Spatial statistics (NMST543)

December 30, 2017

Contents

1. Point process statistics

In this chapter we will deal with the statistical analysis of simple point processes on $\mathbb{R}^d$. First we start with the estimation of summary characteristics. Afterwards we look at the hypothesis testing. Finally, we consider parametric models and discuss the problem of model fitting and diagnostics.

Recall that the point process is defined as a random locally finite counting measure. Simple point process can be viewed also as a random locally finite set. We will use both approaches. Thus, we write $\Phi(B)$ for the number of points (atoms) of the process Φ in the set B. By $X \in \Phi$ we mean that X is a point (atom) of Φ.

1.1 Estimation of summary characteristics

Assume that we have a single realization of Φ in the set $W \in \mathcal{B}_0^d$, so called *observation window*. The window is usually d-dimensional rectangle but it can also have much more complicated shape. Our aim is to estimate summary characteristics of the point process Φ based on the given realization. We provide a list of basic estimates. Most of them are implemented in the R library spatstat [1]. Therefore, we always mention also the corresponding R functions.

Edge effects

When estimating numerical and functional summary characteristics, *edge effects* play an important role. They are caused by the fact that a point process is observed in a bounded window W. For example, we can base the estimate of K-function $K(r)$ on the number of points in balls centred at the points of the process and radius r. However, we are not able to determine this number when the distance of the point to the boundary of W is smaller than r. The situation is depicted in Figure 1 left – true number of points in $b(X, r)$ is 5 but in W we only observe 3 points. For another example we can consider estimation of G-function. We are looking for the nearest neighbour of the point $X \in \Phi$. we would From information inside W in Figure 1 right we would determine Y as the nearest neighbour of X, but in actual fact the nearest neighbour is Z that lies outside the window W. We see that by ignoring edge effects our conclusion about characteristics of the point process could be distorted.

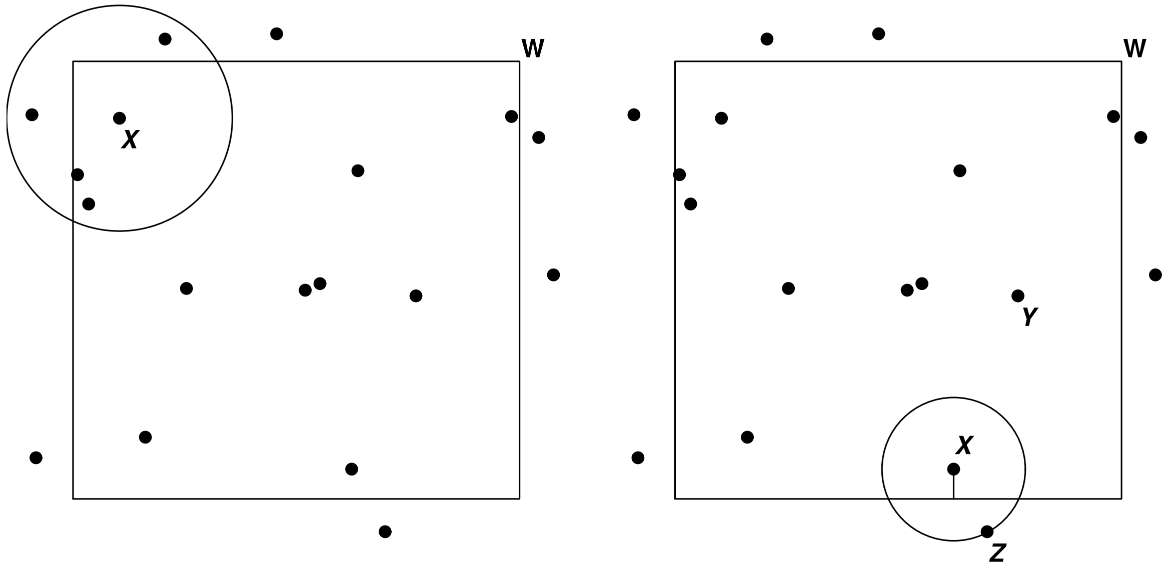

Figure 1. An illustration of edge effects in the case of estimating K-function (left) and G-function (right).

Estimation of intensity function

Let Φ be a stationary point process on $\mathbb{R}^d$ with intensity λ. It follows directly from the definition that

$$\hat{\lambda} = \frac{\Phi(W)}{|W|}$$

is an unbiased estimator of λ. In spatstat package this estimator can be computed using summary.ppp. If Φ is homogeneous Poisson point process, $\hat{\lambda}$ is moreover maximum likelihood estimator. In fact, we may understand Φ on W as a finite point process with density w.r.t. the distribution of unit Poisson process on W and we know that its density has the form

$$p_\lambda(\varphi) = \lambda^{\varphi(W)} \mathrm{e}^{(1-\lambda)|W|}.$$

It is easy to verify that the likelihood function $L(\lambda) = p_\lambda(\varphi)$ is maximized for $\lambda = \varphi(W)/|W|$ (see Exercise class).

For a non-stationary point process with intensity function λ, non-parametric *kernel* estimators are often used. One possibility is to consider the estimator (density.ppp)

$$\hat{\lambda}(x) = \frac{1}{c_{W,b}(x)} \sum_{Y \in \Phi \cap W} k_b(x - Y), \quad x \in W,$$

where k_b is kernel function with *bandwidth* $b > 0$, i.e. $k_b(x) = \frac{k(x/b)}{b^d}$ for some probability density k, and

$$c_{W,b}(x) = \int_W k_b(x - y) \, \mathrm{d}y$$

is the *correction on edge effects*. Another possibility is to use more exact but computationally more demanding estimator (density.ppp with diggle=TRUE)

$$\hat{\lambda}(x) = \sum_{Y \in \Phi \cap W} \frac{k_b(x - Y)}{c_{W,b}(Y)}. \tag{1}$$

The estimator $\hat{\lambda}(x)$ is usually sensitive on the choice of bandwidth while the choice of kernel function doesn't play so important role. For small values of b the estimator is too concentrated around the observed points. On the other hand, for larger b the smoothing could be substantial. Density functions of uniform distribution on a ball or Gaussian distribution are one of the most usual choices of the function k. Often k is chosen as the product of one-dimensional densities: $k(x) = k_1(x_1) \cdots k_d(x_d)$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. A popular example of one-dimensional kernel function is *Epanechnikov kernel*:

$$e(u) = \frac{3}{4}(1 - u^2), \quad u \in [-1, 1].$$

Note that if k is symmetric, then (1) is globally unbiased in the sense that

$$\int_W \hat{\lambda}(x) \, \mathrm{d}x = \Phi(W)$$

, i.e.

$$\mathbb{E} \int_W \hat{\lambda}(x) \, \mathrm{d}x = \int_W \lambda(x) \, \mathrm{d}x.$$

For non-stationary Poisson process we can express the likelihood function because

$$p_\theta(\varphi) = \exp\{|W| - \int_W \lambda_\theta(x) \, \mathrm{d}x\} \prod_{x \in \varphi} \lambda_\theta(x).$$

If we the intensity function $\lambda_\theta(x)$ has parametric form, then the task of finding maximum likelihood estimator of parameter θ has to be solved by some numerical method. We will return to this problem in Subsection 1.3.

Estimation of K-function

Recall that the K-function $K(r)$ of a stationary point process Φ is defined through the relation

$$\lambda K(r) = \mathbb{E}_o^! \Phi(b(o,r)), \quad r > 0,$$

which could be equivalently rewritten as

$$\lambda K(r) = \mathbb{E} \sum_{X \in \Phi \cap A} \frac{\Phi(b(X,r) \setminus \{X\})}{\lambda |A|} = \mathbb{E} \sum_{X,Y \in \Phi}^{\neq} \frac{\mathbf{1}_{[X \in A, \|X-Y\| \leq r]}}{\lambda |A|}, \tag{2}$$

where $A \in \mathcal{B}_0^d$ is an arbitrary bounded Borel set with positive Lebesgue measure ($|A| > 0$).

The following estimators can be obtained by the function Kest in the library spatstat.

0. uncorrected estimate: The equation (2) offers a theoretical unbiased estimator of $\lambda^2 K(r)$ in the form

$$\widehat{\lambda^2 K(r)} = \sum_{X \in \Phi \cap W} \frac{\Phi(b(X,r) \setminus \{X\})}{|W|} = \sum_{X,Y \in \Phi}^{\neq} \frac{\mathbf{1}_{[X \in W, \|X-Y\| \leq r]}}{|W|}.$$

This estimator could be used only if we have additional information from outside of the window W. We need to count also the points Y that are in distance at most r from the points of Φ in W. This situation is known as the *plus sampling*. The problem of inapplicability of the estimate rests in edge effects. We are unable to determine $\Phi(b(X,r) \setminus \{X\})$ only from information inside the window W, see Figure 1 left. If we ignore the edge effects and consider only the points inside W, we get negatively biased estimator

$$\widehat{\lambda^2 K_u(r)} = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[\|X-Y\| \leq r]}}{|W|}.$$

The package spatstat enables to calculate this estimator by setting correction="none" in Kest. However, it is only for instructional reasons. In practice this estimator should not be used. All the following estimators try to compensate the problem of edge effects by including the edge correction factor $e(X,Y,W)$. The estimators of $\lambda^2 K(r)$ then have the form

$$\sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[\|X-Y\| \leq r]}}{|W|} e_{W,r}(X,Y).$$

For $\widehat{\lambda^2 K_u(r)}$ there is no edge correction and the factor $e_{W,r}(X,Y)$ is identically equal to 1.

1. border estimate, correction="border": The simplest way to avoid the edge effects is to consider Φ in smaller window

$$W_{\ominus r} = W \ominus b(o,r) = \{y \in W : b(y,r) \subseteq W\} = \{y \in W : d(y, \partial W) \geq r\},$$

where ∂W denotes the boundary of W. We only consider the points for which we are able to determine the number of neighbours at distance r. This is known as the *minus sampling*. Then

$$\widehat{\lambda^2 K_b(r)} = \sum_{X \in \Phi \cap W_{\ominus r}} \frac{\Phi(b(X,r) \setminus \{X\})}{|W_{\ominus r}|} = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[X \in W_{\ominus r}, \|X-Y\| \leq r]}}{|W_{\ominus r}|}$$

is an unbiased estimator $\lambda^2 K(r)$, as it follows from (2). The edge correction factor is

$$e_{W,r}(X,Y) = \frac{\mathbf{1}_{[X \in W_{\ominus r}]} |W|}{|W_{\ominus r}|}.$$

The estimator $\widehat{\lambda^2 K_b(r)}$ is defined for $r < r_b = \sup\{s > 0 : |W_{\ominus s}| > 0\}$. For example, $r_b = 0.5$ for $W = [0,1]^2$.

2. translation correction, correction="translate": Another possibility is to let the edge correction factor be the function of both X and Y. The *translation correction factor* is

$$e_{W,r}(X,Y) = \frac{|W|}{|W \cap (W + X - Y)|},$$

which leads to the estimator

$$\widehat{\lambda^2 K_t(r)} = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[\|X-Y\| \le r]}}{|W \cap (W + X - Y)|}.$$

Using the Campbell theorem it can be shown that this estimator is unbiased (see Exercises). The estimator is well defined for $r < r_t = \sup\{s > 0 : |W \cap (W + x)| > 0 \ \forall x : \|x\| \le s\}$. For example, $r_t = 1$ for $W = [0,1]^2$. Similarly, we can defined the estimator of reduced second-order moment measure

$$\widehat{\lambda^2 \mathcal{K}_t(B)} = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[X-Y \in B]}}{|W \cap (W + X - Y)|}.$$

The smoothed kernel estimate of the density can be obtained by Kmeasure.

3. Ripley's isotropic correction, correction="isotropic" or correction="Ripley": Another correction factor was suggested by B. D. Ripley [11]. It has the form

$$e_{W,r}(X,Y) = \frac{|\partial b(X, \|X - Y\|)|}{|\partial b(X, \|X - Y\|) \cap W|}$$

and yields the estimator

$$\widehat{\lambda^2 K_R(r)} = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[\|X-Y\| \le r]}}{|W|} \cdot \frac{|\partial b(X, \|X - Y\|)|}{|\partial b(X, \|X - Y\|) \cap W|}.$$

If the process is isotropic, it can be shown that it is an unbiased estimator of $\lambda^2 K(r)$ for $r < r_0 = \inf\{t > 0 : |W^{(t)}| < |W|\}$, where $W^{(t)} = \{x \in W : \partial b(x,t) \cap W \neq \emptyset\}$. Ohser's modification [8] is given by

$$\widehat{\lambda^2 K_O(r)} = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[\|X-Y\| \le r]}}{|W^{(\|X-Y\|)}|} \cdot \frac{|\partial b(X, \|X - Y\|)|}{|\partial b(X, \|X - Y\|) \cap W|}.$$

It extends the definition of $\widehat{\lambda^2 K_R(r)}$ to $r < r^* = \sup\{s > 0 : |W^{(s)}| > 0\}$. For $r < r_0$ we have $\widehat{\lambda^2 K_R(r)} = \widehat{\lambda^2 K_O(r)}$. In case of planar unit square $W = [0,1]^2$, $r_0 = \sqrt{2}/2$, $r^* = \sqrt{2}$ and $W^{(s)} = W$ for all $s \le r_0$.

In order to get the estimator of $K(r)$ itself, we have to divide by the estimate of λ^2. This violates the unbiasedness property. The bias and variance are typically increasing with increasing r. For a planar rectangular window it is recommended to determine the estimators only for r smaller than a 1/4 of the smaller side length of the rectangle. The estimator of λ^2 often has the following form

$$\widehat{\lambda^2} = \frac{\Phi(W)(\Phi(W) - 1)}{|W|^2}.$$

The advantage is that $\widehat{\lambda^2}$ is unbiased estimator of λ^2 in the case of Poisson point process Φ.

Border estimate of the K-function does not have to be monotone function in r (as opposed to the theoretical function). With increasing r and dimension d the loss of information is more essential. The estimators based on translation or isotropic correction factors have statistically better properties. On the other hand, the computation of $\hat{K}_b$ is faster.

Estimation of inhomogeneous K-function

Let Φ be a second-order intensity reweighted stationary point process. Then we define the inhomogeneous K-function as

$$K_{\text{inhom}}(r) = \mathbb{E} \sum_{X,Y \in \Phi}^{\neq} \frac{\mathbf{1}_{[X \in A, \|X-Y\| \leq r]}}{\lambda(X)\lambda(Y)|A|}.$$

Its estimates could be obtained similarly as in the case of stationary processes. For example, the estimator with translation correction has the form

$$\widehat{K}_{\text{inhom}}(r) = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[\|X-Y\| \leq r]}}{\hat{\lambda}(X)\hat{\lambda}(Y)|W \cap (W + X - Y)|},$$

where $\hat{\lambda}(x)$ is the estimator of intensity function $\lambda(x)$. In package spatstat we would use Kinhom with choice correction="translate".

Estimation of pair correlation function

For stationary and isotropic point process the pair correlation function g is related to the K-function by

$$g(r) = \frac{K'(r)}{\sigma_d r^{d-1}}, \quad r > 0,$$

where σ_d is the surface of the unit ball in $\mathbb{R}^d$. The kernel estimator of g is

$$\hat{g}(r) = \frac{1}{\widehat{\lambda^2}} \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{k_b(r - \|X-Y\|)}{\sigma_d r^{d-1}|W|} e_{W,r}(X,Y),$$

where k_b is a suitable kernel function with bandwidth b and $e_{W,r}(X,Y)$ is the edge correction factor. The estimator is obtained by pcf in spatstat. The choices correction="translate" and correction="ripley" correspond to the translation and isotropic edge correction, respectively. In case of second-order intensity reweighted stationary point process, $\widehat{\lambda^2}$ is replaced by $\hat{\lambda}(X)\hat{\lambda}(Y)$ in the denominator of each summand. The computation is performed by pcfinhom.

Another possibility would be to use some estimator of K-function and approximate its derivative by numerical methods (e.g. using splines). This is usually not easy because the estimators of K-function are piecewise constant functions.

Estimation of nearest neighbour distance distribution function

Recall that for stationary point process we define the nearest neighbour distance distribution function as

$$G(r) = P_o^!(\{\varphi \in \mathcal{N} : \varphi(b(o,r)) > 0\}), \quad r > 0.$$

For the computation of the following estimators of G we can use Gest.

0. uncorrected estimate, correction="none": If we would the nearest neighbour distance for each observed point of the process, then we can estimate G classically by the empirical distribution function

$$\hat{G}(r) = \frac{1}{\Phi(W)} \sum_{X \in \Phi \cap W} \mathbf{1}_{[e(X) \leq r]},$$

where $e(x) = d(x, \Phi \setminus \{x\})$ is the distance from the point x to its nearest neighbour. From the Campbell-Mecke theorem it follows that

$$\mathbb{E} \sum_{X \in \Phi \cap W} \mathbf{1}_{[e(X) \leq r]} = \lambda \int_W \int_{\mathcal{N}} \mathbf{1}_{[d(o,\varphi) \leq r]} P_o^!(\mathrm{d}\varphi)\, \mathrm{d}x = \lambda |W| G(r).$$

Hence, we see that $\hat{G}(r)$ is so called *ratio-unbiased* estimator of G. It means that $\hat{G}(r)$ is given as the ratio, where the ratio of the expectations of numerator and denominator gives $G(r)$, i.e.

$$\frac{\mathbb{E} \sum_{X \in \Phi \cap W} \mathbf{1}_{[e(X) \leq r]}}{\mathbb{E}\Phi(W)} = \frac{\lambda |W| G(r)}{\lambda |W|} = G(r).$$

Again due to the edge effects we are not able to get $e(X)$ for each $X \in \Phi \cap W$, see Figure 1 right. We can replace $e(X)$ by the distance $e^*(X) = d(X, (\Phi \setminus \{X\}) \cap W) \geq e(X)$, which we are able to observe in the window. Then we obtain the naive estimator

$$\hat{G}_r(r) = \frac{1}{\Phi(W)} \sum_{X \in \Phi \cap W} \mathbf{1}_{[e^*(X) \leq r]}.$$

Since $e^*(X) \leq r$ implies $e(X) \leq r$, we have $\hat{G}_r(r) \leq \hat{G}(r)$. The estimator $\hat{G}_r(r)$ is not used for practical purposes. However, it can be obtained by the choice correction="none".

1. border estimate, correction="border" or "rs": By restricting again to the eroded window $W_{\ominus r}$, the following ratio-unbiased estimator is obtained,

$$\hat{G}_b(r) = \frac{1}{\Phi(W_{\ominus r})} \sum_{X \in \Phi \cap W_{\ominus r}} \mathbf{1}_{[e(X) \leq r]} = \frac{1}{\Phi(W_{\ominus r})} \sum_{X \in \Phi \cap W_{\ominus r}} \mathbf{1}_{[e^*(X) \leq r]}.$$

2. Kaplan-Meier estimate, correction="km": Edge effects could be understood as a type of censoring (see Subsection 5.1). Therefore, we can introduce Kaplan-Meier type estimator,

$$\hat{G}_{KM}(r) = 1 - \prod_{s \leq r} \left(1 - \frac{\#\{X \in \Phi \cap W : e(X) = s, e(X) \leq c(X)\}}{\#\{X \in \Phi \cap W : e(X) \geq s, c(X) \geq s\}} \right),$$

where $c(x) = d(x, \partial W)$ is the distance of x from the window boundary. If $e(X) \leq c(X)$, we are sure that we observe true distance to the nearest neighbour of x. In the opposite case the distance $e(X)$ is censored by $c(X)$. We only know that $e(X)$ is larger than $c(X)$. However, we don't have information about exact value of $e(X)$. Realize that only information from the window W is sufficient for evaluating of $\hat{G}_{KM}(r)$. As opposed to the classical situation of random censoring we may not expect independence of data and censors. Hence, the optimality of Kaplan-Meier estimator is violated. Nevertheless, it usually gives better results than the border estimate.

