Náhodné kótované množiny a redukce dimenze

Viktor Beneš, Ondřej Šedivý, Jakub Staněk

KPMS MFF UK Praha

ROBUST 2012, Němčičky

Obsah

RMCS - náhodné kótované uzavřené množiny

Molchanov (1984), Ballani, Kabluchko, Schlather (2009)

RMCS je dvojice $(\mathcal{Y}, \Lambda)$ kde Λ je náhodná funkce (kóta) definovaná na náhodné uzavřené množině $\mathcal{Y} \subset \mathbb{R}^d$.

Příklady RMCS:

(i) Kótovaný bodový proces - $\mathcal{Y}$ je bodový proces (dimenze $k = 0$), $\Lambda : \mathcal{Y} \to \mathbb{R}$ kóta, viz Stoyan et al. (1995), Illian et al. (2008).

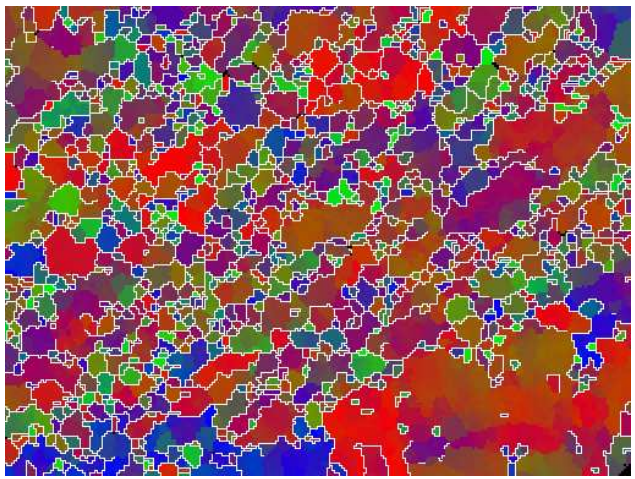
(ii) Úrovňové množiny náhodného pole Λ v $\mathbb{R}^d$ (dimenze $k = d$). Nechť $\mathcal{Y}_t = \{x \in \mathbb{R}^d : \Lambda(x) \geq t\}$, $t \in \mathbb{R}$. Potom $(\mathcal{Y}_t, \Lambda)$ je RMCS.

Cíl přednášky: Studovat k-rozměrné RMCS v $\mathbb{R}^d$ pro $k = 0, 1, \ldots, d-1$.

1.1 Náhodné mozaiky - RMCS různé dimenze

v $\mathbb{R}^3$ $\mathcal{Y}$ systém buněk (dimenze 3), stěn (2), hran (1), vrcholů (0).

Geometrické kóty - objem buněk, plocha povrchu stěn, délka hran, atd.



V materiálovém výzkumu - mikrostruktura zrn v kovech

Orientace krystalografických mřížek buněk, kóta stěn - disorientace mřížek sousedních buněk, ovlivňuje vlastnosti kovů

Charakteristiky náhodných množin $\mathcal{Y}$

$\mathcal{H}^k$ Hausdorffova míra, $\mathcal{Y}$ Hausdorffovy dimenze k

Lokálně konečná míra $\Psi(B) = \mathcal{H}^k(B \cap \mathcal{Y}),\ B \in \mathcal{B}^d$.

Míra intenzity $\varrho(B) = \mathbb{E}\Psi(B)$. Funkce intenzity ρ - hustota ϱ.

Bud' $M(D, B) = \mathbb{E}\left[\int_D \frac{\Psi(dy)}{\rho(y)} \int_B \frac{\Psi(dy)}{\rho(y)}\right],\ B, D \in \mathcal{B}^d,$

Stacionarita Ψ druhého řádu převážená intenzitou (SOIRS) když

$$M(D, B) = M(D + x, B + x) \quad \forall x \in \mathbb{R},\ B, D \in \mathcal{B}^d.$$

Redukovaná druhá momentová míra

$$\mathcal{K}(B) = \frac{1}{|A|}\mathbb{E}\left[\int_A \int \frac{\mathbf{1}_B(x - y)}{\rho(x)\rho(y)}\Psi(dx)\Psi(dy)\right].$$

K-funkce $K(r) = \mathcal{K}(b(0, r))$.

