Prostorová statistika (NMST543)

verze 9. 1. 2015

Obsah

1. Statistika bodových procesů

V této kapitole se budeme zabývat statistickou analýzou jednoduchých bodových procesů na $\mathbb{R}^d$. Nejprve začneme odhady popisných charakteristik, poté se podíváme na testy hypotéz, volbu a diagnostiku parametrického modelu.

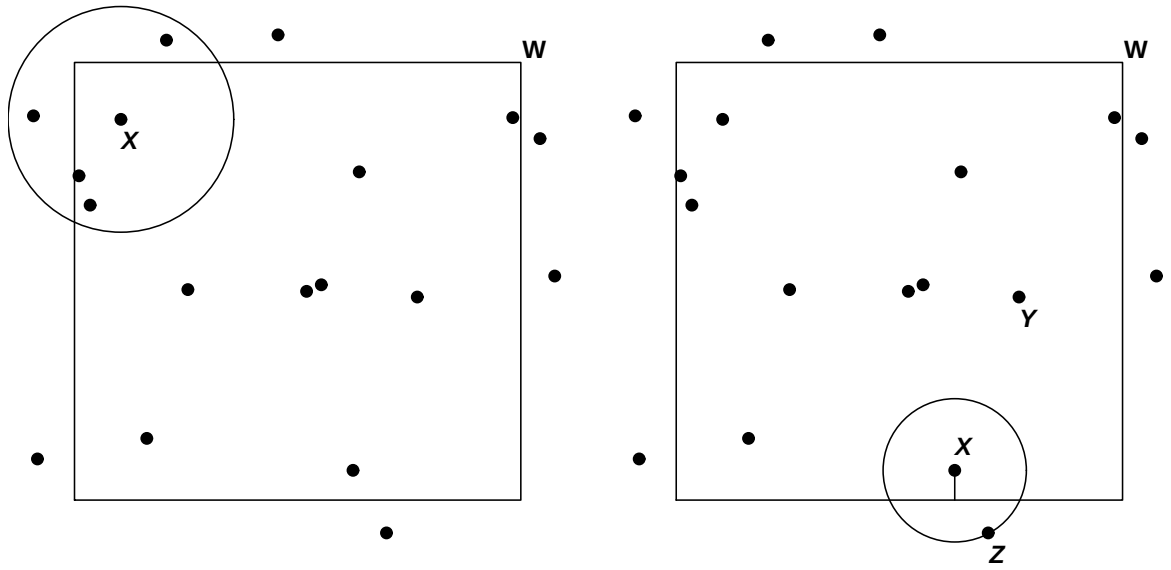
Připomeňme, že bodový proces je definován jako náhodná celočíselná lokálně konečná míra. Na jednoduchý bodový proces můžeme také pohlížet jako na náhodnou lokálně konečnou množinu. Budeme využívat oba přístupy, takže budeme psát $\Phi(B)$ pro počet bodů (atomů) procesu Φ v množině B a $X \in \Phi$ pro situaci, kdy X je bod (atom) procesu Φ.

1.1 Odhady popisných charakteristik

Předpokládejme, že máme realizaci bodového procesu Φ na množině $W \in \mathcal{B}_0^d$, tzv. *okno pozorování*. Okno je obvykle d-rozměrný obdélník, ale může mít i komplikovanější tvar. Cílem je odhadnout charakteristiky procesu Φ na základě dané realizace. Podáme přehled základních odhadů, které jsou vesměs implementovány v balíčku spatstat, a tak zároveň uvedeme příslušné příkazy.

Okrajové efekty

Největší problém při odhadování číselných a funkcionálních popisných charakteristik hrají tzv. *okrajové efekty (edge effects)*, které jsou způsobeny tím, že bodový proces pozorujeme v omezeném okně. Například odhad K-funkce bychom mohli založit na počtech bodů v koulích se středem v bodě procesu a poloměru r. Pokud je ovšem vzdálenost bodu procesu od hranice okna menší než r, tak tento počet z dat nezjistíme. Situace je znázorněna na obrázku 1 vlevo – skutečný počet v kouli $b(X, r)$ je pět, ale z pozorování v okně W vidíme pouze tři body. Jako jiný příklad uveďme situaci při odhadu G-funkce, kdy hledáme nejbližšího souseda bodu X procesu. Z informace, kterou máme z okna W, bychom na obrázku 1 vpravo za nejbližšího souseda bodu X označili bod Y. Ve skutečnosti je však nejbližším sousedem bodu X bod Z, který leží mimo okno W. Je tedy vidět, že zanedbáním okrajových efektů můžeme dostat zkreslené závěry o charakteristikách procesu.



Obrázek 1. Ilustrace okrajových efektů v případě odhadování K-funkce (vlevo) a G-funkce (vpravo).

Odhad funkce intenzity

Nechť Φ je stacionární bodový proces na $\mathbb{R}^d$ s intenzitou λ. Přímo z definice plyne, že

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

je nestranný odhad λ, v balíčku spatstat ho lze zjistit pomocí funkce summary.ppp. Pokud Φ je homogenní Poissonův proces, tak $\hat{\lambda}$ je dokonce maximálně věrohodný odhad. Můžeme totiž Φ na množině W chápat

jako konečný bodový proces s hustotou vzhledem k jednotkovému Poissonovu procesu na W, přitom víme, že hustota je

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

Lehce se zjistí, že věrohodnostní funkce $L(\lambda) = p_\lambda(\varphi)$ nabývá maxima pro $\lambda = \varphi(W)/|W|$.

Pro nestacionární bodové procesy s funkcí intenzity λ se používá *jádrový (kernel)* neparametrický odhad (density.ppp)

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

kde k_b je jádrová funkce se *šířkou pásma (bandwidth)* nebo také *vyhlazovacím okénkem* $b > 0$, tj. $k_b(x) = \frac{k(x/b)}{b^d}$, kde k je nějaká pravděpodobnostní hustota, a

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

je *korekce na okrajové efekty*. Jiná možnost je použít přesnější ale výpočetně náročnější odhad (density.ppp s volbou diggle=TRUE)

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

Odhad $\hat{\lambda}(x)$ je obvykle citlivý na volbu šířky pásma, zatímco volba jádrové funkce není tak důležitá. Pro malé hodnoty b je odhad příliš koncentrován kolem bodů procesu, pro velké hodnoty b dochází k velkému vyhlazení. Mezi nejobvyklejší volby funkce k patří hustota rovnoměrného rozdělení na jednotkové kouli nebo hustota d-rozměrného normálního rozdělení. Často se také volí k jako součin jednorozměrných hustot: $k(x) = k_1(x_1) \cdots k_d(x_d)$ pro $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Oblíbeným příkladem jednorozměrné jádrové funkce je *Epanečnikovovo jádro*:

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

Všimněme si, že pokud je k symetrická, je druhý z uvedených odhadů nestranný v tom smyslu, že platí

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

neboli

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

Pro nestacionární Poissonův proces je možné opět explicitně vyjádřit věrohodnostní funkci:

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

Pokud funkce intenzity $\lambda_\theta(x)$ má parametrický tvar, pak úlohu nalezení maximálně věrohodného odhadu parametru θ je třeba řešit některým numerickým postupem. Ještě se k tomu dostaneme v podkapitole 1.3.

Odhad K-funkce

Připomeňme, že K-funkce $K(r)$ je definována pomocí vztahu

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

který lze ekvivalentně přepsat na

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

kde $A \in \mathcal{B}_0^d$ je libovolná množina s kladnou Lebesgueovou mírou ($|A| > 0$).

Následující odhady lze v knihovně spatstat dostat pomocí funkce Kest.

0. nekorigovaný odhad (uncorrected estimate): Na základě vztahu (1) bychom teoreticky mohli uvažovat nestranný odhad $\lambda^2 K(r)$ tvaru

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

Tento odhad je ale použitelný pouze, pokud máme dodatečnou informaci z vnějšku okna W o bodech Y, které leží ve vzdálenosti nejvýše r od bodů procesu Φ ležících v okně, tzv. *plusový výběr (plus sampling)*. Problém spočívá v okrajových efektech: nejsme schopni vyčíslit $\Phi(b(X,r) \setminus \{X\})$ jenom z informace uvnitř okna W, viz obrázek 1 vlevo. Pokud bychom ignorovali okrajové efekty a uvažovali pouze body uvnitř okna, dostaneme záporně vychýlený odhad:

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

Knihovna spatstat umožňuje spočítat tento odhad volbou correction="none". Je to však jen z instruktážních důvodů. V praxi je tento odhad nepoužitelný.

1. minusový odhad (border estimate), correction="border": Nejjednodušším postupem, jak se vyhnout okrajovým efektům, je uvažovat Φ v menším okně

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

tzv. *minusový výběr (minus sampling)*, kde ∂W značí hranici množiny W. Pak

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

je nestranný odhad $\lambda^2 K(r)$, jak plyne z (1). Tento odhad je definovaný pro $r < r_b = \sup\{s > 0 : |W_{\ominus s}| > 0\}$, například pro $W = [0,1]^2$ je $r_b = 0{,}5$.

2. translačně korigovaný odhad, correction="translate": Jiná možnost je využít korekčních faktorů, což jsou jakési váhy přiřazené tomu, že pozorujeme dva body v určité vzdálenosti. Použijeme-li *translanční korekční faktor*, dostaneme

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

Důkaz nestrannosti tohoto odhadu je založen na Campbellově větě (viz cvičení). Tento odhad je dobře definovaný pro $r < r_t = \sup\{s > 0 : |W \cap (W + x)| > 0 \ \forall x : \|x\| \leq s\}$, například pro $W = [0,1]^2$ je $r_t = 1$. Podobně můžeme definovat odhad redukované momentové míry druhého řádu

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

Odhad jádrově vyhlazené hustoty této míry se dá spočítat příkazem Kmeasure.

3. Ripleyho izotropicky korigovaný odhad, correction="isotropic" nebo correction="Ripley": Jiný korekční faktor navrhl B. D. Ripley [8], vede na odhad

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

Pokud je proces navíc izotropní, lze ukázat, že se jedná o nestranný odhad pro $r < r_0 = \inf\{t > 0 : |W^{(t)}| < |W|\}$, kde $W^{(t)} = \{x \in W : \partial b(x,t) \cap W \neq \emptyset\}$. Ohserova [7] modifikace

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

rozšiřuje definici na $r < r^* = \sup\{s > 0 : |W^{(s)}| > 0\}$. Pro $r < r_0$ je $\widehat{\lambda^2 K_R(r)} = \widehat{\lambda^2 K_O(r)}$. V případě $W = [0,1]^2$ je $r_0 = \sqrt{2}/2$, $r^* = \sqrt{2}$ a $W^{(s)} = W$ pro všechna $s \le r_0$.

Při odhadování $K(r)$ je třeba dělit odhadem druhé mocniny intenzity, což má za následek porušení nestrannosti. Vychýlení a rozptyl se typicky zvětšují s rostoucím r. Pro obdélníkové okno se doporučuje odhady počítat pro r menší než čtvrtina kratší strany obdélníku. Jako odhad druhé mocniny intenzity se často používá

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

který je nestranný pro Poissonův bodový proces Φ.

Minusový odhad K-funkce nemusí být monotónní funkce v r (na rozdíl od teoretické funkce). S rostoucím r a dimenzí prostoru dochází u minusové metody k významné ztrátě informace z dat. Statisticky lepší vlastnosti mají odhady založené na korekčních faktorech. Na druhou stranu výpočet $\hat{K}_b$ je rychlejší.

Odhad nehomogenní K-funkce

Pro po převážení funkcí intenzity slabě stacionární bodové procesy lze provést odhad nehomogenní K-funkce

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

podobnými postupy jako při odhadu K-funkce pro stacionární procesy, např. translačně korigovaný odhad má tvar

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

kde $\hat{\lambda}(x)$ je odhad funkce intenzity v bodě x. V balíčku spatstat bychom použili Kinhom s volbou correction="translate".

Odhad párové korelační funkce

Pro stacionární a izotropní bodový proces souvisí párová korelační funkce g s K-funkcí:

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

Lze použít jádrový odhad s translačním nebo Ripleyho korekčním faktorem:

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

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

kde k_b je vhodná jádrová funkce s vhodnou šířkou pásma b. V balíčku spatstat lze párovou korelační funkci odhadnout pomocí pcf. Odhad $\hat{g}_t$ odpovídá volbě correction="translate", zatímco odhad $\hat{g}_R$ volbě correction="ripley". V případě po převážení funkcí intenzity slabě stacionárních bodových procesů bychom každý člen součtu místo $\widehat{\lambda^2}$ dělili součinem $\hat{\lambda}(X)\hat{\lambda}(Y)$, k výpočtu slouží funkce pcfinhom.

Další možnost je využít některý z odhadů K-funkce a aproximovat derivaci numerickými metodami (např. pomocí splinů). To obvykle není snadné, protože odhad K-funkce je po částech konstantní funkce.

Odhad distribuční funkce vzdálenosti nejbližšího souseda

Připomeňme, že pro stacionární bodový proces definujeme distribuční funkci vzdálenosti nejbližšího souseda jako

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

K výpočtu následujících odhadů funkce G lze použít Gest.

0. nekorigovaný odhad, correction="none": Kdybychom pro každý pozorovaný bod procesu znali vzdálenost k nejbližšímu sousedu, tak můžeme odhadnout distribuční funkci vzdálenosti nejbližšího souseda klasickým způsobem jako empirickou distribuční funkci

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

kde $e(x) = d(x, \Phi \setminus \{x\})$ je vzdálenost bodu x k nejbližšímu sousedu. Z Campbellovy-Meckeho věty plyne, že

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

Odtud vidíme, že odhad $\hat{G}(r)$ je tzv. *podílově nestranný (ratio-unbiased)*, což znamená, že $\hat{G}(r)$ je tvaru zlomku, přičemž podíl středních hodnot čitatele a jmenovatele dává $G(r)$, neboli

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

Opět díky okrajovým efektům nejsme schopni získat $e(X)$ pro každé $X \in \Phi \cap W$, viz obrázek 1 vpravo. Pokud nahradíme $e(X)$ vzdáleností $e^*(X) = d(X, (\Phi \setminus \{X\}) \cap W) \ge e(X)$, kterou jsme schopni pozorovat, dostáváme následující naivní odhad

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

Jelikož $e^*(X) \le r$ implikuje $e(X) \le r$, je $\hat{G}_r(r) \le \hat{G}(r)$.

1. minusový odhad, correction="border" nebo "rs": Přejdeme-li opět k erodovanému oknu $W_{\ominus r}$, dostaneme podílově nestranný odhad

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

2. Kaplanův-Meierův odhad, correction="km": Okrajové efekty lze chápat jako druh cenzorování (viz podkapitola 5.1). Můžeme tak zavést odhad Kaplanova-Meierova typu:

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

kde $c(x) = d(x, \partial W)$ je vzdálenost x od hranice okna. V případě $e(X) \le c(X)$ víme, že pozorujeme skutečnou vzdálenost k nejbližšímu sousedu bodu X. V opačném případě je vzdálenost $e(X)$ cenzorována hodnotou $c(X)$. Máme jistotu, že $e(X)$ bude větší než $c(X)$. Uvědomme si, že k výpočtu odhadu $\hat{G}_{KM}(r)$ nám stačí informace, kterou máme z okna W. Na rozdíl od klasické situace náhodného cenzorování nemůžeme očekávat nezávislost pozorování a cenzorů, a tak je optimalita Kaplanova-Meierova odhadu narušena. Přesto ve srovnání s minusovým odhadem dává obvykle přesnější výsledky.

Pokud máme absolutně spojitou distribuční funkci $H(t)$ s hustotou $h(t)$, tak *riziková funkce (hazard rate)* je definována jako $\lambda_h(t) = h(t)/(1 - H(t))$. Prostorová Kaplanova-Meierova metoda umožňuje odhadnout rizikovou funkci $\lambda_h(r)$ distribuční funkce $G(r)$. Musíme však být opatrní, protože G nemusí mít nutně hustotu. V balíčku spatstat se tento odhad počítá společně s Kaplanovým-Meierovým odhadem G-funkce.

