Jádrové odhady hustoty náhodného vektoru.

Jiří Zelinka, KAM, přírodovědecká fakulta MU, Brno

Jádrové odhady hrají významnou roli mezi neparametrickými odhady, Jejich pojmenování je odvozeno od funkcí zvaných jádra. Tento příspěvek je zaměřen na jádrové odhady hustoty náhodného vektoru. Tyto odhady jsou zobecněním odhadů hustoty náhodné veličiny – viz např. [1].

Mějme náhodný výběr $\mathbf{X}_1, \ldots, \mathbf{X}_n$ z p - rozměrného spojitého rozdělení o hustotě $f(\mathbf{x})$, kde $\mathbf{x} = (x_1, \ldots, x_p)^T$. Tento náhodný výběr použijeme pro konstrukci odhadu $\hat{f}$ hustoty f.

Nechť reálná funkce K je definována na R^p a nechť platí

$$\int_{R^p} K(\mathbf{x})d\mathbf{x} = 1.$$

Tuto funkci budeme nazývat jádrem. Někdy se také požaduje, aby K byla spojitá, nezáporná či symetrická vzhledem k počátku, tyto vlastnosti však pro potřeby tohoto příspěvku vyžadovat nebudeme.

Jádrový odhad hustoty náhodného vektoru f má tvar

$$\hat{f}_h(\mathbf{x}) = \frac{1}{nh^p} \sum_{i=1}^{n} K(\frac{\mathbf{x} - \mathbf{X}_i}{h}). \tag{1}$$

Parametr h se nazývá šířka okna a závisí na n, bylo by tedy přesnější používat označení h_n. Je také možné použít p-rozměrný vektor h, tj. jinou hodnotu pro každou proměnnou. Odhad pak má tvar

$$\hat{f}_h(\mathbf{x}) = \frac{1}{nh_1 \ldots h_p} \sum_{i=1}^{n} K(\frac{x_1 - X_{i,1}}{h_1}, \ldots, \frac{x_p - X_{i,p}}{h_p}). \tag{2}$$

Často se používají jádra s kompaktním nosičem. Pro tento případ budeme používat charakteristickou funkci množiny, kterou označíme symbolem I s indexem charakterizujícím danou množinu.

Mezi nejpoužívanější typy jader patří:

- konstantní jádro $K(\mathbf{x}) = \frac{1}{2^p} I_{max=|x_i| \leq 1, i=1,\ldots,p}$
- gaussovské jádro $K(\mathbf{x}) = (2\pi)^{-p/2} e^{-\frac{1}{2}\mathbf{x}^T\mathbf{x}}$

- Epanečnikovo jádro $K(\mathbf{x}) = \frac{p+2}{2c_d}(1 - \mathbf{x}^T\mathbf{x})I_{\mathbf{x}^T\mathbf{x}\leq 1}$, kde $c_d = \frac{\pi^{p/2}}{\Gamma(d/2+1)}$

Podrobnější přehled používaných jader je možné najít například v [4]. Často se též používá jádro p proměnných ve tvaru součinu p jader jedné proměnné, tj. $K(x_1, \ldots, x_p) = \Pi_{j=0}^{p} K_j(x_j)$. Pro tento typ jader uvedeme tvrzení popisující konzistenci odhadu, nejprve ale zavedeme speciální třídu jader potřebnou pro toto tvrzení a některá označení.

Nechť H_s pro s sudé a kladné označuje třídu funkcí K jedné proměnné definovaných na celé reálné ose a splňujících vlastnosti:

(i) K je sudá a omezená funkce

(ii) $\int_{-\infty}^{\infty} K(x)dx = 1$

(iii) $\int_{-\infty}^{\infty} x^i K(x)dx = 0$, pro $i = 1, \ldots, s-1$

(iv) $\int_{-\infty}^{\infty} x^s K(x)dx \neq 0$

(v) $\int_{-\infty}^{\infty} x^s |K(x)|dx < \infty$.