Nezávislost náhodné množiny a kóty

Bud' Λ' shora polospojité náhodné pole na $\mathbb{R}^d$ a RMCS $(\mathcal{Y}, \Lambda)$ je taková, že $\Lambda = \Lambda'$ na $\mathcal{Y}$.

Když jsou $\mathcal{Y}$ a Λ' nezávislé, píšeme, že $(\mathcal{Y}, \Lambda)$ je IRMCS.

Test H_0 : $(\mathcal{Y}, \Lambda)$ je IRMCS proti H_A : není IRMCS.

a) vyber náhodně m testovacích bodů $x_i \subset \mathcal{Y}$,

b) odhad $\rho(x_i)$, $\Lambda(x_i)$, $i = 1, \ldots, m$,

c) odhad Λ-vážené K funkce:
$\hat{K}_\Lambda(r) = \frac{1}{m} \sum_{x_i} \Lambda(x_i) \int_{\mathcal{Y}} \frac{\mathbf{1}(||x_i - y|| < r)}{\rho(x_i)\rho(y)} \mathcal{H}^k(dy)$,

d) v c) n permutací $\{\Lambda(x_i), i = 1, \ldots m\}$, výpočet $\hat{K}_{max}(r)$, $\hat{K}_{min}(r)$,

e) graf obálek $\hat{L}_{max}(r) - \hat{L}_\Lambda(r)$, $\hat{L}_{min}(r) - \hat{L}_\Lambda(r)$, kde $\hat{L}_\cdot = \left(\frac{\hat{K}_\cdot}{\omega_d}\right)^{\frac{1}{d}}$,

f) H_0 zamítneme, je-li vodorovná osa vně obálek.

Numerická studie RMCS

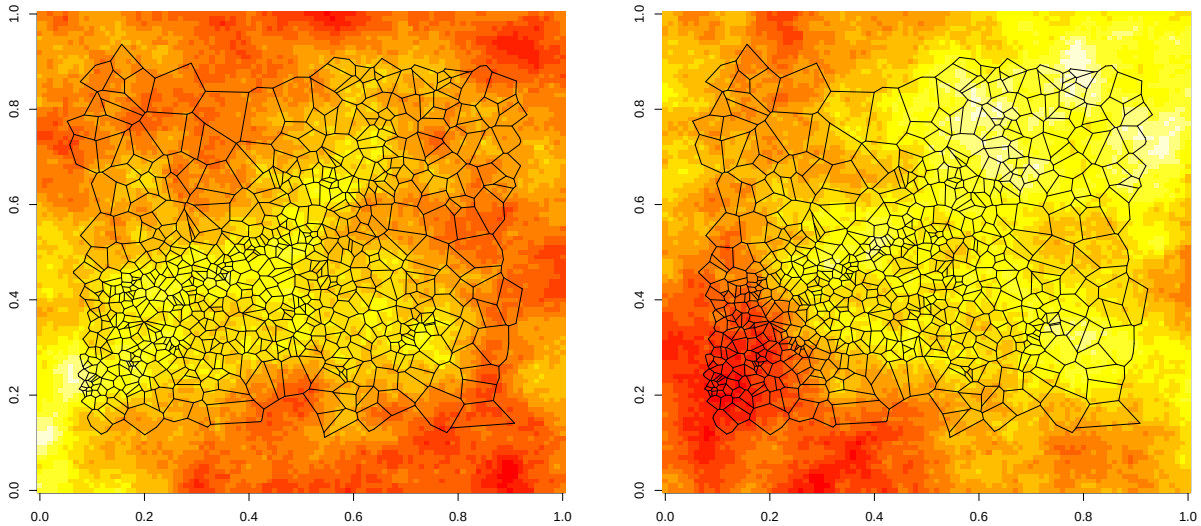
$$X(s) = (X_1(s), X_2(s)),\ s \in [0,1]^2$$