3. Hanischův odhad, correction="han" nebo "Hanisch": Jiné vylepšení minusového odhadu přináší následující korekce na okrajové efekty:

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

kde

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

Tento odhad používá pouze body, které jsou blíže k svému nejbližšímu sousedu než k hranici okna. Na obrázku 1 vpravo by bod X nebyl zahrnut do odhadu, neboť jeho vzdálenost k hranici okna je menší k svému nejbližšímu sousedu. Hanischův odhad je podílově nestranný, o čemž se můžeme snadno přesvědčit z Campbellovy-Meckeho věty, uvědomíme-li si, že $\mathbf{1}_{[e(X) \le c(X)]} = \mathbf{1}_{[X \in W_{\ominus e(X)}]}$.

Odhady G nemusí mít vlastnosti distribuční funkce: $\hat{G}_b$ nemusí být monotónní, $\hat{G}_{KM}$ je neklesající, ale maximální hodnota může být menší než 1.

Odhad kontaktní distribuční funkce

Kontaktní distribuční funkci stacionárního bodového procesu Φ jsme definovali jako

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

kde D je vzdálenost od počátku k nejbližšímu bodu procesu Φ.

V prostoru $\mathbb{R}^d$ zvolme pravidelnou mříž I_a:

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

kde $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$ a $a = (a_1, \ldots, a_d) \in \mathbb{R}^d_+$, tj. $a_i > 0$ pro $i = 1, \ldots, d$. K výpočtu následujících odhadů kontaktní distribuční funkce F v rovinném případě lze použít Fest.

0. nekorigovaný odhad, correction="none": Pro každý bod mříže v okně W najděme nejbližší bod procesu, ten ovšem může ležet mimo okno. Pokud budeme uvažovat pouze body $\Phi \cap W$, dostaneme

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

kde $I_a(W)$ je počet prvků množiny $I_a \cap W$. Tento odhad je záporně vychýlený, tj. $\mathbb{E}\hat{F}_r(r) \le F(r)$, neboť $\mathbf{1}_{[d(x,\Phi \cap W) \le r]} \le \mathbf{1}_{[d(x,\Phi) \le r]}$ a $\mathbb{P}(d(x,\Phi) \le r) = F(r)$ ze stacionarity. Vychýlení je způsobeno okrajovými efekty.

1. minusový odhad, correction="border" nebo "rs": Označme $d(x) = d(x, \Phi)$ vzdálenost x od nejbližšího bodu procesu. Potom

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

je nestranný odhad $F(r)$, neboť vzhledem ke stacionaritě $\mathbb{P}(d(x) \le r) = F(r)$. Spojitá verze tohoto odhadu (když $a \to o$) má tvar

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

kde $\Phi_r = \{x \in \mathbb{R}^d : d(x, \Phi) \le r\} = \cup_{X \in \Phi} b(X, r)$. Opět se jedná o nestranný odhad.

2. Kaplanův-Meierův odhad, correction="km":

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

kde $c(x) = d(x, \partial W)$ je vzdálenost x od hranice okna.

Kontaktní distribuční funkce $F(r)$ stacionárního procesu je absolutně spojitá a riziková funkce $\lambda_h(r)$ existuje. Odhad je založen na Kaplanově-Meierově odhadu $\hat{F}_{KM}(r)$.

3. Chiův-Stoyanův odhad, correction="cs" nebo "Hanisch": Použitím stejné korekce na okrajové efekty jako u Hanischova odhadu G-funkce dostaneme

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

kde

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

Spojitá verze tohoto odhadu má tvar

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

kde

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

Odhady F nemusí mít stejné vlastnosti jako teoretická funkce F: $\hat{F}_b$ nemusí být spojitá ani monotónní, $\hat{F}_{KM}$ je neklesající, ale maximální hodnota může být menší než 1. Minusový odhad je méně vydatný než Kaplanův-Meierův nebo Chiův-Stoyanův odhad.

Odhad J-funkce

V balíčku spatstat je možné J-funkci

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

odhadnout pomocí Jest.

Odhad J-funkce vychází z její definice:

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

Rozlišíme následující odhady (podle toho, jaké odhady G a F použijeme):

- *nekorigovaný (*correction="none"*)*,
- *minusový (*correction="border" *nebo* "rs"*)*,
- *Kaplanův-Meierův (*correction="km"*)*,
- *Hanischův (*correction="Hanisch"*)*.

I když nekorigované odhady $\hat{G}_r$ a $\hat{F}_r$ jsou výrazně vychýlené, tak jejich podílem se dostane přibližně nestranný odhad (alespoň když daný bodový proces je blízký Poissonovu procesu). Výhodou tohoto odhadu je necitlivost vzhledem k okrajovým efektům, měl by se tedy použít, pokud jsou okrajové efekty významné.

Další tři odhady jsou mírně vychýlené (podíl dvou přibližně nestranných odhadů). Logaritmus Kaplanova-Meierova odhadu je nestranný odhad $\log J$.

Knihovna spatstat umožňuje odhadnout čtyři základní sumární statistiky (funkce F, G, J, K) najednou pomocí allstats.

Odhad indexu agregace

Pro každou nezápornou náhodnou veličinu T platí mezi její střední hodnotou $\mathbb{E}T$ a distribuční funkcí $H(t)$ následující vztah (např. [5], lemma 5.7)

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

Máme-li odhad $\hat{G}(t)$ distribuční funkce nejbližšího souseda $G(t)$, můžeme tak dostat odhad Clarkova-Evansova indexu

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

jako

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

V knihovně spatstat je pro odhad CE určena funkce clarkevans.

1.2 Testování hypotéz

Další statistickou úlohou je testování hypotézy, zda pozorovaný bodový vzorek odpovídá zvolenému modelu bodového procesu. Nejdůležitějším případem je testování úplné prostorové náhodnosti. Pokud tuto hypotézu nezamítneme, pak můžeme data modelovat Poissonovým procesem a není třeba uvažovat složitější procesy. Tento postup je jedním ze základních kroků průzkumové analýzy dat.

Rozdělme okno pozorování W na k navzájem disjunktních oblastí stejného obsahu a spočtěme počty bodů v každé z těchto oblastí, označme je $n_1, \ldots, n_k$. Za hypotézy, že data pochází z homogenního Poissonova procesu, jsou tyto počty realizace náhodného výběru z Poissonova rozdělení s parametrem $\lambda|W|/k$. Navíc víme, že za hypotézy je všech $n = n_1 + \cdots + n_k$ bodů nezávislé rovnoměrně rozmístěno ve W. Můžeme tak použít klasický Pearsonův χ^2-test dobré shody. Testová statistika je dána jako

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

a rovná se *indexu disperze*

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

kde $\bar{n} = \frac{1}{k}\sum_{i=1}^{k} n_i = n/k$ je průměrný počet bodů na jednu oblast a $s^2 = \frac{1}{k-1}\sum_{i=1}^{k}(n_i - \bar{n})^2$ je výběrový rozptyl. Index I má přibližně χ^2-rozdělení o $k-1$ stupních volnosti. Praktické doporučení pro dobrou aproximaci je $\bar{n} > 5$. Malé hodnoty I odpovídají menší variabilitě než u Poissonova procesu (známka regularity procesu). Velké hodnoty I naopak svědčí o větší variabilitě v bodovém vzorku (shlukování). V knihovně spatstat existuje tento test pod názvem quadrat.test.

Test založený na indexu disperze je jeden z mála případů v prostorové statistice, kdy (asymptotické) rozdělení testové statistiky je jednoduché. Vzhledem k tomu, že často je rozdělení statistik velmi komplikované, používají se simulační (Monte Carlo) testy. Proto nyní vysvětlíme jejich obecnou myšlenku.

Dejme tomu, že chceme testovat hypotézu H_0, že data odpovídají danému modelu. Uvažujme nějakou vhodnou statistiku T, jejíž odhad na základě dat označíme $\hat{T}$. Provedeme M simulací z modelu platného za H_0 a z každé simulace odhadneme statistiku T. Odhady $\hat{T}_1, \ldots, \hat{T}_M$ uspořádáme podle velikosti od nejmenšího k největšímu, čímž dostaneme pořádkové statistiky $\hat{T}_{(1)} \le \cdots \le \hat{T}_{(M)}$. Za nulové hypotézy jsou $\hat{T}$ a $\hat{T}_1, \ldots, \hat{T}_M$ nezávislé a stejně rozdělené, proto ze symetrie je pravděpodobnost, že $\hat{T}$ bude menší než $\hat{T}_{(q)}$ rovna $q/(M+1)$. Chceme-li testovat hypotézu H_0 na hladině významnosti α, určíme takové q, pro které platí

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

Hypotézu pak zamítáme, jestliže $\hat{T} \notin [\hat{T}_{(q)}, \hat{T}_{(M-q+1)}]$. Tento test označujeme jako *bodový Monte Carlo test (pointwise Monte Carlo test)*.

V bodových procesech se spíše pracuje s funkcionálními charakteristikami než s číselnými. Mějme nějakou funkcionální charakteristiku $S(r)$. Pro pevné r, které je nezávisle na datech předem zvolené, můžeme provést výše popsaný Monte Carlo test s volbou $T = S(r)$. Tím ale využíváme jen zlomek informace, kterou nám dává odhad $S(r)$.

Uvažujme odhad $S(r)$ na intervalu $[s_0, s_1]$, kde $0 \le s_0 < s_1$ jsou předem zvolené konstanty. Odhad $S(r)$ získaný z dat označíme $\hat{S}(r)$ a odhady spočtené z M simulací označíme $\hat{S}_1(r), \ldots, \hat{S}_M(r)$. Pro každé r bychom opět mohli určit $\hat{S}_{(q)}(r)$ a $\hat{S}_{(M-q+1)}(r)$. Spojením hodnot $\hat{S}_{(q)}(r)$ pro různá r dostaneme tzv. *dolní obálku (lower envelope)*, zatímco hodnoty $\hat{S}_{(M-q+1)}(r)$ vytvoří *horní obálku (upper envelope)*. K jejich vykreslení lze použít funkci envelope s parametrem global=FALSE. Musíme si uvědomit, že se jedná o bodové obálky, které pomůžou při grafickém znázornění případných odchylek od nulové hypotézy, ale bylo by chybou interpretovat je tak, že hypotézu zamítneme, pokud se odhad $\hat{S}(r)$ někde dostane mimo obálky. Jedná se o problém vícenásobného testování.

Pro přesný obálkový test předpokládejme, že za nulové hypotézy známe teoretický funkční předpis pro $S(r) = S_0(r)$. Určíme maximální odchylky, o které se odhady liší od teoretické funkce:

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

Hodnoty $D_1, \ldots, D_M$ seřadíme podle velikosti, čímž dostaneme pořádkové statistiky $D_{(1)} \le D_{(2)} \le \cdots \le D_{(M)}$. Nulovou hypotézu zamítáme, jestliže $D > D_{(M-q+1)}$, kde hodnota q je volena podle požadované

hladiny testu: $\alpha = \frac{q}{M+1}$. Tomuto postupu se říká *simultánní Monte Carlo test (simultaneous Monte Carlo test)* a v knihovně spatstat ho lze provést použitím funkce envelope s volbou global=TRUE. Testování si můžeme představit také tak, že zkonstruujeme pás o šířce $2D_{(M-q+1)}$ kolem funkce $S_0(r)$ a pokud $\hat{S}(r)$ leží mimo tento pás (vně obálek) pro nějaké $r \in [s_0, s_1]$, tak hypotézu zamítneme. Jiná možnost je místo supremální vzdálenosti uvažovat integrální. Pro data a pro každou simulaci určíme integrální čtvercové odchylky od teoretické funkce:

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

Nulovou hypotézu zamítáme, jestliže $D > D_{(M-q+1)}$, kde hodnota q je volena podle požadované hladiny testu: $\alpha = \frac{q}{M+1}$. V tomto případě mluvíme o *integrálním Monte Carlo testu*.

Pro test hypotézy úplné prostorové náhodnosti tak můžeme použít některý ze simulačních testů. Jako charakteristika $S(r)$ se obvykle volí F, G, J, K nebo L-funkce. Stejným způsobem lze testovat libovolný model, ze kterého umíme simulovat.

1.3 Odhad parametrů modelu

Další statistickou úlohou je vybrat vhodný model, který popisuje naše data. Dejme tomu, že rozdělení bodového procesu Φ je parametrizováno vektorem θ neznámých parametrů. Naším cílem je najít odhad vektoru θ na základě realizace procesu Φ v omezeném okně $W \in \mathcal{B}_0^d$.

Metoda minimálního kontrastu

V několika málo případech je teoretický tvar nějaké popisné charakteristiky $S(r)$ známý a lze ho vyjádřit jako funkci parametrů modelu: $S(r) = S_\theta(r)$. Příkladem jsou popisné charakteristiky Poissonova procesu nebo párová korelační funkce stacionárního Neymanova-Scottové procesu, která splňuje

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

kde λ_p je intenzita rodičovského procesu a p je hustota bodů shluku. Pokud má funkce p parametrický tvar (jako např. u Thomasové nebo Matérnova shlukového procesu), tak máme $g(x)$ vyjádřenu jako funkci parametrů modelu. Odhad parametrů θ pak můžeme hledat analogicky jako u metody momentů tak, že položíme odhad $\hat{S}(r)$ získaný z dat roven teoretické funkci. Řešením soustavy rovnic $S_\theta(r) = \hat{S}(r)$, kde r nabývá několika různých hodnot, dostaneme odhad θ. Pokud je θ k-rozměrný vektor, potřebujeme vzít alespoň k různých hodnot r, aby soustava $S_\theta(r) = \hat{S}(r)$ mohla mít jednoznačné řešení. Vhodnější se ale zdá hledat θ, pro které budeme minimalizovat odchylku $\hat{S}(r)$ od $S_\theta(r)$ přes nějaký interval $[a, b]$. Definujme

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

kde $0 \leq a < b$ a $p, q > 0$ jsou zvolené konstanty a $w(r)$ je zvolená váhová funkce. Odhad θ pak dostaneme minimalizací funkce $D(\theta)$. Tato metoda se nazývá *metoda minimálního kontrastu (method of minimum contrast)*. Pokud neznáme analytické vyjádření $S_\theta(r)$, můžeme pro pevné θ neznámou hodnotu aproximovat pomocí mnoha simulací z modelu. V knihovně spatstat je pro odhad parametrů metodou minimálního kontrastu (s váhovou funkcí identicky rovnou jedné) k dispozici funkce mincontrast, ve které jsou parametry p a q přednastaveny na hodnoty $p = 2$ a $q = 1/4$ (samozřejmě je možné toto nastavení změnit). Pokud za popisnou charakteristiku $S(r)$ zvolíme K-funkci, umožňuje spatstat hledat odhady metodou minimálního kontrastu pro některé konkrétní parametrické modely bodových procesů pomocí speciálních funkcí lgcp.estK (logaritmicko-gaussovský Coxův proces), matclust.estK (Matérnův shlukový proces) a thomas.estK (Thomasové proces).

Příklad: Nechť Φ je Thomasové bodový proces s parametry λ_p (intenzita rodičovského Poissonova procesu), λ_c (střední počet bodů shluku) a σ^2 (rozptyl normálního rozdělení udávajícího polohu bodu shluku vzhledem k rodičovskému bodu). Potom párová korelační funkce je rovna

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

Lze ji odhadnout pomoci jádrového odhadu s korekcí na okrajové efekty. Máme-li takovýto odhad $\hat{g}(r)$ na intervalu $[a, b]$, potom odhad metodou minimálního kontrastu můžeme založit na funkci

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

Minimalizace tohoto integrálu vyžaduje numerické metody. Všimněme si, že v tomto vyjádření se nevyskytuje parametr λ_c. Ten bychom museli odhadnout jinými postupy.

