Spatial statistics and image analysis. Lecture 7

Mats Rudemo

April 27, 2020

Todays lecture will cover

Support vector machines

Markov random field models

Point processes

Support vector machines

Support vector machines can be thought of as generalizations of linear discrimination.

Suppose we have two classes ω_1 and ω_2 and explaning vector x

In linear discrimination choose class ω_1 if $f(x) > 0$ where

$$f(x) = (\mu_1 - \mu_2)^T C^{-1}(x - \frac{1}{2}(\mu_1 - \mu_2)) - \frac{\pi_2}{\pi_1}) \tag{1}$$

More generally, choose class ω_1 if $f(x) > 0$ with

$$f(x) = \beta_0 + x^T \beta \tag{2}$$

How choose β_0 and β ?

Suggestion in Figure 1: maximize minimal distance to separating hyperplane.

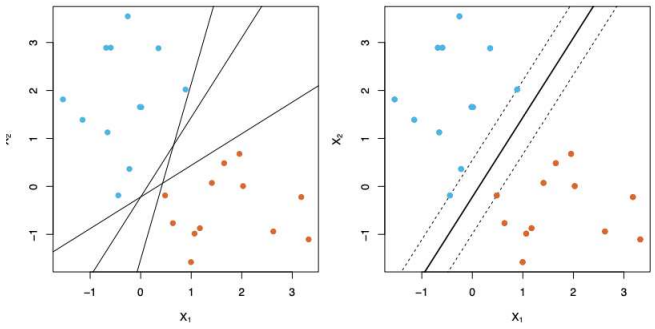

Figure 19.1 Left panel: data in two classes in $\mathbb{R}^2$. Three potential decision boundaries are shown; each separate the data perfectly. Right panel: the optimal separating hyperplane (a line in $\mathbb{R}^2$) creates the biggest margin between the two classes.

Figure 1: Decision boundaries for data with perfect separation, left panel: three potential separating lines, right panel optimal separating hyperplane (line in 2D), from Efron and Hastie (2016)

In general no hyperplane with complete separation

Instead find a minimum with a regularized loss function

$$\min_{\beta_0,\beta} \left\{ \sum_{(x,y)\in\mathcal{T}} [1 - y(\beta_0 + x^T\beta)]_+ + \lambda||\beta||^2 \right\} , \qquad (3)$$

where $y = -1$ or $y = +1$ for the two classes. Increasing λ corresponds to taking account of more and more data points.

Similarly as for neural nets, find an optimal tuning parameter λ by use of a separate validation set or by cross-validation.

See further Section 3.2 in Lecture notes

Statistical image modelling

In Figure 2 two examples of images obtained by simulation from models with independent pixel values.

