf{i}}), \sigma(Z_{\mathbf{j}})\big), \quad k \in \mathbb{N},$$

where, for $\mathcal{B}$ and $\mathcal{C}$ two sub σ-fields of $\mathcal{A}$,

$$\alpha(\mathcal{B},\mathcal{C}) = \sup_{\substack{B\in\mathcal{B} \\ C\in\mathcal{C}}} \Big|\mathbf{P}(B\cap C) - \mathbf{P}(B)\mathbf{P}(C)\Big|$$

and

$$\beta(\mathcal{B},\mathcal{C}) = \frac{1}{2}\sup\Big\{\sum_{i\in I}\sum_{j\in J}\Big|P(B_i\cap C_j) - P(B_i)P(C_j)\Big|\Big\}.$$

- **Kernel estimate of f**

$$f_{\mathbf{n}}(z) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} K\left(\frac{z - Z_{\mathbf{i}}}{h_{\mathbf{n}}}\right), \quad z \in \mathbb{R}^d.$$

- **Theorem**. *Assume that*

(i) *For* $\mathbf{i} \neq \mathbf{j}$, $(Z_{\mathbf{i}}, Z_{\mathbf{j}})$ *has a density and, for some* $p \in]2, \infty]$ *and some constant* D,

$$\sup_{\mathbf{i} \neq \mathbf{j}} \|g_{(Z_{\mathbf{i}}, Z_{\mathbf{j}})}\|_p \leq D.$$

(ii) $\alpha_k \leq R k^{-r}$ *for* $k \geq 1$, *where* R *and* r *are positive constants.*

If $r > 2N(p-1)/(p-2)$ *then, for* $z \in \mathbb{R}^d$,

$$\mathbf{E}\big(f_{\mathbf{n}}(z) - f(z)\big)^2 = \frac{h_{\mathbf{n}}^4}{4} \Delta_d^2(z) + \frac{f(z)}{\hat{\mathbf{n}} h_{\mathbf{n}}^d} \int_{\mathbb{R}^d} K^2(u) \mathrm{d}u + \mathrm{o}\Big(h_{\mathbf{n}}^4 + \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n}}^d}\Big),$$

where

$$\Delta_d(z) = \sum_{1 \leq i,j \leq d} \frac{\partial^2 f}{\partial z_i \partial z_j}(z) \int_{\mathbb{R}^d} u_i u_j K(u) \mathrm{d}u\,.$$

Continuous-space density estimation

- **Basic framework**

  $\rightarrow$ The field $(Z_{\mathbf{x}})_{\mathbf{x}\in\mathbb{R}_+^N}$ is observed on the set

$$\mathcal{I}_{\mathbf{T}} = \{\mathbf{x} \in \mathbb{R}_+^N \,:\, 0 \le x_i \le T_i,\, i = 1, \dots, N\}.$$

  $\rightarrow$ Each $Z_{\mathbf{x}}$ has a density f on $\mathbb{R}^d$.

  $\rightarrow$ The field is **mixing**.

$$\alpha(u) = \sup_{\substack{\mathbf{x},\mathbf{x}' \in \mathbb{R}_+^N \\ \|\mathbf{x}-\mathbf{x}'\|=u}} \alpha\big(\sigma(Z_{\mathbf{x}}), \sigma(Z_{\mathbf{x}'})\big), \quad u \ge 0.$$

- **Kernel estimate of f**

$$f_{\mathbf{T}}(z) = \frac{1}{\hat{\mathbf{T}} h_{\mathbf{T}}^d} \int_{\mathcal{I}_{\mathbf{T}}} K\left(\frac{z - Z_{\mathbf{x}}}{h_{\mathbf{T}}}\right) \mathrm{d}\mathbf{x}\,, \quad z \in \mathbb{R}^d.$$

Optimal rate

- $\mathbf{A}_1(\Gamma, p)$ – There exists a Borel set Γ of $\mathbb{R}_+^N \times \mathbb{R}_+^N$ containing

$$\mathcal{D}_N = \left\{(\mathbf{x},\mathbf{y}) \in \mathbb{R}_+^N \times \mathbb{R}_+^N \,:\, \mathbf{x} = \mathbf{y}\right\}$$

  and $p \in]2, \infty]$ such that

  (a) $g_{\mathbf{x},\mathbf{y}} = f_{(Z_\mathbf{x}, Z_\mathbf{y})} - f \otimes f$ exists for $(\mathbf{x},\mathbf{y}) \notin \Gamma$;