Metoda maximální věrohodnosti

Jiný přístup je založen na maximální věrohodnosti. Předpokládejme, že Φ je konečný bodový proces s hustotou p vzhledem k rozdělení Π jednotkového Poissonova procesu na omezené množině $B \in \mathcal{B}_0^d$, přitom $p(\varphi) = p_\theta(\varphi)$ je parametrizována vektorem θ neznámých parametrů. Pro jednoduchost uvažujme, že okno pozorování W splývá s B. Maximálně věrohodný odhad θ se obdrží maximalizací věrohodnostní funkce $L(\theta) = p_\theta(\varphi)$, kde φ je pozorovaná realizace procesu Φ. Často je výhodnější maximalizovat logaritmus věrohodnostní funkce, tj. $l(\theta) = \log L(\theta)$. Tvar věrohodnostní funkce je známý pro Poissonův proces s funkcí intenzity λ_θ:

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

Jak jsme již zmínili v podkapitole , v homogenním případě ($\lambda_\theta(x) = \lambda$) je tato funkce maximalizována pro $\lambda = \varphi(W)/|W|$. Pro nehomogenní Poissonův proces není vyjádření maximálně věrohodného odhadu analyticky zvládnutelné a musíme přistoupit k numerickým algoritmům (např. Newtonova-Raphsonova metoda) pro maximalizaci věrohodnostní funkce.

Pro jiné procesy než Poissonův je normující konstanta většinou dána komplikovaným integrálem, který není možné spočítat explicitně. V tom případě se dají použít Monte Carlo metody. Nechť hustota bodového procesu má tvar $p_\theta(\varphi) = h_\theta(\varphi)/c_\theta$, kde h_θ je známá funkce a $c_\theta = \mathbb{E} h_\theta(\Phi_P)$ je neznámá normující konstanta (Φ_P je Poissonův proces na B s jednotkovou intenzitou). Potom $l(\theta) = \log h_\theta(\varphi) - \log c_\theta$. Výhodnější bude maximalizovat poměr věrohodností vzhledem k nějakému pevnému parametru θ_0:

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

Pro první člen známe analytické vyjádření, zatímco druhý můžeme aproximovat metodami MCMC (Markov Chain Monte Carlo). Poměr normujících konstant lze totiž rozepsat

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

kde Π_{θ_0} je rozdělení bodového procesu Φ s hustotou p_{θ_0} (neboli skutečný parametr je θ_0). Přitom předpokládáme, že když $h_{\theta_0}(\varphi) = 0$, tak i $h_\theta(\varphi) = 0$, a využíváme úmluvu $0/0 = 1$. Existují různé MCMC algoritmy pro generování procesu s rozdělením Π_{θ_0}. Ty jsou založeny na konstrukci markovského řetězce $\{\Phi^{(n)}\}$, jehož limitní rozdělení je dáno hustotou p_{θ_0} vzhledem k rozdělení Π bodového procesu Φ_P. Pokud střední hodnotu $\mathbb{E}\frac{h_\theta(\Phi)}{h_{\theta_0}(\Phi)}$ nahradíme pomocí výběrového průměru, dostaneme aproximaci poměru logaritmických věrohodností

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

o které se hovoří jako o *aproximaci pomocí výběru na základě důležitosti (importance sampling approximation)*. Maximalizace $l_{\theta_0,n}(\theta)$, pro kterou se používají obvyklé numerické postupy, dává MCMC aproximaci $\hat{\theta}_n$ maximálně věrohodného odhadu $\hat{\theta}$ parametru θ. Tato aproximace je použitelná, pokud θ_0 je blízko $\hat{\theta}$. Jako θ_0 se většinou volí hrubý odhad získaný nějakou jednodušší ale méně efektivní metodou. Celou proceduru lze iterativně opakovat. Existují alternativní aproximace, o kterých se lze podrobněji dočíst v [6], kap. 8.2.4. a 8.2.5.

Metoda maximální pseudověrohodnosti

Protože věrohodnostní funkce je často komplikována, je jiná strategie odhadování parametrů modelu založena na nahrazení věrohodnostní funkce nějakou její aproximací.

Definice 1. Nechť Φ je konečný bodový proces na $B \in \mathcal{B}_0^d$ s Papangelouovou podmíněnou intenzitou $\lambda_\theta^*(x, \varphi)$, kde θ je vektor neznámých parametrů. Mějme dánu realizaci φ procesu v okně pozorování W, o kterém předpokládáme, že splývá s B. Definujeme *pseudověrohodnost (pseudolikelihood)* vztahem

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

Hodnota $\hat{\theta}$, která maximalizuje $\mathrm{PL}(\theta)$, je *odhad metodou maximální pseudověrohodnosti* vektoru θ.

Poznámka 1. Pro Poissonův bodový proces je $\lambda^*(x, \varphi) = \lambda(x)$ a pseudověrohodnost je identická s věrohodností.

Příklad: Straussův proces má Papangelouovu podmíněnou intenzitu $\lambda^*(x, \varphi) = \beta\gamma^{t_R(x,\varphi)}$, kde $t_R(x, \varphi) = \sum_{y \in \varphi} \mathbf{1}_{[0<\|x-y\|\le R]}$. Neznámé parametry jsou $\beta > 0$, $0 \le \gamma \le 1$ a $R > 0$. Logaritmus pseudověrohodnosti Straussova procesu je

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

kde $S_R(\varphi) = \sum_{\{x,y\} \subseteq \varphi} \mathbf{1}_{[0<\|x-y\|\le R]}$. Položíme-li derivace podle β a podle γ rovny nule, dostaneme rovnice

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

Parametr R považujeme za známý a hledáme řešení soustavy dvou rovnic numericky, výsledkem jsou odhady parametrů β a γ. Uvědomíme-li si, že $t_R(x, \varphi)$ nabývá pouze nezáporných celočíselných hodnot, a položíme-li

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

pak má naše soustava tvar

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

Výhoda předpokladu, že R je známý parametr, spočívá v tom, že potom má Papangelouova podmíněná intenzita log-lineární tvar. Mohli bychom zvolit několik různých hodnot $R_1, \ldots, R_K$ parametru R, pro každou spočítat maximálně pseudověrohodné odhady parametrů β a γ a určit takovou hodnotu, pro kterou bude pseudověrohodnost nejvyšší. Tu pak vezmeme jako odhad parametru R.

Složená věrohodnost

Metoda maximální pseudověrohodnosti patří do obecné třídy statistických metod tzv. *složené věrohodnosti (composite likelihood)*, které se používají v případě, kdy metoda maximální věrohodnost je výpočetně nezvládnutelná nebo není dostupná. U složené věrohodnosti je odhad založen na funkci, která je součinem věrohodností jednodušších složek, a to i v případě, že tyto složky nejsou nezávislé. Konkrétní

tvar složek závisí na kontextu. Pro bodové procesy se nabízí uvažovat součin přes příspěvky jednotlivých bodů nebo dvojic bodů.

Mějme stacionární bodový proces Φ na $\mathbb{R}^d$ se součinovou hustotou druhého řádu $\lambda_\theta^{(2)}$, která je parametrizována pomocí vektoru θ neznámých parametrů. Ze stacionarity plyne, že $\lambda_\theta^{(2)}(x,y) = \lambda_\theta^{(2)}(x-y)$. Předpokládáme, že bodový proces Φ pozorujeme v okně W. Potom hustota dvojice bodů procesu ve W je

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

Jednotlivé dvojice bodů samozřejmě nejsou nezávislé, ale i tak můžeme uvažovat součin hustot přes jednotlivé dvojice různých bodů. Po zlogaritmování máme logaritmus složené věrohodnosti tvaru

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

Pro praktické účely nás nezajímají dvojice bodů ve velké vzdálenosti, protože u nich se často projevují jen velmi slabá závislost, takže jejich vynecháním neztrácíme příliš informace. Navíc tím snížíme výpočetní složitost a variabilitu výsledného odhadu. Volíme tedy $R > 0$ a pracujeme s dvojicemi bodů ve vzdálenosti menší než R, dostáváme hustotu

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

a logaritmus složené věrohodnosti

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

Odhad parametru θ získáme maximalizací této funkce. Všimněme si, že ve vyjádření f_θ resp. $\log \mathrm{CL}(\theta)$ můžeme místo součinové hustoty $\lambda_\theta^{(2)}$ použít párovou korelační funkci $g_\theta = \lambda_\theta^{(2)}/\lambda^2$, kde λ je intenzita procesu Φ.

Na rozdíl od předchozích dvou podkapitol nyní pracujeme se stacionárními bodovými procesy. Metoda složené věrohodnosti se využívá především pro Coxovy bodové procesy, kde často máme analytický tvar součinové hustoty druhého řádu. Je-li Φ stacionární Coxův bodový proces s řídící funkcí intenzity Z, jejíž rozdělení závisí na θ, potom $\lambda_\theta^{(2)}(x-y) = \mathbb{E}Z(x)Z(y)$. Dále si předvedeme ještě jednu metodu vhodnou pro Coxovy bodové procesy, která bude založená na jiné charakteristice druhého řádu.

Palmova věrohodnost

Nechť Φ je stacionární bodový proces na $\mathbb{R}^d$ s intenzitou λ a součinovou hustotou druhého řádu $\lambda^{(2)}$. Potom můžeme psát $\lambda^{(2)}(y-x) = \lambda\lambda_o(y-x)$, kde λ_o se nazývá *Palmova intenzita (Palm intensity)*. Faktoriální momentovou míru druhého řádu lze vyjádřit z Campbellovy věty jako

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

Na druhou stranu podle Campbellovy-Meckeho věty je

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

Srovnáním se zjistí, že

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

tj. λ_o je funkce intenzity redukovaného Palmova rozdělení procesu Φ. Odtud je také jasnější, proč se nazývá Palmova intenzita. Uvědomme si, že se jedná o charakteristiku druhého řádu. Palmova intenzita je konstantní pro stacionární Poissonův proces, ale obecně nejde o konstantní funkci.

Budeme uvažovat bodový proces rozdílů pozorovaných bodů procesu Φ v okně W, jejichž vzdálenost je menší než dané R, tj.

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

Zřejmě se jedná o bodový proces v kouli $b(o, R)$. Jeho míra intenzity je

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

a proto má Φ_R funkci intenzity

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

Předpokládáme, že máme parametrický tvar Palmovy intenzity $\lambda_o^\theta(u)$ a chceme odhadnout vektor θ neznámých parametrů. To provedeme tak, že považujeme Φ_R za nehomogenní Poissonův proces s funkcí intenzity $\lambda_R(u)$, kterou aproximujeme tak, že skutečnou intenzitu λ nahradíme pozorovanou intenzitou $\Phi(W)/|W|$ a člen $|W \cap (W - u)|$ nahradíme $|W|$, což je rozumná aproximace pro R podstatně menší než velikost okna. Tím dostaneme pro $\lambda_R(u)$ aproximaci $\Phi(W)\lambda_o(u)$ a pro věrohodnostní funkci aproximaci na základě věrohodnosti pro Poissonův proces. Mluvíme o ní jako o *Palmově věrohodnosti*. Znamená to, že logaritmus Palmovy věrohodnosti je

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

Alternativní způsob, jakým dojdeme k Palmově věrohodnosti, je uvažovat bodové procesy

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

což jsou nehomogenní bodové procesy s funkcí intenzity λ_o. Ignorujeme interakce v procesech $\Phi_X \cap b(o, R)$ a aproximujeme je nehomogenními Poissonovými procesy, jejichž logaritmické věrohodnosti jsou

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

Budeme-li nyní považovat Φ_X, $X \in \Phi \cap W$, za nezávislé stejně rozdělené procesy a ignorujeme okrajové efekty, potom máme logaritmickou Palmovu věrohodnost tvaru

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

což se liší od předchozího vyjádření jen o konstantu.

Takacsova-Fikselova metoda

Při metodě maximální věrohodnosti řešíme soustavu rovnic $l'(\theta) = 0$. U metody momentů zase máme soustavu $S_\theta(r) - \hat{S}(r) = 0$. Oba tyto přístupy se dají zahrnout pod pojem odhadovacích rovnic.

Definice 2. Nechť Φ je bodový proces, jehož rozdělení závisí na parametru θ. Mějme funkci $\psi(\theta, \Phi)$ takovou, že pro každé θ je $\mathbb{E}_\theta \psi(\theta, \Phi) = 0$, kde $\mathbb{E}_\theta$ značí střední hodnotu vzhledem k rozdělení procesu Φ parametrizovaného θ. Pro danou realizaci φ uvažujme rovnici $\psi(\theta, \varphi) = 0$, kterou nazveme *nestrannou odhadovací rovnicí (unbiased estimating equation)* . Jako odhad parametru θ na základě realizace φ vezmeme řešení $\hat{\theta}$ soustavy nestranných odhadovacích rovnic, kterou obdržíme různými volbami funkcí ψ.

Kromě metody momentů a maximální věrohodnosti (či pseudověrohodnosti) je dalším příkladem odhadovacích rovnic pro modely bodových procesů Takacsova-Fikselova metoda. Ta je založena na Georgiiho-Nguyenově-Zessinově identitě

$$\mathbb{E} \sum_{X \in \Phi} h(X, \Phi \setminus \{X\}) = \int_{\mathbb{R}^d} \mathbb{E} h(x, \Phi)\, \Lambda^*(\mathrm{d}x \mid \Phi), \tag{2}$$

kde Λ^* je Papangelouovo jádro. Levá strana je podle Campbellovy-Meckeho věty rovna

$$\int_{\mathbb{R}^d} \mathbb{E}^!_x h(x, \Phi)\, \Lambda(\mathrm{d}x).$$

Je-li míra $\Lambda^*(\cdot \mid \varphi)$ absolutně spojitá vzhledem k Λ s hustotou $\lambda^*(x, \varphi)$ (nazveme podmíněnou intenzitou) a existuje-li funkce intenzity $\lambda(x)$ bodového procesu Φ, pak pro libovolnou měřitelnou g na $\mathbb{R}^d \times \mathcal{N}$ platí

$$\mathbb{E}^!_x g(x, \Phi) = \mathbb{E} \frac{\lambda^*(x, \Phi)}{\lambda(x)} g(x, \Phi).$$

V případě konečného bodového procesu s hustotou p vzhledem k rozdělení jednotkového Poissonova bodového procesu Φ_P na množině $B \in \mathcal{B}^d_0$ je λ^* rovna Papangelouově podmíněné intenzitě.

Předpokládejme, že známe parametrický tvar podmíněné intenzity $\lambda^*_\theta(x, \varphi)$ procesu Φ. Definujeme-li

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

pro libovolně zvolenou funkci h, pak podle Georgiiho-Nguyenovy-Zessinovy identity je $\mathbb{E}_\theta \psi_h(\theta, \Phi) = 0$. Odhad parametru θ dostaneme řešením nestranných odhadovacích rovnic $\psi_h(\theta, \varphi) = 0$. Podobně jako u metody minimálního kontrastu, můžeme zvolit více funkcí h, než je počet neznámých parametrů. Pokud například zvolíme k funkcí $h_1, \ldots, h_k$, můžeme hledat takové θ, které minimalizuje

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

Poznámka 2. Získáme-li odhad $\hat{\theta}$ řešením nestranných odhadovacích rovnic, neznamená to, že se jedná o nestranný odhad parametru θ.

U některých přirozených voleb funkcí h nemusíme být schopni vyjádřit $\psi_h(\theta, \varphi)$ z pozorování φ v omezeném okně W. Problémy totiž mohou způsobovat okrajové efekty. Potom se často nabízí místo $\psi_h(\theta, \varphi)$ vzít odhad $\hat{\psi}_h(\theta, \varphi)$, který uvažuje korekce na okrajové efekty. Například v případě stacionárního bodového procesu je u volby $h(x, \varphi) = \varphi(b(x, r)) \mathbf{1}_{[x \in W]} / |W|$ střední hodnota prvního členu v $\psi_h(\theta, \Phi)$, neboli levá strana identity (2), rovna $\lambda^2 K(r)$. První člen v $\psi_h(\theta, \varphi)$ nejsme schopni spočítat pouze z informace v okně W, a tak ho nahradíme třeba pomocí translačně korigovaného odhadu K-funkce.