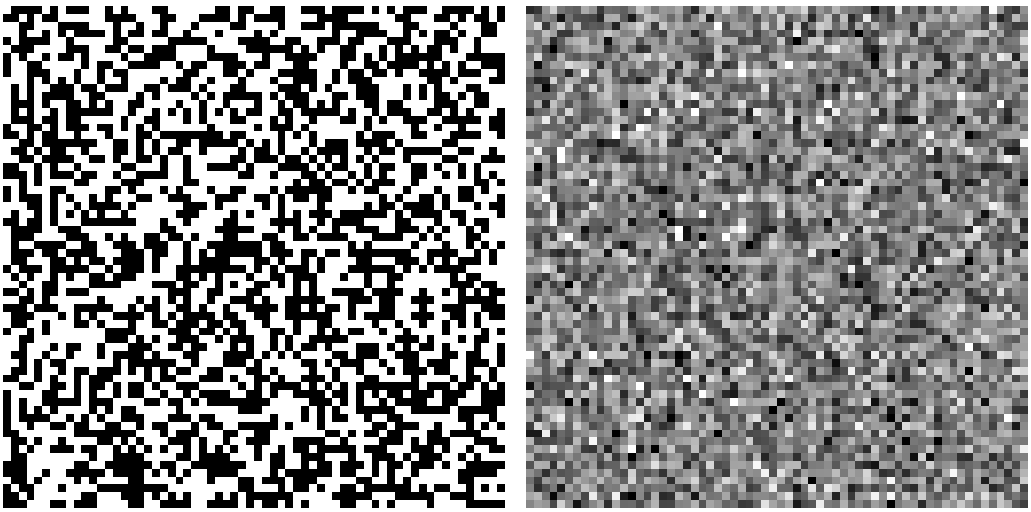

Figure 2: Images of size 64×64 obtained by simulation from models with independent pixel values: to the left a black-and-white image with equal probabilities for the two colours, and to the right a grey-level image with values from a normal distribution with expectation $\mu = 0.5$, a standard deviation $\sigma = 0.2$ and truncated to the interval $[0, 1]$.

How generalize to models with dependence between pixel values?

We will consider Markov random field models defined by a neighbourhood for each pixel

One-dimensional Markov chains

A random sequence X_t with values in a finite or countable set V is a Markov chain if

$$\Pr(X_{t+1} = x | X_s, s \leq t) = \Pr(X_{t+1} = x | X_t), \quad x \in V. \qquad (4)$$

How can this be generalized to processes in the plain?

One can prove that condition (4) is equivalent to the condition

$$\Pr(X_t = x | X_s, s \neq t) = \Pr(X_t = x | X_{t-1}, X_{t+1}), \quad x \in V \qquad (5)$$

Thus, to predict X_t from all $X_s, s \neq t$, it is enough to know X_s in two neighbouring sites with $s = t - 1$ and $s = t + 1$.

Condition (5) can be generalized in a straightforward way to several dimensions

Markov random field models

Regard a random image $X = (X_s, s \in S)$, where S denotes the set of sites (pixel locations).

To $s \in S$ there is a set $N_s \subset S$ of neighbour sites such that:

(i) $s \notin N_s$,

(ii) $t \in N_s$ if and only if $s \in N_t$.

Two often used neighbourhood systems are shown in Figure 3.

To the left a system where $s = (i, j)$ has the neighbourhood

$$N_s = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}. \qquad (6)$$

To the right a system with eight neighbours

Figure 3: Two often used neighbourhood systems: to the left the site s has four neighbours and to the right it has eight neighbours.

Suppose $X = (X_s, s \in S)$ is a set of discrete random variables with values in V.

Then X is a *Markov random field* with respect to the system $(N_s, s \in S)$ of neighbourhoods if

$$\Pr(X_s = x | X_t, t \neq s) = \Pr(X_s = x | X_t, t \in N_s), \quad x \in V, s \in S. \tag{7}$$

Thus to predict the pixel value X_s knowing all other pixel values it is enough to know the pixel values in the neighbourhood N_s.

Highly useful in an iterative sampling method called Gibbs sampling for simulation of a Markov random field.

Neighbourhoods of border sites have to be considered separately.

Suppose that the set of sites is

$$S = \{(i, j) : i = 1, \ldots, m, j = 1, \ldots, n\}. \tag{8}$$

One possibility is to use *periodic boundary conditions*

This means that sites in the leftmost column are considered as neighbours of sites in the rightmost column

Similarly, sites in the top row are considered as neighbours of the bottom row.

Specifically with four neighbourhoods for non-border sites, define for $s = (i, n)$ with $1 < i < m$

$$N_s = \{(i-1, n), (i+1, n), (i, n-1), (i, 1)\}, \tag{9}$$

with similar definitions for other border sites.

Think of periodic boundary conditions as a folding of S like a torus (a doughnut)!

Example 4.1. The Ising model. Let S be a $m \times n$-rectangle with periodic boundary conditions.

In physical applications m and n are large

Suppose X_s can take two possible values, -1 and $+1$.

Let X_s^+ and X_s^- denote the number of neighbours of s that take positive and negative values. Thus $X_s^+ + X_s^- = 4$.

Assume that

$$\Pr(X_s = +1 | X_t, t \in N_s) = \frac{\exp(2\beta(X_s^+ - X_s^-))}{1 + \exp(2\beta(X_s^+ - X_s^-))} \qquad (10)$$

and that $\beta > 0$.

