Odhady párové korelační funkce bodového procesu

Jakub Vondráček

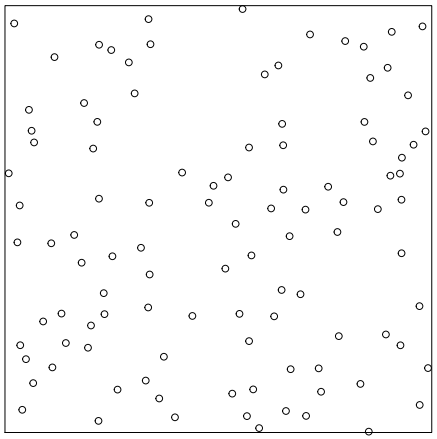
30. března 2020

Osnova

Motivace

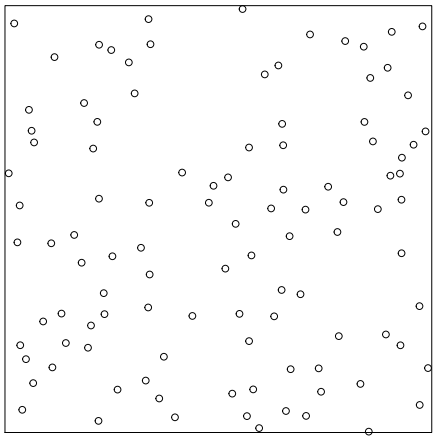
Motivace

- Představte si čtverec $[0,1] \times [0,1]$ a v něm body, viz obrázek. To představuje realizaci bodového procesu v rovině.



Motivace

- Představte si čtverec $[0,1] \times [0,1]$ a v něm body, viz obrázek. To představuje realizaci bodového procesu v rovině.



- Mezi body procesu můžeme pozorovat různé interakce: body se přitahují, odpuzují nebo se vůbec neovlivňují.

Motivace

Motivace

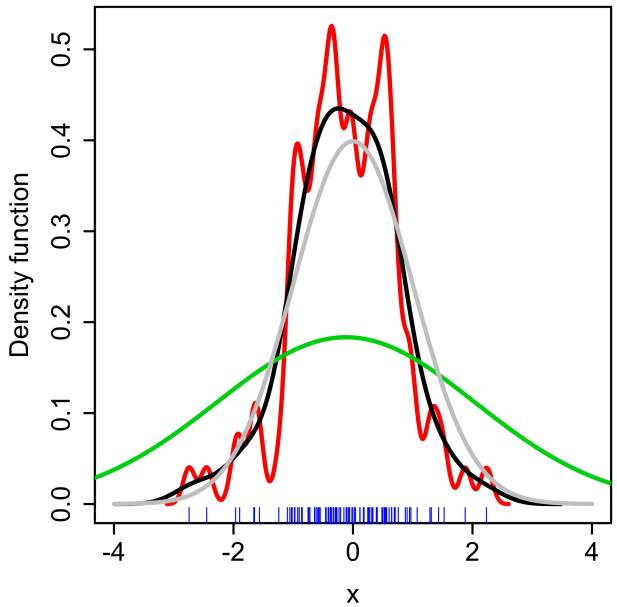
- Teorie bodových procesů nabízí mnoho možností, jak tyto interakce měřit. Jednou z nich je párová korelační funkce.

Motivace

- Teorie bodových procesů nabízí mnoho možností, jak tyto interakce měřit. Jednou z nich je párová korelační funkce.
- Pokud máme reálná data, chceme tuto funkci odhadnout. K tomu slouží metoda jádrových odhadů. Vlastnosti jádrových odhadů jsou silně ovlivněny volbou tzv. „šířky pásma" b které určuje, jak vzdálená pozorování při odhadu ještě bereme do úvahy. Pokud zvolíme b moc malé, dostaneme odhad, který je velmi závislý na pozorovaných datech(graf funkce bude velmi zubatý), zatímco moc velké b může naopak zakrýt důležité informace(graf bude velmi hladký a nebude vidět struktura dat). Správná volba b je hlavním těžištěm této práce.

Odhad hustoty náhodné veličiny: malé vs. velké b

Odhad hustoty náhodné veličiny: malé vs. velké b



- Šedá křivka představuje odhadovanou hustotu, červená odhad s malým b, černá odhad se skoro optimálním b a zelená odhad s moc velkým b.

Bodový proces

Bodový proces

- Přesná definice bodového procesu je relativně dost technická, definuje se obdobně jako náhodná veličina s tím, že jeho hodnoty jsou lokálně konečné podmnožiny $\mathbb{R}^2$.