Příklad: Uvažujme Straussův proces a předpokládejme, že parametr R je známý. Naším cílem je najít odhad parametrů $\theta = (\beta, \gamma)$. Volba $h_1(x, \varphi) = 1$ dává

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

Volbou $h_2(x, \varphi) = t_R(x, \varphi)$ dostaneme

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

Všimněme si, že jsme získali stejnou soustavu dvou rovnici o dvou neznámých jako u metody maximální pseudověrohodnosti.

1.4 Diagnostika modelu

K ověření zvoleného parametrického modelu můžeme použít Monte Carlo testy. Pokud jsme schopni simulovat z námi zvoleného modelu, pak pro každou nasimulovanou realizaci spočteme nějakou popisnou charakteristiku a porovnáme ji se stejnou charakteristikou odhadnutou z dat. Jestliže je model určen správně, neměla by charakteristika odhadnutá z dat vykazovat velké odlišnosti. Problémy tohoto přístupu se začnou projevovat u obecnějších nehomogenních bodových procesů, kde bychom potřebovali vhodnou popisnou charakteristiku.

Podívejme se nyní, jak se dá v situaci bodových procesů využít zobecnění reziduí z klasických regresních lineárních modelů. Obecně jsou rezidua rozdílem pozorovaných hodnot a hodnot získaných podle zvoleného modelu. Jestliže je model správný, měla by být rezidua kolem nuly. Naopak velká odlišnost od nuly může napovědět, co je chybně ve zvoleném modelu (např. špatně odhadnutý trend nebo interakce).

Definice 3. Nechť Φ je bodový proces s podmíněnou intenzitou λ^*. Pro nezápornou měřitelnou funkci h definujme *h-váženou inovaci* jako znaménkovou náhodnou míru

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

Podle Georgiiho-Nguyenovy-Zessinovy identity (2) je $\mathbb{E} I_h(B) = 0$ pro každé $B \in \mathcal{B}^d$.

Příklad: Nechť Φ je Poissonův bodový proces s funkcí intenzity λ a uvažujme následující tři volby funkce h: $h(x, \varphi) = 1$, $h(x, \varphi) = 1/\lambda^*(x, \varphi)$ a $h(x, \varphi) = 1/\sqrt{\lambda^*(x, \varphi)}$. Vzhledem k tomu, že $\lambda^*(x, \varphi) = \lambda(x)$, dostaneme

$$\begin{aligned} I_1(B) &= \Phi(B) - \int_B \lambda(x) \, \mathrm{d}x, \\ I_{1/\lambda^*}(B) &= \sum_{X \in \Phi \cap B} \frac{1}{\lambda(X)} - |B|, \\ I_{1/\sqrt{\lambda^*}}(B) &= \sum_{X \in \Phi \cap B} \frac{1}{\sqrt{\lambda(x)}} - \int_B \sqrt{\lambda(x)} \, \mathrm{d}x. \end{aligned}$$

Lehce se můžeme přímým výpočtem z Campbellovy věty přesvědčit, že $\mathbb{E} I_h(B) = 0$. Pro rozptyly pak máme

$$\begin{aligned} \operatorname{var} I_1(B) &= \int_B \lambda(x) \, \mathrm{d}x, \\ \operatorname{var} I_{1/\lambda^*}(B) &= \int_B \frac{1}{\lambda(x)} \, \mathrm{d}x, \\ \operatorname{var} I_{1/\sqrt{\lambda^*}}(B) &= |B|. \end{aligned}$$

Tyto vztahy přímo plynou z následujícího lemmatu.

Lemma 1. *Nechť Φ je Poissonův bodový proces na $\mathbb{R}^d$ s mírou intenzity Λ. Pak pro libovolnou nezápornou měřitelnou funkci f platí*

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

Důkaz: Podle Campbellovy věty je

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

Druhý moment můžeme rozepsat podle Campbellovy věty druhého řádu, přičemž využijeme, že faktoriální momentová míra druhého řádu Poissonova procesu je $\Lambda \times \Lambda$:

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

Odtud již dostáváme dokazovaný vztah pro rozptyl.

Předpokládejme, že podmíněná intenzita $\lambda^*_\theta(x,\varphi)$ závisí na parametru θ. Dále předpokládejme, že jsme našli odhad $\hat\theta$ (např. některou metodou z podkapitoly 1.3) na základě pozorování procesu v okně $W \in \mathcal{B}^d_0$. Potom odhad podmíněné intenzity je $\lambda^*_{\hat\theta}(x,\varphi)$. U definice inovací připouštíme, že funkce h závisí na $\hat\theta$.

Definice 4. Znaménkovou náhodnou míru

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

nazveme *h-váženou reziduální mírou*.

Jelikož $\mathbb{E}I_h(B) = 0$, očekáváme, že když bude náš model s $\hat\theta$ správný, budou se hodnoty $R_h(B)$ pohybovat kolem nuly (obecně však $\mathbb{E}R_h(B)$ nemusí být rovno nule). Oblasti s extrémními hodnotami $R_h(B)$ indikují oblasti odchýlení od zvoleného modelu. Mezi speciální volby h patří $h = 1$ (*hrubá rezidua (raw residuals)*), $h = 1/\lambda^*$ (*inverzní rezidua (inverse-lambda residuals)*) a $h = 1/\sqrt{\lambda^*}$ (*Pearsonova rezidua*). Rezidua pro tyto tři volby umožňuje spočítat funkce residuals.ppm v knihovně spatstat. Pro $h = 1$ je

$$R_1(B) = \Phi(B) - \int_B \lambda^*_{\hat\theta}(x,\Phi)\,\mathrm{d}x,$$

tedy míra R_1 je dána rozdílem míry s atomy v bodech procesu a míry s hustotou $\lambda^*_{\hat\theta}(x,\Phi)$ vzhledem k Lebesgueově míře.

Příklad: Uvažujme stacionární Poissonův bodový proces s intenzitou λ, kterou na základě pozorování v okně W odhadneme jako $\Phi(W)/|W|$. Potom (pokud $\Phi(W) > 0$)

$$R_1(B) = \Phi(B) - \Phi(W)\frac{|B|}{|W|},$$

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

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

Dá se ukázat, že střední hodnoty těchto tří h-vážených reziduálních měr jsou nula. Také si můžeme všimnout, že $R_1(W) = R_{1/\lambda^*}(W) = R_{1/\sqrt{\lambda^*}}(W) = 0$. To odpovídá situaci v klasické lineární regresi, kdy součet všech reziduí je roven nule.

Pro grafické znázornění reziduí je vhodné provést vyhlazení pomocí jádrové funkce.

Definice 5. Nechť k je pravděpodobnostní hustota na $\mathbb{R}^d$. Realizaci bodového procesu Φ pozorujeme v okně $W \in \mathcal{B}^d_0$ a máme sestrojený odhad $\hat\theta$ parametru θ. Definujeme *vyhlazené reziduální pole (smoothed residual field)* vztahem

$$S(x) = e(x)\int_W k(x-y)\,R_h(\mathrm{d}y),$$

kde

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

je korekce na okrajové efekty.

Poznámka 3. Pro $h = 1$ je

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

2. Statistika kótovaných bodových procesů

Nechť Φ_{m} je kótovaný bodový proces na $\mathbb{R}^d$ s prostorem kót $\mathbb{M}$. Příslušný nekótovaný bodový proces značíme Φ.

2.1 Odhady charakteristik

Budeme předpokládat, že pozorujeme kótovaný bodový proces Φ_{m} v omezeném okně $W \in \mathcal{B}_0^d$. Odhady charakteristik kótovaných bodových procesů jsou většinou buď přímočaré z definice, nebo stačí vhodně modifikovat odhady, které se používají u bodových procesů (podkapitola 1.1).

Uvažujme nejprve případ kvalitativních kót $\mathbb{M} = \{1, \dots, k\}$, kdy Φ_{m} je k-rozměrný bodový proces $(\Phi_1, \dots, \Phi_k)$. Pak pro křížovou G-funkci $G_{ij}(r) = P_o^{!i}(D_j(o) \le r)$ a kondenzovanou G-funkci $G_{i\cdot}(r) = P_o^{!i}(D(o) \le r)$, $r \ge 0$, můžeme například použít Kaplanův-Meierův odhad:

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

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

kde $c(x) = d(x, \partial W)$ je vzdálenost bodu x od hranice okna, $e_j(x) = d(x, \Phi_j \setminus \{x\})$ je vzdálenost bodu x k nejbližšímu bodu procesu Φ_j a $e(x) = d(x, \Phi \setminus \{x\})$ je vzdálenost bodu x k nejbližšímu bodu procesu (bez ohledu na kótu). V knihovně spatstat lze tyto odhady dostat funkcemi Gcross a Gdot.

Křížovou K-funkci K_{ij} jsme definovali vztahem

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

a kondenzovanou K-funkci $K_{i\cdot}$ vztahem

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

Odhad křížové (Kcross) a kondenzované (Kdot) K-funkce uvedeme ve tvaru s translační korekcí (jiné korekce na okrajové efekty by šly rovněž použít):

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

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

Přitom přirozený nestranný odhad intenzity podprocesu Φ_i je $\widehat{\lambda_i} = \Phi_i(W)/|W|$. Všimněte si, že takto definovaný odhad křížové K-funkce splňuje $\widehat{K_{ij}}(r) = \widehat{K_{ji}}(r)$.

Pro kótované bodové procesy s kvantitativními kótami se nejprve věnujme odhadu nenormalizované f-korelační funkce kót, která má tvar $\kappa_f(r) = \frac{\lambda_f^{(2)}(r)}{\lambda^{(2)}(r)}$. Jádrový odhad hustoty $\lambda_f^{(2)}(r)$ faktoriální momentové míry druhého řádu

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

příslušné funkci f má tvar

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

zatímco jádrový odhad součinové hustoty druhého řádu $\lambda^{(2)}(r)$ je

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

kde k_b je zvolená jádrová funkce se šířkou pásma b. V obou případech jsme použili translační korekci na okrajové efekty. Nenormalizovanou f-korelační funkci kót pak můžeme odhadnout jako

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

Dále odhadneme f-korelační funkci kót $k_f(r) = \frac{\kappa_f(r)}{c_f}$ jako

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

kde

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

je odhad $c_f = \int\int f(m_1, m_2)\, \mathbb{Q}(\mathrm{d}m_1)\, \mathbb{Q}(\mathrm{d}m_2)$. Označíme-li $\mu = \mathbb{E}M_0 = \int m\, \mathbb{Q}(\mathrm{d}m)$ střední hodnotu typické kóty, tak pro $f(m_1, m_2) = m_1 m_2$ je $c_f = \mu^2$ a k_f se označuje jako k_{mm}. Pro $f(m_1, m_2) = m_1$ je $c_f = \mu$ a k_f se označuje jako $k_{m\cdot}$. Hodnoty $k_{mm}(r)$ nebo $k_{m\cdot}(r)$ větší než 1 indikují vzájemnou stimulaci ve vzdálenosti r. Oproti tomu hodnoty menší než 1 znamenají inhibici.

Z číselných charakteristik jsme definovali nenormalizovaný korelační index nejbližších sousedů vztahem $\bar{\nu}_f = \mathbb{E}_o^! f(M(o), M(Z_1))$, kde Z_1 je bod procesu, který je nejblíže počátku, a $M(Z_1)$ je jeho kóta. Tento index můžeme přirozeným způsobem odhadnout jako

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

kde $Z_X \in \Phi$ je nejbližší soused bodu X. Odhad normalizovaného korelačního indexu nejbližších sousedů $\bar{n}_f = \frac{\bar{\nu}_f}{c_f}$ dostaneme jako

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

Pro $f(m_1, m_2) = m_1 m_2$ indikují hodnoty $\bar{n}_f > 1$ vzájemnou stimulaci mezi sousedy.

2.2 Testy nezávislosti

Statistická analýza kótovaného bodového procesu většinou začíná testováním hypotézy, zda lze kóty považovat za nezávislé. Pokud ano, tak můžeme použít metody vyvinuté pro nezávislá data. Nejprve se tedy budeme věnovat testování nezávislosti kót. Poté zmíníme možnosti testování nezávislosti kót na polohách. Použijeme simulační testy, jejichž obecný princip byl vyložen v podkapitole 1.2.

Testování nezávislosti kót

Uvažujme nejprve dvourozměrný bodový proces $\Phi_{\mathrm{m}} = (\Phi_1, \Phi_2)$. Nulová hypotéza nezávislosti kót se dá chápat dvěma způsoby:

1. nezávislé kótování – bodům bodového procesu Φ jsou nezávisle náhodně přiřazeny kóty 1 nebo 2,
2. náhodná superpozice (složení) – dva nezávislé bodové procesy Φ_1 a Φ_2 tvoří dvourozměrný bodový proces.

První situace je příklad *aposteriorního kótování* – popisujeme, jak vznikly kóty podmíněně při daných polohách bodů. Příkladem mohou být stromy v lese, které jsou buď nakaženy nějakou nemocí či zničeny větrem (kóta 1), nebo nejsou (kóta 2). V druhém případě jde o *apriorní kótování* – kótovaný bodový proces je vytvořen určitým mechanismem.

Pro testování hypotézy, že Φ_{m} je nezávisle kótovaný bodový proces se používá *metoda náhodného přerozdělení (random allocation)*. Zafixujeme polohy pozorovaných bodů a vytvoříme nové kóty náhodným zpermutováním pozorovaných (v balíčku spatstat pomocí funkce rlabel). Těchto permutací vygenerujeme celkem M a provedeme simultánní Monte Carlo test. Za hypotézy nezávislého kótování platí:

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

kde $K(r)$, $g(r)$, $G(r)$ a $J(r)$ jsou funkcionální charakteristiky nekótovaného bodového procesu Φ. Proto se jako vhodná popisná charakteristika jeví např. $S(r) = K_{1\cdot}(r) - K(r)$, pak totiž $S_0(r) = 0$.

Chceme-li testovat hypotézu náhodné superpozice procesů Φ_1 a Φ_2, můžeme použít *metodu náhodného posunutí (random shift)*. Polohy bodů s kótou 1 zafixujeme a vygenerujeme M realizací podprocesu Φ_2 tak, že všechny jeho body současně náhodně posuneme (funkce rshift) o vektor s předem zvolenou délkou $R > 0$. Z každé takto vzniklé realizace dvourozměrného bodového procesu spočítáme odhad charakteristiky $S(r)$ a aplikujeme simultánní Monte Carlo test. Jako funkci $S(r)$ můžeme použít některou z křížových funkcionálních charakteristik. Za hypotézy náhodného složení platí:

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

V případě procesu s kvantitativními kótami můžeme hypotézu nezávislého kótování opět testovat pomocí metody náhodného přerozdělení. Jako testová statistika se hodí některá z f-korelačních funkcí kót. Pro stacionární a izotropní nezávisle kótované bodové procesy je $k_f(r) = 1$.

Nezávislost kót a poloh

Na závěr této podkapitoly uvedeme dva testy nezávislosti kót a poloh v kótovaných bodových procesech s kvantitativními kótami. V případě nezávislosti lze vyšetřovat kóty a polohy zvlášť, což zjednodušuje statistickou analýzu.

První test pochází z článku [9] a je založen na faktu, že pro stacionární a izotropní kótovaný bodový proces s geostatistickým kótováním jsou funkce $E(r) = \mathbb{E}_{or} M(r)$ a $V(r) = \mathbb{E}_{or}(M(o) - E(r))^2$ konstantní. Pokud se odhady těchto funkcí z dat výrazně odlišují od konstantní funkce, svědčí to proti hypotéze nezávislosti kót a poloh. Pokud dodefinujeme $E(0) = \mathbb{E}M_0$ a $V(0) = \operatorname{var} M_0$, pak můžeme provést simultánní Monte Carlo test s volbou $S(r) = E(r) - E(0)$ nebo $S(r) = V(r) - V(0)$, přitom $S_0(r) = 0$.