Note that if $X_s^+ > X_s^-$ then the probability that X_s shall have a positive value is greater than $1/2$.

An alternative way of specifying the probability distribution of X is as a Gibbs distribution,

$$\Pr(X = x) = \frac{1}{Z} \exp(\beta \sum_{s \sim t} x_s x_t), \tag{11}$$

where Z is a normalizing constant, and $s \sim t$ denotes that s and t are neighbours.

In the right member of (11) we sum over all pairs (s, t) of sites that are neighbours.

In physics the Ising model is used as a model for ferromagnetism and β may be interpreted as inverse temperature.

For temperature below a critical value, that is for $\beta > \beta_c$, there are long range dependencies and possible phase transitions

Then a clear majority of the X_s-values will either be equal to $+1$ or a clear majority will be equal to -1.

For $\beta < \beta_c$ there are no phase transitions and the value of X_s averaged over large sets of sites is close to zero.

A famous computation by Onsager from 1944 gives

$$\beta_c = \frac{1}{2} \log(1 + \sqrt{2}) = 0.44069 \tag{12}$$

Autonormal random field models

Regard a Markov random field models, where $X_s, s \in S$, are continuous real-valued random variables.

The Markov condition needs then a modification to

$$\Pr(X_s \in A | X_t, t \neq s) = \Pr(X_s \in A | X_t, t \in N_s), \quad A \subseteq \mathbb{R}, s \in S, \tag{13}$$

for all considered subsets A of $\mathbb{R}$.

Here we only consider some *autonormal* models where the conditional distribution of X_s given its neighbours is normal with a constant variance σ^2 and an expectation that is a linear combination of the neighbour values.

Consider the neighbourhood system with four neighbours and denote the neighbours of s in the West, North, East and South directions $W(s)$, $N(s)$, $E(s)$, and $S(s)$ Assume that

$$\mathbf{E}(X_s | X_t, t \in N_s) = \mu + \beta_W(X_{W(s)} - \mu) + \beta_N(X_{N(s)} - \mu) + \beta_E(X_{E(s)} - \mu) + \beta_S(X_{S(s)} - \mu). \tag{14}$$

Simulation of Markov random fields

There are several ways of simulating images from Markov random field models. One of the most used methods is Gibbs sampling.

In *Gibbs sampling* sites $s \in S$ are viited in a specified way which may be random or deterministic.

An often used random method is to choose successive sites to be visited independently and in a purely random way from the set of all sites.

And an often used deterministic visiting scheme is to choose sites to be visited row-wise from left to right starting with the first row and proceeding until all sites have been visited.

Such a set of visits is called a sweep. The procedure is iterated a given number of of sweeps.

Example 4.2. The Ising model. Continuation. Consider Gibbs sampling for the Ising model.

Start with a purely random configuration. For a set of β-values we see in Figure 4 binary images obtained by deterministic row-wise sweeps as described above.

The upper two rows correspond to β values under the critical value, that if for high temperature, and the two lower rows correspond to low temperature.

In the middle row we have β very close to the critical value, actually slightly above.

It may be noted that for large β-values (the two lower rows) the number of iterations used in Figure 4 is far too small to arrive at a stationary distribution for the Markov chain formed by the successive iterations.