For an absolutely continuous distribution function $H(t)$ with density $h(t)$, the *hazard rate* is defined as $\lambda_h(t) = h(t)/(1 - H(t))$. The spatial Kaplan-Meier method enables to estimate the hazard rate $\lambda_h(r)$ of the distribution function $G(r)$. However, we have to be cautious because G does not have to admit a density. In library spatstat this estimator can be obtained together with the Kaplan-Meier estimator of G-function.

3. Hanisch estimate, correction="han" or "Hanisch": Another improvement of the border estimate is obtained by the following edge correction:

$$\hat{G}_H(r) = \frac{1}{\hat{\lambda}} \sum_{X \in \Phi \cap W} \frac{\mathbf{1}_{[e(X) \leq c(X)]}}{|W_{\ominus e(X)}|} \mathbf{1}_{[e(X) \leq r]},$$

where

$$\hat{\lambda} = \sum_{X \in \Phi \cap W} \frac{\mathbf{1}_{[e(X) \leq c(X)]}}{|W_{\ominus e(X)}|}.$$

This estimator uses only points that are closer to its neighbour than to the boundary. In Figure 1 right the point X is not involved in the estimator because its distance to the boundary of W is smaller than the distance to its nearest neighbour. The Hanisch estimator is ratio-unbiased as we can easily verify by the Campbell-Mecke theorem if we realize that $\mathbf{1}_{[e(X) \leq c(X)]} = \mathbf{1}_{[X \in W_{\ominus e(X)}]}$.

The estimators of G may not have the properties of distribution function: $\hat{G}_b$ is not necessarily monotone, $\hat{G}_{KM}$ is non-decreasing but its maximal value could be strictly smaller than 1.

Estimation of contact distribution function

The contact distribution function of a stationary point process Φ is defined as

$$F(r) = \mathbb{P}(\Phi(b(o, r)) > 0) = \mathbb{P}(D \leq r), \quad r > 0,$$

where D is the distance from the origin to the nearest point of Φ.

In the space $\mathbb{R}^d$ we choose a regular grid I_a:

$$I_a = y + a\mathbb{Z}^d = \{(y_1 + a_1 z_1, \dots, y_d + a_d z_d) \in \mathbb{R}^d : z_i \in \mathbb{Z}\},$$

where $y = (y_1, \dots, y_d) \in \mathbb{R}^d$ and $a = (a_1, \dots, a_d) \in \mathbb{R}^d_+$, i.e. $a_i > 0$ for $i = 1, \dots, d$. The function Fest can be used for calculation of the following estimators of F.

0. uncorrected estimate, correction="none": For every point of the grid in the window W find the nearest point of the process. This nearest point may lie outside the window. If we only look for the nearest points from $\Phi \cap W$, we get

$$\hat{F}_r(r) = \frac{1}{|I_a \cap W|} \sum_{x \in I_a \cap W} \mathbf{1}_{[d(x, \Phi \cap W) \leq r]},$$

where $|I_a \cap W|$ is the number of points of a finite set $I_a \cap W$. This estimator is negatively biased, i.e. $\mathbb{E}\hat{F}_r(r) \leq F(r)$, because $\mathbf{1}_{[d(x,\Phi\cap W)\leq r]} \leq \mathbf{1}_{[d(x,\Phi)\leq r]}$ and $\mathbb{P}(d(x, \Phi) \leq r) = F(r)$ from stationarity. The bias is caused by edge effects.

1. border estimate, correction="border" or "rs": Let $d(x) = d(x, \Phi)$ be the distance of x from the nearest point of Φ. Then

$$\hat{F}_b(r) = \frac{1}{|I_a \cap W_{\ominus r}|} \sum_{x \in I_a \cap W_{\ominus r}} \mathbf{1}_{[d(x) \leq r]}$$

is unbiased estimator of $F(r)$ because $\mathbb{P}(d(x) \leq r) = F(r)$ by stationarity. The continuous version of this estimator (as $a \to o$) has the form

$$\hat{F}_b(r) = \frac{|W_{\ominus r} \cap \Phi_r|}{|W_{\ominus r}|},$$

where $\Phi_r = \{x \in \mathbb{R}^d : d(x, \Phi) \leq r\} = \cup_{X \in \Phi} b(X, r)$. Again it is an unbiased estimator.

2. Kaplan-Meier estimate, correction="km":

$$\hat{F}_{KM}(r) = 1 - \prod_{s \leq r} \left(1 - \frac{\#\{x \in I_a \cap W : d(x) = s, d(x) \leq c(x)\}}{\#\{x \in I_a \cap W : d(x) \geq s, c(x) \geq s\}}\right),$$

where $c(x) = d(x, \partial W)$ is the distance of x to the boundary of W.

Contact distribution function $F(r)$ of a stationary point process is absolutely continuous and the hazard rate $\lambda_h(r)$ exists. Its estimator is based on the Kaplan-Meier estimator $\hat{F}_{KM}(r)$.

3. Chiu-Stoyan estimate, correction="cs" or "Hanisch": Using the same correction as in the Hanisch estimator of G we obtain

$$\hat{F}_{CS}(r) = \frac{1}{C_a} \sum_{x \in I_a \cap W} \frac{\mathbf{1}_{[d(x) \leq c(x)]}}{|W_{\ominus d(x)}|} \mathbf{1}_{[d(x) \leq r]},$$

where

$$C_a = \sum_{x \in I_a \cap W} \frac{\mathbf{1}_{[d(x) \leq c(x)]}}{|W_{\ominus d(x)}|}.$$

The continuous version of this estimator is

$$\hat{F}_{CS}(r) = \frac{1}{C} \int_W \frac{\mathbf{1}_{[d(x) \leq c(x)]}}{|W_{\ominus d(x)}|} \mathbf{1}_{[d(x) \leq r]} \, \mathrm{d}x,$$

where

$$C = \int_W \frac{\mathbf{1}_{[d(x) \leq c(x)]}}{|W_{\ominus d(x)}|} \, \mathrm{d}x.$$

The estimators of F may not have the properties of distribution function: $\hat{F}_b$ is not necessarily monotone, $\hat{F}_{KM}$ is non-decreasing but its maximal value could be strictly smaller than 1. The border estimator is less efficient than the Kaplan-Meier estimator or the Chiu-Stoyan estimator.

Estimation of J-function

In spatstat it is possible to estimate the J-function

$$J(r) = \frac{1 - G(r)}{1 - F(r)}, \quad r > 0 : F(r) < 1,$$

using Jest.

The estimator of J arises from its definition:

$$\hat{J}(r) = \frac{1 - \hat{G}(r)}{1 - \hat{F}(r)}.$$

We may distinguish the following estimators (depending on the type of estimate of G and F):

- *uncorrected* (correction="none"),
- *border* (correction="border" or "rs"),
- *Kaplan-Meier* (correction="km"),
- *Hanisch* (correction="Hanisch").

Even if the uncorrected estimators $\hat{G}_r$ and $\hat{F}_r$ are substantially biased, taking their ratio gives approximately unbiased estimate (at least when the point process is close to the Poisson process). The advantage of this estimator is that it is insensitive to edge effects. Therefore, it should be used when the edge effects are severe.

Other three estimators are slightly biased (ratio of two approximately unbiased estimators). The logarithm of Kaplan-Meier estimator is an unbiased estimator of $\log J$.

Library spatstat enables to estimate four basic summary characteristics (functions F, G, J, K) at the same time by allstats.

Estimation of aggregation index

The expectation of an arbitrary non-negative random variable T can be expressed using its distribution function $H(t)$ as follows (e.g. [6], Lemma 5.7),

$$\mathbb{E}T = \int_0^\infty (1 - H(t))\,\mathrm{d}t.$$

Having the estimator $\hat{G}(t)$ of the nearest neighbour distance distribution function $G(t)$, we can estimate the Clark-Evans index

$$\mathrm{CE} = \frac{d(\lambda\omega_d)^{1/d}}{\Gamma(1/d)} \mathbb{E}_o^! D$$

as

$$\widehat{\mathrm{CE}} = \frac{d(\hat{\lambda}\omega_d)^{1/d}}{\Gamma(1/d)} \int_0^\infty (1 - \hat{G}(t))\,\mathrm{d}t.$$

In spatstat the function clarkevans is intended for estimation of CE.

1.2 Hypothesis testing

Next statistical task is testing the hypothesis that the observed point patterns corresponds to a given point process model. The most important case is testing of complete spatial randomness. If we do not reject this hypothesis, then we can model the data by Poisson process and it is unnecessary to consider more complicated processes. Complete spatial randomness test is one of the basic steps of exploratory data analysis.

Divide the observation window W to k mutually disjoint regions (so called *quadrats*) of the same volume and count the number of points in each of these quadrats. Denote these counts by $n_1, \ldots, n_k$. Under the hypothesis of homogeneous Poisson point process, these counts should form a realization of random sample from Poisson distribution with parameter $\lambda|W|/k$. Moreover, all $n = n_1 + \cdots + n_k$

points are i.i.d. and have uniform distribution in W. Hence, we can use the well-known Pearson's χ^2 goodness-of-fit test. The test statistics is given as

$$\sum_{i=1}^{k} \frac{(n_i - n/k)^2}{n/k}$$

and it is equal to the *dispersion index*

$$I = \frac{(k-1)s^2}{\bar{n}},$$

where $\bar{n} = \frac{1}{k}\sum_{i=1}^{k} n_i = n/k$ is the average number of points per quadrat and $s^2 = \frac{1}{k-1}\sum_{i=1}^{k}(n_i - \bar{n})^2$ is the sample variance. Index I has approximately χ^2-distribution with $k-1$ degrees of freedom. In order to get reasonable approximation the practical recommendation is $\bar{n} > 5$. Small values of I correspond to smaller variability than for the Poisson process (indication of regularity of the process). On the other hand, larger values I show bigger variability in the point pattern (clustering). In spatstat library this test can be found as quadrat.test.

The test based on the dispersion index is one of few cases in spatial statistics where the (asymptotic) distribution of a test statistic is tractable. The test statistic has often very complicated distribution. Therefore, simulation (Monte Carlo) tests are used. First we explain their general idea.

Suppose that we want to test the hypothesis H_0 that data correspond to a given model. Consider a suitable test statistic T and denote its estimator by $\hat{T}$. We perform M simulations from the null model H_0 and for each simulation we estimate T. We rank the estimators $\hat{T}_1, \ldots, \hat{T}_M$ from the smallest to the largest. So we obtained the ordered sample $\hat{T}_{(1)} \leq \cdots \leq \hat{T}_{(M)}$. Under the null hypothesis, $\hat{T}$ and $\hat{T}_1, \ldots, \hat{T}_M$ are i.i.d. and hence by symmetry every ranking has the same probability. In particular, the probability that $\hat{T}$ is smaller than $\hat{T}_{(q)}$ equals $q/(M+1)$. We want to test H_0 on the significance level α. We determine q such that

$$\alpha = \frac{2q}{M+1}.$$

The hypothesis is rejected if $\hat{T} \notin [\hat{T}_{(q)}, \hat{T}_{(M-q+1)}]$. This test is referred to as the *pointwise Monte Carlo test*. We can also determine the p-value of the test, $p = 2\min(p_+, p_-)$, where

$$p_+ = \frac{1 + \sum_{j=1}^{M} \mathbf{1}\{\hat{T}_j \geq \hat{T}\}}{M+1} \quad \text{and} \quad p_- = \frac{1 + \sum_{j=1}^{M} \mathbf{1}\{\hat{T}_j \leq \hat{T}\}}{M+1}.$$

Then we reject H_0 if $p < \alpha$.

In point processes we rather work with functional than numerical characteristics. Let $S(r)$ be some functional summary characteristics. For fixed r, chosen in advance independently on data, we may carry out above mentioned Monte Carlo test with $T = S(r)$. However, then we only use part of the information given by $S(r)$.

Consider a statistic $S(r)$ evaluated on the interval $[s_0, s_1]$, where $0 \leq s_0 < s_1$ are prescribed real constants. Denote by $\hat{S}(r)$ the estimator of $S(r)$ computed from data and by $\hat{S}_1(r), \ldots, \hat{S}_M(r)$ the estimators computed from M simulations. For each r we would be able to determine $\hat{S}_{(q)}(r)$ and $\hat{S}_{(M-q+1)}(r)$. By joining the values $\hat{S}_{(q)}(r)$ for different r we get so called *lower envelope*, while values $\hat{S}_{(M-q+1)}(r)$ form *upper envelope*. We can draw them by envelope with parameter global=FALSE. We have to realize that these are pointwise envelopes. They could be useful to reveal the deviations from null hypothesis. However, it would be incorrect to use them for testing. We are dealing with the problem of multiple testing. When the curve $\hat{S}(r)$ reaches outside the envelopes for some r, it doesn't mean that H_0 has to be rejected.

In order to get the exact envelope test assume that we know theoretical form of $S(r) = S_0(r)$ under the null hypothesis. Let us determine maximal absolute differences from the theoretical function:

$$D = \sup_{s_0 \leq r \leq s_1} |\hat{S}(r) - S_0(r)|, \quad D_i = \sup_{s_0 \leq r \leq s_1} |\hat{S}_i(r) - S_0(r)|, \ i = 1, \ldots, M.$$

Ordering the values $D_1, \ldots, D_M$ we obtain rank statistics $D_{(1)} \leq D_{(2)} \leq \cdots \leq D_{(M)}$. The null hypothesis is rejected if $D > D_{(M-q+1)}$, where q is chosen according to the required significance level, $\alpha = \frac{q}{M+1}$. This approach is known as *simultaneous Monte Carlo test* In spatstat it can be performed by envelope with global=TRUE. The testing procedure could be represented in the following way. Construct a band of width $2D_{(M-q+1)}$ around the function $S_0(r)$. If $\hat{S}(r)$ lies outside of this band (outside envelopes) for some $r \in [s_0, s_1]$, the hypothesis is rejected. Another possibility is to consider integral deviation measure instead of supremum one. For data and for each simulation we determine integral square deviations from the theoretical function,

$$D = \int_{s_0}^{s_1} (\hat{S}(r) - S_0(r))^2 \,\mathrm{d}r, \quad D_i = \int_{s_0}^{s_1} (\hat{S}_i(r) - S_0(r))^2 \,\mathrm{d}r, \; i = 1, \ldots, M.$$

The null hypothesis is rejected if $D > D_{(M-q+1)}$, where q is chosen according to the required significance level, $\alpha = \frac{q}{M+1}$. In this case we speak about *integral Monte Carlo test.*

We can use any of the simulation tests In order to test complete spatial randomness we can use any of the simulation test. As a statistic $S(r)$ usually F, G, J, K or L-function is used. In the same way we can test arbitrary model, from which we are able to simulate.

1.3 Estimation of model parameters

Another statistical problem is to find a suitable model that describes our data well. Consider that a distribution of a point process Φ is parametrized by a vector θ of unknown parameters. Our aim is to find an estimate of θ based on the realization of Φ in a bounded window $W \in \mathcal{B}_0^d$.

Minimum contrast method

In several cases the theoretical form of $S(r)$ is known and it can be expressed as the function of model parameters: $S(r) = S_\theta(r)$. Examples include statistics of Poisson process or pair correlation function of stationary Neyman-Scott process, which is given by

$$g(x) = 1 + \frac{1}{\lambda_p} \int p(y)p(y-x) \,\mathrm{d}y, \quad x \in \mathbb{R}^d,$$

where λ_p is the intensity of parent point process and p is the density displacement of daughter point from a parent point. If p has a parametric form (as for Thomas or Matérn cluster process), then $g(x)$ is expressed as the function of model parameters. An estimator of θ can be found analogically as by the method of moments. We put the estimator $\hat{S}(r)$ obtained from data equal to the theoretical function $S_\theta(r)$. Solving the equations $S_\theta(r) = \hat{S}(r)$ with several different values of r, we obtain the estimator of θ. If θ is k-dimensional vector, we have to take at least k distinct values $r_1, \ldots, r_k$, so that the equations $S_\theta(r_i) = \hat{S}(r_i)$, $i = 1, \ldots, k$, could have a unique solution. However, it would be more appropriate to look for θ that minimizes the deviation of $\hat{S}(r)$ from $S_\theta(r)$ over some interval $[a, b]$. Define

$$D(\theta) = \int_a^b \left| \hat{S}(r)^q - S_\theta(r)^q \right|^p w(r) \,\mathrm{d}r,$$

where $0 \leq a < b$ and $p, q > 0$ are given constants and $w(r)$ is a weight function. The estimator of θ is attained by minimizing function $D(\theta)$. This method is called *method of minimum contrast.* When the analytic expression of $S_\theta(r)$ is unknown we can approximate $S_\theta(r)$ for fixed θ by many simulations from the model. In library spatstat To compute the parameter estimator by the method of minimum contrast (with weight function identically 1) we can use the function mincontrast in the library spatstat. There, the default choice of parameters p and q is $p = 2$ and $q = 1/4$. If $S(r)$ is the K-function $= K(r)$, spatstat enables to find the estimators for some particular models of point processes using special functions lgcp.estK (log-Gaussian Cox process), matclust.estK (Matérn cluster process) and thomas.estK (Thomas point process).

Example: Let Φ be a Thomas point process with parameters λ_p (intensity of parent Poisson process), λ_c (mean number of cluster points) and σ^2 (variance of normal distribution describing the displacement of a cluster point from a parent point). Then the pair correlation function is

$$g(r) = 1 + \frac{1}{\lambda_p (4\pi\sigma^2)^{d/2}} \exp\left\{ -\frac{r^2}{4\sigma^2} \right\}, \quad r > 0.$$

We can estimate $g(r)$ by the kernel estimator with some edge correction factor. Having such estimator $\hat{g}(r)$ defined on the interval $[a,b]$, we can define the contrast function

$$D(\lambda_p,\sigma^2)=\int_a^b\left(\hat{g}(r)-1-\frac{1}{\lambda_p(4\pi\sigma^2)^{d/2}}\exp\left\{-\frac{r^2}{4\sigma^2}\right\}\right)^2\,\mathrm{d}r.$$

The method of minimum contrast requires the minimization of this integral which has to be done by numerical methods. Notice that the parameter λ_c is not appearing in the contrast function so we have to estimate it by other approaches.

Maximum likelihood method

Another approach is based on the maximum likelihood. Assume that Φ is a finite point process with density p w.r.t. the distribution Π of unit Poisson process on the bounded set $B\in\mathcal{B}_0^d$. The density is parametrized by the vector θ of unknown parameters, $p(\varphi)=p_\theta(\varphi)$. For simplicity we consider that the observation window W coincides with B. The maximum likelihood estimator of θ is obtained by maximizing the likelihood function $L(\theta)=p_\theta(\varphi)$, where φ is observed realization of the process Φ. Often it is more advantageous to maximize the log-likelihood function, $l(\theta)=\log L(\theta)$. The likelihood function is known for the Poisson process with intensity function λ_θ:

$$l(\theta)=|W|-\int_W\lambda_\theta(x)\,\mathrm{d}x+\sum_{x\in\varphi}\log\lambda_\theta(x).$$

As we already mentioned in Subsection 1.1, in homogeneous case ($\lambda_\theta(x)=\lambda$) the argument of maxima is $\lambda=\varphi(W)/|W|$. For inhomogeneous Poisson point process the maximum likelihood estimator is not analytically tractable and we have to proceed to numerical algorithms (e.g. Newton-Raphson method) for maximization of likelihood function.

