



MASARYKOVA UNIVERZITA

Přírodovědecká fakulta

David HAMPEL

NĚKTERÉ STATISTICKÉ METODY

HODNOCENÍ

NEUROFYZIOLOGICKÝCH DAT

Dizertační práce

Školitel: doc. RNDr. Petr LÁNSKÝ, CSc.

Brno, 2009

Bibliografická identifikace

Jméno a příjmení autora: David Hampel

Název dizertační práce: Některé statistické metody hodnocení neurofyziologických dat

Název dizertační práce anglicky: Some statistical methods of the evaluation for neurophysiological data

Studijní program: Matematika

Studijní obor: Pravděpodobnost, statistika a matematické modelování

Školitel: doc. RNDr. Petr Lánský, CSc.

Rok obhajoby: 2010

Klíčová slova v češtině: Aktivita samostatného neuronu, Bodové procesy obnovy, Entropie, Intervalově cenzorovaná data, Neurofyziologická data, Odhad posunutí, Refrakterní perioda

Klíčová slova v angličtině: Single neuron activity, Point renewal processes, Entropy, Interval-censored data, Neurophysiological data, Shift estimation, Refractory period

Poděkování

Touto cestou bych chtěl poděkovat svému školiteli doc. RNDr. Petru Lánskému, CSc. za podnětné rady a připomínky, trpělivost, a vůbec za vše co jsem se od něj naučil. Dále bych chtěl poděkovat za vstřícný přístup všem, kteří mě při tvorbě této dizertační práce přímo či nepřímo podpořili.

Abstrakt

Předložená dizertační práce je věnována některým statistickým metodám aplikovatelným v neurofyziologii, a to odhadu posunutí a diferenciální entropie. Práce sestává ze šesti kapitol a dodatku. V první kapitole jsou stručně nastíněny neurofyziologické základy, zejména popis neuronu a jeho aktivity. V kapitole druhé jsou popsány vybrané matematické modely samostatného neuronu. Ve třetí kapitole jsou položeny statistické základy. Jsou definovány pojmy týkající se charakteristik rozdělení, náhodných procesů, parametrických odhadů rozdělení a diferenciální entropie. Podrobněji jsou rozpracovány techniky neparametrického odhadu hustoty.

Kapitola čtvrtá pojednává o odhadu parametru posunutí rozdělení definovaných na polopřímce. Uvedeny jsou jednak neparametrické odhady (minimální hodnota, odhady odvozené z odhadů nosiče hustoty, navržená korekce minimální hodnoty), jednak odhady parametrické. Z nich je větší prostor věnován odhadům s minimálním rizikem, uvedeny jsou maximálně věrohodné a robustní odhady. Všechny uvedené odhady jsou porovnány pomocí simulací i reálných dat.

Pátá kapitola se zabývá odhadem diferenciální entropie. Diskutují se plug-in odhady založené na odhadu hustoty histogramem, jádrovým odhadem a momentovým odhadem. Je navržen histogram optimální vzhledem k odhadu entropie. Dále jsou uvedeny některé přímé odhady diferenciální entropie, je navržen přímý odhad inspirovaný momentovými odhadem hustoty. Všechny odhady jsou opět porovnány pomocí simulovaných dat.

Poslední kapitola se věnuje odhadu diferenciální entropie z intervalově cenzurovaných dat. Jsou aplikovány dva algoritmy a je navržen postup odhadu entropie na základě jejich výstupů. V dodatku je uveden popis programových implementací vytvořených ve výpočetním systému MATLAB.

Základní literatura

- Cover, T. M., Thomas, J. A., 1991. Elements of information theory. John Wiley, New York.
- Lehmann, E. L., Casella, G., 1998. Theory of Point Estimation, 2nd Edition. Springer-Verlag, New York.

Abstract

Submitted thesis is engaged in some statistical methods applicable in neurophysiology – estimation of shift parameter and differential entropy. This thesis consists of the six chapters and the appendix. In the first chapter, there are briefly outlined neurophysiological basics, particularly description of single neuron and its activity. The second chapter consists of characterization of the chosen mathematical models of a single neuron. In the third chapter, there are stated statistical foundations. The terms related to characteristics of random variable distribution, stochastic processes, parametric estimation of a distribution and differential entropy are defined there. Techniques of nonparametric density estimation are elaborated more deeply.

The fourth chapter deals with estimation of the shift parameter of a distributions defined on half-line. Mentioned are nonparametric estimates (minimum value, estimates derived from density support, proposed correction of minimum value) as well as parametric estimates – minimum risk estimates, maximum likelihood estimates and robust estimates. All the stated estimates are compared with the aid of simulated and real data.

Differential entropy estimates are included in the fifth chapter. There are discussed plug-in estimators based on estimation density by histogram, kernel estimate and moment estimate. There is proposed histogram optimal with respect to entropy estimate. Further, included are some direct entropy estimators, outlined is direct entropy estimator inspired by moment density estimate. Comparison of all stated estimates is done, based on simulated data.

The last chapter is turned to estimation of differential entropy in the case of interval-censored data. There are applied two algorithms for this purpose. The procedure based on theirs results is suggested for entropy estimate. In the appendix, there are stated descriptions of implementations created in computational system MATLAB.

Basic references

- Cover, T. M., Thomas, J. A., 1991. Elements of information theory. John Wiley, New York.
- Lehmann, E. L., Casella, G., 1998. Theory of Point Estimation, 2nd Edition. Springer-Verlag, New York.

Obsah

Přehled značení a základních pojmů	10
Úvod	14
1 Neurofyziologické základy	17
1.1 Axon	17
1.2 Synapse	19
1.3 Tělo a dendrity	20
1.4 Neuron emitující ideální pulsy	20
1.5 Spontánní neuronální aktivita	21
2 Vybrané matematické modely samostatného neuronu	23
2.1 Neurofyziologická data	23
2.2 Přehled neuronálních modelů	24
2.3 Real-time neuronální modely	26
2.3.1 Perfect integrate-and-fire model	26
2.3.2 Leaky integrate-and-fire model	28
2.4 Náhodnost v neuronálních modelech	28
2.5 Neuronální modely použité v dalším textu	29
3 Statistické základy	32
3.1 Charakteristiky rozdělení	32
3.2 Náhodné procesy	34
3.3 Parametrické odhady rozdělení	38
3.3.1 Základy teorie odhadu	38
3.3.2 Maximálně věrohodné odhady	39
3.4 Neparametrické odhady hustoty	42

3.4.1	Odhad hustoty histogramem	42
3.4.2	Jádrové odhady hustoty	44
3.4.3	Momentové odhady hustoty	45
3.5	Diferenciální entropie	48
4	Odhad parametru posunutí	52
4.1	Neparametrické metody odhadu posunutí	52
4.1.1	Odhad minimální hodnotou	52
4.1.2	Odhad posunutí jako součást odhadu nosiče	53
4.1.3	Korigované minimum	54
4.2	Parametrické metody odhadu posunutí	54
4.2.1	Intervalové odhady parametru posunutí	54
4.2.2	Robustní odhad	55
4.2.3	Maximálně věrohodné odhady	56
4.2.4	Odhady posunutí s minimálním rizikem	57
4.3	Porovnání jednotlivých metod odhadu posunutí	62
4.3.1	Parametrický přístup	63
4.3.2	Neparametrický přístup	66
4.3.3	Analýza reálných neurofyziologických dat	69
5	Odhady diferenciální entropie	72
5.1	Odhad entropie na základě histogramu	73
5.1.1	Odhad entropie při ekvidistantním dělení histogramu	74
5.1.2	Odhad entropie při neekvidistantním dělení histogramu	77
5.2	Odhad entropie na základě jádrového odhadu hustoty	78
5.3	Odhad entropie na základě momentového odhadu hustoty	82
5.3.1	Momentový odhad konstruovaný z gamma hustoty	84
5.3.2	Výpočet entropie z momentového odhadu hustoty	85
5.4	Přímé odhady entropie	87
5.4.1	Odhad entropie na základě rozestupu vzorků	87
5.4.2	Odhad entropie podle Vasicka	88

5.4.3	Srovnání odhadu založeného na rozestupu vzorků a Vasic- kových odhadů vzhledem k volbě parametru m	89
5.4.4	Odhad entropie na základě vzdálenosti od nejbližšího souseda	90
5.4.5	Přímý odhad entropie na základě momentových charakteristik	92
5.5	Srovnání všech uvedených odhadů entropie	95
6	Odhad entropie z intervalově cenzorovaných dat	100
6.1	Intervalově cenzorovaná data	100
6.2	Metoda založená na Vardiho algoritmu	101
6.3	Metoda založená na algoritmu McCleanové a Devineho	103
6.4	Odhad hustoty a entropie na základě výstupů předchozích metod . .	105
6.5	Srovnání uvedených metod	106
	Závěr	116
	Literatura	118
	A Popis programových implementací	123

Přehled značení a základních pojmů

Přehled značení

- a ... parametr posunutí
- CV ... koeficient variace
- Δ ... šířka datového okna
- $\gamma_1, \hat{\gamma}_1$... šikmost, resp. její odhad
- $\gamma_2, \hat{\gamma}_2$... špičatost, resp. její odhad
- $\Gamma(b)$... gamma funkce, $\Gamma(b) = \int_0^{\infty} x^{b-1} e^{-x} dx, b > 0$
- IQR ... mezikvartilové rozpětí, $\text{IQR} = t_{0,75} - t_{0,25}$
- n ... rozsah výběru
- \mathbb{N}_0 ... množina přirozených čísel a nuly
- nosič hustoty $f(t)$... množina $\{t | f(t) > 0\}$
- $o(h^k)$... třída konvergence, $\frac{o(h^k)}{h^k} \rightarrow 0$ pro $h \rightarrow 0$.
- Ω ... základní prostor
- $\Psi(b)$... digamma funkce, $\Psi(b) = \frac{\partial}{\partial b} \ln \Gamma(b)$
- \mathbb{R} ... množina reálných čísel
- S_t resp. s_t ... výběrová směrodatná odchylka, resp. její realizace

- $\sigma, \hat{\sigma}$... odchylka, resp. její odhad
- $\sigma^2, \hat{\sigma}^2$... rozptyl, resp. jeho odhad
- $[t]$... celá část čísla t
- (T_1, \dots, T_n) resp. (t_1, \dots, t_n) ... náhodný výběr rozsahu n , resp. jeho realizace
- $T_{(1)} < \dots < T_{(n)}$ resp. $t_{(1)} < \dots < t_{(n)}$... pořadová statistika uspořádaná na základě náhodného výběru rozsahu n , resp. její realizace
- t_α ... α -kvantil
- Θ ... parametrický prostor
- $\theta, \boldsymbol{\theta}$... skalární resp. vektorový parametr
- \mathbb{Z} ... množina celých čísel

Základní pojmy a vzorce

- Indexní funkce $I_{[a,b)}(t)$ je definována jako

$$I_{[a,b)}(t) = \begin{cases} 1 & \text{pro } t \in [a, b), \\ 0 & \text{pro } t \notin [a, b). \end{cases}$$

- Charakteristickou funkci $\psi(x)$ náhodné veličiny T definujeme pro $\forall x \in \mathbb{R}$ vzorcem

$$\psi(x) = E(e^{ixT}).$$

- Symbolem \mathcal{D}^j budeme značit diferenciální operátor

$$\mathcal{D}^j f(t) = \frac{\partial^j f(t)}{\partial t^j}.$$

- Mějme posloupnost náhodných veličin T_1, T_2, \dots a náhodnou veličinu T . Potom řekneme, že T_n konverguje k T

– skoro jistě, jestliže

$$P\{\omega : T_n(\omega) \rightarrow T(\omega)\} = 1;$$

– podle pravděpodobnosti, jestliže pro každé $\epsilon > 0$ platí

$$P\{\omega : |T_n(\omega) - T(\omega)| > \epsilon\} \rightarrow 0;$$

– podle středu, když při $ET_n^2 < \infty$ pro $n = 1, 2, \dots$ platí

$$E(T_n - T)^2 \rightarrow 0;$$

– v distribuci, když T_n má distribuční funkci $F_n(t)$, T má distribuční funkci $F(t)$ a $F_n(t)$ konverguje k $F(t)$ v každém bodě t , v němž je $F(t)$ spojitá.

Rozdělení použitá v dalším textu

- Exponenciální rozdělení $E(\lambda; a)$ s parametry $\lambda > 0$ a a dané hustotou

$$f(t; \lambda, a) = \lambda e^{-\lambda(t-a)}, \quad t \geq a. \quad (1)$$

- Gamma rozdělení $G(\lambda; b; a)$ s parametry $\lambda > 0$, $b > 0$ a a dané hustotou

$$f(t; \lambda, b, a) = \frac{\lambda(\lambda(t-a))^{b-1}}{\Gamma(b)} e^{-\lambda(t-a)}, \quad t \geq a. \quad (2)$$

- Alfa rozdělení $A(\lambda; a)$ s parametry $\lambda > 0$ a a dané hustotou

$$f(t; \lambda; a) = \lambda^2(t-a)e^{-\lambda(t-a)}, \quad t \geq a. \quad (3)$$

- Normální rozdělení $N(\mu; \sigma^2)$ s parametry μ a $\sigma > 0$ dané hustotou

$$f(t; \mu; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}, \quad t \in (-\infty, \infty). \quad (4)$$

- Weibullovo rozdělení $W(a; b; c)$ s parametry $b > 0$, $c > 0$ a a dané hustotou

$$f(t; a; b; c) = cb(t-a)^{b-1} e^{-c(t-a)^b}, \quad t \geq a. \quad (5)$$

- Logaritmicko normální rozdělení $LN(\mu; \sigma^2; a)$ s parametry μ , $\sigma > 0$ a a dané

hustotou

$$f(t; \mu; \sigma; a) = \frac{1}{(t-a)\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(t-a)-\mu}{\sigma}\right)^2}, \quad t \geq a. \quad (6)$$

- Inverzní Gaussovo rozdělení $IG(\mu; \lambda; a)$ s parametry $\mu > 0$, $\lambda > 0$ a a dané hustotou

$$f(t; \mu; \lambda; a) = \sqrt{\frac{\lambda}{2\pi(t-a)^3}} e^{-\lambda\frac{((t-a)-\mu)^2}{2\mu^2(t-a)}}, \quad t \geq a. \quad (7)$$

Úvod

V současnosti se v neurofyziologických vědách vynakládá mimořádné úsilí ve výzkumu v oblasti přenosu informace mezi neurony. Většina přístupů k této otázce je do značné míry založena na přesné identifikaci času generování neuronálního pulsu. Proto je důležité charakterizovat posloupnosti akčních potenciálů formálním popisem, který by ovšem odpovídal experimentálním poznatkům. Mezi dobře známé a ověřené vlastnosti neuronů patří i refrakternost. Nyní uvedeme několik příkladů ilustrujících jak důležité může být využití refrakterní periody.

Při záznamu neuronální aktivity jednoduchou extracelulární elektrodou je obvyklé zaznamenávat akční potenciály více neuronů. Při analýze takového záznamu vyvstává otázka identifikace aktivního neuronu. Pokud budeme „chybně“ předpokládat, že v datech není obsažena žádná refrakterní perioda, nezískáme analýzou společného záznamu žádné či jen velmi vágní informace o jednotlivých akčních potenciálech (viz Cox (1962), strana 79). V práci Meunier et al. (2003) je vyvinuta metoda pro odhad frekvencí akčních potenciálů dvou neuronů emitujících nerozlišitelné akční potenciály. Tato metoda je založena na existenci tzv. „tiché periody“ v posloupnosti akčních potenciálů, refrakterní periodu lze též v tomto smyslu použít.

V článku Lehky (2004) je analyzován Poissonův proces. Pro odhad reakce neuronu na stimul je zde vypracována metoda postavená na bayesovském základě. Bayesovský odhad rozptylu parametru symbolizujícího rozptýlení je mnohem nižší než obvyklé odhady, což umožňuje vytvářet pomocí diskutované bayesovské metody přesnější intervalové odhady odpovědí neuronu na stimul. Existence refrakterní periody umožňuje redukovat rozptyl počtu akčních potenciálů a následně rozptyl frekvencí pozorovaných akčních potenciálů. Korekční faktor založený na refrakterní periodě lze též použít k redukci rozptylu věrohodnostní funkce a následně ke snížení vychýlení odhadů charakteristik spojených s akčními potenciály.

Kullback-Leiblerova vzdálenost, kterou se okrajově zabýváme i v této práci, je důležitým nástrojem pro výzkum neuronálního kódování informace (viz Kostal a

Lansky (2006); Kostal a Lansky (2007)). Existence refrakterní periody má významný teoretický dopad na určení Kullback-Leiblerovy vzdálenosti. Navíc, tato vzdálenost je často odhadována pomocí diferenciálních entropií jednotlivých hustot. Kvalitní odhad refrakterní periody pomůže vylepšit odhady diferenciální entropie, a to zejména v případě použití tzv. plug-in odhadů.

V práci Berry II a Meister (1998) je modelována odezva na světlo u sítnicového ganglionu pomocí náhodných akčních potenciálů kombinovaných s refrakterní periodou. Dle jejich teorie je momentální frekvence akčních potenciálů důsledkem „volné frekvence akčních potenciálů“, která závisí pouze na stimulu, a „obnovovací funkci“, která závisí pouze na času od posledního akčního potenciálu. Tato obnovovací funkce je nulová pro absolutní refrakterní periodu a poté postupně roste až do 1. Autoři ukazují, že refrakterní vlastnost, ačkoliv často považována za limitující faktor výkonnosti neuronu, může ve skutečnosti přispívat k neuronální stabilitě. Obecně řečeno, správné určení refrakterní periody podstatně ovlivní správnost posuzování frekvence akčních potenciálů (viz Barbieri et al. (2001) a Brown et al. (2002)). Tento problém je také často předmětem zájmu v rámci aplikace neuronálních modelů, což je diskutováno například v pracích Albano et al. (2007), Buonocore et al. (2002), Buonocore et al. (2003), Ricciardi a Esposito (1966) a Teich et al. (1978).

Metoda použitá v práci Guillory et al. (2006) využívá pro modelování posloupnosti akčních potenciálů Poissonův model s proměnlivým časem společně s rozšířením o refrakterní periodu. Na rozdíl od jiných přístupů používaných pro Poissonův proces je formulace prezentovaná v jejich práci nezávislá na definici dělicích intervalů a vyžaduje zadání několika uživatelem specifikovaných parametrů. Pro metodu odhadu jsou diskutovány také problémy týkající se statistické závislosti a efekty stacionarity.

V práci Fiore et al. (2005) je navržena procedura pro srovnání aktivity jednoho neuronu stimulovaného v opakujících se cyklech s nehomogenním Poissonovým procesem. Hustota pravděpodobnosti mezipulsních intervalů predikovaná pro tento proces je vypočtena na základě experimentální hustoty akčních potenciálů a jsou vyhodnoceny odlišnosti od rozdělení experimentálních intervalů. Vyvinutý postup byl aplikován na posloupnosti akčních potenciálů, které byly indukovány otevíráním a zavíráním distálního článku nohy humra. Dle předpokladů ležela hustota krátkých mezipulsních intervalů (o délce méně než 20–40 ms) těsně pod úrovní predikovanou pro nehomogenní Poissonův proces, výskyt refrakterní periody byl zahrnut.

Ve shodě s výsledky dalších prací data použitá v práci Schaeffer et al. (2005) naznačují, že neuronální refrakterní vlastnost silně ovlivňuje variabilitu posloupnosti

akčních potenciálů. Autoři zkoumají stochastický model generování akčních potenciálů, který zahrnuje vliv refrakterní periody prostřednictvím obnovovací funkce. Vzhledem k tomu, že data použitá v jejich práci jsou ve shodě s definicí procesu obnovy, může být obnovovací funkce odvozena z histogramu mezispulsních intervalů získaných při konstantní stimulaci. Výsledný popis poskytuje kvantitativně přesné predikce variability odpovědi na celém rozsahu frekvencí akčních potenciálů, pro konstantní stejně jako pro amplitudově modulovaný zvukový stimul. Modelové parametry získané pro konstantní stimulování mohou být využity pro predikci variability odpovědi na dynamický stimul.

Z výše uvedených příkladů plyne, že refrakterní perioda podstatně ovlivňuje kódování informace přenášené posloupnostmi akčních potenciálů. Z experimentálního hlediska je tedy odhad její délky důležitým úkolem.

Pomocí konceptu diferenciální entropie můžeme identifikovat změny v chování neuronu (podobně jako pomocí frekvenčních charakteristik) a srovnat chování dvou či více neuronů za různých experimentálních situací. Pro tento účel je třeba mít k dispozici kvalitní odhady diferenciální entropie. Z definice diferenciální entropie plyne, že otázka kvality odhadu entropie je spjata s problémem identifikace hustoty realizací náhodných veličin. Odhad diferenciální entropie lze použít k výpočtu Kullback-Leiblerovy vzdálenosti, pomocí níž lze určit případné změny v chování neuronu či rozdíly mezi neurony.

V praxi je potřebné odhadnout diferenciální entropii v relativně krátkém čase a navíc na základě malého rozsahu dat. K tomuto účelu by mohly posloužit rychlé plug-in odhady založené na momentovém odhadu hustoty. Tyto odhady hustoty trpí několika významnými nedostatky. Jejich nasazení je třeba prověřit.

Experimentální situace při záznamu dat je komplikována faktem, že je nemožné pořídit časově neomezený záznam neuronální aktivity. Konkrétně, pokud zaznamenáváme neuronální aktivitu po nějakém stimulu, je reakce neuronu pouze dočasná. Abychom získali dostatečné množství dat, musíme tento experiment několikrát opakovat. Další podobná situace nastane, pokud sledujeme větší množství neuronů najednou a chceme identifikovat možnou neuronální aktivitu ihned po stimulu.

V obou případech dostaneme větší množství relativně krátkých intervalově cenzorovaných záznamů neuronální aktivity namísto jednoho dlouhého záznamu jak je obvyklé. Z těchto vzorků chceme odhadnout diferenciální entropii (a poté případně spočítat Kullback-Leiblerovu vzdálenost od rozdělení neuronální aktivity bez stimulu).

Kapitola 1

Neurofyziologické základy

V této kapitole stručně shrneme neurofyziologické základy fungování neuronu. Informace zde uvedené pochází především z knih Sampath a Srinivasan (1977) a Gerstner a Kistler (2002). Detailnější informace lze získat například v knihách Ochs (1965) a Katz (1966).

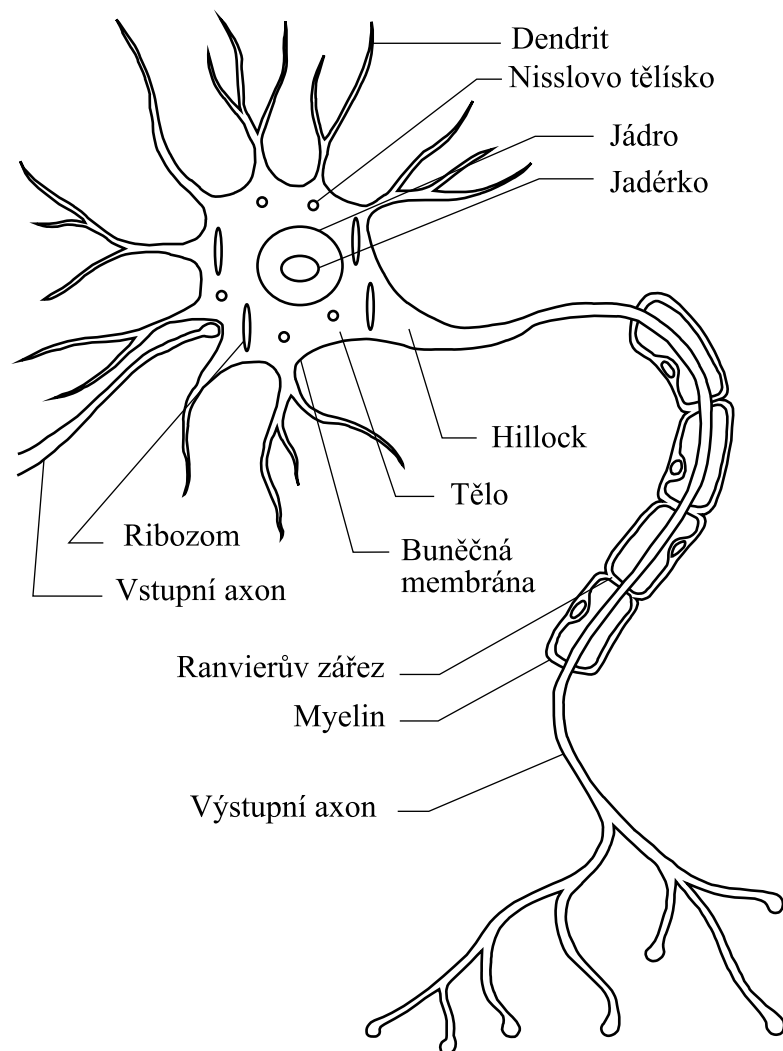
Existuje mnoho typů neuronů s širokým rozpětím strukturálních a funkčních odlišností. Zhruba je můžeme rozdělit na

1. senzorické neurony,
2. interneurony,
3. motoneurony a
4. neurony v centrálním nervovém systému.

Nicméně všechny neurony se skládají z následujících čtyř částí: těla, dendritů, jedné či více synapsí a axonu, který může být pokryt myelinovým pouzdrém. Schéma obecného neuronu je uvedeno na obr. 1.1 a 1.2.

1.1 Axon

Se značným zjednodušením můžeme axon považovat za cylindrickou polopropustnou membránu obsahující axoplasmu a obklopenou extracelulární tekutinou. Vnitřní i vnější kapaliny obsahují soli, zejména chlorid draselný a chlorid sodný, jež jsou ionizovány v roztoku. Koncentrace draselných ionů uvnitř je mnohem vyšší než vně. Toto vede k přesunu přebytečných draselných ionů ven. Pomocí této ionové difuze vznikne



Obr. 1.1: Schéma obecného neuronu.

elektrické pole, které je v protíváze „chemickému poli“. Rovnováhy je dosaženo, když se tyto síly vyrovnají. Potom řekneme, že membrána je „polarizována“ v klidovém stavu.

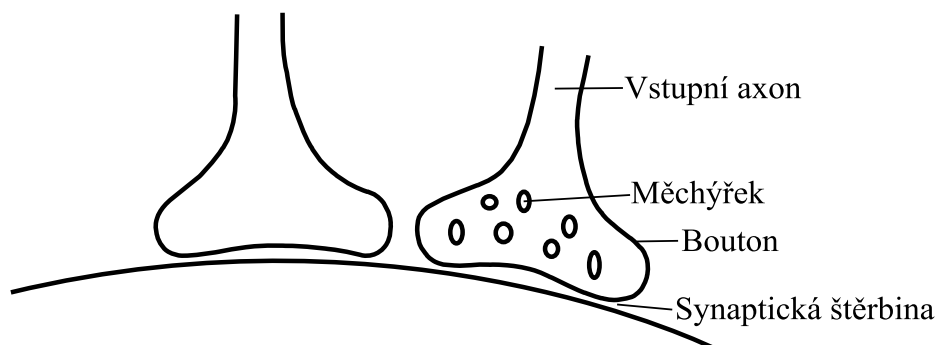
Rovnováha může být narušena buď změnou extracelulární koncentrace ionů anebo vnějším působením elektrického potenciálu směřujícího přes membránu. Poznamenejme, že rozdíl potenciálů dvou bodů se rovná elektrickému napětí mezi těmito dvěma body. Pokud je výsledkem vyšší kladný membránový potenciál, mluvíme o

depolarizaci. Když je tato depolarizace dostatečně velká (pokud dosáhne určitého prahu), je přes stimulovanou část emitován puls¹ a potenciál membrány se vrátí do klidové úrovně. Předpokládáme tedy, že axon může být pouze ve dvou stavech: excitovaném a neexcitovaném.