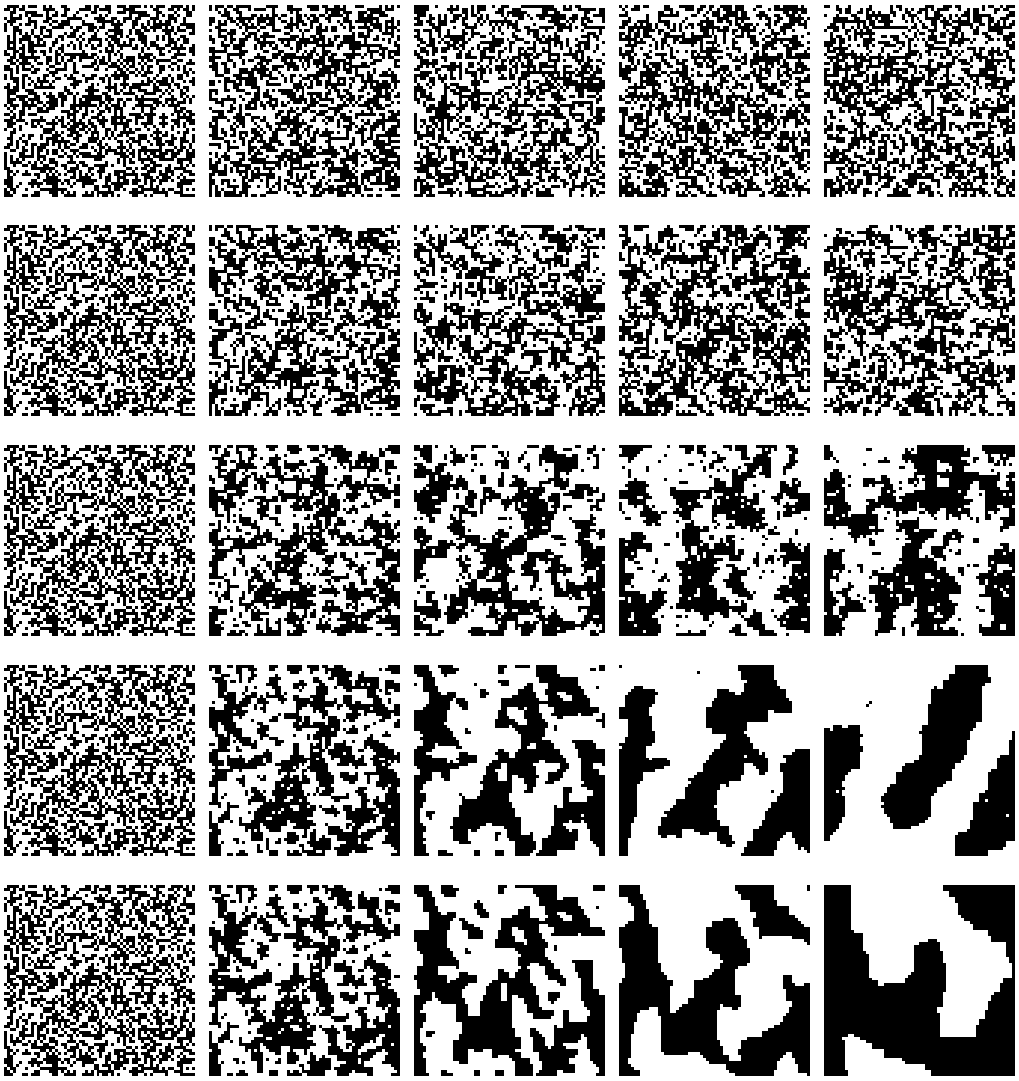

Figure 4: Binary images obtained by simulation for the Ising model with $\beta = 0.11$, 0.22, 0.4407, 0.88 and 1.76 in rows 1 to 5, respectively. In the columns we have to the left a purely random start configuration and then the result after 1 sweep, after 4 sweeps, after 16 sweeps and after 64 sweeps, respectively.

