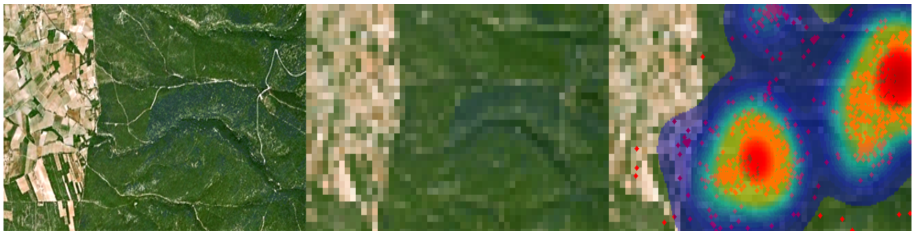
Ninth French-Danish Workshop on Spatial Statistics and Image Analysis in Biology

Avignon, May 9-11, 2012



Scientific Coommittee Kasper Berthelsen, Bjarne Ersbøll, Kiên Kiêu, Antti Penttinen

Organization Coommittee Denis Allard, Delphine Blanke, Florent Bonneu, Edith Gabriel, Rachid Senoussi, Samuel Soubeyrand

Unité Biostatistique et Processus Spatiaux
Institut National de la Recherche Agronomique
Avignon, France
www.biosp.org

Equipe de Statistique
Laboratoire d'Analyse non Linéaire et Géométrie
Université d'Avignon et des Pays de Vaucluse
math.univ-avignon.fr

Foreword

This 9th edition of the French-Danish Workshop on Spatial Statistics and Image Analysis in Biology (SSIAB) is jointly organized by the Biostatistics and Spatial Processes (BioSP) research unit at INRA and the statistics group from the Department of Mathematics, LANLG at the University of Avignon.

Previous issues of this workshop were alternatively organized in France and Denmark. Attendees are from Denmark and France, of course, but also from other Scandinavian countries, Holland, Czesh Republic, Italy, Spain and the United States. Applications are numerous, but oriented towards biology in very broad sense.

We wish to thank Sylvie Jouslin for her contribution to the preparation of this workshop.

The organization committee

Denis Allard
INRA, Biostatistics and Spatial Processes (BioSP)
allard@avignon.inra.fr

Viktor Benes
Charles University Prague, Faculty of Mathematics and Physics
benesv@karlin.mff.cuni.cz

Kasper Berthelsen
Aalborg University, Department of Matematical Sciences
kkb@math.aau.dk

Jean-Michel Billiot
Université de Grenoble, Laboratoire Jean Kuntzmann
jean-michel.billiot@upmf-grenoble.fr

Delphine Blanke
Univeristé d'Avignon et des Pays de Vaucluse, Laboratoire de Mathématiques d'Avignon
delphine.blanke@univ-avignon.fr

Mathieu Bonneau
INRA, Mathématique et Informatique Appliquée Toulouse
mbonneau@toulouse.inra.fr

Florent Bonneu
Université d'Avignon et des Pays de Vaucluse Laboratoire de Mathématiques d'Avignon
florent.bonneu@univ-avignon.fr

Line Clemmensen
Technical University of Denmark, Informatics and Mathematical Modelling
lhc@imm.dtu.dk

Rémy Drouilhet
Université de Grenoble, Laboratoire Jean Kuntzmann
remy.drouilhet@upmf-grenoble.fr

Edith Gabriel
Université d'Avignon et des Pays de Vaucluse,Laboratoire de Mathématiques d'Avignon
edith.gabriel@univ-avignon.fr

Marc Genton
Texas A& M University, Department of Statistics
genton@stat.tamu.edu

Mohammad Ghorbani
Aalborg University, Department of Mathematical Sciences
ghorbani@math.aau.dk

Yongtao Guan
University of Miami, Management Science
yguan@miami.edu

Gilles Guillot
Technical University of Denmark, Informatics
gilles.b.guillot@gmail.com

Jan-Otto Hooghoudt
Aalborg University, Mathematics
janotto@math.aau.dk

Robert Jacobsen
Aalborg University, Department of Mathematical Sciences
robert@math.aau.dk

Kiên Kiêu,
INRA, UR 341 Mathématiques et Informatique Appliquées
kien.kieu@jouy.inra.fr

Etienne Klein
INRA, Biostatistics and Spatial Processes (BioSP)
etienne.klein@avignon.inra.fr

Frédéric Lavancier
Université de Nantes, Laboratoire de Mathématiques Jean Leray
frederic.lavancier@univ-nantes.fr

David Legland
INRA, UMR GMPA
david.legland@grignon.inra.fr

Jesper Møller
Aalborg University, Department of mathematical Statistics
jm@math.aau.dk

Nadia Morsli
Université Joseph Fourier
m_nadia_99@yahoo.fr

Tomas Mrkvicka
University of South Bohemia, Department of Applied Mathematics and Informatics
mrkvicka@prf.jcu.cz

Mari Myllymäki
Aalto University, Dep. of Biomedical Eng. and Computational Science
mari.myllymaki@aalto.fi

Nathalie Peyrard
INRA, Mathématique et Informatique Appliquées Toulouse
nathalie.peyrard@toulouse.inra.fr

Emilio Porcu
Universidad Castilla la Mancha, Statistics
emilio.porcu@uclm.es

Jakob Rasmussen
Aalborg University, Department of Mathematical Sciences
jgr@math.aau.dk

Anne Ruiz-Gazen
Toulouse School of Economics, Department of Mathematics
anne.ruiz-gazen@tse-fr.eu

Aila Särkkä
Chalmers University, Mathematical Sciences
aila@chalmers.se

Farzaneh Safavimanesh
Aalborg University, Department of Mathematical Sciences
farzaneh@math.aau.dk

Rachid Senoussi
INRA, Biostatistics and Spatial Processes (BioSP)
senoussi@avignon.inra.fr

Samuel Soubeyrand
INRA, Biostatistics and Spatial Processes (BioSP)
samuel.soubeyrand@avignon.inra.fr

Katerina Stankova Helisova
Czech Technical University in Prague, Department of Mathematics (Fac. of El. Engineering)
helisova@math.feld.cvut.cz

Radu Stoica
Universite Lille 1, Laboratoire Paul Painleve
stoica@math.univ-lille1.fr

Christine Thomas-Agnan
Université Toulouse 1 Capitole, Toulouse School of Economics
Christine.Thomas@tse-fr.eu

Giovanni Luca Torrisi
CNR, Istituto per le Applicazioni del Calcolo
torrisi@iac.rm.cnr.it

Marie-Colette van Lieshout CWI, PNA2
Marie-Colette.van.Lieshout@cwi.nl

Rasmus Waagepetersen
Aalborg University, Department of Mathematical Sciences
rw@math.aau.dk

Dimension reduction in random marked sets

V. Beneš, O. Šedivý, J. Staněk

Charles University in Prague, Faculty of Mathematics and Physics, Sokolovskaá 83, 18675 Praha 8, Czech Republic, benesv@karlin.mff.cuni.cz

The talk deals with spatial point, fibre and surface processes and multivariate Gaussian random fields (GRF) as covariates. Alternatively the concept of a random marked set from Ballani et al. (2009) will be used. The aim is to study the dependencies between these objects and the dimension reduction problem of covariates. We consider a generalization of the sufficient dimension reduction paradigm for inhomogeneous spatial point processes developed in Guan and Wang (2010). Among inverse regression techniques we concentrate on the basic method called the sliced inverse regression. Guan (2008) claimed SIR to be hardly applicable in point processes because of non-existence of natural slicing. We show that slicing can be realized in spatial processes, based on suitable geometrical marks. Basic theorems on the structure of dimension reduction subspaces are derived for SIR. Statistical tests of independence and estimation of the sufficient dimension are investigated. Guan and Wang (2010) refined the analysis defining the k-th order central intensity subspace and they studied the case $k = 1$ in detail. In the present paper moreover the case $k = 2$ is investigated in detail and an approach to slicing is suggested.

The methods developed are demonstrated in simulations of different models from stochastic geometry. We estimate directions in the central subspace, hypotheses about its dimension are tested. The quality of estimators is quantified, the power of the tests can be evaluated in repeated simulations. First the intensity of an inhomogeneous Poisson point process is proportional to a function of a component of a multivariate GRF in $\mathbb{R}^2$ or $\mathbb{R}^3$. Slicing is based on the nearest neighbour distance. Further a Poisson Voronoi tesselation is generated by the point process and the fibre, surface process of its edges, faces is considered, respectively. Slicing is based on the length of edges or the surface area of faces.

Secondly the second order central intensity subspace is estimated in simulations of a hard-core point process of Matern type in $\mathbb{R}^2$. The thinning rule is determined by the attached GRF. Slicing of the set of pairs of events is based on the second-order intensity.

References:

F. Ballani, Z. Kabluchko, M. Schlather (2009) Random marked sets. arXiv: 0903.2388v1 [math.PR]

Y. Guan (2008) On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. J. Amer. Stat. Asoc. 103, 483, 1238–1247.

Y. Guan, H. Wang (2010) Sufficient dimension reduction for spatial point processes directed by Gaussian random fields. J. R. Statist. Soc. B, 72, 3, 367–87.

O. Šedivý, J. Staněk, B. Kratochvílová, V. Beneš (2011) Sliced inverse regression and independence in random marked sets with covariates, submitted.

EBSpat a R package dedicated to simulation and estimation in the framework of Gibbs nearest-neighbour point processes

Rémy Drouilhet

LJK, 1251 avenue centrale, BP 47, 38040 Grenoble Cedex 9.

The class of nearest-neighbour Gibbs point processes is interesting because it can allow us to expand from the superstable class of Gibbs point processes introduced by Ruelle whose point interactions are naturally based on the complete graph. It is now possible to consider point interactions based on nearest-neighbour graphs (Delaunay graph, for example). Many theoretical results obtained for this kind of processes have been proposed in the recent years. The first step was to establish the existence (see [1,4] for the main contributions) of these stationary Gibbs point processes when defined in $\mathbb{R}^d$ (d being the dimension). The series of papers (among which [2,3]) dealing with the estimation of stationary point processes (applicables in the framework of nearest-neighbour Gibbs point processes) have then been submitted.

Together with the first theoretical advancement on the existence of this type of Gibbs point processes, E. Bertin and R. Drouilhet had developed a `C` software program to simulate these point processes, especially for the Delaunay, the k-nearest neighbours and Gabriel graphs. In memory of E. Bertin, a `R` package, named `EBSpat`, is being developed in `R` to offer on the one hand simulation tools for the nearest-neighbour point processes and, on the other hand, estimation tools based on the pseudo-likelihood and Takacs-Fiksel methods.

This package now completes the comprehensive `R` package `spatstat` whose main contributor is Adrian Baddeley. In a near future, the development of `EBSpat` will try to embrace the spirit of the second version of the `spatstat` package (which presumably be called `spatstat2`).

To have a better idea of the main functions of `EBSpat`, let us now discuss an example of simulation and estimation for a Delaunay pair interaction point process. The next few lines of `R` code allow the simulation of a realization of a Strauss model based on the Delaunay graph.

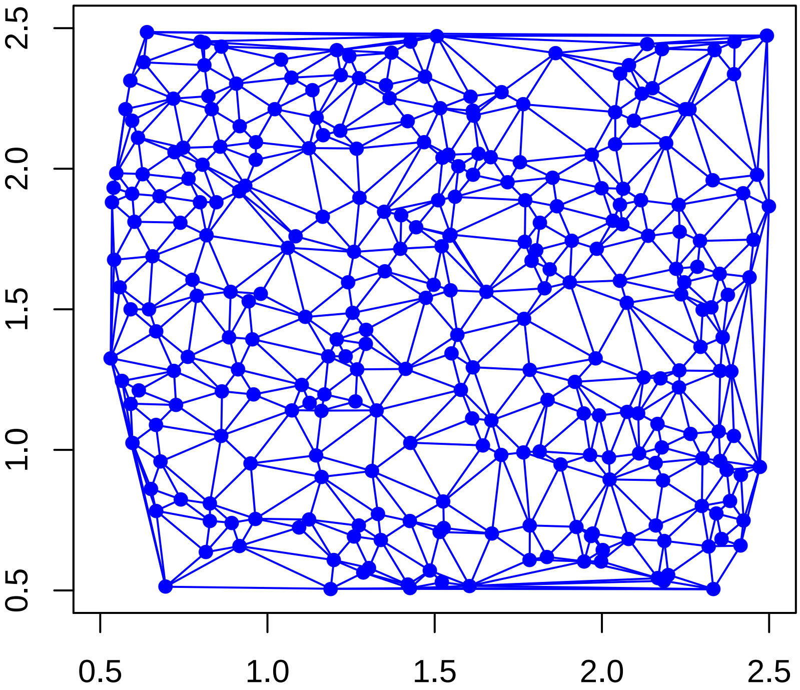
```
> gD <- EBGibbs(~ (-4.61) + Del2(th*(l2<=0.0025),th=0.69),
+               center=c(1.5,1.5),size=2,sizeIn=1.5)
> run(gD) # equivalent here to "simulate"
```

The object `gd` has the `EBGibbs` class. The quantity `l2` is predefined and corresponds to the squared length of the Delaunay edge. The sizes of the interior and exterior domains are respectively fixed at 1.5 and 2. The singleton potential is fixed at -4.61 and the interaction function is that of the Strauss model based on the Delaunay graph, with the range fixed at 0.05 and the step value at 0.69. Even with this simple example, it is noticeable that the user can easily define the shape of the interaction function that will be applied to each Delaunay edge. A general Delaunay pair interaction of the form:

$$\sum_{\xi \in Del_2(\varphi)} f(\xi, \theta_1, \cdots, \theta_p) \tag{1}$$

is then declared in the `R` system by inserting an additive term `Del2(f(...),theta1=,...,thetaP=)` in the `R` formula provided as the first and main argument of the `EBGibbs` function. `f(...)` is an `R` expression entered by the user (corresponding to the function f in the Equation (1)) which can possibly depend on several parameters `theta1,...,thetaP` (related in Equation (1) to $\theta_1, \cdots, \theta_p$) and the following predefined characteristics :

- `x`: points coordinates (ex: `x[[1]]` and `x[[2]][2]`)
- `v`: marks list (ex: `v[[1]]$m` and `v[[2]]$m2`)
- `a`, `da`: neigbouring Voronoi cells areas and the absolute value of their difference (ex: `a[1]`)
- `l`, `l2`: Delaunay edge length and squared length.
- `ol`, `ol2`: dual Delaunay edge length and squared length.



Now, it is now possible to offer the estimations of the two parameters of the model. Below are the instructions for the estimations using the pseudo-likelihood method.

```
> peD <- EBPseudoExpo(gD~Del2(l2<=0.0025),domainSize=1.5)
> run(peD,c(0,0),update=TRUE)
[1] -4.769941 0.8691488
```

The `R` function `EBPseudoExpo` generates a `peD` object. The `run` method provides the resulting estimations. The syntax to declare the interaction differs from the one used for the simulation since the model is here considered as an exponential family model. In this framework, the syntax used to provide the exhaustive statistics is: `Del2(f1(...),...,fP(...))` where `f1(...)`,···,`fP(...)` are the p `R` expressions of these exhaustive statistics which depend on the same characteristics (i.e. `x`, `v`, `a`, `da`, `l`, `l2`, `ol` et `ol2`) introduced in the `EBGibbs` function. Let us notice that the singleton parameter is always in the first position.