Posloupnost emitovaných pulsů se obvykle nazývá posloupností akčních potenciálů. Krátkou dobu po emitování akčního potenciálu nemůže být oblast stimulace znovu excitována. Tato vlastnost se nazývá refrakternost². Bezprostředně po generování akčního potenciálu nemůže být oblast stimulace excitována jakkoliv silným stimulem. Zde mluvíme o absolutní refrakterní periodě. Následuje tzv. relativní refrakterní perioda, kdy je excitace možná pouze na základě velmi silného stimulu.

1.2 Synapse

Interakce neuronů se uskutečňuje v synaptické oblasti (viz obr. 1.2) přenosem aktivity jednoho neuronu na další neurony axonem. Konce axonu rozšířené do vy-



Obr. 1.2: Schéma synapse.

bouleniny zvané bouton přiléhají k buněčné membráně nebo k dendritu. Bouton nemá s membránou fyzický kontakt, mezi nimi je štěrbina. V boutonu se nacházejí malé zásobníky nazývané „vesicles“. Tyto obsahují tzv. „přenašeč“, který hraje roli zprostředkovatele neuronální aktivity neuronu mezi axonem a tělem buňky. V zásadě je různými způsoby modifikována prostupnost membrány různým ionům.

¹V tomto smyslu lze též použít neurofyziologické termíny „akční potenciál“ či „impuls“ a statistický termín „událost“.

²Překlad anglického refractoriness. Používá se též výraz refrakčnost, refrakční.

Pokud je výsledná změna potenciálu kladná, mluvíme o excitačním postsynaptickém potenciálu, a pro zápornou změnu potenciálu o inhibičním postsynaptickém potenciálu. Za nepřítomnosti jakéhokoli vstupu směřuje membránový potenciál ke klidové úrovni.

Jsou známy i synapse, které neobsahují chemický přenašeč, nýbrž zde působí přímý elektrický účinek. Nazývají se „gap junctions“ a nacházejí se obvykle podél chemických synapsí. Specializované membránové proteiny zajišťují přímé elektrické spojení mezi dvěma neurony. Předpokládá se, že „gap junctions“ jsou zapojeny do synchronizace neuronů.

1.3 Tělo a dendrity

Jaderná a cytoplazmatická struktura i metabolické funkce neuronu se neliší od jiných buněk. Jedinou výjimkou je, že neurony se obvykle nedělí. Výstupy sousedících neuronů, axony, se připojují na tělo nebo na dendrity. Postsynaptické potenciály jsou indukovány emitovanými akčními potenciály na několika místech membrány. V těle neuronu se načítají. Tyto jevy se označují jako prostorové a časové sečítání. Pokud načítaný výsledek přesáhne práh excitace membrány, neuron emituje prostřednictvím axonu puls. Část axonu připojená k tělu, hillock, má nižší práh než ostatní části membrány, a proto je oblastí generující puls.

Dendrity představují rozšířenou plochu pro připojení příchozích axonů, které se větví pro vytvoření mnoha synapsí. V dendritech může též vzniknout akční potenciál. Úlohy a chování dendritů nejsou dosud zcela pochopeny.

1.4 Neuron emitující ideální pulsy

Předcházející text můžeme zhruba shrnout následujícím způsobem: Dendrity hrají roli „vstupního zařízení“, shromažďují signály přicházející od okolních neuronů a přenáší je do těla neuronu. Tělo můžeme považovat za „vyhodnocovací jednotku“ zajišťující generování výstupního signálu v případě, že celkový vstup dosáhne jistého prahu. Tento výstupní signál je převzat axonem, „výstupním zařízením“, který doručí signál dalším neuronům.

Signálem v nervovém systému rozumíme postupující akční potenciál. Zpracování signálu v takto složité síti je komplexní záležitostí a nemůže být popsáno jedno-

duchými prostředky. Nicméně informace je vždy kódována pomocí posloupností akčních potenciálů. Přesněji řečeno jsou „neuronální zprávy“ kódovány do posloupností pulsů. Oprávněně je možné předpokládat, že tato zpráva je obsažena ve frekvenci či v časovém uspořádání pulsů, nikoliv ve tvaru akčních potenciálů.

V tomto kontextu se spíše než označení akční potenciál užívá výrazu puls³. Pulsy mají amplitudu kolem 100 mV a obvykle trvají 1–2 ms. Posloupnost pulsů emitovanou samostatným neuronem⁴ budeme považovat za řadu nerozlišitelných událostí, které nastávají v pravidelných nebo nepravidelných intervalech. Pulsy jsou v této posloupnosti obvykle dostatečně separovány (díky refrakterní vlastnosti) a je tedy možné je zaznamenávat pomocí jemné mikroelektrody umístěné poblíž axonu či těla neuronu (pro ilustraci viz obr. 1.3).



Obr. 1.3: Záznam neuronální aktivity pomocí mikroelektrody.

1.5 Spontánní neuronální aktivita

V polovině minulého století byly poprvé pozorovány malé náhodné depolarizační potenciály za absence jakéhokoliv stimulu. Nazýváme je „miniaturní koncové potenciály“ a vysvětlujeme je náhodným uvolňováním zásobníků acetylcholinu ze synaptických „vesicles“. Tento fakt znamenal konec dlouhou dobu udržovaného přesvědčení, že se periferní neuronální systémy chovají deterministicky. Neurony tedy emitují pulsy i spontánně, bez externího stimulu. Proto za identických podmínek pro realizace můžeme říci, že posloupnosti akčních potenciálů mají charakter sto-

³V anglických textech se používá výraz „spike“, který však není možné překládat přímo.

⁴Anglicky „spike train“.

chastických bodových procesů (náhodně rozdělených posloupností nerozlišitelných událostí).

Proces generování pulsů je navíc komplikován tím, že neurony jsou vzájemně propojeny, což činí teoretický rozbor velmi náročným. Nicméně za jistých předpokladů je možno modely aktivity samostatného neuronu konstruovat. V kapitole 2 jsou některé neuronální modely diskutovány.

Kapitola 2

Vybrané matematické modely samostatného neuronu

V této kapitole uvedeme několik matematických modelů samostatného neuronu. Výjimečně se zde budeme odkazovat na definice uvedené v kapitole Statistické základy. Čerpáme zde především z knih Gerstner a Kistler (2002), Holden (1976) a Sampath a Srinivasan (1977).

2.1 Neurofyziologická data

Pozorovaná posloupnost akčních potenciálů může být popsána mnoha způsoby. Na jedné straně může být celá posloupnost popsána výčtem délek časových intervalů mezi jednotlivými pulsy. Na straně druhé ji můžeme popsat jediným číslem, průměrnou frekvencí pulsů, což ovšem představuje značnou kondenzaci dat. Obvykle se pro porovnávání dat s určitým modelem používají statistické charakteristiky patřící mezi tyto dva extrémní případy. Jedním z takových popisů je histogram časových intervalů mezi pulsy.

Obecně se má za to, že posloupnosti intervalů mezi spontánními akčními potenciály jsou stacionární ve smyslu definice 3.6, třebaže jsou známy výjimky. Pokud bychom připustili, že všechny intervaly mezi pulsy bodového procesu mají stejné rozdělení a jsou navzájem nezávislé, dostáváme se k procesu obnovy (viz definice 3.10). V tomto případě je proces generování akčních potenciálů zcela popsán hustotou pravděpodobnosti mezipulsních intervalů. Myšlenka, že posloupnost akčních potenciálů je procesem obnovy, se jeví jako velmi zjednodušující. Nicméně

statistické testy ukazují, že posloupnosti akčních potenciálů pozorované na dlouhém časovém úseku tuto vlastnost často mají.

Řada autorů, např. Fetz a Gerstein (1963), Levick a Williams (1964), Poggio a Vinerstein (1964) a Skvaril et al. (1971) používá při modelování posloupnosti intervalů mezi akčními potenciály nejrůznějších hustot pravděpodobnosti: často bývá použita hustota exponenciálního rozdělení (1) s $a = 0$ i s $a > 0$, normálního rozdělení (4) či gamma rozdělení (2) (popř. jeho speciálního případu $b = 2$, tj. alfa rozdělení (3)); používají se jak rozdělení nesymetrická tak symetrická, unimodální i multimodální; uplatňují se tzv. rozdělení s těžkým chvostem. Rozmanitost používaných hustot ukazuje na rozdílnost skupin neuronů. Komplexní neuronální model musí generovat mnoho různých rozdělení mezipulsních intervalů a současně obsahovat co nejvíce vlastností reálného neuronu.

2.2 Přehled neuronálních modelů

Na začátku rozboru jednotlivých neuronálních modelů se musíme zmínit o tom, že procesy obnovy představují pouze jednu část aktivity neuronu. Neuron, který v průběhu konkrétní periody produkuje pulsy ve shodě s procesem obnovy se může „přepnout“ do stavu aktivity, kdy produkuje vysoce korelovaný či periodický výstup. Pro modelování takového výstupu musí být použity jiné přístupy než proces obnovy. Nicméně obecný popis procesu jako bodového procesu a jeho charakterizace pomocí hustot bývá užitečná i k teoretickému vyjádření chování neuronu, jehož výstup nesplňuje podmínky procesu obnovy.

Neuronální modely se v literatuře klasifikují mnoha způsoby (viz Sampath a Srinivasan (1977)). Pokud o nich budeme uvažovat jako o matematicky definovaném procesu, můžeme je rozdělit na

- superpoziční modely,
- interaktivní modely,
- difuzní modely,
- načítací modely,
- modely diskrétní náhodné procházky a
- modely spojité náhodné procházky.

Realističtější klasifikaci je možné založit na daném fyziologickém mechanismu. Zde bude neuron vystupovat jako zařízení s jistým prahem, které přijímá sledy excitačních a inhibičních impulsů. Tyto sledy se mohou ovlivňovat různými způsoby:

1. a) presynapticky, s potlačeným inhibičním efektem a
b) postsynapticky, s excitačním impulsem zvyšujícím membránový potenciál a s inhibičním impulsem působícím na snížení membránového potenciálu či způsobujícím návrat na klidovou úroveň potenciálu;
2. a) nezávisle na sobě, tj. určitá posloupnost akčních potenciálů nezávisí na ostatních, nebo
b) jedna posloupnost je řízena jinými.

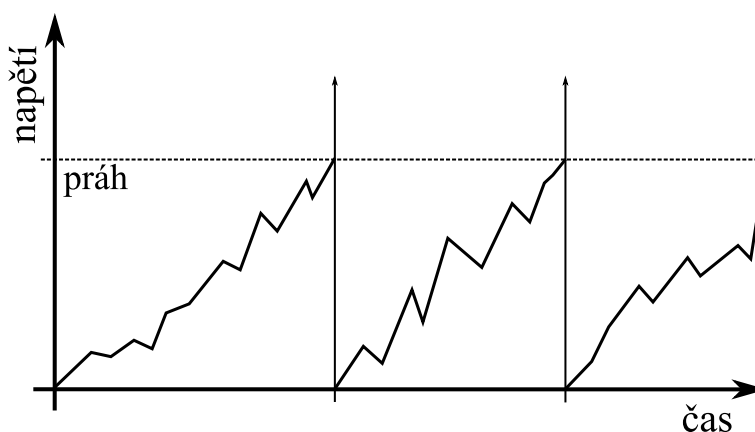
Kromě vlastností jednotlivých neuronálních modelů je důležité vymezit charakteristiku hypotetického neuronu. V následujícím výčtu jsou sumarizovány některé vlastnosti takového neuronu.

1. Předpokládáme, že neuron je izolován od ostatních neuronů, propojení neuronů do sítě je ignorováno.
2. V zásadě máme dvě vstupní posloupnosti: „excitační“ a „inhibiční“. Tato označení se vztahují k efektu, který má puls v synaptické oblasti.
3. Posloupnosti akčních potenciálů splňují podmínky kladené na procesy obnovy.
4. Změna membránového potenciálu zapříčiněná příchozím pulsem je okamžitá.
5. Dendrity a jejich efekty se neuvažují.
6. Prostorová sumace je ignorována, nicméně její efekt je obvykle zachycen prostřednictvím superpozice vstupních posloupností.

V průběhu zkoumání neuronální aktivity bylo vyvinuto velké množství modelů neuronu, viz například publikace Gerstner a Kistler (2002), Holden (1976) a Sampath a Srinivasan (1977). V další podkapitole uvedeme relativně jednoduché, avšak prakticky použitelné modely třídy real-time.

2.3 Real-time neuronální modely

Princip nejjednoduššího real-time neuronálního modelu je následující: V neuronu je načítán potenciál (který může v průběhu načítání i poklesnout), a když je dosaženo prahu θ , tak je emitován puls. Poté je potenciál okamžitě vybit a znova začíná načítání. Schéma tohoto modelu je zobrazeno na obr. 2.1. Tento real-time model



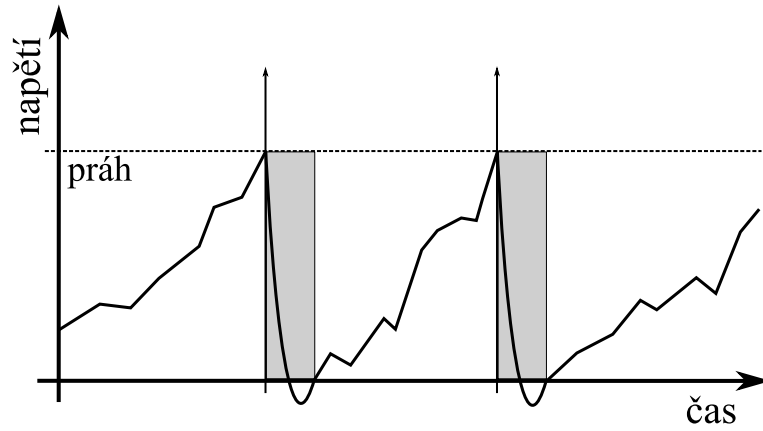
Obr. 2.1: Schéma real-time neuronálního modelu.

ale není realistický, neboť předpokládá, že potenciál je vybit bezprostředně po emitování pulsu. Reálně je pro vybití potenciálu potřeba relativně krátký časový okamžik (v souladu s refrakterní vlastností). Na obr. 2.2 je do real-time modelu začleněna refrakterní perioda. Následující dvě podkapitoly popisují konkrétní neuronální modely ze třídy real-time modelů.

2.3.1 Perfect integrate-and-fire model

Perfect integrate-and-fire model načítá vstupy bez jakéhokoliv poklesu až do doby, kdy je dosaženo prahového napětí θ . Bezprostředně poté je generován akční potenciál a membránový potenciál je resetován do klidové hodnoty. Pokud je $x(\tau)$ vstup daný jako funkce času, a předpokládáme, že předchozí puls byl emitován v čase 0, tak je podprahový generátor potenciálu dán jako

$$V(\tau) = \int_0^{\tau} x(u) du.$$



Obr. 2.2: Schéma real-time neuronálního modelu s refrakterní periodou. Šedá pole vymezují trvání tohoto časového intervalu.

Pokud byl poslední akční potenciál emitován v čase τ_{i-1} , následující akční potenciál bude emitován v čase τ_i takovém, že

$$\tau_i = \min \left\{ \tau \mid \tau > \tau_{i-1}; \quad \theta = \int_{\tau_{i-1}}^{\tau} x(u) du; \quad \lim_{\tau_0 \rightarrow \tau^+} V(\tau_0) = 0 \right\}.$$

Pokud je vstupem konstantní funkce $x(\tau) = c$, frekvence generování akčních potenciálů je $r = \frac{1}{\tau_i - \tau_{i-1}} = \frac{c}{\theta}$ a je tedy úměrná ke vstupu. Perfect integrator model s konstantním prahem je možné využít jako jednoduchý analytický model pro generování náhodných posloupností akčních potenciálů, kdy jako vstupní funkci vezmeme realizace náhodné veličiny.

Vlastnosti tohoto modelu závisí na typu vstupu. Pokud je vstupní funkce sérií kladných pulsů, $V(\tau)$ poroste ve formě schodovité funkce dokud nebude dosažen práh θ . Pokud je vstupní funkce Poissonův proces o intenzitě λ (viz příklad 3.8), má výstup gamma rozdělení řádu b , jež se rovná celé části čísla $\frac{\theta}{A}$. Symbolem A zde značíme amplitudu změny v napětí $V(\tau)$ zapříčiněné jedním vstupním pulsem.

Pokud je vstup $x(\tau)$ modelu perfect integrate-and-fire posloupností kladných a záporných pulsů, výslednou trajektorií je náhodná procházka. Pokud je vstupem bílý šum, nebo pokud frekvence kladných a záporných pulsů konverguje k nekonečnu, když jejich amplituda $A \rightarrow 0_+$, je generován Wienerův proces.

2.3.2 Leaky integrate-and-fire model

V tzv. leaky integrate-and-fire modelu (nebo také Lapiqueově či leaky integrator modelu) se v podprahovém režimu vstupy lineárně načítají a potenciál exponenciálně klesá na základě časové konstanty ϕ . Pokud napětí dosáhne prahu θ je generován akční potenciál a napětí je okamžitě resetováno na klidovou hodnotu, která může být transformována na 0. Model můžeme popsat rovnicí

$$V(\tau) = \int_0^{\tau} x(u) e^{-\frac{(\tau-u)}{\phi}} du \text{ pro } 0 < V(\tau) < \theta, \quad (2.1)$$

kde předpokládáme emitování předchozího pulsu v čase 0. Čas generování pulsu τ_i za předpokladu generování předcházejícího pulsu v čase τ_{i-1} můžeme zjistit z rovnice

$$\tau_i = \min \left\{ \tau \mid \tau > \tau_{i-1}; \quad \theta = \int_{\tau_{i-1}}^{\tau} x(u) e^{-\frac{(\tau_i-u)}{\phi}} du; \quad \lim_{\tau_0 \rightarrow \tau^+} V(\tau_0) = 0 \right\}. \quad (2.2)$$

Časy generování pulsů, které jsou odpovědí na konstantní vstup $x(u) = c$ jsou dány jako řešení rovnice (2.2), tedy

$$\tau_i = \tau_{i-1} - \phi \ln \left(1 - \frac{\theta}{c\phi} \right), \quad (2.3)$$

kde $c\phi > \theta$. Refrakterní perioda, přestože není přímo zabudována v modelu, může být jednoduše přičtena k délce $t_i = (\tau_i - \tau_{i-1})$ každého mezipulsního intervalu. Tímto způsobem získáme mezipulsní intervaly o délkách

$$t_i + a, \quad i \geq 1,$$

kde a je doba trvání absolutní refrakterní periody.

2.4 Náhodnost v neuronálních modelech

Modely zmíněné v předchozí podkapitole jsou v zásadě deterministické. Náhodná posloupnost akčních potenciálů je produkována pouze tehdy, když jsou přítomny náhodné fluktuace ve vstupu nebo v generovaném potenciálu. Dále je možné uvažovat neuronální modely s nestálým prahem anebo s nestabilními parametry – obě tyto

situace způsobí, že výstup neuronu bude stochastický.

Jedním z možných vysvětlení proměnlivosti prahu či parametrů je neurofyziologický fenomén zvaný akomodace. Tato vlastnost neuronu je založena na skutečnosti, že pokud je stimul aplikován v dostatečně nízké frekvenci, neuron není excitován přestože bylo dosaženo jeho předpokládaného prahu. Neuron se sám přizpůsobil změně.¹

Další otázkou je, zda neurony pracují náhodně samy o sobě. Možnými zdroji náhodnosti při generování pulsu mohou být

1. termální šum, kdy teplota ovlivní potenciál neuronální membrány,
2. konečný počet ionových kanálů v membráně neuronu, kdy elektrická vodivost membrány závisí na počtu otevřených kanálů a
3. efekty související se zesíťováním neuronů.

Obecně řečeno, pokud jsou neurony řízeny vstupním napětím o známé závislosti na čase, chovají se víceméně deterministicky. Zda se neurony budou chovat téměř deterministicky nebo spíše náhodně závisí na stimulu.

2.5 Neuronální modely použité v dalším textu

V dalším textu budeme předpokládat, že posloupnosti akčních potenciálů splňují požadavky kladené na procesy obnovy. V takovém případě jsou pozorované mezipulsní intervaly realizacemi stejně rozdělených náhodných veličin T_1, \dots, T_n s hustotou $f(t)$. Tato hustota je obvykle definována na intervalu $[0, \infty)$, zanedbává se doba trvání akčního potenciálu i refrakterní perioda. To je ovšem v případě refrakterní periody značné zjednodušení, reálně by měly být náhodné veličiny T_i definovány na intervalu $[a, \infty)$, kde $a > 0$ je konstanta.

Nejjednodušším stochastickým modelem neuronální aktivity je Poissonův proces. Pokud je posunut o konstantu a , mají mezipulsní intervaly exponenciální rozdělení (1). Univerzálnějším modelem je gamma rozdělení (2). Ve článku Reeke a Coop (2004) se konstatuje, že pomocí hustoty gamma rozdělení lze dobře aproximovat hustotu mezipulsních intervalů pro libovolný model, neboť pro $b \geq 1$ je gamma rozdělení velmi tvárné, od exponenciálního k téměř normálnímu. Připomeňme též

¹Co je však zajímavější, pokud stimul dostatečně pomalu zvyšujeme, neuron zůstane neexcitovaný za jakkoliv silného stimulu. Horním limitem zde bude pouze průrazné napětí membrány.

vztah gamma rozdělení k perfect integrate-and-fire modelu se vstupním Poissonovým procesem, viz str. 27. Exponenciální rozdělení (1) je zvláštním případem gamma rozdělení (2) s $b = 1$. Pro $b = 2$ v rovnici (2) dostáváme hustotu alfa rozdělení (3). Hustota alfa rozdělení reflektuje existenci obou refrakterních period, absolutní i relativní, stejně jako hustota gamma rozdělení.

K modelování posloupnosti intervalů mezi akčními potenciály použijeme i dalších rozdělení: normální s hustotou (4), kde sice je $P(t \leq 0) > 0$, ovšem při vhodné volbě parametrů ($\mu > 0$, σ relativně malé) zanedbatelná; Weibullovo (5), kde se volbou $b = 1$ dostáváme k exponenciálnímu rozdělení; logaritmicko normální (6); inverzní Gaussovo (7).

Charakteristiky vybraných rozdělení při posunutí $a = 0$ jsou uvedeny v tab. 2.1. Při neurofyziologické interpretaci má střední hodnota mezipulsních intervalů souvislost s frekvencí generování pulsů a rozptyl s variabilitou posloupnosti akčních potenciálů. Koeficient variace nám poslouží k výběru konkrétních parametrů daných rozdělení pro simulační studie.

Tab. 2.1: Střední hodnota E, rozptyl D a koeficient variace CV vybraných rozdělení.

Rozdělení	E	D	CV
Normální	μ	σ^2	$\frac{\sigma}{\mu}$
Gamma	$\frac{b}{\lambda}$	$\frac{b}{\lambda^2}$	$\frac{1}{\sqrt{b}}$
Exponenciální	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	1
Weibullovo	$\Gamma(\frac{b+1}{b})c^{-1/b}$	$[\Gamma(\frac{b+2}{b}) - \Gamma^2(\frac{b+1}{b})]c^{-2/b}$	$\sqrt{\frac{2b\Gamma(\frac{2}{b})}{\Gamma^2(\frac{1}{b})} - 1}$
Logaritmicko normální	$e^{\mu+\sigma^2/2}$	$e^{2(\mu+\sigma^2/2)}(e^{\sigma^2} - 1)$	$\sqrt{e^{\sigma^2} - 1}$
Inverzní Gaussovo	μ	μ^3/λ	$\sqrt{\mu/\lambda}$

Dalším modelem, se kterým budeme pracovat, je Ornstein-Uhlenbeckův proces generování pulsů, který je jedním z nejznámějších neuronálních modelů (viz např. Gerstner a Kistler (2002)). Tento model je speciálním případem leaky integrate-and-fire modelu (2.1), kde vstup $X(\tau)$ je Gaussovský bílý šum. Alternativní popis tohoto modelu může být proveden následujícím způsobem.

Předpokládejme náhodný proces $X(\tau)$, popisující depolarizaci membrány, která generuje akční potenciál vždy při prvním dosažení určité úrovně. Model depolarizace membrány je založen na předpokladu, že vstup neuronu je načítán v čase až do

dosažení jistého prahu. V tomto momentě je generován akční potenciál a načítání začíná znova. Pokud je zde refrakterní perioda o délce a , jsou mezipulsní intervaly determinovány časem T_S prvního dosažení prahu S procesem $X(\tau)$, což můžeme zapsat jako

$$T_S = \inf\{\tau \geq a, X(\tau) > S | X(a) = x_0 = 0 < S\} \quad (2.4)$$

a proces $X(\tau)$ je popsán rovnicí

$$dX(\tau) = \left(-\frac{X(\tau)}{\phi} + \mu \right) d\tau + \sigma dW(\tau), \quad X(a) = 0, \quad (2.5)$$

kde $W(\tau)$ představuje Wienerův proces (integrál z δ korelovaného šumu s nulovou střední hodnotou a amplitudou rovnou jedné, viz příklad 3.9), ϕ je časová konstanta modelu. Amplituda šumu je dána odchylkou $\sigma > 0$, a parametr μ , který zde reprezentuje signál, je konstantou způsobující v procesu $X(\tau)$ růst. Hustota náhodné veličiny (2.4) není známa v uzavřené formě (viz Lansky a Sato (1999)), a proto musí být při práci s tímto modelem použita simulační technika.

Konvence pro další kapitoly

V dalším textu, pokud to nebude bezpodmínečně nutné, již nebudeme používat neurofyziologické termíny, ale jejich ekvivalenty se statistickým významem. Zejména namísto spojení „mezipulsní interval“ budeme užívat pojmu náhodná veličina, resp. realizace a pokud nebude uvedeno jinak značit T , resp. t . Ze statistického hlediska je absolutní refrakterní perioda ekvivalentní posunutí (relativní refrakterní periodou se nebudeme zabývat), proto budeme v dalším textu používat tento termín a označovat symbolem a . Při odhadu entropie z cenzorovaných dat budeme namísto pojmu puls užívat obecnější termíny „událost“ či „obnova“.

Kapitola 3

Statistické základy

V této kapitole shrneme statistické definice, metody a postupy používané, analyzované a rozpracováváné v následujících kapitolách.

3.1 Charakteristiky rozdělení

Definice 3.1. Obecný s -tý moment α_s náhodné veličiny T kolem bodu c je dán vztahem

$$\alpha_s = E(T - c)^s, \quad s = 0, 1, 2, \dots$$

Řekneme, že moment α_s existuje, je-li $E|T - c|^s < \infty$.