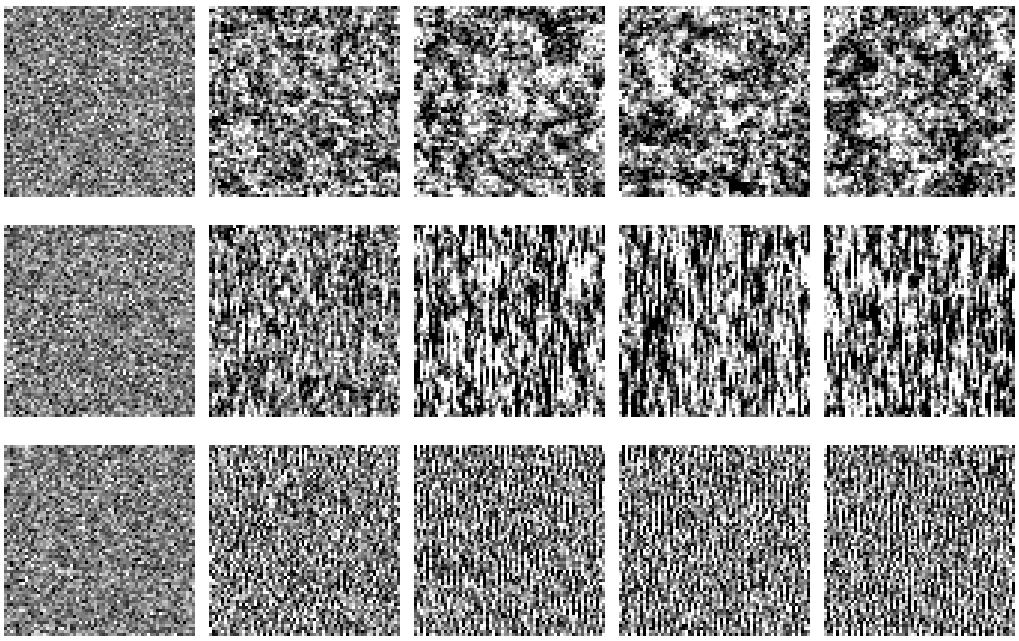

Figure 5: Grey-scale images obtained by simulation for autonormal models. In the columns we have to the left a purely random start configuration and then the result after 1 sweep, after 16 sweeps, after 128 sweeps and after 256 sweeps, respectively. The parameters in (14) are in the upper row $\beta_W = \beta_E = \beta_N = \beta_S = 0.24$, in the second row $\beta_W = \beta_E = 0$ and $\beta_N = \beta_S = 0.48$, and in the third row $\beta_W = \beta_E = -0.24$ and $\beta_N = \beta_S = 0.24$. In all three rows we have $\mu = 0.5$ and the residual standard deviation $\sigma = 0.3$.

Example 4.3. Simulation of an autonormal model. Consider Gibbs sampling for the autonormal model with conditional expectations (14) and constant conditonal variance given the neighbour values. For three sets of parameters we obtain results shown in Figure 5. □

Bayesian analysis of images

Common approach in Bayesian image analysis: assume that we start with a random image X given by a Markov random field.

Then we observe a distorted image Y and one basic problem is to reconstruct X from Y.

A simple model for the observed image $Y = (Y_s, s \in S)$ is to assume that given X the Y_s-variables are independent and that the distribution of Y_s only depends on X_s, that is

$$\Pr(Y = y|X) = \prod_{s \in S} \Pr(Y_s = y_s|X_s). \tag{15}$$

Reconstruction of X from Y is difficult computionally.

Iterative algorithms have been developed for this, most of them based on Markov chain Monte Carlo algorithms.

Bayesian models for image reconstruction by use of Markov random field models as priors for the unobserved image X has generally suffered from the problem that it seems difficult to specify realistic priors for images typically found in applications.

An interesting approach developed by David Mumford and Song Chun Zhu, see Section 4.5 in Lecture notes for details.

See further Section 4.7 for Markov chain Monte Carlo, particularly the Metropolis-hastings algorithm.

Point processes. Poisson processes.

Let A be a subset of $\mathbb{R}^2$ with finite and positive area $|A|$.

Consider a random subset X of A consisting of finitely many points, and call X a point process on A.

If $B \subseteq A$ we let $X(B)$ denote the number of points in X that belong to B.

The point process X is said to be *stationary* if the probability distribution of X is invariant under any translation of the sets B where we regard the point process.

Further, X is *isotropic* if the process is stationary and the distribution of X is invariant under any rotation of such sets B.

Consider a stationary point process X on A such that $X(A)$ has finite expectation. One can then show that

$$\mathbf{E}(X(B)) = \lambda |B| \tag{16}$$

for some constant λ called the intensity of the point process.

Poisson process with constant intensity.

A point process X is called a Poisson process with constant intensity $\lambda \geq 0$ on A if

(i) $X(B_1)$ and $X(B_2)$ are independent for disjoint subsets B_1 and B_2 of A

(ii) $X(B)$ is Poisson distributed with expectation $\lambda|B|$ for a subset $B \subseteq A$ with area $|B|$, that is

$$\Pr(X(B) = n) = \frac{(\lambda|B|)^n}{n!} \exp(-\lambda|B|). \tag{17}$$

A Poisson process with constant intensity is stationary and isotropic.