[1] Bertin, E., Billiot, J.M. et Drouilhet, R. (1999) *Existence of "Nearest-Neighbour" Gibbs Point Models*, Adv. Appl. Prob., 31, 895–909.
[2] Billiot, J.-M., Coeurjolly, J.-F and Drouilhet, R. (2008) *Maximum pseudolikelihood estimator for exponential family models of marked Gibbs point processes*, Electronic Journal of Statistics, 2 234-264.
[3] Coeurjolly J.-F., Dereudre, D., Drouilhet, R. et Lavancier, F. (2010) *Takacs Fiksel method for stationary marked Gibbs point processes*, HAL, numéro hal-00502004 (To appear in Scandinavian Journal of Statistics).
[4] Dereudre, D., Drouilhet, R. et Georgii, H.-0. (2010) *Existence of Gibbsian point processes with geometry-dependent interactions*, Probab. Theory Related Fields 150 (2011).

Analysis of the spatial organisation of maize vascular bundles

David Legland[1,2], Marie-Françoise Devaux[3], Fabienne Guillon[3]

[1] INRA, UMR 0782 Génie et Microbiologie des Procédés Alimentaire, Thiverval-Grignon, F-78850, France, [2] Agroparistech, UMR 0782 Génie et Microbiologie des Procédés Alimentaire, Thiverval-Grignon, F-78850, France, [3] INRA UR1268 Biopolymères, Interaction et Assemblages, F-44300 Nantes, France

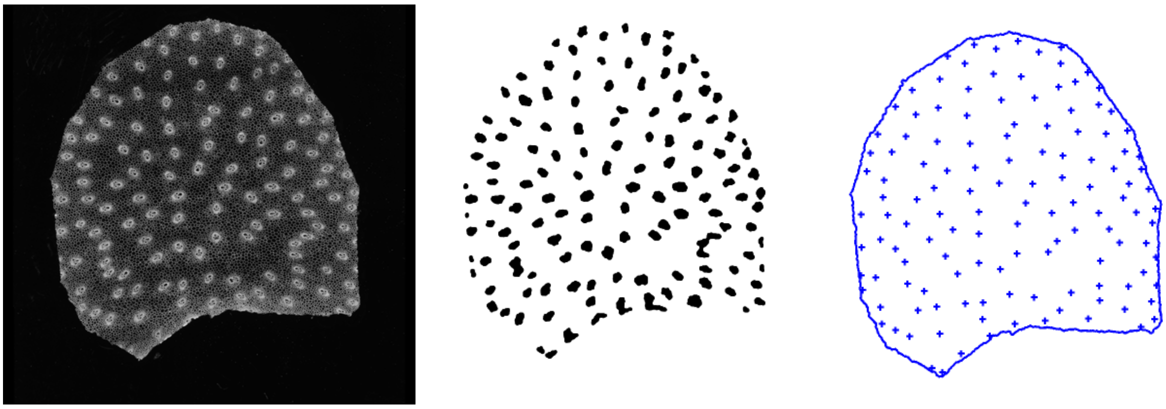
david.legland@grignon.inra.fr, [devaux, guillon]@nantes.inra.fr

Introduction

Crop species like maize are of interest for cattle feeding or for bio ethanol production. They are transformed to energy or fuel after several mechanical, biochemical and/or enzymatic processes. The degradability of plant material depends on its biochemical composition, and on the structure and organisation of plant tissues. This cellular structure is usually investigated through morphology (size, shape, orientation) of cells. The cellular organisation within the stem of biological structures such as the vascular bundles is however rarely investigated. The aim of this study was to quantify the spatial organisation of vascular bundles within maize stems using methods issued from spatial statistics.

Image Processing

Images of maize slices were obtained using a macroscopy imaging system [2]. Images had a size of approximately 4500*4500 pixels and a pixel resolution equal to 3.62 µm (Fig. 1-a). Vascular bundles were enhanced using alternate sequential filters [3], and segmented by detecting extended maxima (Fig. 1-b). The position of each bundle was defined as the centroid of the corresponding maxima. The contour of each slice was obtained by filtering and applying a threshold to the original image. The resulting polygon was simplified to reduce further computation time. The position of vascular bundles as well as the contour of the stem yielded a bounded point pattern observation for each slice (Fig. 1-c).



(a) (b) (c)

Figure 1: Identification of vascular bundle positions. (a) Example of input image. (b) Segmentation of vascular bundles using alternate sequential filters and extended maxima. (b) Resulting point pattern and bounding contour.

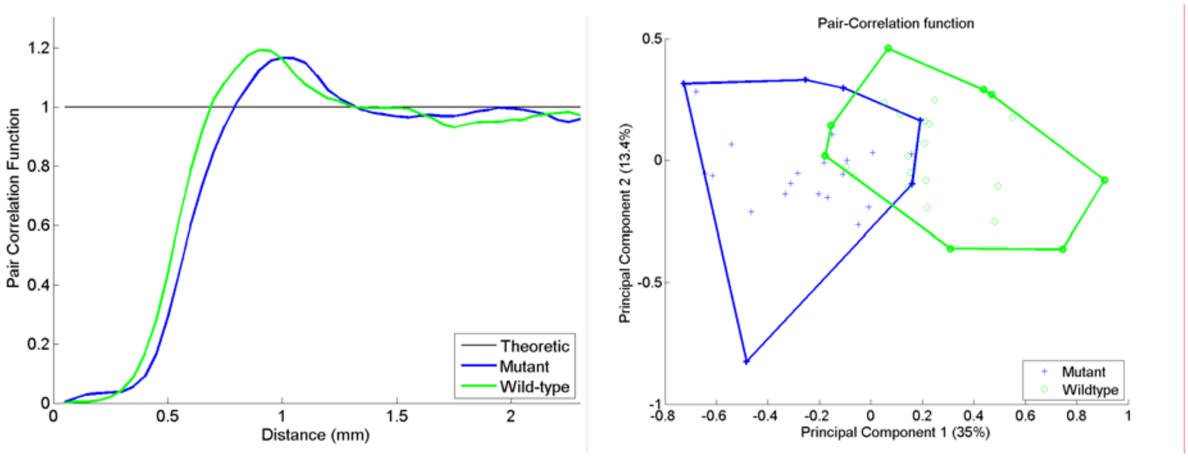
Spatial organisation analysis

Several descriptive functions were computed for each pattern: Ripley's K-function, F-Function, pair-correlation function [1]. All functions were computed using the same set of input distances, allowing group wise analysis of the patterns.

Principal component analysis was applied to enhance revealed specific interaction distances. Analyses of variance were used to detect differences between genotypes and/or slice position within the stem.

Results

A global difference in bundle numerical density was observed between genotypes. The pair-correlation function revealed an inhibition before 300-400 microns (Fig. 2-a). A characteristic interaction distance was observed, approximately equal to 0.9 mm for the wild type, and to 1 mm for the mutant.



(a) (b)

Figure 2 (a) Computation of average pair-correlation functions computed on all wild-type (green) and mutant (blue) slices. (b) Classification of the images with respect to the first two principle components of the pair correlation functions.

By applying principle components analysis on array formed by the pair correlation functions, it was possible to discriminate genotypes (Fig. 2-b). The difference in principal vectors reflects the difference in the interaction for distances around 1 mm.

Perspectives

Current works focus on the use of an estimate of local bundles density to improve the estimation of descriptive functions, and on modelling of the point spatial distribution of the within the stem using Strauss or Hardcore models. Results will be compared and coupled with those obtained from other analytical methods, such as physico-chemical analyses. The integration with mechanical models is also envisioned.

References

[1] A. Baddeley, R. Turner, “spatstat: An R Package for Analyzing Spatial Point Patterns”. Journal of Statistical Software, Vol. 12 (6), 2005, pp 1-42.

[2] M.-F Devaux, B. Bouchet, D. Legland, F. Guillon, M. Lahaye, “Macro-vision and grey level granulometry for quantification of tomato pericarp structure”. Postharvest Biol. Technol., Vol. 47, 2008, pp. 199-209

[3] P. Soille, Morphological Image Analysis, 2nd edition, Springer, 2003

A Sequential Point Process Model for Spatial Point Patterns with Linear Structures

Jakob Rasmussen
Aalborg University, Department of Mathematical Sciences

Many observed spatial point patterns contain points placed roughly on line segments. Point patterns exhibiting such structures can be found for example in archaeology (locations of bronze age graves in Denmark) and geography (locations of mountain tops). We consider a particular class of point processes whose realizations contain such linear structures. This point process is constructed sequentially by placing one point at a time. The points are placed in such a way that new points are often placed close to previously placed points, and the points form roughly line shaped structures. We consider Markov chain Monte Carlo based estimation for this class of point processes in a Bayesian setup. This is exemplified by real data.

Analysis of spatial structure of epidermal nerve entry point patterns based on replicated data

Aila Särkkä
(joint with Mari Myllymäki and Ioanna Panoutsopoulou)

Epidermal nerve fiber (ENF) density and morphology are used to diagnose small fiber involvement in diabetic and other small fiber neuropathies. ENF density and summed length of ENFs per epidermal surface area are reduced, and ENFs may appear more clustered within the epidermis in subjects with small fiber neuropathy compared to healthy subjects. Therefore, it is important to understand the spatial behavior of ENFs in healthy and diseased subjects. We have investigated the spatial structure of ENF entry points, which are the locations where the nerves enter the epidermis (the outmost living layer of the skin). The study is based on suction skin blister specimens from two body locations of 25 healthy subjects. The ENF entry points are regarded as a realization of a spatial point process and Ripley's K function is used to investigate the effect of covariates (gender, age and body mass index) on the degree of clustering of ENF entry points. The effects of covariates and individual variation are characterized by a mixed model approach.

Statistical aspects of determinantal point processes

Jesper Møller, Aalborg University

Determinantal point processes are largely unexplored in statistics, though they possess a number of appealing properties and have been studied in mathematical physics, combinatorics, and random matrix theory. In this talk we consider statistical aspects of determinantal point processes defined on $\mathbb{R}^d$, with a focus on $d = 2$.

Determinantal point processes are defined by a function C satisfying certain regularity conditions and they possess the following properties:

(a) Determinantal point processes are flexible models for repulsive interaction.

(b) All orders of moments of a determinantal point process are described by certain determinants of matrices with entries given in terms of C.

(c) A one-to-one smooth transformation or an independent thinning of a determinantal point process is also a determinantal point process.

(d) A determinantal point process can easily be simulated, since it is a mixture of 'determinantal projection processes'.

(e) A determinantal point process restricted to a compact set has a density (with respect to a Poisson process) which can be expressed in closed form including the normalizing constant.

In contrast Gibbs point processes, which constitute another flexible class of models for repulsive interaction, do not in general have moments that are expressible in closed form, the density involves an intractable normalizing constant, and rather time consuming Markov chain Monte Carlo methods are needed for simulations and approximate likelihood inference.

In the talk we describe how to simulate determinantal point processes in practice and investigate how to construct parametric models. Furthermore, different inferential approaches based on both moments and the likelihood are studied.

The work has been carried out in collaboration with Ege Rubak, Aalborg University, and Frédéric Lavancier, University of Nantes.

Definition and existence

In this abstract we just state the definition of a determinatal point process on $\mathbb{R}^d$ and conditions ensuring its existence. Due to lack of space, the simulation procedure, the density expression, and the statistical aspects of determinatal point processes are deferred to the talk.

Consider a simple locally finite spatial point process X on $\mathbb{R}^d$, i.e. we can view X as a random locally finite subset of $\mathbb{R}^d$. We refer to the elements (or points) of X as events. The following basic notions are needed before defining when X is a determinantal point process.

Recall that for an integer $n > 0$, X has n'th order product density function $\rho^{(n)} : \mathbb{R}^{nd} \to [0, \infty)$ if this function is locally integrable (with respect to Lebesgue measure) and for any Borel function $h : \mathbb{R}^{nd} \to [0, \infty)$,

$$\mathrm{E} \sum_{x_1,\dots,x_n \in X}^{\neq} h(x_1, \dots, x_n) = \int \cdots \int \rho^{(n)}(x_1, \dots, x_n) h(x_1, \dots, x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \quad (0.1)$$

where $\neq$ over the summation sign means that $x_1, \dots, x_n$ are pairwise distinct events. Intuitively, for any pairwise distinct points $x_1, \dots, x_n \in \mathbb{R}^d$, $\rho^{(n)}(x_1, \dots, x_n) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n$ is the probability that for each $i = 1, \cdots, n$, X has a point in an infinitesimally small region around x_i of volume $\mathrm{d}x_i$. Clearly, $\rho^{(n)}$ is only uniquely defined up to a Lebesgue nullset. We shall henceforth require that $\rho^{(n)}(x_1, \dots, x_n) = 0$ if $x_i = x_j$ for some $i \neq j$. This convention becomes consistent with Definition 0.1 below.

In particular, $\rho = \rho^{(1)}$ is the intensity function and $g(x, y) = \rho^{(2)}(x, y)/[\rho(x)\rho(y)]$ is the pair correlation function, where we set $g(x, y) = 0$ if $\rho(x)$ or $\rho(y)$ is zero. By our convention above, $g(x, x) = 0$ for all $x \in \mathbb{R}^d$. The terminology 'pair correlation function' may be confusing, but it is commonly used by spatial statisticians. In fact, for disjoint bounded Borel sets $A, B \subset \mathbb{R}^d$, if $N(A)$ denotes the number of events falling in A, then the covariance between $N(A)$ and $N(B)$ is the integral over $A \times B$ of the covariance function given by $c(x, y) = \rho(x)\rho(y)(g(x, y) - 1)$ for $x \neq y$. For a Poisson point process with an intensity function ρ, and for $x \neq y$, we have $c(x, y) = 0$, while $g(x, y) = 1$ if both $\rho(x) > 0$ and $\rho(y) > 0$. In spatial statistics and stochastic geometry, g is more commonly used than c, and we shall also pay attention to g.

Let $\mathbb{C}$ denote the complex plane. For a complex number $z = z_1 + \mathrm{i}z_2$ (where $z_1, z_2 \in \mathbb{R}$ and $\mathrm{i} = \sqrt{-1}$), we denote $\overline{z} = z_1 - \mathrm{i}z_2$ the complex conjugate and $|z| = \sqrt{z_1^2 + z_2^2}$ the modulus.

For any function $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$, let $[C](x_1, \dots, x_n)$ be the $n \times n$ matrix with (i, j)'th entry $C(x_i, x_j)$. For a square complex matrix A, let $\det A$ denote its determinant.