Definice 3.2. Pokud v definici 3.1 položíme $c = ET$, budeme mluvit o centrálních momentech

$$\mu_s = E(T - ET)^s, \quad s = 0, 1, 2, \dots$$

Poznámka. Mezi obecnými momenty a centrálními momenty platí vztahy

$$\mu_0 = 1,$$

$$\mu_1 = 0,$$

$$\mu_2 = \alpha_2 - \alpha_1^2,$$

$$\mu_3 = \alpha_3 - 3\alpha_1\alpha_2 + 2\alpha_1^3,$$

$$\mu_4 = \alpha_4 - 4\alpha_1\alpha_3 + 6\alpha_1^2\alpha_2 - 3\alpha_1^4.$$

Definice 3.3. Kumulanty jsou lineární kombinace momentů volené tak, aby kumulant součtu nezávislých náhodných veličin byl součtem kulantů jednotlivých sčítanců. Označíme-li s -tý kumulant jako κ_s , má tedy pro nezávislé náhodné veličiny T_1, \dots, T_n platit

$$\kappa_s(T_1 + \dots + T_n) = \kappa_s(T_1) + \dots + \kappa_s(T_n).$$

Tuto vlastnost mají charakteristiky definované jako

$$\begin{aligned}\kappa_1 &= ET \\ \kappa_2 &= \mu_2 \\ \kappa_3 &= \mu_3 \\ \kappa_4 &= \mu_4 - 3\mu_2^2 \\ \kappa_5 &= \mu_5 - 10\mu_2\mu_3 \\ \kappa_6 &= \mu_6 - 15\mu_2\mu_4 - 10\mu_3^2 + 30\mu_2^3.\end{aligned}$$

Poznámka. Existuje-li k -tý moment kolem počátku, tj. α_k pro $c = 0$, lze charakteristickou funkci rozvinout pro malé t v řadu

$$\psi(t) = 1 + \sum_{j=1}^k \frac{\alpha_j}{j!} (it)^j + o(t^k).$$

Položíme-li

$$z = \sum_{j=1}^k \frac{\alpha_j}{j!} (it)^j + o(t^k)$$

a použijeme dalšího Taylorova rozvoje

$$\ln(1+z) = z - \frac{1}{2}z^2 + \frac{1}{3}z^3 - \dots + (-1)^{k+1} \frac{1}{k} z^k + o(z^k),$$

dostaneme

$$\ln \psi(t) = \sum_{j=1}^k \frac{\kappa_j}{j!} (it)^j + o(t^k). \quad (3.1)$$

Jelikož se při nezávislosti charakteristické funkce násobí (a tudíž jejich logaritmy se

sčítají) vyplývá z (3.1) aditivní vlastnost kumulantů při nezávislosti.

Poznámka. Mezi základní charakteristiky hustoty $f(t)$ patří míra její polohy, rozptýlenosti, šikmosti a špičatosti. Jako míru polohy používáme střední hodnotu μ , jako míru variability směrodatnou odchylku σ , jako míru šikmosti (informaci o symetrii či asymetrii) používáme vztah

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\kappa_3}{\sigma^3}$$

a pro míru špičatosti předpis

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 = \frac{\kappa_4}{\sigma^4},$$

který zaručí, že míra špičatosti je nulová pro normální rozdělení.

3.2 Náhodné procesy

V této podkapitole uvedeme některé definice a charakteristiky vztahující se k náhodným procesům. Čerpat zde budeme ze knih Anděl (1976), Cox (1962) a Prášková a Lachout (2001).

Definice 3.4. Nechť (Ω, \mathcal{A}, P) je pravděpodobnostní prostor, nechť $\mathcal{T} \subset \mathbb{R}$. Systém reálných náhodných veličin $\{X_\tau, \tau \in \mathcal{T}\}$ definovaných na (Ω, \mathcal{A}, P) se nazývá náhodný proces.

Poznámka. V případě, že $\mathcal{T} = \mathbb{Z}$ nebo $\mathcal{T} = \mathbb{N}_0$, mluvíme o procesu s diskrétním časem nebo časové řadě. Pokud $\mathcal{T} = [a, b]$, kde $-\infty \leq a < b \leq \infty$, říkáme, že $\{X_\tau, \tau \in \mathcal{T}\}$ je proces se spojitým časem.

Poznámka. Náhodný proces se podle okolností označuje také $\{X_\tau(\omega)\}$, $\{X_\tau\}$ anebo $\{X(\tau, \omega)\}$; symbol $\tau \in \mathcal{T}$ se vynechává, je-li ze souvislosti zřejmé, o jakou množinu \mathcal{T} jde.

Poznámka. Náhodný proces $\{X_\tau(\omega), \tau \in \mathcal{T}\}$ můžeme chápat jako funkci dvou proměnných ω, τ . Pro pevné $\tau \in \mathcal{T}$ je $X_\tau = X_\tau(\cdot)$ náhodná veličina definovaná na Ω ; pro pevné $\omega \in \Omega$ je $X_{(\cdot)} = X_{(\cdot)}(\omega)$ reálnou funkcí proměnné τ . Této funkci říkáme trajektorie procesu $\{X_\tau(\omega), \tau \in \mathcal{T}\}$.

Poznámka. Každé konečné podmnožině $\{\tau_1, \dots, \tau_n\} \subset \mathcal{T}$, $n \in \mathbb{N}$, lze přiřadit systém náhodných veličin $\{X_{\tau_1}, \dots, X_{\tau_n}\}$, které mají simultánní rozdělení s distribuční funkcí

$$F_{\tau_1, \dots, \tau_n}(x_1, \dots, x_n) = P(X_{\tau_1} \leq x_1, \dots, X_{\tau_n} \leq x_n).$$

Definice 3.5. Necht' $\{X_\tau, \tau \in \mathcal{T}\}$ je náhodný proces takový, že pro každé $\tau \in \mathcal{T}$ existuje střední hodnota EX_τ . Potom funkce

$$u_\tau = EX_\tau$$

definovaná na \mathcal{T} se nazývá střední hodnota procesu $\{X_\tau\}$. Jestliže platí $E|X_\tau|^2 < \infty$ pro všechna $\tau \in \mathcal{T}$, potom funkce dvou proměnných definovaná na $\mathcal{T} \times \mathcal{T}$ předpisem

$$R(s, \tau) = E(X_s - u_s)(X_\tau - u_\tau)$$

se nazývá autokovarianční funkce procesu $\{X_\tau\}$. Hodnota $R(\tau, \tau)$ se nazývá rozptyl procesu v čase τ .

Definice 3.6. Řekneme, že náhodný proces $\{X_\tau, \tau \in \mathcal{T}\}$ je striktně stacionární, jestliže pro libovolné $n \in \mathbb{N}$, pro libovolná reálná x_1, \dots, x_n a pro libovolná τ_1, \dots, τ_n a h taková, že $\tau_k \in \mathcal{T}$, $\tau_k + h \in \mathcal{T}$, $1 \leq k \leq n$ platí

$$F_{\tau_1, \dots, \tau_n}(x_1, \dots, x_n) = F_{\tau_1+h, \dots, \tau_n+h}(x_1, \dots, x_n).$$

Poznámka. Z definice striktní stacionarity mimo jiné plyne, že všechny náhodné veličiny X_τ mají stejné rozdělení a že základní charakteristiky jako střední hodnota a autokovarianční funkce se nemění při posunutí v čase.

Příklad 3.7. Bílý šum je proces $\{X_\tau, \tau \in \mathbb{Z}\}$ nekorelovaných náhodných veličin s nulovou střední hodnotou a stejným konečným rozptylem. Pokud jsou náhodné veličiny X_τ nezávislé, mluvíme o striktním bílém šumu. Název je odvozen ze spektrálních vlastností tohoto procesu a analogie s fyzikálními vlastnostmi bílého světla.

Příklad 3.8. Předpokládejme, že v intervalu $(\tau, \tau + h]$, a to nezávisle na τ , dojde k výskytu jedné události s pravděpodobností $\lambda h + o(h)$, $\lambda > 0$, více než jedné události s pravděpodobností $o(h)$, a že počty událostí, které se vyskytnou v disjunktních časových intervalech, jsou vzájemně nezávislé náhodné veličiny. Necht' N_τ značí počet událostí v intervalu $[0, \tau]$. Potom $\{N_\tau, \tau \geq 0\}$ je náhodný proces, který se nazývá

Poissonův. Počty událostí N_τ mají Poissonovo rozdělení s parametrem $\lambda\tau$,

$$P(N_\tau = k) = \frac{e^{-\lambda\tau}(\lambda\tau)^k}{k!}, \quad k = 0, 1, \dots$$

Konstanta λ se nazývá intenzita Poissonova procesu.

Příklad 3.9. Wienerův proces (nebo také Brownův pohyb) je náhodný proces $\{W_\tau, \tau \geq 0\}$, který má tyto vlastnosti:

1. $W_0 = 0$ a $\{W_\tau, \tau \geq 0\}$ má spojité trajektorie.
2. Pro libovolné časové okamžiky $0 \leq \tau_1 < \tau_2 < \dots < \tau_n$ jsou přírůstky procesu $W_{\tau_2} - W_{\tau_1}, W_{\tau_3} - W_{\tau_2}, \dots, W_{\tau_n} - W_{\tau_{n-1}}$ nezávislé náhodné veličiny.
3. Pro libovolné $0 \leq \tau < s$ mají přírůstky $W_s - W_\tau$ normální rozdělení s nulovou střední hodnotou a rozptylem $\sigma^2(s - \tau)$, kde σ^2 je kladná konstanta.

Definice 3.10. Nechť $\{T_n, n \in \mathbb{N}_0\}$ je posloupnost nezávislých náhodných veličin, které nabývají pouze kladných hodnot; dále předpokládejme, že T_1, T_2, \dots mají stejné rozdělení s distribuční funkcí F a se střední hodnotou μ . Položme

$$S_n = \sum_{k=0}^n T_k, \quad n \geq 0.$$

Potom proces náhodných veličin $\{N_\tau, \tau \geq 0\}$ takových, že

$$N_\tau = \sup\{n : S_n \leq \tau\},$$

se nazývá proces obnovy.

Poznámka. Náhodná veličina S_n značí čas, kdy dojde k výskytu $(n + 1)$ -ní události (obnovy), N_τ je počet událostí v intervalu $[0, \tau]$. Je-li $T_0 = S_0 = 0$, považujeme počátek za čas obnovy. Někdy v tomto případě mluvíme o čistém procesu obnovy. Náhodné veličiny T_1, T_2, \dots jsou doby mezi událostmi.

Věta 3.11. Intervalů mezi událostmi u Poissonova procesu $\{N_\tau, \tau \geq 0\}$ o intenzitě λ mají exponenciální rozdělení (1) s parametrem $a = 0$.

Důkaz. Viz Tuckwell (1988). □

Definice 3.12. Dobu mezi τ a následující událostí, tj.

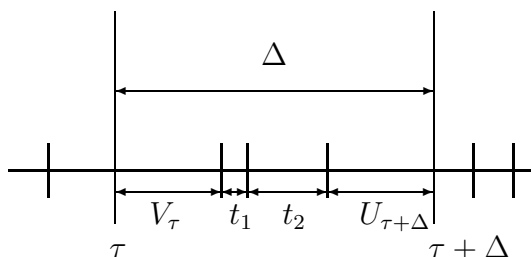
$$V(\tau) = S_{N_\tau} - \tau,$$

označujeme jako dopředný rekurenční čas. O době mezi poslední událostí a časem τ , tedy

$$U(\tau) = \tau - S_{N_{\tau-1}}, \quad \tau \geq S_0,$$

mluvíme jako o zpětném rekurenčním čase.

Poznámka. Rekurenční časy jsou ilustrovány na obr. 3.1.



Obr. 3.1: Ilustrace rekurenčních časů.

Definice 3.13. Nechť $f(t)$ je hustota intervalů mezi událostmi. Potom funkci

$$\mathcal{F}(t) = \int_t^\infty f(u) du$$

nazveme funkcí přežití. Dále nechť výraz $E(N_{\tau, \tau + \Delta})$ je střední hodnota počtu událostí na intervalu $(\tau, \tau + \Delta)$. Potom definujeme tzv. hustotu obnovy jako funkci

$$h(\tau) = \lim_{\Delta \rightarrow 0^+} \frac{E(N_{\tau, \tau + \Delta})}{\Delta}.$$

Věta 3.14. Nechť $f_1(t)$ označuje hustotu prvního intervalu mezi událostmi (předpokládáme, že první událost nastane v čase 0) a Δ je šířka intervalu na kterém sledujeme nastoupení událostí. Potom hustota U_τ je dána jako

$$f_{U_\tau}(t) = h(\tau - t)\mathcal{F}(t) \tag{3.2}$$

a hustota V_τ je rovna

$$f_{V_\tau}(t) = f_1(\tau + t) + \int_0^\tau h(\tau - u)f(u + t)du. \quad (3.3)$$

Dále, pokud uvažujeme limitní rozdělení pro $\Delta \rightarrow \infty$, pak limitní hustota je pro oba rekurenční časy dána jako

$$f_{lim} = \frac{\mathcal{F}(t)}{\mu}, \quad (3.4)$$

kde

$$\frac{1}{\mu} = \lim_{\Delta \rightarrow \infty} h(\Delta - t).$$

Důkaz. Viz Cox (1962). □

Příklad 3.15. Pro Poissonův proces o intenzitě λ platí $\mathcal{F}(t) = e^{-\lambda t}$, $h(t) = \lambda$ a $f_{V_\tau}(t) = f_{U_\tau}(t) = \lambda e^{-\lambda t}$. Rozdělení obou rekurenčních časů je zde tedy stejné jako rozdělení intervalů mezi událostmi.

3.3 Parametrické odhady rozdělení

Tato kapitola je zpracována na základě knih Anděl (1978), Lehmann a Casella (1998) a Machek (1980).

3.3.1 Základy teorie odhadu

Předpokládejme, že náhodný vektor $\mathbf{T} = (T_1, \dots, T_n)$ má hustotu $f(\mathbf{t}, \boldsymbol{\theta})$, přičemž $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ je neznámý parametr. Na základě \mathbf{T} je třeba získat co nejlepší odhad parametru $\boldsymbol{\theta}$, o kterém víme pouze to, že náleží do parametrického prostoru Θ . V případě bodového odhadu, kterým se budeme dále zabývat, to znamená najít zobrazení g takové, aby náhodný vektor $\mathbf{U} = g(\mathbf{T})$ v nějakém rozumném smyslu co nejlépe approximoval hodnotu $\boldsymbol{\theta}$.

Definice 3.16. Řekneme, že odhad \mathbf{U} parametru $\boldsymbol{\theta}$ je nestranný, platí-li $E\mathbf{U} = \boldsymbol{\theta}$ pro každé $\boldsymbol{\theta} \in \Theta$. Řekneme, že odhad \mathbf{U} parametru $\boldsymbol{\theta}$ je asymptoticky nestranný, platí-li $\lim_{n \rightarrow \infty} E\mathbf{U} = \boldsymbol{\theta}$ pro každé $\boldsymbol{\theta} \in \Theta$. Platí-li $E\mathbf{U} = \boldsymbol{\theta} + \mathbf{b}(\boldsymbol{\theta})$, kde funkce \mathbf{b} není identicky rovna nule na množině Θ , nazývá se odhad \mathbf{U} vychýlený. Vektoru $\mathbf{b}(\boldsymbol{\theta})$ se pak říká vychýlení odhadu \mathbf{U} v bodě $\boldsymbol{\theta}$.

Definice 3.17. Nechť θ je jednorozměrný parametr a předpokládejme, že T_1, T_2, \dots je výběr z rozdělení závislého na θ . Pro každé přirozené číslo n mějme definován odhad $U_n = g_n(T_1, \dots, T_n)$. Řekneme, že odhad U_n je konzistentní, jestliže $U_n \rightarrow \theta$ podle pravděpodobnosti při $n \rightarrow \infty$.

3.3.2 Maximálně věrohodné odhady

Nechť náhodný vektor T_1, \dots, T_n má hustotu $f(t_1, \dots, t_n; \theta)$. Parametr $\theta \in \Theta$, kde Θ je interval na přímce, respektive k -rozměrný vektor parametrů $\theta \in \Theta$, kde Θ je část k -rozměrného euklidovského prostoru.

Definice 3.18. Funkci parametru θ definovanou pro všechny možné realizace $(T_1, \dots, T_n) = (t_1, \dots, t_n)$ vztahem

$$L(\theta|t_1, \dots, t_n) = f(t_1, \dots, t_n; \theta)$$

nazveme funkcí věrohodnosti.

Metoda maximální věrohodnosti spočívá v tom, že se jako odhad parametru θ volí odhad $\hat{\theta}(t_1, \dots, t_n)$ takový, že platí

$$L(\hat{\theta}(t_1, \dots, t_n)|t_1, \dots, t_n) = \sup_{\theta \in \Theta} L(\theta|t_1, \dots, t_n), \quad (3.5)$$

pokud funkce $L(\theta|t_1, \dots, t_n)$ v některém bodě parametrického prostoru svého suprema nabývá a pokud je tento bod skutečně závislý na t_1, \dots, t_n .

Definice 3.19. Statistika $\hat{\theta}(t_1, \dots, t_n)$, splňující (3.5), se nazývá maximálně věrohodný odhad parametru θ .

Poznámka. Zpravidla je výhodnější pracovat s logaritmem funkce věrohodnosti místo s původní funkcí věrohodnosti. Logaritmus funkce věrohodnosti,

$$l(\theta|t_1, \dots, t_n) = \ln L(\theta|t_1, \dots, t_n), \quad L(\theta|t_1, \dots, t_n) > 0,$$

se nazývá logaritmická funkce věrohodnosti. Protože je logaritmus ryze rostoucí funkcí, můžeme podmínku (3.5) nahradit obdobnou podmínkou pro logaritmickou věrohodnostní funkci: maximálně věrohodným odhadem bude odhad $\hat{\theta}(t_1, \dots, t_n)$

splňující podmínku

$$l(\hat{\theta}(t_1, \dots, t_n) | t_1, \dots, t_n) = \sup_{\theta \in \Theta} l(\theta | t_1, \dots, t_n). \quad (3.6)$$

Poznámka. Pokud

1. funkce $l(\theta | t_1, \dots, t_n)$ nabývá svého suprema v některém vnitřním bodě parametrického prostoru;
2. množina hodnot t_1, \dots, t_n , pro které $f(t_1, \dots, t_n; \theta) > 0$ nezávisí na θ ;
3. funkce $l(\theta | t_1, \dots, t_n)$ má derivaci podle θ , popř. při vektorovém parametru $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ první parciální derivace podle všech složek parametru $\boldsymbol{\theta}$,

lze maximálně věrohodný odhad parametru θ hledat obvyklými metodami pro stanovení extrému, tedy řešením rovnice

$$\frac{\partial l(\theta | t_1, \dots, t_n)}{\partial \theta} = 0$$

při reálném parametru, a řešením soustavy rovnic

$$\frac{\partial l(\boldsymbol{\theta} | t_1, \dots, t_n)}{\partial \theta_i} = 0, \quad i = 1, \dots, k$$

při vektorovém parametru.

Příklad 3.20. Pro exponenciální rozdělení (1) postupně získáme

$$f(t_1, \dots, t_n; \lambda, a) = \lambda^n e^{n\lambda a} e^{-\lambda(t_1 + \dots + t_n)} = L(\lambda, a | t_1, \dots, t_n)$$

a

$$l(\lambda, a | t_1, \dots, t_n) = n \ln(\lambda) + n\lambda a - \lambda(t_1 + \dots + t_n).$$

Pokud uvažujeme parametr a jako pevně daný, dostaneme odhad parametru λ jako

$$\lambda_{MLE} = \frac{1}{\frac{1}{n}(t_1 + \dots + t_n) - a}.$$

Pro současný odhad obou parametrů stanovíme soustavu rovnic

$$\begin{aligned} \frac{n}{\lambda} - (t_1 + \dots + t_n) + na &= 0 \\ \lambda n &= 0, \end{aligned}$$

která však nemá řešení, což je v souladu s předešlou poznámkou – podmínka 2. je porušena.

Maximálně věrohodný odhad parametru a odvodíme následující úvahou: bez ohledu na parametr λ hodnota věrohodnostní funkce poroste se zvětšujícím se parametrem a . Přírozenou hranicí, kam až může a růst na základě realizací t_1, \dots, t_n je hodnota $\min\{t_1, \dots, t_n\}$, která je tedy maximálně věrohodným odhadem parametru a . Maximálně věrohodný odhad obou parametrů pak dostaneme jako

$$\hat{a} = T_{(1)} \quad \text{a} \quad \hat{\lambda} = \frac{1}{\frac{1}{n} \sum_{i=1}^n [T_i - T_{(1)}]}. \quad (3.7)$$

Poznámka. Systém věrohodnostních rovnic se nám obvykle podaří sestavit. Většinou však není možné získat jeho řešení analyticky, je třeba maximalizovat věrohodnostní funkci numericky.

Maximálně věrohodné odhady a parametr posunutí

Přítomnost parametru posunutí může zásadním způsobem ovlivnit maximálně věrohodné odhady zbylých parametrů i jejich charakteristiky. Pro konstrukci charakteristik maximálně věrohodných odhadů jsme využili tvrzení z knihy Anděl (1978). V případě exponenciálního rozdělení (1) s $a = 0$ můžeme odvodit rozdělení pro maximálně věrohodný odhad Z parametru λ ,

$$Z = \frac{1}{\frac{1}{n} \sum_{i=1}^n T_i},$$

jako hustotu

$$f_Z(z) = \frac{(n\lambda)^n}{(n-1)!} \frac{1}{z^{n+1}} e^{-\lambda n \frac{1}{z}}.$$

Jeho střední hodnota je

$$E(Z) = \lambda \frac{n}{n-1}$$

a rozptyl

$$D(Z) = \lambda^2 \frac{n^2}{(n-1)^2(n-2)}.$$

Nyní lze provádět úsudky o nestrannosti či konzistenci získaného odhadu a popř. navrhnout v tomto směru kvalitnější odhad.

V případě, že uvažujeme exponenciální rozdělení (1) s $a \neq 0$, se situace změní.

Dostaneme

$$f_Z(z) = \frac{(n\lambda)^n}{(n-1)!} \frac{(\frac{1}{z} - a)^{n-1}}{z^2} e^{-\lambda n(\frac{1}{z} - a)}, \text{ kde } z > na.$$

Nyní vyjádříme střední hodnotu jako

$$E(Z) = (n\lambda)^n a^{n-1} \Gamma(-(n-1), an\lambda).$$

Odvozená střední hodnota obsahuje neúplnou Gamma funkci se záporným prvním parametrem, takže ji není možné převést na nekonečnou řadu, pouze aproximovat. Nelze analyticky vypočítat integrál pro $E(Z^2)$, proto nelze vyjádřit rozptyl. Je tedy vidět, že přítomnost posunutí u rozdělení (1) zásadně mění statistické vlastnosti odhadu vhodného v případě $a = 0$.

3.4 Neparаметrické odhady hustoty

3.4.1 Odhad hustoty histogramem

Histogram je základním neparаметrickým odhadem hustoty. Jeho konstrukce je relativně snadná, přesto se zde objevují problémy s nastavením volitelných parametrů histogramu.

Definice 3.21. Pro naše účely používáme histogram definovaný jako

$$\hat{f}_{hist}(t) = \sum_{j=2}^m \frac{1}{nd_{j-1}} I_{[z_{j-1}, z_j)}(t) \left(\sum_{i=1}^n I_{[z_{j-1}, z_j)}(t_i) \right) + \frac{1}{nd_m} I_{[z_m, z_{m+1}]}(t) \left(\sum_{i=1}^n I_{[z_m, z_{m+1}]}(t_i) \right), \quad t \in \mathbb{R} \quad (3.8)$$

s m ekvidistantními nebo neekvidistantními dělicími intervaly o délkách $d_j = z_{j+1} - z_j$, $j = 1, \dots, m$, a hranicích $z_1 < z_2 < \dots < z_m < z_{m+1}$. V zásadě můžeme říci, že při konstrukci histogramu rozložíme interval $[z_1; z_{m+1}] \in \mathbb{R}$ na podintervaly a na každém z nich neznámou hustotu odhadneme konstantou.

Poznámka. Můžeme zvolit ekvidistantní dělení, kde $d_j = d \forall j$ je konstanta. Pro určení d pak platí vztah

$$d = \frac{z_{m+1} - z_1}{m}.$$

Jiné smysluplné (neekvidistantní) dělení dostaneme, když požadujeme, aby v každém

námi určeném intervalu ležel stejný počet pozorování (až třeba na poslední dělicí interval).

Poznámka. Histogram je definován tak, aby splňoval

$$\int_{-\infty}^{\infty} \hat{f}_{hist}(t) dt = 1.$$

Volba „parametrů“ histogramu

Histogram při své konstrukci vyžaduje zadání „vnitřních parametrů“. Je to počet dělicích intervalů m (u ekvidistantního dělení ekvivalent s šířkou intervalu; u dělení se stejným počtem pozorování v dělicím intervalu ekvivalent s počtem pozorování v intervalu) a vymezení intervalu $[z_1, z_{m+1}]$ na němž histogram konstruujeme. Optimální volba závisí v obou případech nejen na velikosti náhodného výběru n , ale také na typu rozdělení a na hodnotách jejich parametrů.

Vhodnou volbou počtu dělicích intervalů m se snažíme najít kompromis mezi dvěma protichůdnými požadavky: mít co nejvíce dělicích intervalů pro přesný popis odhadované hustoty a zároveň mít co nejvíce pozorování v každém dělicím intervalu pro dobrý odhad hustoty na tomto intervalu. V literatuře se doporučují následující volby počtu ekvidistantních intervalů:

- Sturgersovo pravidlo $m \approx 1 + 3,3 \log n$ (viz Sturgers (1926));
- Doanovo pravidlo pro ne-normální data $m \approx 3 + [\ln n] + [\ln(1 + \hat{\gamma}_1 \sqrt{n/6})]$ (viz Doane (1976)).

Doporučené volby šířky ekvidistantních intervalů jsou

- $d \approx 3,49 \hat{\sigma} n^{-\frac{1}{3}}$ (vhodné zejména pro normální rozdělení, viz Scott (1979)) a
- $d \approx 2 \text{IQR} n^{-\frac{1}{3}}$ (viz Freedman a Diaconis (1981)).

Jednou z možných voleb hranic histogramu je $z_1 = \min\{t_1, \dots, t_n\}$ a $z_{m+1} = \max\{t_1, \dots, t_n\}$. Tato volba ale nerespektuje zejména nekonečné definiční obory hustot. Při stanovování okrajů histogramu bude tedy vhodné zhodnotit možný interval, ve kterém se bude náš náhodný výběr realizovat.

Pro výběry z normálního rozdělení můžeme použít volbu $z_1 = \min\{t_1, \dots, t_n\} - s_t$ a $z_{m+1} = \max\{t_1, \dots, t_n\} + s_t$, kde s_t je odhad směrodatné odchylky pro realizace

náhodného výběru. V našem případě, kdy se zabýváme především rozděleními definovanými na polopřímce, bude opodstatněná volba $z_1 = 0$ či $z_1 = k < \min\{t_1, \dots, t_n\}$ a $z_{m+1} = \max\{t_1, \dots, t_n\} + l$, kde k a l budou kladné konstanty.

3.4.2 Jádrové odhady hustoty

Sofistikovanějším odhadem hustoty je jádrový odhad.

Definice 3.22. Jádrem nazýváme funkci $K : \mathbb{R} \rightarrow (0, \infty)$, která je symetrická, ohraničená a pro niž platí

$$\int_{-\infty}^{\infty} K(t)dt = 1 \quad \text{a} \quad \lim_{t \rightarrow \pm\infty} |t|K(t) = 0.$$

Poznámka. Přehled často užívaných jader je uveden v tab. 3.1.

Tab. 3.1: Přehled obvyklých jader.

Gaussovo	$\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}$	Epanečnikovo	$\frac{3}{4}(1-t^2)I_{[-1,1]}(t)$
Dvojitě vážené	$\frac{15}{16}(1-t^2)^2I_{[-1,1]}(t)$	Trojité vážené	$\frac{35}{32}(1-t^2)^3I_{[-1,1]}(t)$
Trojúhelníkové	$(1- t)I_{[-1,1]}(t)$	Obdélníkové	$\frac{1}{2}I_{[-1,1]}(t)$
Kosinové	$\frac{1}{2}\cos(t)I_{[-1,1]}(t)$	Laplaceovo	$\frac{1}{2}e^{- t }$

Definice 3.23. Nechť K je jádro a $h > 0$ je konstanta. Potom jádrový odhad hustoty je dán vztahem

$$\hat{f}_{jad}(t) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{t-t_i}{h}\right) \quad t \in \mathbb{R}. \quad (3.9)$$

Konstanta h hraje roli měřítka, kterým se mění proporce jádra. Klíčovou otázkou je volba nejvhodnější konstanty h . Většina doporučení vychází z volby vhodné míry kvality (přesnosti) odhadu, jako je např. střední čtvercová chyba, maximum vychýlení atp., a úvah o její asymptotické minimalizaci. Jejich společnou nevýhodou je potřeba znalosti mnoha apriorních informací o odhadované hustotě.

