Level sets estimation of random compact sets

P. Heinrich, R. S. Stoica and C. V. Tran
Université Lille 1 - Laboratoire Paul Painlevé

Workshop in Spatial Statistics and Image Analysis in Biology
Avignon, May 9-11 2012

Introduction : motivating example

Level sets : a tool for compact random sets averaging

Estimation of level sets

Examples of application

Conclusions and perspectives

A practical application (1)

Pattern detection in spatial data :

- the data $\mathbf{d}$: image analysis, epidemiology, galaxy catalogues
- detect and characterise the pattern "hidden" in the data : objects, cluster pattern or filamentary network
- hypothesis : the pattern is the outcome γ of a stochastic process Γ
- possible solution in this context : probabilistic modelling and maximisation

A practical application (2)

Gibbs modelling framework

- Markov random fields, marked point processes, etc.
- general structure of the probability density :

$$h(\gamma|\theta) = \frac{\exp\left[-U_{\mathbf{d}}(\gamma|\theta) - U_i(\gamma|\theta)\right]}{\alpha(\theta)}$$

and also the necessary mathematical details so that everything is well defined ...

A practical application (3)

Gibbs modelling framework (continued)

- $U_{\mathbf{d}}(\gamma|\theta)$: this term is related to the objects location in the data field (inhomogeneous process)
- $U_i(\gamma|\theta)$: this term is related to the object interaction and to the morphology of the pattern (prior model, regularisation term)
- $\alpha(\theta)$: normalisation constant (not always available analytically)
- pattern estimator :

$$\widehat{\gamma} = \arg\max_{\gamma\in\Omega}\{h(\gamma|\theta)\} = \arg\min_{\gamma\in\Omega}\{U_{\mathbf{d}}(\gamma|\theta) + U_i(\gamma|\theta)\} \qquad (1)$$

A practical application (4)

Some concluding remarks

- simulated annealing algorithm : convergence towards the uniform distribution on the solution sub-space given by (1)
- the model parameters are not always known ...
- the convergence is difficult to be stated
- ... or the solution is not always unique (continuous models and/or priors on the model parameters)
- $\Rightarrow$ a real need to average the obtained solution in order to obtain a much more robust solution

Idea : use level sets as a tool for averaging random patterns

Level sets : basic notions and definitions (1)

Random compact sets and coverage function :

- $(\Omega, \mathcal{A}, \mathbb{P})$: probability space
- $(W = [0,1]^d, \mathcal{B}, \nu)$: measure space (... where the data field leaves) with $\mathcal{B}$ the corresponding Borel $\sigma-$algebra and ν the Lebesgue measure
- $\mathcal{C}$: the class of compact sets in W

A random compact set Γ in W is a random map from Ω to $\mathcal{C}$ that is measurable in the sense

$$\forall C \in \mathcal{C}, \quad \{\omega : \Gamma(\omega) \cap C \neq \emptyset\} \in \mathcal{A}$$

The coverage function is given by :

$$p(w) = \mathbb{P}(w \in \Gamma)$$

Level sets : basic notions and definitions (2)

Level or Quantile sets : for $\alpha \in [0,1]$ the (deterministic) $\alpha-$level set is

$$Q_\alpha = \{w \in W : p(w) > \alpha\}$$

or for simplicity $\{p > \alpha\}$.
Vorob'ev expectation : the Borel set $\mathbb{E}_V \Gamma$ such that

$$\nu\left(\mathbb{E}_V \Gamma\right) = \mathbb{E}\left[\nu\left(\Gamma\right)\right]$$

and

$$\{p > \alpha^*\} \subset \mathbb{E}_V \Gamma \subset \{p \geq \alpha^*\},$$

where

$$\alpha^* = \inf\{\alpha \in [0,1] : \nu\left(Q_\alpha\right) \leq \mathbb{E}\left[\nu\left(\Gamma\right)\right]\}.$$

The Vorob'ev expectation is the α^*-level set that matches the mean volume of Γ.

Some known results and properties (1)

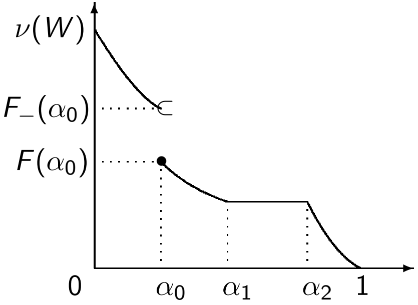


Figure: Behaviour of function $F(\alpha) = \nu(Q_\alpha)$

Remarks :

- F is càdlàg with constant regions (plateaux)
- constant regions of $p(w) \Rightarrow$ discontinuities of $\nu(Q_\alpha)$
- constant regions of $\nu(Q_\alpha) \Rightarrow$ discontinuities of $p(w)$

Some known results and properties (2)

Vorob'ev expectation :

- it is unique provided $F(\alpha) = \nu(Q_\alpha) = \nu(\{p > \alpha\})$ is continuous at α^* ; then we have

$$\mathbb{E}_V \Gamma = \{p \geq \alpha^*\}$$

- it minimises

$$B \to \mathbb{E}\left[\nu(B \triangle \Gamma)\right]$$

under the constraint $\nu(B) = \mathbb{E}\left[\nu(\Gamma)\right]$, where $\triangle$ is the symmetric difference (Molchanov, 05).