Definition 0.1. *Suppose that a simple locally finite spatial point process X has product density functions*

$$\rho^{(n)}(x_1, \dots, x_n) = \det[C](x_1, \dots, x_n), \quad (x_1, \dots, x_n) \in \mathbb{R}^{nd}, \quad n = 1, 2, \dots \quad (0.2)$$

Then X is called a determinantal point process *with kernel C, and we write $X \sim$ DPP(C).*

Existence of a determinantal point process is ensured by the following assumptions on C, where $S \subset \mathbb{R}^d$ denotes a generic compact set. Let $C : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ be Hermitian, i.e. $C(x, y) = \overline{C(y, x)}$ for all $x, y \in \mathbb{R}^d$. In addition, assume that C is continuous. Denote $L^2(S)$ the space of square-integrable functions $h : S \to \mathbb{C}$ and define the integral operator $T_S : L^2(S) \to L^2(S)$ by

$$T_S(h)(x) = \int_S C(x, y) h(y) \, \mathrm{d}y, \quad x \in S.$$

By Mercer's theorem, for any compact set $S \subset \mathbb{R}^d$, C restricted to $S \times S$ has a spectral representation,

$$C(x, y) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \overline{\phi_k(y)}, \quad (x, y) \in S \times S, \tag{0.3}$$

with absolute and uniform convergence of the series, and where

- the set of eigenvalues $\{\lambda_k\}$ is unique, each non-zero eigenvalue is real and has finite multiplicity, and the only possible accumulation point of the eigenvalues is 0;
- the eigenfunctions $\{\phi_k\}$ form an orthonormal basis of $L^2(S)$, i.e.

  $$T_S(\phi_k) = \lambda_k \phi_k, \quad \int_S \phi_k(x)\overline{\phi_l(x)}\, \mathrm{d}x = \begin{cases} 1 & \text{if } k = l, \\ 0 & \text{if } k \neq l, \end{cases} \tag{0.4}$$

  and any $h \in L^2(S)$ can be written as $h = \sum_{k=1}^{\infty} \alpha_k \phi_k$ where $\alpha_k \in \mathbb{C}$, $k = 1, 2, \ldots$. Moreover, ϕ_k is continuous if $\lambda_k \neq 0$.

When we need to stress that the eigenvalue λ_k depends on S, we write λ_k^S. We say that C (or T_S) is of local trace class if $\mathrm{tr}(C) = \int_S C(x, x)\, \mathrm{d}x$ is finite, i.e.

$$\mathrm{tr}(C) = \sum_{k=1}^{\infty} |\lambda_k^S| < \infty \quad \text{for all compact } S \subset \mathbb{R}^d. \tag{0.5}$$

Finally, we introduce the following conditions (C1) and (C2), recalling that C is a complex covariance function if and only if it is Hermitian and non-negative definite:

(C1) C is a continuous complex covariance function;

(C2) $\lambda_k^S \leq 1$ for all compact $S \subset \mathbb{R}^d$ and all k.

Theorem 0.2. *Under (C1), existence of DPP(C) is equivalent to (C2).*

Usually, for statistical models of covariance functions, (C1) is satisfied, and so (C2) becomes the essential condition. As discussed in the talk, (C2) simplifies in the stationary case of X.

Continuum Percolation in the β-skeleton Graph

J.-M. Billiot, F. Corset and E. Fontenas
LJK, FIGAL Team, BSHM, Université Pierre Mendès France,
1251 avenue centrale, B.P. 47, 38040 Grenoble cedex 9, France
Jean-Michel.Billiot@upmf-grenoble.fr

Percolation theory is very useful to describe various physical phenomena. In particular, there are important connections with phase transition problems [4, 3, 1].

The interest for percolation problems has grown rapidly during the last decades: see Lyons and Peres [7] for percolation on trees and networks and Meester and Roy [8] for continuum percolation and the references therein. In 1996, Häggström and Meester [5] proposed results for continuum percolation problems for the k-nearest neighbor graph under Poisson process. In a recent paper, Balister and Bollobás [2] give bounds on k for the k-nearest neighbor graph for percolation with several possible definitions.

Kirkpatrick and Radke [6] defined a parameterized family of neighborhood graphs called β-skeletons. The neighborhood $U_{p,q}(\beta)$, (p and q two vertices of the graph) is defined for any fixed $\beta \geq 1$ as the intersection of two spheres:

$$U_{p,q}(\beta) = B(p + \beta/2(q-p), \beta\delta(p,q)/2) \cap B(q + \beta/2(p-q), \beta\delta(p,q)/2)$$

where $\delta(p,q)$ is the distance between p and q.
The (lune-based) β-skeleton of V (the set of the vertices of the graph) is a neighborhood graph with the set of edges defined as follows :

$$(p,q) \text{ is an edge} \Leftrightarrow U_{p,q}(\beta) \cap V = \emptyset.$$

These graphs includes the Gabriel graph ($\beta = 1$) and the relative neighborhood graph (RNG) ($\beta = 2$) which are important for many applications.

We study the percolation for the previous graphs when the points are distributed under a stationary Poisson point process with unit intensity in the plane. We adapted to our case a method of the rolling ball proposed by [2] relying on 1-independent bond percolation on $\mathbb{Z}^2$.

References

[1] O. Riordan B. Bollobas. *Percolation*. Cambridge University Press, Cambridge, 2006.

[2] P.N. Balister and B. Bollobás. Percolation in the k-nearest neighbor graph. *Manuscript*, 2010.

[3] H.-O. Georgii, O. Häggström, and C. Maes. The random geometry of equilibrium phases. In C. Domb and J.L. Lebowitz, editors, *Phase Transitions and Critical Phenomena*, volume 18, pages 1–142, London, 2001. Academic Press.

[4] G.R. Grimmet. *Percolation, Second Edition*. Springer, New York, 1999.

[5] O. Häggström and R. Meester. Nearest Neighbor and Hard Sphere Models in Continuum Percolation. *Rand. Struct. Algorithms*, 9(3):295–315, 1996.

[6] D. G. Kirkpatrick and J.D. Radke. A framework for computational morphology. *In G. Toussaint, editor. Computational Geometry*, North-Holland:217–248, 1985.

[7] R. Lyons and Y. Peres. *Probability on Trees and Networks*. Cambridge University Press, Cambridge, 2002.

[8] R. W. J. Meester and R. Roy. *Continuum Percolation*. Cambridge University Press, Cambridge, 1996.

Simulation of chromosome theoretical models using Langevin-Hastings algorithms

Kiên Kiêu[1], Katarzyna Adamczyk[1], Clémence Kress[2]

[1] UR 341 Mathématiques et Informatique Appliquées
[2] UR 1196 Génomique et Physiologie de la Lactation
INRA, Jouy-en-Josas

Chromosomes, which carry the genetic information, are highly folded inside the cell nucleus. For instance, in human cells the DNA fibers extend on about 2 meters contained in a nucleus of a few microns diameter. Moreover it is well-known that chromosome compaction is far from being spatially uniform. The interplay between the spatial heterogeneity of chromosomes and the main biological functions involving DNA (replication, repair and RNA transcription) is an active domain of research.

By modelling, it is possible to investigate the spatial organization of chromosomes at the level of the whole genome. So-called coarse-grain models are based on theoretical models of polymers developped by physicists (see e.g. [5, 2, 3]). In a typical model, chromosomes are discretized into loci (from several hundred to several thousand loci). The joint distribution of locus spatial positions is represented as a Gibbs model involving several types of interactions. Most commonly represented interactions are springs keeping consecutive loci on the same chromosome close to each other and volume exclusion (pairwise repulsion).

For simulating such models, one resorts on random walk Metropolis algorithms. However due to the high dimension of the model, obtained simulations seem to show poor mixing and hardly converge. For instance, Kreth et al. [2] observed that their simulated chromosome arrangements were dependent on the initial configuration.

We consider an alternative simulation algorithm namely Langevin-Hastings algorithm [4] which is expected to perform better for target distributions with high dimensions. Langevin proposals try to decrease energy using its current gradient. The algorithm involves a single parameter: a variance/covariance matrix which controls the amplitude of displacements. Since the efficiency of Langevin-Hastings is highly dependent of the tuning variance/covariance matrix, we consider also the adaptive strategy proposed by Atchadé [1]. Simulation algorithms are compared through numerical experiments and based on their autocorrelation functions.

References

[1] Y. F. Atchadé. An adaptive version for the Metropolis adjusted Langevin algorithm with a truncated drift. *Methodol. Comput. Appl. Probab.*, 8:235–254, 2006.

[2] G. Kreth, J. Finsterle, J. von Hase, M. Cremer, and C. Cremer. Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. *Biophysical Journal*, 86(5):2803–12, May 2004.

[3] J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. M. Manders, P. J. Verschure, M. H. G. Indemans, H. J. Gierman, D. W. Heermann, R. van Driel, and S. Goetze. Spatially confined folding of chromatin in the interphase nucleus. *Proceedings of the National Academy of Sciences of the United States of America*, 106(10):3812–7, Mar. 2009.

[4] G. O. Roberts and R. L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

[5] R. K. Sachs, G. van Den Engh, B. Trask, H. Yokota, and J. E. Hearst. A random-walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 92(7):2710–2714, 1995.

Weeds Sampling for Map Reconstruction: a Markov Random Field Approach

M. Bonneau[1,2], S. Gaba[2], N. Peyrard[1], R. Sabbadin[1]
[1]Unité de Biométrie et Intelligence Artificielle - UR 875
INRA Toulouse - France
[2]UMR Biologie et Gestion des Adventices
INRA Dijon - Université de Dijon - France

1 Introduction

In the past 15 years, there has been a growing interest for the study of the spatial repartition of weeds in crops, mainly because this is a prerequisite to herbicides use reduction. There has been a large variety of statistical methods developped for this problem ([5], [7], [10]). However, one common point of all of these methods is that they are based on in situ collection of data about weeds spatial repartition. A crucial problem is then to choose where, in the field, data should be collected. Since exhaustive sampling of a field is too costly, a lot of attention has been paid to the development of spatial sampling methods ([12], [4], [6] [9]). Classical spatial stochastic model of weeds counts are based on Cox processes [3] or kriging [7]. In this work we propose to deal with abundance classes and to adopt a Markov Random Field (MRF) framework.

In a companion paper [2], we present an approach for spatial sampling which is based on MRF. This approach relies on an a priori model of the repartition of weeds in crops. It also relies on a model of sampling costs (time spent to sample), in order to mimic field constraints. The goal of this talk is to present the modelling choices that we have made in order to apply the approach [2] to the sampling and reconstruction problem for a real case study with a large data set of partial samples, in various conditions (weeds, crop, date...). In section 2 we present the model selection study that we have performed in order to build the a priori MRF model of weeds repartition. Then, in section 2, we present the sampling (time) cost model that we have built. Finally, in section 4 we discuss the use of the sampling approach [2] for weeds sampling in crop fields.

2 A MRF model of the abundance repartition of a weed species

2.1 Candidates pairwise MRF models

Let us recall briefly the definition of a pairwise MRF distribution. Let $X = (X_1, \dots, X_n)$ be discrete random variables taking values in $\Omega^n = \{0, \dots, K\}^n$. $V = \{1, \dots, n\}$ is the set of indices of the vector X and an element $i \in V$ will be called a *site*. If $G = (V, E)$ is the graph associated with the MRF, then $\forall(x_1, \ldots, x_n) \in \Omega^n$,

$$\mathbb{P}(x_1, \ldots, x_n) \propto \prod_{i \in V} e^{\psi_i(x_i)} \prod_{(i,j) \in E} e^{\psi_{ij}(x_i, x_j)}.$$

For modeling the map distribution of a particular weed species, we define G as a regular rectangular grid representing a decomposition of the field into quadrats (which are also the sampling units). We considered a first order neighbourhood (2 closest neighbours in each field direction). The variable X_i is the abundance class on quadrat i. For example using Barralis classes [1] : $\Omega = \{0, \ldots, 6\}$ with 0 corresponding to the absence of the species. The choice of an appropriate MRF model for mapping weed abundance classes distribution amounts to the choice of adapted potential functions ψ_i and ψ_{ij}. We considered several options : Potts model with or without external field and with or without anisotropy. The more complex is the Potts model with external field and with aniisotropy :

$$\forall(i,j) \in E, \forall k, l \in \Omega \begin{cases} \psi_i(k) & = \alpha_k, \\ \psi_{ij}(k,l) & = \beta_s \mathbb{1}_{\{k=l\}} \text{ if edge } (i,j) \text{ is along tillage direction} \\ \psi_{ij}(k,l) & = \beta_o \mathbb{1}_{\{k=l\}} \text{ if edge } (i,j) \text{ is orthogonal to tillage direction} \end{cases}$$

where α_k, β_s and β_o are real valued parameters. The three other models are derived by setting all the αs equal and/or $\beta_s = \beta_o$. We also considered an alternative to the Potts model, where we impose a smooth spatial variation of the abundance classes : the order-2 potentials are modified as follows

$$\forall(i,j) \in E, \forall k, l \in \Omega \begin{cases} \psi_{ij}(k,l) & = \beta_s(1 - \frac{|k-l|}{K}) \text{ if edge } (i,j) \text{ is along tillage direction} \\ \psi_{ij}(k,l) & = \beta_o(1 - \frac{|k-l|}{K}) \text{ if edge } (i,j) \text{ is orthogonal to tillage direction} \end{cases}$$

2.2 Model selection

The analysis was performed on 6 species, sampled in different cropping systems, at different periods of the year. For each situation, the data available consist of samples of abundance classes of the weed species within a crop field. We used variational versions of the EM algorithm and the BIC criterion [8] to estimate the parameters of each of the eight candidate models and estimate their BIC score. We obtained the following conclusions : *i*) for a large majority of situations, the isotropic Potts model without external field is the best candidate to represent abudance maps distribution, and *ii*) the MRF model with smooth variation is clearly not adapted. The latter conclusion is in coherence with results from the litterature which claim that variations of weeds abundances are often abrupt within a crop field.

3 Cost of sampling

Sampling is adaptive and divided into H steps. One quadrat is sampled at each step. The cost incured to sampling plan A defines the effort necessary for executing this sampling plan. From discussions with experts, we defined this cost based on the time spent to execute A. If $A = \{a_1, \ldots, a_H\}$ are the indices of observed quadrats, we suppose that the overall time cost, denoted $c(A)$, is the sum of times spent for observing each quadrat. That is $c(A) = \sum_{i=1}^{H} c(a_i)$. We propose a linear model which expresses the time spent for observing a quadrat as a function of variables $(Z_1, \ldots, Z_5)$, representing respectively : the period of observation, the number of weed individuals, the number of species, crop and farming practices. Period of observation is a binary variable with value {*favorable, unfavorable*} depending of the recovery stage of the crop. We consider five different farming practices, depending on the quantity of pesticide used. For fitting the parameters of this model we use a 18300 length dataset which is a result of a nine-years experiment in Dijon-Epoisses. Eight different crops have been tested. Coefficients of the linear model were fitted using a linear regression with the software R.

4 Applying LSDP to weeds sampling

Once a model of the abundance repartition of a weed species is established, it can be used for finding new sampling policies which realise a trade-off beetween quality of the reconstructed map and cost of sampling. One way to compute this policy is based on the LSDP algorithm described in [2]. The main constraint for applying LSDP to weeds sampling problems is the large number of quadrats within a field. For now the LSDP algorithm gives interesting results for problems with 100 quadrats which is much less than the possible number of quadrats within a field. For exemple [11] report that the size of experimental fields usually varies between 0.019 and $173ha$. The same authors report that a quadrat size varies usually between 0.025 and $1.46m^2$. For solving this problem one solution is to divide the overall field into subfields and solve the sampling problem into each subfield. Another possibility is to combine heuristic(s) strategy(ies) and the LSDP algorithm for solving the sampling problem. This two approaches are currently investigated.