V praxi se k této úloze přistupuje zpravidla empiricky. Odhad se vypočte pro různé hodnoty h_n a za optimální se zvolí to h_n , pro něž je výsledný odhad přijatelně hladký. Další možností je použít metodu křížového ověřování. Hledá se konstanta h , jež maximalizuje věrohodnost

$$L(h) = \prod_{i=1}^n f_{ni}(t_i),$$

kde

$$f_{ni}(t) = \frac{1}{nh_n} \sum_{(j=1), (j \neq i)}^n K\left(\frac{t-t_j}{h_n}\right), \quad t \in \mathbb{R} \quad (3.10)$$

a h se vezme jako optimální šířka jádra.

3.4.3 Momentové odhady hustoty

V této kapitole se budeme zabývat málo používanými momentovými odhady hustoty. Nechť $f(t)$ je hustota s kumulanty $\kappa_1, \kappa_2, \dots$ a $e_j, j = 1, 2, \dots$ jsou konstanty. Pak funkce

$$g(t) = \exp\left(\sum_{j=1}^{\infty} e_j \frac{(-\mathcal{D})^j}{j!}\right) f(t) \quad (3.11)$$

má kumulanty $\kappa_1 + e_1, \kappa_2 + e_2, \dots$. Exponenciálu ve vztahu (3.11) lze rozvinout v řadu a psát

$$g(t) = \sum_{i=0}^{\infty} \frac{\left(\sum_{j=1}^{\infty} e_j \frac{(-\mathcal{D})^j}{j!}\right)^i}{i!} f(t). \quad (3.12)$$

Uvažujeme-li standardizovanou veličinu a jako $f(t)$ vezmeme hustotu normálního rozdělení, dostaneme

$$e_1 = 0, \quad e_2 = 0, \quad e_3 = \gamma_1 \quad \text{a} \quad e_4 = \gamma_2.$$

Potom volbou $j = 1, 2, 3, 4, i = 0, 1$ v (3.12) dostaneme tzv. Gram-Charlierův odhad

$$\hat{f}_{GC}(y) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2} \left[1 + \frac{\gamma_1}{6}(y^3 - 3y) + \frac{\gamma_2}{24}(y^4 - 6y^2 + 3) \right]. \quad (3.13)$$

Volbou $j = 1, 2, 3, 4$, $i = 0, 1, 2$ v (3.12) a vynecháním členů řádu menšího než $1/n$ (viz tab. 3.2) dostaneme tzv. Edgeworthův odhad

$$\begin{aligned} \hat{f}_{Edg}(y) = & (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2} \left[1 + \frac{\gamma_1}{6}(y^3 - 3y) + \frac{\gamma_2}{24}(y^4 - 6y^2 + 3) \right. \\ & \left. + \frac{\gamma_1^2}{72}(y^6 - 15y^4 + 45y^2 - 15) \right]. \end{aligned} \quad (3.14)$$

Známe-li tedy μ , σ , γ_1 a γ_2 , můžeme hustotu $f(t)$ aproximovat jako

$$f(t) \approx \frac{1}{\sigma} \tilde{f}\left(\frac{t - \mu}{\sigma}\right),$$

kde \tilde{f} je dáno vztahem (3.14) pro Edgeworthův odhad hustoty respektive vztahem (3.13) pro Gram-Charlierův odhad hustoty. Ačkoliv Edgeworthův odhad obsahuje více členů než Gram-Charlierův odhad, není obecně lepší než Gram-Charlierův odhad.

Edgeworthův odhad (3.14) můžeme odvodit i následujícím způsobem. Předpokládejme, že máme k dispozici charakteristiky μ , σ , γ_1 , γ_2 a chceme z nich přibližně zrekonstruovat hustotu $f(t)$. Vyjdeme od obecnějšího problému, jak využít charakteristik

$$\kappa_j / \sigma^j, \quad 3 \leq j \leq k.$$

Budeme uvažovat náhodnou veličinu T , která je součtem n nezávislých stejně rozdělených náhodných veličin T_1, \dots, T_n . Tyto veličiny mají charakteristiky μ' , σ'^2 , κ'_3 , κ'_4 atd. Náhodná veličina T potom bude mít tyto charakteristiky:

$$\begin{aligned} \mu &= n\mu', \\ \sigma^2 &= n\sigma'^2, \\ \kappa_j &= n\kappa'_j, \quad 3 \leq j \leq k. \end{aligned}$$

Můžeme tedy psát

$$\gamma_{j-2} = \frac{\kappa_j}{\sigma^j} = n^{1-\frac{j}{2}} \frac{\kappa'_j}{\sigma'^j}.$$

Seřadíme-li charakteristiky γ_{j-2} a jejich mocniny podle toho, jakého jsou řádu v mocninách n , dostaneme tab. 3.2.

Tab. 3.2: Charakteristiky γ seřazené dle mocnin n .

řád	charakteristika
$n^{-\frac{1}{2}}$	γ_1
n^{-1}	γ_2, γ_1^2
$n^{-\frac{3}{2}}$	$\gamma_3, \gamma_2\gamma_1^2, \gamma_1^3$
\vdots	\vdots

Provedeme rozvoj charakteristické funkce $\psi(t)$ v řadu a výše uvedená tabulka nám bude vodítkem při zanedbávání členů. S použitím (3.1) dostáváme

$$\psi(t) = e^{it\mu - \frac{1}{2}t^2\sigma^2} \exp \left\{ \sum_{j=3}^k \frac{\kappa_j}{j!} (it)^j + o(t^k) \right\}.$$

Nyní přejdeme ke standardizované náhodné veličině

$$\tilde{T} = \frac{T - \mu}{\sigma},$$

kteřá má charakteristickou funkci

$$\tilde{\psi}(t) = e^{-\frac{1}{2}t^2} \exp \left\{ \sum_{j=3}^k \frac{\kappa_j}{j!\sigma^j} (it)^j + o(t^k/\sigma^k) \right\}.$$

Rozvineme-li druhou exponenciálu v Taylorovu řadu a vynecháme členy řádu menšího než $1/n$, respektive obsahující $o(t^k/\sigma^k)$, dostaneme při $\gamma_{j-2} = \kappa_j/\sigma^j$

$$\tilde{\psi}(t) \approx e^{-\frac{1}{2}t^2} \left[1 + \frac{\gamma_1}{3!}(it)^3 + \frac{\gamma_2}{4!}(it)^4 + \frac{1}{2} \left(\frac{\gamma_1}{3!} \right)^2 (it)^6 \right]. \quad (3.15)$$

Principem Edgeworthova rozvoje nyní je, že hustotu $f(t)$ aproximujeme hustotou, pro niž je charakteristická funkce rovna pravé straně vztahu (3.15). Jelikož platí

$$\int_{-\infty}^{\infty} e^{ity} \varphi^{(k)}(y) dy = (-it)^k e^{-\frac{1}{2}t^2},$$

kde $\varphi^{(k)}$ je k -tá derivace normované normální hustoty

$$\varphi(y) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}y^2}, \quad -\infty < y < \infty,$$

odpovídá pravá strana rovnice (3.15) funkci

$$\tilde{f}(y) = \varphi(y) - \frac{\gamma_1}{6}\varphi^{(3)}(y) + \frac{\gamma_2}{24}\varphi^{(4)}(y) + \frac{\gamma_1^2}{72}\varphi^{(6)}(y).$$

Vypočteme-li derivace $\varphi^{(k)}$, můžeme psát

$$\begin{aligned} \tilde{f}(y) = & (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2} \left[1 + \frac{\gamma_1}{6}(y^3 - 3y) + \frac{\gamma_2}{24}(y^4 - 6y^2 + 3) \right. \\ & \left. + \frac{\gamma_1^2}{72}(y^6 - 15y^4 + 45y^2 - 15) \right], \end{aligned}$$

což je právě Edgeworthův odhad hustoty (3.14).

3.5 Diferenciální entropie

Definice 3.24. Entropii diskrétní náhodné veličiny T s pravděpodobnostní funkcí $p_i = P(T = i) > 0$ definujeme jako

$$H(T) = - \sum_i p_i \log_2 p_i. \quad (3.16)$$

Poznámka. Definice 3.24 je ve shodě se Shannonovou formulací entropie, kde se míra entropie signálu odvozuje z odhadů pravděpodobností, s nimiž jsou jednotlivé signály pozorovány (za určitých daných podmínek). Entropie je měřena v bitech a je mírou „náhodnosti“ rozdělení (viz Cover a Thomas (1991)).

Definice 3.25. Nechť S je nosič hustoty $f(t)$. Rozšíření definice 3.24 pro spojitá rozdělení, tzv. diferenciální entropie, je dána jako

$$H(f) = - \int_S f(t) \ln f(t) dt, \quad (3.17)$$

viz Cover a Thomas (1991). Zejména kvůli usnadnění výpočtů zde použijeme přirozený logaritmus. Je vidět, že hodnota diferenciální entropie nezávisí na posunutí.

Existuje přímý vztah mezi diskrétní a diferenciální entropií. Definiční obor hustoty $f(t)$ můžeme rozdělit na intervaly o délce Δ . Předpokládejme, že hustota je na

každém dělicím intervalu spojitá. Pak existuje hodnota t_i taková, že platí

$$f(t_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(t)dt.$$

Uvažujme diskretizovanou náhodnou veličinu

$$T^\Delta = t_i \quad \text{pro} \quad i\Delta \leq T < (i+1)\Delta.$$

Entropie diskretizované náhodné veličiny je rovna

$$H(T^\Delta) = - \sum_{-\infty}^{\infty} f(t_i)\Delta \ln(f(t_i)\Delta) = - \sum_{-\infty}^{\infty} f(t_i)\Delta \ln(f(t_i)) - \ln \Delta.$$

Jestliže je hustota $f(t)$ Riemannovsky integrabilní, pak platí

$$H(T^\Delta) + \ln \Delta \rightarrow H(T) \quad \text{pro} \quad \Delta \rightarrow 0.$$

Diferenciální entropii můžeme vyjádřit také pomocí distribuční funkce $F(t) = P(T \leq t)$ (viz Vasicek (1976)) jako

$$H(f) = \int_0^1 \ln \left\{ \frac{d}{dt} F^{-1}(t) \right\} dt. \quad (3.18)$$

Přestože předpis (3.17) vypadá jako analogie vzorce (3.16), interpretace a vlastnosti diferenciální entropie se od entropie definované pro diskrétní náhodnou veličinu liší (např. diferenciální entropie může být záporná). Pro interpretaci diferenciální entropie uvažujme následující množinu hustot M :

- všechny hustoty z této množiny mají stejný nosič;
- střední hodnota je stejná pro všechny hustoty z této množiny.

Z této množiny vybereme hustotu $g(t)$ s nejvyšší diferenciální entropií. Potom můžeme říci, že vysoké hodnoty diferenciální entropie pro hustotu $f(t) \in M$ vyjadřují vysoký stupeň přiblížení rozdělení s hustotou $f(t)$ k rozdělení s hustotou $g(t)$. Přehled rozdělení s maximální entropií pro různé nosiče a fixované charakteristiky je uveden v tab. 3.3.

Tab. 3.3: Rozdělení s maximální entropií (viz Kagan et al. (1973))

Nosič	Fixace střední hodnoty výrazu	Hustota s maximální entropií
$(0, 1)$	-	$f(t) = 1$ (stejněměrné)
$(0, 1)$	$\ln T$ a $\ln(1 - T)$	Beta rozdělení
$(0, \infty)$	T	Exponenciální rozdělení
$(0, \infty)$	T a $\ln T$	Gamma rozdělení
$(-\infty, \infty)$	T^2	Normální rozdělení
$(-\infty, \infty)$	$ T $	Laplaceovo rozdělení

V tab. 3.3 vidíme, že mezi všemi rozděleními se stejnou střední hodnotou definovanými na polopřímce maximalizuje diferenciální entropii exponenciální rozdělení¹ (1). V dalším textu, pokud nebude uvedeno jinak, používáme termín „entropie“ ve smyslu „diferenciální entropie“. Přehled entropie vybraných rozdělení je shrnut v tab. 3.4.

Tab. 3.4: Přehled entropie vybraných rozdělení.

Rozdělení	Entropie
Normální	$\frac{1}{2} - \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$
Gamma	$b + (1 - b)\Psi(b) - \ln(\lambda) + \ln \Gamma(b)$
Exponenciální	$1 - \ln \lambda$
Weibullovo	$-\ln(cb) - \frac{(1-b)(0,577215+\ln(c))}{b} + 1$
Logaritmicko normální	$\ln(\sigma\sqrt{2\pi}) + \mu + \frac{1}{2}$

Pro odstranění potíží s interpretací diferenciální entropie se zavádí Kullback-Leiblerova informační vzdálenost (někdy též zvaná vzájemná či relativní entropie).

Definice 3.26. Nechť $f(t)$ resp. $g(t)$ je hustota s nosičem S_f resp. S_g . Kullback-Leiblerovu informační vzdálenost potom definujeme předpisem

$$KL(f(t), g(t)) = \begin{cases} \int_0^{\infty} f(t) \ln \frac{f(t)}{g(t)} dt & \text{pro } S_f \subseteq S_g, \\ \infty & \text{pro } S_f \not\subseteq S_g. \end{cases} \quad (3.19)$$

¹Můžeme tedy říci, že exponenciálně rozdělené mezipulsní intervaly jsou generovány „nejnáhodnějším“ způsobem. Tato situace může být interpretována jako „stav klidové komunikace“ mezi neurony. Transportována je pouze minimální informace a přijímací neuron je tak udržován ve stavu „maximální pozornosti“ nejvyšší náhodností přijímaných pulsů.

Kullback-Leiblerova vzdálenost je „mírou“ podobnosti mezi dvěma hustotami. Je nezáporná, v případě, že nosič hustoty g obsahuje nosič hustoty f je konečná² (viz Cover a Thomas (1991), Kostal a Lansky (2006)). Kullback-Leiblerova vzdálenost není metrikou, protože není symetrická.

Jak již bylo řečeno, významnou roli pro nás hraje vzdálenost hustot rozdělení definovaných na kladné polopřímce od hustoty exponenciálního rozdělení (1). Kullback-Leiblerova vzdálenost mezi hustotou exponenciálního rozdělení f_{Exp} a hustotou f libovolného rozdělení se stejným nosičem a se střední hodnotou $E(T)$ se dá vyjádřit³ jako

$$KL(f, f_{Exp}) = \lambda E(T) - \ln \lambda - H(f). \quad (3.20)$$

V případě, že střední hodnoty hustot f a f_{Exp} jsou shodné ($E(T) = E(T_{Exp}) = \frac{1}{\lambda}$), dostaneme z rovnice (3.20) vztah

$$KL(f, f_{Exp}) = 1 - \ln \lambda - H(f) = H(f_{Exp}) - H(f).$$

To znamená, že otázka „jak určit Kullback-Leiblerovu vzdálenost“ je v tomto případě redukována na výpočet diferenciální entropie. Navíc obecně bude pro hustoty f_A a f_B platit

$$KL(f_B, f_{Exp}) - KL(f_A, f_{Exp}) = H(f_A) - H(f_B).$$

Výpočet rozdílu Kullback-Leiblerových vzdáleností hustot f_A a f_B od exponenciální hustoty se tak v tomto případě stane ekvivalentní výpočtu entropií hustot f_A a f_B , popř. jejich odhadu při reálném problému.

²Při neurofyziologických aplikacích může dojít k situaci $KL(f(t), g(t)) = \infty$, když budeme srovnávat rozdělení intervalů mezi akčními potenciály neuronů s rozdílnou refrakterní periodou.

³Tento výraz můžeme interpretovat jako „míru redukce náhodnosti“ nebo „míru zisku informace“ ve srovnání s klidovou komunikací.

Kapitola 4

Odhad parametru posunutí

V této kapitole se budeme zabývat parametrickými a neparametrickými odhady parametru posunutí. Nejprve jednotlivé metody uvedeme, poté je srovnáme na základě výsledků odhadu ze simulovaných i reálných dat.

4.1 Neparametrické metody odhadu posunutí

4.1.1 Odhad minimální hodnotou

Základním neparametrickým odhadem a je nejmenší hodnota, tedy

$$\hat{a} = T_{(1)}. \quad (4.1)$$

Tento odhad je výpočetně velmi rychlý, jednoduchý, a dává asymptoticky nestranný odhad ($E(T_{(1)})$ konverguje k a pro rostoucí velikost výběru n), ale systematicky nadhodnocuje parametr posunu ($a \leq T_{(1)}$ pro všechna n). Například pro exponenciální model (1) platí

$$E(T_{(1)}) = a + \frac{1}{n\lambda},$$

nicméně

$$\lim_{n \rightarrow \infty} E(T_{(1)}) = a.$$

Výhodou odhadu (4.1) je nezávislost na použitém modelu. Na straně druhé, kvalita tohoto odhadu na něm silně závisí, zejména na tvaru levého cípu hustoty neboli pravděpodobnosti realizace náhodné veličiny blízké k a . Například u exponenciálního rozdělení (1) s parametrem $\lambda = 1$ dojde k realizaci hodnoty blízké parametru posu-

nutí s mnohem větší pravděpodobností než u Weibullova rozdělení (5) s parametry $b = 0,5$ a $c = 10$. Pro ilustraci viz tab. 5.4 na straně 95, kde jsou zobrazeny hustoty předmětných rozdělení.

4.1.2 Odhad posunutí jako součást odhadu nosiče

Pro naše účely jsme použili část metodologie pro odhad nosiče hustoty rozdělení. Odhad je možno konstruovat přímo jako odhad levé hranice nosiče, anebo jako množinu splňující předepsané vlastnosti (jako odhad parametru posunutí vezmeme její levou hranici). Ve článku Cooke (1979) je odhadnuta levá hranice nosiče jako

$$\hat{a} = 2T_{(1)} - \sum_{i=1}^n [(1 - (i-1)/n)^n - (1 - i/n)^n] T_{(i)}, \quad (4.2)$$

pro rozsáhlá pozorování autor navrhuje

$$\hat{a} = 2T_{(1)} - (e - 1) \sum_{i=1}^n e^{-i} T_{(i)}. \quad (4.3)$$

V práci Baillo et al. (2000) je nosič S odhadnut jako

$$\hat{S}_n = \bigcup_{i=1}^n I(t_i; \epsilon_n), \quad (4.4)$$

kde $I(x, r)$ značí uzavřený interval se středem v x a poloměrem r a ϵ_n je posloupnost vyhlazovacích parametrů. Tato posloupnost zde hraje analogickou roli jako parametr šířky vyhlazení v neparametrickém jádrovém vyhlazování. Na základě znalosti, že S je souvislá množina, může být vyhlazovací parametr určen následujícím způsobem: ϵ_n je minimální takový, že nosič není přerušen. Takže pro parametr ϵ_n v rovnici (4.4) platí $\epsilon_n = \max\{t_{(i+1)} - t_{(i)}\}$ a odhad posunutí dostaneme ve formě

$$\hat{a} = T_{(1)} - \epsilon_n. \quad (4.5)$$

Tento odhad může být zásadním způsobem ovlivněn odlehlými pozorováními, proto autoři doporučují pro určení parametru ϵ_n křížovou validaci:

1. Nosič $\hat{S}_{n,i}(\epsilon)$, $i = 1, \dots, n$ se odhadne pro data $t_1, \dots, t_{i-1}, t_{i+1}, \dots, t_n$ a konstantu $\epsilon > 0$.

2. Vyčíslíme počet případů kdy $t_i \notin \hat{S}_{n,i}(\epsilon)$, $i = 1, \dots, n$ a označíme jej m .
3. Nakonec určíme ϵ'_n minimalizací rozdílu $|\hat{P}_n(\epsilon) - \alpha|$, kde $\hat{P}_n(\epsilon) = m/n$ a α je malá konstanta (obvykle $\alpha = 0,05$) vzhledem k ϵ .

Pomocí této procedury dostaneme téměř spojitý nosič, a odhad posunutí je v tomto případě dán jako

$$\hat{a} = T_{(1)} - \epsilon'_n. \quad (4.6)$$

4.1.3 Korigované minimum

Jedním z našich cílů je upravit odhad minimem (4.1) zejména ve směru systematické vychýlenosti (tedy zabránit systematickému nadhodnocování skutečného parametru posunu). Přirozeně přitom požadujeme, aby nejmenším možným odhadem posunu byla 0, neboť se zabýváme pouze kladnými náhodnými veličinami. Formálně můžeme psát, že hledaný odhad bude ve tvaru

$$\hat{a} = \max\{0, T_{(1)} - \text{”korekce”}\}.$$

Pro základ nového odhadu použijeme opět minimální hodnotu $T_{(1)}$ a budeme chtít snížit její hodnotu o výraz obsahující rozptyl a počet pozorování. S rostoucím počtem pozorování n budeme požadovat, aby byl korekční člen blízký nule. Pokud vezmeme v úvahu tyto vlastnosti, uvažujeme nový odhad posunutí ve formě

$$\hat{a} = \max\left\{0, T_{(1)} - \sqrt{\frac{S^2}{n}}\right\}. \quad (4.7)$$

Vzorec (4.7) byl inspirován parametrickou procedurou pro posunuté exponenciální rozdělení, viz rovnice (4.24). Narozdíl od (4.24) byla použita \sqrt{n} , neboť takto je odhad použitelný pro širší škálu rozdělení.

4.2 Parametrické metody odhadu posunutí

4.2.1 Intervalové odhady parametru posunutí

V knize Hátle a Likeš (1974) lze nalézt intervaly spolehlivosti pro oba parametry exponenciálního rozdělení (1). Nejdříve předpokládejme, že parametr λ je znám.

Potom platí, že veličina

$$2\lambda n(T_{(1)} - a) \sim \chi^2(2)$$

a $100(1 - \alpha)\%$ intervalem spolehlivosti pro parametr posunutí a je interval

$$T_{(1)} - \frac{\chi_{1-\alpha_2}^2(2)}{2\lambda n} < a < T_{(1)} - \frac{\chi_{\alpha_1}^2(2)}{2\lambda n},$$

kde $\alpha_1 + \alpha_2 = \alpha$. Nejkratším dvoustranným $100(1 - \alpha)\%$ intervalem spolehlivosti pro parametr a je interval

$$T_{(1)} - \frac{\chi_{1-\alpha}^2(2)}{2\lambda n} < a < T_{(1)}.$$

Pokud parametr λ není znám, tak pro $2 \leq q \leq n$ platí

$$\frac{n(q-1)(T_{(1)} - a)}{v_1} \sim F(2, 2q-2),$$

kde

$$v_1 = \sum_{i=2}^q (T_{(i)} - T_{(1)}) + (n-q)(T_{(q)} - T_{(1)}).$$

$100(1 - \alpha)\%$ intervalem spolehlivosti pro parametr posunutí a je potom interval

$$T_{(1)} - \frac{v_1}{n(q-1)} F_{1-\alpha_2}(2, 2q-2) < a < T_{(1)} - \frac{v_1}{n(q-1)} F_{\alpha_1}(2, 2q-2),$$

kde opět $\alpha_1 + \alpha_2 = \alpha$. Nejkratším dvoustranným $100(1 - \alpha)\%$ intervalem spolehlivosti pro parametr a je interval

$$T_{(1)} - \frac{v_1}{n(q-1)} F_{1-\alpha}(2, 2q-2) < a < T_{(1)}.$$

Intervaly spolehlivosti pro parametry exponenciálního rozdělení zde uvádíme jako přirozený doplněk bodových odhadů prezentovaných v dalším textu. Nicméně analogické intervalové odhady nelze získat pro další rozdělení, která jsou předmětem našeho zájmu, proto se jimi dále nebudeme zabývat.

4.2.2 Robustní odhad

V práci Rousseeuw a Croux (1993) je diskutována metoda pro robustní odhady rozptýlení dat. Její aplikací můžeme získat systém rovnic pro odhad parametrů zkou-

maného rozdělení. Ve zmíněné práci se definuje tzv. mediánová absolutní odchylka jako

$$\text{mad}(T_1, \dots, T_n) = \text{med}(|T_i - \text{med}(T_1, \dots, T_n)|, i = 1, \dots, n), \quad (4.8)$$

kde $\text{med}(T_1, \dots, T_n)$ značí medián původního výběru. Tento přístup je obdobou momentového odhadu parametrů, ovšem za použití robustních charakteristik. Využijeme jej pouze pro model (1), kde jsme schopni vyjádřit

$$\text{med}(T_1, \dots, T_n) = a + \log(2)/\lambda$$

a

$$\text{mad}(T_1, \dots, T_n) \approx \frac{1}{2,0781\lambda},$$

a odhad parametru a můžeme psát ve formě

$$\hat{a} = \text{med}(T_1, \dots, T_n) - 2,0781 \log(2) \text{mad}(T_1, \dots, T_n). \quad (4.9)$$

4.2.3 Maximálně věrohodné odhady

Analyticky spočtený maximálně věrohodný odhad posunutí jsme schopni získat pouze pro model (1), a to $\hat{a} = T_{(1)}$ (viz komentář k (3.7), strana 41). V ostatních případech je třeba pro získání odhadů metodou maximální věrohodnosti použít numerických postupů.

Při vlastní numerické maximalizaci je nutná transformace parametru posunutí (nejedná se o transformaci náhodných veličin). Jedna z možných transformací je

$$a' = -\ln \left(\frac{t_{(1)} - a - \varepsilon}{t_{(1)}} \right), \quad (4.10)$$

kde ε je velmi malé číslo (nám se osvědčila hodnota 10^{-8}). Tato transformace zaručí, že budeme hledat parametr posunutí pouze na intervalu $(-\infty, t_{(1)} + \varepsilon]$, zatímco parametr a' se bude moci hledat bez omezení. Zanedbání této transformace zapříčiní, že parametr posunutí poroste v průběhu maximalizační procedury do nekonečna. Jako startovací bod pro iterativní hledání maximálně věrohodných odhadů jsme použili počáteční odhad parametrů získaný fitováním funkce příslušné hustoty pravděpodobnosti na data.

4.2.4 Odhady posunutí s minimálním rizikem

V knize Lehmann a Casella (1998) je rozpracován teoretický koncept tzv. odhadů s minimálním rizikem použitelný i pro odhad parametru posunutí. Vzhledem k tomu, že odhady s minimálním rizikem nejsou běžně používány, uvedeme zde tuto problematiku podrobněji. Nejdříve definujeme ztrátovou funkci a rizikovou funkci.

Definice 4.1. Ztrátová funkce $L(a, \hat{a})$ vyjadřuje míru závažnosti chyby při odhadu parametru a hodnotou \hat{a} . Nejčastěji používanou ztrátovou funkcí je kvadratická chyba $L(a, \hat{a}) = (a - \hat{a})^2$ či absolutní chyba $L(a, \hat{a}) = |a - \hat{a}|$.

Definice 4.2. Riziková funkce, neboli krátce riziko, je definována jako

$$R(a, d(\mathbf{T})) = E_a[L(a, d(\mathbf{T}))],$$

kde $d(\mathbf{T})$ je odhad parametru posunutí a , $\mathbf{T} = (T_1, \dots, T_n)$ a symbolem E_a značíme střední hodnotu za podmínky, že parametr posunutí je roven a .

Definice 4.3. Řekneme, že třída hustot $f(t; a)$, kde a je parametr, a ztrátová funkce $L(a, d)$, kde d je odhad parametru a , jsou invariantní vzhledem k posunutí, pokud $f(t'; a') = f(t; a)$ resp. $L(a, d) = L(a', d')$, přičemž

$$t' = t + c, \tag{4.11}$$

$$a' = a + c \tag{4.12}$$

a

$$d' = d + c \tag{4.13}$$

pro všechna $c \in \mathbb{R}$. Pokud je invariantní vzhledem k posunutí hustota i ztrátová funkce, potom řekneme, že problém odhadu parametru a je invariantní k posunutí vzhledem k transformacím (4.11), (4.12) a (4.13).