I druhý test je založen na Monte Carlo testování. Byl navržen v článku [2]. Mějme realizaci $\varphi_{\mathrm{m}} = \{(x_1, m_1), \ldots, (x_n, m_n)\}$ kótovaného bodového procesu Φ_{m} pozorovaného v okně W. Předpokládejme, že data jsou uspořádána v jistém pevném pořadí. Nechť $\delta(x_i) = d(x_i, \{x_{i+1}, \ldots, x_n\})$, $i = 1, \ldots, n$, značí vzdálenost bodu x_i k nejbližšímu bodu procesu s vyšším indexem. Pro dané $r > 0$ vybereme ty body, pro které $\delta(x_i) \leq r$. Počet takto vybraných bodů označme n_r. Je doporučeno volit r malé v porovnání se vzdálenostmi nejbližších sousedů. Protože výběr bodů nezávisí na kótách, za nulové hypotézy by průměr kót n_r vybraných bodů měl být blízko průměru jakýchkoli jiných n_r kót vybraných z celkového počtu n. Oproti tomu, pokud je hodnota kóty závislá na přítomnosti bodů v blízkém okolí, tak průměry kót bodů vybraných podle zavedeného kritéria a vybraných náhodně se budou významně lišit. Vygenerujeme M různých náhodných výběrů kót o rozsahu n_r a pro každý určíme průměr. Samotný test pak probíhá stejně jako u klasického Monte Carlo testu popsaného v podkapitole 1.2 (za T bereme průměr n_r kót).

3. Geostatistika

Geostatistika je část prostorové statistiky, ve které jsou data tvořena konečným počtem měření sledované veličiny v pevně rozmístěných místech v prostoru.

Pro modelování geostatistických dat používáme náhodné pole $\{Z(x) : x \in D\}$, kde $D \subseteq \mathbb{R}^d$ má kladnou d-rozměrnou Lebesgueovu míru. Připomeňme, že vnitřně stacionární náhodné pole splňuje podmínky $\mathbb{E}(Z(x) - Z(y)) = 0$ a $\operatorname{var}(Z(x) - Z(y)) = 2\gamma(x - y)$, přičemž funkce $2\gamma(h) = \operatorname{var}(Z(x + h) - Z(x)) = \mathbb{E}(Z(x + h) - Z(x))^2$ se nazývá variogram. Naším prvním cílem bude odhadnout variogram na základě pozorování $Z(x_1), \ldots, Z(x_n)$, kde $x_1, \ldots, x_n \in D$ jsou pevně dány.

3.1 Odhad variogramu

Neparametrické odhady

Pro první představu o variogramu můžeme vykreslit čtverce rozdílů pozorovaných hodnot $(Z(x_i) - Z(x_j))^2$ oproti hodnotám $x_i - x_j$ nebo $\|x_i - x_j\|$. Takový graf se nazývá *empirický mrak variogramu (empirical variogram cloud)* a v knihovně geoR ho lze dostat pomocí funkce variog s volbou option="cloud". Tento graf není často příliš přehledný, protože možných dvojic $\{x_i, x_j\}$ různých bodů může být velký

počet, konkrétně je to $\binom{n}{2}$. Užitečnější informaci dostaneme zprůměrováním hodnot odpovídajících stejnému rozdílu $x_i - x_j$. Máme tak následující nestranný odhad variogramu:

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(x_i) - Z(x_j))^2, \tag{3}$$

kde $N(h) = \{(x_i, x_j) : x_i - x_j = h, i, j = 1, \ldots, n\}$ a $|N(h)|$ je počet různých dvojic v $N(h)$. Jedná se vlastně o odhad získaný momentovou metodou. Snadno vidíme, že tento odhad má následující vlastnosti: $\hat{\gamma}(h) \geq 0$, $\hat{\gamma}(o) = 0$ a $\hat{\gamma}(h) = \hat{\gamma}(-h)$. Zachovává tedy základní teoretické vlasnosti variogramu. Symetrie odhadu je splněna, i když $N(h) \neq N(-h)$. Pro malý rozsah dat nebo nepravidelně rozmístěné body $x_1, \ldots, x_n$, ve kterých probíhá měření, bude počet různých dvojic v $N(h)$ velmi malý a dostaneme hodně variabilní odhad hodnoty $2\gamma(h)$. Praktické doporučení je používat h, pro která je $|N(h)| \geq 30$. Pokud tuto podmínku nemůžeme zaručit, rozdělíme (podobně jako u tvorby histogramu) dvojice bodů do několika skupin s podobnými hodnotami rozdílů $x_i - x_j$ a spočítáme průměr veličin $(Z(x_i) - Z(x_j))^2$ v každé skupině. V knihovně RandomFields toho dosáhneme funkcí EmpiricalVariogram, v knihovně gstat příkazem variogram. Jiná možnost je použít vyhlazení pomocí vhodné jádrové funkce k_b s šířkou pásma b:

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

Vyhlazený i histogramový odhad spočteme v balíčku geoR pomocí variog.

U odhadů založených na $(Z(x_i) - Z(x_j))^2$ se může projevit velký vliv odlehlých pozorování, protože velkou hodnotu rozdílu ještě umocňujeme na druhou. Předpokládejme, že $\{Z(x) : x \in D\}$ je gaussovské náhodné pole. Potom $(Z(x+h) - Z(x))^2$ má rozdělení $2\gamma(h) \cdot \chi^2_1$, které je velmi zešikmené. Jako vhodná transformace, která z tohoto rozdělení vytvoří rozdělení „blízké“ normálnímu, se nabízí čtvrtá odmocnina. Místo $(Z(x_i) - Z(x_j))^2$ bychom tak pracovali s $|Z(x_i) - Z(x_j)|^{1/2}$. To nás vede k robustní verzi odhadu variogramu:

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

kde $B(h) = 0{,}457 + 0{,}494/|N(h)|$. Umocnění na čtvrtou musíme provést v rámci zachování správného měřítka. Touto transformací se poruší nestrannost odhadu, a tak je přidán člen $B(h)$, který představuje korekci vychýlení a zaručuje přibližně nestranný odhad. Robustní odhad se v balíčku geoR spočte volbou estimator.type="modulus" ve funkci variog. Kromě zmírnění vlivu odlehlých pozorování je další výhodou robustního odhadu to, že sčítanci jsou méně korelováni než v případě klasického odhadu (3).

Pokud předpokládáme, že náhodné pole je slabě stacionární, můžeme také pracovat s kovariančí funkcí $C(h) = \mathrm{cov}(Z(x), Z(x+h))$. V geostatistice se pro kovariančí funkci používá název *kovariogram (covariogram)*. Mezi semivariogramem a kovariogramem je potom vztah

$$\gamma(h) = C(o) - C(h). \tag{4}$$

Klasický výběrový odhad kovariančí funkce je

$$\widehat{C}(h) = \frac{1}{|N(h)|} \sum_{N(h)} (Z(x_i) - \bar{Z})(Z(x_j) - \bar{Z}), \tag{5}$$

kde $\bar{Z} = \frac{1}{n} \sum_{j=1}^{n} Z(x_j)$ je klasický výběrový odhad střední hodnoty μ. Nevýhodou je, že musíme odhadovat střední hodnotu, což způsobuje vychýlení odhadu (5). Z tohoto důvodu se variogram jeví jako lepší způsob charakterizace závislostí než kovariančí funkce, ta je však daleko rozšířenější a častěji používanější. Odhad kovariančí funkce je symetrický ($\widehat{C}(h) = \widehat{C}(-h)$) a pro $h = o$ máme odhad rozptylu náhodného pole:

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

Rozepíšeme-li odhad (3) tak, že v závorce každého sčítance přičteme a odečteme $\bar{Z}$, dostaneme

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

Obecně tedy $2\hat{\gamma}(h) \neq 2(\widehat{C}(o) - \widehat{C}(h))$, takže vztah (4) není zachován při přechodu k momentovým odhadům. O tom, že by nebylo vhodné variogram odhadovat výrazem $2(\widehat{C}(o) - \widehat{C}(h))$ neboli dosazením výběrových kovariancí do (4), svědčí také to, že takto můžeme dostat záporné hodnoty.

Často předpokládáme, že náhodné pole je izotropní, variogram je pak funkce vzdálenosti $\|h\|$, čehož lze využít při jeho odhadování, např. u odhadu na bázi histogramu uvažujeme skupiny dvojic bodů s blízkými vzdálenostmi mezi sebou nebo v jádrově vyhlazeném odhadu pokládáme $k_b(\|x_i - x_j\| - \|h\|)$, kde k_b je jednorozměrná jádrová funkce.

Nevýhodou neparametrických odhadů je jejich velký rozptyl a také to, že odhady variogramu a kovarianční funkce nemusí dávat platný variogram nebo kovarianční funkci. Jak víme, každý variogram musí být podmíněně negativně definitní a každá kovarianční funkce pozitivně semidefinitní funkce, ale $\hat{\gamma}$ a $\widehat{C}$ už tyto vlastnosti mít nemusí. Proto budeme uvažovat parametrické metody odhadu variogramu a kovarianční funkce.

Parametrické metody

Zvolíme parametrický model variogramu $2\gamma_\theta(h)$ případně kovariogramu $C_\theta(h)$, kde $\boldsymbol{\theta} \in \Theta$ je vektor neznámých parametrů. Například můžeme uvažovat mocninný model variogramu

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

kde $c_0 \geq 0$ je zbytkový rozptyl, $b \geq 0$ a $0 \leq \alpha < 2$.

Nejmenší čtverce

První možnost odhadu $\boldsymbol{\theta}$ z dat je založena na neparametrickém odhadu spočteném v několika bodech h_k a proložení křivky definující model variogramu body $(h_k, 2\hat{\gamma}(h_k))$, $k = 1, \ldots, K$. Kdybychom proložení křivky provedli klasickou metodou nejmenších čtverců, tak ignorujeme korelace mezi odhady $2\hat{\gamma}(h_k)$ a nestejný rozptyl těchto odhadů. Položme $2\hat{\gamma}(\boldsymbol{h}) = (2\hat{\gamma}(h_1), \ldots, 2\hat{\gamma}(h_K))$ a $2\gamma_\theta(\boldsymbol{h}) = (2\gamma_\theta(h_1), \ldots, 2\gamma_\theta(h_K))$ a uvažujme statistický model tvaru

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

kde předpokládáme, že vektor chyb $e(\boldsymbol{h}) = (e(h_1), \ldots, e(h_K))$ má nulovou střední hodnotu a varianční matici $\boldsymbol{V}(\boldsymbol{\theta})$, která může záviset na $\boldsymbol{\theta}$. Nyní můžeme provést metodu zobecněných nejmenších čtverců, která spočívá v minimalizaci výrazu

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

přes $\theta \in \Theta$. Problém je, jak získat matici $\boldsymbol{V}(\theta)$.

Uvažujme případ, kdy $\{Z(x) : x \in D\}$ je gaussovské náhodné pole. Potom

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

Abychom vyjádřili kovarianci, použijeme toho, že pro náhodný vektor $(X, Y)^{\mathrm{T}}$ s dvourozměrným normálním rozdělením takovým, že $\operatorname{var} X = \operatorname{var} Y = 1$ a $\operatorname{cov}(X, Y) = \varrho$, platí $\operatorname{cov}(X^2, Y^2) = 2\varrho^2$. Odtud

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

Rozptyl odhadu (3) je

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

Jednoduchou aproximací tohoto rozptylu je

$$\operatorname{var} 2\hat{\gamma}(h_k) \approx \frac{2(2\gamma_\theta(h_k))^2}{|N(h_k)|}, \tag{6}$$

která je přesná v případě, že $(Z(x_i) - Z(x_j))^2$ jsou nekorelované. Nahradíme-li matici $\boldsymbol{V}(\theta)$ diagonální maticí $\boldsymbol{\Delta}(\theta)$, jejíž prvky jsou dány vztahem (6), získáme vážený součet čtverců

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

Odhad θ metodou vážených nejmenších čtverců dostaneme minimalizací tohoto součtu.

Maximální věrohodnost

Druhá možnost je hledat odhad parametrů metodou maximální věrohodnosti. Pro gaussovské náhodné pole se střední hodnotou μ a kovarianční funkcí C_θ má logaritmická věrohodnostní funkce založena na datech $\boldsymbol{z} = (z(x_1), \dots, z(x_n))^{\mathrm{T}}$ po vynásobení -2 tento tvar:

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

kde $\mathbf{1} = (1, \dots, 1)^{\mathrm{T}}$ a $\boldsymbol{C}_n(\theta)_{ij} = C_\theta(x_i - x_j)$ závisí na vektoru parametrů kovarianční funkce. Pro dané θ je $L(\mu, \theta)$ maximální pro

$$\tilde{\mu} = (\mathbf{1}^{\mathrm{T}} \boldsymbol{C}_n(\theta)^{-1} \mathbf{1})^{-1} \mathbf{1}^{\mathrm{T}} \boldsymbol{C}_n(\theta)^{-1} \boldsymbol{z}. \tag{7}$$

Jedná se o odhadu metodou zobecněných nejmenších čtverců. Dosazením $\tilde{\mu}$ do $L(\mu, \theta)$ dostaneme funkci θ (tzv. *profilová věrohodnost (profile likelihood)*), kterou je třeba maximalizovat (většinou numericky). Odhad μ pak dostaneme dosazením odhadu θ do (7). Populární variantou maximální věrohodnosti je REML – odhad metodou *reziduální maximální věrohodnosti (residual/restricted maximal likelihood)*. U této metody není věrohodnost aplikována přímo na data, ale na rezidua. Spočívá v tom, že najdeme vhodnou matici $\boldsymbol{A}$, kterou lineárně transformujeme data $\boldsymbol{Z} = (Z(x_1), \dots, Z(x_n))^{\mathrm{T}}$ na $\boldsymbol{Z}^* = \boldsymbol{A}\boldsymbol{Z}$ tak, že rozdělení $\boldsymbol{Z}^*$ nezávisí na μ. Parametr θ pak odhadneme metodou maximální věrohodnosti aplikovanou na transformovaná data $\boldsymbol{Z}^*$. Volba matice $\boldsymbol{A}$ není jednoznačná. Například pro matici $\boldsymbol{A}$ typu $(n-1) \times n$ s prvky $a_{ij} = \mathbf{1}_{[i=j]} - 1/n$ dostaneme $\boldsymbol{A}\boldsymbol{Z} = (Z(x_1) - \bar{Z}, \dots, Z(x_{n-1}) - \bar{Z})^{\mathrm{T}}$ vektor $n-1$ rozdílů od výběrového průměru $\bar{Z}$, čímž se zbavíme závislosti na μ. Odhad θ získáme minimalizací funkce

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

Vložením tohoto odhadu do (7) dostaneme odhad μ. K praktickému určení odhadů parametrů se hodí funkce likfit a variofit v knihovně geoR nebo funkce fitvario v balíčku RandomFields.

Složená věrohodnost

Metodu složené věrohodnosti jsme již zmínili při odhadu parametrů u bodových procesů. Podobně se dá využít pro odhad parametrů variogramu. Předpokládejme, že rozdíly $Z(x_i) - Z(x_j)$ mají normální rozdělení. Součtem příspěvků logaritmických věrohodností přes různé dvojice bodů dostaneme logaritmus složené věrohodnosti:

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

Hledáme θ, které maximalizuje $\mathrm{CL}(\theta)$, a tak zderivujeme podle jednotlivých složek θ_k a položíme rovno nule:

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

Validace modelu

Mějme zvolený parametrický modelu variogramu a odhadnuty jeho parametry. Zajímá nás, jestli takto získaný model $2\gamma_{\hat{\theta}}$ dobře popisuje data. V dalším podkapitole uvidíme, jak lze získat predikci $\hat{Z}(x_0)$ hodnoty $Z(x_0)$ spolu s chybou predikce $\sigma^2(x_0)$. Ta závisí na nafitovaném variogramu, datech a polohách $x_0, x_1, \dots, x_n$. Pokud jsme schopni získat $Z(x_0)$ – např. dodatečným měřením nebo jsme dopředu nechali nějaká data pro validaci modelu – můžeme porovnat rozdíl mezi $Z(x_0)$ a $\hat{Z}(x_0)$. Pokud je variogram zvolen správně, měly by si být hodnoty $Z(x_0)$ a $\hat{Z}(x_0)$ blízké.