5 Discussion

In this work, we propose an alternative to classical kriging approaches or point processes models for representing the spatial distribution of weeds abundance. This seems to be more adapted to the observed non smooth spatial variation of weeds abundance. We are currently testing this hypothesis by extending our model selection work to a new candidate : the log normal Cox process [3]. It could also be worth investigating the adaptation of our Reinforcement-Learning approach [2] to propose sampling strategies relying on this latter model.

Then, the combination of simple heuristic strategies and more complex ones (like the LSDP one) should lead to a promising avenue for designing spatial sampling strategies for weeds with a satisfying trade-off between evaluation complexity and map reconstruction quality.

Références

[1] G. Barralis. Méthode d'étude des groupements adventices des cultures annuelles. In *Colloque International sur l'écologie et la Biologie des Mauvaise herbes*, pages 59–68, Dijon-France, 1976.

[2] M. Bonneau, N. Peyrard, and R. Sabbadin. A reinforcement-learning algorithm for sampling design in markov random fields. In *Proc. of SSIAB 2012*.

[3] A. Bourgeois, S. Gaba, N. Munier-Jolain, B. Borgy, P. Monestiez, and S. Soubeyrand. Inferring weed spatial distribution from multi-type data. *Ecological Modelling*, 226 :92–98, 2012.

[4] R.D. Cousens, R.W. Brown, A.B. McBratney, B. Whelan, and M. Moerkerk. Sampling strategy is important for producing weed maps : a case study using kriging. *Weed science*, 50(4) :542–546, July 2002.

[5] M.R.T. Dale, P. Dixon, M.J. Fortin, P. Legendre, D.E. Myers, and M.S. Rosenberg. Conceptual and mathematical relationships among methods for spatial analysis. *Ecography*, 25(5) :558–577, October 2002.

[6] J. de Gruijter, D. Brus, M. Bierkens, and K. Knotters. *Sampling for Natural Resource Monitoring.* Springer, 2006.

[7] J.A. Dille, M. Milner, J.J. Groeteke, D.A. Mortensen, and M.M. Williams. How good is your weed map ? a comparaison of spatial interpolators. *Weed science*, 51(1) :44–55, January 2003.

[8] F. Forbes and N. Peyrard. Hidden markov random field model selection criteria based on mean field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 :1089–1101, 2003.

[9] WG Müller. *Collecting spatial Data.* Springer Verlag : Heidelberg, 2007. 3rd ed.

[10] J.N. Perry, M. Liebhold, S. Rosenberg, J. Dungan, M. Miriti, A. Jakomulska, and S. Citron-Pousty. Illustrations and guidelines for selecting statistical methods for quatifying spatial pattern in ecological data. *Ecography*, 25 :578–600, 2002.

[11] L.J. Rew and R.D. Cousens. Spatial distribution of weeds in arable crops : are current sampling and analytical methods appropriate ? *Blackwell Science Ltd Weed Research*, 41 :1–8, 2001.

[12] L.J. Wiles. Sampling to make maps for site specific weed management. *Weed science*, 53(2) :228–235, March 2005.

A Reinforcement-Learning Algorithm for Sampling Design in Markov Random Fields

M. Bonneau, N. Peyrard, R. Sabbadin
Unité de Biométrie et Intelligence Artificielle - UR 875
INRA Toulouse - France

1 Introduction

Optimal sampling in spatial random fields is a complex problem, which mobilizes several research fields in spatial statistics and artificial intelligence. An active stream of research about optimal spatial sampling is dedicated to the study of the case of real-valued observations (e.g. temperature or pollution monitoring). Models and efficient algorithms have been proposed, mainly based on the geostatistical framework of Gaussian random fields and kriging. Much less attention has been paid to the case of discrete-valued observations. However, this problem is ubiquitous in many studies about biological systems. Discrete-valued observations can be species abundance classes, disease severity classes, presence/absence values...

Solving optimal sampling problems in discrete-valued random fields is a difficult question admitting no universally accepted solution, so far. We propose, similarly to [2, 5], to define the optimal spatial sampling problem within the framework of Markov random fields (MRF), classically used in image analysis. We consider the case of adaptive sampling, where the set of sampled sites is chosen sequentially, taking into account observations from previous sampling steps. Simple heuristics have been proposed [8, 1, 5] to design adaptive sampling strategies. However, it is difficult to evaluate their quality since there is no efficient exact method to compare to. In this paper, we design a new Reinforcement-Learning (RL, [7]) algorithm which improves classical heuristic and RL approaches, thus providing a reference algorithm. The algorithm, named LSDP (Least Square Dynamic Programing) uses an encoding of the optimal adaptive sampling problem as a finite-horizon Markov Decision Process (MDP, [6]).

The MRF formalization of the optimal adaptive spatial sampling problem is introduced in Section 2. Then, we describe the LSDP algorithm in Section 3. We present an empirical comparison between heuristic approaches, classical RL algorithms and LSDP in Section 4. Conclusions are drawn in Section 5.

2 Problem statement

Let $X = (X_1, \ldots, X_n)$ be discrete random variables taking values in $\Omega^n = \{1, \ldots, K\}^n$. $V = \{1, \ldots, n\}$ is the set of indices of the vector X and an element $i \in V$ will be called a *site*. The distribution $\mathbb{P}$ of X is that of a Markov Random Field (MRF) with associated graph $G = (V, E)$ where $E \subseteq V^2$ is a set of undirected edges. $x = (x_1, \ldots, x_n)$ is a realization of X and we adopt the following notation: $x_B = \{x_i\}_{i \in B}, \forall B \subseteq V$.

In order to reconstruct the vector X on a specified subset $R \subseteq V$ of sites of interest, we can acquire a limited number of observations within a subset $O \subseteq V$ of observable sites. We will assume that $R \cup O = V$ and intersection between O and R can be non-empty. Our objective is to choose sequentially $A \subseteq O$ so that the updated distribution $\mathbb{P}(\cdot|x_A)$ becomes as informative as possible (in expectation over all possible sample outputs x_A).

Adaptive sampling policy. The sampling plan is divided into H steps and The choice of sample $A^h \subseteq O$ depends on the previous sample outputs. An adaptive sampling policy $\delta = (\delta^1, \ldots, \delta^H)$ is then defined by an initial sample A^1 and functions δ^h specifying the sample chosen at step $h \geq 2$, depending on previous observations: $\delta^h((A^1, x_{A^1}), \ldots, (A^{h-1}, x_{A^{h-1}})) = A^h$. A history is a trajectory $(A^1, x_{A^1}), \ldots, (A^H, x_{A^H})$ followed when applying policy δ. The set of all histories which can be followed by policy δ is τ_δ.

Quality of a sampling policy. We first define the quality of a history $((A_h, x_{A_h}))_{h=1..H}$ as a function of (A, x_A), where $A = \cup_h A_h$:

$$U(A, x_A) = \sum_{i \in R} \left[\max_{x_i \in \Omega} \Big\{ \mathbb{P}(x_i \mid x_A) \Big\} \right]. \tag{1}$$

The quality of a sampling policy δ is then defined as an expectation over all possible histories: $V(\delta) = \sum_{((A_h, x_{A_h}))_h \in \tau_\delta} \mathbb{P}(x_A) U(A, x_A)$.

Optimal adaptive sampling in MRF. Finally the problem of optimal adaptive sampling amounts to finding the policy of highest quality, subjet to a cost contraint bounding the sum of the costs of all sampled sites by B, for all trajectories in τ_δ :

$$\delta^* = \arg \max_{\delta, c(\delta) \leq B} V(\delta). \tag{2}$$

3 The LSDP algorithm

The optimisation problem (2) can be solved using tools from the field of Markov Decision Processes (MDP, [6]) and Reinforcement Learning (RL, [7]). In a MDP, at each time step t an agent takes a decision d^t and the system moves from state s^t to state s^{t+1} according to $p(s^{t+1} \mid s^t, d^t)$. A reward r^t is obtained. In the spatial sampling problem, the state is $s^t = \{s^t_1, \ldots, s^t_{|O|}\}$ with $s^t_i = x_i$ if site i has been sampled, and 0 otherwise. Decision d^t is the next set of sites to sample. Rewards r^t are null, except r^{H+1} which is equal to $\sum_{i \in R} \left[\max_{x_i \in \Omega} \Big\{ \mathbb{P}(x_i \mid x_A) \Big\} \right]$. Then, solving (2) amounts[1] to finding the optimal admissible policy δ^* such that $V^*(s,t) \equiv V^{\delta^*}(s,t) \geq V^\delta(s,t), \forall \delta, s, t$, with $V^\delta(s,t) = \mathbb{E}_\delta \left[\sum_{t'=t}^{H+1} r^{t'} \mid s \right], \forall (s,t) \in S \times T.$.

The backwards induction algorithm [6] can be applied to compute the optimal policy. However exact dynamic programming is inapplicable to large problems. Therefore, we have to look for sub-optimal policies. Classical approaches are based on RL (TD(λ) [7], LSPI [4]): the idea is to use repeated simulated transitions (s^t, d^t, r^t, s^{t+1}), instead of apply dynamic programming updates, in order to estimate the optimal $Q-function$ defined as $Q^*(s,d,t) = r^t(s,d) + \sum_{s'} p^t(s'|s,d) V^*(s', t+1$. Then $V^*(s,t) = \max_d Q^*(s,d,t)$ and $\delta^{*,t}(s) = \pi^*(s,t) = \arg\max_d Q^*(s,d,t)$. Still, classical RL algorithms cannot handle problems as complex as the optimisation problem (2).

The originality of the algorithm we propose is to combine three elements in order to approximate Q^*: i) a parametrized representation of the Q-function with time dependent weights, ii) dynamic programming iterations, iii) and simulation of histories. Namely, we consider an approximation of Q^* as a linear combination of n arbitrary *features* derived from the heuristic BP-max sampling method proposed in [7]. Then, a batch of maps is simulated off-line and trajectories (sequences of transitions) are simulated on-line. Finally, to compute the features and the quality of a history (1), marginal distributions are approximated, using the Belief Proparation (BP) algorithm.

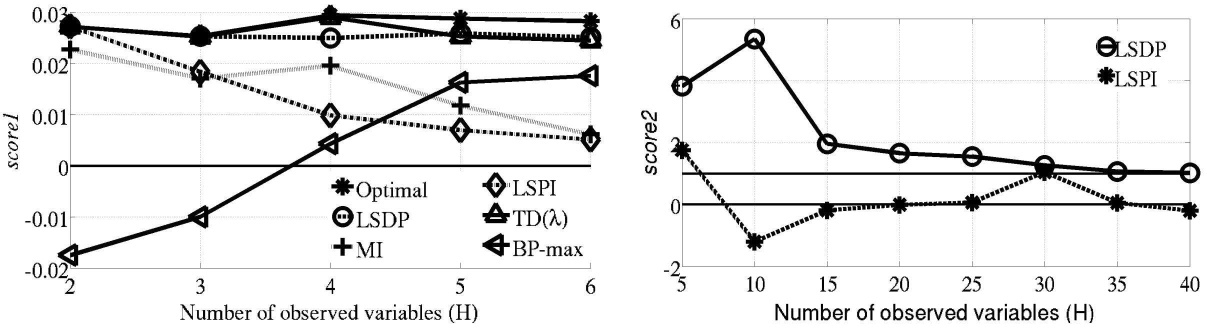
[1] It can be rigorously established that the two optimisation problems are equivalent

4 Experimental evaluation

We compared LSDP to the random heuristic, TD(λ) algorithm with tabular representation of the Q-function, LSPI, and the BP-max policy of [7]. The BP-max policy consists in sampling at each step the sites with the most uncertain marginals. We also compared LSDP to a greedy algorithm based on the Mutual Information (MI) criterion [3].

The sampling problem considered is the following. The graph G is a regular grid and $R = O = V$. One variable is observed at each decision step and sampling costs are null. We considered the following Potts model distribution: $\forall\, x \in \{1,2\}^n\ \mathbb{P}(x) \propto \exp\left(\frac{1}{2}\sum_{(i,j)\in E} \mathbb{1}_{\{x_i=x_j\}}\right)$.

4×4 grid. We were able to compute the optimal policy solution of (2) and the exact value of any policy. The first conclusion is that the absolute difference between the values of all policies is small: an absolute increase of the percentages of 2.2 at most. We also compared the policies in terms of $score1(\delta) = \frac{V(\delta)-V(\delta_R)}{V(\delta^*)-V(\delta_R)}$ (see Figure 1 (a)). Among RL algorithms, TD(λ) is the best and LSDP gives very similar results. In comparaison, LSPI shows a poor behaviour, always returning dominated policies. Surprisingly the relative value of the MI policy decreases with the number of observed variables, while the opposite behavior is observed for the BP-max heuristic. Indeed with few observed sites, all sites have similar marginal probabilities, leading to a purely random choice of samples with BP-max.



(a) (b)

Figure 1: Comparison of LSDP, classical RL and heuristic policies (a) on a problem with 16 variables ($score1$), (b) on a problem with 100 variables ($score2$).

10×10 grid. For this problem size, only LSDP, LSPI, BP-max and the random policy can be computed. We compared them using $score2(\delta) = \frac{V(\delta)-V(\delta_R)}{|V(\delta_{BP-max})-V(\delta_R)|}$. We observed again poor performance of the LSPI algorithm. On the contrary, LSDP performs quite better than the BP-max heuristic for small sample sizes (see Figure 1 (b)). LSDP also performs better than LSPI, in terms of computation time: for $H = 40$, an iteration takes about 7 seconds for LSDP, 77 seconds for LSPI.

Constrained moves problem. Finally, we compared LSDP, BP-max and random policies on a more realistic sampling problem, involving constrained moves on the grid for observing sites. After having observed a site, the agent can only move to distance-2 sites for the following observation. We again observed a small absolute difference between all policies. LSPI still showed poor behaviour. As we expected, the gain provided by LSDP in terms of relative improvement of the random policy ($H \leq 20$) is significant when the sample size is small (Figure 2).

5 Conclusion

We proposed a MDP representation for the problem of optimal adaptive sampling of spatial processes expressed in the Markov random field framework. This allowed us to propose an adapted

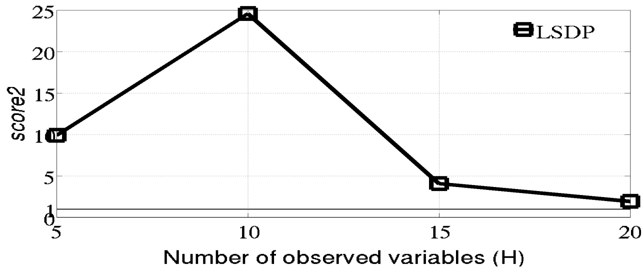


Figure 2: Constrained moves problem with 100 variables: $score2$ of LSDP policy.

simulation-based solution algorithm, LSDP, combination of a parametrized representation of the Q-function and Dynamic Programming principles.