stacionární Gaussovské náhodné pole s nezávislými složkami a

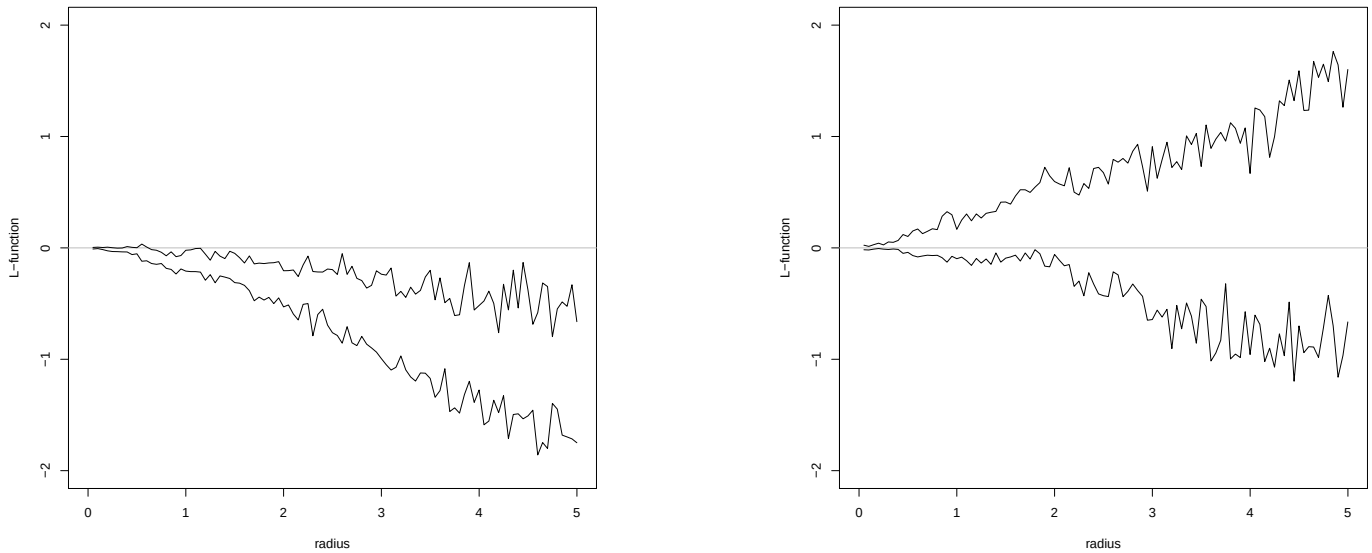
$$\mathbb{E}X = 0,\ R(s,t) = \alpha \exp(-\sigma||s-t||).$$

Necht' $\lambda(s) = a \exp(X_1(s)),\ a > 0$ je řídící funkce intenzity Coxova bodového procesu Φ.

$\mathcal{Y}$ je sjednocení hran 2D Voronoiovy mozaiky generované procesem Φ.



Výsledky testu IRMCS $(\mathcal{Y}, V_i)$, $V_i = \exp X_i$, $i = 1, 2$.



Vlevo ($i = 1$) H_0 zamítáme, $(\mathcal{Y}, V_1)$ není IRMCS. Vpravo ($i = 2$) H_0 nelze zamítnout.

l-tá momentová míra náhodné míry Ψ je

$$\mu^{(l)}(A_1 \times \cdots \times A_l) = \mathbb{E}[\Psi(A_1) \dots \Psi(A_l)]$$

for $A_1, \dots, A_l \in \mathcal{B}^d$.

Funkce intenzity l-tého řádu λ_l

$$\mu^{(l)}(ds_1 \times \cdots \times ds_l) = \lambda_l(s_1, \dots, s_l) ds_1 \dots ds_l.$$

Redukce dimenze ve stochastické geometrii

RMCS (Y, Γ), Y je $\mathcal{H}^k$-množina v $k = 0, 1, d-1$ v $\mathbb{R}^d$

Doprovodné proměnné $X = (X_1, \ldots, X_p)$

Definice: Necht' X je p-rozměrné náhodné pole na $\mathbb{R}^d$.