For other processes than Poisson the normalizing constant is typically given by complicated integral which is impossible to compute explicitly. In that case we can use Monte Carlo methods. Let the density of the point process have the form $p_\theta(\varphi)=h_\theta(\varphi)/c_\theta$, where h_θ is known function and $c_\theta=\mathbb{E}h_\theta(\Phi_P)$ is unknown normalizing constant (Φ_P is Poisson point process on B with unit intensity). Then $l(\theta)=\log h_\theta(\varphi)-\log c_\theta$. It will be more advantageous to maximize likelihood ratio w.r.t. some fixed parameter θ_0,

$$l(\theta)-l(\theta_0)=\log\frac{p_\theta(\varphi)}{p_{\theta_0}(\varphi)}=\log\frac{h_\theta(\varphi)}{h_{\theta_0}(\varphi)}-\log\frac{c_\theta}{c_{\theta_0}}.$$

For the first term we have an analytic expression while the second term may be approximated by MCMC (Markov Chain Monte Carlo) methods. The ratio of normalizing constants could be written as

$$\begin{aligned}\frac{c_\theta}{c_{\theta_0}}&=\frac{1}{c_{\theta_0}}\int h_\theta(\varphi)\,\Pi(\mathrm{d}\varphi)=\int\frac{h_\theta(\varphi)}{h_{\theta_0}(\varphi)}\frac{h_{\theta_0(\varphi)}}{c_{\theta_0}}\,\Pi(\mathrm{d}\varphi)\\&=\int\frac{h_\theta(\varphi)}{h_{\theta_0}(\varphi)}p_{\theta_0}(\varphi)\,\Pi(\mathrm{d}\varphi)=\int\frac{h_\theta(\varphi)}{h_{\theta_0}(\varphi)}\,\Pi_{\theta_0}(\mathrm{d}\varphi)=\mathbb{E}_{\theta_0}\frac{h_\theta(\Phi)}{h_{\theta_0}(\Phi)},\end{aligned}$$

where Π_{θ_0} is the distribution of Φ with density p_{θ_0} (i.e. true parameter is θ_0) and $\mathbb{E}_{\theta_0}$ is the expectation w.r.t. this distribution. Here we assume that $h_{\theta_0}(\varphi)=0$ implies $h_\theta(\varphi)=0$ and use the convention $0/0=1$. There exist different MCMC algorithms for generating process with distribution Π_{θ_0}. They are based on the construction of Markov chain $\{\Phi^{(n)}\}$, whose limiting distribution is given by the density p_{θ_0} w.r.t. the distribution Π of point process Φ_P. Replacing the expectation $\mathbb{E}\frac{h_\theta(\Phi)}{h_{\theta_0}(\Phi)}$ by its sample mean we get the approximation of log-likelihood ratio

$$l_{\theta_0,n}(\theta)=\log\frac{h_\theta(\varphi)}{h_{\theta_0}(\varphi)}-\log\frac{1}{n}\sum_{i=0}^{n-1}\frac{h_\theta(\Phi^{(i)})}{h_{\theta_0}(\Phi^{(i)})}.$$

This approximation is called *importance sampling approximation*. Maximization of $l_{\theta_0,n}(\theta)$ gives MCMC approximation $\hat{\theta}_n$ of maximum likelihood estimator $\hat{\theta}$ of parameter θ. This approximation is usable if θ_0 is close to $\hat{\theta}$. As θ_0 we usually take some rough estimator obtained by a simpler and less effective method. Whole procedure can be iteratively repeated. There are alternative approximations, the details can be found in [7], Subsections 8.2.4. and 8.2.5.

Maximum pseudolikelihood method

Since the likelihood function is often complicated, another strategy to estimate the model parameters is based on approximation of the likelihood function by some simpler variant.

Definition 1. Let Φ be a finite point process on $B \in \mathcal{B}_0^d$ with Papangelou conditional intensity $\lambda_\theta^*(x, \varphi)$, where θ is the vector of unknown parameters. A realization φ of Φ is observed in the window W. We assume that W coincides with B. We define *pseudolikelihood* by the relation

$$\mathrm{PL}(\theta) = \exp\left\{ |W| - \int_W \lambda_\theta^*(x, \varphi)\, \mathrm{d}x \right\} \prod_{x \in \varphi} \lambda_\theta^*(x, \varphi \setminus \{x\}).$$

The estimator $\hat{\theta}$ that maximizes $\mathrm{PL}(\theta)$ is called *maximum pseudolikelihood estimator* of θ.

Remark 1. For Poisson point process $\lambda^*(x, \varphi) = \lambda(x)$ and thus the pseudolikelihood coincides with the likelihood.

Example: Strauss point process has Papangelou conditional intensity

$$\lambda^*(x, \varphi) = \beta \gamma^{t_R(x, \varphi)},$$

where $t_R(x, \varphi) = \sum_{y \in \varphi} \mathbf{1}_{[0 < \|x - y\| \leq R]}$. Unknown parameters are $\beta > 0$, $0 \leq \gamma \leq 1$ and $R > 0$. The logarithm of pseudolikelihood is

$$\begin{aligned} \log \mathrm{PL}(\beta, \gamma, R) &= |W| - \int_W \beta \gamma^{t_R(x, \varphi)}\, \mathrm{d}x + \sum_{x \in \varphi} (\log \beta + t_R(x, \varphi \setminus \{x\}) \log \gamma) \\ &= |W| - \int_W \beta \gamma^{t_R(x, \varphi)}\, \mathrm{d}x + \varphi(W) \log \beta + 2 S_R(\varphi) \log \gamma, \end{aligned}$$

where $S_R(\varphi) = \sum_{\{x,y\} \subseteq \varphi} \mathbf{1}_{[0 < \|x - y\| \leq R]}$. If we put the derivatives w.r.t. β and γ equal to zero, we get the equations

$$\begin{aligned} \varphi(W) &= \beta \int_W \gamma^{t_R(x, \varphi)}\, \mathrm{d}x, \\ 2 S_R(\varphi) &= \beta \int_W t_R(x, \varphi) \gamma^{t_R(x, \varphi)}\, \mathrm{d}x. \end{aligned}$$

The parameter R is considered to be known and we search for the solution numerically. In this way the estimators of β and γ are obtained. We realize that $t_R(x, \varphi)$ take only non-negative integer values and denote

$$m_k = \int_W \mathbf{1}_{[t(x, \varphi) = k]}\, \mathrm{d}x, \quad k \in \mathbb{N}_0.$$

Then our system of equations has the form

$$\begin{aligned} \varphi(W) &= \beta \sum_{k=0}^{\infty} \gamma^k m_k, \\ 2 S_R(\varphi) &= \beta \sum_{k=0}^{\infty} k \gamma^k m_k. \end{aligned}$$

The advantage of assuming R known is that the Papangelou conditional intensity has log-linear form. We can choose several different values $R_1, \ldots, R_K$ of parameter R. For each value we calculate maximum pseudolikelihood estimates of β and γ. Then we determine such R_i, $i = 1, \ldots, K$, for which the pseudolikelihood is the largest. This value is taken as the estimator of R.

Second-order composite likelihood

The maximum pseudolikelihood method belongs to a more general class of statistical methods that are based on so called *composite likelihood*. These methods are used when the maximum likelihood method is computationally very demanding or inaccessible. Composite likelihood is a function obtained by multiplying a collection of likelihoods of simpler components. These components may not be independent. The particular form depends on the context. In the setting of point processes it was suggested to consider the product over the contributions of individual points or pairs of points.

Let Φ be a stationary point process on $\mathbb{R}^d$ with second-order product density $\lambda_\theta^{(2)}$ that is parametrized by vector θ. From stationarity it follows that $\lambda_\theta^{(2)}(x,y) = \lambda_\theta^{(2)}(x-y)$. Again we assume that Φ is observed in the window W. Then the density of pairs of points in W is

$$f_\theta(x,y) = \frac{\lambda_\theta^{(2)}(x-y)}{\int_W \int_W \lambda_\theta^{(2)}(u-v)\,\mathrm{d}u\,\mathrm{d}v}, \quad x,y \in W.$$

Of course, the distinct pairs of points are not independent. However, we consider the product of the densities $f_\theta(x,y)$ over all observed pairs. After taking the logarithm, we have

$$\log \mathrm{CL}(\theta) = \sum_{X,Y\in\Phi\cap W}^{\neq} \left[\log \lambda_\theta^{(2)}(X-Y) - \log \int_W \int_W \lambda_\theta^{(2)}(u-v)\,\mathrm{d}u\,\mathrm{d}v\right].$$

For practical purposes we disregard pairs at larger distances because for them the interactions are typically low. Therefore, we do not lose much information if we omit them. Moreover, in this way we reduce computational complexity and variability of the resulting estimator. Let us choose $R > 0$ and work with pairs of points in the distance smaller than R. We get the density

$$f_\theta(x,y) = \frac{\lambda_\theta^{(2)}(x-y)\mathbf{1}\{\|X-Y\| < R\}}{\int_W \int_W \lambda_\theta^{(2)}(u-v)\mathbf{1}\{\|u-v\| < R\}\,\mathrm{d}u\,\mathrm{d}v}, \quad x,y \in W,$$

and the logarithm of composite likelihood

$$\log \mathrm{CL}(\theta) = \sum_{X,Y\in\Phi\cap W : \|X-Y\|<R}^{\neq} \left[\log \lambda_\theta^{(2)}(X-Y) - \log \int_W \int_W \lambda_\theta^{(2)}(u-v)\mathbf{1}\{\|u-v\| < R\}\,\mathrm{d}u\,\mathrm{d}v\right].$$

The estimator of θ is obtained by maximizing this function. Note that in the expression of f_θ or $\log \mathrm{CL}(\theta)$ we are allowed to replace the product density $\lambda_\theta^{(2)}$ by the pair correlation function $g_\theta = \lambda_\theta^{(2)}/\lambda^2$, where λ is the intensity of Φ.

As opposed to previous two subsections we are now working with stationary point processes. The composite likelihood method is used mainly for Cox point processes where we often have analytic form of the second-order product density. If Φ is a stationary Cox point process with driving intensity function Z having the distribution depending on θ, then $\lambda_\theta^{(2)}(x-y) = \mathbb{E}Z(x)Z(y)$. Next we will demonstrate another method suitable for Cox point processes. This method is based on another second-order characteristics.

Palm likelihood

Let Φ be a stationary point process on $\mathbb{R}^d$ with intensity λ and second-order product density $\lambda^{(2)}$. Then we can write $\lambda^{(2)}(y-x) = \lambda\lambda_o(y-x)$, where λ_o is called *Palm intensity*. The second-order factorial moment measure can be expressed from the Campbell theorem as

$$\alpha^{(2)}(A\times B) = \mathbb{E}\sum_{X,Y\in\Phi}^{\neq} \mathbf{1}_{[X\in A, Y\in B]} = \int_A \int_B \lambda^{(2)}(y-x)\,\mathrm{d}y\,\mathrm{d}x = \lambda \int_A \int_{B-x} \lambda_o(u)\,\mathrm{d}u\,\mathrm{d}x.$$

On the other hand, by the Campbell-Mecke theorem we have

$$\alpha^{(2)}(A\times B) = \mathbb{E}\sum_{X\in\Phi\cap A} \Phi(B\setminus\{X\}) = \lambda \int_A \int \varphi(B-x)\,P_o^!(\mathrm{d}\varphi)\,\mathrm{d}x.$$

Comparing these two expression we find out that

$$E_o^! \Phi(B) = \int_B \lambda_o(u)\,\mathrm{d}u, \quad B \in \mathcal{B}^d,$$

i.e. λ_o is the intensity function of the reduced Palm distribution of Φ. Now it is also clearer why λ_o is called the Palm intensity. Realize that it is a second-order characteristics. For Poisson point process this function is constant. However, generally λ_o is not constant.

We will consider the point process of differences of observed points of Φ in W with the distance smaller than R, i.e.

$$\Phi_R = \{Y - X : X \neq Y \in \Phi \cap W, \|Y - X\| < R\}.$$

Obviously it is a point process in the ball $b(o, R)$. Its intensity measure is (by Campbell theorem)

$$\begin{aligned}\mathbb{E}\Phi_R(A) &= \mathbb{E} \sum_{X,Y \in \Phi \cap W}^{\neq} \mathbf{1}_{[Y-X \in A]} = \int_W \int_W \mathbf{1}_{[y-x\in A]} \lambda^{(2)}(y-x)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int \int \mathbf{1}_{[x \in W, x+u \in W]} \mathbf{1}_{[u \in A]} \lambda^{(2)}(u)\,\mathrm{d}x\,\mathrm{d}u = \lambda \int_A |W \cap (W-u)| \lambda_o(u)\,\mathrm{d}u.\end{aligned}$$

Therefore, the intensity function of Φ_R is

$$\lambda_R(u) = \lambda\lambda_o(u)|W \cap (W-u)|, \quad u \in b(o,R).$$

We assume that a parametric form $\lambda_o^\theta(u)$ of the Palm intensity is given. We want to estimate the vector θ of unknown parameters. To do this we consider Φ_R as an inhomogeneous Poisson process with intensity function $\lambda_R(u)$, which is approximated so that unknown true intensity is replaced by the observed intensity $\Phi(W)/|W|$ and the term $|W \cap (W-u)|$ is replaced by $|W|$, which is a reasonable approximation for R substantially smaller than the window side. Altogether we approximate $\lambda_R(u)$ as $\Phi(W)\lambda_o(u)$. The likelihood function is approximated by the likelihood function for the Poisson process with this approximated intensity function. Such likelihood function is referred to as the *Palm likelihood*. It means that the logarithm of Palm likelihood is

$$\log L_P(\theta) = \sum_{X,Y \in \Phi \cap W}^{\neq} \mathbf{1}_{[\|Y-X\|<R]} \log \Phi(W)\lambda_o^\theta(Y-X) + |b(o,R)| - \Phi(W) \int_{b(o,R)} \lambda_o^\theta(u)\,\mathrm{d}u.$$

Alternative way how to get to the Palm likelihood is to consider point processes

$$\Phi_X = \{Y - X : X \neq Y \in \Phi\}, \quad X \in \Phi \cap W,$$

which are inhomogeneous point processes with intensity function λ_o. We ignore interactions in the processes $\Phi_X \cap b(o,R)$ and approximate them by inhomogeneous Poisson processes whose log-likelihoods are

$$\sum_{Y \in \Phi} \mathbf{1}_{[0<\|X-Y\|<R]} \log \lambda_o^\theta(Y-X) + |b(o,R)| - \int_{b(o,R)} \lambda_o^\theta(u)\,\mathrm{d}u.$$

Now we regard Φ_X, $X \in \Phi \cap W$, as independent identically distributed point processes, and we ignore edge effects. Then the logarithmic Palm likelihood has the form

$$\log L_P(\theta) = \sum_{X,Y \in \Phi \cap W} \mathbf{1}_{[0<\|X-Y\|<R]} \log \lambda_o^\theta(Y-X) + \Phi(W)|b(o,R)| - \Phi(W) \int_{b(o,R)} \lambda_o^\theta(u)\,\mathrm{d}u,$$

which differs from the previous expression only by a constant.

Takacs-Fiksel method

In order to find the maximum likelihood estimate one solves the equation $l'(\theta) = 0$. Method of moments is based on the relation $S_\theta(r) - \hat{S}(r) = 0$. Both these approaches could be included in the notion of estimating equations.

Definition 2. Let Φ be a point process with distribution Q_θ depending on unknown parameter $\theta \in \Theta$. Consider a function $\psi : \Theta \times \mathcal{N} \to \mathbb{R}$ such that $\mathbb{E}_\theta \psi(\theta, \Phi) = 0$ for each $\theta \in \Theta$. Here, $\mathbb{E}_\theta$ denotes the expectation w.r.t. Q_θ. For given realization φ the equation $\psi(\theta, \varphi) = 0$ is called *unbiased estimating equation*. By different choices of ψ we obtain a system of equations. Its solution $\hat{\theta}$ is used as an estimate of θ based on φ.

Besides the method of moments and maximum likelihood (or pseudolikelihood) method, another example of estimating equations for point process models is given by the Takacs-Fiksel method. This method is based on the Georgii-Nguyen-Zessin identity

$$\mathbb{E} \sum_{X \in \Phi} h(X, \Phi \setminus \{X\}) = \int_{\mathbb{R}^d} \mathbb{E} h(x, \Phi) \lambda^*(x, \Phi) \, \mathrm{d}x, \tag{3}$$

where λ^* is the conditional intensity of Φ.

In the case of finite point process with density p w.r.t. the distribution of unit Poisson point process Φ_P on the set $B \in \mathcal{B}_0^d$, λ^* is the Papangelou conditional intensity.

Assume that we know the parametric form of the conditional intensity $\lambda_\theta^*(x, \varphi)$. We define

$$\psi_h(\theta, \varphi) = \sum_{x \in \varphi \cap W} h(x, \varphi \setminus \{x\}) - \int_W h(x, \varphi) \lambda_\theta^*(x, \varphi) \, \mathrm{d}x$$

for arbitrary function h. Then by the Georgii-Nguyen-Zessin identity we have $\mathbb{E}_\theta \psi_h(\theta, \Phi) = 0$. The estimator of θ is obtained as the solution of unbiased estimating equations $\psi_h(\theta, \varphi) = 0$. Similarly as in the method of minimum contrast, we can choose more functions h than the number of unknown parameters. For example, if we have k function $h_1, \dots, h_k$, we may search for θ which minimizes

$$\sum_{i=1}^{k} \psi_{h_i}(\theta, \varphi)^2.$$

Remark 2. When we obtain the estimator $\hat{\theta}$ as the solution of unbiased estimating equation, it doesn't mean that it is an unbiased estimator of θ.

For some natural choices of h it may be impossible to determine $\psi_h(\theta, \varphi)$ only from the observation φ in a bounded window W. The problems with edge effects may arise. Then instead of $\psi_h(\theta, \varphi)$ we can take the estimate $\hat{\psi}_h(\theta, \varphi)$ which considers corrections on edge effects. For example, in the case of stationary point process put $h(x, \varphi) = \varphi(b(x, r)) \mathbf{1}_{[x \in W]} / |W|$. Then the expectation of the first term in $\psi_h(\theta, \Phi)$, i.e. the left-hand side in (3), is equal to $\lambda^2 K(r)$. However, we are not able to determine the first term in $\psi_h(\theta, \varphi)$ only from information inside W. The solution is to replace it by some estimator of $\lambda^2 K(r)$, e.g. the estimator with translation edge correction.

Example: Consider Strauss point process and assume that the parameter R is known. Our goal is to estimate $\theta = (\beta, \gamma)$. The choice $h_1(x, \varphi) = 1$ gives

$$\psi_{h_1}(\theta, \varphi) = \varphi(W) - \beta \int_W \gamma^{t_R(x, \varphi)} \, \mathrm{d}x.$$

Taking $h_2(x, \varphi) = t_R(x, \varphi)$ we get

$$\psi_{h_2}(\theta, \varphi) = 2 S_R(\varphi) - \beta \int_W t_R(x, \varphi) \gamma^{t_R(x, \varphi)} \, \mathrm{d}x.$$

Note that we have obtained the same two equations with two unknown parameters and as in the case of maximum pseudolikelihood method.

1.4 Model diagnostics

In order to verify that the fitted parametric model is appropriate we can exploit the idea of Monte Carlo tests. If we are able to simulate from our fitted model, then we can determined some summary characteristics for each simulated realization of the model. The results from simulations could be compared with the characteristics estimated from data. This should not show any substantial deviations when the parameters of the model are determined correctly. The problems with this approach will begin to exhibit for more general inhomogeneous point processes where we would need appropriate summary characteristics.