V případě, že všechna data byla použita na fitování variogramu a nelze provést dodatečná měření, nabízí se provedení *křížové validace (cross-validation)*. Vypustíme polohu x_j a spočteme predikci $\hat{Z}_{-j}(x_j)$ z $n-1$ zbylých pozorování a zvoleného variogramu $2\gamma_{\hat{\theta}}(h)$. Příslušnou chybu predikce označíme $\sigma^2_{-j}(x_j)$. Toto provedeme pro každé $j = 1, \dots, n$ a spočteme standardizovaná rezidua

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

Jejich výběrový průměr by měl být kolem 0 a jejich výběrový druhý moment kolem 1. Z histogramu standardizovaných reziduí pak můžeme detekovat případné extrémní hodnoty reziduí.

3.2 Krigování

Naším cílem je nalezení predikce $\hat{Z}(x_0)$ hodnoty $Z(x_0)$ náhodného pole v místě $x_0 \in D$ na základě vektoru $\boldsymbol{Z} = (Z(x_1), \dots, Z(x_n))^{\mathrm{T}}$. Pro metody prostorové predikce založené na minimalizaci střední kvadratické chyby se používá název *krigování (kriging)*. Ten je odvozen od D. G. Krigeho, jehož práce [4] zabývající se odhadem zásob rudy je v geostatistice považována za průkopnickou.

Jednoduché krigování

Je dobře známo, že za předpokladu konečných druhých momentů je střední kvadratická chyba $\mathbb{E}[Z(x_0) - \hat{Z}(x_0)]^2$ minimalizována podmíněnou střední hodnotou $\mathbb{E}[Z(x_0) \mid \boldsymbol{Z}]$ a chyba predikce je $\mathbb{E}[Z(x_0) - \hat{Z}(x_0)]^2 = \mathbb{E}\,\mathrm{var}[Z(x_0) \mid \boldsymbol{Z}]$, viz např. [5], věta 7.15. V praxi však bývá obtížné podmíněnou střední hodnotu určit. Pro jednoduchost proto uvažujme lineární predikci $\hat{Z}(x_0) = \alpha + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Z}$. Chceme odhadnout $\alpha \in \mathbb{R}$ a $\boldsymbol{\beta} \in \mathbb{R}^n$ tak, aby střední kvadratická chyba byla minimální. Z teorie lineárních modelů víme, že řešením je

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

kde $\boldsymbol{\mu}_n = \mathbb{E}\boldsymbol{Z} = (\mu(x_1), \dots, \mu(x_n))^{\mathrm{T}}$ je vektor středních hodnot $\boldsymbol{Z}$, $\mu(x_0) = \mathbb{E}Z(x_0)$,

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

je varianční matice vektoru $\boldsymbol{Z}$ a $\boldsymbol{c}_n = (C(x_0, x_1), \dots, C(x_0, x_n))^{\mathrm{T}}$. Tedy

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

a chyba predikce je

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

Technika obdržení této prostorové predikce se nazývá *jednoduché krigování (simple kriging)*. I když jsme to nepožadovali, tak predikce $\hat{Z}(x_0)$ je nestranná, tj. $\mathbb{E}\hat{Z}(x_0) = \mathbb{E}Z(x_0)$. Všimněme si, že chyba predikce nezávisí na datech. Pokud x_0 je jedna z poloh $x_1, \dots, x_n$, tak $\hat{Z}(x_0) = Z(x_0)$, prostorová predikce tedy interpoluje data. Platí totiž

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

což lze vektorově zapsat jako

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

Pro gaussovské náhodné pole je jednoduché krigování optimální.

Lemma 2. *Nechť $\{Z(x) : x \in D\}$ je gaussovské náhodné pole. Nejlepší lineární predikce $\hat{Z}(x_0) = \mu(x_0) + \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_n)$ je nejlepší predikce $Z(x_0)$ a platí*

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

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

Důkaz: Sdružené rozdělení $(Z(x_0), \boldsymbol{Z})^{\mathrm{T}}$ je $(n+1)$-rozměrné normální. Podmíněná rozdělení v mnohorozměrném normálním rozdělení jsou opět normální. V našem případě je podmíněné rozdělení $Z(x_0) \mid \boldsymbol{Z}$ normální se střední hodnotou $\mu(x_0) + \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}(\boldsymbol{Z} - \boldsymbol{\mu}_n)$ a rozptylem $\mathrm{var}\, Z(x_0) - \boldsymbol{c}_n^{\mathrm{T}}\boldsymbol{C}_n^{-1}\boldsymbol{c}_n$. Nejlepší (ne nutně lineární) predikce $Z(x_0)$ je podmíněná střední hodnota $\mathbb{E}[Z(x_0) \mid \boldsymbol{Z}]$.

□

I když je lineární predikce optimální v případě gaussovského modelu, může mít špatné vlastnosti, pokud jsou porušeny předpoklady normálního rozdělení. K vyrovnání se s tímto problémem se ve statistice často používá transformace dat vedoucí na normální rozdělení. Příkladem je tzv. *Boxova-Coxova transformace*

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

Existují různé metody, jak odhadnout parametr λ. Rovněž je možné volit bayesovský přístup a považovat λ za náhodné.

Vyjádřili jsme nejlepší lineární predikci. Problém je, že závisí na hodnotách $\boldsymbol{\mu}_n$, $\mu(x_0)$, $\boldsymbol{c}_n$ a $\boldsymbol{C}_n$, které jsou v praxi neznámé. Obecně se jedná o $(n+1) + n + \binom{n+1}{2}$ neznámých parametrů, které by bylo třeba odhadnout pouze z n dat. Proto dodáme některé předpoklady.

Obyčejné krigování

Předpokládejme nyní, že náhodné pole má konstantní konečnou střední hodnotu μ. Budeme hledat lineární predikci tvaru

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

kde složky $\lambda_1, \ldots, \lambda_n$ vektoru $\boldsymbol{\lambda}$ jsou neznámé reálné koeficienty. Podmínka, že jejich součet je roven jedné, zaručuje, že nestrannost predikce: $\mathbb{E}\hat{Z}(x_0) = \boldsymbol{\lambda}^{\mathrm{T}} \mu \mathbf{1} = \mu = \mathbb{E}Z(x_0)$. Metoda pro hledání prostorové predikce za těchto předpokladu se označuje jako *obyčejné krigování (ordinary kriging)*.

Pro vnitřně stacionární náhodné pole můžeme rozptyl lineární kombinace s nulovým součtem koeficientů rozepsat pomocí semivariogramu γ:

$$\begin{aligned} \mathbb{E}(Z(x_0) - \hat{Z}(x_0))^2 &= \mathbb{E}(Z(x_0) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{Z})^2 = \operatorname{var}(Z(x_0) - \boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{Z}) \\ &= -\sum_{i,j} \lambda_i \lambda_j \gamma(x_i - x_j) + 2 \sum_{i=1}^{n} \lambda_i \gamma(x_i - x_0). \end{aligned} \tag{8}$$

K určení predikce $\hat{Z}(x_0)$ tedy nepotřebujeme znát střední hodnotu μ. Pro minimalizaci (8) za podmínky $\boldsymbol{\lambda}^{\mathrm{T}} \mathbf{1} = 1$ se dá užít Lagrangeova věta o multiplikátorech. Pro jednodušší zápis vynásobíme multiplikátor m dvěma a minimalizujeme

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

kde $\boldsymbol{\Gamma}_n = (\gamma(x_i - x_j))_{i,j=1,\ldots,n}$ a $\boldsymbol{\gamma}_n = (\gamma(x_1 - x_0), \ldots, \gamma(x_n - x_0))^{\mathrm{T}}$. Zderivujeme Q podle $\boldsymbol{\lambda}$ a m a položíme rovno nule, dostaneme soustavu rovnic

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

Řešením je

$$\begin{aligned} \boldsymbol{\lambda}^{\mathrm{T}} &= \left(\boldsymbol{\gamma}_n + \mathbf{1} \frac{1 - \mathbf{1}^{\mathrm{T}} \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n}{\mathbf{1}^{\mathrm{T}} \boldsymbol{\Gamma}_n^{-1} \mathbf{1}} \right)^{\mathrm{T}} \boldsymbol{\Gamma}_n^{-1}, \\ m &= \frac{1 - \mathbf{1}^{\mathrm{T}} \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n}{\mathbf{1}^{\mathrm{T}} \boldsymbol{\Gamma}_n^{-1} \mathbf{1}}. \end{aligned} \tag{9}$$

Predikce má tak tvar

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

Koeficienty λ_i jsou složky vektoru (9) a označují se jako *predikční váhy (prediction weights)*. Typicky bývají predikční váhy odpovídající bodům blízkým x_0 velké, ale jejich přesná hodnota závisí na polohách x_i a kovarianční struktuře dat. Může se stát, že λ_i bude záporné nebo větší než 1. Chyba predikce je

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

Podobně lze pro slabě stacionární náhodná pole přepsat $\hat{Z}(x_0)$ pomocí kovariogramu:

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

Chyba predikce je

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

Vidíme, že tato chyba je větší než v případě jednoduchého krigování, protože poslední člen je kladný. Větší chyba je způsobena tím, že neznáme střední hodnotu μ.

Univerzální krigování

V následujícím odstavci se budeme zabývat situací, kdy střední hodnota $\mu(x) = \mathbb{E}Z(x)$ není konstantní. Nejjednodušším způsobem je pak použít lineární model

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

kde $f_0(x), \ldots, f_p(x)$ jsou známé pozorované hodnoty funkcí f_j v místě $x \in D$ a $\beta_0, \ldots, \beta_p$ jsou neznámé reálné parametry. Za f_0 se typicky volí konstantní funkce rovna jedné, β_0 je pak absolutní člen. Častou volbou $f_i(x)$ je polynom prostorových souřadnic polohy x. Takto je možné modelovat například lineární trend. Jiná možnost je, že $f_i(x)$ představuje pozorovanou vysvětlující proměnnou. Označme $\boldsymbol{f} = (f_0(x_0), \ldots, f_p(x_0))^{\mathrm{T}}$ a $\boldsymbol{F}$ matici typu $n \times (p+1)$, jejíž prvky jsou $f_j(x_i)$, $i = 1, \ldots, n$, $j = 0, \ldots, p$. Pokud uvažujeme predikci tvaru

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

mluvíme o *univerzálním krigování (universal kriging)*. Volba $\boldsymbol{\lambda}^{\mathrm{T}} \boldsymbol{F} = \boldsymbol{f}^{\mathrm{T}}$ zaručí, že tato predikce je nestranná, neboť

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

Optimální predikci (minimalizující střední čtvercovou chybu) lze opět hledat pomocí Lagrangeových multiplikátorů. Podobně jako u obyčejného krigování se dá ukázat, že optimální predikční váhy mají tvar

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

Chyba predikce je

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

Pomocí kovariancí lze predikční váhy přepsat jako

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

a chybu predikce jako

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

Pomocí zobecněných nejmenších čtverců lze pak odhadnout i parametr $\boldsymbol{\beta}$:

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

Predikce se tak dá zapsat ve tvaru

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

V případě, že veličina $Z(x_0)$ je nekorelovaná s daty, tak predikce $\hat{Z}(x_0)$ splývá s nejlepším lineárním nestranným odhadem střední hodnoty, který je roven $\boldsymbol{f}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$. Obecně jsou však predikce $Z(x_0)$ a odhad $\mathbb{E}Z(x_0)$ odlišné.

Další možnosti

Předpokládejme, že místo predikce $Z(x_0)$ z dat $Z(x_1), \ldots, Z(x_n)$ nás zajímá předpověď průměrné hodnoty v nějaké oblasti (bloku) B:

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

Analogie obyčejného krigování vede k tzv. *blokovému krigování (block kriging)*, hledáme predikci tvaru

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

kde $\sum_{i=1}^{n} \lambda_i = 1$. Optimální váhy mají tvar

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

kde $\boldsymbol{c}_B = (\mathrm{cov}(Z(B), Z(x_1)), \ldots, \mathrm{cov}(Z(B), Z(x_n)))^{\mathrm{T}}$. Vyjádření pomocí variogramu by vypadalo analogicky.

Podobně nás může zajímat predikce $g(Z(x_0))$, kde g je daná funkce. Nejlepší predikce je $\mathbb{E}[g(Z(x_0)) \mid \boldsymbol{Z}]$.

Dalším častým příkladem je úloha odhadu pravděpodobnosti $\mathbb{P}(Z(x_0) \leq y \mid \boldsymbol{Z})$, kde y je daná reálná hodnota. Mluví se pak o *indikátorovém krigování (indicator kriging)*.

V knihovně geoR je pro krigování určená funkce krige.conv, zatímco v balíčku gstat to je krige.

3.3 Vliv odhadů kovariančních parametrů

Vzorečky pro prostorovou predikci odvozené v minulé podkapitole závisí na hodnotách kovariogramu nebo variogramu, které jsou v praxi typicky neznámé a je třeba je nějakým způsobem odhadnout. Již byly zmíněny základní postupy, kterými lze odhadovat parametry variogramu nebo kovariogramu. Dosazením odhadů parametrů za neznámé parametry do parametrického tvaru příslušné funkce dostaneme tzv. *plug-in* odhad. Postup tedy probíhá v následujících krocích:

1. vybereme parametrický model pro variogram $\gamma_\theta(h)$ nebo kovariogram $C_\theta(h)$,
2. odhadneme parametry θ,
3. upravíme statistickou inferenci vzhledem k tomu, že místo konstanty θ pracujeme s náhodnou veličinou $\hat{\theta}$.

U obyčejného krigování plug-in predikce

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

už není nejlepší nestranná lineární predikce (BLUP = best linear unbiased prediction) prvku $Z(x_0)$, je to pouze odhad této predikce (tzv. EBLUP = estimated best linear unbiased prediction). Zatímco chyba predikce $\hat{Z}(x_0)$ je

$$C(o) - \boldsymbol{c}_n(\theta)^{\mathrm{T}} \boldsymbol{C}_n(\theta)^{-1} \boldsymbol{c}_n(\theta) + \frac{(1 - \mathbf{1}^{\mathrm{T}} \boldsymbol{C}_n(\theta)^{-1} \boldsymbol{c}_n(\theta))^2}{\mathbf{1}^{\mathrm{T}} \boldsymbol{C}_n(\theta)^{-1} \mathbf{1}}, \tag{10}$$

tak chybu predikce $\hat{\hat{Z}}(x_0)$ neznáme. Dosadíme-li $\hat{\theta}$ do (10), dostaneme odhad chyby predikce $\hat{Z}(x_0)$, tedy jiné predikce než ve skutečnosti používáme. Takto získaný odhad chyby predikce má tendenci podhodnocovat skutečnou chybu predikce $\hat{\hat{Z}}(x_0)$, protože nepostihujeme fakt, že náhodné $\hat{\theta}$ vnáší další variabilitu do odhadu predikce.