Je-li Y podmíněně nezávislé na X při daném $B^T X$ pro nějakou $p \times c$ matici B, $c \leq p$,

potom $\mathcal{S}(B)$ (lineární podprostor generovaný sloupci matice B) je postačující podprostor redukce dimenze (SDRS).

$\mathcal{S}_{Y|X} = \cap_{SDRS} \mathcal{S}(B)$ je centrální podprostor (CS).

Zjemněná definice redukce dimenze

Nechť $l \in \mathbb{N}$ a platí

$$\lambda_l(s_1, \ldots, s_l) = f_l(B^T X(s_1), \ldots, B^T X(s_l)),$$

pro nějakou měřitelnou funkci f_l a $p \times c$ matici B, $c \leq p$.

Potom $\mathcal{S}(B)$ se nazývá postačující podprostor redukce dimenze pro intenzitu $l-$tého řádu

$\mathcal{S}_l = \cap_B \mathcal{S}(B)$ se nazývá centrální podprostor pro intenzitu $l-$tého řádu.

Guan (2010) pro bodové procesy: $\mathcal{S}_{Y|X} = \cup_{l \geq 1} \mathcal{S}_l$,

Problém:
Odhad centrálního podprostoru (generátoru B) a jeho dimenze c.

Metody redukce dimenze

Na bodové procesy adaptovány metody SAVE (Cook, Weisberg, 1991), směrová regrese (Li, Wang, 2007), z vektorových dat

Zde se omezíme na plátkovanou inverzní regresi (SIR, Li 1991)

(i) plátkování náhodné množiny Y pomocí vhodné kóty Γ,

(ii) výpočet plátkových průměrů náhodného pole X,

(iii) aplikace metody hlavních komponent na plátkové průměry.

Obor hodnot Γ rozdělíme na m disjunktních intervalů $J_1, \ldots, J_m$ - plátků.

Indukuje dělení Y na m disjunktních množin $(Y^1, \ldots, Y^m)$ s funkcemi intenzity $\lambda^{(i)}$,

$$\int_A \lambda^{(i)}(s)ds = \mathbb{E}\Psi_{Y^i}(A),\ A \in \mathcal{B}^d,\ i = 1, \ldots, m.$$

Analýza $\mathcal{S}_1$

Předpoklady: $\lambda(s) = f(B^T X(s))$ pro matici B typu $p \times c$, $c \leq p$,

X je $p-$rozměrné stacionární normované Gaussovské náhodné pole.

Y, $\{X(s), s \in Y\}$ jsou ergodické

Lemma

Necht' kóta Γ *je podmíněně nezávislá na* X *při daném* $B^T X$. *Potom* $\lambda^{(j)}(s) = f^j(B^T X(s))$ *pro nějaké nezáporné měřitelné funkce* f^j, $j = 1, \ldots, m$.

Theorem

$\mathcal{S}(V_1) \subset \mathcal{S}_1$ *pro*

$$V_1 = \frac{1}{\mathbb{E}[\lambda(.)]} \sum_{j=1}^{m} \frac{\mathbb{E}[\lambda^{(j)}(.)X(.)]\mathbb{E}[\lambda^{(j)}(.)X(.)]^T}{\mathbb{E}[\lambda^{(j)}(.)]}.$$

Konvexní kompaktní okno $C \subset \mathbb{R}^d$, statistika

$$\hat{V}_1 = \frac{1}{\Psi_Y(C)} \sum_{j=1}^{m} \frac{1}{\Psi_{Y^j}(C)} \int_{Y^j \cap C} X(s)\mathcal{H}^k(ds)[\int_{Y^j \cap C} X(s)\mathcal{H}^k(ds)]^T.$$