  (b) $\delta_p(\Gamma) = \sup_{(\mathbf{x},\mathbf{y})\notin\Gamma} \|g_{\mathbf{x},\mathbf{y}}\|_p < \infty$;

  (c) There exists a positive constant l_Γ such that

$$\limsup_{\hat{\mathbf{T}}\to\infty} \frac{1}{\hat{\mathbf{T}}} \int_{(\mathcal{I}_\mathbf{T}\times\mathcal{I}_\mathbf{T})\cap\Gamma} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} = l_\Gamma < \infty \, .$$

- $\mathbf{A}_2(\gamma, \beta)$ – $\alpha\big(\|\mathbf{x}-\mathbf{y}\|\big) \leq \gamma \, \|\mathbf{x}-\mathbf{y}\|^{-\beta}$ for $(\mathbf{x},\mathbf{y}) \notin \Gamma$, where $\gamma > 0$ and $\beta > 0$.

- **Lemma**. *Let* $z \in \mathbb{R}^d$*. If* $\mathbf{A}_1(\Gamma, p)$ *and* $\mathbf{A}_2(\gamma, \beta)$ *hold for some* Γ, p, γ, β *with* $\beta \geq 2N(p-1)/(p-2)$*, then*

$$\mathrm{Var}\ f_{\mathbf{T}}(z)$$

$$\leq \frac{1}{\hat{\mathbf{T}} h_{\mathbf{T}}^d} \mathbf{E}\left[\frac{1}{h_{\mathbf{T}}^d} K^2\left(\frac{z - Z_0}{h_{\mathbf{T}}}\right)\right] \cdot \frac{1}{\hat{\mathbf{T}}} \int_{(\mathcal{I}_{\mathbf{T}} \times \mathcal{I}_{\mathbf{T}}) \cap \Gamma} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}$$

$$+ \left(V(\mathcal{B}_N)\, \delta_p(\Gamma)\, \|K\|_q^2 + \frac{4V(\mathcal{S}_{N-1})\, \gamma\, \|K\|_\infty^2}{\beta - N}\right) \frac{h_{\mathbf{T}}^\eta}{\hat{\mathbf{T}} h_{\mathbf{T}}^d},$$

where $q = p/(p-1)$ *and* $\eta = 2d(1 - N/\beta)/q - d \geq 0$.

- **Theorem**. *If* f *is bounded and twice continuously differentiable with bounded second derivatives, if* $h_{\mathbf{T}} = c_{\mathbf{T}} \hat{\mathbf{T}}^{-1/(4+d)}$ *where* $c_{\mathbf{T}} \to c' > 0$*, then*

$$\limsup_{\mathbf{T} \to \infty} \sup_{z \in \mathbb{R}^d} \hat{\mathbf{T}}^{4/(4+d)} \mathbf{E}\big(f_{\mathbf{T}}(z) - f(z)\big)^2 \leq C_1\,.$$

Parametric rate

- **Lemma**. *Let $z \in \mathbb{R}^d$. If $g_{\mathbf{x},\mathbf{y}} = g_{\|\mathbf{x}-\mathbf{y}\|}$ exists for $\mathbf{x} \neq \mathbf{y}$, if the mapping*

$$(z_1, z_2) \mapsto \int_0^\infty u^{N-1} \big| g_u(z_1, z_2) \big| \mathrm{d}u$$

is defined, bounded and continuous at (z, z), then

$$\limsup_{\mathbf{T} \to \infty} \hat{\mathbf{T}} \, \mathrm{Var} \; f_{\mathbf{T}}(z) \leq V(\mathcal{S}_{N-1}) \, \|K\|_1^2 \int_0^\infty u^{N-1} \big| g_u(z, z) \big| \mathrm{d}u \, .$$

- **Theorem**. *Assume that $g_{\mathbf{x},\mathbf{y}} = g_{\|\mathbf{x}-\mathbf{y}\|}$ exists for $\mathbf{x} \neq \mathbf{y}$ and that the mapping $u \mapsto u^{N-1} \|g_u\|_\infty$ is integrable over $]0, \infty)$. If f is twice continuously differentiable with bounded second derivatives, then the choice*

$$h_{\mathbf{T}} = \mathrm{o}\big(\hat{\mathbf{T}}^{-1/4}\big)$$

leads to

$$\limsup_{\mathbf{T} \to \infty} \sup_{z \in \mathbb{R}^d} \hat{\mathbf{T}} \, \mathbf{E}\big(f_{\mathbf{T}}(z) - f(z) \big)^2 \leq V(\mathcal{S}_{N-1}) \, \|K\|_1^2 \int_0^\infty u^{N-1} \|g_u\|_\infty \mathrm{d}u \, .$$