Vraťme se k situaci univerzálního krigování, kdy uvažujme model $Z(x) = \boldsymbol{F}(x)^{\mathrm{T}} \boldsymbol{\beta} + e(x)$, kde $\boldsymbol{F}(x) = (f_0(x), \ldots, f_p(x))^{\mathrm{T}}$ a $\{e(x) : x \in D\}$ je vnitřně stacionární náhodné pole s variogramem parametrizovaným pomocí θ. Pro odhad parametru θ není rozumné použít empirický odhad z dat $\boldsymbol{Z} = (Z(x_1), \ldots, Z(x_n))^{\mathrm{T}}$, protože ten je vychýlený. Vychýlení odhadu (3) je způsobeno tím, že $Z(x)$ nemá konstantní střední hodnotu, a proto $\mathbb{E}(Z(x_i) - Z(x_j))^2 = \mathrm{var}(Z(x_i) - Z(x_j)) + (\mu(x_i) - \mu(x_j))^2$. Potřebovali bychom odhad variogramu $\{e(x) : x \in D\}$, ovšem náhodné pole chyb $\{e(x) : x \in D\}$ nepozorujeme. Pokud by však bylo β známé, pak $e(x) = Z(x) - \boldsymbol{F}(x)^{\mathrm{T}} \boldsymbol{\beta}$ a z hodnot $\boldsymbol{e} = (e(x_1), \ldots, e(x_n))^{\mathrm{T}}$ bychom mohli odhadnout θ. Jenže parametr $\boldsymbol{\beta}$ je neznámý. Pokud je pole $\{e(x) : x \in D\}$ slabě stacionární s kovarianční funkcí $C_\theta(h)$, tak dostaneme odhad $\boldsymbol{\beta}$ pomocí metody zobecněných nejmenších čtverců

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^{\mathrm{T}} \boldsymbol{C}_n(\theta)^{-1} \boldsymbol{F})^{-1} \boldsymbol{F}^{\mathrm{T}} \boldsymbol{C}_n(\theta)^{-1} \boldsymbol{Z}. \tag{11}$$

Tento odhad však vyžaduje znalost parametru θ. Znamená to, že nemůžeme rozumně odhadnout θ bez znalostí $\boldsymbol{\beta}$ a na druhou stranu k odhadu $\boldsymbol{\beta}$ potřebujeme odhad θ. O této kruhové situaci se někdy mluví jako hře kočky s myší univerzálního krigování.

Možné řešení nabízí metoda IRWGLS (zkratka z anglického výrazu *iteratively re-weighted generalized least squares*):

1. získáme počáteční odhad parametru $\boldsymbol{\beta}$ nezávisle na θ, např. metodou obyčejných nejmenších čtverců: $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{Z}$,
2. spočteme rezidua $\boldsymbol{r} = \boldsymbol{Z} - \boldsymbol{F}\hat{\boldsymbol{\beta}}$,
3. odhadneme parametrický model variogramu nebo kovariogramu reziduí a dostaneme $\hat{\theta}$,
4. spočteme nový odhad $\hat{\boldsymbol{\beta}}$ jako $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n(\hat{\theta})^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{C}_n(\hat{\theta})^{-1}\boldsymbol{Z}$,
5. opakujeme kroky 2.-4., až dokud relativní změny v odhadech $\boldsymbol{\beta}$ a θ jsou malé.

Odhad variogramu je vychýlený, ale tentokrát není vychýlení způsobeno nekonstantní střední hodnotou, ale tím, že odhadujeme variogram reziduí a ne variogram pole $\{e(x) : x \in D\}$.

Studium chování této procedury je složité. Není zaručeno, že odhady budou konvergovat k teoretickým hodnotám.

Jiná možnost je použít metodu maximální věrohodnosti k odhadu $\boldsymbol{\beta}$ a θ současně. Například pokud předpokládáme, že $\{Z(x) : x \in D\}$ je gaussovské náhodné pole, pak logaritmus věrohodnostní funkce má tvar

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

Při daném θ je tato funkce maximalizována pro $\boldsymbol{\beta}$ dané vztahem (11) s vektorem $\boldsymbol{Z}$ nahrazeným pozorovanými daty $\boldsymbol{z}$. Dosazením do $\log L(\boldsymbol{\beta}, \theta)$ dostaneme funkci θ (profilovou věrohodnost), kterou je nutné maximalizovat numericky.

3.4 Bayesovský přístup

Na základě pozorovaných dat $\boldsymbol{z}$ je v klasickém přístupu nejlepší predikce $\mathbb{E}[Z(x_0) \mid \boldsymbol{Z} = \boldsymbol{z}]$ a její chyba je rovna $\mathbb{E}\,\mathrm{var}[Z(x_0) \mid \boldsymbol{Z} = \boldsymbol{z}]$. Často nás však spíše než střední hodnota nebo rozptyl zajímá celé podmíněné rozdělení $Z(x_0)$ za podmínky $\boldsymbol{Z} = \boldsymbol{z}$, tzv. *prediktivní rozdělení (predictive distribution)*. V bayesovském přístupu je prediktivní rozdělení rovno aposteriornímu rozdělení $Z(x_0)$.

Připomeňme, že v bayesovské statistice se parametry modelu považují za náhodné. Znamená to, že není rozdíl mezi predikcí a odhadem parametrů. Bayesovský přístup je založen na kombinaci historické informace o neznámých parametrech $\boldsymbol{\theta}$ a pozorovaných dat $\boldsymbol{z}$. Informace o parametrech je určena *apriorním rozdělením* s hustotou $p(\boldsymbol{\theta})$ vzhledem k σ-konečné míře ν na parametrickém prostoru Θ. Pokud má $\boldsymbol{Z}$ při daném $\boldsymbol{\theta}$ hustotu $f(\boldsymbol{z} \mid \boldsymbol{\theta})$, tak hustota *aposteriorního rozdělení* $\boldsymbol{\theta}$ za podmínky $\boldsymbol{Z} = \boldsymbol{z}$ je dána Bayesovou větou

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

pokud je jmenovatel nenulový. Tento vztah se většinou zkráceně zapisuje jako

$$p(\boldsymbol{\theta} \mid \boldsymbol{z}) \propto f(\boldsymbol{z} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{12}$$

symbol $\propto$ značí rovnost až na multiplikativní konstantu.

Prostorová predikce pomocí bayesovského přístupu se označuje jako *bayesovské krigování*. Pro predikci $Z(x_0)$ dostaneme *prediktivní hustotu* vyintegrováním přes $\boldsymbol{\theta}$:

$$f(z_0 \mid \boldsymbol{z}) = \int_\Theta f(z_0, \boldsymbol{\theta} \mid \boldsymbol{z})\,\nu(\mathrm{d}\boldsymbol{\theta}) = \int_\Theta f(z_0 \mid \boldsymbol{z}, \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \boldsymbol{z})\,\nu(\mathrm{d}\boldsymbol{\theta}). \tag{13}$$

Pokud by byl vektor parametrů $\boldsymbol{\theta}$ známý, dostaneme stejný výsledek jako v klasickém případě. Výhoda bayesovského přístupu je v tom, že zahrnuje do úvahy nejistotu o parametrech modelu. Tvar (13) prediktivní hustoty je převážně velmi komplikovaný. Proto se používají MCMC metody, které umožňují generovat posloupnost $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(T)}$ z aposteriorního rozdělení s hustotou $p(\boldsymbol{\theta} \mid \boldsymbol{z})$. Potom průměr

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

dává aproximaci prediktivní hustoty $f(z_0 \mid \boldsymbol{z})$. V praxi se obvykle výpočet této aproximace obchází tak, že pro každé $\boldsymbol{\theta}^{(i)}$ vygenerujeme $z_0^{(i)}$ z rozdělení s hustotou $f(z_0 \mid \boldsymbol{z}, \boldsymbol{\theta}^{(i)})$. Pak $z_0^{(1)}, \ldots, z_0^{(T)}$ je výběr z prediktivního rozdělení a vykreslení příslušného histogramu nebo jádrového odhadu hustoty dává přibližný tvar prediktivní hustoty. Jiný možný přístup pro určení prediktivní hustoty se nabízí

v případech, kdy jsme schopni spočítat aposteriorní hustoty $p(\boldsymbol{\theta} \mid \boldsymbol{z})$ a $p(\boldsymbol{\theta} \mid \boldsymbol{z}, z_0)$, potom lze použít vztah

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

Příklad: Uvažujme lineární model

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

kde $\boldsymbol{F}(x) = (f_0(x), \ldots, f_p(x))^{\mathrm{T}}$ je vektor vysvětlujících proměnných, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ je vektor regresních parametrů s apriorním rozdělením $N_{p+1}(\boldsymbol{m}, \boldsymbol{Q})$ a $\{e(x) : x \in D\}$ je slabě stacionární centrované gaussovské náhodné pole s kovarianční funkcí $C(h)$. Předpokládáme, že známe vektor $\boldsymbol{m}$, matici $\boldsymbol{Q}$ i funkci C. Cílem je prostorová predikce $Z(x_0)$ na základě dat $\boldsymbol{Z} = (Z(x_1), \ldots, Z(x_n))^{\mathrm{T}}$. Označme $\boldsymbol{C}_n$ matici s prvky $C(x_i - x_j)$, $i, j = 1, \ldots, n$, a $\boldsymbol{F}$ matici typu $n \times (p+1)$, jejíž prvky jsou $f_j(x_i)$, $i = 1, \ldots, n$, $j = 0, \ldots, p$. Dále předpokládejme, že $\boldsymbol{Q}$ i $\boldsymbol{F}^{\mathrm{T}} \boldsymbol{C}_n^{-1} \boldsymbol{F}$ mají plnou hodnost. Díky konjugovanosti normálního rozdělení je aposteriorní rozdělení $\boldsymbol{\beta} \mid \boldsymbol{Z}$ mnohorozměrné normální $N_{p+1}(\boldsymbol{m}^*, \boldsymbol{Q}^*)$, kde

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

Sdružené rozdělení $(\boldsymbol{Z}, Z(x_0))^{\mathrm{T}}$ je mnohorozměrné normální $N_{n+1}(\boldsymbol{F}_{n0} \boldsymbol{\beta}, \boldsymbol{C}_{n0})$, kde

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

a

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

$\boldsymbol{c}_n = (C(x_0 - x_1), \ldots, C(x_0 - x_n))^{\mathrm{T}}$. Prediktivní hustotu dostaneme z vyjádření

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

přitom $p(\boldsymbol{\beta} \mid \boldsymbol{z})$ je hustota $N_{p+1}(\boldsymbol{m}^*, \boldsymbol{Q}^*)$ a $f(z_0 \mid \boldsymbol{z}, \boldsymbol{\beta})$ je hustota normálního rozdělení se střední hodnotou $\boldsymbol{F}(x_0)^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{c}_n^{\mathrm{T}} \boldsymbol{C}_n^{-1} (\boldsymbol{z} - \boldsymbol{F}\boldsymbol{\beta})$ a rozptylem $C(o) - \boldsymbol{c}_n^{\mathrm{T}} \boldsymbol{C}_n^{-1} \boldsymbol{c}_n$, jak víme z lemmatu 2. Po přímočarém (i když poněkud zdlouhavém) výpočtu se zjistí, že prediktivní rozdělení je normální se střední hodnotou

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

a rozptylem

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

V praxi většinou neznáme funkci C. Můžeme však použít některý z parametrických modelů (např. Whittleův-Matérnův), specifikovat vhodné apriorní rozdělení pro parametry kovarianční funkce a odvodit aposteriorní rozdělení.

Geostatistické modely, které jsme uvažovali, se dají chápat jako dvoustupňové *hierarchické modely*. V prvním stupni hierarchie popisujeme, jak data závisí na náhodných efektech. Konkrétně je specifikováno náhodné pole $Z = \{Z(x) : x \in D\}$, které generuje data, podmíněně při $e = \{e(x) : x \in D\}$. Ve druhém stupni modelujeme rozdělení náhodných efektů, tedy nepozorovaného náhodného pole e.

V bayesovkém přístupu máme tři základní náhodné objekty, kromě Z a e je to ještě vektor neznámých parametrů $\boldsymbol{\theta}$. Dostáváme tak třístupňový hierarchický model:

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

Konkrétním příkladem je model popsaný u univerzálního krigování a použitý také v předchozím příkladě:

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

Předpokládejme, že reziduální náhodné pole $\{e(x) : x \in D\}$ je centrované stacionární gaussovské náhodné pole a lze rozepsat jako součet prostorové složky a bílého šumu:

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

kde $W = \{W(x) : x \in D\}$ je centrované stacionární gaussovské náhodné pole s kovarianční funkcí $C_W(h; \sigma^2, \phi) = \sigma^2\rho(h; \phi)$ a $\{\epsilon(x) : x \in D\}$ jsou nekorelované normálně rozdělené náhodné veličiny s nulovou střední hodnotou a rozptylem τ^2. Znamená to, že semivariogram náhodné pole e je tvaru

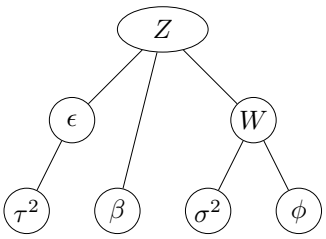
$$\gamma_e(h; \sigma^2, \tau^2, \phi) = \tau^2\mathbf{1}_{[h\neq o]} + \sigma^2\left(1 - \rho(h; \phi)\right),$$

který je parametrizován pomocí zbytkového rozptylu τ^2, částečného prahu σ^2 a korelačního parametru ϕ, který vystupuje v korelační funkci $\rho(h; \phi)$ náhodného pole W. Vektor neznámých parametrů modelu je tak $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \sigma^2, \tau^2, \phi)^{\mathrm{T}}$. Pak Z za podmínky $\boldsymbol{\theta}$ a W je gaussovské náhodné pole se střední hodnotou $\boldsymbol{F}(x)^{\mathrm{T}}\boldsymbol{\beta} + W(x)$ a kovarianční funkcí $C_{Z|W}(h; \tau^2) = \tau^2\mathbf{1}_{[h=0]}$. Všimněme si, že $Z \mid \boldsymbol{\theta}, W$ vůbec nezávisí na σ^2 a ϕ. Ve druhém stupni hierarchie specifikujeme W, které podmíněně při $\boldsymbol{\theta}$ je centrované stacionární gaussovské náhodné pole s kovarianční funkcí $C_W(h; \sigma^2, \phi)$, tedy nezávisí na $\boldsymbol{\beta}$ a τ^2. Třetí stupeň vyžaduje stanovení vhodného apriorního rozdělení pro vektor $\boldsymbol{\theta}$. Grafické znázornění hierarchického modelu je na obrázku 2. Obvykle se volí jednotlivé složky $\boldsymbol{\theta}$ apriorně nezávislé neboli apriorní hustota je tvaru

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

Vhodní kandidáti na volbu marginálních apriorních rozdělení jsou mnohorozměrné normální rozdělení pro $\boldsymbol{\beta}$, inverzní Γ-rozdělení pro σ^2 a τ^2 (neboli $1/\sigma^2$ a $1/\tau^2$ mají Γ-rozdělení). Volba pro ϕ samozřejmě závisí na tvaru variogramu, např. pro exponenciální model $\rho(h; \phi) = \exp\{-\phi\|h\|)\}$ se často bere Γ-rozdělení jako apriorní pro ϕ. Popsaný model můžeme samozřejmě formulovat také jako dvoustupňový, když využijeme toho, že $Z \mid \boldsymbol{\theta}$ je gaussovské náhodné pole se střední hodnotou $\boldsymbol{F}(x)^{\mathrm{T}}\boldsymbol{\beta}$ a kovarianční funkcí

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



Obrázek 2. Znázornění třístupňového hierarchického modelu.