Now let us examine how we can use the generalization of residuals from the classical linear regression models to the context of point processes. Generally, the residuals are differences between observed and fitted values. If the fitted model is correct, the residuals should fluctuate around zero. On the contrary, large deviations from zero may indicate what is wrong in the fitted model (e.g. incorrectly estimated trend or interactions).

Definition 3. Let Φ be a point process with conditional intensity λ^*. For some non-negative measurable function h we define *h-weighted innovation* as the signed random measure

$$I(B,h,\lambda^*) = \sum_{X\in\Phi\cap B} h(X,\Phi\setminus\{X\}) - \int_B h(x,\Phi)\lambda^*(x,\Phi)\,\mathrm{d}x.$$

According to Georgii-Nguyen-Zessin identity (3) we have $\mathbb{E}I(B,h,\lambda^*)=0$ for any $B\in\mathcal{B}^d$.

Example: Let Φ be Poisson point process with intensity function λ and consider the following three choices of h: $h(x,\varphi)=1$, $h(x,\varphi)=1/\lambda^*(x,\varphi)$ and $h(x,\varphi)=1/\sqrt{\lambda^*(x,\varphi)}$. Since $\lambda^*(x,\varphi)=\lambda(x)$, we get

$$\begin{aligned}
I(B,1,\lambda) &= \Phi(B) - \int_B \lambda(x)\,\mathrm{d}x,\\
I(B,1/\lambda^*,\lambda) &= \sum_{X\in\Phi\cap B}\frac{1}{\lambda(X)} - |B|,\\
I(B,1/\sqrt{\lambda^*},\lambda) &= \sum_{X\in\Phi\cap B}\frac{1}{\sqrt{\lambda(X)}} - \int_B\sqrt{\lambda(x)}\,\mathrm{d}x.
\end{aligned}$$

It is easy to verify that $\mathbb{E}I(B,h,\lambda)=0$ by direct computation from Campbell theorem. For the variances we have

$$\begin{aligned}
\operatorname{var} I(B,1,\lambda) &= \int_B \lambda(x)\,\mathrm{d}x,\\
\operatorname{var} I(B,1/\lambda^*,\lambda) &= \int_B \frac{1}{\lambda(x)}\,\mathrm{d}x,\\
\operatorname{var} I(B,1/\sqrt{\lambda^*},\lambda) &= |B|.
\end{aligned}$$

These relations follow directly from the following lemma.

Lemma 1. *Let Φ be Poisson point process on $\mathbb{R}^d$ with intensity measure Λ. Then for arbitrary non-negative measurable function f,*

$$\operatorname{var}\sum_{X\in\Phi} f(X) = \int f(x)^2\,\Lambda(\mathrm{d}x).$$

Proof: From Campbell's theorem,

$$\mathbb{E}\sum_{X\in\Phi} f(X) = \int f(x)\,\Lambda(\mathrm{d}x).$$

The second moment could rewritten using second-order Campbell's theorem. Moreover, we make use of the fact that the second-order factorial moment measure of the Poisson point process is $\Lambda\times\Lambda$,

$$\begin{aligned}
\mathbb{E}\left(\sum_{X\in\Phi} f(X)\right)^2 &= \mathbb{E}\sum_{X,Y\in\Phi} f(X)f(Y) = \mathbb{E}\sum_{X\in\Phi} f(X)^2 + \mathbb{E}\sum_{X,Y\in\Phi}^{\neq} f(X)f(Y)\\
&= \int f(x)^2\,\Lambda(\mathrm{d}x) + \int\int f(x)f(y)\,\Lambda(\mathrm{d}x)\,\Lambda(\mathrm{d}y) = \int f(x)^2\,\Lambda(\mathrm{d}x) + \left(\int f(x)\,\Lambda(\mathrm{d}x)\right)^2.
\end{aligned}$$

From this we already obtained the assertion.

□

Assume that the conditional intensity $\lambda^*_\theta(x,\varphi)$ depends on the parameter θ. Furthermore, suppose that we found the estimator $\hat\theta$ (e.g. by one of the methods from Subsection 1.3) based on the observation of Φ in the window $W \in \mathcal{B}^d_0$. Then the estimator of conditional intensity is $\widehat{\lambda^*}(x,\varphi) = \lambda^*_{\hat\theta}(x,\varphi)$. In the definition of innovation we admit that h depends on $\hat\theta$. Put $\hat h(x,\varphi) = h_{\hat\theta}(x,\varphi)$.

Definition 4. A random signed measure

$$R(B,\hat h,\hat\theta) = I(B,\hat h,\widehat{\lambda^*}) = \sum_{X\in\Phi\cap B} h_{\hat\theta}(X,\Phi\setminus\{X\}) - \int_B h_{\hat\theta}(x,\Phi)\lambda^*_{\hat\theta}(x,\Phi)\,\mathrm{d}x$$

is called *h-weighted residual measure.*

Since $\mathbb{E}I(B,h,\lambda^*) = 0$, we expect that $R(B,\hat h,\hat\theta)$ is around zero when the model with $\hat\theta$ is correct. Note that the expectation $\mathbb{E}R(B,\hat h,\hat\theta)$ does not have to be zero but it should be approximately zero when the model is true. Regions B with extreme values of $R(B,\hat h,\hat\theta)$ may indicate regions of irregularity. The usual choices for h include $h = 1$ (*raw residuals*), $h = 1/\lambda^*$ (*inverse-lambda residuals*) and $h = 1/\sqrt{\lambda^*}$ (*Pearson residuals*). Residuals for these three choices may be computed by residuals.ppm in package spatstat. For $h = 1$ we have

$$R(B,1,\hat\theta) = \Phi(B) - \int_B \lambda^*_{\hat\theta}(x,\Phi)\,\mathrm{d}x.$$

It means that raw residual measure is given as the difference of an atomic measure with atoms in the observed points and a measure with density $\lambda^*_{\hat\theta}(x,\Phi)$ w.r.t. Lebesgue measure.

Example: Consider a stationary Poisson point process Φ with intensity λ, which could be estimated as $\hat\lambda = \Phi(W)/|W|$. Then (provided that $\Phi(W) > 0$)

$$R(B,1,\hat\lambda) = \Phi(B) - \Phi(W)\frac{|B|}{|W|},$$

$$R(B,1/\widehat{\lambda^*},\hat\lambda) = |W|\frac{\Phi(B)}{\Phi(W)} - |B|,$$

$$R(B,1/\sqrt{\widehat{\lambda^*}},\hat\lambda) = \Phi(B)\sqrt{\frac{|W|}{\Phi(W)}} - |B|\sqrt{\frac{\Phi(W)}{|W|}}.$$

It can be shown that the expectations of these three h-weighted residual measures are 0. We can also notice that $R(W,1,\hat\lambda) = R(W,1/\widehat{\lambda^*},\hat\lambda) = R(W,1/\sqrt{\widehat{\lambda^*}},\hat\lambda) = 0$. This corresponds to the situation in the classical linear regression, where the sum of all residuals is 0.

For the graphical representation of the residuals it is convenient to perform kernel smoothing.

Definition 5. Let k be a probability density on $\mathbb{R}^d$. The realization of a point process Φ is observed in $W \in \mathcal{B}^d_0$. We have constructed the estimator $\hat\theta$ of θ. Define *smoothed residual field* by the relation

$$S(x) = e(x)\int_W k(x-y)\,R(\mathrm{d}y,\hat h,\hat\theta),$$

where

$$e(x) = \left(\int_W k(x-y)\,\mathrm{d}y\right)^{-1}$$

is the edge correction.

Remark 3. For $h = 1$ we have

$$S(x) = e(x)\sum_{Y\in\Phi\cap W} k(x-Y) - e(x)\int_W k(x-y)\lambda^*_{\hat\theta}(y,\Phi)\,\mathrm{d}y.$$

2. Statistics of marked point processes

Let Φ_{m} be a marked point process on $\mathbb{R}^d$ with mark space $\mathbb{M}$. The corresponding unmarked point process is denoted by Φ.

2.1 Estimation of summary characteristics

We will assume that a marked point process Φ_{m} is observed in a bounded window $W \in \mathcal{B}_0^d$. The estimates of summary characteristics are mostly either straightforward from the definition or it is enough to suitably modify estimator that were defined for point processes (Subsection 1.1).

First consider the case of qualitative marks $\mathbb{M} = \{1, \dots, k\}$. It means that Φ_{m} is k-dimensional point process $(\Phi_1, \dots, \Phi_k)$. Then for cross G-function $G_{ij}(r) = P_o^{!i}(D_j(o) \le r)$ and condensed G-function $G_{i\cdot}(r) = P_o^{!i}(D(o) \le r)$, $r \ge 0$, we may use, for example, Kaplan-Meier estimator:

$$\widehat{G_{ij}}(r) = 1 - \prod_{s \le r} \left(1 - \frac{\#\{X \in \Phi_i : e_j(X) = s, e_j(X) \le c(X)\}}{\#\{X \in \Phi_i : e_j(X) \ge s, c(X) \ge s\}} \right),$$

$$\widehat{G_{i\cdot}}(r) = 1 - \prod_{s \le r} \left(1 - \frac{\#\{X \in \Phi_i : e(X) = s, e(X) \le c(X)\}}{\#\{X \in \Phi_i : e(X) \ge s, c(X) \ge s\}} \right),$$

where $c(x) = d(x, \partial W)$ is the distance of x from the window boundary, $e_j(x) = d(x, \Phi_j \setminus \{x\})$ is the distance of x from the nearest point of Φ_j and $e(x) = d(x, \Phi \setminus \{x\})$ is the distance of x from the nearest point of the process (regardless of the mark). In library spatstat we can obtain these estimators using Gcross and Gdot.

The cross K-function K_{ij} is defined by the relation

$$\lambda_j K_{ij}(r) = \mathbb{E}_o^{!i} \Phi_j(b(o, r))$$

and the condensed K-function $K_{i\cdot}$ by the relation

$$\lambda K_{i\cdot}(r) = \mathbb{E}_o^{!i} \Phi(b(o, r)).$$

Here λ_i is the intensity Φ_i and its natural estimator is $\widehat{\lambda_i} = \Phi_i(W)/|W|$. We mention the estimates of cross (Kcross) and condensed (Kdot) K-function in the form with translation correction (other edge corrections could be considered as well):

$$\widehat{K_{ij}}(r) = \frac{1}{\widehat{\lambda_i}\widehat{\lambda_j}} \sum_{X \in \Phi_i \cap W, Y \in \Phi_j \cap W}^{\neq} \frac{\mathbf{1}_{[\|X - Y\| \le r]}}{|W \cap (W + X - Y)|},$$

$$\widehat{K_{i\cdot}}(r) = \frac{1}{\widehat{\lambda_i}\widehat{\lambda}} \sum_{X \in \Phi_i \cap W, Y \in \Phi \cap W}^{\neq} \frac{\mathbf{1}_{[\|X - Y\| \le r]}}{|W \cap (W + X - Y)|}.$$

Notice that this estimator of the cross K-function satisfies $\widehat{K_{ij}}(r) = \widehat{K_{ji}}(r)$.

For marked point processes with quantitative marks we first deal with the estimation of a non-normalized f-mark correlation function. It is defined as $\kappa_f(r) = \frac{\lambda_f^{(2)}(r)}{\lambda^{(2)}(r)}$, where $\lambda_f^{(2)}(r)$ is the density of second-order f-weighted factorial moment measure

$$\alpha_f^{(2)}(B_1 \times B_2) = \mathbb{E} \sum_{(X_1, M_1), (X_2, M_2) \in \Phi_{\mathrm{m}}}^{\neq} \mathbf{1}_{[X_1 \in B_1, X_2 \in B_2]} f(M_1, M_2).$$

The kernel estimator of $\lambda_f^{(2)}(r)$ is

$$\widehat{\lambda_f^{(2)}}(r) = \sum_{X, Y \in \Phi \cap W}^{\neq} \frac{f(M(X), M(Y)) k_b(\|X - Y\| - r)}{\sigma_d r^{d-1} |W \cap (W + X - Y)|},$$

while analogous estimator of the second-order product density $\lambda^{(2)}(r)$ is

$$\widehat{\lambda^{(2)}}(r) = \sum_{X, Y \in \Phi \cap W}^{\neq} \frac{k_b(\|X - Y\| - r)}{\sigma_d r^{d-1} |W \cap (W + X - Y)|},$$

where k_b is a kernel function with bandwidth b. In both cases we have used translation correction on the edge effects. Then the non-normalized f-mark correlation function can be estimated as

$$\widehat{\kappa_f}(r) = \frac{\widehat{\lambda_f^{(2)}}(r)}{\widehat{\lambda^{(2)}}(r)}, \quad r > 0.$$

The function $\kappa_f(r)$ can be also expressed by two-point mark distribution, $\kappa_f(r) = \mathbb{E}_{or} f(M(o), M(r))$. This conditional expectation could be estimated by arithmetic mean of f-values for marks corresponding to the points at distance r. The number of pairs exactly at distance r will be usually small. Therefore, we take $\varepsilon > 0$ and put

$$\widehat{\kappa_f}(r) = \frac{1}{N_f(\varepsilon)} \sum_{X,Y \in \Phi \cap W : |\|X-Y\|-r| < \varepsilon/2}^{\neq} f(M(X), M(Y)),$$

where $N_f(\varepsilon) = \#\{X, Y \in \Phi \cap W : |\|X - Y\| - r| < \varepsilon/2\}$.

Furthermore, we estimate f-mark correlation function $k_f(r) = \frac{\kappa_f(r)}{c_f}$ by

$$\widehat{k_f}(r) = \frac{\widehat{\kappa_f}(r)}{\widehat{c_f}}, \quad r > 0,$$

where

$$\widehat{c_f} = \frac{1}{\Phi(W)^2} \sum_{X,Y \in \Phi \cap W} f(M(X), M(Y))$$

is the estimator of $c_f = \int \int f(m_1, m_2)\, \mathbb{Q}(\mathrm{d}m_1)\, \mathbb{Q}(\mathrm{d}m_2)$. Denote $\mu = \mathbb{E}M_0 = \int m\, \mathbb{Q}(\mathrm{d}m)$ the mean typical mark. Then for $f(m_1, m_2) = m_1 m_2$ we have $c_f = \mu^2$ and k_f is denoted as k_{mm}. For $f(m_1, m_2) = m_1$, $c_f = \mu$ and k_f is denoted as $k_{m\cdot}$. The values $k_{mm}(r)$ or $k_{m\cdot}(r)$ larger than 1 indicated mutual stimulation in the distance r. On the other hand, the values smaller than 1 correspond to the inhibition.

The functions $\kappa_f(r)$ and $k_f(r)$ are examples of non-cumulative summary characteristics. A cumulative analogue is the f-weighted K-function $K_f(r)$, which generalizes K-function for point processes in the following way

$$\lambda K_f(r) = \frac{1}{c_f} \mathbb{E}_o^! \sum_{(X, M(X)) \in \Phi_\mathrm{m}} f(M(o), M(X)) \mathbf{1}_{[X \in b(o,r)]},$$

where λ is the intensity of a stationary process Φ_m. In particular, for $f(m_1, m_2) = m_1 m_2$ we get the function $K_{mm}(r)$, for $f(m_1, m_2) = m_1$ the function $K_{m\cdot}(r)$ and for $f(m_1, m_2) = m_2$ the function $K_{\cdot m}(r)$. An unbiased estimator of $\lambda^2 c_f K_f(r)$ using translation correction has the form

$$\widehat{\lambda^2 c_f K_f}(r) = \sum_{X,Y \in \Phi \cap W}^{\neq} \frac{f(M(X), M(Y)) \mathbf{1}\{\|X - Y\| \le r\}}{|(W - X) \cap (W - Y)|}.$$

In order to estimate $K_f(r)$ we have to divide by estimators of λ^2 and c_f.

As a representative of numerical characteristics we consider non-normalized nearest-neighbour correlation index defined as $\bar{\nu}_f = \mathbb{E}_o^! f(M(o), M(Z_1))$, where Z_1 is the point of the process that is the closest to the origin and $M(Z_1)$ is the mark at Z_1. This index can be naturally estimated by

$$\widehat{\bar{\nu}_f} = \frac{1}{\Phi(W)} \sum_{X \in \Phi \cap W} f(M(X), M(Z_X)),$$

where $Z_X \in \Phi$ is the nearest neighbour of X. A natural estimator of the normalized nearest-neighbour correlation index $\bar{n}_f = \frac{\bar{\nu}_f}{c_f}$ is obtained as

$$\widehat{\bar{n}_f} = \frac{\widehat{\bar{\nu}_f}}{\widehat{c_f}}.$$

For $f(m_1, m_2) = m_1 m_2$ the values $\bar{n}_f > 1$ indicate mutual stimulation between neighbours.

2.2 Tests of independence

Statistical analysis of a marked point process mostly starts with the test of hypothesis of independent marks. If the marks may be considered to be independent, then we can use the methods developed for independent data. First we will deal with test of independent marks. Afterwards we mention some approaches for testing independence of marks and locations. We pursue non-parametric approach and use simulation tests, whose principle was explained in Subsection 1.2.

Testing independent marking

First consider a two-dimensional point process $\Phi_{\mathrm{m}} = (\Phi_1, \Phi_2)$. The null hypothesis of independent marks may have two different interpretations:

1. independent marking (random labelling) – to the points of Φ independently randomly either mark 1 or mark 2 is assigned,
2. random superposition – two independent point processes Φ_1 and Φ_2 form a bivariate point process.

The first situation is an example of *posterior marking* – we describe how the marks were created conditionally on given locations of points. This is appropriate model when the points are tree locations and the trees could be affected by some disease or catastrophe (mark 1) or not (mark 2). In the second situation we have *prior marking* – a marked point process is formed by certain mechanism, namely union of two independent populations.

For testing the hypothesis that Φ_{m} is independently marked point process a method of *random allocation* is used. We fix the locations and create new marks by random permutation of observed marks (rlabel in spatstat). We generate M such permutations and carry out the corresponding Monte Carlo test. Under the hypothesis of independent marking,

$$\begin{aligned} K(r) &= K_{11}(r) = K_{22}(r) = K_{12}(r) = K_{1\cdot}(r), \\ g(r) &= g_{11}(r) = g_{22}(r) = g_{12}(r), \\ G(r) &= G_{1\cdot}(r), \\ J(r) &= J_{1\cdot}(r), \end{aligned}$$

where $K(r)$, $g(r)$, $G(r)$ and $J(r)$ are functional characteristics of unmarked point process Φ. Therefore, a useful test statistic could be e.g. $S(r) = K_{1\cdot}(r) - K(r)$, which is equal to $S_0(r) = 0$ if the null hypothesis is true.

When we want to test the hypothesis of random superposition of point processes Φ_1 and Φ_2, we can use method of *random shift*. The locations of points with mark 1 are fixed. We generate M realizations of the subprocess Φ_2 so that all its points are simultaneously shifted (rshift) by a vector with prescribed length $R > 0$. For each of M realizations of a bivariate point process we calculate estimator of $S(r)$ and apply simultaneous Monte Carlo test. As a function $S(r)$ we may use one of the cross functional characteristics. Under the hypothesis of random superposition the following relations hold:

$$\begin{aligned} K_{12}(r) &= \omega_d r^d, \\ g_{12}(r) &= 1, \\ G_{12}(r) &= F_2(r), \\ J_{12}(r) &= 1. \end{aligned}$$

In the case of process with quantitative marks we can again test the hypothesis of independent marking by the method of random allocation. It means that the locations are kept fixed and the marks are assigned by permuting the observed marks (sampling without replacement). In this way all M simulations lead to the same empirical mark distribution. Another possibility is to generate marks from the empirical mark distribution (sampling with replacement). The test statistic could be one of the f-mark correlation functions. For stationary and isotropic independently marked point processes we have $k_f(r) = 1$.