Comparison of the LSDP algorithm with heuristic algorithms and classical RL algorithms enables us to draw the following conclusions. First we notice that the performance of a purely random strategy is quite close to that of the best available solution. However, in real-life applications of sampling for mapping, small gains in the reconstruction of maps are important since they can lead to significant reduction in management costs.

Second, for large problems only BP-max heuristic and the LSDP algorithm provide good results. BP-max is less costly to apply than LSDP. However, its performance depends on which form of sampling costs are considered. We can also predict poor performances when the set of observable variables differs from the set of variables of interest. In contrast, LSDP can handle different cost functions. It can also easily be adapted to other definitions of policy value, provided that they can be estimated efficiently from a batch of trajectories. Furthermore, the LSDP algorithm can be applied to general factored finite-horizon MDP, and not only to spatial sampling problems.

References

[1] J. de Gruijter, D. Brus, M. Bierkens, and K. Knotters. *Sampling for Natural Resource Monitoring*. Springer, 2006.

[2] A. Krause and C. Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, 35:557–591, 2009.

[3] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in Gaussian processes: theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

[4] M. Lagoudakis and R. Parr. Least-squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.

[5] N. Peyrard, R. Sabbadin, D. Spring, R. Mac Nally, and B. Brook. Model-based adaptive spatial sampling for occurrence map construction. *Statistics and Computing (to appear)*, 2012.

[6] M. Puterman. *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc, 1994.

[7] R. S. Sutton and A.G. Barto. *Reinforcement Learning : An Introduction.* MIT Press, 1998.

[8] S. Thompson and G. Seber. *Adaptive sampling*. Series in Probability and Statistics. Wiley, New York, 1996.

Estimating parameters in spatio-temporal Quermass-interaction process

Kateřina Staňková Helisová [1]

joint work with Markéta Zikmundová [2] *and Viktor Beneš* [2]

[1] Czech Technical University in Prague, Faculty of Electrical Engineering, Department of Mathematics, Technická 2, 16627 Prague 6, Czech Republic
Email: helisova@math.feld.cvut.cz

[2] Charles University in Prague, Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 18675 Prague 8, Czech Republic
Emails: zikmundm@karlin.mff.cuni.cz, benesv@karlin.mff.cuni.cz

Abstract

Consider a random set observed in a bounded window $W \subset \mathbf{R}^2$ in discrete times $k = 0, 1, ..., T$. The set is given by a union of interacting discs and it is developed in time so that the discs appear and disappear (but do not grow). In each time k, the set is described by the Quermass-interaction process, i.e. the probability density of any finite configuration $\mathbf{x} = (x_1, ..., x_n)$ of the discs $x_1, ..., x_n$ with respect to the probability measure of a stationary random-disc Boolean model is given by

$$f_{\theta^{(k)}}(\mathbf{x}) = \frac{\exp\{\theta_1^{(k)} A(U_\mathbf{x}) + \theta_2^{(k)} L(U_\mathbf{x}) + \theta_3^{(k)} \chi(U_\mathbf{x})\}}{c_{\theta^{(k)}}}, \tag{1}$$

where $A(U_\mathbf{x})$ denotes the area, $L(U_\mathbf{x})$ the perimeter and $\chi(U_\mathbf{x})$ the Euler-Poincare characteristic of the union $U_\mathbf{x}$ composed of the discs from the configuration $\mathbf{x}$. Further, for each time k, $\theta^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$ is a vector of parameters and $c_{\theta^{(k)}}$ is a normalizing constant.

The temporal evolution of the random set is given by the evolution of the parameters according to the relation

$$\theta^{(k)} = \theta^{(k-1)} + \eta^{(k)}, \ k = 1, 2 \ldots, T, \tag{2}$$

where $\theta^{(0)}$ fixed is given and $\eta^{(k)}$ are iid random vectors with Gaussian distribution $\mathcal{N}(a, \sigma^2 I)$, where $a \in \mathbf{R}^3, \sigma^2 > 0$ and I is the unit matrix.

The temporal dependence in the random set is defined within its simulation algorithm. We start the simulation so that we choose a fixed $\theta^{(0)}$ and according to (2), we simulate parameter vectors $\theta^{(k)}, k = 1, 2, \ldots, T$. Further, using classical birth-death Metropolis-Hastings algorithm MCMC (see [1]), we simulate a realization $\mathbf{x}_0$ from the density (1) with $\theta^{(0)}$. Then we simulate realizations $\mathbf{x}_k$, $k = 1, 2 \ldots, T$ from the density (1) with $\theta^{(k)}$ and the birth-death Metropolis-Hastings algorithm is used again, but with a special way of adding a disc: since the realizations are aimed to be dependent, the choice of a newly added disc in the algorithm depends on the previously simulated configuration $\mathbf{x}_{k-1}$ so that the proposal distribution of the newly added disc $Prop_k$ at time k is a mixture

$$Prop_k = (1 - \beta) \cdot Prop^{(RP)} + \beta \cdot Prop_{k-1}^{(emp)}, \quad \beta \in (0, 1),$$

where $Prop^{(RP)}$ is a distribution of the reference process, $Prop_{k-1}^{(emp)}$ is the empirical distribution obtained from the configuration $\mathbf{x}_{k-1}$ and β is a chosen constant describing power of time dependence. It means that $(\beta \times 100)\%$ of the added discs are taken from the previous configuration and the remaining discs are simulated randomly, so the dependence is stronger when β is bigger.

In this contribution, different methods for estimating the parameters $\theta^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)})$, $a = (a_1, a_2, a_3)$ and σ^2 will be described. More precisely, combination of MCMC maximum likelihood method described in [2] with regression methods and particle filter studied in [3] and [4] will be shown.

References

[1] Møller J., Helisová K. (2008): *Power diagrams and interaction processes for unions of discs.* Advances in Applied Probability **40**(2), 321–347.

[2] Møller J., Helisová K. (2010): *Likelihood inference for unions of interacting discs.* Scandinavian Journal of Statistics **37**(3), 365–381.

[3] Zikmundová M., Staňková Helisová K., Beneš V. (2012): *Spatio-temporal model for a random set given by a union of interacting discs.* Methodology and Computing in Applied Probability. Submitted.

[4] Zikmundová M., Staňková Helisová K., Beneš V. (2012): *On the use of particle MCMC in the parameter estimation of a spatio-temporal random set.* In preparation.

Level sets estimation of random compact sets

P. Heinrich [1]

R. S. Stoica [2]

V. C. Tran [3]

Université Lille 1
Laboratoire Paul Painlevé
59655 Villeneuve d'Ascq Cedex, France

Abstract :

The present work proposes a ready to use estimator based on level sets. This estimator approximates the mean shape or the expectation of a random set. The concrete motivation behind this is given by pattern recognition applications arising in applied domains such as astronomy, epidemiology or image processing.

There is no canonical definition for the mean shape of a random set. One possible approach is the so-called Vorob'ev expectation $\mathbb{E}_V(X)$, which is closely related to quantile sets. The estimator for $\mathbb{E}_V(X)$ we propose is consistent and is built from independent copies of X using spatial discretization. The control of discretization errors is handled using a mild regularity assumption on the boundary of X: a not too large 'box counting' dimension. Some examples are developed and applications to epidemiological and cosmological data are presented.

Key Words :

stochastic geometry, random closed sets, level sets and Vorob'ev expectation

[1] philippe.heinrich@math.univ-lille1.fr
[2] radu.stoica@math.univ-lille1.fr
[3] chi.tran@math.univ-lille1.fr

Pseudo Bayesian inference for intensity-dependent point processes

Kasper K. Berthelsen, Aalborg University, Denmark
Mari Myllymäki, Aalto University, Finland

We consider a marked point process models, which extend the models of Berthelsen & Møller (2008), Ho & Stoyan (2008) and Myllymäki & Penttinen (2009). Specifically we consider a marked pairwise interaction point process where interaction depends on the mark. The mark distribution in turn depends on the intensity of the point process, where we a priori assume that first order term of the process is given by a shot noise process.

Regarding inference we adopt a Bayesian approach. Posterior distributions are explored using Markov chain Monte Carlo (MCMC). Since the normalising constant is unknown, using conventional MCMC algorithms will require evaluating ratios of unknown normalising constants. There exists several solution to this problem including the approaches of Murray et al. (2006) and Møller et al. (2006) — both of which rely on perfect simulation of the model under consideration. As it turns out perfect simulation is not feasible, when analysing data of practical interest in our setting. As an alternative we use a conventional MCMC algorithm for sampling the posterior where the likelihood has been replaced by the pseudolikelihood.

In the outset, pseudolikelihood requires integrating over the product space of the location space and the mark space which usually involves some form of numerical integration. Fortunately it is possible and feasible to perform the integral over the (unbounded) mark space. We notice, as also pointed out by Rubak & Coeurjolly (2012), that the standard Berman-Turner approach considered by Baddeley & Turner (2000) for implementing the pseudoliklihood is intrinsically biased.

We apply our model to both real and simulated data sets.

Literature

Baddeley, A. and Turner, R. (2000). Practical maximum pseudolikelihood for spatial point patterns. *Aust. N. Z. J. Stat.* **42**(3), 283–322.

Berthelsen, K. and Møller, J. (2008). Non-parametric Bayesian inference for inhomogeneous Markov point processes. *Aust. N. Z. J. Stat.* **53**(3), 257–272.

Ho, L. P. and D. Stoyan (2008), Modelling marked point patterns by intensity-marked Cox processes, *Stat. & Prob. Letters* **78**, 1194–1199.

Møller, J., A. N. Pettitt, R. Reeves, and K. K. Berthelsen (2006). An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* **93**, 451—458.

Murray, I., Z. Ghahramani, and D. J. C. MacKay. (2006). MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Arlington, Virginia. AUAI Press.

Myllymäki, M. and A. Penttinen (2009). Conditionally heteroscedastic intensity-dependent marking of log Gaussian Cox processes. *Stat. Neerlandica*, **63** (4), 450–473.

Statistical analysis for the Johnson-Mehl germination-growth model

March 15, 2012

Jesper Møller and **Mohammad Ghorbani**
Department of Mathematical Sciences, Aalborg University

Abstract

Second-order summary statistics play a fundamental role for analyzing spatial and spatio-temporal point process data-sets and for model checking. In this paper the Johnson-Mehl germination-growth model defined by a secondary space-time point process Ψ obtained by a dependent thinning of a primary space-time inhomogeneous Poisson process Φ is considered. We estimate parameters of specific parametric models for the conditional intensity of Ψ using the maximum likelihood method, and we check goodness-of-fit of the estimated model using new functional summary statistics related to the inhomogeneous K-function and to the Palm distribution of the typical Johnson-Mehl cell. Our methodology is illustrated using simulated and real data-set.

Key words: spatio-temporal functional summary statistics; K-function; Palm distribution, typical Johnson-Mehl cell, Voronoi tessellation, pair correlation function.

A unified framework for measuring industry location characteristics based on marked spatial point processes

Florent Bonneu and Christine Thomas-Agnan

March 25, 2012

We propose a unified framework for defining measures of industrial concentration based on micro-geographic data. These encompass the Duranton-Overman and the Marcon-Puech indices. We discuss the basic requirements for such measures introduced by Duranton and Overman (2005) and we propose five additional requirements. We describe several types of concentration depending on the second order characteristics of the marginal patterns of positions and of marks but also on their mutual dependence. We also discuss the null assumptions classically used for testing aggregation of a particular sector. The framework we propose is based on some second order characteristics of marked spatial point processes discussed in Illian et al. (2008). The general measure involves a cumulative and a non-cumulative version. This allows us to propose an alternative version of the Duranton-Overman index with a proper baseline as well as a cumulative version of this index.

Krugman's theory of economic geography states that "instead of spreading out evenly around the world, production will tend to concentrate in a few countries, regions, or cities, which will become densely populated but also have higher levels of income."

There are numerous motivations for studying the geographic concentration of economic sectors. Such a measure allows to understand the determinants of localization, compare different sectors with respect to agglomeration/dispersion and predict the evolutions of localization. A similar question is that of co-localization and interactions between sectors for which measures can be generally derived from the former. Another related issue is cluster detection but we do not include this problem in the present paper.

Until 2000, all studies about geographic concentration of economic activity use areal data for measuring spatial concentration. The localization of firms is not available and data consists only in counts aggregated on administrative zones. There is a large literature on this topic and many measures including the Herfindahl index, the locational Gini index (which is the Gini index of the localization ratio), the Ellison-Glaeser index, the Maurel-Sédillot's index and many others. However these measures depend upon the aggregation level (Modifiable Areal Unit Problem) and most importantly they do not take geography into account: a permutation of the sites does not affect the measure !

A new vein of this literature arises in 2002 considering the treatment of micro-geographic data. This type of data usually consists in the precise location of firms together with a size measure such as the number of employees. Duranton, G. and Overman, H.G. (2005) introduce a measure based on the distribution of inter-distances between firms. Marcon, E. and Puech, F. (2002, 2010) introduce another measure based on Ripley's K-function. Combes P-J., Meyer T., and Thisse J-F. (2008) survey this literature. Espa, Giuliani and Arbia (2010) use a model-based approach to assess concentration. Duranton et Overman (2002) list five properties that a good measure of industrial concentration should satisfy. In this paper, we introduce five additional requirements BTA1 to BTA5.

Note that the Duranton-Overman index as well as the Marcon-Puech index are both inspired from the marked point process theory. However none of them corresponds to a well identified statistical parameter. This weakness relates to the absence of clear definition of the theoretical meaning of spatial concentration only introduced through empirical measures. We intend to fill this gap and make progress in the understanding of spatial concentration. None of the cited measures satisfies two of our new requirements. The last one is only satisfied by Marcon-Puech and not by Duranton-Overman. Our index satisfied the ten requirements.

References

[1] Duranton, G. and Overman, H.G. (2005) Testing for localization using micro-geographic data. *Review of Economic Studies* **72** 1077-1106.

[2] Illian, J., Illian P, Stoyan H. and Stoyan D. (2008) Statistical analysis and modelling of spatial point patterns, Wiley, Statistics in practice.

[3] Marcon, E. and Puech, F. (2002) A new method to evaluate spatial economic activity and its application to two french areas, preprint.

[4] Marcon, E. and Puech, F. (2010) Measures of the geographic concentration of industries : improving distance-based methods. *Journal of Economic Geography* **10(5)** 745-762.

[5] Moller, J. et Waagepetersen, R.P. (2004) *Statistical inference and simulation for spatial point processes.* vol. 100. Chapman & HallCRC.

[6] Schlather, M. (2001) On the second-order characteristics of marked point processes. *Bernoulli* **7**(1) 99-117.