Definice 4.4. Odhad d splňující

$$d(T_1 + c, \dots, T_n + c) = d(T_1, \dots, T_n) + c \quad \text{pro } \forall c \in \mathbb{R}$$

nazveme ekvivariantním vzhledem k transformacím (4.11), (4.12) a (4.13) nebo též lokačně ekvivariantním.

Příklad 4.5. Odhad parametru posunutí minimem (4.1) je ekvivariantním odhadem, neboť platí $\min\{T_1 + c, \dots, T_n + c\} = \min\{T_1, \dots, T_n\} + c$ pro všechna $c \in \mathbb{R}$.

Lemma 4.6. Ztrátová funkce L splňuje

$$L(a', d') = L(a, d) \quad (4.14)$$

pro všechna a právě tehdy, pokud závisí pouze na diferencích $d - a$, tedy

$$L(a, d) = \rho(d - a), \quad (4.15)$$

kde ρ je vhodná funkce.

Důkaz. Je zřejmé, že z (4.15) plyne (4.14). Opačnou implikaci dokážeme položením $a = -a$ v (4.14) a stanovením $\rho(d - a) = L(0, d - a)$. \square

Věta 4.7. Nechť $\mathbf{T} = T_1, \dots, T_n$ má simultánní hustotu

$$f(\mathbf{t} - a) = f(t_1 - a, \dots, t_n - a), \quad -\infty < a < \infty, \quad (4.16)$$

kde funkce f je známá a a je neznámý parametr posunutí. Dále nechť d je ekvivariantním odhadem parametru a se ztrátovou funkcí (4.15). Potom vychýlení, riziko a rozptyl odhadu δ jsou konstanty (tj. nezávisí na a).

Důkaz. Pokud má \mathbf{T} hustotu $f(\mathbf{t})$ (tj. $a = 0$), pak $\mathbf{T} + a$ má hustotu (4.16). Potom můžeme psát vychýlení jako

$$b(a) = E_a[d(\mathbf{T})] - a = E_0[d(\mathbf{T} + a)] - a = E_0[d(\mathbf{T})],$$

což nezávisí na a . Důkazy pro riziko a rozptyl jsou analogické. \square

Definice 4.8. Pokud při problému odhadu invariantního vzhledem k posunutí ve smyslu definice 4.3 existuje lokačně ekvivariantní odhad, který minimalizuje riziko, potom jej nazveme odhadem s minimálním rizikem (MRE).

Lemma 4.9. Pokud je d_0 libovolný ekvivariantní odhad, potom nutnou a dostatečnou podmínkou pro ekvivarianci odhadu d je

$$d(\mathbf{t}) = d_0(\mathbf{t}) + u(\mathbf{t}), \quad (4.17)$$

kde $u(\mathbf{t})$ je funkce splňující

$$u(\mathbf{t} + c) = u(\mathbf{t}) \quad \text{pro } \forall \mathbf{t}, c. \quad (4.18)$$

Důkaz. Nejdříve předpokládejme, že platí (4.17) a (4.18). Potom $d(\mathbf{t} + c) = d_0(\mathbf{t} + c) + u(\mathbf{t} + c) = d_0(\mathbf{t}) + c + u(\mathbf{t}) = d(\mathbf{t}) + c$, tedy d je ekvivariantní odhad.

Naopak, pokud je d ekvivariantní odhad, nechť

$$u(\mathbf{t}) = d(\mathbf{t}) - d_0(\mathbf{t}).$$

Potom

$$\begin{aligned} u(\mathbf{t} + c) &= d(\mathbf{t} + c) - d_0(\mathbf{t} + c) \\ &= d(\mathbf{t}) + c - d_0(\mathbf{t}) - c = u(\mathbf{t}) \end{aligned}$$

takže (4.17) a (4.18) platí. □

Lemma 4.10. Funkce u splňuje (4.18) tehdy a jen tehdy, pokud je funkcí diferencí $y_i = t_i - t_n$, $i = 1, \dots, n - 1$, $n \geq 2$; pro $n = 1$ tehdy a jen tehdy pokud je to konstantní funkce.

Důkaz. Postup je v podstatě stejný jako u Lemmatu 4.6. □

Na základě předchozích dvou Lemmat můžeme formulovat následující větu:

Věta 4.11. Pokud je d_0 libovolný ekvivariantní odhad, potom nutnou a dostatečnou podmínkou pro ekvivarianci odhadu d je existence funkce v o $n - 1$ parametrech, pro kterou platí

$$d(\mathbf{t}) = d_0(\mathbf{t}) - v(\mathbf{y}) \quad \text{pro } \forall \mathbf{t}.$$

Věta 4.12. Nechť vektor $\mathbf{T} = (T_1, \dots, T_n)$ má rozdělení s hustotou (4.16), dále nechť $Y_i = T_i - T_n$, $i = 1, \dots, n - 1$ a $\mathbf{Y} = (Y_1, \dots, Y_n)$. Předpokládejme, že ztrátová funkce je dána vztahem (4.15) a existuje ekvivariantní odhad d_0 parametru a s konečným rizikem. Dále předpokládejme, že pro každé \mathbf{y} existuje číslo $v(\mathbf{y}) = v^*(\mathbf{y})$, které minimalizuje výraz

$$E_0(\rho(d_0(\mathbf{T}) - v(\mathbf{y}))|\mathbf{y}). \quad (4.19)$$

Potom lokačně ekvivariantní odhad d parametru a existuje a je dán vztahem

$$d(\mathbf{T}) = d_0(\mathbf{T}) - v^*(\mathbf{Y}). \quad (4.20)$$

Důkaz. Dle Věty 4.11 se ekvivariantní odhad s minimálním rizikem nalezne určením funkce v minimalizující

$$R_a(d) = E_a(\rho(d_0(\mathbf{T}) - v(\mathbf{y}) - a)).$$

Protože je riziková funkce na parametru a nezávislá, je dostačující minimalizovat

$$\begin{aligned} R_0(d) &= E_0(\rho(d_0(\mathbf{T}) - v(\mathbf{y}))) \\ &= \int E_0(\rho(d_0(\mathbf{T}) - v(\mathbf{y}))|\mathbf{y})dP_0(\mathbf{y}). \end{aligned}$$

Integrál je minimalizován minimalizací integrandu, tedy vlastě (4.19), pro každé \mathbf{y} . Protože má odhad d_0 konečné riziko $E_0(\rho_0(\mathbf{T})|\mathbf{y}) < \infty$ skoro všude, je minimalizace (4.19) smysluplná. Konečné tvrzení nyní plyne přímo z předpokladů věty. \square

Poznámka. Za předpokladů Věty 4.12 platí, že pro $\rho(d - a) = (d - a)^2$ je

$$v^*(\mathbf{y}) = E_0(d_0(\mathbf{T})|\mathbf{y}) \quad (4.21)$$

a ekvivariantní odhad s minimálním rizikem je dán jako

$$d(\mathbf{T}) = d_0(\mathbf{T}) - E_0(d_0(\mathbf{T})|\mathbf{Y}). \quad (4.22)$$

Věta 4.13. Za předpokladů Věty 4.12, s $L(a, d) = (d - a)^2$, můžeme odhad (4.22) vyjádřit jako

$$d(\mathbf{t}) = \frac{\int_{-\infty}^{\infty} u f(t_1 - u, \dots, t_n - u) du}{\int_{-\infty}^{\infty} f(t_1 - u, \dots, t_n - u) du}. \quad (4.23)$$

Tento vztah se nazývá Pitmanův odhad parametru a .

Důkaz. Nechť $d_0(\mathbf{T}) = T_n$. Pro výpočet $E_0(T_n|\mathbf{y})$ položíme

$$y_i = t_i - t_n (i = 1, \dots, n - 1); \quad y_n = t_n.$$

Jakobián této transformace je roven jedné. Simultánní hustota (Y_1, \dots, Y_n) je potom

$$p_Y(y_1, \dots, y_n) = f(y_1 + y_n, \dots, y_{n-1} + y_n, y_n)$$

a podmíněná hustota Y_n za daného $\mathbf{y} = (y_1, \dots, y_{n-1})$ je

$$\frac{f(y_1 + y_n, \dots, y_{n-1} + y_n, y_n)}{\int f(y_1 + k, \dots, y_{n-1} + k, k) dk}.$$

Z toho plyne, že

$$E_0(T_n | \mathbf{y}) = E_0(Y_n | \mathbf{y}) = \frac{\int k f(y_1 + k, \dots, y_{n-1} + k, k) dk}{\int f(y_1 + k, \dots, y_{n-1} + k, k) dk}.$$

Tento výraz můžeme vyjádřit pomocí t_i jako

$$E_0(T_n | \mathbf{y}) = \frac{\int k f(t_1 - t_n + k, \dots, t_{n-1} - t_n + k, k) dk}{\int f(t_1 - t_n + k, \dots, t_{n-1} - t_n + k, k) dk}$$

nebo po transformaci $u = t_n - k$ jako

$$E_0(T_n | \mathbf{y}) = t_n - \frac{\int u f(t_1 - u, \dots, t_n - u) du}{\int f(t_1 - u, \dots, t_n - u) du},$$

což kompletuje důkaz. □

Poznámka. Integrály v Pitmanově vzorci je pro většinu hustot velmi náročné či nemožné vyřešit, obvykle je nutno použít numerického řešení.

Pro exponenciální model (1) se známým parametrem λ můžeme odvodit vzorec pro odhad parametru posunutí přímo ze vzorce (4.20). Pokud položíme $d_0 = T_{(1)}$, z Basuovy věty (viz Lehmann a Casella (1998), strana 42) plyne, že $v(y) = v$ je určeno minimalizací $E_0[L_1(T_{(1)} - v)]$. Pokud je ztrátovou funkcí kvadratická chyba, výsledkem minimalizace je $v = E_0[T_{(1)}]$. Nyní spočteme

$$E_0[T_{(1)}] = \int_0^{\infty} t \lambda n e^{-\lambda n(t-0)} dt = \frac{1}{\lambda n},$$

a dostaneme odhad parametru posunutí s minimálním rizikem jako

$$\hat{a} = T_{(1)} - \frac{1}{n\lambda}. \tag{4.24}$$

Situace kdy jsou parametry rozdělení kromě parametru posunutí známy je spíše výjimečná. V knize Lehmann a Casella (1998) je diskutován i případ, že nejsou známy žádné parametry rozdělení, jehož posun odhadujeme. Pro exponenciální model (1), kde parametr λ není znám, je uveden odhad s minimálním rizikem

$$\hat{a} = T_{(1)} - \frac{1}{n^2} \sum_{i=1}^n (T_i - T_{(1)}). \quad (4.25)$$

Pro gamma rozdělení (2) s neznámými parametry λ a b není odhad s minimálním rizikem dostupný. Dokonce ani pro známé parametry λ a b jej není možné analyticky vyjádřit. Je nutno použít vztah (4.23) a integrály v něm vystupující vyhodnotit numericky pro partikulární hodnoty parametrů λ a b .

4.3 Porovnání jednotlivých metod odhadu posunutí

Nyní budeme testovat výše uvedené metody odhadu posunutí na simulovaných datech a reálných neurofyziologických datech. Simulovaná data jsme získali následujícím způsobem: Generovali jsme 100 sérií simulací různého rozsahu pozorování n (50, 100, 250, 500, 1000) modelů (1), (2) a (2.4). Parametry neposunutých verzí modelů (1) a (2) jsme stanovili tak, aby střední hodnota mezipulsních intervalů byla rovna 100 ms (pro exponenciální model $\lambda = 0,01$ a pro gamma model $\lambda = 0,025$ a $b = 2,5$). Pro Ornstein-Uhlenbeckův model (2.4) jsme položili $\phi = 20$ ms, $S = 10$ mV, $\sigma = 1 \sqrt{\text{mV}}/\text{ms}$. Dále jsme stanovili dvě úrovně parametru μ , a to $\mu_1 = 0,50$ mV/ms a $\mu_2 = 0,65$ mV/ms. Odpovídající střední hodnoty realizací byly aproximativně 36 ms a 24 ms. Ke každé generované sérii jsme přičetli posunutí o velikosti 5 % střední hodnoty.

Jako kritérium kvality odhadu jsme zvolili střední kvadratickou chybu (MSE)

$$\frac{1}{100} \sum_{i=1}^{100} (\hat{a}_i - a)^2,$$

kde \hat{a}_i je odhad posunutí a získaný pro i -tou sérii. Dále jsme počítali střední chybu (ME)

$$\frac{1}{100} \sum_{i=1}^{100} (\hat{a}_i - a)$$

pro ilustraci případného vychýlení odhadu. Zvláště jsme ověřovali parametrické a neparametrické metody.

4.3.1 Parametrický přístup

Při apriorní znalosti rozdělení dat v případě exponenciálního modelu (1) nebo gamma modelu (2) můžeme odhadovat parametr posunutí pomocí metody maximální věrohodnosti či pomocí odhadů s minimálním rizikem. Pro exponenciální model můžeme použít robustní odhad (4.9), neboť pro tento model jsme schopni stanovit soustavu rovnic řešitelnou vzhledem k parametru posunutí. Dále budeme testovat, zda konkrétní předpoklad o rozdělení aplikovaný na data generovaná v souladu s Ornstein-Uhlenbeckovým modelem povede k použitelným výsledkům či nikoliv. V práci Reeke a Coop (2004) se uvádí, že tento model lze vyhovujícím způsobem aproximovat pomocí gamma rozdělení (2).

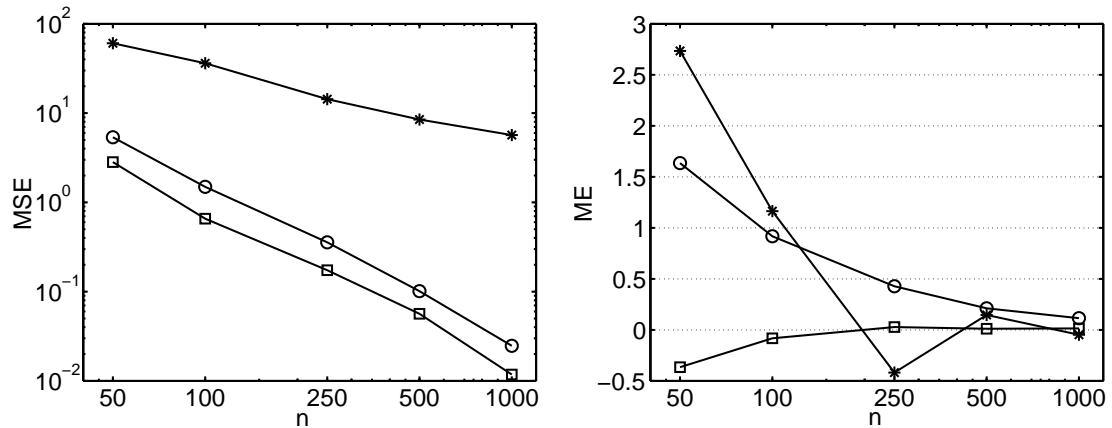
K parametrickým odhadům lze obvykle spočítat několik doprovodných charakteristik určených ke zhodnocení odhadu, například vydatnost. V případě odhadu posunutí jsou výpočty sice technicky proveditelné, ale výsledky jsou nepoužitelné, protože zde nejsou splněny podmínky regularity (viz např. Anděl (1978)). Konkrétně se jedná o podmínku, že odhadovaný parametr neovlivní nosič hustoty pravděpodobnosti. Pro rozdělení definovaná na polopřímce je parametr posunutí přímo levou hranicí nosiče hustoty.

Exponenciální model

Maximálně věrohodný odhad parametru posunutí je v tomto případě dán vztahem (3.7). Tento odhad nezávisí na parametru λ . Oproti tomu při odhadu s minimálním rizikem dostaneme různé odhady pro případ, že parametr λ je znám a pro případ že parametr λ neznáme. Pro námi analyzovaná data jsme získali v obou případech téměř stejné výsledky (s přesností na 3 platné číslice).

Výsledky všech tří testovaných metod jsou srovnány na obr. 4.1. Vidíme, že odhad s minimálním rizikem je lepší než maximálně věrohodný odhad pro všechna n a že robustní odhad je podstatně horší než oba zbývající. Tedy robustnost je kompenzována nižší efektivitou. Metoda odhadu s minimálním rizikem dává odhady nejen s nejnižší MSE, ale je také méně vychýlená než obě zbývající metody. Na obr. 4.1 vidíme, že pouze relativně velký počet pozorování zajistí v tomto případě kvalitní odhad, což ale nestačí pro dostatečně kvalitní robustní odhad. Můžeme shrnout, že

pokud víme, že data pochází z exponenciálního rozdělení, je odhad s minimálním rizikem nejvhodnějším přístupem pro odhad parametru posunutí.



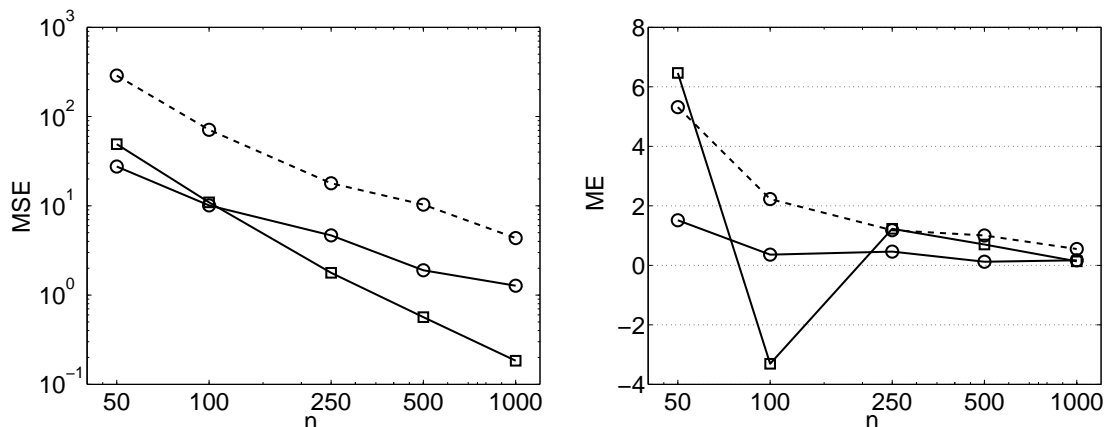
Obr. 4.1: Srovnání odhadů parametru posunutí v případě předpokládaného exponenciálního modelu (1) známého až na parametry, MSE v [ms²] (levý graf) a ME v [ms] (pravý graf), ○ maximálně věrohodný odhad, □ odhad s minimálním rizikem a * robustní odhad. Správná hodnota posunutí je $a = 5$ ms.

Gamma model

Ve srovnání s exponenciálním modelem máme u gamma modelu opačnou situaci. Odhad s minimálním rizikem pro neznámé parametry λ a b není dostupný, navíc pro posunuté gamma rozdělení (2) se známými parametry λ a b není možné analyticky vyjádřit odhad analogický k odhadu získanému pro posunuté exponenciální rozdělení. Je nutno použít vztah (4.23) a integrály v něm vystupující vyhodnotit numericky pro partikulární hodnoty parametrů λ a b .

Maximálně věrohodný odhad lze odvodit v obou situacích – pro známé i neznámé parametry λ a b . Porovnání jednotlivých odhadů je ilustrováno na obr. 4.2. Jak lze očekávat, maximálně věrohodný odhad se známými parametry λ a b je vždy lepší než maximálně věrohodný odhad s neznámými parametry λ a b . S rostoucím rozsahem výběru n mají odhady s minimálním rizikem viditelně menší MSE než maximálně věrohodné odhady se známými parametry λ a b . Na druhou stranu, vychýlení odhadů s minimálním rizikem je větší než pro maximálně věrohodné odhady se známými parametry λ a b , zejména pro malý rozsah výběru. Na obr. 4.2 můžeme vidět pokles MSE analogický jako v případě exponenciálního modelu, viz obr. 4.1. Obecně můžeme říci, že metody dávající nejlepší výsledky odhadu parametru posunutí jsou

v tomto případě založeny na apriorní znalosti zbývajících parametrů, což není příliš reálné.



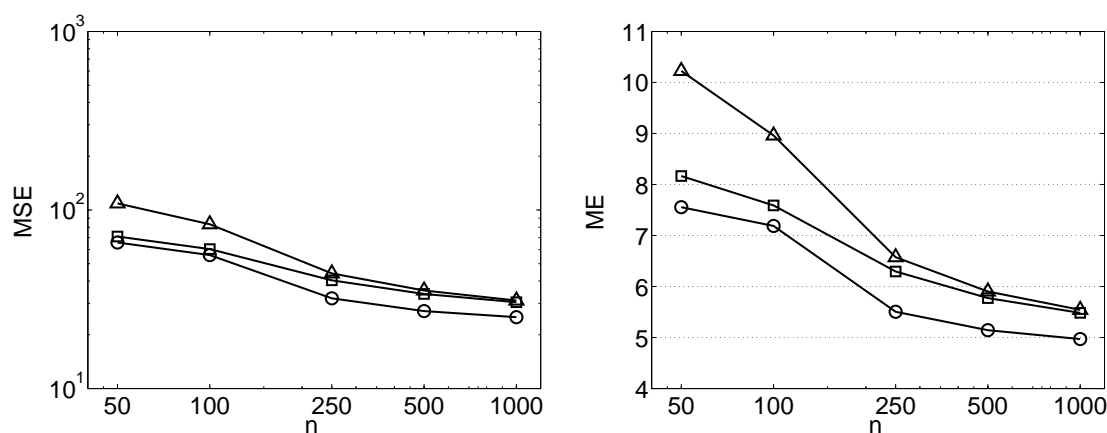
Obr. 4.2: Srovnání odhadů parametru posunutí v případě předpokládaného gamma modelu (2) známého až na parametry, MSE v [ms²] (levý graf) a ME v [ms] (pravý graf), ○ plná čára – maximálně věrohodný odhad se známými parametry $\lambda = 0,025$ a $b = 2,5$, ○ čárkovaná čára – maximálně věrohodný odhad s neznámými parametry λ a b a □ odhad s minimálním rizikem se známými parametry λ a b . Správná hodnota posunutí je $a = 5$ ms.

Ornstein-Uhlenbeckův model

Data pro následující analýzu byla generována ve shodě s Ornstein-Uhlenbeckovým modelem v nadprahovém režimu ($\mu\phi > S$) a v prahovém režimu ($\mu\phi = S$). Odhady posunutí jsou založeny na „chybném“ předpokladu, že data pocházejí z exponenciálního (1), alfa (3) nebo gamma (2) rozdělení. Dle předpokladů (viz obr. 4.3, $\mu = 0,65$ mV/ms) gamma model dává odhady se systematicky nižší MSE a ME než odhady získané fitováním alfa či exponenciálního rozdělení. Odhady dosažené fitováním exponenciálního rozdělení jsou vzhledem k ME podstatně horší než oba ostatní, zejména pro malé rozsahy výběru.

Poznamenejme, že ve všech diskutovaných situacích jsou hodnoty MSE a ME relativně vysoké. Je to zapříčiněno skutečností, že data generovaná s naší volbou parametrů Ornstein-Uhlenbeckova modelu jsou „přirozeně posunutá“¹. Vše je patrné

¹To znamená, že prakticky nejsou generovány relativně krátké mezipulsní intervaly ani v situaci bez posunutí. Naše volba parametrů je interpretovatelná a akceptovatelná z neurofyziologického hlediska, měla by tedy vést k simulacím relativně blízkým reálným datům. Je možno najít takové parametry Ornstein-Uhlenbeckova modelu, aby byl uvedený problém eliminován, ovšem jejich interpretace bude nesmyslná.



Obr. 4.3: Srovnání odhadů parametru posunutí v případě, kdy data byla generována ve shodě s Ornstein-Uhlenbeckovým modelem (2.4) s parametrem $\mu = 0,65$ mV/ms a fitována pomocí různých rozdělení, MSE v [ms²] (levý graf) a ME v [ms] (pravý graf), □ maximálně věrohodný odhad s alfa rozdělením, ○ maximálně věrohodný odhad s gamma rozdělením, Δ maximálně věrohodný odhad s exponenciálním rozdělením. Správná hodnota posunutí je $a = 1,2$ ms.

na pravém grafu obr. 4.3, kde ME konverguje k hodnotě pět namísto k nule jako na obr. 4.1 a 4.2. Můžeme tedy konstatovat, že parametrické modely posun výrazně nadhodnocují. Situace při odhadu posunu na simulacích Ornstein-Uhlenbeckova modelu s parametrem $\mu = 0,50$ mV/ms je téměř identická (na grafech není viditelný rozdíl, proto je neuvádíme). Oproti práci Reeke a Coop (2004) ukazujeme, že gamma rozdělení není uspokojivou náhradou Ornstein-Uhlenbeckova modelu vzhledem k odhadu posunutí.

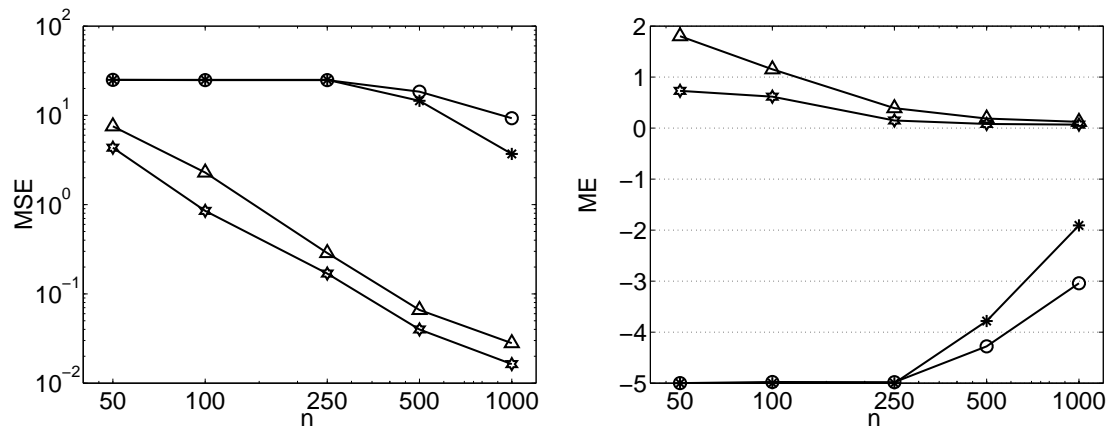
4.3.2 Neparametrický přístup

Situace, kdy nemáme žádné informace o rozdělení dat, je obvyklá. K ověření kvality neparametrických metod odhadu posunutí použijeme stejné série dat jako v minulé podkapitole.

Exponenciální model

Na obr. 4.4 lze vidět, že metody odhadu se vzhledem k MSE dělí do dvou skupin. Skupina metod s viditelně nižší MSE zahrnuje Cookův odhad a odhad minimem. Vzhledem k tomu, že odhad minimem je v tomto případě vlastně maximálně věrohodným odhadem, je kvalita Cookova odhadu překvapující. Druhá skupina me-

tod, tj. korigované minimum a Baillův odhad, má viditelně vyšší MSE i ME než první skupina metod. Ukazuje se tedy, že „korekce“ minimální hodnoty nevede vždy k lepšímu výsledku. Zde tkví příčina v exponencialitě analyzovaných dat. Srovnáním obr. 4.1 a 4.4 se ukazuje, že apriorní znalost exponenciálního rozdělení nepřináší pro odhad posunutí významný profit.