Odhad B

$\tilde{X}$ Gaussovské náhodné pole
pozorované v okně $W \subset \mathbb{R}^d$ s Lebesgue mírou $0 < |W| < \infty$

$$\bar{X} = \frac{1}{|W|}\int_W \tilde{X}(s)ds, \quad \hat{\Sigma} = \frac{1}{|W|}\int_W [\tilde{X}(s) - \bar{X}][\tilde{X}(s) - \bar{X}]^T ds$$

Normované náhodné pole

$$X(s) = \hat{\Sigma}^{-1/2}[\tilde{X}(s) - \bar{X}]$$

RMCS (Y, Γ),

$$p_j = \mathbb{P}(x \in Y^j | x \in Y),\ j = 1, \ldots, m,$$

odhadnuto jako

$$\hat{p}_j = \frac{\mathcal{H}^k(Y^j)}{\mathcal{H}^k(Y)}.$$

Odhad B pokračování

$$m_j = \mathbb{E}(X(s)|s \in Y^j)$$

vážená kovarianční matice $V = \sum_{j=1}^{m} p_j m_j m_j^T$

odhad pomocí množiny $\{z\}_j$ testovacích bodů v $j-$tém plátku Y^j, $j = 1, \ldots, m$,

$$\hat{m}_j = \frac{1}{card\{z\}_j} \sum_{\{z\}_j} X(z).$$

$$\hat{V} = \sum_{j=1}^{m} \hat{p}_j \hat{m}_j \hat{m}_j^T,$$

najdi c největších vlastních čísel $\hat{V}$, odpovídající vlastní vektory $\hat{\eta}_i$ definují

$$\hat{\beta}_i = \hat{\Sigma}^{-1/2} \hat{\eta}_i, \ i = 1, \ldots, c.$$

Pak $\hat{\beta}_i$ patří do centrálního podprostoru.

Test ortogonality, simulační studie

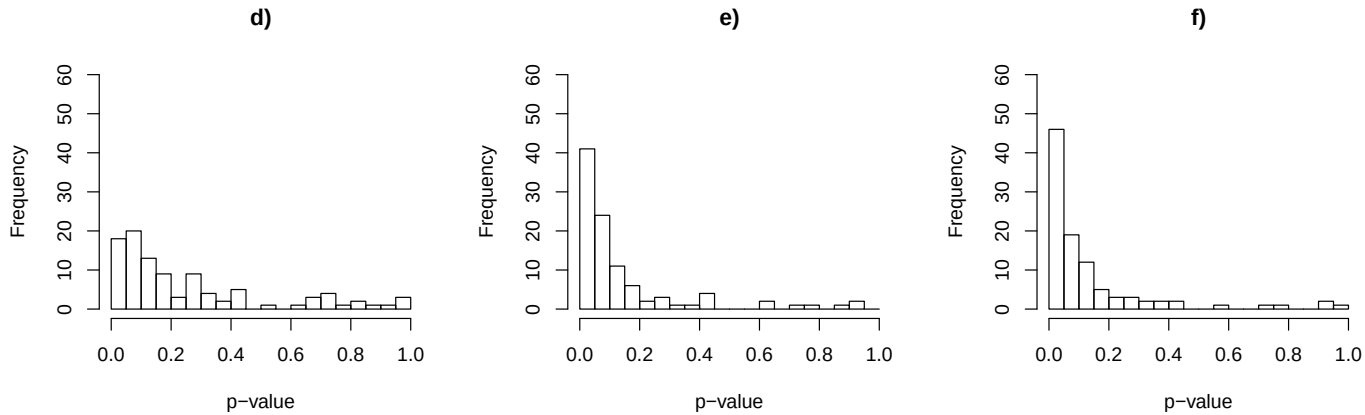
$$H_0 : \ c = 0 \quad \text{proti} \quad H_A : \ c > 0$$

H_0 ekvivalentní nezávislosti X a Y, implikuje $R^2 = 0$, kde

$$R^2(\beta_i) = max_{\beta \in \mathcal{S}_{Y|X}} \frac{(\beta_i^T \beta)^2}{\beta_i^T \beta_i \beta^T \beta},$$

a) výpočet $R^2(\hat{\beta}_i) = R_e^2$ z pozorování Y a X,

b) výpočet $R^2(\hat{\beta}_i) = R_j^2$, $j = 1, \ldots, n$, z pozorování Y a n simulovaných realizací X, c) p-hodnota pro test $H_0 : \ R^2 = 0$ je $\frac{card\{R_j^2 \geq R_e^2\}+1}{n+1}$.