Semi-parametric estimation of the Poisson intensity parameter for stationary Gibbs point processes

Nadia Morsli[1] and Jean-François Coeurjolly[2].
[1] Laboratory Jean Kuntzmann, Grenoble University, France.
nadia.morsli@imag.fr
[2] Laboratory Jean Kuntzmann, Grenoble University, France.
jean-francois.coeurjolly@upmf-grenoble.fr

In this work, we consider stationary Gibbs point process (sGpp) in $\mathbb{R}^d$. More precisely, we consider Gibbs models such that the Papangelou conditional intensity can be written for $u \in \mathbb{R}^d$ and $x \in \Omega$ (space of locally finite configurations in $\mathbb{R}^d$) as

$$\lambda(u, x; \beta^*) = \beta^\star \, \widetilde{\lambda}(u, x),$$

where $\beta^\star \geq 0$ is the "Poisson intensity" parameter and where $\widetilde{\lambda}$ is a function from $\mathbb{R}^d \times \Omega$ to $\mathbb{R}^+$. We propose to estimate $\beta^\star$ independently of $\widetilde{\lambda}$. For this, we only assume that the sGpp has a finite range $R < +\infty$ (i.e. the associated Papangelou conditional intensity satisfies $\lambda(u, x; \beta^*) = \lambda(u, x_{B(u,R)}; \beta^*)$, for any $u \in \mathbb{R}^d$ and $x \in \Omega$), and that $\widetilde{\lambda}(u, \emptyset) = 1$. Based on a single observation of a sGpp, denoted X, in a domain Λ_n (a cube aimed at growing up to $+\infty$ as $n \to +\infty$), we show that the estimate $\widehat{\beta}_n(X) := N_{\Lambda_n}(X; R)/V_{\Lambda_n}(X; R)$, where $N_{\Lambda_n}(X; R)$ (resp. $V_{\Lambda_n}(X; R)$) is the number (resp. volume) of data points (resp. points in Λ_n) having no R-close neighbors in X_{Λ_n}, is a consistent estimate of $\beta^\star$ for any $R \geq \widetilde{R}$. Moreover, we prove that $|\Lambda_n|^{1/2}(\widehat{\beta}_n - \beta^\star)$ converges towards a Gaussian distribution as $n \to +\infty$ with explicit variance, variance for which we propose two consistent estimates.

We illustrate the interest and efficiency of the simple estimate we propose in a simulation study. This work may have two further developments:

1. A stationary and isotropic pairwise interaction point process is defined by $-\log \lambda(u, X) = -\log \beta^\star + \sum_{v \in X} g(\|v - u\|)$. Our estimate of $\beta^\star$ is the first step to estimate non parametrically (using *e.g.* a kernel method) the function g.

2. Baddeley and Dereudre (REF) developed recently a promising method based on a variational approach to identify the model $-\log \lambda(u, X) = -\log(\beta^\star) + \theta^t \mathbf{v}(u, X)$. The main interest of their method is that an estimate of θ can be derived without using any optimization procedure. The drawback is that $\beta^\star$ cannot be estimated. The estimate $\widehat{\beta}_n$ we propose could potentially be combined to the estimate $\widehat{\theta}_n$ proposed by REF. we should be able to derive the asymptotic normality of $(\widehat{\beta}_n, \widehat{\theta}_n)$.

The following references (and the references therein) are used in this contribution.

References

[1] A. Baddeley, R.T urner, J. Moller and M. Hazelton. *Residual analysis for spatial point processes. Journal of the Royal Statistical Society (series B), 67:1–35, 2005.*

[2] A. Baddeley, D. Dereudre. Variational estimators for the parameters of Gibbs point process models. Bernoulli journal, 2010

[3] J.F. Coeurjolly and F. Lavancier. Resuduals and goodness-of-fit tests for stationary marked Gibbs point processes. submitted, 2012.

[4] J.F. Coeurjolly and E. Rubak. Fast estimation of covariances for spatial Gibbs point processes. preprint, 2011.

[5] J.F. Coeurjolly, D. Dereudre, R.Drouilhet and F. Lavancier. Takacs-Fiksel method for stationary marked Gibbs point processes. Scandinavian Journal of Statistics, 2012.

Two step estimation for Neyman-Scott point process with inhomogeneous cluster centers

T. Mrkvička[1], M. Muška[2], J. Kubečka[2]

This work has been motivated by ecological studies of spatial distribution of fish population in an inland reservoirs. Outstanding questions that can be addressed are how the fish interact with each other on a small scale and how fish density is influenced by recorded covariates.

We model fish positions by the inhomogeneous Neyman-Scott process. The clusters correspond to fish families or shoals which keep together (Pitcher, 1979). The models of optimal shoal size suggest that they are homogeneous under similar environmental conditions (Bertram, 1978). Therefore sizes of shoals are assumed to be homogeneous in the investigated part of the reservoir, but the occurrence of fish shoals is assumed to be inhomogeneous. Therefore this situation is modeled by inhomogeneous cluster centers.

Since the likelihood-based inference for inhomogeneous point process models is computationally very demanding and not straightforward to implement (Møller & Waagepetersen, 2004, 2007, Waagepetersen, 2007), we focus on two-step estimation methods. This algorithm for inhomogeneous Neyman-Scot process with second order intensity reweighted stationarity (Baddeley et. al., 2000) is described in (Waagepetersen, 2007) and (Waagepetersen & Guan, 2009). The possibilities for two step estimation procedures applied to models with different types of inhomogeneity were further discussed in (Prokešová, 2010).

In the first step, the inhomogenity parameters are estimated by a Poisson log likelihood score function (Schoenberg, 2005). In the second step, the clustering parameters are estimated. Since our model is not second order intensity reweighted stationary, the method described in (Waagepetersen, 2007) can not be applied and we thus introduce four other estimation methods.

The first method is the minimum contrast method, where the contrast is measured on the K-function which is modified to be homogeneous under our model.

The second method maximize the composite likelihood (Guan, 2006).

The third method is based on Bayesian principles (Møller & Waagepetersen, 2004, 2007). The MCMC algorithm is used to sample from the posterior distribution of the clustering parameters. In the algorithm, the process of cluster centers is updated in each step and the clustering parameters are also updated in each step.

Finally, the fourth method uses the Bayesian information criterion (BIC) to choose an appropriate model. The different models are represented by different clustering parameters. The likelihoods of each model, which are used in BIC, are obtained from the MCMC algorithm. In this algorithm, only the process of cluster centers is updated in each step.

[1] Department of Applied Mathematics and Informatics, Faculty of Economics, University of South Bohemia, Studentská 13, 37005 České Budějovice, Czech Republic (mrkvicka@prf.jcu.cz)

[2] Biology center of the AS CR, Institute of Hydrobiology, Na Sádkách 7, 37005 České Budějovice, Czech Republic

The performance of the proposed methods is tested by a simulation study. The most precise method is then applied onto real data derived from fisheries.

Acknowledgments: The work was supported by the Grant Agency of Czech Republic, Projects Nos. P201/10/0472 and 206/07/1392. The access to the MetaCentrum computing facilities, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic, is acknowledged.

References

[1] Baddeley A., Møller J. and Waagepetersen R. P. (2000). Non- and semiparametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, **54**, 329–350.

[2] Bertram, B. C. R. (1978). *Living in groups: predators and prey, in Behavioural Ecology, 1st edn (eds. J.R.Krebs and N.B.Davies).* Blackwell, Oxford, U.K., pp. 64-96.

[3] Guan, Y. (2006). A Composite Likelihood Approach in Fitting Spatial Point Process Models. *Journal of American Statistical Sociaty* **101**, 1502-1512.

[4] Møller, J., Waagepetersen, R. P. (2004). *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall/CRC, London.

[5] Møller, J., Waagepetersen, R. P. (2007). Modern Statistics for Spatial Point Processes. *Scandinavian Journal of Statistics.* **34/4**, 643–684.

[6] Pitcher, T. J. (1979). Sensory information and the organisaton of behaviour of shoaling cyprinid. *Anim. Behav.* **27**, 126–149.

[7] Prokešová, M. (2010). Inhomogeneity in Spatial Point Processes - Geometry Versus Tractable Estimation. *Image Analysis and Stereology.* **29**, 133–141.

[8] Schoenberg, F. P. (2005): Consistent parametric estimation of the intensity of a spatial-temporal point processes. *Journal of Statistical Planning and Inference* **128** (79-93).

[9] Waagepetersen, R. P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics* **63/1**, 252–258.

[10] Waagepetersen, R. P., Guan, Y. (2009). Two-step estimation for inhomogeneous spatial point processes. *Journal of the Royal Statistical Sociaty Series B* **71/3**, 685-702.

Testing of mark independence for marked point patterns by envelope and deviation tests

Pavel Grabarnik[a], Mari Myllymäki[b], and Dietrich Stoyan[c]

[a]*Institute of Physico-Chemical and Biological Problems in Soil Science, the Russian Academy of Sciences, Pushchino, 142290 Moscow Region, Russia*
[b]*Department of Biomedical Engineering and Computational Science, Aalto University, P.O. Box 12200, FI-00076 Aalto, Finland*
[c]*Institut für Stochastik, TU Bergakademie Freiberg, D-09596 Freiberg, Germany*

One of the key issues in the statistical application of marked point processes is the question of spatial correlations of the marks. Thus, mark independence tests play a fundamental role in spatial statistics and modeling and have been thoroughly considered in the statistical literature, see e.g. Diggle (2003), Illian et al. (2008). Nevertheless, some popular established tests are not fully satisfactory and need improvement. The present talk tries to contribute to this issue. It is inspired by Loosmore and Ford (2006), who considered goodness-of-fit tests for non-marked point patterns and strongly recommended the use of deviation tests instead of the popular envelope tests, which may lead to unreasonably high type I error probabilities.

Both deviation and simulation envelope tests are Monte Carlo significance tests (Besag and Diggle, 1977) based on some summary function $F(r)$, where r in the context of the applications considered in this talk denotes distance. It is a common situation in spatial statistics that the distribution of $F(r)$ is unknown, and the use of Monte Carlo simulations is the only way to test hypotheses.

A deviation test summarizes information on $F(r)$ in a *single* number and compares it with some reference value, obtained from simulations of the model corresponding to the null hypothesis. This Monte Carlo test (Barnard, 1963) is based on the rank of a test statistic and provides the exact type I error probability. Here “exact” means that the null hypothesis is declared as false, when it is true, precisely with the prescribed probability. In contrast, in an envelope test the values of $F(r)$ are inspected for a range of distances simultaneously. Thus, the statistician enters the field of the “multiple testing problem”, and the determination of the type I error probability is difficult. Already Ripley (1977), who introduced envelope tests, mentioned that the frequency of committing the type I error in the goodness-of-fit test based on range of distances may be higher than for the single-distance test.

If $F(r)$ were of interest only for a single distance, one could proceed as in the deviation test. However, "single-distance" tests are rarely applied, since prior knowledge of a single "interesting" distance r is untypical in practice.

Following the approach of Diggle (1979, 2003), Loosmore and Ford (2006) adopted the deviation test, which was introduced as an alternative to the envelope test, since it gives an easy way to adjust the testing for multiple scales. In order to demonstrate the difficulties of the simulation envelope test, Loosmore and Ford (2006) estimated the type I error probability by simulation for the complete spatial randomness (CSR) hypothesis based on the nearest neighbor distance distribution function $G(r)$. They concluded that it may be dangerous to make inference using conventional envelope-based tests because the probability that the test will reject the null hypothesis may be too high. While Loosmore and Ford (2006) considered the non-marked case, the present talk treats the marked case where the marks can be quantitative (continuous) as well as qualitative (discrete). In contrast to Loosmore and Ford (2006) who did not recommend the envelope test for inference, we show that it can be used as well as the deviation test when it is coupled with controlling the type I error.

The talk considers both envelope and deviation tests. It demonstrates the weakness of the envelope method and shows how it can be refined to provide a correct statistical test. Further, deviation tests are described for mark-independence hypotheses, and advantages and disadvantages of the tests are discussed. The methods are illustrated by examples from forest ecology.

REFERENCES

Barnard, G.A., 1963. Discussion of paper by M.S. Bartlett. J. R. Stat. Soc. B25, 294.

Besag, J., Diggle,P. J., 1977. Simple Monte Carlo tests for spatial pattern. Appl. Stat. 26, 327–333.

Diggle, P. J. 1979. On parameter estimation and goodness-of-fit testing for spatial point patterns. Biometrics 35, 87–101.

Diggle, P. J. 2003. Statistical Analysis of Spatial Point Patterns. 2nd Edition. Arnold, London.

Grabarnik, P., Myllymäki, M. and Stoyan, D., 2011. Correct testing of mark independence for marked point patterns. Ecological Modelling 222, 3888–3894.

Illian, J., Penttinen, A., Stoyan, H., Stoyan, D., 2008. Statistical Analysis and Modelling of Spatial Point Patterns. Wiley, Chichester.

Loosmore, N.B., Ford, E.D., 2006. Statistical inference using the G or K point pattern spatial statistics. Ecology 87, 1925–1931.

Ripley, B.D., 1977. Modelling of spatial patterns. J. R. Stat. Soc. B39, 172–192.

Unbiased approximate pseudo-likelihood for spatial point processes

Rasmus Waagepetersen
Department of Mathematical Sciences
Fredrik Bajersvej 7G
DK-9220 Aalborg
rw@math.aau.dk

For spatial Gibbs and Markov point processes, popular options for parameter estimation include maximum likelihood, maksimum pseudo-likelihood and Takacs-Fiksel estimation. However, in practice approximate versions of these estimation methods are always used. The pseudo-likelihood e.g. contains a two or three-dimensional integral which is approximated using numerical quadrature.

The R-package spatstat implements a particular approximation of pseudo-likelihood using the Berman-Turner device which allows the approximate pseudo-likelihood to be maximized using standard procedures for generalized linear models (GLMs). The implementation in spatstat is very flexible and computationally efficient and makes statistical inference for complex spatial Markov point process models feasible also for users who are not experts in spatial statistics.

When an unbiased estimating function is approximated using numerical quadrature the approximate estimating function is typically not unbiased and the resulting bias in the parameter estimates is difficult to quantify. In this paper we suggest an unbiased Monte Carlo approximation of the pseudo-likelihood estimating function. Our approach has several advantages. First, the resulting estimates are unbiased. Second, the unbiased approximate pseudo-likelihood estimating function takes the form of a logistic regression score and can thus like the estimating function in spatstat easily be fitted using existing software for GLMs. Third, the variance of the parameter estimates can straightforwardly be decomposed into the sum of the variance of the (exact) pseudo-likelihood estimate and the additional variance due to the Monte Carlo approximation. By the third property the user can establish how large a Monte Carlo sample is needed in order to achieve a certain level of accuracy.

The presentation is based on joint work with Adrian Baddeley, Jean-Francois Coeurjolly and Ege Rubak.

Modeling group dispersal of particles with a spatiotemporal point process

Samuel Soubeyrand
Biostatistics and Spatial Processes, INRA Avignon

A stochastic model describing the dispersal of group of particles will be built and its properties will be analyzed. This model can be viewed like a generalization of propagation models based on explicit dispersal kernels and used in ecology and epidemiology. It can also be viewed like a doubly stochastic spatiotemporal point process (spatiotemporal Cox process). However, because of the application field of interest, namely population dynamics, the properties to be determined may be different from those generally studied in stochastic geometry. Thus, in addition to study the ability of the model to generate clusters (secondary foci), we will also calculate the probability distribution of the furthest point and describe the resulting consequences on the invasion speed of the population modeled by the points.