Odhad hustoty f budeme uvažovat ve tvaru

$$\hat{f}_{h_n}(\mathbf{x}) = \frac{1}{nh_n^p} \sum_{i=1}^{n} \Pi_{j=1}^{p} K_j\left(\frac{x_j - X_{i,j}}{h_n}\right). \tag{3}$$

Integrální střední kvadratickou chybu IMSE příslušnou šířce okna h_n označme $U(h_n)$. Tato chyba je definována vztahem

$$U(h_n) = \int_{R^p} E[\hat{f}_{h_n}(\mathbf{x}) - f(\mathbf{x})]^2 d\mathbf{x}. \tag{4}$$

A nyní můžeme uvést tvrzení popisující konzistenci odhadu:

Věta. Nechť hustota f náhodného vektoru $\mathbf{X}$ má spojité a omezené parciální derivace až do řádu s včetně, přičemž tyto derivace jsou třídy $L_2(\mathrm{R}^p)$. Dále nechť jádra K_j, $j = 1, \ldots, p$ jsou z třídy H_s, $\alpha_j = \int_{-\infty}^{\infty} u^s K(u)du$. Konečně nechť h_n je posloupnost jdoucí k nule, přičemž $\lim_{n\to\infty} nh_n^p = \infty$. Pak pro $n \to \infty$ platí

$$U(h_n) = \frac{1}{nh_n^p} \int K^2(\mathbf{x})d\mathbf{x} + \frac{h_n^{2s}}{(s!)^2} \int \left[\sum_{j=1}^{p} \alpha_j \frac{\partial^{(s)} f(\mathbf{x})}{\partial x_j^s}\right]^2 d\mathbf{x} + O\left(\frac{1}{nh^p} + h_n^{2s}\right). \tag{5}$$

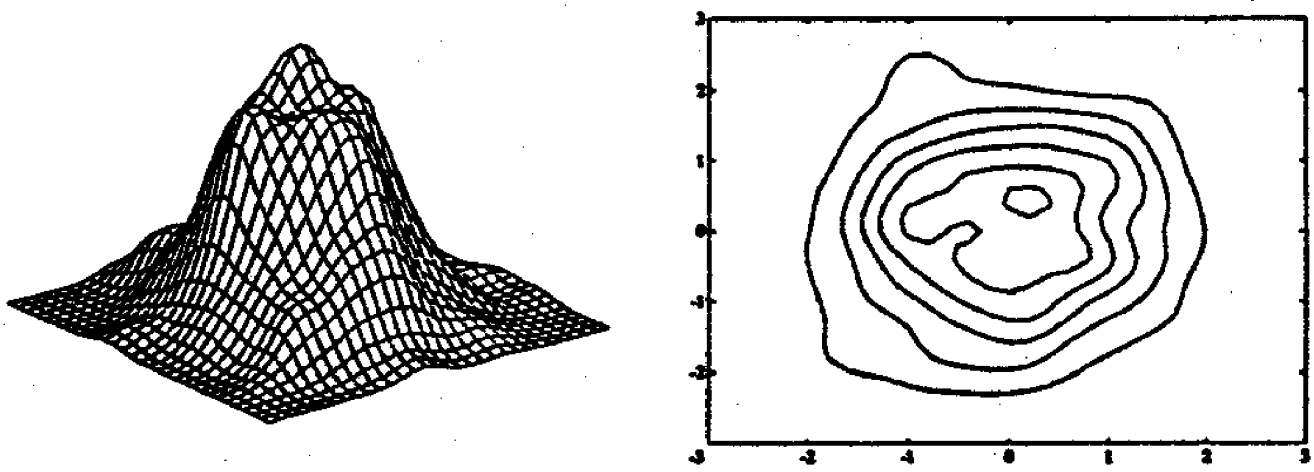
Optimální posloupnost h_n^o minimalizující $U(h_n)$ je tvaru