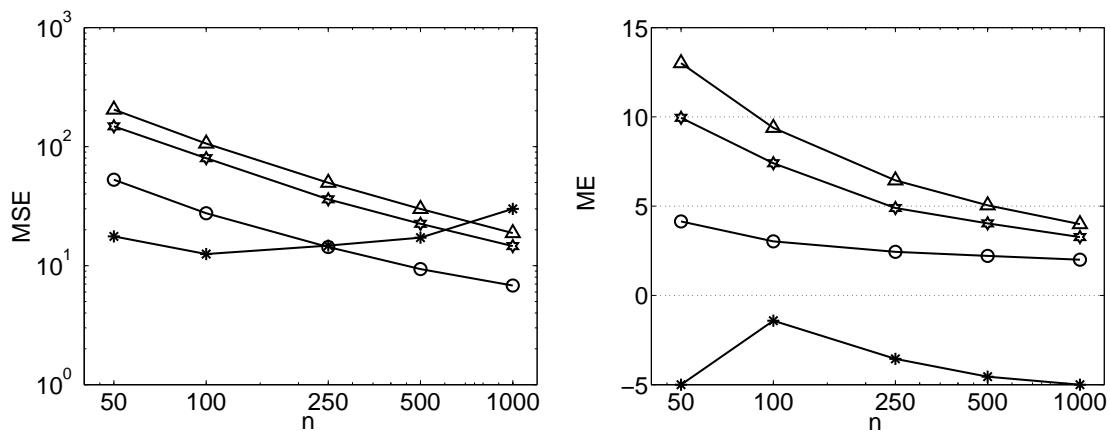
Obr. 4.4: Srovnání neparametrických odhadů parametru posunutí pro data pocházející z exponenciálního modelu (1), MSE v [ms²] (levý graf) a ME v [ms] (pravý graf), Δ minimum, ○ vylepšené minimum, ☆ Cookův odhad a * Baillův odhad. Správná hodnota posunutí je $a = 5$ ms.

Gamma model

Výsledky neparametrického odhadu parametru posunutí pro gamma rozdělení jsou ilustrovány na obr. 4.5. Vidíme, že kvalita odhadu korigovaným minimem, Cookeovým odhadem a minimem se s vyšším počtem pozorování vylepšuje. Korigované minimum je zde prakticky nejlepším odhadem. Baillův odhad dává dobré výsledky pro malé rozsahy výběru, nicméně s rostoucím n se odhad překvapivě horší. Žádná z metod nedává příliš přesné odhady, a to ani při velkých rozsazích výběru, jak můžeme vidět ze srovnání obr. 4.5 s obr. 4.2.

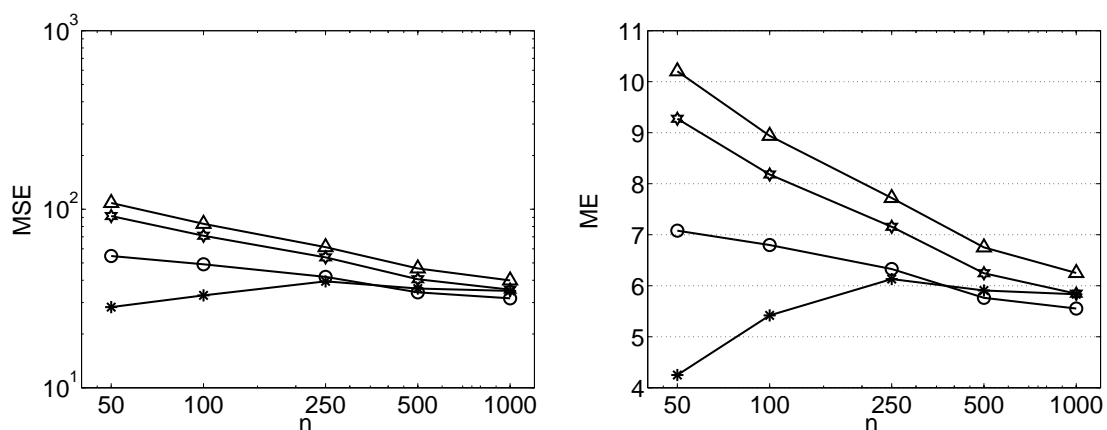
Ornstein-Uhlenbeckův model

Na obr. 4.6 jsou vykresleny MSE a ME neparametrických metod odhadu posunutí u Ornstein-Uhlenbeckova modelu (2.4) s parametrem $\mu = 0,65$ mV/ms. Nejnižších hodnot MSE je dosaženo pro Baillův odhad v případě malého rozsahu výběru. Přestože analyzované metody dávají pro malý počet pozorování odlišné hodnoty



Obr. 4.5: Srovnání neparametrických odhadů parametru posunutí pro data pocházející z gamma modelu (2), MSE v $[\text{ms}^2]$ (levý graf) a ME v $[\text{ms}]$ (pravý graf), Δ minimum, \circ vylepšené minimum, \star Cookeův odhad a $*$ Baillův odhad. Správná hodnota parametru posunutí je $a = 5$ ms.

MSE, s rostoucím rozsahem výběru n dostaneme prakticky stejnou MSE i ME. Opět, stejně jako v případě gamma modelu a zejména pro velké rozsahy výběru, nejsou výsledky uspokojivé, ale naopak jsou srovnatelné s výsledky získanými pomocí neadekvátně použitých parametrických metod (srovnejme obr. 4.6 a obr. 4.3). Stejně jako v předcházející podkapitole zde není podstatných rozdílů v průběhu MSE a ME odhadu posunutí pro Ornstein-Uhlenbeckův model s parametrem $\mu = 0,50$ mV/ms a s parametrem $\mu = 0,65$ mV/ms.



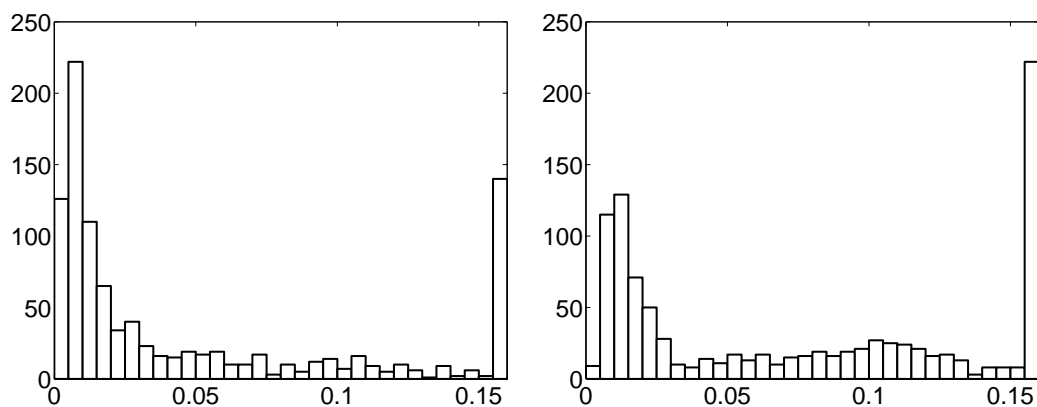
Obr. 4.6: Srovnání neparametrických odhadů parametru posunutí pro data pocházející z Ornstein-Uhlenbeckova modelu (2.4) s parametrem $\mu = 0,65$, MSE v $[\text{ms}^2]$ (levý graf) a ME v $[\text{ms}]$ (pravý graf), Δ minimum, \circ vylepšené minimum, \star Cookeův odhad a $*$ Baillův odhad. Správná hodnota parametru posunutí je $a = 1,2$ ms.

4.3.3 Analýza reálných neurofyzilogických dat

Nejprve ukážeme asymptotické vlastnosti metod odhadu a poté se budeme zajímat o rozdíly mezi odhady posunutí z dat získaných za působení určitého stimulu a bez něj.

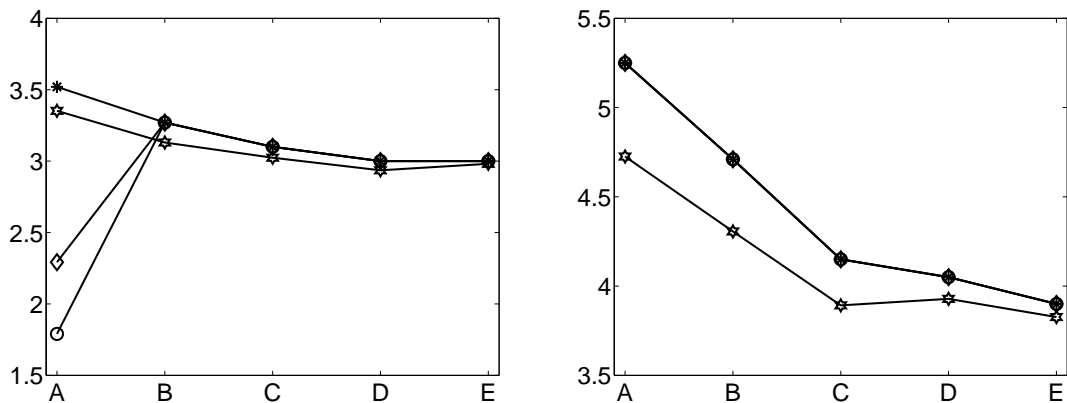
Data o velkém rozsahu

Pro analýzu dlouhodobě zaznamenávaných dat jsme použili dva příklady aktivity hippocampálních CA1 místních buněk zaznamenávaných při pohybu zvířete hledajícího potravu ve standardizovaném kruhovém válci. Data byla poskytnuta Prof. Fentonem, jejich záznam byl proveden standardními technikami (viz Fenton et al. (2000)). Obě série dat obsahovaly původně 1000 hodnot. Postupně jsme je rozdělili na 2, 4, 10 a 20 stejně dlouhých intervalů, ze kterých jsme odhadovali posunutí. Všechna data jsme testovali na stacionaritu v tom smyslu, že průměr a odchylka se časem neměnily. Při testu exponenciality byla tato zamítnuta na hladině $\alpha = 0,05$ v obou případech, použili jsme Kolmogorova-Smirnova testu. Histogramy obou sérií dat jsou zobrazeny na obr. 4.7.



Obr. 4.7: Histogramy datových sérií o velkém rozsahu.

Výsledky odhadu jsou prezentovány jako průměr odhadnutého posunu z adekvátních částí dat. Odhady (obr. 4.8) jsme získali pouze pro Cookeovu metodu (4.2), minimum (4.1) a numericky vyhodnocenou metodu maximální věrohodnosti pro exponenciální a gamma modely (1) a (2). Zbývající metody odhadu diskutované výše daly nulové či záporné odhady parametru posunutí, což je v našem případě neakceptovatelné.



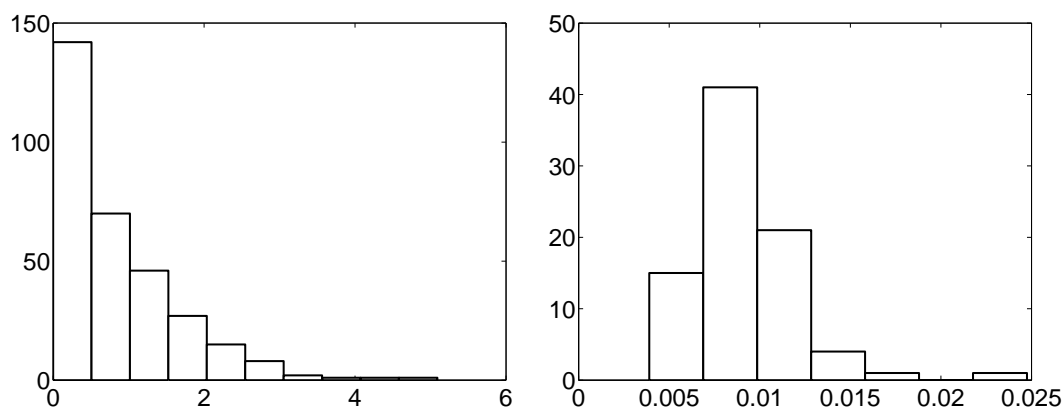
Obr. 4.8: Odhad refrakterní periody z první série dat (levý graf) a z druhé série dat (pravý graf). Osa y: odhadnutá refrakterní perioda v [ms] – průměrné odhady získané pomocí výsledků z jednotlivých částí dat. Osa x: data rozdělená na A – 20 částí, B – 10 částí, C – 4 částí, D – 2 částí, E – kompletní série. Symbol * označuje minimum, ☆ Cookeův odhad, ◇ maximálně věrohodný odhad pro exponenciální rozdělení (numericky) a ○ maximálně věrohodný odhad pro gamma rozdělení (numericky).

Na grafech vidíme, že minimum a numericky spočtené maximálně věrohodné odhady pro exponenciální a gamma rozdělení dávají téměř totožné výsledky. Jediná výjimka je patrná u první série dat rozdělených na 20 částí, kde jsou oba numericky spočtené maximálně věrohodné odhady výrazně menší než minimum. Tato odchylka je však dána spíše numerickými než statistickými příčinami. Výsledky Cookeova odhadu jsou systematicky nižší než výsledky zbývajících metod.

Spontánní a stimulovaná data

V tomto odstavci odhadujeme posunutí ze spontánních dat (312 pozorování) a stimulovaných dat (83 pozorování). Histogramy těchto datových sérií jsou zobrazeny na obr. 4.9. Data pochází z intracelulárního záznamu sluchových sensorických neuronů morčete (details viz Lansky et al. (2006)). Zamítli jsme exponencialitu pro stimulovanou posloupnost časových intervalů, avšak exponencialita nebyla zamítnuta pro spontánní data. Výsledky odhadu posunutí z obou záznamů jsou shrnuty v tab. 4.1.

Můžeme zde vidět, že pro mezipulsní intervaly zaznamenané během spontánní aktivity je odhadnutý posun významně větší než je reálné trvání refrakterní periody. Tento výsledek prezentuje omezení všech diskutovaných metod. Není překvapivé, podobné výsledky jsme získali z předchozích simulací, konkrétní ilustrace viz obr. 4.1, kde jsme se zabývali právě exponenciálními daty. Výsledek získaný pomocí alfa



Obr. 4.9: Histogram spontánních dat (levý graf) a histogram stimulovaných dat (pravý graf).

Tab. 4.1: Výsledky odhadu posunutí (v [ms]) pro spontánní a stimulovanou posloupnost časových intervalů. Gamma – maximálně věrohodný odhad pro gamma model, Alfa – maximálně věrohodný odhad pro alfa model, Minimum – odhad minimální hodnotou (4.1) (tj. maximálně věrohodný odhad pro exponenciální model), Kor. minimum – odhad (4.7), Cooke – Cookeův odhad (4.2), Baillo – Baillův odhad (4.6) a Robust – robustní odhad parametru posunutí exponenciálního rozdělení (4.9).

Data	n	Gamma	Alfa	Minimum	Kor. minimum	Cooke	Baillo	Robust
Spontánní	312	88,2	0	88,5	44,9	86,4	48,5	84,9
Stimulovaná	83	2,5	3,8	3,9	3,6	3,8	0	6,2

modelu je sice odlišný, avšak opět nereálný, což je dáno exponenciálním charakterem dat, který alfa model nemůže vystihnout.

Celá situace se zásadně změní při analýze stimulovaných dat. S výjimkou Baillova odhadu, který je záporný, je třeba se rozhodnout mezi třemi přijatelnými úrovněmi odhadu posunutí. Na základě našich výsledků na simulacích Ornstein-Uhlenbeckova modelu, který by měl být blízký realitě, bychom nejspíše přijali odhad korigovaným minimem. Nicméně lze říci, že odhady jsou si velmi podobné.

Kapitola 5

Odhady diferenciální entropie

K odhadu entropie můžeme přistoupit více strategiemi. Odhad entropie může být založený na parametrické či neparametrické metodě. Jiné dělení může vymezit metody přímé, kdy máme přímý vztah pro odhad entropie, a na metody nepřímé (tzv. plug-in odhady), kdy nejdříve musíme odhadnout hustotu či distribuční funkci a až poté entropii. Dalším možným dělením jsou metody založené na odhadu hustoty popř. z něj odvozené, metody založené na odhadu distribuční funkce a na ostatní metody.

Při odhadu entropie je také možno vyjít ze vztahu mezi entropií (3.16) a diferenciální entropií (3.17). Můžeme použít aproximativní techniky odhadu diferenciální entropie založené na vzorkování (viz Paninsky (2003)). Tento přístup používat nebudeme, zaměříme se na metody určené přímo pro odhad diferenciální entropie.

Při užití parametrického odhadu na základě předpokladu o rozdělení dat odhadujeme parametry $\theta_1, \dots, \theta_k$ tohoto rozdělení. Pomocí vhodné metody (většinou metoda maximální věrohodnosti) dostaneme odhad parametrů $\hat{\theta}_1, \dots, \hat{\theta}_k$. Jejich užitím a na základě rovnice (3.17) spočteme odhad entropie analyticky či numericky z rovnice

$$\hat{H}(T) = - \int_S \hat{f}(t) \ln \hat{f}(t) dt, \quad (5.1)$$

kde $\hat{f}(t)$ je odhad hustoty s nosičem S v případě $f(t; \hat{\theta}_1, \dots, \hat{\theta}_k)$. Odhad entropie bude vyhovující v situaci, kdy předpoklad o rozdělení byl správný a bude se zlepšovat v závislosti na rostoucím počtu pozorování.

V případě, že nemůžeme vyslovit žádný předpoklad o rozdělení, použijeme některou z neparametrických metod a místo parametrů rozdělení budeme odhadovat přímo hustotu. Odhad entropie (přesněji plug-in odhad entropie) se poté získá opět pomocí vztahu (5.1).

V dalším textu budou krátce a obecně zmíněny metody parametrické, podrobně rozebrány některé plug-in metody a nakonec uvedeny některé přímé metody. Pro každý typ odhadu uvedeme zhodnocení provedené na základě simulovaných dat. Odhadnutá entropie se bude srovnávat s teoretickou entropií použitých rozdělení, přehled je uveden v tab. 5.1.

Tab. 5.1: Teoretická entropie rozdělení použitých pro hodnocení metod odhadu entropie. N - normální rozdělení (4), E - exponenciální rozdělení (1), W - Weibullovo rozdělení (5), G - gamma rozdělení (2).

N(0;1)	1,4189	E(0,3)	2,2040	W(5;2)	-0,2093	G(5;20)	1,2905
N(0;0,16)	0,5026	E(1)	1,0000	W(0,5;10)	-0,7138	G(7;2)	-0,3687
N(0;9)	2,5176	E(10)	-1,3026	W(2;2)	0,2489	G(0,6;2)	2,0880

5.1 Odhad entropie na základě histogramu

V této kapitole se zaměříme na odhad entropie pomocí histogramu (3.8). Optimální volba vnitřních parametrů histogramu (okrajů a délek dělicích intervalů) z hlediska odhadu entropie závisí jak na rozsahu výběru n , tak na typu rozdělení a jeho parametrech. Pro pevně zvolený počet dělicích intervalů m roste s rostoucí šířkou intervalu $[z_1, z_{m+1}]$ odhad entropie, neboť se více přiblížíme rovnoměrnému spojitému rozdělení, které má na daném intervalu maximální entropii.

Na základě simulací jsme spočetli odhady entropie z různě konstruovaných histogramů. Při hodnocení odhadu entropie z histogramu rozlišíme dvě varianty: variantu A s okraji histogramu $z_1 = \min\{t_1, \dots, t_n\}$ a $z_{m+1} = \max\{t_1, \dots, t_n\}$ a variantu B s okraji histogramu $z_1 = \min\{t_1, \dots, t_n\} - s_t$; $z_{m+1} = \max\{t_1, \dots, t_n\} + s_t$ pro normální rozdělení a $a = 0$; $b = \max\{t_1, \dots, t_n\} + s_t$ pro ostatní rozdělení.

Pro obě varianty byl zkonstruován histogram pro $m = 2, \dots, n$ dělicích intervalů. Pro dané rozdělení, parametry, rozsah výběru a variantu okrajů histogramu je možno odhad entropie uvažovat jako funkci proměnné m a na základě průběhu této funkce v konfrontaci s teoretickou entropií vybrat nejlepší m . Tuto funkci si označíme M .

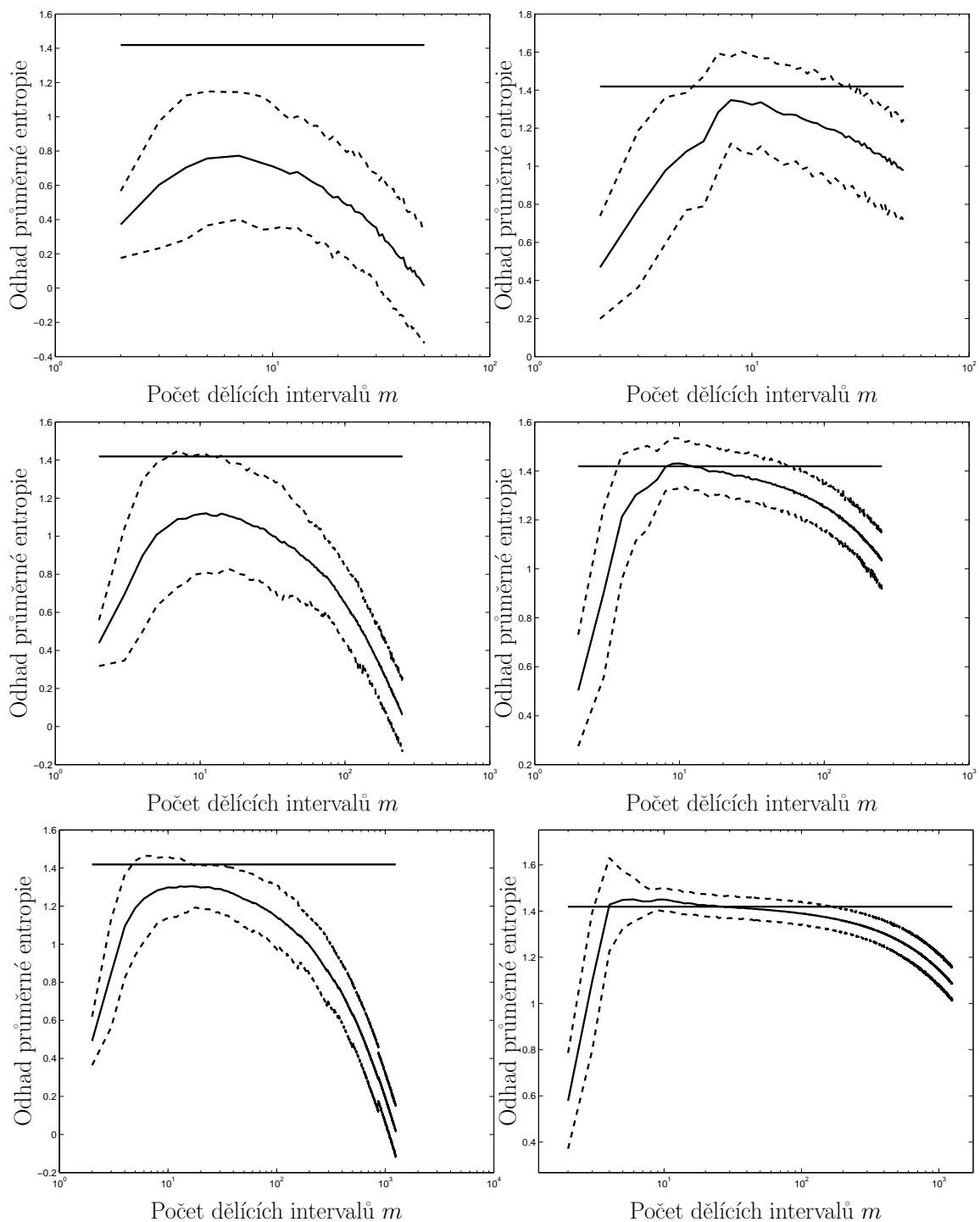
5.1.1 Odhad entropie při ekvidistantním dělení histogramu

Při simulacích z normálního rozdělení pro variantu A platí, že odhad entropie je systematicky podhodnocený. Nejlepší volba m je vždy v maximu M . Varianta B se liší v tom, že odhad pro daný rozsah výběru n je posunutý nahoru. Maximum funkce M pro zhruba $n < 220$ leží pod teoretickou entropií, pro vyšší n nad teoretickou entropií. Když vezmeme jako nejlepší odhad v maximu funkce M , dá varianta B mnohem lepší výsledky. Je pravděpodobné, že lze nalézt korekci okrajů histogramu varianty B v závislosti na n takovou, že průměrný odhad entropie v maximu funkce M by odpovídal teoretické entropii. Pro $\sigma^2 = 0,01$ a vyšší se jeví, že nejlepší odhad průměrné entropie v závislosti na počtu dělicích intervalů je zhruba v maximu. Korigováním okrajů pomocí počtu pozorování n by zřejmě šlo docílit toho, aby nejlepší odhad byl vždy přímo v maximu. Pro $\sigma^2 < 0,01$ by nejlepší odhad entropie mohl být vždy v inflexním bodě funkce M . Srovnání varianty A a B pro výběr z $N(0;1)$ je uvedeno na obr. 5.1.

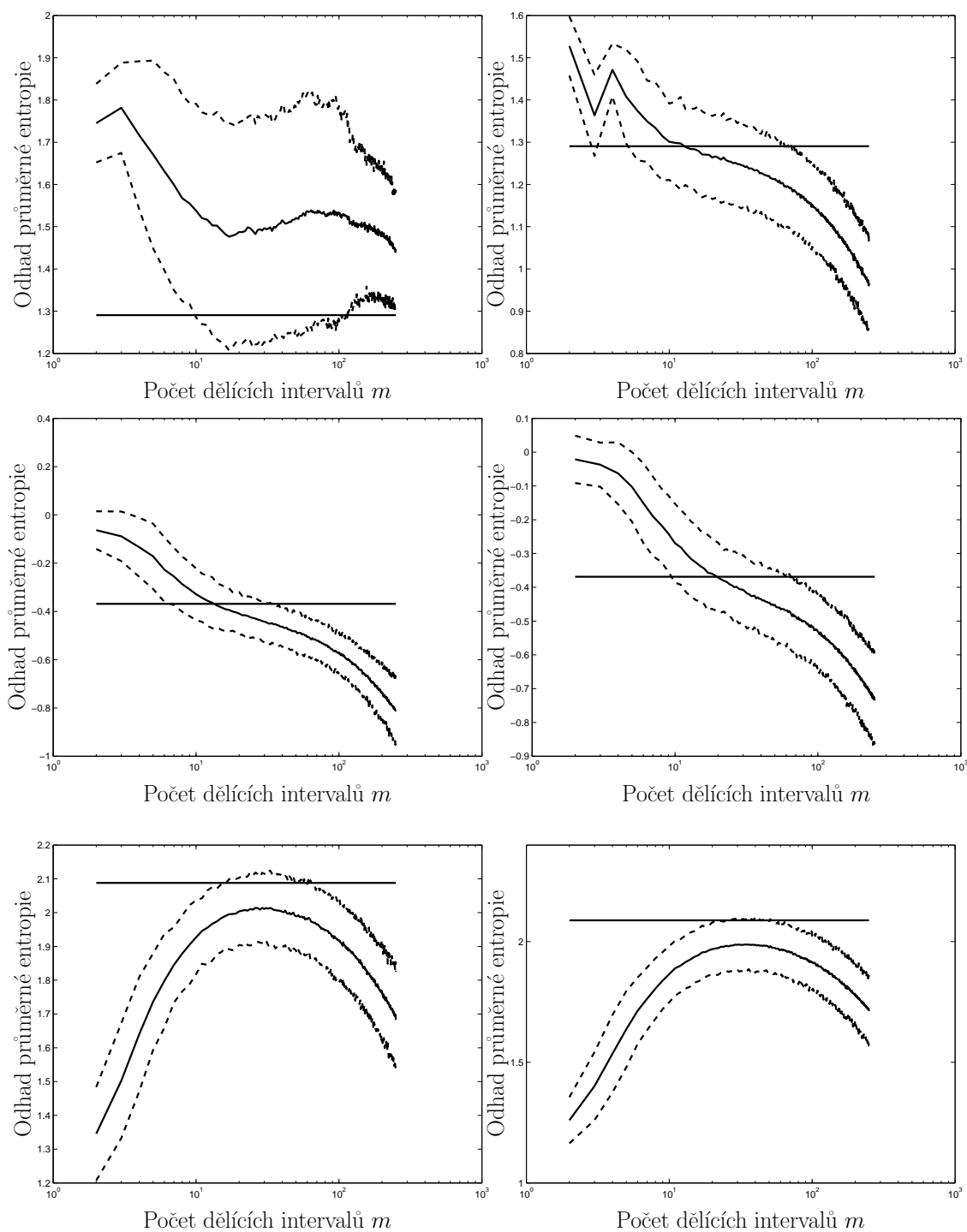
Pro exponenciální rozdělení je odhad entropie systematicky podhodnocený zhruba pro $\lambda = 1,5$ a nižší. Při větších λ a zejména tehdy, když je odhad entropie záporný, lze najít optimální m . Volba m bude ale záviset na parametru (na odhadu momentů). Pro varianty A i B má funkce M velmi podobný průběh.