Comparisons of discriminant analysis techniques for correlated data

Line H. Clemmensen

Informatics and Mathematical Modelling
Technical University of Denmark
Richard Petersens Plads 305/123
Lyngby, Denmark

I will compare a range of newly proposed techniques for performing discriminant analysis on high-dimensional data. In particular, the techniques differ in performance depending on the correlation structures present in data. Highly correlated data is for example common in image analysis where pixels values are closely related or in spectral or temporal data. The differences in the techniques are not the various optimization criteria or variable selection techniques, which often are emphasized in the original papers, but the choice of estimate of the within-class covariance matrix. All of the methods build on linear discriminant analysis (Fischer, 1936), and to further cope with the high-dimensionality of the data, they introduce sparseness in the feature space and/or regularization of the within-class covariance matrix.

The two methods: nearest shrunken centroids (Tibshirani et al., 2003) and penalized linear discriminant analysis (Witten and Tibshirani, 2011), assume independence between the variables by using a diagonal estimate of the within-class covariance matrix. The three methods: regularized discriminant analysis (Guo et al., 2007), sparse discriminant analysis (Clemmensen et al., 2011), and sparse linear discriminant analysis (Shao et al., 2011), are able to estimate the correlation structures in the off-diagonal of the within-class covariance matrix. Performances are illustrated on simulated data.

References

- L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll, Sparse discriminant analysis, Technometrics, **53**(4): 406-413 (2011).
- R. Fisher, The Use of Multiple Measurements in Axonomic Problems, Annals of Eugenics **7**: 179-188 (2007).
- Y. Guo, T. Hastie, and R. Tibshirani, Regularized linear discriminant analysis and its applications in microarrays, Biostatistics **8**: 86-100 (2007).
- J. Shao, G. Wang, X. Deng, and S. Wang, Sparse linear discriminant analysis by thresholding for high dimensional data, The Annals of Statistics **39**: 1241-1265 (2011).
- R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, Class prediction by nearest shrunken centroids, with applications to DNA microarrays, Statistical Science **18**: 104-11 (2003).
- D. Witten and R. Tibshirani, Penalized classifiation using Fisher's linear discriminant, Journal of the Royal Statistical Society, Series B **73**(5): 753-772 (2011).

On the simple and partial Mantel tests in presence of spatial auto-correlation

Gilles Guillot
Technical University of Denmark, Department of Informatics

For the detection of clustering of cancer cases in space and time, Mantel (1967) introduced a test based on permutations. He concluded his article by claiming that this method was general - a claim later relayed by Sokale (1979) - and could be used whenever one has to assess the significance of the correlation between the values of two square matrices containing distances relative to pairs of individuals. Smouse (1986) proposed an extension of the test, referred to as partial Mantel test, and aimed at assessing the dependence between two matrices of distances while "controlling" the effect of a third distance matrix. Since then, and despite the fact that (or perhaps because) none of these four original methodological papers stated the null hypothesis explicitly, the simple and partial Mantel tests have known a tremendous popularity.

The simple Mantel test is for example used routinely to assess the significance of the association between a matrix of genetic measurements and a matrix of phenotypic measurements relative to the same individuals. It is also intensively used in ecology to assess how a matrix of genetic or phenotypic distances relates to a matrix of geographical distances. The latter may contain plain geographical (Euclidean) distances between pairs of sampling sites but it may alternatively contain values that attempt to reflect the actual cost for an individual to move across the area (accounting e.g. for the presence of barriers or hostile areas). In the latter case, the distance is known in ecology as "cost distance". It may not enjoy the properties of a mathematical distance but it is in general correlated with the Euclidean distance. A classical analysis consists in assessing the significance of the dependence between genetic (or phenotypic) distances and cost distances while controlling for the "effect" of geographical distances through the partial Mantel test.

In view of the tasks above, the Mantel tests have a number of appealing features. First they allow one to synthesize information contained in multivariate data in a single index and hence in a single test ; second they allow one to deal with the case outlined above where the "distance" between individuals cannot be expressed as a difference (or combination of differences) between one or several variables (e.g. case of a cost distance) ; finally, they do not seem to rely on any parametric assumption.

The aim for this talk will be to illustrate that the Mantel tests are not valid statistical procedure as soon as the data display spatial auto-correlation. Alternative strategies will be also discussed.

Planar Markov fields

M.N.M. van Lieshout
CWI, Amsterdam, The Netherlands

We introduce a class of Gibbs-Markov random fields built on regular tessellations that can be understood as discrete counterparts of Arak-Surgailis polygonal fields and generalise the bivariate fields of Schreiber and Van Lieshout (2010). We focus first on consistent polygonal fields, for which we show consistency, Markovianity, and solvability by means of dynamic representations. Next, we develop simulation dynamics for their general Gibbsian modifications, which cover most lattice-based Gibbs-Markov random fields.

Inference of within cell protein interactions and spatial structure using Fluorescence Resonance Energy Transfer microscopy

Jan-Otto Hooghoudt
Dpt. of Mathematics, Aalborg University, Denmark

Fluorescence resonance energy transfer (FRET) microscopy has become one of the preferred tools to obtain information concerning the distribution of proteins throughout living cells. FRET is an electrodynamic phenomenon that can be explained using classical physics. FRET occurs between a donor molecule in the excited state and an acceptor molecule in the ground state. The main parameter describing FRET is the FRET efficiency. It is defined as the fraction of photon energy absorbed by donors that is transfered to acceptors.

Although the interactions between donors and acceptors are on the molecular level (1-10 nm), the pixel resolution of FRET microscopy is typical of the order of 100x100 nm^2. This implies that in general a large number of proteins are confined within one pixel and therefor no direct information concerning the spatial distribution of the proteins on an inner pixel level can be inferred. Previous studies have made some progress by qualitatively state whether in some micro-domains proteins are distributed in clusters or randomly. However, no information is available concerning typical cluster sizes, number of clusters and type of clustering.

Zimet et al. (1995) suggested stochastic models for the physical process behind FRET and Berney and Danuser (2003) demonstrated the validity of these models by showing that FRET data simulated using these models agree well with experimental FRET data. An extensive quantitative survey, has been carried out by Corry et al. (2005).

In order to obtain a complete stochastic model for FRET data we combine the FRET efficiency model with stochastic point process models for the underlying protein configurations (e.g. multi-Strauss hardcore model). Due to the complicated nature of the stochastic models involved the corresponding likelihood function is intractable.

In this talk I will give a short introduction to FRET, show results obtained from simulations and discuss approaches to do statistical inference on experimentally obtained FRET datasets in order to estimate parameter values for the supposed underlying theoretical point process model.

Detecting fake paintings

Robert Jacobsen
Aalborg University, Department of Mathematical Sciences

It was probably not long after people started buying paintings that a business in forging paintings was born which initiated the need for detecting forgeries. This task of determining if a painting is indeed painted by the claimed artist, i.e. is authentic, is called authentication.

Authentication has traditionally been performed by connoisseurs which employ a wide range of tools to establish physical properties of the painting, e.g. the age of the materials and the type of paint used. Furthermore, visual characterisations, e.g. determining if the sweeps of the brush resembles those in other works of the painter, are sought to see if the painting "fits" into the style of the painter. In many cases the brushstrokes/lines in a painting/drawing are believed to be characteristic of an artist, like an artistic signature.

The judgment of the visual characteristics are mainly subjective, as they are decided by a few experts who are steeped in the style of the concerned artist. It is therefore of interest to aid the traditional authentication of paintings by unbiased and objective analyses. The topic of this presentation is to present recent results in the field of authentication based on analysing digital reproductions of paintings by mathematical methods.

Using tools from harmonic analysis we can extract high frequency details from digital reproductions that includes lines and brushstrokes. By modelling such details appropriately we are able to distinguish authentic images from forgeries.

Simulating the tail of the interference in a Poisson network model

Giovanni Luca Torrisi* and Emilio Leonardi†

1 Extended Abstract

1.1 Introduction

Mutual interference among simultaneous transmissions constitutes the main limitation factor to the performance of dense wireless networks, severely reducing the capacity of the whole system (see [14], [16], [17], [19] and [20].)

The availability of efficient analytical/numerical techniques to tightly characterize the interference produced by transmitting nodes operating over the same channel is a key ingredient to better predict performance of such complex systems as well as to design new Medium Access Control (MAC) protocols and more advanced transmission schemes that better exploit the system bandwidth. Just as matter of example, we shall explain how the tail of the interference is directly related to the probability that the communication does not succeed, in the case when a single input/single output transmission scheme is adopted.

In this talk we shall consider a simple wireless network setting in which nodes are placed according to a Poisson process on the plane and employ a simple ALOHA MAC protocol (see [1], [3], [4], [5], [6], [8], [9], [10] and [15]). We propose a provably efficient numerical methodology to estimate the tail of the interference, under natural assumptions on fading and attenuation. If the tail of the interference is not too small, one may exploit a crude Monte Carlo approach to evaluate the complementary of the cumulative distribution function of the interference. However, when the tail of the interference is small the crude Monte Carlo method becomes inefficient, and different numerical techniques are needed. The methodology we shall use is based on (state-dependent) importance sampling (see e.g. [2] and [7].) Despite of the fact that a significant body of work has attempted a characterization of the interference in large wireless networks (see [1], [3], [4], [5], [6], [8], [9], [10], [11] and [15]), we are not aware of previous work proposing provably efficient numerical algorithms to estimate the tail of the interference, assuming that the fading has a light-tail distribution and the attenuation decays sub-exponentially with the distance. Actually, most of the existing literature on the subject focuses on analytical characterizations of either the interference distribution or the outage probability, under specific assumptions on fading and attenuation. For instance, if the attenuation is of the form $\|x\|^{-\alpha}$, $x \in \mathbb{R}^2$, $\alpha > 2$, where the symbol $\| \cdot \|$ denotes the Euclidean norm, and the fading is constant (i.e. there is a purely geometric attenuation) or distributed according to a Rayleigh law, closed form expressions for the Laplace transform of the interference are derived e.g. in [1], [4] and [15]. However, only in exceptional cases the Laplace transform may be inverted to obtain the law of the interference. This is possible, for instance, if $\alpha = 4$ and the system is subjected to a purely geometric attenuation [11]. Under more general assumptions on fading and attenuation, explicit bounds on the tail of the interference may be found in [11]. In [10] a large deviations approach is employed to study the asymptotic behavior

*Istituto per le Applicazioni del Calcolo "Mauro Picone", CNR, Via dei Taurini 19, I-00185 Roma, Italia. e-mail: torrisi@iac.rm.cnr.it

†Dipartimento di Elettronica, Politecnico di Torino, Corso Duca degli Abruzzi 24, I-10129 Torino, Italia. e-mail: leonardi@polito.it

of the logarithm of the tail of the interference, for a quite general fading (possibly heavy-tail) and ideal Hertzian propagation, i.e. of the form $\max(R, \|x\|)^{-\alpha}$, $R > 0$, $\alpha > 2$. The results in [10] constitute the starting point to build provably efficient numerical algorithms to estimate the tail of the interference.

Under general assumptions on the nodes distribution, the fading distribution and the attenuation function, asymptotic estimates for the outage probability, as the intensity of the nodes goes to zero, are derived in [12] and [13]. Finally, a Monte Carlo algorithm to estimate the density of the interference for a quite general wireless network model has been proposed in [18].

The methodology proposed in this talk complements the previously mentioned results, providing an efficient and accurate Monte Carlo algorithm to compute the tail of the interference in cases where the analytical approach is not feasible. We believe that the proposed methodology may yield hints for a successive development of Monte Carlo procedures that allow fast and accurate evaluations of the tail of the interference when the transmitting nodes are distributed according to more general point processes models.

1.2 The system model and organization of the talk

We consider the following simple model of multi-hop wireless network, which accounts for interference effects that arise when several nodes transmit at the same time.

Suppose that transmitting nodes (antennas) are located according to a Poisson process $\{X_k\}_{k\geq 1}$ on the plane with a locally integrable intensity function $\lambda(x)$, $x \in \mathbb{R}^2$, i.e. X_n is the location of node n. Denote by $P_n \in (0, \infty)$ the transmission power of node n. Assume that a new receiver is added at the origin and that a new transmitter is added at $x \in \mathbb{R}^2$.

Let w be a positive constant which describes the thermal noise at the receiver, and suppose that the physical propagation of the signal is described by a measurable positive function $L : \mathbb{R}^2 \to (0, \infty)$, which gives the attenuation or path-loss of the signal. In addition, the signal undergoes random fading (due to occluding objects, reflections, multi-path interference, etc.). We denote by H_n the random fading between node n and the receiver, and define $Y_n := P_n H_n$. Thus $Y_n L(X_n)$ is the received power at the origin due to node n. Similarly, we denote by $YL(x)$, the received power at the origin due to the transmitter at x. We assume that $\{Y, \{Y_k\}_{k\geq 1}\}$ is a sequence of independent and identically distributed (i.i.d.) random variables (r.v.'s), independent of locations. In the following (with an abuse of terminology) we shall call the r.v.'s Y_k signals.

In this talk we shall provide a computationally efficient (state-dependent) importance sampling algorithm for the characterization of the total interference at the origin, which is given by the Poisson shot noise r.v. $V := \sum_{k\geq 1} Y_k L(X_k)$. We emphasize that a tight characterization of the tail of the interference $\psi(\beta) := P(V > \beta)$ is needed to predict the performance of large scale wireless networks. In particular, the tail of the interference is related to the probability of successfully decoding the signal from the transmitter at x. Indeed, given the adopted modulation and encoding scheme, we can claim that the receiver at the origin can successfully decode the signal from the transmitter at x if the Signal to Interference plus Noise Ratio (SINR) is greater than a given threshold, say $\tau > 0$ (which depends on the adopted scheme), i.e.

$$\frac{YL(x)}{w + V} \geq \tau, \quad \text{almost surely (a.s..)}$$

So, conditional to the event $\{Y = y\}$, the probability that the communication succeeds is given by

$$P\left(\frac{YL(x)}{w + V} \geq \tau \,\Big|\, Y = y\right) = P\left(\frac{yL(x)}{w + V} \geq \tau\right) = P(V \leq yL(x)\tau^{-1} - w). \tag{1}$$

The attenuation function is often taken to be isotropic (i.e. rotation invariant) and of the form $L(x) = \ell(\|x\|) = \|x\|^{-\alpha}$ or $(1 + \|x\|)^{-\alpha}$ or $\max(R, \|x\|)^{-\alpha}$, where $\alpha > 2$ and $R > 0$ are positive constants. Setting $\tau = \theta\tau'$ in (1), where $\theta > 0$ and $\tau' > 0$ are two positive constants, we have

$$P\left(\frac{YL(x)}{\tau'(w+V)} \geq \theta \,\middle|\, Y = y\right) = P\left(\frac{yL(x)}{\tau'(w+V)} \geq \theta\right) = P\left(V \leq \frac{yL(x)}{\theta}\tau'^{-1} - w\right).$$

The high-reliability regime corresponds to the high-SINR regime, i.e. the regime where $\tau' \to 0$ (see [12] and [13] for the analysis of the high-SINR regime as the intensity of the nodes goes to zero.) Thus, for large values of β, the probability $\psi(\beta)$ is also related to the outage probability in the high-SINR regime.