Bodový proces

- Přesná definice bodového procesu je relativně dost technická, definuje se obdobně jako náhodná veličina s tím, že jeho hodnoty jsou lokálně konečné podmnožiny $\mathbb{R}^2$.
  - Řekneme, že M je lokálně konečná, pokud pro každou B omezenou borelovskou množinu je počet bodů v $M \cap B$ konečný.

Bodový proces

- Přesná definice bodového procesu je relativně dost technická, definuje se obdobně jako náhodná veličina s tím, že jeho hodnoty jsou lokálně konečné podmnožiny $\mathbb{R}^2$.
  - Řekneme, že M je lokálně konečná, pokud pro každou B omezenou borelovskou množinu je počet bodů v $M \cap B$ konečný.
- Můžeme tedy bodový proces považovat za lokálně konečnou náhodnou podmnožinu $\mathbb{R}^2$. S touto interpretací si bez technických detailů pro pochopení teorie a výsledků vystačíme.

Definice 1

Definice 1

Nechť X je bodový proces. Nechť $\mathcal{B}^2$ je množina borelovských množin na $\mathbb{R}^2$ a $n(A)$ značí počet bodů v množině A.

Definice 1

Nechť X je bodový proces. Nechť $\mathcal{B}^2$ je množina borelovských množin na $\mathbb{R}^2$ a $n(A)$ značí počet bodů v množině A.

Míra intenzity, intenzita.

Nechť $\mu(\cdot)$ je míra na $\mathbb{R}^2$ splňující

$$\mu(B) = \mathbb{E}n(X_B), B \in \mathcal{B}^2.$$

X_B značí $X \cap B$. Potom se $\mu(\cdot)$ nazývá míra intenzity. Pokud existuje hustota $\mu(\cdot)$ vzhledem k Lebesgueově míře, říkáme, že to je funkce intenzity procesu a značíme ji $p(\cdot)$.

Definice 1

Nechť X je bodový proces. Nechť $\mathcal{B}^2$ je množina borelovských množin na $\mathbb{R}^2$ a $n(A)$ značí počet bodů v množině A.

Míra intenzity, intenzita.

Nechť $\mu(\cdot)$ je míra na $\mathbb{R}^2$ splňující

$$\mu(B) = \mathbb{E}n(X_B), B \in \mathcal{B}^2.$$

X_B značí $X \cap B$. Potom se $\mu(\cdot)$ nazývá míra intenzity. Pokud existuje hustota $\mu(\cdot)$ vzhledem k Lebesgueově míře, říkáme, že to je funkce intenzity procesu a značíme ji $p(\cdot)$.

Pokud je p konstantní, lze interpretovat jako střední počet bodů procesu na jednotkovou plochu.

Definice 2

Definice 2

Faktoriální momentová míra druhého řádu

Řekneme, že $\alpha^2(\cdot)$ je faktoriální momentová míra druhého řádu, pokud splňuje:

$$\alpha^2(B) = \mathbb{E} \sum\nolimits^{\neq}_{x_1,x_2 \in X} \mathbb{1}_{(x_1,x_2 \in B)}\ , B \in (\mathcal{B}^2)^2,$$

kde horní index $\neq$ u sumy znamená, že sčítáme přes dvojice různých bodů $x_1 \neq x_2$. Dvojice jsou uspořádané. $\mathbb{1}$ značí indikátorovou funkci.

Definice 2

Faktoriální momentová míra druhého řádu

Řekneme, že $\alpha^2(\cdot)$ je faktoriální momentová míra druhého řádu, pokud splňuje:

$$\alpha^2(B) = \mathbb{E} \sum\nolimits^{\neq}_{x_1,x_2 \in X} \mathbb{1}_{(x_1,x_2 \in B)}\ , B \in (\mathcal{B}^2)^2,$$

kde horní index $\neq$ u sumy znamená, že sčítáme přes dvojice různých bodů $x_1 \neq x_2$. Dvojice jsou uspořádané. $\mathbb{1}$ značí indikátorovou funkci.

Součinová hustota druhého řádu

Pokud má $\alpha^2(\cdot)$ hustotu $p^{(2)}(\cdot,\cdot)$ vzhledem k Lebesgueově míře, nazýváme $p^{(2)}(\cdot,\cdot)$ součinovou hustotou druhého řádu.