U Weibullova rozdělení (když nebudeme uvažovat jeho speciální samostatnou variantu exponenciálního rozdělení) je možné identifikovat optimální m . Prakticky se dá usoudit, že optimální m bude ležet v inflexním bodě funkce M . Obě varianty, A i B, dávají velmi podobné výsledky, varianta B o málo lepší (výrazněji je vidět, že teoretická entropie protíná funkci M v jejím inflexním bodě).

Nakonec se budeme zabývat gamma rozdělením. Pro variantu A platí, že situace se výrazně liší v závislosti na kombinaci parametrů. Optimální m lze určit pouze pro záporný odhad entropie, odhad entropie je někdy systematicky nadhodnocený, jindy systematicky podhodnocený. Varianta B dává u některých kombinací parametrů lepší výsledky, viz obr. 5.2.



Obr. 5.1: Odhad průměrné entropie pro $N(0;1)$. Levý sloupec varianta A, pravý sloupec varianta B. Shora dolů 50 pozorování, 250 pozorování a 1250 pozorování. Plnou čarou je vyneseno odhad průměrné entropie (průměr pro každé m z 30 realizací) a čárkovaně pás z průměrné entropie plus minus dvakrát směrodatná odchylka. Rovnou plnou čarou je vyznačena teoretická entropie. Osa x má logaritmické měřítko.



Obr. 5.2: Odhad průměrné entropie shora dolů pro $G(5;20)$, $G(7;2)$ a $G(0,6;2)$. Levý sloupec varianta A, pravý sloupec varianta B. Vždy 250 pozorování. Plnou čarou je vyneseno odhad průměrné entropie (průměr pro každé m z 30 realizací) a čárkovaně pás z průměrné entropie plus minus dvakrát směrodatná odchylka. Rovnou plnou čarou je vyznačena teoretická entropie. Osa x má logaritmické měřítko.

5.1.2 Odhad entropie při neekvidistantním dělení histogramu

Na základě rozsáhlé simulační studie se nám mezi histogramy s neekvidistantním dělením nejlépe jeví takový histogram, který bude tvořen intervaly se stejným počtem pozorování ve všech intervalech až na poslední. Průběh funkce M definované výše je velmi podobný pro všechny typy rozdělení. Varianty A a B se liší, ale není důvod hodnotit variantu B jako lepší. Se zvětšujícím se intervalem pro konstrukci histogramu se funkce M posouvá o něco nahoru a napřimuje se její pravý cíp. Ilustrace průběhu funkce M pro exponenciální rozdělení, variantu A, je uvedena na obr. 5.3.

Vysvětlení průběhu funkce M je následující. Pro málo pozorování v jednotlivých intervalech začne od určitého m závislého na n růst počet pozorování v posledním intervalu. Počet pozorování v posledním intervalu bude vysoký, postupně se bude snižovat (entropie klesat), protože si z něj „ukrojí“ další a další malé intervaly. Pak se zase počet pozorování v posledním intervalu náhle zvýší (entropie se také náhle zvýší, bude zde dlouhý interval s konstantní hodnotou, tj. přiblížení k rovnoměrnému spojitému rozdělení) atd.

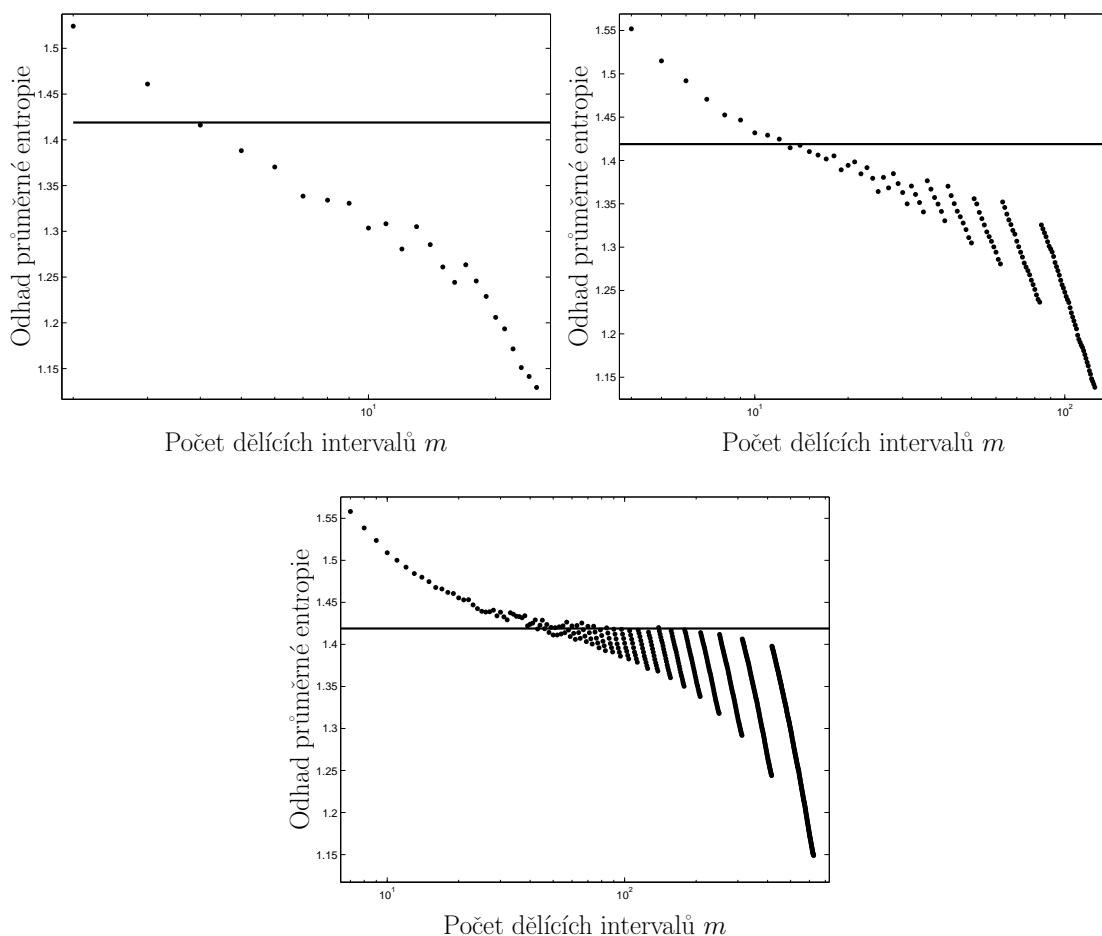
Na základě empirických výsledků jsme stanovili jako optimální počet dělicích intervalů m se stejným počtem pozorování ve všech intervalech až na poslední to m , pro které bude platit vztah

$$2[m/n] > n - [m/n](m - 1). \quad (5.2)$$

Pro 50 pozorování bude $m = 8$, pro 250 $m = 18$ a pro 1250 $m = 42$.

Závěry pro odhad entropie na základě histogramu

Můžeme říci, že není univerzální předpis pro m , který obecně dává nejlepší odhad entropie. Histogram s ekvidistantními dělicími intervaly dává odlišný výstup než histogram s dělicími intervaly se stejným počtem pozorování. Pro histogram s ekvidistantními dělicími intervaly by mohla nejlepší volba m být ve vnitřním extrému či inflexním bodě funkce M . Pro histogram s dělicími intervaly se stejným počtem pozorování by mohla být nejlepší volba m v úseku, kdy M přestane klesat tak strmě a ještě není příliš rozkmitaná.



Obr. 5.3: Odhad entropie na základě histogramu s neekvidistantním dělením. Varianta A, $N(0;1)$, průměry ze 30 realizací. Vlevo nahoře 50 pozorování, vpravo nahoře 250 pozorování a dole 1250 pozorování. Plnou čarou je vynesena odhad průměrné entropie (průměr pro každé m z 30 realizací) a čárkovaně pás z průměrné entropie plus minus dvakrát směrodatná odchylka. Rovnou plnou čarou je vyznačena teoretická entropie. Osa x má logaritmické měřítko.

5.2 Odhad entropie na základě jádrového odhadu hustoty

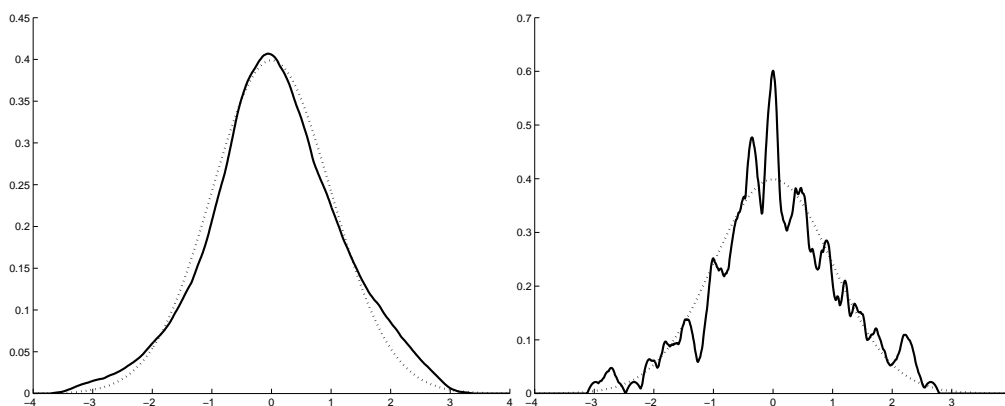
V této podkapitole se budeme zabývat určením optimálního (vyhlazovacího) parametru jádrového odhadu hustoty (3.9) vzhledem k odhadu entropie. Pro normální rozdělení je na základě průběhu odhadu průměrné entropie v závislosti na vyhlazovacím parametru h možné říci, že lze najít vyhlazovací parametr h optimální vzhledem k odhadu entropie. Srovnání průměrného vyhlazovacího parametru optimálního

vzhledem k odhadu entropie a parametru daného metodou křížového ověřování je uvedeno v tab. 5.2. Parametr optimální pro odhad entropie je v průměru systematicky menší než parametr daný metodou křížového ověřování (je vždy zhruba dvakrát menší).

Tab. 5.2: Srovnání průměrného vyhlazovacího parametru optimálního vzhledem k odhadu entropie h_e a průměrného vyhlazovacího parametru daného metodou křížového ověřování h_k . Je uveden odpovídající odhad entropie (pro h_e se jedná o teoretickou entropii). Použito Epanečnikovo jádro, 250 hodnot v jedné realizaci.

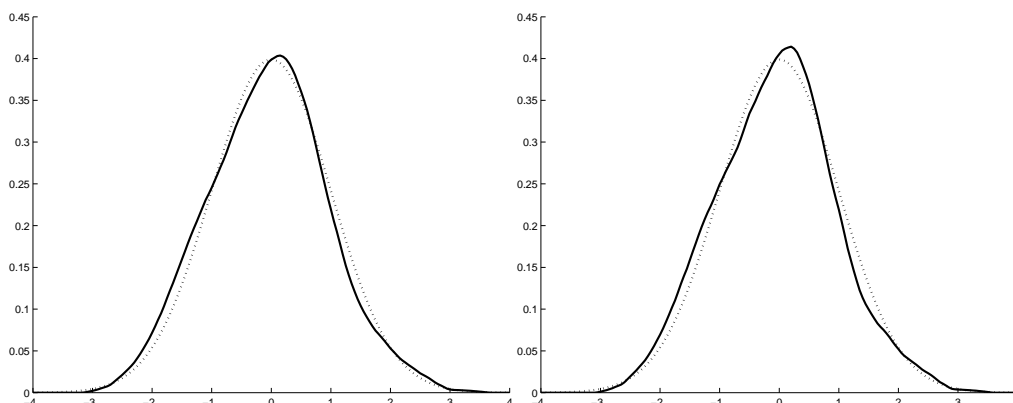
	N(0;0,16) 50 realizací		N(0;1) 30 realizací		N(0;9) 15 realizací	
	h_e	h_k	h_e	h_k	h_e	h_k
	0,1846	0,3136	0,3677	0,8111	1,2090	2,5656
Entropie	0,5026	0,5481	1,4189	1,4817	2,5176	2,5765

Když se však zaměříme na jednotlivé realizace, můžeme pozorovat případy odpovídající průměrné charakteristice (obr. 5.4), případy dobrého odhadu entropie při použití křížového ověřování (obr. 5.5) i případy, kdy jádrový odhad s optimální entropií je hladší než podle křížového ověřování (obr. 5.6). Odhady entropie s různými jádry z tab. 3.1 se příliš neliší.

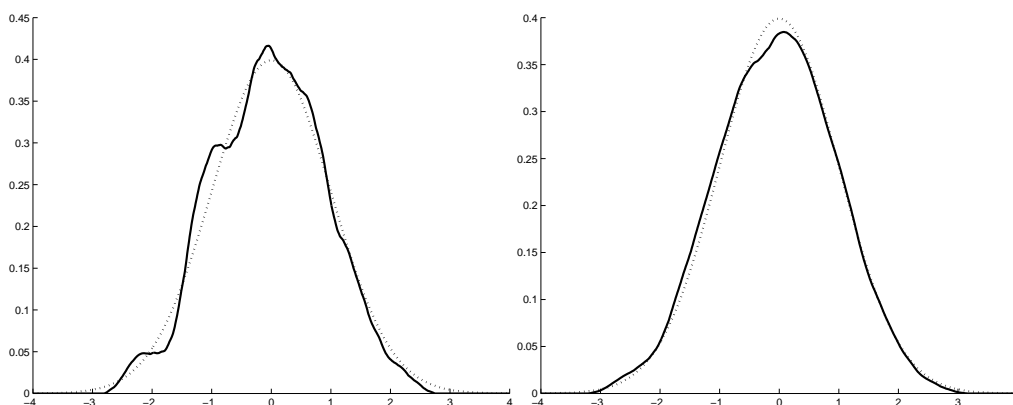


Obr. 5.4: Srovnání jádrového odhadu hustoty dle křížového ověřování (vlevo) a s optimální entropií (vpravo). Použito Epanečnikovo jádro, simulováno 250 hodnot z rozdělení $N(0; 1)$. Tečkovaně teoretická hustota, plnou čarou odhad. $h_k = 0,7406$ (entropie 1,5006), $h_e = 0,14$ (entropie 1,4183).

Pro asymetrická rozdělení, tj. v našem případě exponenciální, Weibullovo a gamma rozdělení totéž co o rozdělením normálním říci nemůžeme. Na základě průběhu odhadu průměrné entropie v závislosti na vyhlazovacím parametru h se ukázalo, že ne vždy lze najít vyhlazovací parametr h optimální vzhledem k odhadu entropie.



Obr. 5.5: Srovnání jádrového odhadu hustoty dle křížového ověřování (vlevo) a s optimální entropií (vpravo). Použito Epanečnikovo jádro, simulováno 250 hodnot z rozdělení $N(0; 1)$. Tečkovaně teoretická hustota, plnou čarou odhad. $h_k = 0,8488$ (entropie 1,4325), $h_e = 0,7655$ (entropie 1,4182).

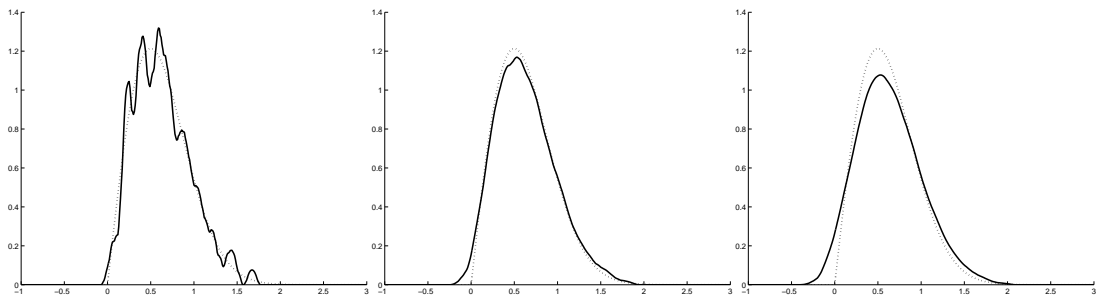


Obr. 5.6: Srovnání jádrového odhadu hustoty dle křížového ověřování (vlevo) a s optimální entropií (vpravo). Použito Epanečnikovo jádro, simulováno 250 hodnot z rozdělení $N(0; 1)$. Tečkovaně teoretická hustota, plnou čarou odhad. $h_k = 0,4379$ (entropie 1,3562), $h_e = 0,843$ (entropie 1,4183).

Špatný odhad entropie bude dán tím, že pro nesymetrické hustoty použijeme symetrické jádro, které může relativně dobře popsat buď jeden nebo druhý cíp grafu nesymetrické hustoty, ale ne oba zároveň (viz obr. 5.7). Otázkou je, zda pro hustotu exponenciálního rozdělení je vhodné vůbec použít jádrové odhady jako takové.

Ilustrační srovnání průměrného vyhlazovacího parametru optimálního vzhledem k odhadu entropie a parametru daného metodou křížového ověřování je pro Weibullovo rozdělení uvedeno v tab. 5.3. Parametr optimální pro odhad entropie je v průměru systematicky menší než parametr daný metodou křížového ověřování.

Při jádrovém odhadu hustoty a potažmo entropie z konkrétních dat můžeme



Obr. 5.7: Jádrový odhad hustoty pro $W(2;2)$. Použito Epanečnikovo jádro, 250 pozorování. Vlevo relativně dobře popsán levý cíp hustoty ($h = 0,0763$, entropie 0,2489), uprostřed odhad dle křížového ověřování ($h = 0,2449$, entropie 0,3266) a vpravo hladký (přehlazený) odhad hustoty ($h = 0,4$, entropie 0,4172). V levém grafu je použit parametr vyhlazovacího jádra $h_e = 0,0763$ dávající optimální entropii. V tomto případě si entropie vynucuje správné vymezení hustoty na kladnou polopřímku.

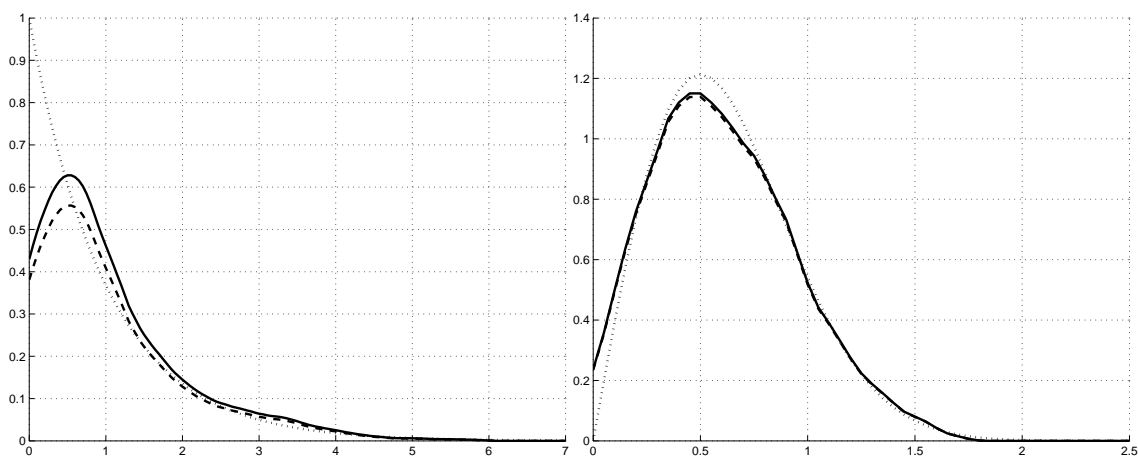
Tab. 5.3: Srovnání průměrného vyhlazovacího parametru optimálního vzhledem k odhadu entropie h_e a průměrného vyhlazovacího parametru daného metodou křížového ověřování h_k . Je uveden odpovídající odhad entropie (pro h_e se jedná o teoretickou entropii). Použito Epanečnikovo jádro, 250 hodnot v jedné realizaci.

	W(2;2) 15 realizací		W(5;2) 15 realizací		W(0,5;10) 15 realizací	
	h_e	h_k	h_e	h_k	h_e	h_k
	0,0978	0,2825	0,0582	0,1521	0,0506	0,0868
Entropie	0,2489	0,3500	-0,2093	-0,1253	-0,7138	-0,6600

využít předpokladů o rozdělení a omezit odhad hustoty, který přesáhne do záporných hodnot, na správný definiční obor. Následně odhad renormalizujeme. Odhad entropie se ve většině případů mírně zlepší, viz obr. 5.8. Při posunutí hustoty do kladných hodnot problém s přesahem (tj. špatným odhadem cípu hustoty) přetrvává.

Na základě srovnání průběhu odhadu entropie v závislosti na vyhlazovacím parametru a srovnání vyhlazovacího parametru optimálního vzhledem k odhadu entropie a vyhlazovacího parametru daného metodou křížového ověřování můžeme konstatovat, že odhad entropie na základě jádrových odhadů není příliš dobrý, i pro normální rozdělení se může dojít k velmi špatnému výsledku. Odhad entropie pomocí jádrových odhadů může být pro konkrétní případy o něco lepší než odhad pomocí histogramu, zejména v tom, že závisí jen na jednom parametru.

U asymetrických hustot dochází k okrajovému problému jádrového odhadu hustoty, kdy není respektován definiční obor. Tento problém je do jisté míry řešitelný useknutím a renormalizací odhadu. Existují i pokročilé techniky které tento problém řeší, například variabilní volba vyhlazovacího parametru (Terrell a Scott, 1992), me-



Obr. 5.8: Jádrový odhad hustoty vymezený na kladnou polopřímku. Tečkovaně teoretická hustota, čárkovaně původní jádrový odhad a plnou čarou renormalizovaný odhad. Vždy 250 pozorování. Vlevo $E(1)$, původní odhad entropie 1,2470, po renormalizaci 1,1016. Vpravo $W(2;2)$, původní odhad entropie 0,3191, po renormalizaci 0,2877.

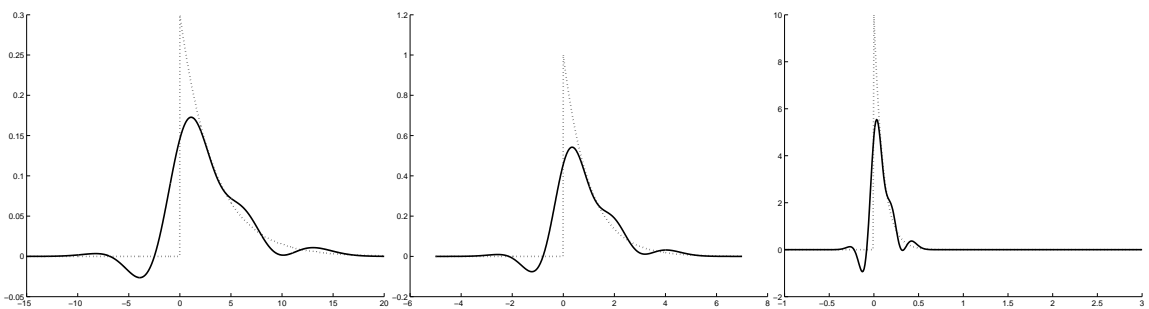
toda zrcadlení či metoda navržená v práci Koláček a Karunamuni (2009).

5.3 Odhad entropie na základě momentového odhadu hustoty

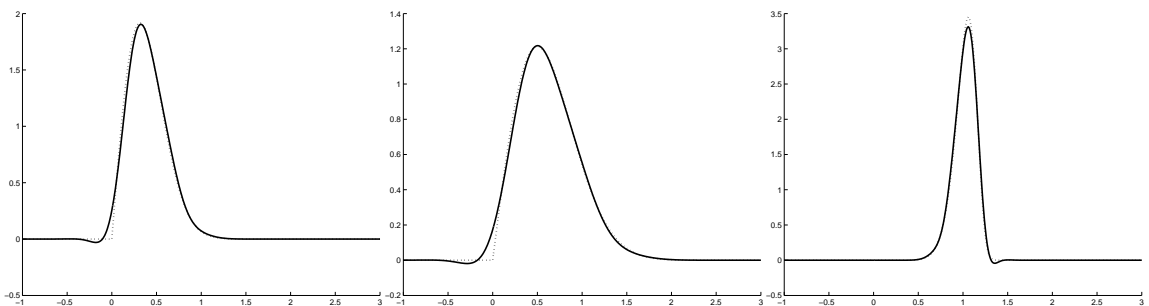
V této kapitole se budeme zabývat plug-in odhady entropie stanovenými na základě momentových odhadů hustoty. Hlavním nedostatkem momentových odhadů hustoty je fakt, že odhady založené na momentových charakteristikách nemusí splňovat vlastnosti hustoty, tedy $\int_{-\infty}^{\infty} f(t)dt = 1$ (odchylky při výpočtech ale nebyly velké) a $f(t) \geq 0 \forall t$. Druhý nedostatek je zásadní. Pro smysluplný odhad hustoty a následný výpočet entropie je nutná renormalizace odhadnuté hustoty. Odhadnutá hustota se renormalizuje ve dvou krocích:

1. Eliminují se (položí se rovny 0) ty části odhadnuté hustoty, které jsou záporné. Jako nulové se též položí kladné oscilace mimo hlavní odhad hustoty (není důvod domnívat se, že by odhadnutá hustota měla být vícevrcholová, jejich existence vyplývá z metody odhadu). Nakonec se takto upravený odhad vydělí svým obsahem. Upravený odhad má již vlastnosti hustoty.
2. V případech, že odhad hustoty je nenulový i mimo předpokládaný definiční obor, odhad na tento definiční obor omezíme a opět vydělíme jeho obsahem.