Note that whenever $V < \infty$ a.s. (a sufficient condition for this is e.g. $\mathrm{E}[V] < \infty$, i.e. $\mathrm{E}[Y_1] < \infty$ and $\int_{\mathbb{R}^2} L(x)\lambda(x)\,\mathrm{d}x < \infty$), $\psi(\beta) \to 0$, as $\beta \to +\infty$, so the event $\{V > \beta\}$ is rare as β increases, and this rises questions about the numerical estimation of the small probabilities $\psi(\beta)$ via a Monte Carlo algorithm. The importance sampling technique which will be proposed in this talk can be successfully used to obtain accurate estimates of $\psi(\beta)$ for values of β that correspond to small $\psi(\beta)$ (note that such values of β may be moderately large.) This permits to unveil how different system's parameters, such as the intensity of the nodes, the path-loss exponent and the fading distribution, impact on the system performance. For these reasons, we believe that our approach is complementary with respect to the previously proposed analytical approaches that capture either the asymptotic behavior, as $\beta \to \infty$, of the tail of the interference ([10]) or the asymptotic behavior, as the intensity of the nodes goes to zero, of the outage probability ([12, 13].)

The talk will be structured as follows: (*i*) we shall describe networks with nodes distributed according to a stationary Poisson process on $\mathbb{R}^2$ with intensity function $\lambda(\cdot)$ and attenuation function of the form $L(x) = \ell(\|x\|) = \max(R, \|x\|)^{-\alpha}$, $\alpha > 2$, $R > 0$; (*ii*) we shall describe the importance sampling methodology in this context; (*iii*) we shall provide asymptotically admissible simulation laws for $\psi(\beta)$, as $\beta \to +\infty$, under a quite general light tail assumption on the distribution of the signals; (*iv*) we shall give asymptotically efficient simulation laws for $\psi(\beta)$, as $\beta \to +\infty$, when the signals are bounded, Weibull super-exponential or Exponential; (*v*) we shall discuss some numerical illustrations.

References

[1] Ali, O.B., Cardinal, C. and Gagnon, F. (2010). Performance of optimum combining in a Poisson field of interferes and Rayleigh fading channels. *IEEE Transactions on Wireless Communications* **9**, 2461 –2467.

[2] Asmussen, S. and Glynn, P. (2007). *Stochastic Simulation*, Springer, New York.

[3] Baccelli, F. and Błaszczyszyn, B. (2001). On a coverage process ranging from the Boolean model to the Poisson-Voronoi tessellation with applications to wireless communications. *Advances in Applied Probability* **33**, 293–323.

[4] Baccelli, F. and Blaszczyszyn, B. (2009). *Stochastic Geometry and Wireless Networks*, Foundations and Trends in Networking, NoW Publishers, Part I: Theory.

[5] Baccelli, F. and Blaszczyszyn, B. (2009). *Stochastic Geometry and Wireless Networks*, Foundations and Trends in Networking, NoW Publishers, Part 2: Applications.

[6] Baccelli, F., Blaszczyszyn., B. and Muhlethaler, P. (2006). An Aloha protocol for multihop mobile wireless networks. *IEEE Transactions on Information Theory* **52**, 421–436.

[7] Bucklew, J.A. (2004). *Introduction to Rare Event Simulation*, Springer, New York.

[8] Dousse, O., Baccelli, F. and Thiran, P. (2005). Impact of interferences on connectivity in ad hoc networks. *IEEE/ACM Transaction on Networking* **13**, 425–436.

[9] Dousse, O., Franceschetti, M., Macris, N., Meester, R. and Thiran, P. (2006). Percolation in the signal to interference ratio graph. *Journal of Applied Probability* **43**, 552–562.

[10] Ganesh, A. and Torrisi, G.L. (2008). Large deviations of the interference in a wireless communication model. *IEEE Transactions on Information Theory* **54**, 3505–3517.

[11] Ganti, R.K. and Haenggi, M. (2009). Interference and outage in clustered wireless ad hoc networks. *IEEE Transactions on Information Theory* **55**, 4067–4086.

[12] Ganti, R.K., Andrews, J.G. and Haenggi, M. (2011). High-SIR transmission capacity of wireless networks with general fading and node distribution. *IEEE Transactions on Information Theory* **57**, 3100–3116.

[13] Giacomelli, R., Ganti, R.K. and Haenggi, M. (2011). Outage probability of general ad hoc networks in the high-reliability regime. *IEEE/ACM Transactions on Networking* **19**, 1151–1163.

[14] Gupta, P. and Kumar, P.R. (2000). The capacity of wireless networks. *IEEE Transactions on Information Theory* **46**, 388–404.

[15] Haenggi, M. and Ganti, R. K., (2008) *Interference in Large Wireless Networks*, Foundations and Trends in Networking, NoW Publishers.

[16] Hunter, A. Andrews, J. and Weber, S. (2008). Transmission capacity of ad hoc networks with spatial diversity. *IEEE Transactions on Wireless Communications* **7**, 5058–5071.

[17] Ozgur, A., Leveque, O. and Tse, D. (2002). Hierarchical cooperation achieves optimal capacity scaling in ad hoc networks. *IEEE Transactions on Information Theory* **53**, 3549–3572.

[18] Privault, N. and Torrisi, G.L. (2011). Density estimation of functionals of spatial point processes with application to wireless networks. *SIAM Journal on Mathematical Analysis*, **43**, 1311-1344.

[19] Weber, S., Andrews, J.G. and Jindal, N. (2007). The effect of fading, channel inversion, and threshold scheduling on ad hoc networks. *IEEE Transactions on Information Theory* **53**, 4127–4149.

[20] Weber, S., Yang, X., Andrews, J.G. and de Veciana, G. (2005). Transmission capacity of wireless ad hoc networks with outage constraints. *IEEE Transactions on Information Theory* **51**, 4091–4102.

Functional Median Polish, with Climate Applications

Marc G. Genton
Department of Statistics, Texas A&M University

We propose functional median polish, an extension of univariate median polish, for one-way and two-way functional analysis of variance (ANOVA). The functional median polish estimates the functional grand effect and functional main factor effects based on functional medians in an additive functional ANOVA model assuming no interaction among factors. A functional rank test is used to assess whether the functional main factor effects are significant. The robustness of the functional median polish is demonstrated by comparing its performance with the traditional functional ANOVA fitted by means under different outlier models in simulation studies. The functional median polish is illustrated on various applications in climate science, including one-way and two-way ANOVA when functional data are either curves or images. Specifically, Canadian temperature data, U.S. precipitation observations and outputs of global and regional climate models are considered. This is based on joint work with Ying Sun.

Inferring epidemiological parameters of space-time dynamics of apple scab in orchards

Senoussi R.[1], Parisi L.[2] and Gros C.[2]
[1]BIOSP, Inra Avignon. [2]UERI, Inra, Gotheron

Summary

Apple is a major fruit crop worldwide and scab, caused by *Venturia inaequalis*, is its most common fungal disease. Its control is based on intensive use of fungicide sprays. Pertinent mixtures of resistant and susceptible cultivars could help to reduce fungicide use.

In 9 experimental apple orchards in Gotheron (South of France), 3 of pure (susceptible) and 6 of mixed (susceptible + resistant), leaf infections were observed at random times during nearly 2 months after an initial artificial inoculation was performed. We consider here a time-space model in vegetal epidemiology for fungus dispersal within the two types of orchards. Fungus epidemics depend on many factors and particularly on climatic conditions and spatial characteristics. Climatic conditions, mainly temperature and humidity, can be viewed as the effective variables controlling the specific time schedule of fungus dynamics. Similarly, spatial components must be given taking into account the field characteristics ((void space, vegetal screens, etc…).

To describe the *Venturia inaequalis* epidemics, we first discretize space into square regular cells allotted with specific characteristics whereas time is divided into climatic units. We, then, proposed a time-Markovian and multidimensional (for space cells) process for statistical inference. More specifically, an orchard is considered as set of **N** cells $\mathbf{C_i}$ centered on $\mathbf{c_i=(x_i , y_i)}$ and characterized by epidemic index (**V**=void, **S**=susceptible region, **R**=resistant region). Conditional distributions simply assert that if a fungus spot located at $\mathbf{c_i}$ with intensity $\mathbf{I_i}$ is present at a proper "time" t, it supposedly delivers during an epidemiological time unit a Poisson number of spores on a site $\mathbf{C_j}$ with a mean defined via a dispersal kernel.
After an incubation period, spores give rise to new fungus spots that disperse additively with the same rules and independently of each others, etc…

Considering first the space time variables, we considered the intrinsic observation times which were irregular and random as Markov stopping times defined by climatic conditions. Then, we defined for fungus dispersal, the epidemiological spatial distance between two points as the sum of the weighted lengths of the segments delimited by the different crossed cells. We then adopted a classical parametric model for statistical inference. We introduced 1 parameter $\boldsymbol{\alpha_0}$ for a uniform base intensity for leaf infection, 2 parameters ($\boldsymbol{\alpha_S}$ and $\boldsymbol{\alpha_R}$) to describe the cell characteristics, ie the resistivity to displacement ($\boldsymbol{\alpha_V}$=1 being the reference) and the epidemiological distance, 1 parameter for the dispersal range of a spot and 1 parameter $\boldsymbol{\alpha_I}$ for the dispersal intensity which roughly describes the mean number of effective spores attaining a cell at one distance unit, and finally 1 parameter $\boldsymbol{\alpha_T}$ for time/climate unit. This set of parameters makes it possible to test some sensible epidemiologic hypotheses such as the pertinence of alternating resistant and susceptible cultivars in orchards and to estimate quantitatively the effect of climatic variates.

Identification of local multivariate outliers

Peter Filzmoser[1], Anne Ruiz-Gazen[2] and Christine Thomas-Agnan[3]

Multivariate outlier detection belongs to the most important tasks for the statistical analysis of multivariate data. Multivariate outliers behave differently than the majority of observations which are assumed to follow some underlying model like a multivariate normal distribution. The deviations of outlying observations from the majority of data points can also be understood in an exploratory context, for example by visualizing a measure describing outlyingness and inspecting possible deviations or gaps in the resulting plot.

The most commonly used measure of outlyingness is the Mahalanobis distance. This multivariate distance measure assigns each observation a distance to the center, taking into account the multivariate covariance structure. Practically, for obtaining a reliable distance measure for multivariate data it is crucial to estimate robustly the center and the covariance from the data. A frequently used robust estimator of multivariate location and scatter is the Minimum Covariance Determinant estimator which looks for a subset of observations with smallest determinant of the sample covariance matrix.

The distance measure for multivariate outlier detection does not account for any spatial dependence among the observations. Moreover, it is limited to identify overall, "global" outliers that differ from the main bulk of the data, but not outliers in a local neighborhood. Interestingly, spatial or "local" outliers are most often also outlying according to the spatial dependence. Usually, it turns out that spatial data sets contain positive spatial autocorrelation which means that observations with high (respectively low) values for an attribute are surrounded by neighbors which are also associated with high (respectively low) values. Thus, in a positive autocorrelation scheme, observations that differ from their neighbors do not follow the same process of spatial dependence as the main bulk of the data. Graphics such as the variogram cloud and the Moran scatterplot are interesting tools for detecting local outliers in a univariate framework. However, up to our knowledge very few proposals have been made in the multivariate context.

The main objective of the present paper is to introduce new exploratory tools in order to detect outliers in multivariate spatial data sets. Our purpose is also to illustrate that if global outliers are present in the data set, they are usually also local outliers and they can completely mask other local outliers. The exploratory tools we introduce do not only detect both kinds of outliers but also give an insight into their global or/and local nature.

[1] Vienna University of Technology, Department of Statistics and Probability Theory. Email: P.Filzmoser@tuwien.ac.at

[2] Toulouse School of Economics, France. Email: anne.ruiz-gazen@tse-fr.eu

[3] Toulouse School of Economics, France. Email:christine.thomas@tse-fr.eu

Combining probabilities with log-linear pooling : application to spatial data

Denis Allard[1], Alessandro Comunian[2,3], Philippe Renard[3], Dimitri D'Or[4]

[1] *INRA, UR 546 Biostatistique et Processus Spatiaux (BioSP), Site Agroparc, 84914 Avignon, France.* allard@avignon.inra.fr
[2] *National Centre for Groundwater Research and Training (NCGST), The University of New South Wales, Sydney, Australia.* a.comunian@unsw.edu.au
[3] *CHYN, Université de Neuchâtel, Neuchâtel, Suisse.* Philippe.Renard@unine.ch
[4] *Ephesia-Consult, Brussels, Belgium.* dimitri.dor@ephesia-consult.com

The need of combining in a probabilistic framework different sources of information is a frequent task in management, environmental and earth sciences and spatial statistics. The problem of aggregating these different conditional probability distributions into a single conditional distribution arises as an approximation to the inaccessible genuine conditional probability given all information, since building a full probabilistic model is in general impossible. This paper makes a formal review of most aggregation methods proposed in the literature with a particular focus on their mathematical properties. Calibration fo the agregated probability distribution is of particular importance. It is known that linear agregation operators are not calibrated. Here, we show that the log-linear agregation operator with parameters estimated from maximum likelihood is calibrated. An application with spatial data illustrate the performance of these operators.

Bibliography

[1] Allard, D., Comunian, A. & Renard, P. (2012) Probability aggregation methods in geoscience *Mathematical Geoscience* (In press).
[2] Bordley, R. F. (1982). A multiplicative formula for aggregating probability assessments. *Management Science*, 28(10) :1137–1148.
[3] Clemen, R. T. and Winkler, R. L. (2007). Aggregating probability distributions. In *Advances in Decision Analysis.* Edwards, W and Miles, R.F. and von Winterfeldt, D.
[4] Genest, C. and Zidek, J. V. (1986). Combining probability distributions : A critique and an annotated bibliography. *Statistical Science*, 1(1) :114–148.
[5] Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72(1) :71–91.

Some contributions for second order scalar or vector valued random fields

Emilio Porcu

Universidad Castilla la Mancha, Department of Statistics

Scalar and Vector–valued random fields (RFs) have received an increasing interest in the last thrty years from several scientific areas. When the finite dimensional distribution of the RF is Gaussian, then only second order properties matter for statistical inference. In particular, weakly and intrinsically stationary Gaussian RF's properties are specified through the properties of their associated (respectively) covariance or variogram functions. In this talk we shall present a personal selection of contributions from the last four years or research. In particular, for space–time scalar valued RFs, we show (a) that Gneiting's (2002) conditions are not only sufficient, but also necessary; (b) we relax the hypothesis on the involved functions and (c) we give necessary conditions when the generator of the space–time covariance is compactly supported on the unit sphere. For the case of scalar–valued RFs, we show that some classes of variograms are closed under product, completing a part of the literature built, for over 40 years, that variograms do not preserve permissibility under product. For vector–valued RFs, we show some important criteria of construction that are proved to be useful for implementing new matrix–valued covariance and correlation functions. We also find a class of functions that is compactly supported and another that allows for hole effects.