Independence of marks and locations

We are going to present three methods for testing independence of marks and locations in marked point processes with quantitative marks. If the marks and locations are independent (so called geostatistical marking) we may investigate both components separately, which simplifies the statistical inference.

The first method is based on the summary characteristics $K_f(r)$, in particular special cases $K_{m\cdot}(r)$ and $K_{\cdot m}(r)$. Both these functions are equal to the K-function $K(r)$ of unmarked point process Φ if the process is geostatistically marked. The test works conditionally on the locations and it is based on the random allocation. We will generate M realizations by random permutations of marks (or by random sample from empirical mark distribution). Since the locations are fixed, the estimators of $K(r)$ will be the same for all M simulations as well as for data. We use this estimator of $K(r)$ as the statistics $S_0(r)$ in Monte Carlo test. Further, we estimate $K_{m\cdot}(r)$ or $K_{\cdot m}(r)$ from data and also from M simulations. Under the null hypothesis all these $M+1$ functions should look approximately like $S_0(r)$. By simultaneous or integral Monte Carlo test we find out whether the estimate from data significantly differs. This approach ignores correlations among marks, this can cause that the hypothesis is rejected not because of dependence between marks and locations but because of dependencies within marks.

The second method originates from the paper [12]. It is based on the fact that the functions $E(r) = \mathbb{E}_{or}M(r)$ and $V(r) = \mathbb{E}_{or}(M(o)-E(r))^2$ are constant for stationary and isotropic geostatistically marked point process If the estimates of these functions from data significantly differ from constant function, it gives evidence against the null hypothesis. Defining $E(0) = \mathbb{E}M_0$ and $V(0) = \operatorname{var} M_0$, we can perform simultaneous Monte Carlo test with the choice $S(r) = E(r) - E(0)$ or $S(r) = V(r) - V(0)$. Under the null hypothesis, $S_0(r) = 0$.

Also the third test is based on the principle of simultaneous tests. It was proposed in the paper [3]. Let $\varphi_{\mathrm{m}} = \{(x_1, m_1), \dots, (x_n, m_n)\}$ be a realization of stationary and isotropic marked point process Φ_{m} observed in the window W. Assume that the data are save in some fixed order. Let $\delta(x_i) = d(x_i, \{x_{i+1}, \dots, x_n\})$, $i = 1, \dots, n$, denote the distance of point x_i to the nearest point of the process with larger index. For given $r > 0$ we choose those points with $\delta(x_i) \leq r$. The number of selected points will be n_r. It is recommended to choose r small in comparison with distance of nearest neighbours. Since the selection of points does not depend on marks, the mean of marks of n_r points should be under the null hypothesis close to the mean of marks of arbitrary randomly selected n_r points out of n points. On the other hand, if a mark is dependent on the presence of other points in the vicinity of its location, then the means of marks selected according to the proposed criterion and the means of randomly selected marks should differ significantly. We generate M different random samples of marks and for each such sample of size n_r we compute its mean. The test itself works in the same way as the classical Monte Carlo test described in Subsection 1.2 (T is the mean of n_r marks).

3. Geostatistics

Geostatistics is a part of spatial statistics dealing with data formed by finitely many observations of a given variable in some fixed spatial locations.

The geostatistical data are modeled by a random field $\{Z(x) : x \in D\}$, where $D \subseteq \mathbb{R}^d$ has positive d-dimensional Lebesgue measure. Recall that intrinsically stationary random field satisfies the conditions $\mathbb{E}(Z(x) - Z(y)) = 0$ and $\operatorname{var}(Z(x) - Z(y)) = 2\gamma(x - y)$. The function $2\gamma(h) = \operatorname{var}(Z(x+h) - Z(x)) = \mathbb{E}(Z(x+h) - Z(x))^2$ is called variogram. Our first aim is to estimate the variogram from the observations $Z(x_1), \dots, Z(x_n)$, where $x_1, \dots, x_n \in D$ are fixed deterministic points.

3.1 Variogram estimation

Non-parametric estimators

For first idea about the variogram we can depict squares of differences of observed data $(Z(x_i) - Z(x_j))^2$ against $x_i - x_j$ or $\|x_i - x_j\|$. Such graph is called *empirical variogram cloud* and can be obtained in library geoR [10] by variog with option="cloud". This graph often does not give clear picture because the number of possible pairs $\{x_i, x_j\}$ of distinct points could be quite large, Specifically, it is $\binom{n}{2}$. More useful information is obtained by averaging the values corresponding to same difference $x_i - x_j$. Then we have the following unbiased estimator of the variogram,

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(x_i) - Z(x_j))^2, \tag{4}$$

where $N(h) = \{(x_i, x_j) : x_i - x_j = h, i, j = 1, \dots, n\}$ and $|N(h)|$ is the cardinality of $N(h)$. It is in fact the estimator obtained by the method of moments. The following properties of the estimator can be easily seen: $\hat{\gamma}(h) \geq 0$, $\hat{\gamma}(o) = 0$ and $\hat{\gamma}(h) = \hat{\gamma}(-h)$. Thus, the estimator preserves basic theoretical properties of the variogram. The symmetry of the estimator is satisfied even if $N(h) \neq N(-h)$. For small sample size or irregularly scattered points $x_1, \dots, x_n$, where the measurements are taken, the number of pairs in $N(h)$ will be very small and the estimator of $2\gamma(h)$ will have large variance. The practical recommendation is to use h, for which $|N(h)| \geq 30$. If we are unable to assure this condition, we divide (similarly as in the construction of histogram) pairs of points into several groups with similar differences $x_i - x_j$. We calculate the mean of variables $(Z(x_i) - Z(x_j))^2$ in each group. In library RandomFields we can perform this by function EmpiricalVariogram, in library gstat [9] by variogram. Another possibility is to use kernel smoothing with kernel function k_b and bandwidth b:

$$2\hat{\gamma}(h) = \frac{\sum_{i \neq j} (Z(x_i) - Z(x_j))^2 k_b(x_i - x_j - h)}{\sum_{i \neq j} k_b(x_i - x_j - h)}.$$

Both smoothed and histogram-based estimator can be computed in package geoR using variog.

The estimates based on the squared differences $(Z(x_i) - Z(x_j))^2$ are very sensitive to outlying observations because for them large values are squared making them even larger. Assume that $\{Z(x) : x \in D\}$ is a Gaussian random field. Then $(Z(x+h) - Z(x))^2$ has distribution $2\gamma(h) \cdot \chi_1^2$, which is very skewed. The fourth root is a suitable transformation that creates a distribution "close" to the normal distribution, see Figure 2. Instead of $(Z(x_i) - Z(x_j))^2$ we can thus work with $|Z(x_i) - Z(x_j)|^{1/2}$. This leads us to the robust version of the variogram estimator:

$$2\bar{\gamma}(h) = \left(\frac{1}{|N(h)|} \sum_{N(h)} |Z(x_i) - Z(x_j)|^{1/2} \right)^4 \Big/ B(h),$$

where $B(h) = 0{,}457 + 0{,}494/|N(h)|$. The fourth power is there to preserve the proper scale. This transformation breaks the unbiasedness of the estimator, and so the term $B(h)$ is added. This term represents the bias correction and ensures approximately unbiased estimator. The robust estimator is computed in package geoR by the choice estimator.type="modulus" in variog. Except of reducing the influence of outliers another advantage of the robust estimator is that the summands are less correlated than in the case of classical estimator (4).

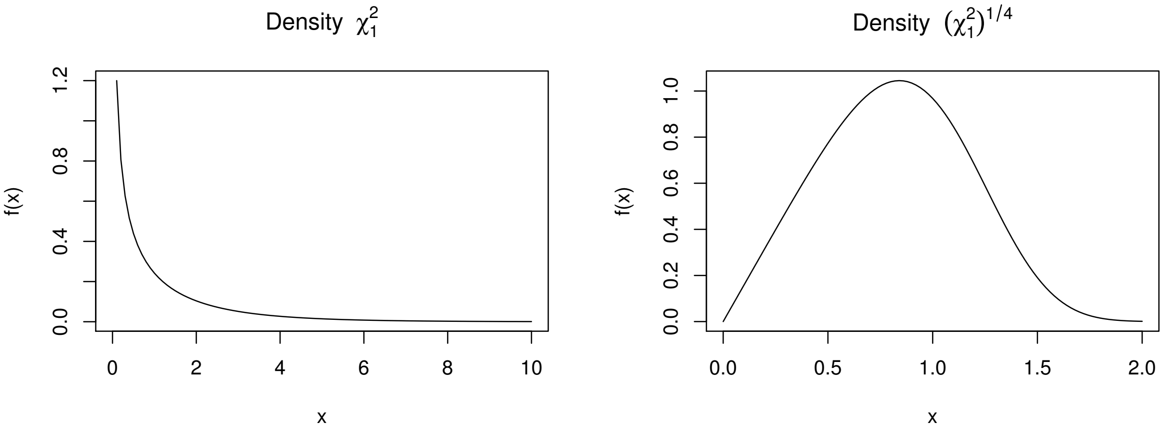

Figure 2. The density of a random variable X with χ_1^2-distribution (left) and the density of a random variable $X^{1/4}$ (right).

If we assume that the random field is weakly stationary, we may also work with autocovariance function $C(h) = \operatorname{cov}(Z(x), Z(x+h))$. In geostatistics, the term *covariogram* is usually used for the autocovariance function. Then there exists a relation between semivariogram and covariogram,

$$\gamma(h) = C(o) - C(h). \tag{5}$$

The classical empirical estimator of the autocovariance function is

$$\widehat{C}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(x_i) - \bar{Z})(Z(x_j) - \bar{Z}), \tag{6}$$

where the sample mean $\bar{Z} = \frac{1}{n} \sum_{j=1}^{n} Z(x_j)$ estimates the mean μ. The disadvantage is that we have to estimate μ which causes bias of the estimator (6). From this reason the variogram seems to provide better characterization of dependence than the autocovariance function. However, the autocovariance function is much widely used. The estimator of autocovariance function is symmetric ($\widehat{C}(h) = \widehat{C}(-h)$) and for $h = o$ we have the estimator of variance:

$$\widehat{C}(o) = \frac{1}{n} \sum_{i=1}^{n} (Z(x_i) - \bar{Z})^2.$$

Rewriting (4) so that we add and subtract $\bar{Z}$ in each summand, we get

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} \left[(Z(x_i) - \bar{Z})^2 + (Z(x_j) - \bar{Z})^2 \right] - 2\widehat{C}(h).$$

Therefore, $2\hat{\gamma}(h) \neq 2(\widehat{C}(o) - \widehat{C}(h))$, and so the relation (5) is not preserved if we pass to the moment estimates. It would be unreasonable to estimate variogram by $2(\widehat{C}(o) - \widehat{C}(h))$, i.e. by plugging sample covariances into (5), because negative values could be obtained.

Often we assume that the random field is also isotropic. Then the variogram is the function of distance $\|h\|$. We can exploit this in the construction of estimators. In the histogram-based estimator we can consider groups of pairs of points with similar mutual distances. In the kernel estimator we use $k_b(\|x_i - x_j\| - \|h\|)$, where k_b is an one-dimensional kernel function.

The disadvantage of non-parametric estimates is their larger variance and also the resulting estimators do not have to be valid variograms or covariograms. We know that every variogram must be conditionally negative definite and every covariogram must be positive semidefinite. However, the estimators $\hat{\gamma}$ and $\widehat{C}$ do not necessarily have these properties. Therefore, we are now going to study parametric methods for estimation of variogram and autocovariance function.

Parametric methods

We select a parametric model for the variogram $2\gamma_{\boldsymbol{\theta}}(h)$ or the covariogram $C_{\boldsymbol{\theta}}(h)$, where $\boldsymbol{\theta} \in \Theta$ is the vector of unknown parameters. For example, we may consider power model for variogram,

$$2\gamma_{\boldsymbol{\theta}}(h) = c_0 + b\|h\|^{\alpha}, \quad \boldsymbol{\theta} = (c_0, b, \alpha)^{\mathrm{T}},$$

where $c_0 \geq 0$ is the nugget, $b \geq 0$ and $0 \leq \alpha < 2$. Our aim is to estimate $\boldsymbol{\theta}$ from data.

Least squares

The first possibility is a curve-fitting method of some non-parametric estimator computed in several values h_k, $k = 1, \ldots, K$. The simplest approach would be to minimize

$$\sum_{k=1}^{K} \left(2\hat{\gamma}(h_k)) - 2\gamma_{\boldsymbol{\theta}}(h_k)\right)^2.$$

This is the ordinary least squares method. It disregards the correlations among the estimates $2\hat{\gamma}(h_k)$ and their unequal variances. Put $\boldsymbol{h} = (h_1, \ldots, h_K)^{\mathrm{T}}$, $2\hat{\gamma}(\boldsymbol{h}) = (2\hat{\gamma}(h_1), \ldots, 2\hat{\gamma}(h_K))^{\mathrm{T}}$ and $2\gamma_{\boldsymbol{\theta}}(\boldsymbol{h}) = (2\gamma_{\boldsymbol{\theta}}(h_1), \ldots, 2\gamma_{\boldsymbol{\theta}}(h_K))^{\mathrm{T}}$, and consider the statistical model in the form

$$2\hat{\gamma}(\boldsymbol{h}) = 2\gamma_{\boldsymbol{\theta}}(\boldsymbol{h}) + e(\boldsymbol{h}),$$

where we assume that $e(\boldsymbol{h}) = (e(h_1), \ldots, e(h_K))^{\mathrm{T}}$ has zero mean and variance matrix $\boldsymbol{V}(\boldsymbol{\theta})$, which may depend on $\boldsymbol{\theta}$. Now we can apply the method of generalized least squares and minimize

$$(2\hat{\gamma}(\boldsymbol{h}) - 2\gamma_{\boldsymbol{\theta}}(\boldsymbol{h}))^{\mathrm{T}} \boldsymbol{V}(\boldsymbol{\theta})^{-1} (2\hat{\gamma}(\boldsymbol{h}) - 2\gamma_{\boldsymbol{\theta}}(\boldsymbol{h}))$$

over $\boldsymbol{\theta} \in \Theta$. The problem is how to get the matrix $\boldsymbol{V}(\boldsymbol{\theta})$.

Let $\{Z(x) : x \in D\}$ be a Gaussian random field. Then

$$\mathbb{E}(Z(x_1 + h_1) - Z(x_1))^2 = 2\gamma(h_1) \quad \text{and} \quad \operatorname{var}(Z(x_1 + h_1) - Z(x_1))^2 = 2(2\gamma(h_1))^2.$$

In order to express the covariance we use that $\operatorname{cov}(X^2, Y^2) = 2\varrho^2$ holds for a random vector $(X, Y)^{\mathrm{T}}$ with two-dimensional normal distribution such that $\operatorname{var} X = \operatorname{var} Y = 1$ and $\operatorname{cov}(X, Y) = \varrho$. Hence,

$$\begin{aligned}\operatorname{cov}((Z(x_1 + h_1) - Z(x_1))^2, (Z(x_2 + h_2) - Z(x_2))^2) = 2\big(&\gamma(x_1 - x_2 + h_1) + \gamma(x_1 - x_2 - h_2)\\ &- \gamma(x_1 - x_2 + h_1 - h_2) - \gamma(x_1 - x_2)\big)^2.\end{aligned}$$

The variance of the estimator (4) is

$$\begin{aligned}\operatorname{var} 2\hat{\gamma}(h_k) &= \frac{1}{|N(h_k)|^2} \operatorname{var} \sum_{N(h_k)} (Z(x_i) - Z(x_j))^2\\ &= \frac{1}{|N(h_k)|^2} \sum_{i,j} \sum_{l,m} \operatorname{cov}((Z(x_i) - Z(x_j))^2, (Z(x_l) - Z(x_m))^2).\end{aligned}$$

A simple approximation of this variance is

$$\operatorname{var} 2\hat{\gamma}(h_k) \approx \frac{2(2\gamma_{\boldsymbol{\theta}}(h_k))^2}{|N(h_k)|}. \tag{7}$$

This approximation is precise if $(Z(x_i) - Z(x_j))^2$ are uncorrelated. We replace the matrix $\boldsymbol{V}(\boldsymbol{\theta})$ by the diagonal matrix $\boldsymbol{\Delta}(\boldsymbol{\theta})$ with elements given by the relation (7). Then we obtain weighted sum of squares

$$(2\hat{\gamma}(\boldsymbol{h}) - 2\gamma_{\boldsymbol{\theta}}(\boldsymbol{h}))^{\mathrm{T}} \boldsymbol{\Delta}(\boldsymbol{\theta})^{-1} (2\hat{\gamma}(\boldsymbol{h}) - 2\gamma_{\boldsymbol{\theta}}(\boldsymbol{h})) = \sum_{k=1}^{K} \frac{|N(h_k)|}{2\gamma_{\boldsymbol{\theta}}(h_k)^2} \left(\hat{\gamma}(h_k) - \gamma_{\boldsymbol{\theta}}(h_k)\right)^2.$$

The estimator of $\boldsymbol{\theta}$ by the method of weighted least squares is obtained by the minimization of this sum.

Maximum likelihood

The second possibility is to look for an estimator by the maximum likelihood method. For Gaussian random field with mean μ and autocovariance function $C_{\boldsymbol{\theta}}$, the log-likelihood based on data $\boldsymbol{z}_n = (z(x_1), \ldots, z(x_n))^{\mathrm{T}}$ has after multiplying by -2 this form:

$$-2 \log L(\mu, \boldsymbol{\theta}) = n \log 2\pi + \log \det(\boldsymbol{C}_n(\boldsymbol{\theta})) + (\boldsymbol{z}_n - \mu \mathbf{1}_n)^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} (\boldsymbol{z}_n - \mu \mathbf{1}_n),$$

where $\mathbf{1}_n = (1, \ldots, 1)^{\mathrm{T}}$ and $\boldsymbol{C}_n(\boldsymbol{\theta})_{ij} = C_{\boldsymbol{\theta}}(x_i - x_j)$ depend on the vector $\boldsymbol{\theta}$ of covariance parameters. For given $\boldsymbol{\theta}$, $L(\mu, \boldsymbol{\theta})$ is maximized for

$$\tilde{\mu} = (\mathbf{1}_n^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} \boldsymbol{z}_n. \tag{8}$$

It is the generalized least squares estimator. Plugging $\tilde{\mu}$ into $L(\mu, \boldsymbol{\theta})$ we get the function of $\boldsymbol{\theta}$ (so called *profile likelihood*), which has to be maximized (mostly by numerical methods). The estimator of μ is then given by (8) with the estimate of $\boldsymbol{\theta}$ inserted.