A Poisson process on A with intensity λ can be generated in the following way:

(i) Let first N be Poisson distributed with expectation $\lambda|A|$.

(ii) Given that $N = n$, generate $X_1, \ldots, X_n$ as independent and identically distributed variables, each with a uniform distribution over A.

Then let X consist of the points $X_1, \ldots, X_n$, that is $X = \{X_1, \ldots, X_n\}$.

In Figure 6 we see two examples of such generation of a Poisson process in the unit square with the constant intensity $\lambda = 50$.

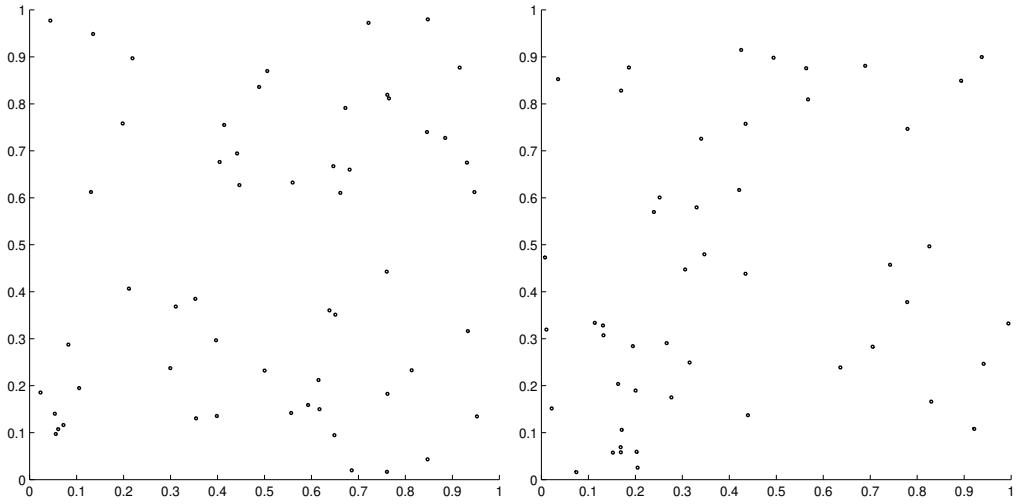

Figure 6: Two examples of Poisson point processes generated in the unit square with $\lambda = 50$. The generated number of points is to the left $N = 55$ and to the right $N = 49$.

Poisson process with varying intensity.

A point process X is called a Poisson process with intensity function $\lambda(s), s \in A$, if

(i) $X(B_1)$ and $X(B_2)$ are independent for disjoint subsets B_1 and B_2 of A

(ii) $X(B)$ is Poisson distributed with expectation $\int_B \lambda(s)\, ds$ for $B \subseteq A$.

A Poisson process with intensity function $\lambda(s), s \in A$, can be generated in the following way

(i) Let first N be Poisson distributed with expectation $\int_A \lambda(s)\, ds$.

(ii) Given that $N = n$, generate $X_1, \ldots, X_n$ as independent and identically distributed variables, each with a distribution specified by

$$\Pr(X_i \in B) = \frac{\int_B \lambda(s)\, ds}{\int_A \lambda(s)\, ds} \text{ for } B \subseteq A. \tag{18}$$

Then put $X = \{X_1, \ldots, X_n\}$.

The Neyman-Scott process, a point processes with clustering

(i) Consider a Poisson process with constant intensity λ, and regard the points of this process as mother points.

(ii) From each mother point we generate daughter points such that the number of daughter points from the mother points are all independent and identically distributed.

(iii) Two-dimensional vectors from a mother point to the daughter points are all independent and identically distributed. This distribution we call the scattering distribution.

The process of daughter points is called a Neyman-Scott process.

Suppose that we want to generate a Neyman-Scott process. If the daughter process is regarded on a set A we need to start by generating the mother point process on a set larger than A, in fact so large that (essentially) all points from which daughters can get scattered into A are included.