Test ortogonality, RMCS hran mozaiky v $\mathbb{R}^2$. Histogramy p-hodnot pro počet plátků 1, 2, 4, z $q = 100$ simulací.

Další numerické výsledky

Veličina

$$\triangle(B, \widehat{B}) = ||B(B^T B)^{-1} B^T - \widehat{B}(\widehat{B}^T \widehat{B})^{-1} \widehat{B}^T||_{\max}$$

k porovnání odhadnuté a skutečné matice centrálního podprostoru.

počet plátků	$\triangle(B, \widehat{B})$
1	0.582
2	0.343
4	0.195
8	0.167
16	0.159
32	0.160

Výběrové průměry $\triangle(B, \widehat{B})$ z $q = 100$ simulací, $p = 3$,
$Z_1 = a(\arctan(X_1(s)) + \frac{\pi}{2}),\ a = 0.02$

Analýza $\mathcal{S}_2$

Necht' $\lambda_2(s,t) = f_2(B^T X(s), B^T X(t))$ pro matici B typu $p \times c,\ c \leq p$,

Y a $\{X(s)X(t)^T,\ s,t \in Y\}$ jsou ergodické, $C \subset \mathbb{R}^d$ konvexní kompaktní okno. Pak

$$\hat{M}_2 = \frac{\int_{s,t \in Y \cap C} X(s)X(t)^T \mathcal{H}^k(ds)\mathcal{H}^k(dt)}{\Psi_Y(C)^2} \to \frac{\int\int \mathbb{E}[\lambda_2(s,t)X(s)X(t)^T]dsdt}{\int\int \mathbb{E}[\lambda_2(s,t)]dsdt} =$$

$= M_2$ když $C \uparrow \mathbb{R}^d$. Kdy je $\mathcal{S}(M_2) \subset \mathcal{S}_2$?

Theorem

Bud' $P_B = B(B^T B)^{-1} B^T$ projekční matice, $Q_B = I_p - P_B$, pak

$$M_2 = M_2^P + M_2^Q = \frac{P_B \int\int \mathbb{E}[f_2(B^T X(s), B^T X(t))X(s)X(t)^T]dsdtP_B}{\int\int \mathbb{E}[f_2(B^T X(s), B^T X(t))]dsdt} +$$

$$+ \frac{\int\int \mathbb{E}[f_2(B^T X(s), B^T X(t))]\mathbb{E}[Q_B X(s)X(t)^T Q_B]dsdt}{\int\int \mathbb{E}[f_2(B^T X(s), B^T X(t))]dsdt}.$$

Příklad: Determinantový bodový proces

$$\lambda_2(s,t) = \begin{vmatrix} C_0(0) & C_0(s-t) \\ C_0(t-s) & C_0(0) \end{vmatrix}$$