Example

Let $(Z_{\mathbf{x}}, \mathbf{x} \in \mathbb{R}_+^N)$ be a **stationary** and **isotropic** spatial Gaussian process with zero mean and autocorellation function

$$\rho(u) = 1 - au^\theta + \mathrm{o}(u^\theta) \quad \text{as } u \to 0(+), \tag{1}$$

where $a, \theta > 0$. Then it is easy to verify that

$$\big|g_u(z_1, z_2)\big| \le a\big|\rho(u)\big|\mathbf{1}_{u>b} + (c + d\,u^{-\theta/2})\mathbf{1}_{u\le b}\,, \quad u > 0.$$

Therefore, the required conditions are satisfied as soon as $u \mapsto u^{N-1}\rho(u)$ is integrable on $[b, \infty)$ and $\theta < 2N$. Gaussian fields with $\theta < 2N$ in (1) have "irregular" sample paths in contrast to the more regular case $\theta \ge 2N$.

Irregular sampling

The field $(Z_{\mathbf{x}}, \mathbf{x} \in \mathbb{R}_+^N)$ is observed on a discrete grid

$$\mathcal{G}_{\mathbf{n}} = \{t_{1,1}, \ldots, t_{1,n_1}\} \times \ldots \times \{t_{N,1}, \ldots, t_{N,n_N}\}.$$

Suppose moreover that there exists a constant $m > 0$ such that, for all $k = 1, \ldots, N$,

$$\min_{1 \le j \le n_k - 1} \big(t_{k,j+1} - t_{k,j}\big) \ge m.$$

The corresponding kernel estimate at the point $z \in \mathbb{R}^d$ reads

$$\overline{f}_{\mathbf{n}}(z) = \frac{1}{\hat{\mathbf{n}} h_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{G}_{\mathbf{n}}} K\left(\frac{z - Z_{\mathbf{i}}}{h_{\mathbf{n}}}\right).$$

Admissible sampling

The field $(Z_{\mathbf{x}}, \mathbf{x} \in \mathbb{R}_+^N)$ is observed on a discrete grid

$$\mathcal{I}_{\underline{\mathbf{n}}} = \{\delta_n, 2\delta_n, \ldots, n\delta_n\}^N,$$

$\delta_n \to 0(+)$ and $T_n = n\delta_n \to \infty$ as $n \to \infty$.

$$f_{\underline{\mathbf{n}}}^*(z) = \frac{1}{n^N h_n^d} \sum_{\mathbf{i} \in \mathcal{I}_{\underline{\mathbf{n}}}} K\left(\frac{z - Z_{\mathbf{i}}}{h_n}\right), \quad z \in \mathbb{R}^d.$$

(1) $g_{\mathbf{x},\mathbf{y}} = g_{\|\mathbf{x}-\mathbf{y}\|}$ exists for $\mathbf{x} \neq \mathbf{y}$ and $\|g_u\|_\infty \leq \pi(u), u > 0$, where

$$(1 + u + \ldots + u^{2N-1})\pi(u)$$

is integrable over $]0, \infty)$, $u^{2N-1}\pi(u)$ is ultimately decreasing and $u^k\pi(u)$ is bounded for

$$k = N, \ldots, 2N - 1.$$

Furthermore, g_u is continuous over the diagonal of $\mathbb{R}_+^{2d}$ and,

(2) for $k = 0, \ldots, N - 1$,

$$\sup_{(z_1,z_2) \in \mathbb{R}_+^{2d}} \left| \int_0^\infty u^k g_u(z_1, z_2) - \sum_{i=1}^\infty \delta_n (i\delta_n)^k g_{i\delta_n}(z_1, z_2) \right|$$

goes to 0 as $\delta_n \to 0$.