A Neyman-Scott plant process with 2D normal scattering.

We want simulate a Neyman-Scott process of mother and daughter plants within the unit square $[0,1]\times[0,1]$ with

(i) Intensity $\lambda = 10$ for the Poisson process of mother points

(ii) A number of daughter points that is binomial (n, p) with $n = 8$ and $p = 0.5$

(iii) A 2D scattering distribution that is $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ with $\mu_1 = \mu_2 = \sigma_1 = \sigma_2 = 0.1$ and $\rho = 0.5$ corresponding to wind spread of seeds with a main wind direction from south-west.

We start by simulating the Poisson mother plant point process in the axis-parallell quadrat with south-west and north-east corners in $(-0.5, -0.5)$ and $(1.3, 1.3)$, respectively. The result of the simulation is shown in Figure 7.

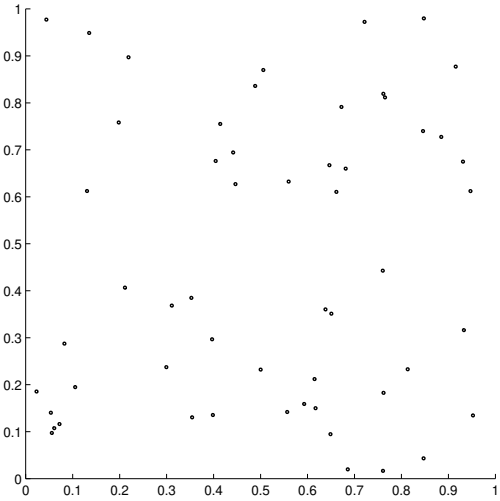

Figure 7: A simulation of a Neyman-Scott process with mother points as circles and daughter points as dots. OBS OBS a new figure must be generated.

A hard-core inhibition point process

In the cluster point process in the previous section the occurrence of a point typically increases the intensity of points in a neighborhood of this point.

We will now describe a point processes with inhibition, suggested 1960 by Matérn, which has the opposite property: the occurrence of a point inhibits other points within a certain distance.

(i) Start by generating a Poisson point process with intensity λ on a bounded set A.

(ii) To each point $X_i, i = 1, \ldots, N$, associate a random mark consisting of random variable U_i, which is uniformly distributed on the interval $(0, 1)$ and such that the U_i's are indendent, mutually and of the X_i's. We can think of U_i as the birth time of the point X_i.

(iii) Then we thin the X-process by deleting each point X_i for which there exists an older point X_j of the original point process closer than a distance d, that is a point X_j satisfying $|X_i - X_j| < d$ and $U_j < U_i$. The distance d is called the hard core distance.

The K-function, a diagnostic tool for detecting clustering and inhibition

Consider an isotropic point process with intensity λ

Let x be a point of the point process X

$\|y - z\|$ is the distance between two points y and z in $\mathbb{R}^2$
Define the K-function of X as follows

$$K(r) = \frac{1}{\lambda}\mathbf{E}(\text{number of further points of } X \text{ within distance } r \text{ from } x | x \in X) \tag{19}$$

More precisely

$$K(r) = \frac{1}{\lambda}\mathbf{E}(X(C_x(r)|x \in X), \tag{20}$$

where $C_x(r) = \{y : 0 < \|y - x\|) \leq r\}$

For a stationary Poisson process

$$K(r) = \pi r^2. \tag{21}$$

Regard also $L(r) = (K(r))^{1/2}$, for a Poisson process

$$L(r) = \sqrt{\pi} r \tag{22}$$

For a point process with clustering we can expect that the K-function will lie above the K-function for a Poisson process

while with inhibition $K(r)$ should lie below

Point processes operations such as thinning, displacement and superposition

Thinning: simplest case, points are deleted independently with a probability $1 - p$, and retained with retention probablity p, $1 \leq p \leq 1$.

A p-thinned Poisson process with constant intensity λ is a Poisson process with intensity $p\lambda$.

Displacement: simplest case, points are displaced independently with one displacement dsitribution

Superposition: simplest case, superposition $X \cup Z$ of two point processes X and Z on a given set A