A popular variant of maximum likelihood is REML – estimator by the method of *residual/restricted maximal likelihood*. This method does not apply the likelihood directly to data but to the residuals. It relies on finding an appropriate matrix $\boldsymbol{A}$, which linearly transforms data $\boldsymbol{Z}_n = (Z(x_1), \ldots, Z(x_n))^{\mathrm{T}}$ to $\boldsymbol{Z}^* = \boldsymbol{A}\boldsymbol{Z}_n$ so that the distribution of $\boldsymbol{Z}^*$ is not depending on μ. The parameter $\boldsymbol{\theta}$ is then estimated by the maximum likelihood method applied to the transformed data $\boldsymbol{Z}^*$. The choice of matrix $\boldsymbol{A}$ is not unique. For example, for matrix $\boldsymbol{A}$ of type $(n-1) \times n$ with entries $a_{ij} = \mathbf{1}_{[i=j]} - 1/n$, we get $\boldsymbol{A}\boldsymbol{Z}_n = (Z(x_1) - \bar{Z}, \ldots, Z(x_{n-1}) - \bar{Z})^{\mathrm{T}}$ the vector of $n-1$ differences from the sample mean $\bar{Z}$. In this way we get rid of dependence on μ. The estimator of $\boldsymbol{\theta}$ minimizes the function

$$\log \det(\boldsymbol{A}\boldsymbol{C}_n(\boldsymbol{\theta})\boldsymbol{A}^{\mathrm{T}}) + \boldsymbol{z}_n^{\mathrm{T}} \boldsymbol{A}^{\mathrm{T}} (\boldsymbol{A}\boldsymbol{C}_n(\boldsymbol{\theta})\boldsymbol{A}^{T})^{-1} \boldsymbol{A}\boldsymbol{z}_n.$$

Plugging this estimator to (8) we get the estimator of μ. For practical determination of parameter estimators one can use functions likfit and variofit in library geoR or fitvario in library RandomFields.

Composite likelihood

Composite likelihood method was already mentioned when dealing with parameter estimation in point processes. Similarly, it can be used for estimation of variogram parameters. Assume that the differences $Z(x_i)-Z(x_j)$ have normal distribution. Summing the contributions of log-likelihood over pairs of distinct points we get the logarithm of composite likelihood:

$$\log \mathrm{CL}(\boldsymbol{\theta}) = \sum_{i,j=1,\dots,n}^{\neq} \left[-\frac{1}{2} \log 4\pi\gamma_{\boldsymbol{\theta}}(x_i - x_j) - \frac{1}{4\gamma_{\boldsymbol{\theta}}(x_i - x_j)}(z(x_i) - z(x_j))^2 \right].$$

We are looking for $\boldsymbol{\theta}$ that maximizes $\mathrm{CL}(\boldsymbol{\theta})$. So we differentiate w.r.t. components θ_k and put equal to zero:

$$\sum_{i,j=1,\dots,n}^{\neq} \frac{\partial\gamma_{\boldsymbol{\theta}}(x_i - x_j)}{\partial\theta_k} \frac{1}{4\gamma_{\boldsymbol{\theta}}(x_i - x_j)^2} \left[(z(x_i) - z(x_j))^2 - 2\gamma_{\boldsymbol{\theta}}(x_i - x_j) \right] = 0.$$

Model validation

Once we have already chosen a parametric model of variogram and estimated its parameters, we are interested how well the obtained model $2\gamma_{\hat{\boldsymbol{\theta}}}$ describes the data. In the next subsection we will see how to obtain the prediction $\hat{Z}(x_0)$ of $Z(x_0)$ together with the prediction error $\sigma^2(x_0)$. It depends on the fitted variogram, data and the locations $x_0, x_1, \dots, x_n$. If we are able to get $Z(x_0)$, e.g. by additional measurement or by remaining data that we left for model validation, we may compare the difference between $Z(x_0)$ and $\hat{Z}(x_0)$. These values should be close to each other if the variogram is chosen correctly.

If all data were used for variogram fitting and it is impossible to execute additional measurement, we can accomplish the *cross-validation*. We omit the location x_j and calculate the prediction $\hat{Z}_{-j}(x_j)$ from $n-1$ remaining observations and fitted variogram $2\gamma_{\hat{\boldsymbol{\theta}}}(h)$. The corresponding prediction error is denoted by $\sigma^2_{-j}(x_j)$. We make this procedure for each $j = 1, \dots, n$ and calculate the standardized residua

$$\frac{Z(x_j) - \hat{Z}_{-j}(x_j)}{\sigma_{-j}(x_j)}.$$

Their arithmetic mean has to be around 0 and their sample second moment around 1. From the histogram of standardized residua we can detect possible extreme values of residuals.

3.2 Kriging

Again we assume that the observed data form a vector $\boldsymbol{Z}_n = (Z(x_1), \dots, Z(x_n))^{\mathrm{T}}$. Our aim is to find the prediction $\hat{Z}(x_0)$ of the value $Z(x_0)$ that random field attains in some further location $x_0 \in D$. For the methods of spatial prediction based on the minimization of mean squared error the term *kriging* is used. It is named after a South African mining engineer D. G. Krige. His paper [5] dealing with mineral resources is a pioneering paper in geostatistics.

Simple kriging

It is well-known that the mean squared error $\mathbb{E}[Z(x_0) - \hat{Z}(x_0)]^2$ is under the assumption of finite second moments minimized by the conditional expectation $\mathbb{E}[Z(x_0) \mid \boldsymbol{Z}_n]$ and the minimum value is $\mathbb{E}[Z(x_0) - \hat{Z}(x_0)]^2 = \mathbb{E}\operatorname{var}[Z(x_0) \mid \boldsymbol{Z}_n]$, see e.g. [6], Theorem 7.15. In practice the conditional expectation is difficult to determine. Therefore, for simplicity we consider linear prediction $\hat{Z}(x_0) = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}_n$. Our aim is to estimate $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^n$ so that mean squared error is minimal. From the theory of linear models we know that the solution is

$$\boldsymbol{\beta}_0 = \boldsymbol{C}_n^{-1}\boldsymbol{c}_n, \quad \alpha_0 = \mu(x_0) - \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\mu}_n,$$

where $\boldsymbol{\mu}_n = \mathbb{E}\boldsymbol{Z}_n = (\mu(x_1), \dots, \mu(x_n))^{\mathrm{T}}$, $\mu(x_0) = \mathbb{E}Z(x_0)$,

$$\boldsymbol{C}_n = (\operatorname{cov}(Z(x_i), Z(x_j)))_{i,j=1,\dots,n}$$

is the variance matrix of $\boldsymbol{Z}_n$ and $\boldsymbol{c}_n = (C(x_0, x_1), \dots, C(x_0, x_n))^{\mathrm{T}}$. Hence,

$$\hat{Z}(x_0) = \mu(x_0) + \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}(\boldsymbol{Z}_n - \boldsymbol{\mu}_n)$$

and the prediction error is

$$\sigma^2(x_0) = \mathbb{E}(Z(x_0) - \hat{Z}(x_0))^2 = \operatorname{var} Z(x_0) - \boldsymbol{c}_n^{\mathrm{T}} \boldsymbol{C}_n^{-1} \boldsymbol{c}_n.$$

The technique for obtaining this spatial prediction is called *simple kriging*. Even if we haven't require it, the prediction $\hat{Z}(x_0)$ is unbiased, i.e. $\mathbb{E}\hat{Z}(x_0) = \mathbb{E}Z(x_0)$. Notice that the prediction error does not depend on data. If x_0 is one of the locations $x_1, \ldots, x_n$, then $\hat{Z}(x_0) = Z(x_0)$, i.e. spatial prediction interpolates data. To make sure that it is true, note that

$$\hat{Z}(x_j) = \mu(x_j) + (C(x_j, x_1), \ldots, C(x_j, x_n))^{\mathrm{T}} \boldsymbol{C}_n^{-1}(\boldsymbol{Z}_n - \boldsymbol{\mu}_n), \quad j = 1, \ldots, n,$$

which rewritten for the vectors becomes

$$(\hat{Z}(x_1), \ldots, \hat{Z}(x_n))^{\mathrm{T}} = \boldsymbol{\mu}_n + \boldsymbol{C}_n \boldsymbol{C}_n^{-1}(\boldsymbol{Z}_n - \boldsymbol{\mu}_n) = \boldsymbol{Z}_n.$$

Simple kriging is optimal for Gaussian random fields.

Lemma 2. *Let $\{Z(x) : x \in D\}$ be a Gaussian random field. The best linear prediction $\hat{Z}(x_0) = \mu(x_0) + \boldsymbol{c}_n^{\mathrm{T}} \boldsymbol{C}_n^{-1}(\boldsymbol{Z}_n - \boldsymbol{\mu}_n)$ is the best prediction of $Z(x_0)$ and*

$$Z(x_0) \mid \boldsymbol{Z}_n \sim N(\hat{Z}(x_0), \mathbb{E}(Z(x_0) - \hat{Z}(x_0))^2),$$

where $\mathbb{E}(Z(x_0) - \hat{Z}(x_0))^2 = \operatorname{var} Z(x_0) - \boldsymbol{c}_n^{\mathrm{T}} \boldsymbol{C}_n^{-1} \boldsymbol{c}_n$.

Proof: The joint distribution of $(Z(x_0), \boldsymbol{Z}_n)^{\mathrm{T}}$ is $(n+1)$-dimensional normal. Conditional distributions in a multi-dimensional normal distribution are again normal. In our case the conditional distribution $Z(x_0) \mid \boldsymbol{Z}_n$ is normal with mean $\mu(x_0) + \boldsymbol{c}_n^{\mathrm{T}} \boldsymbol{C}_n^{-1}(\boldsymbol{Z}_n - \boldsymbol{\mu}_n)$ and variance $\operatorname{var} Z(x_0) - \boldsymbol{c}_n^{\mathrm{T}} \boldsymbol{C}_n^{-1} \boldsymbol{c}_n$. The best (not necessarily linear) prediction of $Z(x_0)$ is the conditional expectation $\mathbb{E}[Z(x_0) \mid \boldsymbol{Z}_n]$.

□

The best linear prediction is optimal in case of Gaussian model. However, the best linear prediction may have bad properties when the assumptions of normal distribution are violated. In statistics this problem is often settled up with the transformation of data leading to normal distribution An example is so called *Box-Cox transformation*

$$g_\lambda(z) = \begin{cases} \frac{z^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log z, & \lambda = 0. \end{cases}$$

There exist different methods to estimate parameter λ. It is also possible to follow Bayesian approach and consider λ as random.

We have expressed the best linear prediction. However, it depends on the values $\boldsymbol{\mu}_n$, $\mu(x_0)$, $\boldsymbol{c}_n$ and $\boldsymbol{C}_n$, which are unknown in practice. In general we have $(n+1) + n + \binom{n+1}{2}$ unknown parameters that would have to be estimated from n observations. Therefore, we add some further assumptions.

Ordinary kriging

Assume now that the random field has constant and finite mean μ. We look for the linear prediction in the form

$$\hat{Z}(x_0) = \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{Z}_n, \quad \text{where} \sum_{j=1}^{n} \lambda_j = \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{1}_n = 1,$$

where the components $\lambda_1, \ldots, \lambda_n$ of vector $\boldsymbol{\lambda}$ are unknown real coefficients. The condition that their sum is one ensures that the prediction is unbiased: $\mathbb{E}\hat{Z}(x_0) = \boldsymbol{\lambda}^{\mathrm{T}} \mu \mathbf{1}_n = \mu = \mathbb{E}Z(x_0)$. The method for finding the spatial prediction under these assumptions is named *ordinary kriging*.

For intrinsically stationary random field with semivariogram γ we can express the variance of a linear combination with zero sum coefficients as follows:

$$\begin{aligned} \mathbb{E}(Z(x_0) - \hat{Z}(x_0))^2 &= \mathbb{E}(Z(x_0) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{Z}_n)^2 = \operatorname{var}(Z(x_0) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{Z}_n) \\ &= -\sum_{i,j} \lambda_i \lambda_j \gamma(x_i - x_j) + 2 \sum_{i=1}^{n} \lambda_i \gamma(x_i - x_0). \qquad (9) \end{aligned}$$

It means that we don't need to know μ in order to determine the prediction $\hat{Z}(x_0)$. To find the minimum of (9) under the condition $\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{1}_n = 1$ we can apply the method of Lagrange multipliers. For simpler notation we multiply the multiplier by 2 and minimize

$$Q = \operatorname{var}(Z(x_0) - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{Z}_n) - 2m(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{1}_n - 1) = -\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Gamma}_n\boldsymbol{\lambda} + 2\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\gamma}_n - 2m(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{1}_n - 1),$$

where $\boldsymbol{\Gamma}_n = (\gamma(x_i - x_j))_{i,j=1,\dots,n}$ and $\boldsymbol{\gamma}_n = (\gamma(x_1 - x_0), \dots, \gamma(x_n - x_0))^{\mathrm{T}}$. Differentiate Q w.r.t. $\boldsymbol{\lambda}$ and m, set the derivative equal to zero, and obtain

$$\begin{aligned} \frac{\partial Q}{\partial \lambda} &= -2\boldsymbol{\Gamma}_n\boldsymbol{\lambda} + 2\boldsymbol{\gamma}_n - 2m\mathbf{1}_n = 0, \\ \frac{\partial Q}{\partial m} &= -2(\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{1}_n - 1) = 0. \end{aligned}$$

The solution is

$$\begin{aligned} \boldsymbol{\lambda}^{\mathrm{T}} &= \left(\boldsymbol{\gamma}_n + \mathbf{1}_n\frac{1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\mathbf{1}_n}\right)^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}, \qquad (10) \\ m &= -\frac{1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\mathbf{1}_n}. \end{aligned}$$

Hence, the prediction has the form

$$\hat{Z}(x_0) = \left(\boldsymbol{\gamma}_n + \mathbf{1}_n\frac{1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\mathbf{1}_n}\right)^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{Z}_n = \lambda_1 Z(x_1) + \dots + \lambda_n Z(x_n).$$

The coefficients λ_i are components of vector (10) and they are called *prediction weights*. The prediction weights are typically large for points close to x_0. Nevertheless, their precise value depends on the locations x_i and covariance structure of data. It can happen that λ_i is negative or larger than 1. The prediction error is

$$\sigma^2(x_0) = \mathbb{E}(Z(x_0) - \hat{Z}(x_0))^2 = 2\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\gamma}_n - \boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\Gamma}_n\boldsymbol{\lambda} = \boldsymbol{\gamma}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n - \frac{(1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n)^2}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\mathbf{1}_n}.$$

Similarly we can rewrite $\hat{Z}(x_0)$ for weakly stationary random field using autocovariance function:

$$\hat{Z}(x_0) = \left(\boldsymbol{c}_n + \mathbf{1}_n\frac{1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\mathbf{1}_n}\right)^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{Z}_n.$$

The prediction error is

$$\sigma^2(x_0) = \operatorname{var} Z(x_0) - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n + \frac{(1 - \mathbf{1}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n)^2}{\mathbf{1}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\mathbf{1}_n}.$$

We see that this error is larger than in the case of simple kriging because the last term is positive. Larger error is caused by the fact that we don't know mean μ.

Universal kriging

In this part we deal with situation when the mean $\mu(x) = \mathbb{E}Z(x)$ is not constant. The simplest approach is to use a linear model

$$\mu(x) = \sum_{j=0}^{p}\beta_j f_j(x),$$

where $f_0(x), \dots, f_p(x)$ are known observed values of functions f_j in points $x \in D$ and $\beta_0, \dots, \beta_p$ are unknown real parameters. A usual choice for f_0 is the constant function equal to 1, then β_0 is an absolute term. For $f_i(x)$ one can consider a polynomial of spatial coordinates of the location x. In this way it is possible to model for example the linear trend. Another possibility is that $f_i(x)$ represents some covariate. Denote $\boldsymbol{f} = (f_0(x_0), \dots, f_p(x_0))^{\mathrm{T}}$ and let $\boldsymbol{F}$ be a matrix of type $n \times (p+1)$ with elements $f_j(x_i)$, $i = 1, \dots, n$, $j = 0, \dots, p$. If we require the prediction in the form

$$\hat{Z}(x_0) = \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{Z}_n, \quad \text{where } \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{F} = \boldsymbol{f}^{\mathrm{T}},$$

we speak about *universal kriging*. The condition $\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{F} = \boldsymbol{f}^{\mathrm{T}}$ ensures that this prediction is unbiased because

$$\mathbb{E}\hat{Z}(x_0) = \boldsymbol{\lambda}^{T} \mathbb{E}\boldsymbol{Z}_n = \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{F}\boldsymbol{\beta} = \boldsymbol{f}^{\mathrm{T}}\boldsymbol{\beta} = \mu(x_0) = \mathbb{E}Z(x_0).$$

The optimal prediction (minimizing mean squared error) can be again found by applying the method of Lagrange multipliers. Analogously as in the case of ordinary kriging we can show that the optimal prediction weights have the form

$$\boldsymbol{\lambda}^{\mathrm{T}} = \left(\boldsymbol{\gamma}_n + \boldsymbol{F}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{F})^{-1}(\boldsymbol{f} - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n)\right)^{\mathrm{T}} \boldsymbol{\Gamma}_n^{-1}.$$

The corresponding prediction error is

$$\boldsymbol{\gamma}_n^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n + (\boldsymbol{f} - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n)^{\mathrm{T}}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{F})^{-1}(\boldsymbol{f} - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{\Gamma}_n^{-1}\boldsymbol{\gamma}_n).$$

Using covariance the prediction weights could be written as

$$\boldsymbol{\lambda}^{\mathrm{T}} = \left(\boldsymbol{c}_n + \boldsymbol{F}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})^{-1}(\boldsymbol{f} - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n)\right)^{\mathrm{T}} \boldsymbol{C}_n^{-1}$$

and the prediction error as

$$\sigma^2(x_0) = C(o) - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n + (\boldsymbol{f} - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n)^{\mathrm{T}}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})^{-1}(\boldsymbol{f} - \boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n).$$

By generalized least squares we can also estimate the parameter $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{Z}_n.$$

The prediction could be also written as

$$\hat{Z}(x_0) = \boldsymbol{f}^{\mathrm{T}}\widehat{\boldsymbol{\beta}} + \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}(\boldsymbol{Z}_n - \boldsymbol{F}\widehat{\boldsymbol{\beta}}).$$

If $Z(x_0)$ is uncorrelated with data, the prediction $\hat{Z}(x_0)$ coincides with the best linear unbiased estimator of the mean, which is equal to $\boldsymbol{f}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}$. However, generally the prediction of $Z(x_0)$ and the estimator of $\mathbb{E}Z(x_0)$ are distinct.

Other possibilities

Assume that instead of prediction of $Z(x_0)$ we are interested in the prediction of mean value in some block B,

$$Z(B) = \frac{1}{|B|}\int_B Z(x)\,\mathrm{d}x.$$

An analogy of ordinary kriging leads to so called *block kriging*. We look for the prediction in the form

$$\hat{Z}(B) = \sum_{i=1}^{n} \hat{\lambda}_i Z(x_i),$$

where $\sum_{i=1}^{n} \lambda_i = 1$. The optimal prediction weights have the form

$$\boldsymbol{\lambda}^{\mathrm{T}} = \left(\boldsymbol{c}_B + \boldsymbol{1}_n \frac{1 - \boldsymbol{1}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_B}{\boldsymbol{1}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{1}_n}\right)^{\mathrm{T}} \boldsymbol{C}_n^{-1},$$

where $\boldsymbol{c}_B = (\operatorname{cov}(Z(B), Z(x_1)), \dots, \operatorname{cov}(Z(B), Z(x_n)))^{\mathrm{T}}$. The expression using variogram would look analogously.

Similarly we may be interested in the prediction of $g(Z(x_0))$, where g is given function. The best prediction is $\mathbb{E}[g(Z(x_0)) \mid \boldsymbol{Z}_n]$.

Another frequent case is the task of estimating probabilities $\mathbb{P}(Z(x_0) \le y \mid \boldsymbol{Z}_n)$, where y is given real number. We speak about *indicator kriging*.

To accomplish kriging techniques in R we can use krige.conv in library geoR or krige in library gstat.