- **Lemma**. *Consider the grid* $\{\delta_n, 2\delta_n, \ldots, n\delta_n\}^N$. *Then the total number of pairs of points on this grid whose components are separated by a distance of* $i\delta_n$, *say* $P_{n,i}$, *is a polynomial of the form*

$$P_{n,i} = n^N \Bigg[\sum_{k=0}^{N-2} i^k \mathrm{O}_k(1) + i^{N-1}\Big(\Delta + \mathrm{O}_{N-1}\Big(\frac{1}{n}\Big)\Big) + \sum_{k=N}^{2N-1} i^k \mathrm{O}_k\Big(\frac{1}{n^{k-N+1}}\Big)\Bigg],$$

where $\Delta = N2^N$.

- **Theorem**. *If assumptions (1) and (2) hold, if* f *is twice continuously differentiable with bounded second derivatives, then the sampling* $\delta_n = T_n^{-d/4}$ *is admissible provided* $h_n = n^{-N/(4+d)}$. *More precisely, for* $z \in \mathbb{R}^d$,

$$\mathbf{E}\big(f_n^*(z) - f(z)\big)^2 = \mathrm{O}\Big(\frac{1}{T_n^N}\Big)$$

and the sequence (δ_n) *is optimal.*

Spatial prediction

- **Basic framework**
  - → The **strictly stationary** field $(Z_{\mathbf{i}})_{\mathbf{i}\in\mathbb{N}^N}$ is observed on some set $\mathcal{S}_{\mathbf{n}}$ contained in

$$\mathcal{I}_{\mathbf{n}} = \{\mathbf{i} \in \mathbb{N}^N \,:\, 1 \le i_k \le n_k,\, k = 1, \ldots, N\}.$$

  - → We wish to interpolate the value of the field at the fixed point $\mathbf{j}_0 \in \mathcal{I}_{\mathbf{n}} - \mathcal{S}_{\mathbf{n}}$.
  - → A generic value $Z_{\mathbf{j}}$ only depends of the values in a vicinity of $\mathbf{j}$ not containing $\mathbf{j}$, say $\mathcal{V}_{\mathbf{j}}$.
  - → $(Z_{\mathbf{i}})_{\mathbf{i}\in A}$ admits a density for any bounded A.
  - → $\mathbf{E}\exp\big(a|Z_{\mathbf{i}}|^b\big) < \infty$.
  - → The field is **geometrically mixing**.

$$\alpha\big(\mathcal{B}(E), \mathcal{B}(E')\big) \le \psi(\text{Card } E, \text{Card } E')\,\chi\big(\text{dist}(E, E')\big),$$

with $\chi(t) = \mathrm{O}(e^{-\xi t})$.

- **Best L_2 prediction**

$$\mathbf{r}(\tilde{Z}_{\mathbf{j}_0}) = \mathbf{E}[Z_{\mathbf{j}_0} | Z_{\mathbf{i}}, \mathbf{i} \in \mathcal{V}_{\mathbf{j}_0}] = \mathbf{E}[Z_{\mathbf{j}_0} | \tilde{Z}_{\mathbf{j}_0}].$$

- **Nonparametric predictor**

$$\hat{r}(\tilde{Z}_{\mathbf{j}_0}) = \frac{\displaystyle\sum_{\substack{\mathbf{i} \in \mathcal{S}_{\mathbf{n}} \\ \mathcal{V}_{\mathbf{i}} \subset \mathcal{S}_{\mathbf{n}}}} Z_{\mathbf{i}} \, K\Big(\frac{\tilde{Z}_{\mathbf{j}_0} - \tilde{Z}_{\mathbf{i}}}{h_{\mathbf{n}}}\Big)}{\displaystyle\sum_{\substack{\mathbf{i} \in \mathcal{S}_{\mathbf{n}} \\ \mathcal{V}_{\mathbf{i}} \subset \mathcal{S}_{\mathbf{n}}}} K\Big(\frac{\tilde{Z}_{\mathbf{j}_0} - \tilde{Z}_{\mathbf{i}}}{h_{\mathbf{n}}}\Big)} \, .$$

- **Theorem**. *Assume that f and r are continuous and that $Z_{\mathbf{i}}$ is bounded. Then, under the condition*

$$\frac{\hat{\mathbf{n}}\, h_{\mathbf{n}}^d}{(\log \hat{\mathbf{n}})^{1+N+2/b}} \to \infty\,,$$

one has, with probability 1,

$$\hat{r}\big(\tilde{Z}_{\mathbf{j}_0}\big) \to r\big(\tilde{Z}_{\mathbf{j}_0}\big).$$

Set $v(x) = \mathbf{E}[Z_{\mathbf{i}}^2 | \tilde{Z}_{\mathbf{i}} = x] - \mathbf{E}^2[Z_{\mathbf{i}} | \tilde{Z}_{\mathbf{i}} = x]$.