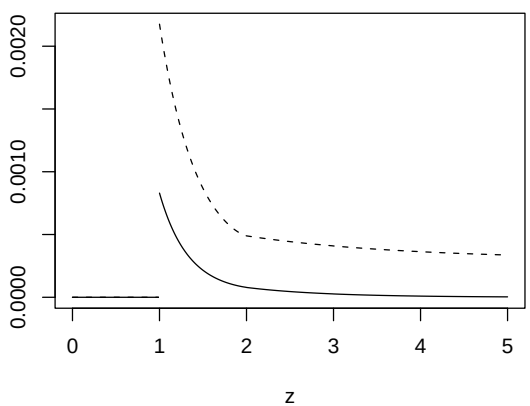
kde C_0 je kovarianční funkce, volme

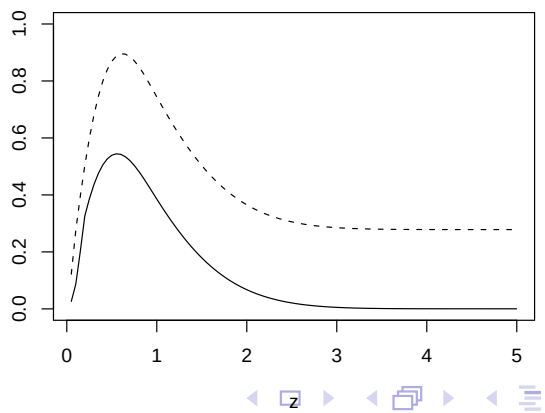
$$C_0(x) = \frac{2\rho}{\pi}\left(\arccos\frac{||x||}{\alpha} - \frac{||x||}{\alpha}\sqrt{1-\left(\frac{||x||}{\alpha}\right)^2}\right)\mathbf{1}_{[||x||<\alpha]}.$$

Parametr ρ je znáhodněn:

$$\rho = \frac{4}{\pi^2\alpha^2}(\arctan(X_1(s)X_1(t)) + \frac{\pi}{2}).$$

Obrázek vpravo: Integrand v čitateli M_2^P čárkovaně, M_2^Q plná čára.





Plátkování pro odhad $\mathcal{S}_2$

Kótování se aplikuje na Kartézský součin $Y \times Y$

$$\begin{array}{lll} \mathcal{Y} = Y \times Y, & \Psi_{\mathcal{Y}}(C) = \Psi_Y(C)^2, & k > 0, \\ \mathcal{Y} = \{(s,t); s \in Y, t \in Y, s \neq t\}, & \Psi_{\mathcal{Y}}(C) = \Psi_Y(C)(\Psi_Y(C) - 1), & k = 0. \end{array}$$

Kóta $\Gamma : \mathcal{Y} \to \mathbb{R}$, $\Gamma(s,t) = \Gamma(t,s)$ pro $(s,t) \in \mathcal{Y}$.

Obor hodnot Γ rozdělen na m disjunktních plátků

Indukuje rozklad $\mathcal{Y}$ na $(\mathcal{Y}^1, \ldots, \mathcal{Y}^m)$, bud'

$$\Psi_{\mathcal{Y}^j}(C) = \int \int_{\mathcal{Y}^j \cap C^2} \mathcal{H}^k(ds)\mathcal{H}^k(dt).$$

plátkové průměry $o_j = \mathbb{E}(X(s)X(t)^T | (s,t) \in \mathcal{Y}^j)$

$q_j = P((s,t) \in \mathcal{Y}^j | (s,t) \in \mathcal{Y})$, $j = 1, \ldots, m$.

Matice $U_2 = \sum_{j=1}^m q_j o_j o_j^T$, metoda hlavních komponent.

Literatura

F. Ballani, Z. Kabluchko, M. Schlather (2009) Random marked sets. arXiv:0903.2388v1 [math.PR]
V. Beneš, J. Rataj (2004) Stochastic Geometry: Selected Topics. Kluwer Acad. Publ. Dordrecht.
R.D. Cook (1998) Regression Graphics. Wiley, New York.
Y. Guan (2008) On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. JASA 103, 483, 1238–1247.
Y. Guan, H. Wang (2010) Sufficient dimension reduction for spatial point processes directed by Gaussian random fields. J. R. Statist. Soc. B, 72, 3, 367–87.
J. Illian, A. Penttinen, H. Stoyan, D. Stoyan (2008) Statistical Analysis and Modelling of Spatial Point Patterns. Wiley, New York.
F. Lavancier, J. Møller, E. Rubak (2012) Statistical aspects of determinantal point processes, submitted.
K.-C. Li (1991) Sliced inverse regression for dimension reduction. JASA 86, 316–327.