Definice 3

Definice 3

Párová korelační funkce

Pokud $p(\cdot)$ i $p^{(2)}(\cdot,\cdot)$ existují, pak definujeme párovou korelační funkci jako

$$g(x,y) = \frac{p^2(x,y)}{p(x)p(y)} \geq 0.$$

Definice 3

Párová korelační funkce

Pokud $p(\cdot)$ i $p^{(2)}(\cdot,\cdot)$ existují, pak definujeme párovou korelační funkci jako

$$g(x,y) = \frac{p^2(x,y)}{p(x)p(y)} \geq 0.$$

Obecně, hodnoty g(x,y)$>$ 1 indikují přitažlivé interakce mezi body a hodnoty g(x,y)$<$ 1 odpudivé interakce. Hodnoty kolem 1 reprezentují vzájemnou nezávislost bodů. Často se pracuje s předpoklady ohledně rozdělení X, za kterých $g(x,y)$ závisí pouze na $x-y$ či $\|x-y\|$(zde a v dalším textu $\|\cdot\|$ značí eukleidovskou normu).

Jádrové odhady

Jádrové odhady

- Tuto funkci chceme odhadovat neparametrickou metodou jádrových odhadů. Ta se používá například pro odhadování hustot náhodných veličin. Myšlenka je následující: Nechť g je odhadovaná hustota náhodné veličiny Z.

Jádrové odhady

- Tuto funkci chceme odhadovat neparametrickou metodou jádrových odhadů. Ta se používá například pro odhadování hustot náhodných veličin. Myšlenka je následující: Nechť g je odhadovaná hustota náhodné veličiny Z.
  - Vezmeme nějakou „jádrovou" funkci k, obecně se tato funkce volí jako nějaká jednorozměrná symetrická hustota(tzn. $\int_{-\infty}^{\infty} k(x)\,\mathrm{d}x = 1$) s omezeným nosičem(většinou [-1,1]).

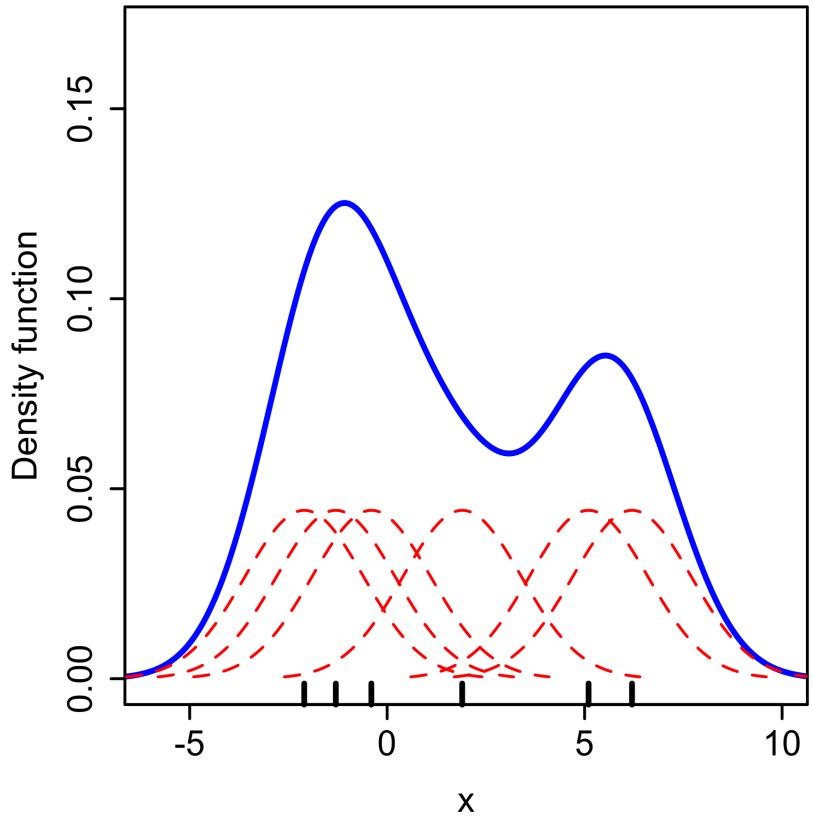
Jádrové odhady

- Tuto funkci chceme odhadovat neparametrickou metodou jádrových odhadů. Ta se používá například pro odhadování hustot náhodných veličin. Myšlenka je následující: Nechť g je odhadovaná hustota náhodné veličiny Z.
  - Vezmeme nějakou „jádrovou" funkci k, obecně se tato funkce volí jako nějaká jednorozměrná symetrická hustota(tzn. $\int_{-\infty}^{\infty} k(x)\,\mathrm{d}x = 1$) s omezeným nosičem(většinou [-1,1]).
  - Vydělíme k počtem pozorováním n, tím dostaneme funkci $k\frac{1}{n}$ která má integrál $1/n$.