- **Theorem**. *Assume that f and r are Lipschitzian, v is continuous and $f(x)v(x) > 0$ for all $x \in \mathbb{R}^d$. Suppose that $\hat{\mathbf{n}}\, h_{\mathbf{n}}^{d+2} \to 0$ and that there exists $\delta \in]0, 1/2[$ such that*

$$\hat{\mathbf{n}}\, h_{\mathbf{n}}^{d(1+2\delta N)} (\log \hat{\mathbf{n}})^{-8N/b-2N} \to \infty.$$

Assume moreover that the field $(Z_{\mathbf{i}})_{\mathbf{i}\in\mathbb{N}^N}$ is such that

$$\sup\big\{|\mathbf{P}(A|B) - \mathbf{P}(A)|\big\} \le \varphi\big(\mathrm{dist}(E, E')\big).$$

Then,

$$\frac{\sqrt{\hat{\mathbf{n}}\, h_{\mathbf{n}}^d}}{\sigma(\tilde{Z}_{\mathbf{j}_0})}\big(\hat{r}(\tilde{Z}_{\mathbf{j}_0}) - r(\tilde{Z}_{\mathbf{j}_0})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1),$$

where $\sigma^2(x) = v(x) \int_{\mathbb{R}^d} K^2(u)\mathrm{d}u / f(x)\,.$

Sketch of proof

For $x \in \mathbb{R}^d$, let

$$\zeta_{\mathbf{n}}(x) = \frac{1}{\hat{\mathbf{n}}\, h_{\mathbf{n}}^d} \sum_{\mathbf{i} \in \mathcal{I}_{\mathbf{n}}} Y_{\mathbf{i}}\, K\Big(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{n}}}\Big)$$

and

$$\Delta_{\mathbf{i}}(x) = \frac{1}{\hat{\mathbf{n}}\, h_{\mathbf{n}}^d} \bigg(Y_{\mathbf{i}} K\Big(\frac{x - X_{\mathbf{i}}}{h_{\mathbf{n}}}\Big) - \mathbf{E}\Big[Y K\Big(\frac{x - X}{h_{\mathbf{n}}}\Big)\Big]\bigg).$$

Fix two integers $p, q \geq 1$ and assume that for some integers $t_1, \ldots, t_N$,

$$n_i = t_i(p+q), \quad i = 1, \ldots, N.$$

Let

$$U(1, \mathbf{n}, x, \mathbf{j}) = \sum_{\substack{i_k = j_k(p+q)+1 \\ k=1,\ldots,N}}^{j_k(p+q)+p} \Delta_{\mathbf{i}}(x)\,,$$

$$U(2, \mathbf{n}, x, \mathbf{j}) = \sum_{\substack{i_k = j_k(p+q)+1 \\ k=1,\ldots,N-1}}^{j_k(p+q)+p} \; \sum_{i_N = j_N(p+q)+p+1}^{(j_N+1)(p+q)} \Delta_{\mathbf{i}}(x)\,,$$

and so on. Finally

$$U(2^N, \mathbf{n}, x, \mathbf{j}) = \sum_{\substack{i_k = j_k(p+q)+p+1 \\ k=1,\ldots,N}}^{(j_k+1)(p+q)} \Delta_{\mathbf{i}}(x)\,.$$

Setting $\mathcal{T} = \{0, \ldots, t_1 - 1\} \times \ldots \times \{0, \ldots, t_N - 1\}$ define, for each integer $i = 1, \ldots, 2^N$,

$$T(\mathbf{n}, x, i) = \sum_{\mathbf{j} \in \mathcal{T}} U(i, \mathbf{n}, x, \mathbf{j}) \, .$$

Then, with these notations, we obtain the decomposition

$$\zeta_{\mathbf{n}}(x) - \mathbf{E}\zeta_{\mathbf{n}}(x) = \sum_{i=1}^{2^N} T(\mathbf{n}, x, i) \, .$$

Perspectives

- Prediction in continuous-space.
- Comparison with kriging.
- Practical implementation.
- The problem of the dimension.