3.3 Estimating covariance parameters

The formulas for spatial prediction derived in previous subsection depend on the values of a covariogram or a variogram that are typically unknown in practice and so they must be somehow estimated. We have already mentioned basic approaches for estimation of parameters for variogram or covariogram. We put estimates of the parameters into the parametric formula of the corresponding function. In this way we obtain so called *plug-in* estimators. The procedure goes in the following steps:

1. we select a parametric model for variogram $\gamma_{\boldsymbol{\theta}}(h)$ or covariogram $C_{\boldsymbol{\theta}}(h)$,
2. we estimate the parameter $\boldsymbol{\theta}$,
3. we adjust statistical inference to take into account that instead of constant $\boldsymbol{\theta}$ we work with random variable $\hat{\boldsymbol{\theta}}$.

The plug-in prediction for ordinary kriging has the form

$$\hat{\hat{Z}}(x_0) = \left(\boldsymbol{c}_n(\hat{\boldsymbol{\theta}}) + \mathbf{1}_n \frac{1 - \mathbf{1}_n^{\mathrm{T}} \boldsymbol{C}_n(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{c}_n(\hat{\boldsymbol{\theta}})}{\mathbf{1}_n^{\mathrm{T}} \boldsymbol{C}_n(\hat{\boldsymbol{\theta}})^{-1} \mathbf{1}_n} \right)^{\mathrm{T}} \boldsymbol{C}_n(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{Z}_n.$$

It is not longer the best linear unbiased prediction (BLUP) of $Z(x_0)$. It is just the estimator of this prediction (i.e. EBLUP = estimated best linear unbiased prediction). While the prediction error of $\hat{Z}(x_0)$ is

$$C(o) - \boldsymbol{c}_n(\boldsymbol{\theta})^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} \boldsymbol{c}_n(\boldsymbol{\theta}) + \frac{(1 - \mathbf{1}_n^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} \boldsymbol{c}_n(\boldsymbol{\theta}))^2}{\mathbf{1}_n^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} \mathbf{1}_n}, \tag{11}$$

the prediction error of $\hat{\hat{Z}}(x_0)$ is unknown. If we plug $\hat{\boldsymbol{\theta}}$ into (11), we get the estimate of prediction error of $\hat{Z}(x_0)$, i.e. of different prediction than we in fact use. This estimated prediction error has tendency to undervalue true prediction error of $\hat{\hat{Z}}(x_0)$ because we affect the fact that random $\hat{\boldsymbol{\theta}}$ introduces further variability into the EBLUP.

Return back to the case of universal kriging, where we consider the model $Z(x) = \boldsymbol{F}(x)^{\mathrm{T}} \boldsymbol{\beta} + e(x)$, where $\boldsymbol{F}(x) = (f_0(x), \dots, f_p(x))^{\mathrm{T}}$ and $\{e(x) : x \in D\}$ is intrinsically stationary random field with variogram parametrized by $\boldsymbol{\theta}$. It would be unreasonable to use empiric estimate of $\boldsymbol{\theta}$ from data $\boldsymbol{Z}_n = (Z(x_1), \dots, Z(x_n))^{\mathrm{T}}$ because it is substantially biased. The bias of (4) is caused by the fact that $Z(x)$ does not have constant mean. Therefore, $\mathbb{E}(Z(x_i) - Z(x_j))^2 = \mathrm{var}(Z(x_i) - Z(x_j)) + (\mu(x_i) - \mu(x_j))^2$. We would need a variogram estimator for $\{e(x) : x \in D\}$, but the error random field $\{e(x) : x \in D\}$ is unobservable. If $\boldsymbol{\beta}$ was known, then $e(x) = Z(x) - \boldsymbol{F}(x)^{\mathrm{T}} \boldsymbol{\beta}$ and we would be able to estimate $\boldsymbol{\theta}$ from $\boldsymbol{e}_n = (e(x_1), \dots, e(x_n))^{\mathrm{T}}$. we would be able to estimate $\boldsymbol{\theta}$. However, the parameter $\boldsymbol{\beta}$ is unknown. If the field $\{e(x) : x \in D\}$ is weakly stationary with autocovariance function $C_{\boldsymbol{\theta}}(h)$, we get the estimator of $\boldsymbol{\beta}$ using method of generalized least squares,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} \boldsymbol{F})^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} \boldsymbol{Z}_n. \tag{12}$$

Nevertheless, this estimator requires knowledge of parameter $\boldsymbol{\theta}$. It means that we are not able to reasonably estimate $\boldsymbol{\theta}$ without knowledge of $\boldsymbol{\beta}$ and on the other hand to estimate $\boldsymbol{\beta}$ we need an estimator of $\boldsymbol{\theta}$. This situation is referred to as the cat-and-mouse-game of universal kriging.

A possible solution is the *iteratively re-weighted generalized least squares* method. It is defined by the following steps.

1. obtain an initial estimator of $\boldsymbol{\beta}$, independent of $\boldsymbol{\theta}$, e.g. by ordinary least squares method: $\widehat{\boldsymbol{\beta}} = (\boldsymbol{F}^{\mathrm{T}} \boldsymbol{F})^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{Z}_n$,
2. calculate residuals $\boldsymbol{r} = \boldsymbol{Z}_n - \boldsymbol{F} \widehat{\boldsymbol{\beta}}$,
3. estimate parametric model of the variogram or covariogram of residuals and obtain $\hat{\boldsymbol{\theta}}$,
4. determine new estimator $\widehat{\boldsymbol{\beta}}$ as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{F}^{\mathrm{T}} \boldsymbol{C}_n(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{F})^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{C}_n(\hat{\boldsymbol{\theta}})^{-1} \boldsymbol{Z}_n$,
5. repeat steps 2.–4. until relative changes of estimators $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are small.

The estimator of variogram is biased though the bias is not caused by non-constant mean but by estimating variogram of residuals and not variogram of $\{e(x) : x \in D\}$.

The theoretical study of this procedure is intricate. It is not assured that the estimates converge to theoretical parameters.

Another possibility is to use maximum likelihood method to estimate both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ simultaneously. For example, for Gaussian random field $\{Z(x) : x \in D\}$ the log-likelihood has the form

$$\log L(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \det \boldsymbol{C}_n(\boldsymbol{\theta}) - \frac{1}{2} (\boldsymbol{z}_n - \boldsymbol{F}\boldsymbol{\beta})^T \boldsymbol{C}_n(\boldsymbol{\theta})^{-1} (\boldsymbol{z}_n - \boldsymbol{F}\boldsymbol{\beta}).$$

For fixed $\boldsymbol{\theta}$ this function is maximized for $\boldsymbol{\beta}$ given by (12) with vector $\boldsymbol{Z}_n$ replaced by observed data $\boldsymbol{z}_n$. Substituting this to $\log L(\boldsymbol{\beta}, \boldsymbol{\theta})$ we get a function of $\boldsymbol{\theta}$ (profile likelihood), which has to be maximized numerically.

3.4 Bayesian approach

In the classical approach the best prediction based on the observed data $\boldsymbol{z}_n$ is $\mathbb{E}[Z(x_0) \mid \boldsymbol{Z}_n = \boldsymbol{z}_n]$ and its error is $\mathbb{E}\operatorname{var}[Z(x_0) \mid \boldsymbol{Z}_n]$. Often we are rather interested in the whole conditional distribution of $Z(x_0)$ given $\boldsymbol{Z}_n = \boldsymbol{z}_n$ than only its mean and variance. This distribution is known as *predictive distribution*. In Bayesian approach the predictive distribution is equal to the posterior distribution of $Z(x_0)$.

Recall that in Bayesian statistics parameters are considered to be random. It means that there is no difference between prediction and parameter estimation. Bayesian approach is based on the combination of historical information about unknown parameters $\boldsymbol{\theta}$ and observed data $\boldsymbol{z}_n$. Information about parameters is given in so called *prior distribution* with density $p(\boldsymbol{\theta})$ w.r.t. σ-finite measure ν on the parametric space Θ. Let $\boldsymbol{Z}_n$ given $\boldsymbol{\theta}$ have a density $f(\boldsymbol{z}_n \mid \boldsymbol{\theta})$. Then the *posterior distribution* of $\boldsymbol{\theta}$ given $\boldsymbol{Z}_n = \boldsymbol{z}_n$ is given by the Bayes theorem

$$p(\boldsymbol{\theta} \mid \boldsymbol{z}_n) = \frac{f(\boldsymbol{z}_n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_\Theta f(\boldsymbol{z}_n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \nu(\mathrm{d}\boldsymbol{\theta})},$$

provided that the denominator is positive. This relation is shortly written as

$$p(\boldsymbol{\theta} \mid \boldsymbol{z}_n) \propto f(\boldsymbol{z}_n \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{13}$$

The symbol $\propto$ denotes equality up to a multiplicative constant.

Spatial prediction using Bayesian approach is denoted as *Bayesian kriging*. For prediction of $Z(x_0)$ we get the *predictive density* by integrating over $\boldsymbol{\theta}$:

$$f(z_0 \mid \boldsymbol{z}_n) = \int_\Theta f(z_0, \boldsymbol{\theta} \mid \boldsymbol{z}_n) \, \nu(\mathrm{d}\boldsymbol{\theta}) = \int_\Theta f(z_0 \mid \boldsymbol{z}_n, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{z}_n) \, \nu(\mathrm{d}\boldsymbol{\theta}). \tag{14}$$

For known $\boldsymbol{\theta}$ the result is the same as in the classical approach. The advantage of Bayesian approach is that it involves uncertainty about model parameters into consideration. The form (14) of predictive density is mostly quite complicated. Therefore, the MCMC method are used. They enable to generate a sequence $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ from posterior distribution with density $p(\boldsymbol{\theta} \mid \boldsymbol{z}_n)$. Then the average

$$\hat{f}(z_0 \mid \boldsymbol{z}_n) = \frac{1}{T} \sum_{i=1}^{T} f(z_0 \mid \boldsymbol{z}_n, \boldsymbol{\theta}^{(i)})$$

gives an approximation of the predictive density $f(z_0 \mid \boldsymbol{z}_n)$. In practice the calculation of this approximation is usually accomplished in the following way. For each $\boldsymbol{\theta}^{(i)}$ generate $z_0^{(i)}$ from distribution with density $f(z_0 \mid \boldsymbol{z}_n, \boldsymbol{\theta}^{(i)})$. Then $z_0^{(1)}, \ldots, z_0^{(T)}$ is a sample from the predictive distribution and depiction of the corresponding histogram or the kernel density estimator gives approximate shape of predictive density. Another possible approach for the determination of predictive density suggests itself when we are able to calculate posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{z}_n)$ and $p(\boldsymbol{\theta} \mid \boldsymbol{z}_n, z_0)$. Then we can exploit the relation

$$f(z_0 \mid \boldsymbol{z}_n) = f(z_0 \mid \boldsymbol{z}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{\theta} \mid \boldsymbol{z}_n)}{p(\boldsymbol{\theta} \mid \boldsymbol{z}_n, z_0)}.$$

Example: Consider a linear model

$$Z(x) = \boldsymbol{F}(x)^{\mathrm{T}} \boldsymbol{\beta} + e(x), \quad x \in D,$$

where $\boldsymbol{F}(x) = (f_0(x), \ldots, f_p(x))^{\mathrm{T}}$ is the vector of covariates, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of regression parameters with prior distribution $N_{p+1}(\boldsymbol{m}, \boldsymbol{Q})$ and $\{e(x) : x \in D\}$ is weakly stationary zero mean Gaussian random field with autocovariance function $C(h)$. Assume that we know the vector $\boldsymbol{m}$, matrix $\boldsymbol{Q}$ and function C. Our aim is the spatial prediction of $Z(x_0)$ based on the data $\boldsymbol{Z}_n = (Z(x_1), \ldots, Z(x_n))^{\mathrm{T}}$. Denote $\boldsymbol{C}_n$ the matrix with elements $C(x_i - x_j)$, $i, j = 1, \ldots, n$, and $\boldsymbol{F}$ the matrix of type $n \times (p+1)$, with elements $f_j(x_i)$, $i = 1, \ldots, n$, $j = 0, \ldots, p$. Further assume that both $\boldsymbol{Q}$ and $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F}$ have full rank. Since the normal distribution is a conjugate prior for a normally distributed data, the posterior distribution is multidimensional distribution. More precisely, $\boldsymbol{\beta} \mid \boldsymbol{Z}_n$ is distributed according to $N_{p+1}(\boldsymbol{m}^*, \boldsymbol{Q}^*)$, where

$$\boldsymbol{m}^* = (\boldsymbol{Q}^{-1} + \boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})^{-1}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{Z}_n + \boldsymbol{Q}^{-1}\boldsymbol{m}), \quad \boldsymbol{Q}^* = (\boldsymbol{Q}^{-1} + \boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})^{-1}.$$

The joint distribution $(\boldsymbol{Z}_n, Z(x_0))^{\mathrm{T}}$ is multidimensional normal $N_{n+1}(\boldsymbol{F}_{n0}\boldsymbol{\beta}, \boldsymbol{C}_{n0})$, where

$$\boldsymbol{F}_{n0} = \begin{pmatrix} \boldsymbol{F} \\ \boldsymbol{F}(x_0)^{\mathrm{T}} \end{pmatrix}$$

and

$$\boldsymbol{C}_{n0} = \begin{pmatrix} \boldsymbol{C}_n & \boldsymbol{c}_n \\ \boldsymbol{c}_n^{\mathrm{T}} & C(o) \end{pmatrix},$$

$\boldsymbol{c}_n = (C(x_0 - x_1), \ldots, C(x_0 - x_n))^{\mathrm{T}}$. The predictive density may be obtained from the expression

$$f(z_0 \mid \boldsymbol{z}_n) = \int f(z_0 \mid \boldsymbol{z}_n, \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \boldsymbol{z}_n) \, \mathrm{d}\boldsymbol{\beta},$$

where $p(\boldsymbol{\beta} \mid \boldsymbol{z}_n)$ is the density of $N_{p+1}(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ and $f(z_0 \mid \boldsymbol{z}_n, \boldsymbol{\beta})$ is the normal density with mean $\boldsymbol{F}(x_0)^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}(\boldsymbol{z}_n - \boldsymbol{F}\boldsymbol{\beta})$ and variance $C(o) - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n$, as we know from Lemma 2. After straightforward (even if somewhat lengthy) calculation we find out that the predictive distribution is normal with mean

$$(\boldsymbol{F}(x_0)^{\mathrm{T}} - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})\boldsymbol{Q}^*\boldsymbol{Q}^{-1}\boldsymbol{m} + \left[\boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1} + (\boldsymbol{F}(x_0)^{\mathrm{T}} - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})\boldsymbol{Q}^*\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n^{-1}\right]\boldsymbol{Z}_n$$

and variance

$$C(o) - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n + (\boldsymbol{F}(x_0)^{\mathrm{T}} - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})\boldsymbol{Q}^*(\boldsymbol{F}(x_0)^{\mathrm{T}} - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{F})^{\mathrm{T}}.$$

In practice we don't know the function C. However, we may use some of the parametric models (e.g. Whittle-Matérn). Then we specify appropriate prior distribution for parameters of autocovariance and derive the corresponding posterior distribution.

Geostatistical models that we have considered could be understood as two-stage *hierarchical models*. In the first stage of hierarchy we model the dependence of data on random effects. Specifically, a random field $\boldsymbol{Z} = \{Z(x) : x \in D\}$ is prescribed conditionally on $\boldsymbol{e} = \{e(x) : x \in D\}$. In the second stage of hierarchy we model the distribution of random effect, i.e. unobserved random field $\boldsymbol{e}$.

In the Bayesian approach we have three basic random objects, except $\boldsymbol{Z}$ and $\boldsymbol{e}$ it is also the vector of unknown parameters $\boldsymbol{\theta}$. We get three-stage hierarchical model:

1. $\boldsymbol{Z} \mid \boldsymbol{\theta}, \boldsymbol{e}$,
2. $\boldsymbol{e} \mid \boldsymbol{\theta}$,
3. $\boldsymbol{\theta}$.

A particular example is the model described at universal kriging and used also in the previous example:

$$Z(x) = \boldsymbol{F}(x)^{\mathrm{T}}\boldsymbol{\beta} + e(x).$$

Assume that the residual random field $\{e(x) : x \in D\}$ is centred stationary Gaussian random field and can be written as the sum of a spatial component and a white noise:

$$e(x) = W(x) + \epsilon(x),$$

where $\boldsymbol{W} = \{W(x) : x \in D\}$ is centred stationary Gaussian random field with autocovariance function $C_W(h; \sigma^2, \phi) = \sigma^2\rho(h; \phi)$ and $\{\epsilon(x) : x \in D\}$ are uncorrelated random variables having normal distribution with zero mean and variance τ^2. It means that the semivariogram of random field $\boldsymbol{e}$ is

$$\gamma_e(h; \sigma^2, \tau^2, \phi) = \tau^2 \mathbf{1}_{[h \neq o]} + \sigma^2 \left(1 - \rho(h; \phi)\right).$$

It is parametrized by nugget τ^2, partial sill σ^2 and correlation parameter ϕ, which appears in the correlation function $\rho(h;\phi)$ of random field $\boldsymbol{W}$. The vector of unknown parameters is thus $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2, \tau^2, \phi)^{\mathrm{T}}$. Then $\boldsymbol{Z}$ given $\boldsymbol{\theta}$ and $\boldsymbol{W}$ is a Gaussian random field with mean $\boldsymbol{F}(x)^{\mathrm{T}}\boldsymbol{\beta} + W(x)$ and autocovariance function $C_{Z|W}(h;\tau^2) = \tau^2\mathbf{1}_{[h=0]}$. Notice that $\boldsymbol{Z} \mid \boldsymbol{\theta}, \boldsymbol{W}$ is not depending on σ^2 and ϕ at all. In the second stage of hierarchy we specify $\boldsymbol{W}$ that conditionally on $\boldsymbol{\theta}$ has centred stationary Gaussian random field with autocovariance function $C_W(h;\sigma^2,\phi)$, i.e. $\boldsymbol{W}$ is not depending on $\boldsymbol{\beta}$ and τ^2. Third stage requires determination of suitable prior distribution for $\boldsymbol{\theta}$. The graphical illustration of this hierarchical model is shown in Figure 3. Usually the components of $\boldsymbol{\theta}$ are taken to be a priori independent, i.e. the prior density is

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\sigma^2)p(\tau^2)p(\phi).$$

Appropriate candidates for the choice of marginal prior distributions are multivariate normal distribution for $\boldsymbol{\beta}$, inverse Γ-distribution for σ^2 and τ^2 (i.e. $1/\sigma^2$ and $1/\tau^2$ have Γ-distribution). The choice for ϕ certainly depends on the form of variogram, e.g. for exponential model $\rho(h;\phi) = \exp\{-\phi\|h\|\}\}$ a prior distribution for ϕ is often taken to be Γ. The described model may be also formulated as two-stage. We utilize that $\boldsymbol{Z} \mid \boldsymbol{\theta}$ is Gaussian random field with mean $\boldsymbol{F}(x)^{\mathrm{T}}\boldsymbol{\beta}$ and autocovariance function

$$C_Z(h;\sigma^2,\tau^2,\phi) = \tau^2\mathbf{1}_{[h=0]} + \sigma^2\rho(h;\phi).$$

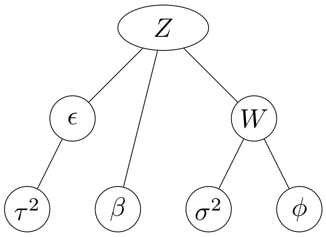

Figure 3. Representation of three-stage hierarchical model.