$$h_n^o = \Theta(f) n^{-1/(2s+p)}, \tag{6}$$

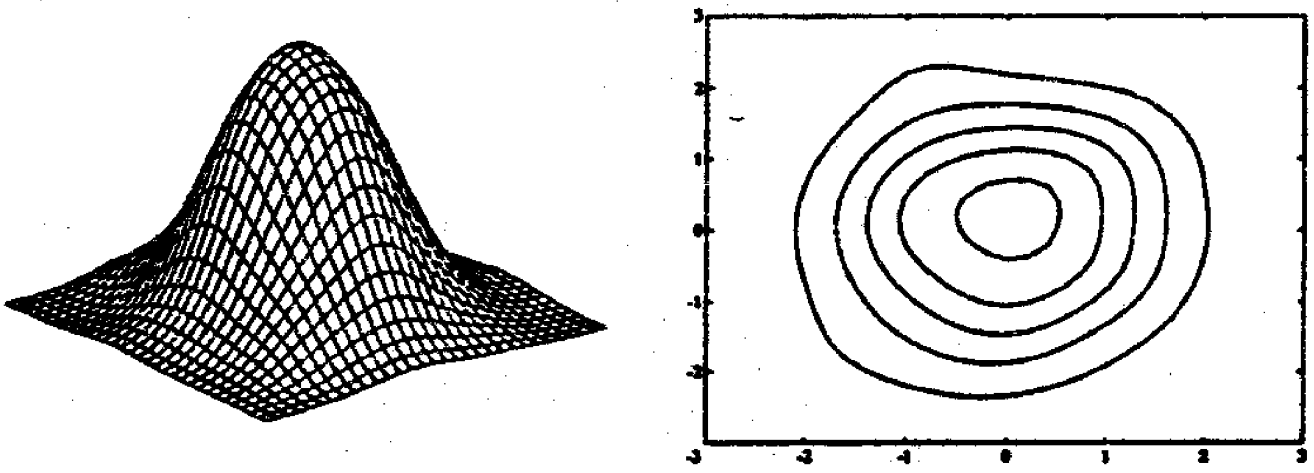
kde Θ je funkcionál hustoty f.

Zbylá část příspěvku bude věnována praktickým výpočtům. První dva obrázky demonstrují odhad hustoty dvojrozměrného náhodného vektoru s normálním rozdělením o nulové střední hodnotě a jednotkové varianční matici. Náhodný výběr v rozsahu 500 prvků byl vygenerován generátorem systému MATLAB, v tomtéž systému byl zpracován také celý výpočet. Šířka okna byla stejná v obou směrech a byla volena postupně zkusmo. První obrázek ukazuje odhad pro $h = 0.3$, druhý pro $h = 0.5$. V obou případech bylo použito gaussovské jádro. K průběhu odhadu je připojen obrázek ukazující vrstevnicové řezy výslednými funkcemi.

Odhady hustoty normálního rozdělení.



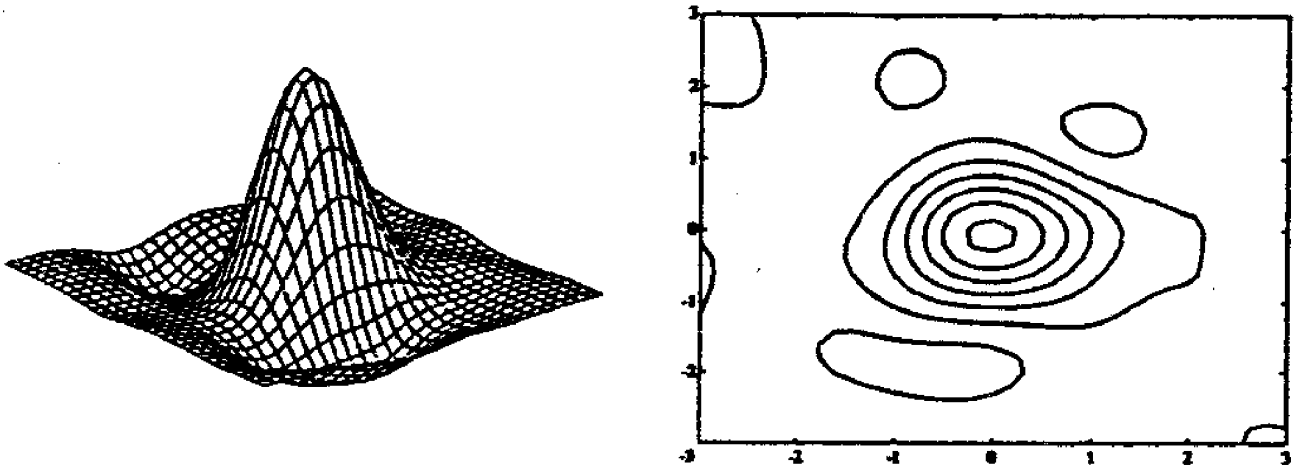
Obr.1 - odhad hustoty pro šířku okna 0.3.