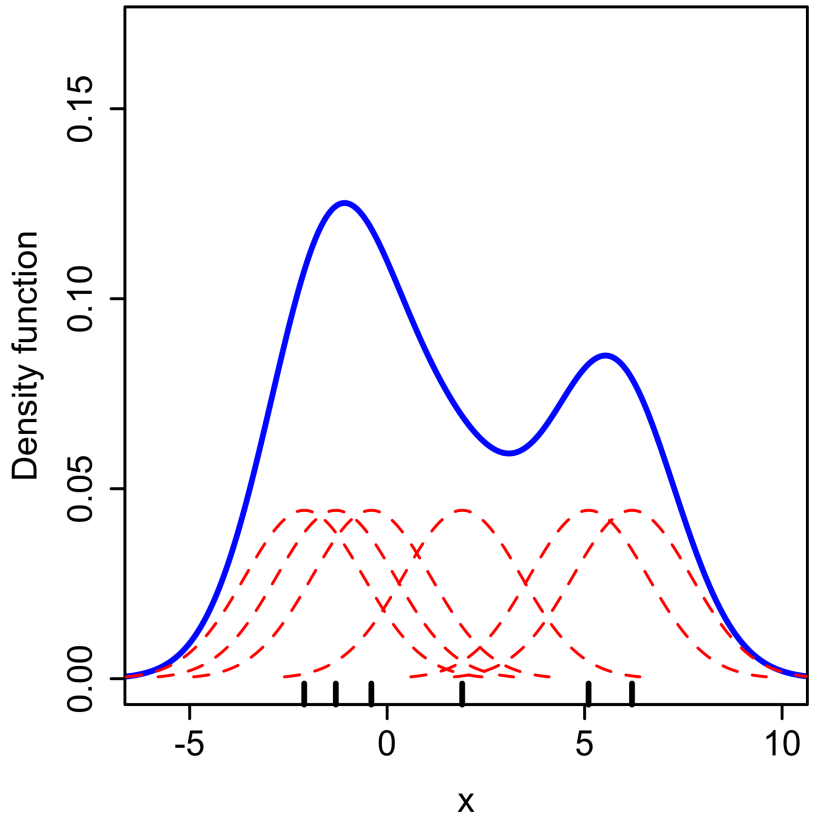
Jádrové odhady

- Tuto funkci chceme odhadovat neparametrickou metodou jádrových odhadů. Ta se používá například pro odhadování hustot náhodných veličin. Myšlenka je následující: Nechť g je odhadovaná hustota náhodné veličiny Z.
  - Vezmeme nějakou „jádrovou" funkci k, obecně se tato funkce volí jako nějaká jednorozměrná symetrická hustota(tzn. $\int_{-\infty}^{\infty} k(x)\,\mathrm{d}x = 1$) s omezeným nosičem(většinou [-1,1]).
  - Vydělíme k počtem pozorováním n, tím dostaneme funkci $k^{\frac{1}{n}}$ která má integrál $1/n$.
  - Na každé jednotlivé pozorování z_i vycentrujeme $k^{\frac{1}{n}}$ a jednotlivé funkce sečteme. Ve výsledku dostaneme funkci $\hat{g} = \sum_{i=1}^{n} k^{\frac{1}{n}}[z_i]$(zde $k^{\frac{1}{n}}[z_i]$ značí funkci $k^{\frac{1}{n}}$ vycentrovanou na z_i), která má integrál rovný 1 a to je náš odhad hustoty.

Jádrový odhad hustoty



Jádrový odhad hustoty



- Modrá plná křivka reprezentuje odhad $\hat{g}$, červené křivky reprezentují $k\frac{1}{n}[z_i]$.

Šířka pásma

Šířka pásma

- Nosič k ovšem nemusí být jenom $[-1, 1]$. Pomocí přeškálování $k_b(x) = \frac{k(x/b)}{b}$, $b > 0$ se nosič změní na $[-b, b]$. To umožní $k^{\frac{1}{n}}$ více „rozprostřít" kolem jednotlivých pozorování.

Šířka pásma

- Nosič k ovšem nemusí být jenom $[-1, 1]$. Pomocí přeškálování $k_b(x) = \frac{k(x/b)}{b}, b > 0$ se nosič změní na $[-b, b]$. To umožní $k^{\frac{1}{n}}$ více „rozprostřít" kolem jednotlivých pozorování.
- Parametr b se nazývá šířka pásma a ukazuje se, že jeho volba při odhadování je mnohem důležitější, než volba funkce k.