More generally, on level sets :

- $p(w)$ not always available in an analytical closed form
- the level sets cannot be computed for all the points $w \in W$ $\Rightarrow$ discretisation should be considered

Plug-in estimation (1)

Definition

- consider n i.i.d. copies $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$ of Γ
- the empirical counterpart of $p(w)$

$$p_n(w) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{w \in \Gamma_i\}}$$

- the plug-in estimator

$$Q_{n,\alpha} = \{p_n > \alpha\}$$

Plug-in estimation (2)

Properties :

the problem was deeply studied in the literature

- some references : (Molchanov, 87, 90, 98), (Cuevas, 97, 06) and many others
- L^1-consistency under weak assumptions $\to$ $p(w)$ does not need to be continuous
- Hausdorff distance : similar consistency results using some extra assumptions
- rates of convergence and asymptotic normality : regularity conditions on $p(w)$

Aim of our work

- plug-in estimator that takes into account the discretisation effects
- estimator for the Vorob'ev expectation $\to$ its definition contains another quantity that need approximation ...

A new level-set estimator (1)

Discretisation : for any Borel set B in W and $r \in 2^{-\mathbb{N}}$, its corresponding grid approximation is

$$B^r = \bigsqcup_{w \in B \cap r\mathbb{Z}^d} [w, w+r)^d.$$

Regularity : the "upper box counting dimension" of ∂B is

$$\overline{\dim_{\mathrm{box}}}(\partial B) = \limsup_{r \to 0} \frac{\log N_r(\partial B)}{-\log r},$$

with

$$N_r(\partial B) = \mathrm{Card}\{w \in r\mathbb{Z}^d : [w, w+r)^d \cap \partial B \neq \emptyset\}.$$

A new level-set estimator (2)

Proposition

Assume that $\overline{\dim}_{\mathrm{box}}(\partial B) < d$*. For all* $\varepsilon > 0$*, there exists* r_ε *such that*

$$0 < r < r_\varepsilon \Rightarrow \nu\left(B^r \triangle B\right) \leq r^{d - \overline{\dim}_{\mathrm{box}}(\partial B) - \varepsilon}.$$

Proposition

Assume that $\overline{\dim}_{\mathrm{box}}(\partial \Gamma) \leq d - \kappa$ *with probability one for some* $\kappa > 0$*. For all* α *such that* $\nu\left(\{p = \alpha\}\right) = 0$*,*
(i) with probability 1*,*

$$\lim_{\substack{r \to 0 \\ n \to \infty}} \nu\left(Q^r_{n,\alpha} \triangle Q_\alpha\right) = 0$$

(ii) for all $\varepsilon > 0$*,*

$$\mathbb{E}\left[\nu\left(Q^r_{n,\alpha} \triangle Q_\alpha\right)\right] \leq r^\kappa + 2\mathrm{e}^{-2n\varepsilon^2} + F(\alpha - \varepsilon) - F(\alpha + \varepsilon).$$

The proof is an extension of the result in (Cuevas, 06).

Vorob'ev expectation estimator (1)

Kovyazin's mean : the empirical counter-part of the Vorob'ev expectation. That is the Borel set K_n such that

$$\nu(K_n) = \frac{1}{n} \sum_{i=1}^{n} \nu(\Gamma_i)$$

and

$$\{p_n > \alpha_n^*\} \subset K_n \subset \{p_n \geq \alpha_n^*\},$$

where

$$\alpha_n^* = \inf\{\alpha \in [0,1] : \nu(\{p_n > \alpha\}) \leq \nu(K_n\}.)$$

Theorem

Assume that $\nu(\{p = \alpha^*\}) = 0$. Then, with probability one,

$$\lim_{n \to \infty} \nu(K_n \triangle \mathbb{E}_V \Gamma) = 0.$$

The proof revisits the result given by (Kovyazin, 86).

Vorob'ev expectation estimator (2)

Grid approximation of K_n : this is the estimator we propose. That is the Borel set $K_{n,r}$ such that

$$\{p_n > \alpha^*_{n,r}\}^r \subset K_{n,r} \subset \{p_n \geq \alpha^*_{n,r}\}^r,$$

where

$$\alpha^*_{n,r} = \inf\{\alpha \in [0,1] : \nu\left(\{p_n > \alpha\}^r\right) \leq \nu\left(K_n\right)\}.$$

Some remarks :

- quite strong assumption : $\nu\left(K_n\right)$ is computed exactly ...
- an alternative idea may consider directly the discretisation of Γ or K_n, but this does not guarantee a mean volume equal to $\nu\left(K_n\right)$ ...
- still, in practice ...