Obr.2 - odhad hustoty pro šířku okna 0.5.

Pro názornou představu o rozdílu mezi skutečnou hustotou a jejím odhadem pro $h = 0.5$ je připojen následující obrázek, který zobrazuje rozdíl těchto dvou funkcí. Maximální rozdíl byl v počátku a činil přibližně 0.053, přičemž hodnota

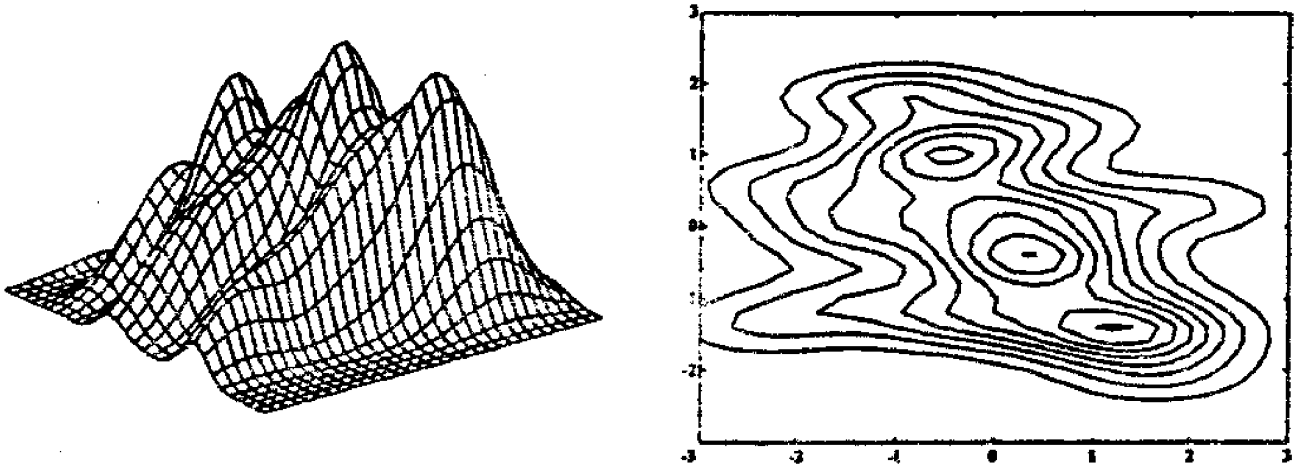
hustoty v počátku je asi 0.159.



Obr.3.

Protože jsme předem znali hustotu rozdělení, bylo možné porovnat číselné výsledky s teoretickými. Numericky spočítaný integrál z druhé mocniny rozdílu obou funkcí měl hodnotu asi 0.0043, hodnota $U(h_n)$ z (5) při zanedbání třetího členu vyšla přibližně 0.0031. Řádově si hodnoty odpovídají, abychom dostali přesnější výsledky by bylo potřeba porovnat odhady pro výběry s různým rozsahem.

Poslední obrázek byl pořízen pro reálná data. Jedná se o třicet hodnot průměrných letních teplot a výnosů obilí na jistém území ve Švédsku v letech 1913 až 1942. Příklad je převzat z [2]. V tomto případě bylo třeba použít rozdílnou šířku okna v jednotlivých směrech, protože jednotlivé složky vektoru se řádově lišily. Poněvadž výběrový soubor měl malý rozsah, výsledný odhad hustoty má spíše jen demonstrativní charakter.



Obr.4.

Literatura:

[1] Antoch J., Odhady hustoty, sborník Robust 82

[2] Cramér H., Mathematical Methods of Statistics, Princeton University Press, Princeton, 1946

[3] Izenman A. J., Recent Developments in Nonparametric Density Estimation, J. Amer. Statist. Assoc., 86, 1991, 205-224

[4] Nadaraya E. A., Nonparametric Estimation of Probability Densities and Regression Curves, Dordrecht/Boston/London, Kluwer Academic Publishers, 1989