Assume that we observe the vector $\boldsymbol{z}_n = (z(x_1),\dots,z(x_n))^{\mathrm{T}}$. Bayesian estimate of the parameter is then obtained from posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{z}_n)$ given by (13), where in this case $f(\boldsymbol{z}_n \mid \boldsymbol{\theta})$ is the density of n-dimensional normal distribution with mean $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\beta}$ and variance matrix $\tau^2\boldsymbol{I}_n + \sigma^2\boldsymbol{H}(\phi)$. Here, $\boldsymbol{F} = (f_j(x_i))_{i,j}$ is the matrix of type $n \times (p+1)$, $\boldsymbol{I}_n$ is the identity matrix of size n and $\boldsymbol{H}(\phi)$ is the matrix of size n with elements $\rho(x_i - x_j;\phi)$, $i,j = 1,\dots,n$. We may as well use three-stage hierarchical model and express the posterior density as

$$p(\boldsymbol{\theta}, \boldsymbol{w}_n \mid \boldsymbol{z}_n) \propto f(\boldsymbol{z}_n \mid \boldsymbol{\theta}, \boldsymbol{w}_n)p(\boldsymbol{w}_n \mid \boldsymbol{\theta})p(\boldsymbol{\theta}),$$

where $f(\boldsymbol{z}_n \mid \boldsymbol{\theta}, \boldsymbol{w}_n)$ is the density of n-dimensional distribution with mean $\boldsymbol{F}^T\boldsymbol{\beta} + \boldsymbol{w}_n$ and variance matrix $\tau^2\boldsymbol{I}_n$ and $p(\boldsymbol{w}_n \mid \boldsymbol{\theta})$ is the density of n-dimensional normal distribution with zero mean and variance matrix $\sigma^2\boldsymbol{H}(\phi)$. However, in this way the number of parameters increases by n components of vector $\boldsymbol{w}_n = (w(x_1),\dots,w(x_n))^{\mathrm{T}}$. In practice MCMC methods (in particular, Gibbs sampler) are used. The form (13) is preferable because the variance matrix $\tau^2\boldsymbol{I}_n + \sigma^2\boldsymbol{H}(\phi)$ behaves better than the variance matrix $\sigma^2\boldsymbol{H}(\phi)$. This could be illustrated on the situation when the points x_i and x_j are close together. Then the matrix $\sigma^2\boldsymbol{H}(\phi)$ is close to a singular matrix while $\tau^2\boldsymbol{I}_n + \sigma^2\boldsymbol{H}(\phi)$ is not.

Estimation of parameters $\boldsymbol{w}_n$ corresponds to the reconstruction of spatial surface $\boldsymbol{W}$ in measurement points $x_1,\dots,x_n$. Similarly, we can be interested in the prediction of $W(x_0)$ for distinct choices of x_0. According to the relation

$$p(\boldsymbol{w}_n \mid \boldsymbol{z}_n) = \int\int p(\boldsymbol{w}_n \mid \sigma^2,\phi)p(\sigma^2,\phi \mid \boldsymbol{z}_n)\,\mathrm{d}\sigma^2\,\mathrm{d}\phi,$$

we may obtain the posterior distribution of $\boldsymbol{W}_n = (W(x_1),\dots,W(x_n))^{\mathrm{T}}$ from the posterior distribution of (σ^2,ϕ). Recall that in our case, $p(\boldsymbol{w}_n \mid \sigma^2,\phi)$ is the density of n-dimensional centred normal distribution with variance matrix $\sigma^2\boldsymbol{H}(\phi)$. Let $((\sigma^2)^{(t)},\phi^{(t)})$ be the output of MCMC algorithm, which generates samples from distribution with posterior density $p(\sigma^2,\phi \mid \boldsymbol{z})$. Then it suffices to generate vector $\boldsymbol{w}_n^{(t)}$ from the distribution with density $p(\boldsymbol{w}_n \mid (\sigma^2)^{(t)},\phi^{(t)})$, which gives us output from the distribution with density $p(\boldsymbol{w}_n \mid \boldsymbol{z}_n)$.

4. Lattice data

4.1 Modelling and estimation for areal data

By areal data we mean that there are recorded values associated with some geographical region (county, district, country, etc.). It is convenient to model these data using random fields on a lattice. The sites of lattice L correspond to individual regions. The neighbourhood relation $\sim$ may be defined in such a way that two regions are in the relation $\sim$ if and only if they share the common boundary. The data are often formed by counts of certain event (e.g. number of infected people, number of criminal acts). The modelling of discrete spatial data may be based on generalized linear models.

Let $\boldsymbol{Z} = \{Z_i : i \in L\}$ and $\boldsymbol{W} = \{W_i : i \in L\}$ be random fields on lattice L. Assume that conditionally on $\boldsymbol{W}$, the Z_i are independent random variables with mean $\mathbb{E}(Z_i \mid \boldsymbol{W}) = \mu_i$. Next consider function h (so called *link function*) and assume that

$$h(\mu_i) = \boldsymbol{F}_i^{\mathrm{T}} \boldsymbol{\beta} + W_i,$$

where $\boldsymbol{F}_i = (f_{0i}, \ldots, f_{pi})^{\mathrm{T}}$ is the vector of region-specific covariates and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ is the vector of regression parameters. This enables non-linear relationships between data and covariates. For binary data the usual choice of h is the logit function $h(\mu) = \log \frac{\mu}{1-\mu}$. Random field $\boldsymbol{W}$ models spatial variation. It captures spatial dependence present in the data. We can use e.g. one of the Gaussian models (CAR, SAR, SMA, SARMA).

The most common approach when modelling the number of events is to use the Poisson model. Assume that $Z_i \mid \boldsymbol{W}$ has Poisson distribution with parameter $E_i\theta_i$, $i \in L$. Here, E_i is supposed to be known and it represents the expected number of events in the region i. As a link function we use the logarithm and we obtain the linear model for θ_i:

$$\log \theta_i = \boldsymbol{F}_i^{\mathrm{T}} \boldsymbol{\beta} + W_i.$$

Spatial epidemiology is one of the main fields, where this model is used. In this case, Z_i represents observed number of cases of some disease in region i and E_i is the expected number of cases, which can be known from some additional information about the problem or may be some known function of n_i people at risk of the disease. For example, we can have $E_i = rn_i$, where r is the overall infection rate in the whole population. It can be estimated by the ration

$$\frac{\sum_{i \in L} Z_i}{\sum_{i \in L} n_i}.$$

This choice means that we expect the same infection rate in all regions. The value θ_i is the region-specific relative risk. It gives the true relative risk of the infection in region i. As a covariate we can imagine e.g. the level of air pollution, which will have an important contribution when studying respiratory diseases. We can also view the whole situation as the hierarchical model and use Bayesian methods to make the statistical inference.

Now we consider a different model that specifies the joint distribution. Let $\{Z_i : i \in L\}$ be a Markov random field with density $p(\boldsymbol{z}; \boldsymbol{\theta})$ parametrized by finite-dimensional vector $\boldsymbol{\theta}$. For discrete data the density is equal to the joint probabilities $\mathbb{P}(Z_i = z_i, i \in L)$, $\boldsymbol{z} = (z_i, i \in L)$. For continuous data it is the joint density w.r.t. n-dimensional Lebesgue measure.

The maximum likelihood method is one of the most popular statistical methods for estimating the parameters of a model. We find a value $\hat{\boldsymbol{\theta}}$, at which the likelihood function $L(\boldsymbol{\theta}) = p(\boldsymbol{z}; \boldsymbol{\theta})$ attains its maximum. Here $\boldsymbol{z} = (z_i, i \in L)$ are observed data.

If we have a Markov random field with Gibbs distribution, then

$$L(\boldsymbol{\theta}) = p(\boldsymbol{z}; \boldsymbol{\theta}) = \exp\left\{ -\sum_{C \in \mathcal{C}} \Phi_C(\boldsymbol{z}_C, \boldsymbol{\theta}) \right\} = \frac{\exp\left\{ -\sum_{C \in \mathcal{C} : C \neq \emptyset} \Phi_C(\boldsymbol{z}_C, \boldsymbol{\theta}) \right\}}{\int \exp\left\{ -\sum_{C \in \mathcal{C} : C \neq \emptyset} \Phi_C(\boldsymbol{z}_C, \boldsymbol{\theta}) \right\} \nu(\mathrm{d}\boldsymbol{z})},$$

where $\mathcal{C}$ is the system of cliques (subsets of L, for which any two sites are neighbours). The problem is that the normalizing constant depends on $\boldsymbol{\theta}$ and it usually has very complicated form. There exist methods for approximation of the normalizing constant by means of simulations (mostly MCMC methods). Then we maximize this approximated likelihood function.

More similar procedure is to consider so called *pseudolikelihood*

$$L_P(\boldsymbol{\theta}) = \prod_{i\in L} p(z_i \mid \boldsymbol{z}_{\partial i}; \boldsymbol{\theta}) = \prod_{i\in L} \frac{\exp\left\{-\sum_{C\in\mathcal{C}:C\neq\emptyset, i\in C} \Phi_C(\boldsymbol{z}_C, \boldsymbol{\theta})\right\}}{c(\boldsymbol{z}_{\partial i}, \boldsymbol{\theta})}.$$

The normalizing constant $c(\boldsymbol{z}_{\partial i}, \boldsymbol{\theta})$ is often easier to express (in discrete case it is the sum of $|S|$ terms, where S is the state space of the random field). If we enumerate the elements of L by $1, \dots, n$, then the likelihood could be written as

$$L(\boldsymbol{\theta}) = p(z_1 \mid z_2, \dots, z_n; \boldsymbol{\theta}) p(z_2 \mid z_3, \dots, z_n; \boldsymbol{\theta}) \cdots p(z_{n-1} \mid z_n; \boldsymbol{\theta}) p(z_n; \boldsymbol{\theta}).$$

Replacing conditional densities $p(z_k \mid z_{k+1}, \dots, z_n; \boldsymbol{\theta})$ by full conditional densities $p(z_k \mid \boldsymbol{z}_{-k}; \boldsymbol{\theta})$, which are equal to $p(z_k \mid \boldsymbol{z}_{\partial k}; \boldsymbol{\theta})$ thanks to the Markov property, we obtain the pseudolikelihood $L_P(\boldsymbol{\theta})$.

The maximum pseudolikelihood estimators belongs to the class of estimators that are known in statistics as M-estimators. Generally, an M-estimator of $\boldsymbol{\theta}$ is obtained as the maximum of contrast function $\varrho(\boldsymbol{Z}, \boldsymbol{\theta})$. In the classical situation of the maximum likelihood estimation for the sequence of i.i.d. random variables we have

$$\varrho(\boldsymbol{z}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(z_i; \boldsymbol{\theta}).$$

In our case we get

$$\varrho(\boldsymbol{z}, \boldsymbol{\theta}) = \sum_{i\in L} \log p(z_i \mid \boldsymbol{z}_{\partial i}; \boldsymbol{\theta}).$$

4.2 Testing of spatial autocorrelation

Recall that for the random field $\boldsymbol{Z} = \{Z_i : i \in L\}$ with constant mean $\mathbb{E} Z_i = \mu$ and constant variance $\operatorname{var} Z_i = \sigma^2$ we have defined *Moran's I* by the relation

$$I = \frac{n}{w} \frac{\sum_{i\in L}\sum_{j\in L} w_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i\in L}(Z_i - \bar{Z})^2}$$

and *Geary's c* as

$$c = \frac{n-1}{2w} \frac{\sum_{i\in L}\sum_{j\in L} w_{ij}(Z_i - Z_j)^2}{\sum_{i\in L}(Z_i - \bar{Z})^2},$$

where w_{ij} are spatial proximity weights (i.e. we require $w_{ij} = 0$ if $i = j$ or $i \not\sim j$) and $w = \sum_{i\in L}\sum_{j\in L} w_{ij}$. For the computation of Moran and Geary index one can use functions moran and geary, respectively, in R library spdep.

These characteristics can be used as test statistics for testing the hypothesis of no spatial autocorrelation in data. Denote by M one of these test statistics (either Moran's I and Geary's c) and by M_{obs} this test statistic computed from data.

Let us consider two different assumptions that correspond to the null hypothesis:

1. randomness assumption: all $n!$ permutations of observed values at n sites of L have equal probability $1/n!$,
2. Gaussianity assumption: random field $\boldsymbol{Z}$ is formed by independent random variables with normal distribution $N(\mu, \sigma^2)$.

One of the following three approaches is usually used for testing under the randomness assumption.

1. *permutation test*: The null hypothesis H_0 means that the observed values Z_i, $i \in L$, are assigned completely at random. For n sites we have $n!$ possible assignments. If we compute M for all $n!$ possibilities, we obtain the distribution of M under H_0. Then we can determine the probability that the value M_{obs} is exceeded. Both large and small values of this probability indicate against H_0 (if we consider two-sided test).
2. *Monte Carlo test*: Even if n is not very large the corresponding number of permutations could be huge. Instead of calculating M for all permutations we can generate k random permutations and

construct the empirical distribution of M under H_0. Larger k means better approximation of the true distribution under H_0. We take together M_{obs} with k values of M from generated permutations and order them from the smallest to the largest. For extreme rank values of M_{obs} the null hypothesis should be rejected. For example, if $k = 999$ we reject H_0 on the level 5% when the rank of M_{obs} is between 1 and 25 or between 976 and 1 000.

3. *asymptotic test*: Denote by $\mathbb{E}_r M$ and $\operatorname{var}_r M$ the expectation and variance of M under H_0 and randomness assumption, respectively. These first two moments can be determined analytically. Since one can often show the asymptotic normality of M, it suffices to compare
$$\frac{M_{obs} - \mathbb{E}_r M}{\sqrt{\operatorname{var}_r M}}$$
with quantiles of the standard normal distribution $N(0, 1)$.

Under the assumption of Gaussianity it is not difficult to express the expectation and variance of M when H_0 holds. Denote these moments by $\mathbb{E}_g M$ and $\operatorname{var}_g M$, respectively. Again we can consider the asymptotic test and compare
$$\frac{M_{obs} - \mathbb{E}_g M}{\sqrt{\operatorname{var}_g M}}$$
with quantiles of the standard normal distribution $N(0, 1)$.

It can be shown that $\mathbb{E}_g I = \mathbb{E}_r I = -\frac{1}{n-1}$ and $\mathbb{E}_g c = \mathbb{E}_r c = 1$. The expectations are the same when assuming randomness and Gaussianity. However, the formulas for variances already differ (see [2]). Moran's and Geary's statistic can be interpreted as follows: if $I > \mathbb{E}I$ or $c < \mathbb{E}c$, then the lattice site has the tendency to be connected to the site with similar value of the random field. It corresponds to positive spatial autocorrelation. On the contrary, if $I < \mathbb{E}I$ or $c > \mathbb{E}c$, the values at neighbouring sites have tendency to be dissimilar. Therefore, we can also consider one-sided variants of the test against the alternative that the spatial autocorrelation is positive (or negative).

5. Appendix

5.1 Random censoring

Assume that $T_1, \dots, T_n$ are independent and identically distributed non-negative random variables with distribution function F. Our aim is to estimate F. If we observe all values of $T_1, \dots, T_n$, the most natural estimator of F is the empirical distribution function
$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[T_i \le t]}, \quad t \ge 0.$$
However, in some situations we don't have information about all T_i. In particular, T_i could represent the times to some event and their observation is prematurely interrupted. A typical example is the medical study of the influence of some treatment to the survival of the group of patients. Some of the observations are incomplete because the patient moved away or the time reserved for the study expired. Another example comes from the reliability theory where we measure the times to the breakdown of some product. Except of the random variables T_i (so called *survival times* or *life times*) we also consider random variables $C_1, \dots, C_n$ (so called *censoring times*). We observe the random sample $(\tilde{T}_1, D_1), \dots, (\tilde{T}_n, D_n)$, where $\tilde{T}_i = \min(T_i, C_i)$ are censored survival times and $D_i = \mathbf{1}_{[T_i \le C_i]}$ are indicators of non-censoring. For $D_i = 1$ we observe the true time T_i. while for $D_i = 0$ the censoring happened and we have only partial information about T_i, namely $T_i \ge \tilde{T}_i$. In the case of random censoring we assume that $C_1, \dots, C_n$ are i.i.d. random variables and independent of $T_1, \dots, T_n$. Then the non-parametric maximum likelihood estimator of F is the Kaplan-Meier estimator introduced in [4] and defined as
$$\hat{F}_{KM}(t) = 1 - \prod_{s \le t} \left(1 - \frac{\#\{i : \tilde{T}_i = s, D_i = 1\}}{\#\{i : \tilde{T}_i \ge s\}} \right).$$
The product is effectively consisting only of finitely many terms that correspond to the times s, at which some life time is realized. The estimator $\hat{F}_{KM}(t)$ is always a non-decreasing and right-continuous

function. Its limit as $t \to \infty$ could be strictly smaller than 1. This happens when the largest observed value is censored.

The intuitive explanation of the Kaplan-Meier estimator is the following. Divide the interval $[0, t)$ into smaller intervals $[0, t_1), [t_1, t_2), \dots, [t_k, t)$. Then

$$1 - F(t) = \mathbb{P}(T_1 > t) = \mathbb{P}(T_1 > t \mid T_1 \geq t_k) \cdot \mathbb{P}(T_1 \geq t_k \mid T_1 \geq t_{k-1}) \cdots \mathbb{P}(T_1 \geq t_2 \mid T_1 \geq t_1) \cdot \mathbb{P}(T_1 \geq t_1),$$

where the conditional probabilities

$$\mathbb{P}(T_1 \geq t_j \mid T_1 \geq t_{j-1}) = 1 - \mathbb{P}\left(T_1 \in [t_{j-1}, t_j) \mid T_1 \geq t_{j-1}\right)$$

could be estimated by

$$1 - \frac{\#\{i : \tilde{T}_1 \in [t_{j-1}, t_j), D_i = 1\}}{\#\{i : \tilde{T}_1 \geq t_{j-1}\}}.$$

Making the intervals $[t_{j-1}, t_j)$ smaller we get in the limit the expression $\hat{F}_{KM}(t)$.

References

[1] A. Baddeley and R. Turner (2005): Spatstat: an R package for analyzing spatial point patterns, *J. Stat. Softw.* **12**, 1–42.

[2] A. D. Cliff and J. K. Ord (1981): *Spatial Processes; Models and Applications*, Pion Limited, London.

[3] Y. Guan (2006): Tests for independence between marks and points of a marked point process, *Biometrics* **62**, 126–134.

[4] E. L. Kaplan and P. Meier (1958): Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.* **53**, 457–481.

[5] D. G. Krige (1951): A statistical approach to some basic mine valuation problems on the Witwatersrand, *J. Chem. Metal. Min. Soc. S. Afr.* **52**, 119–139.

[6] P. Lachout (2004): *Teorie pravděpodobnosti*, second edition, in Czech, Karolinum, Praha.

[7] J. Møller and R. P. Waagepetersen (2003): *Statistical Inference and Simulation for Spatial Point Processes*, Chapman & Hall/CRC, Boca Raton.

[8] J. Ohser (1983): On estimators for the reduced second-moment measure of point processes, *Math. Operationsf. Statist., Ser. Statistics* **14**, 63–71.

[9] E. J. Pebesma (2004): Multivariable geostatistics in S: the gstat package, *Computers & Geosciences* **30**, 683–691.

[10] P. J. Ribeiro Jr and P. J. Diggle (2001): geoR: a package for geostatistical analysis, *R-NEWS* **1**, 15–18.

[11] B. D. Ripley (1976): The second-order analysis of stationary point processes, *J. Appl. Probab.* **13**, 255–266.

[12] M. Schlather, P. J. Ribeiro Jr. and P. J. Diggle (2004): Detecting dependence between marks and locations of marked point processes, *J. R. Statist. Soc. B* **66**, 79–93.