Consistency of the Vorob'ev estimator

Theorem

Assume that $\overline{\dim}_{\mathrm{box}}(\partial\Gamma) \leq d - \kappa$ *with probability one for some* $\kappa > 0$, *and that* $\nu(\{p = \alpha^*\}) = 0$ *and* $\nu(\{p = \beta^*\}) = 0$ *with* $\beta^* = \sup\{\alpha \in [0,1] : \nu(\{p > \alpha\}) \geq \mathbb{E}[\nu(\Gamma)]\}$. *Then, we have almost surely*

$$\lim_{\substack{r \to 0 \\ n \to \infty}} \nu(K_{n,r} \triangle \mathbb{E}_V \Gamma) = 0.$$

Proof.

We write that

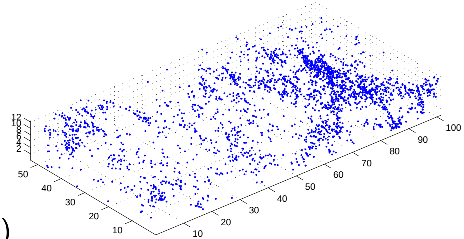
$$\nu(K_{n,r} \triangle \mathbb{E}_V \Gamma) \leq \nu(K_{n,r} \triangle K_n) + \nu(K_n \triangle \mathbb{E}_V \Gamma)$$

and use Theorem 1 and two lemmas to conclude. For technical details, a draft is available on demand ... □

Cosmic filaments : simulated annealing detection

(Stoica, Martinez and Saar, 07,10)

a)



b)

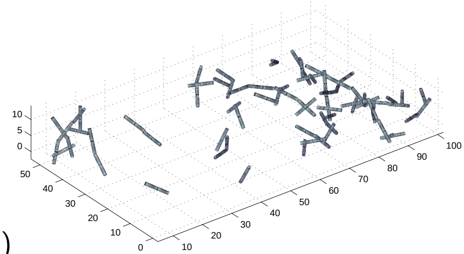


Figure: a) Original data. b) Cylinder configuration detected.

Cosmic filaments : level sets averaging

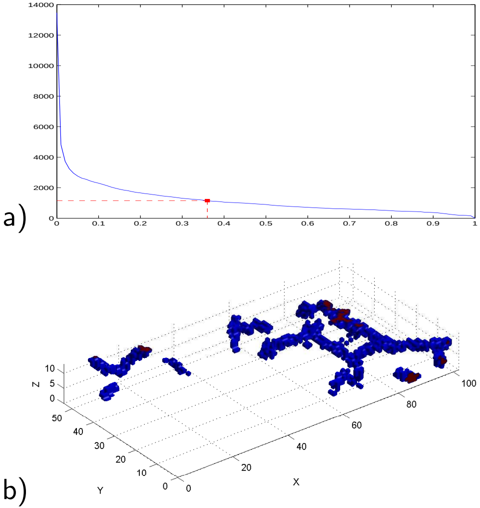


Figure: a) Behaviour of the level set volume. b) Estimated Vorob'ev expectation.

Epidemiology (veterinary context)

Disease : sub-clinical mastitis for diary herds

- points → farms location
- to each farm → disease score (continuous variable)
- clusters pattern detection : regions where there is a lack of hygiene or rigour in farm management

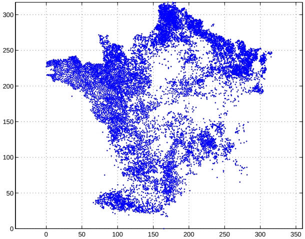


Figure: The spatial distribution of the farms outlines almost the entire French territory (INRA Avignon).

Epidemiology : sub-clinical mastitis data

(Stoica, Gay and Kretzschmar, 07)

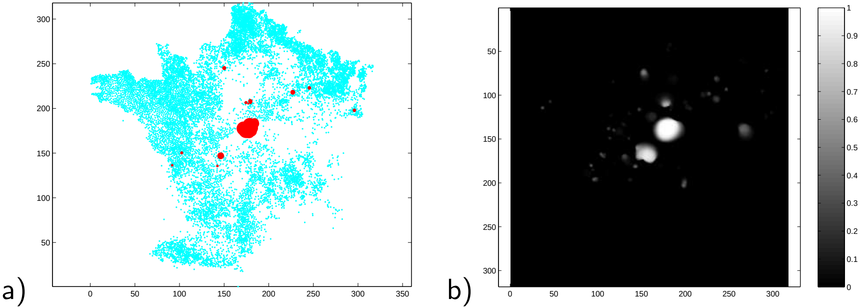


Figure: Disease data scores and coordinates for the year 1996 : a) cluster pattern (disk configuration) detected ; b) Level sets.

Conclusion :

- estimator including the discretisation effects
- averaging the shape of the pattern ...

Perspectives :

- ... provided the model is correct ...
- relax hypotheses
- what is the variance of the pattern ?

Acknowledgements :

this work was done together with wonderful co-authors and also with help of some very generous people ... Some of them are today with us :) ...

GDR Géométrie Stochastique

Aim :

- network of scientists
- no obligations at all ...
- joining mathematicians sharing common research interests but not only ... also the scientists from the corresponding application domains ...

Contact :

- Pierre.Calka@univ-rouen.fr
- web page : http://gdr-geostoch.math.cnrs.fr/