Šířka pásma

- Nosič k ovšem nemusí být jenom $[-1, 1]$. Pomocí přeškálování $k_b(x) = \frac{k(x/b)}{b}, b > 0$ se nosič změní na $[-b, b]$. To umožní $k^{\frac{1}{n}}$ více „rozprostřít" kolem jednotlivých pozorování.
- Parametr b se nazývá šířka pásma a ukazuje se, že jeho volba při odhadování je mnohem důležitější, než volba funkce k.
- Pro volbu b pro odhad párové korelační funkce existují různá doporučení založená na aproximaci, zkušenostech a simulacích, ale neexistuje „přesný vzoreček", který by dal optimální b. Obecně se b volí tak, aby minimalizovalo střední čtvercovou chybu

$$\mathbb{E}(\hat{g}(r, b) - g(r))^2, r \geq 0.$$

Odhady

Odhady

- Jádrové odhady párové korelační funkce mají tvar náhodných sum jako třeba
$$\hat{g}(r,b) = \sum_{u,v \in X_W}^{\neq} k_b(r - \|v-u\|)\Phi(u,v), r \geq 0$$
pro nějakou funkci $\Phi:(\mathbb{R}^2)^2 \to \mathbb{R}$, kde W značí pozorovací okno(třeba čtverec $[0,1]\times[0,1]$) a X_W značí $X \cap W$.

Odhady

- Jádrové odhady párové korelační funkce mají tvar náhodných sum jako třeba
$$\hat{g}(r,b) = \sum_{u,v \in X_W}^{\neq} k_b(r - \|v-u\|)\Phi(u,v), r \geq 0$$
pro nějakou funkci $\Phi{:}(\mathbb{R}^2)^2 \to \mathbb{R}$, kde W značí pozorovací okno(třeba čtverec $[0,1]\times[0,1]$) a X_W značí $X \cap W$.
- Momenty takových odhadů lze počítat pomocí Campbellovy věty:

Campbellova věta

Pro integrovatelnou funkci $h{:}(\mathbb{R}^2)^2 \to \mathbb{R}$ platí
$$\mathbb{E}\sum_{x_1,x_2 \in X}^{\neq} h(x_1,x_2) = \int_{(\mathbb{R}^2)^2} h(x_1,x_2)\,\mathrm{d}\alpha^2(x_1,x_2).$$

Cíle

Cíle

- Odvodit vychýlení a rozptyl(a z nich střední čtvercovou chybu) odhadů párové korelační funkce pomocí Campbellovy věty.

Cíle

- Odvodit vychýlení a rozptyl(a z nich střední čtvercovou chybu) odhadů párové korelační funkce pomocí Campbellovy věty.
- Pro vybrané modely spočítat b jako

$$argmin_{b>0}\, \mathbb{E}(\hat{g}(r,b) - g(r))^2, r \geq 0.$$

Cíle

- Odvodit vychýlení a rozptyl(a z nich střední čtvercovou chybu) odhadů párové korelační funkce pomocí Campbellovy věty.
- Pro vybrané modely spočítat b jako

$$argmin_{b>0}\, \mathbb{E}(\hat{g}(r,b) - g(r))^2, r \geq 0.$$

  - Poznámka: r se ve funkci výše může zvolit pevné, nebo se místo střední čtvercové chyby minimalizuje tzv. střední integrovaná čtvercová chyba definovaná jako

$$MISE(b) = \int_0^r MSE(r,b)\,\mathrm{d}r$$

kde $MISE(b)$ je střední integrovaná čtvercová chyba a $MSE(r,b)$ je střední čtvercová chyba.

Cíle

- Odvodit vychýlení a rozptyl(a z nich střední čtvercovou chybu) odhadů párové korelační funkce pomocí Campbellovy věty.
- Pro vybrané modely spočítat b jako

$$argmin_{b>0}\, \mathbb{E}(\hat{g}(r,b) - g(r))^2, r \geq 0.$$

  - Poznámka: r se ve funkci výše může zvolit pevné, nebo se místo střední čtvercové chyby minimalizuje tzv. střední integrovaná čtvercová chyba definovaná jako

$$MISE(b) = \int_0^r MSE(r,b)\,\mathrm{d}r$$

  kde $MISE(b)$ je střední integrovaná čtvercová chyba a $MSE(r,b)$ je střední čtvercová chyba.
- V simulační studii porovnat naše výsledky s doporučeními získanými z literatury na několika různých modelech bodových procesů.

Díky za pozornost!