Předpokládejme, že máme vektor pozorovaných dat $\boldsymbol{z} = (z(x_1), \ldots, z(x_n))^{\mathrm{T}}$. Bayesovský odhad parametrů se pak dostane z aposteriorní hustoty $p(\boldsymbol{\theta} \mid \boldsymbol{z})$ dané vztahem (12), kde v tomto případě $f(\boldsymbol{z} \mid \boldsymbol{\theta})$ je hustota n-rozměrného normálního rozdělení se střední hodnotou $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{\beta}$ a varianční maticí $\tau^2\boldsymbol{I} + \sigma^2\boldsymbol{H}(\phi)$, přičemž $\boldsymbol{F} = (f_j(x_i))_{i,j}$ je matice typu $n \times (p + 1)$, $\boldsymbol{I}$ je jednotková matice typu $n \times n$ a $\boldsymbol{H}(\phi)$ je matice typu $n \times n$, jejíž prvky jsou $\rho(x_i - x_j; \phi)$, $i, j = 1, \ldots, n$. Mohli bychom rovněž využít třístupňového modelu a vyjádřit aposteriorní hustotu jako

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

kde $f(\boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{w})$ je hustota n-rozměrného normálního rozdělení se střední hodnotou $\boldsymbol{F}^T\boldsymbol{\beta} + \boldsymbol{w}$ a varianční maticí $\tau^2\boldsymbol{I}$ a $p(\boldsymbol{w} \mid \boldsymbol{\theta})$ je hustota n-rozměrného normálního rozdělení s nulovou střední hodnotou a varianční maticí $\sigma^2\boldsymbol{H}(\phi)$. Tímto se nám ovšem zvýší počet parametrů o n složek vektoru $\boldsymbol{w} = (w(x_1), \ldots, w(x_n))^{\mathrm{T}}$. V praxi se používají MCMC metody (konkrétně Gibbsův výběrový plán) a tvar (12) je upřednostňován, protože varianční matice $\tau^2\boldsymbol{I} + \sigma^2\boldsymbol{H}(\phi)$ se chová lépe než varianční matice $\sigma^2\boldsymbol{H}(\phi)$. To lze ilustrovat na situaci, kdy body x_i a x_j jsou od sebe velmi málo vzdáleny, pak matice $\sigma^2\boldsymbol{H}(\phi)$ bude blízká singulární matici, zatímco $\tau^2\boldsymbol{I} + \sigma^2\boldsymbol{H}(\phi)$ ne.

Odhad parametrů $\boldsymbol{w}$ odpovídá rekonstrukci prostorového povrchu W v bodech měření $x_1, \ldots, x_n$. Podobně nás může zajímat predikce $W(x_0)$ pro různé volby x_0. Vzhledem ke vztahu

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

můžeme aposteriorní rozdělení $\boldsymbol{W} = (W(x_1), \ldots, W(x_n))^{\mathrm{T}}$ získat z aposteriorního rozdělení (σ^2, ϕ). Připomeňme, že v našem případě je $p(\boldsymbol{w} \mid \sigma^2, \phi)$ hustota n-rozměrného centrovaného normálního rozdělení s varianční maticí $\sigma^2 \boldsymbol{H}(\phi)$. Když $((\sigma^2)^{(t)}, \phi^{(t)})$ je výstup MCMC algoritmu, který generuje z rozdělení s aposteriorní hustotou $p(\sigma^2, \phi \mid \boldsymbol{z})$, pak stačí vygenerovat vektor $\boldsymbol{w}^{(t)}$ z rozdělení s hustotou $p(\boldsymbol{w} \mid (\sigma^2)^{(t)}, \phi^{(t)})$, čímž dostaneme výstup z rozdělení s hustotou $p(\boldsymbol{w} \mid \boldsymbol{z})$.

4. Regionální data

4.1 Modely pro diskrétní regionální data

U regionálních dat je sledována určitá veličina vztažená ke geografické oblasti (okres, kraj, stát apod.). K jejich popisu se hodí použít náhodná pole na mříži. Vrcholy mříže L představují jednotlivé oblasti. Relace sousedství $\sim$ může být definována například tak, že dva regiony jsou v relaci, pokud mají společnou hranici. Data jsou často tvořena zaznamenanými počty nějaké události v dané oblasti (např. počet nakažených, počet trestných činů). Modelování diskrétních prostorových dat pak může být založeno na zobecněných lineárních modelech.

Nechť $\boldsymbol{Z} = \{Z_i : i \in L\}$ a $\boldsymbol{W} = \{W_i : i \in L\}$ jsou náhodná pole na mříži L. Předpokládejme, že podmíněně při $\boldsymbol{W}$ jsou $\{Z_i\}$ nezávislé náhodné veličiny se střední hodnotou $\mathbb{E}(Z_i \mid \boldsymbol{W}) = \mu_i$. Dále mějme dánu funkci h (tzv. *spojovací funkce*) a předpokládejme, že

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

kde $\boldsymbol{F}_i = (f_{0i}, \ldots, f_{pi})^{\mathrm{T}}$ je vektor vysvětlujících proměnných a $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^{\mathrm{T}}$ je vektor regresních parametrů. Pro binární data se obvykle používá logit, tj. $h(\mu) = \log \frac{\mu}{1-\mu}$. Náhodné pole $\boldsymbol{W}$ modeluje prostorovou závislost, můžeme pro něj použít některý z gaussovských modelů (CAR, SAR, SMA, SARMA).

Pro modelování počtů událostí se nejčastěji uvažuje Poissonův model. Předpokládejme, že pro každé $i \in L$ má $Z_i \mid \mu_i$ Poissonovo rozdělení s parametrem $E_i \theta_i$, kde E_i je známý očekávaný počet události v regionu i. Jako spojovací funkci použijeme logaritmus a dostaneme lineární model pro θ_i:

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

Jedním z oborů, kde se tento model uplatňuje, je epidemiologie. V tom případě Z_i představuje pozorovaný počet případů daného onemocnění v regionu i a E_i je očekávaný počet případů onemocnění, ten může být znám z nějaké dodatečné informace o problému nebo můžeme uvažovat, že je to nějaká známá funkce počtu n_i lidí ohrožených onemocněním. Například může být $E_i = r n_i$, kde r je celková míra nakažení v celé populaci a můžeme ji odhadnout podílem

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

Tato volba odpovídá tomu, že ve všech oblastech očekáváme stejnou míru onemocnění. Hodnota θ_i udává skutečné relativní riziko nemoci v regionu i. Vysvětlující proměnné mohou být třeba míra znečištění ovzduší, což nejspíš bude mít vliv u chorob dýchacích cest. Na celou situaci lze také pohlížet jako na hierarchický model a pro statistické vyhodnocení použít bayesovské metody.

4.2 Odhad parametrů

Uvažujme markovské náhodné pole $\{Z_i : i \in L\}$ s hustotou $p(\boldsymbol{z}; \boldsymbol{\theta})$ parametrizovanou konečně rozměrným vektorem $\boldsymbol{\theta}$. Pro diskrétní data je hustota $p(\boldsymbol{z}; \boldsymbol{\theta})$ rovna sdružené pravděpodobnosti $\mathbb{P}(Z_i = z_i, i \in L)$, $\boldsymbol{z} = (z_i, i \in L)$. Pro spojitá data jde o sdruženou hustotu vzhledem k n-rozměrné Lebesgueově míře.

Pro odhad parametrů je ve statistice nejpopulárnějším postupem metoda maximální věrohodnosti. Pro daná data $\boldsymbol{z} = (z_i, i \in L)$ hledáme hodnotu $\hat{\boldsymbol{\theta}}$, která maximalizuje věrohodnostní funkci $L(\boldsymbol{\theta}) = p(\boldsymbol{z}; \boldsymbol{\theta})$.

Pro markovské náhodné pole s Gibbsovým rozdělením je

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

Problém je, že normující konstanta závisí na $\boldsymbol{\theta}$ a obvykle je velmi komplikovaná. Existují metody, kterými je možné normující konstantu aproximovat pomocí simulací (většinou MCMC) a potom maximalizovat takto aproximovanou věrohodnost.

Jednodušší možnost je uvažovat tzv. *pseudověrohodnost*

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

Tentokrát lze normující konstantu $c(\boldsymbol{z}_{\partial i}, \boldsymbol{\theta})$ často vyjádřit (v diskrétním případě jde o sumu $|S|$ členů, kde S je stavový prostor). Pokud bychom očíslovali prvky L pomocí $1, \ldots, n$, tak věrohodnost lze zapsat jako

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

Když nahradíme podmíněné hustoty $p(z_k \mid z_{k+1}, \ldots, z_n; \boldsymbol{\theta})$ plně podmíněnými hustotami $p(z_k \mid \boldsymbol{z}_{-k}; \boldsymbol{\theta})$, které jsou díky markovské vlastnosti rovny $p(z_k \mid \boldsymbol{z}_{\partial k}; \boldsymbol{\theta})$, dostaneme pseudověrohodnost $L_P(\boldsymbol{\theta})$.

Odhad metodou maximální pseudověrohodnosti patří do třídy odhadů, které jsou ve statistice označovány jako M-odhady. Obecně je M-odhad parametru $\boldsymbol{\theta}$ řešením úlohy maximalizace kontrastní funkce $\varrho(\boldsymbol{Z}, \boldsymbol{\theta})$. V klasické situaci maximální věrohodnosti pro posloupnost nezávislých stejně rozdělených náhodných veličin je

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

V našem případě je

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

4.3 Testování prostorové autokorelace

Připomeňme, že pro náhodné pole $\{Z_i : i \in L\}$ s konstantní střední hodnotou $\mathbb{E}Z_i = \mu$ a konstantním rozptylem $\operatorname{var} Z_i = \sigma^2$ jsme definovali *Moranův index (Moran's I)* předpisem

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

a *Gearyho index (Geary's c)* vztahem

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

kde w_{ij} jsou prostorové váhy blízkosti a $w = \sum_{i \in L} \sum_{j \in L} w_{ij}$. K výpočtu Moranova a Gearyho indexu jsou v balíčku spdep k dispozici funkce moran a geary.

Tyto veličiny se používají při testování hypotézy, že v datech jsou nulové prostorové autokorelace. Označme M jednu z testových statistik (Moranův nebo Gearyho index) a M_{obs} tuto testovou statistiku spočtenou z dat. Pro testování se obvykle používá některý z následujících čtyřech přístupů.

1. *permutační test*: Nulovou hypotézu nepřítomnosti prostorových autokorelací chápeme tak, že pozorovaným hodnotám Z_i, $i \in L$, jsou body mříže přiřazeny zcela náhodně. Pro n vrcholů mříže máme celkem $n!$ možných přiřazení. Spočteme-li veličinu M pro $n!$ přiřazení, dostaneme její rozdělení za nulové hypotézy. Potom je možné zjistit pravděpodobnost, že hodnota M_{obs} bude překročena. Velké

nebo malé hodnoty této pravděpodobnosti svědčí proti nulové hypotéze (při oboustranném testu). U tohoto *přístupu založeného na znáhodnění* také můžeme z rozdělení testové statistiky za nulové hypotézy určit její střední hodnotu a rozptyl, označme je $\mathbb{E}_r M$ a $\mathrm{var}_r\, M$.

2. *Monte Carlo test*: I pro ne moc velká n je počet možných permutací velký. Místo výpočtu pro všechna přiřazení můžeme generovat k náhodných přiřazení a sestrojit empirické rozdělení M za nulové hypotézy. Čím větší k, tím lepší je aproximace rozdělení za nulové hypotézy. Spojíme k získaných hodnot s M_{obs} a spočteme pořadí M_{obs}. Pro extrémní hodnoty pořadí je hypotéza nekorelovanosti zamítnuta. Například pro $k = 999$ zamítneme hypotézu na hladině 5%, pokud pořadí M_{obs} je mezi 1 a 25 nebo mezi 976 a 1 000.
3. *Asymptotický test založený na normálním rozdělení*: Rozdělení M je možné určit za předpokladu, že známe rozdělení Z. Typickým předpokladem je normální rozdělení s konstantní střední hodnotou a konstantním rozptylem. Uvažujeme nulovou hypotézu tvaru $\mathrm{cov}(Z_i, Z_j) = 0$ pro $i \neq j$. Pak není těžké vyjádřit střední hodnotu a rozptyl M za nulové hypotézy, označme je $\mathbb{E}_g M$ a $\mathrm{var}_g\, M$. Vzhledem k tomu, že často lze ukázat asymptotickou normalitu zvolené testové statistiky, stačí porovnat
$$\frac{M_{obs} - \mathbb{E}_g M}{\sqrt{\mathrm{var}_g\, M}}$$
s kvantily normovaného normálního rozdělení.
4. *Asymptotický test založený na znáhodnění*: V tomto případě porovnáváme
$$\frac{M_{obs} - \mathbb{E}_r M}{\sqrt{\mathrm{var}_r\, M}}$$
s kvantily normovaného normálního rozdělení. Přitom $\mathbb{E}_r M$ a $\mathrm{var}_r\, M$ jsou získané z přístupu založeného na znáhodnění (jako v permutačním nebo Monte Carlo testu).

Dá se ukázat, že $\mathbb{E}_g I = \mathbb{E}_r I = -\frac{1}{n-1}$ a $\mathbb{E}_g c = \mathbb{E}_r c = 1$. Vyjádření rozptylu za předpokladu normality a znáhodnění se už ovšem liší (viz [1]). Moranovu a Gearyho statistiku lze interpretovat následovně: pokud $I > \mathbb{E}I$ nebo $c < \mathbb{E}c$, tak vrchol mříže má tendenci být spojen s vrcholem mříže, který má podobnou hodnotu pole, prostorová autokorelace je kladná. Naopak, pokud $I < \mathbb{E}I$ nebo $c > \mathbb{E}c$, mají hodnoty ve dvou sousedních vrcholech mříže tendenci být odlišné. Můžeme tak uvažovat jednostranné testy proti alternativě, že autokorelace je kladná nebo naopak záporná.

5. Dodatky

5.1 Náhodné cenzorování

Předpokládejme, že $T_1, \ldots, T_n$ jsou nezávislé stejně rozdělené nezáporné náhodné veličiny s distribuční funkcí F. Naším cílem je získat odhad distribuční funkce F. V případě, kdy pozorujeme všechny sledované veličiny $T_1, \ldots, T_n$, je přirozeným odhadem F empirická distribuční funkce
$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[T_i \leq t]}, \quad t \geq 0.$$

V některých situacích, ale nemáme k dispozici pozorování všech T_i. Může tomu být například proto, že některá měření sledované veličiny byla předčasně ukončena. Příkladem je medicínská studie, ve které se zkoumá vliv určité léčby na přežití skupiny pacientů, některá pozorování nejsou úplná, protože se pacient odstěhoval nebo čas vyhrazený pro studii uplynul. Jiný příklad pochází z teorie spolehlivosti, kde je měřena doba do poruchy určitého výrobku. Kromě náhodných veličin T_i (tzv. časy přežití nebo doby života) ještě uvažujeme náhodné veličiny $C_1, \ldots, C_n$ (tzv. časové cenzory). Přitom pozorujeme jen náhodný výběr $(\tilde{T}_1, D_1), \ldots, (\tilde{T}_n, D_n)$, kde $\tilde{T}_i = \min(T_i, C_i)$ jsou cenzorované časy přežití a $D_i = \mathbf{1}_{[T_i \leq C_i]}$ jsou indikátory necenzorování. Pro $D_i = 1$ máme k dispozici skutečné pozorování T_i. Naopak pokud je $D_i = 0$ (nastalo cenzorování), máme pouze částečnou informaci, že $T_i \geq \tilde{T}_i$. V případě náhodného cenzorování předpokládáme, že $C_1, \ldots, C_n$ jsou nezávislé stejně rozdělené náhodné veličiny a nezávislé na $T_1, \ldots, T_n$. Potom neparametrickým maximálně věrohodným odhadem F je Kaplanův-Meierův odhad zavedený v [3] a definovaný jako
$$\hat{F}_{KM}(t) = 1 - \prod_{s \leq t} \left(1 - \frac{\#\{i : \tilde{T}_i = s, D_i = 1\}}{\#\{i : \tilde{T}_i \geq s\}}\right).$$

Do součinu ve skutečnosti efektivně přispívá jenom konečně mnoho členů, které odpovídají těm časům s, kdy je pozorován konec doby života. Odhad $\hat{F}_{KM}(t)$ je neklesající, zprava spojitá funkce, která může mít limitu pro $t \to \infty$ menší než 1 (když největší pozorovaná hodnota je cenzorována).

Intuitivní vysvětlení Kaplanova-Meierova odhadu je následující. Rozdělme interval $[0, t)$ na menší intervaly $[0, t_1), [t_1, t_2), \ldots, [t_k, t)$. Potom

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

přičemž podmíněné pravděpodobnosti

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

odhadujeme pomocí

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

Zjemňováním intervalů $[t_{j-1}, t_j)$ dostáváme limitní tvar $\hat{F}_{KM}(t)$.

Literatura

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

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

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

[4] 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.

[5] P. Lachout (2004): *Teorie pravděpodobnosti*, druhé vydání, Karolinum, Praha.

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

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

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

[9] 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.