Nyní k jednotlivým sledovaným rozdělením. U normálního rozdělení je momentový odhad hustoty (3.13) či (3.14) vzhledem ke konstrukci odhadu kvalitativně srovnatelný k maximálně věrohodnému odhadu parametrů dosazených do teoretické hustoty normálního rozdělení, tj. velmi dobrý. Vzhledem ke tvaru hustoty exponenciálního rozdělení se nedá očekávat, že by odhad v kombinaci s polynomm byl dobrý. Odhad hustoty je rozvlněný, s přesahy do záporných hodnot i do záporného definičního oboru, viz obr. 5.9. U Weibullova rozdělení je odhad hustoty nečekaně dobrý. Problém je u strmého cípu, který není popsán přesně a jeho nástupu předchází propad do záporných hodnot, viz obr. 5.10. Pro gamma rozdělení platí, že odhad pozvolnějšího cípu hustoty je rozvlněný, strmý cíp není popsán dobře a předchází mu propad do záporných hodnot, viz obr. 5.11. Celkově můžeme říci, že

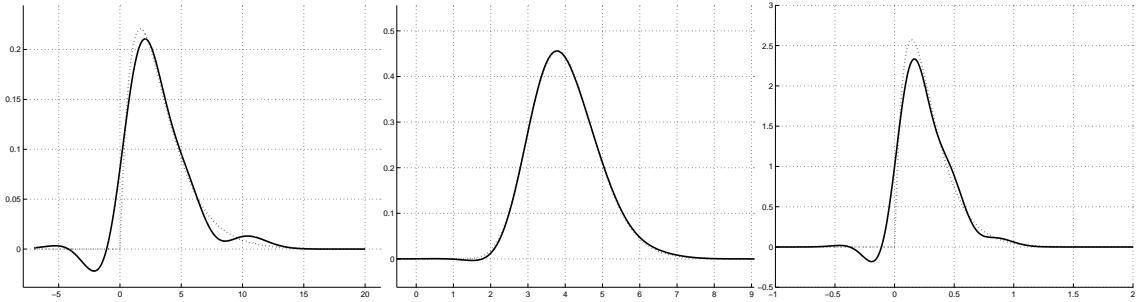


Obr. 5.9: Momentový odhad hustoty. 2500 hodnot, tj. prakticky při teoretických parametrech a charakteristikách. Zleva doprava $E(0,3)$, $E(1)$ a $E(10)$. Plnou čarou odhad, tečkovaně teoretická hustota.



Obr. 5.10: Momentový odhad hustoty. 2500 hodnot, tj. prakticky při teoretických parametrech a charakteristikách. Zleva doprava $W(5;2)$, $W(2;2)$ a $W(0,5;10)$. Plnou čarou odhad, tečkovaně teoretická hustota.

odhad hustoty byl dobrý u všech sledovaných rozdělení až na exponenciální rozdělení a případy strmé výrazně asymetrické hustoty u dalších rozdělení.



Obr. 5.11: Momentový odhad hustoty. 2500 hodnot, tj. prakticky při teoretických parametrech a charakteristikách. Zleva doprava $G(0,6;2)$, $G(5;20)$ a $G(7;2)$. Plnou čarou odhad, tečkovaně teoretická hustota.

5.3.1 Momentový odhad konstruovaný z gamma hustoty

Vzhledem k charakteru námi zkoumaných rozdělení se nám zdálo vhodné odvodit momentový odhad hustoty konstruovaný na základě rozdělení definovaného na polopřímce, zvolili jsme gamma rozdělení (2) s $a = 0$.

Při odvozování nejprve upravíme veličiny T vynásobením konstantou $k = \frac{E(T)}{D(T)}$, dostaneme tak transformované veličiny $X = kT$. Platí pro ně $E(X) = D(X) = c$. Odhad hustoty veličiny X dostaneme jako

$$\hat{g}(x) = f(x) + e_3 d_3 f(x) + e_4 d_4 f(x), \quad (5.3)$$

kde

$$f(x) = G(1; c) = \frac{e^{-x} x^{c-1}}{\Gamma(c)},$$

a

$$d_i f(x) = \frac{\partial^i f(x)}{\partial x^i},$$

tedy

$$d_3 f(x) = \frac{e^{-x} x^{c-4}}{\Gamma(c)} (x^3 + x^2(3 - 3c) + x(3c^2 - 9c + 6) - c^3 + 6c^2 - 11c + 6)$$

a

$$d_4 f(x) = \frac{e^{-x} x^{c-5}}{\Gamma(c)} (x^4 + x^3(4 - 4c) + x^2(12 + 6c^2 - 18c) + x(24 - 4c^3 + 24c^2 - 44c) + 24 - 50c + 35c^2 - 10c^3 + c^4),$$

dále

$$e_3 = \frac{1}{6}(\mu_3 - 2c) \quad \text{a} \quad e_4 = \frac{1}{24}(\mu_4 - 18\mu_2^2c).$$

Odhad hustoty původní veličiny T získáme jako

$$\hat{g}_T\left(\frac{x}{k}\right) = k\hat{g}(x).$$

V práci Bowers (1966) je odvození provedeno pomocí rozvoje hustoty gamma rozdělení ve tvaru

$$f(t; c, \beta) = \frac{e^{-\frac{t}{\beta}} t^{c-1}}{\beta^c \Gamma(c)} \quad c > 0, \quad \beta > 0, \quad t \geq 0$$

s využitím Laguerrových polynomů (viz např. Abramowitz a Stegun (1965)), výsledkem je odhad

$$\begin{aligned} g(x) \approx & f(x; c, 1)[1 - A + B] + f(x; c + 1, 1)[3A - 4B] + f(x; c + 2, 1)[6B - 3A] \\ & + f(x; c + 3, 1)[A - 4B] + f(x; c + 4, 1)B, \end{aligned} \quad (5.4)$$

kde

$$A = \frac{\mu_3 - 2c}{6}$$

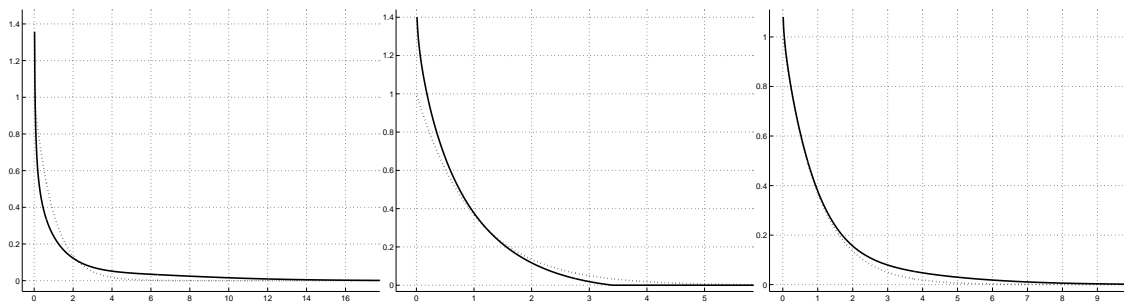
a

$$B = \frac{\mu_4 - 12\mu_3 - 3c^2 + 18c}{24}.$$

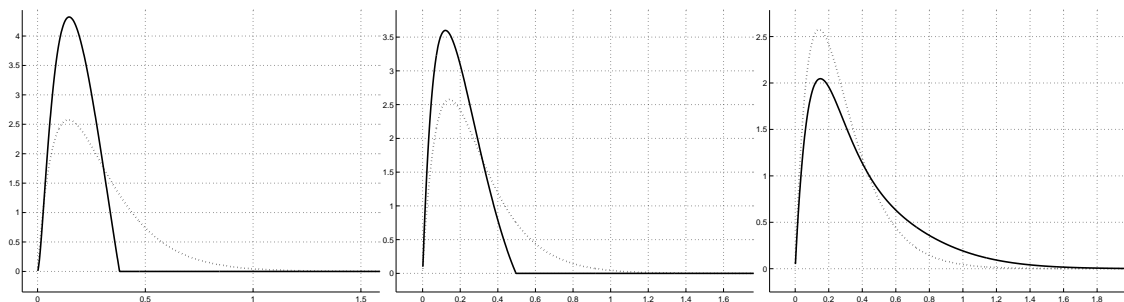
Ačkoliv se od sebe odhady (5.4) a (5.3) formálně liší, ve výsledku jsou ekvivalentní. Na simulacích se ukázalo, že tyto odhady jsou špatné. Důvodem bude velká citlivost na odhady momentů, takže pro námi zkoumané rozsahy pozorování jsou odhady založené na hustotě gamma rozdělení nepoužitelné. Jak ilustrujeme na obr. 5.12 a 5.13, odhad je odvozen dobře, nicméně aproximace může dát i pro velké rozsahy výběru naprosto chybný odhad hustoty i entropie. V dalším textu tyto odhady používat nebudeme.

5.3.2 Výpočet entropie z momentového odhadu hustoty

Entropii nelze spočítat přímou integrací odhadnuté hustoty, vypočteme ji numericky dle vzorce (5.1) dosazením odhadnuté renormalizované hustoty. I při špatném odhadu hustoty (případ exponenciálního rozdělení) dostaneme relativně velmi dobrý

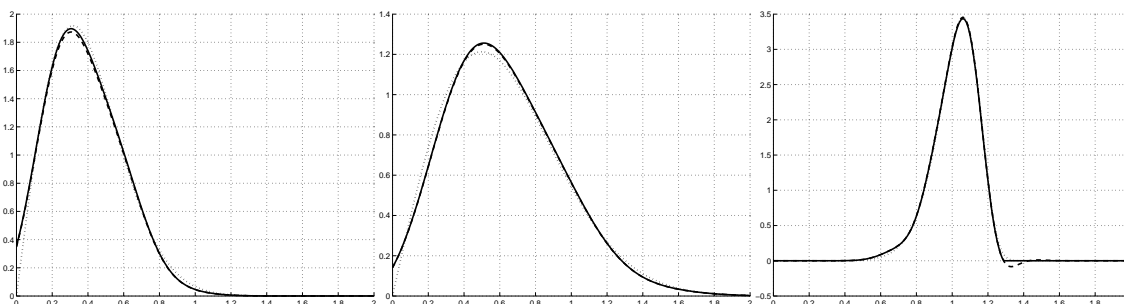


Obr. 5.12: Momentový odhad hustoty $E(1)$ (podobně pro jiné parametry) kombinací gamma hustot. Teoretická entropie je rovna 1. Zleva doprava vykreslen odhad pro 50 (entropie 1,9004), 250 (entropie 0,6895) a 1250 pozorování (entropie 1,3769), plnou čarou normalizovaný odhad hustoty, tečkovaně teoretická hustota.

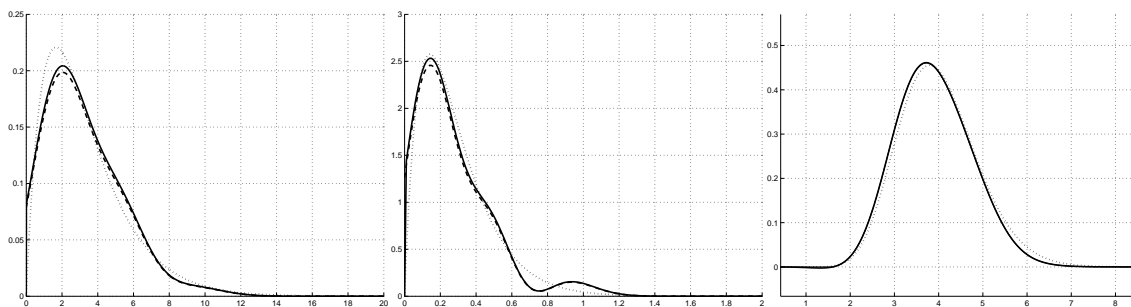


Obr. 5.13: Momentový odhad hustoty $G(7;2)$ kombinací gamma hustot. Teoretická entropie je rovna $-0,3687$. Zleva doprava vykreslen odhad pro 50 (entropie $-1,1384$), 250 (entropie $-0,9046$) a 1250 pozorování (entropie $-0,0275$), plnou čarou normalizovaný odhad hustoty, tečkovaně teoretická hustota.

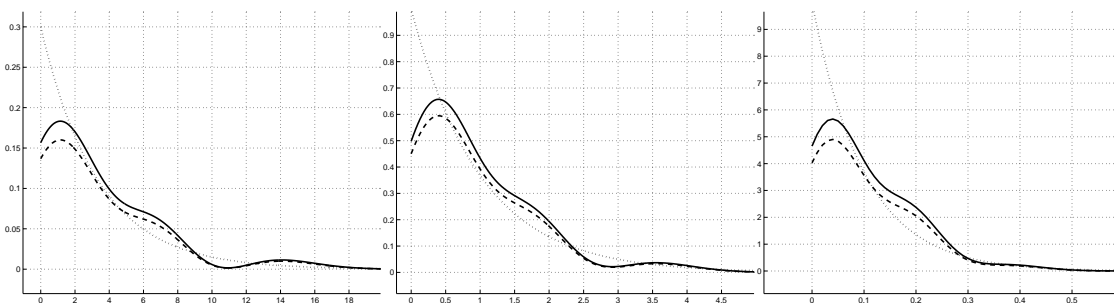
odhad entropie (viz obr. 5.14 až 5.16). Pouze pokud je odhad hustoty na první pohled nepřijatelný, odhad entropie je nepoužitelný. Odhad je oproti jádrovému odhadu vypočítán mnohem rychleji.



Obr. 5.14: Momentový odhad hustoty, vždy 250 hodnot. Zleva doprava vykreslen odhad pro $W(5;2)$ (entropie $-0,2063$), $W(2;2)$ (entropie 0,2394) a $W(0,5;10)$ (entropie $-0,7109$), plnou čarou normalizovaný odhad hustoty, čárkovaně původní odhad hustoty, tečkovaně teoretická hustota.



Obr. 5.15: Momentový odhad hustoty, vždy 250 hodnot. Zleva doprava vykreslen odhad pro $G(0,6;2)$ (entropie 2,0827), $G(7;2)$ (entropie -0,3568) a $G(5;20)$ (entropie 1,2414). Plnou čarou normalizovaný odhad hustoty, čárkovaně původní odhad hustoty, tečkovaně teoretická hustota.



Obr. 5.16: Momentový odhad hustoty, vždy 250 hodnot. Zleva doprava vykreslen odhad pro $E(0,3)$ (entropie 2,3218), $E(1)$ (entropie 1,0022) a $E(10)$ (entropie -1,2203). Plnou čarou normalizovaný odhad hustoty, čárkovaně původní odhad hustoty, tečkovaně teoretická hustota.

5.4 Přímé odhady entropie

5.4.1 Odhad entropie na základě rozestupu vzorků

Prvním přímým odhadem entropie, kterým se budeme zabývat, je odhad entropie založený na „rozestupu vzorků“ (viz Beirlant et al. (1997)). Rozestupem řádu m neboli m -rozestupem ($1 \leq i < i + m \leq n$, kde n je rozsah výběru) rozumíme $t_{i+m} - t_i$. Na základě rozestupu vzorků je možné konstruovat odhad hustoty

$$\hat{f}(t) = \frac{m}{n} \frac{1}{t_{im} - t_{(i-1)m}}$$

pro $t \in [t_{(i-1)m}, t_{im}]$. Tento odhad hustoty je konzistentní, jestliže pro $n \rightarrow \infty$ platí

$$m_n \rightarrow \infty, \quad m_n/n \rightarrow 0.$$

Z odhadu hustoty založeného na rozestupu vzorků můžeme pro pevné m odvodit m -rozestupový odhad entropie

$$H_s = \frac{1}{n} \sum_{i=1}^{n-m} \ln \left(\frac{n}{m} (t_{(i+m)} - t_{(i)}) \right) + \Psi(m) + \ln m. \quad (5.5)$$

Tímto odhadem se budeme dále zabývat. Odpovídající odhad hustoty sice není konzistentní (m je pevné, nesplňuje podmínky), ale dá se ukázat, že odhad entropie za určitých podmínek konzistentní je (viz Beirlant et al. (1997)). Kvůli snížení asymptotického rozptylu se někdy používá m_n -rozestupový odhad pro $m_n \rightarrow \infty$

$$H_n = \frac{1}{n} \sum_{i=1}^{n-m_n} \ln \left(\frac{n}{m_n} (t_{i+m_n} - t_i) \right).$$

5.4.2 Odhad entropie podle Vasicka¹

Odhad entropie vyjádřený pomocí distribuční funkce (3.18) může být zkonstruován nahrazením distribuční funkce F empirickou distribuční funkcí F_n a použitím diferenciálního operátoru místo diferenciálního. Derivace $F^{-1}(p)$ je pak odhadnuta výrazem

$$\frac{n(t_{(i+m)} - t_{(i-m)})}{2m}$$

pro

$$\frac{i-1}{n} < p \leq \frac{i}{n}, \quad i = m+1, m+2, \dots, n-m,$$

kde m je malé přirozené číslo menší než $n/2$. Jednostranné diference typu $t_{(i+m)} - t_{(1)}$ nebo $t_{(n)} - t_{(i-m)}$ se použijí namísto $t_{(i+m)} - t_{(i-m)}$ když $p \leq m/n$ respektive $p >$

¹Oldřich Vasicek (66), původním jménem Oldřich Alfons Vašíček, je dosud jediným čechem, který získal celosvětové renomé v oblasti ekonomie. Od roku 1968 žije s manželkou Silvou v San Francisku. V Praze absolvoval jadernou fyziku, při práci v Československé akademii věd vystudoval matematicko-fyzikální fakultu. V USA nastoupil do výzkumného oddělení banky Wells Fargo. Tam se setkal s několika finančními teoretiky, kteří byli později oceněni Nobelovou cenou (Markovitz a Sharpe za model oceňování kapitálových aktiv, Merton a Scholes za model oceňování opčních kontraktů). Učil na prestižních vysokých školách, např. na Kalifornské univerzitě. Jeho „Vasicek model“, popisující chování úrokových měr, je dnes součástí všech moderních kursů ekonomie. V České republice neudrhuje žádné kontakty, nicméně každoročně přispívá značnou částkou na obnovu zchátralých historických staveb. Zdroj: časopis Týden, 38/2005, str. 34–37.

$(n - m)/n$. Takto dostaneme odhad H_{mn} entropie $H(f)$

$$H_{mn} = n^{-1} \sum_{i=1}^n \log \left\{ (t_{(i+m)} - t_{(i-m)})n / (2m) \right\}, \quad (5.6)$$

kde $t_{(i)} = t_{(1)}$, $i < 1$ a $t_{(i)} = t_{(n)}$, $i > n$. Vasicek tuto sumu rozepisuje do tří komponent, ukazuje vlastnosti těchto komponent, a vzhledem k eliminaci chyby odhadu doporučuje upravený odhad entropie

$$H'_{mn} = H_{mn} - \left(\log n - \log 2m + \left(1 - \frac{2m}{n}\right) \psi(2m) - \psi(n+1) + \frac{2}{n} \sum_{i=1}^m \psi(i+m-1) \right). \quad (5.7)$$

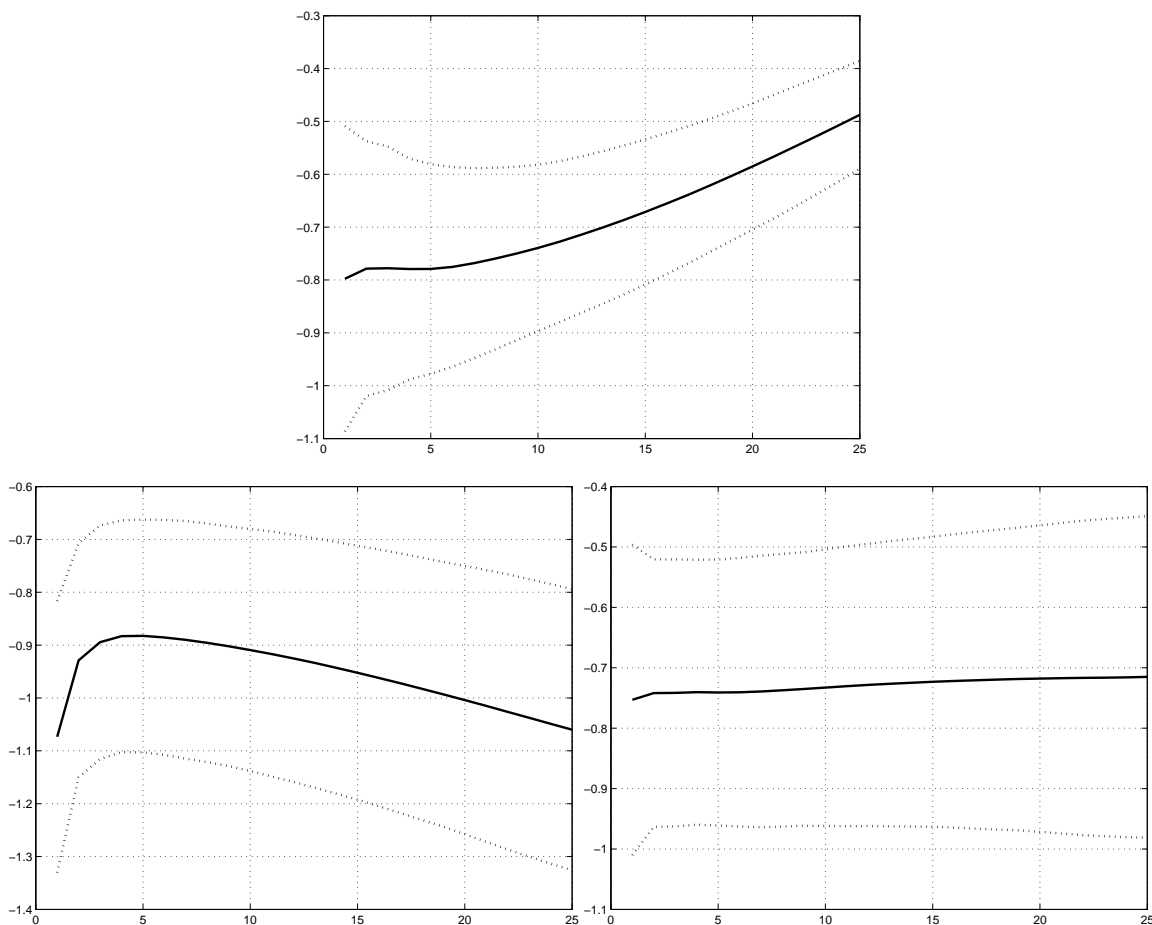
5.4.3 Srovnání odhadu založeného na rozestupu vzorků a Vasickových odhadů vzhledem k volbě parametru m

Při hodnocení dosud uvedených přímých odhadů se ukazuje, že je zde silná závislost na volbě parametru m . Na obr. 5.17 až 5.19 je ilustrován odhad entropie v závislosti na tomto parametru.

Odhad (5.5) založený na rozestupu vzorků dával nejlepší výsledky pro $m = 1$ ve všech případech, kdy odhadnutá entropie byla kladná. V případě, že odhadnutá entropie byla záporná, bylo m mnohem vyšší úměrně počtu pozorování. Průběh odhadu entropie v závislosti na m nebyl v tomto případě lineární, vzít jiné m mezi 1 a optimálním m by nebylo dobré (viz obr. 5.18 horní graf). Pro $m = 1$ vždy dostaneme přijatelný odhad.

Odhad entropie podle Vasicka (5.6) bude z těchto odhadů nejméně kvalitní. Pro malý počet pozorování se i pro nejlepší volbu m může od teoretické hodnoty lišit výrazně. Optimální m bylo pro 50 pozorování 5 (mimo exponenciálního), pro exponenciální rozdělení více než 25. Pro 250 pozorování bylo m v rozsahu od 15 do 100, v průměru 25 (mimo exponenciálního), pro exponenciální 70. Pro 1250 pozorování se optimální hodnoty pohybovaly od 40 do 150, průměrně 60. Exponenciální rozdělení se neodlišovalo.

Vylepšený Vasickuv odhad (5.7) dává jednoznačně lepší výsledky než neupravený Vasickuv odhad. Pro 50 pozorování se optimální m pohybovalo od 1 do 25, ale vzhledem k průběhu odhadu entropie v závislosti na m (viz obr. 5.17 pravý dolní graf) to není velký problém. Pro 250 pozorování se optimální hodnoty pohybovaly

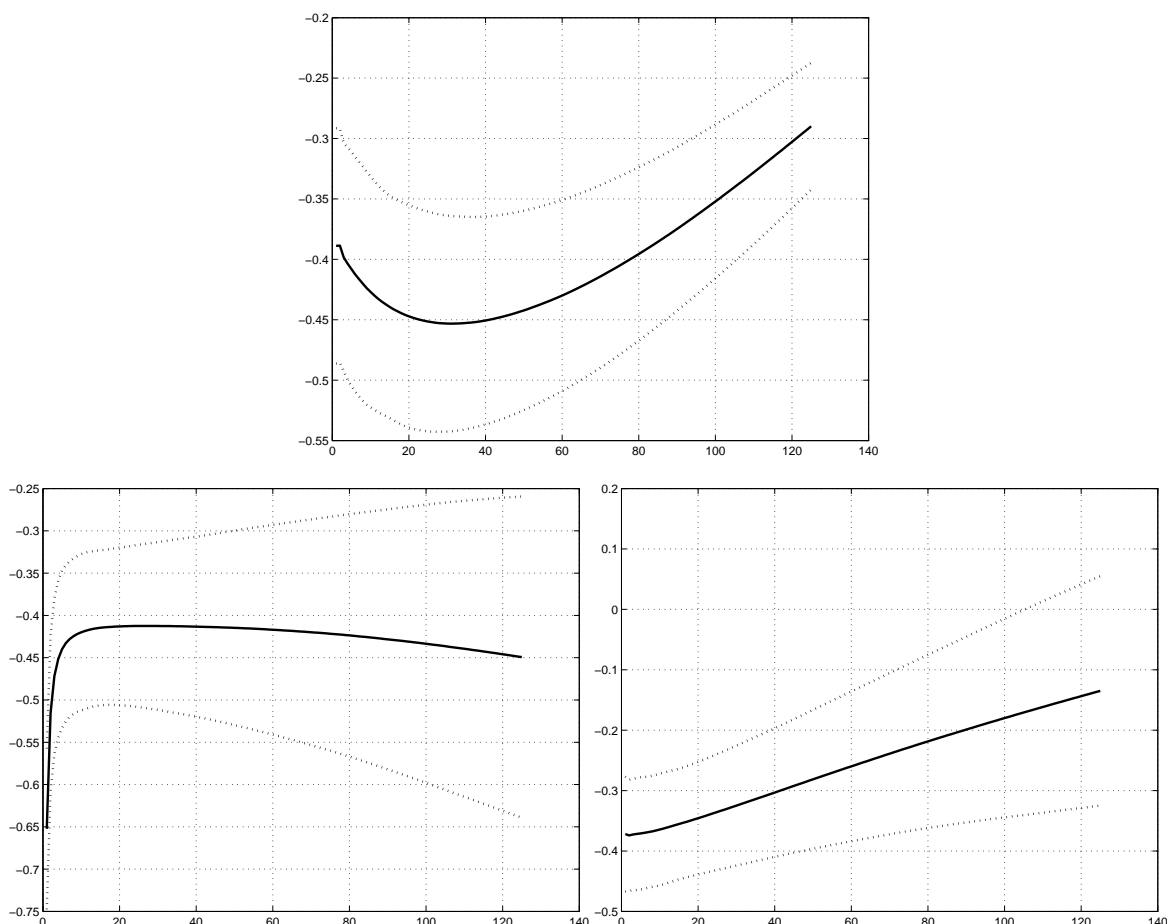


Obr. 5.17: Průběh odhadu entropie v závislosti na parametru m pro rozdělení $W(0,5;10)$, 50 pozorování. Vykreslen je průměrný odhad ze 100 realizací (plná čára) a kolem něj pás z výběrové směrodatné odchylky (tečkovaně). Teoretická entropie je rovna $-0,7138$. Nahoře odhad založený na rozestupu vzorků (5.5) (optimální $m = 12$, entropie $-0,7145$), vlevo dole Vasickuv odhad (5.6) (optimální $m = 5$, entropie $-0,882$) a vpravo dole vylepšený Vasickuv odhad (5.7) (optimální $m = 25$, entropie $-0,7151$).

od 1 do 13 (průměr 4), pro 1250 pozorování od 1 do 22 (průměr 4). Původní Vasickuv odhad je systematicky menší než vylepšený Vasickuv odhad.

5.4.4 Odhad entropie na základě vzdálenosti od nejbližšího souseda

Dalším přímým odhadem entropie je odhad založený na „vzdálenosti od nejbližšího souseda“ (viz Beirlant et al. (1997)). Nechť $\rho_{n,i}$ je vzdálenost t_i od nejbližšího souseda definovaná jako $\rho_{n,i} = \min_{j \neq i, j \leq n} |t_i - t_j|$. Odhad entropie na základě vzdálenosti od



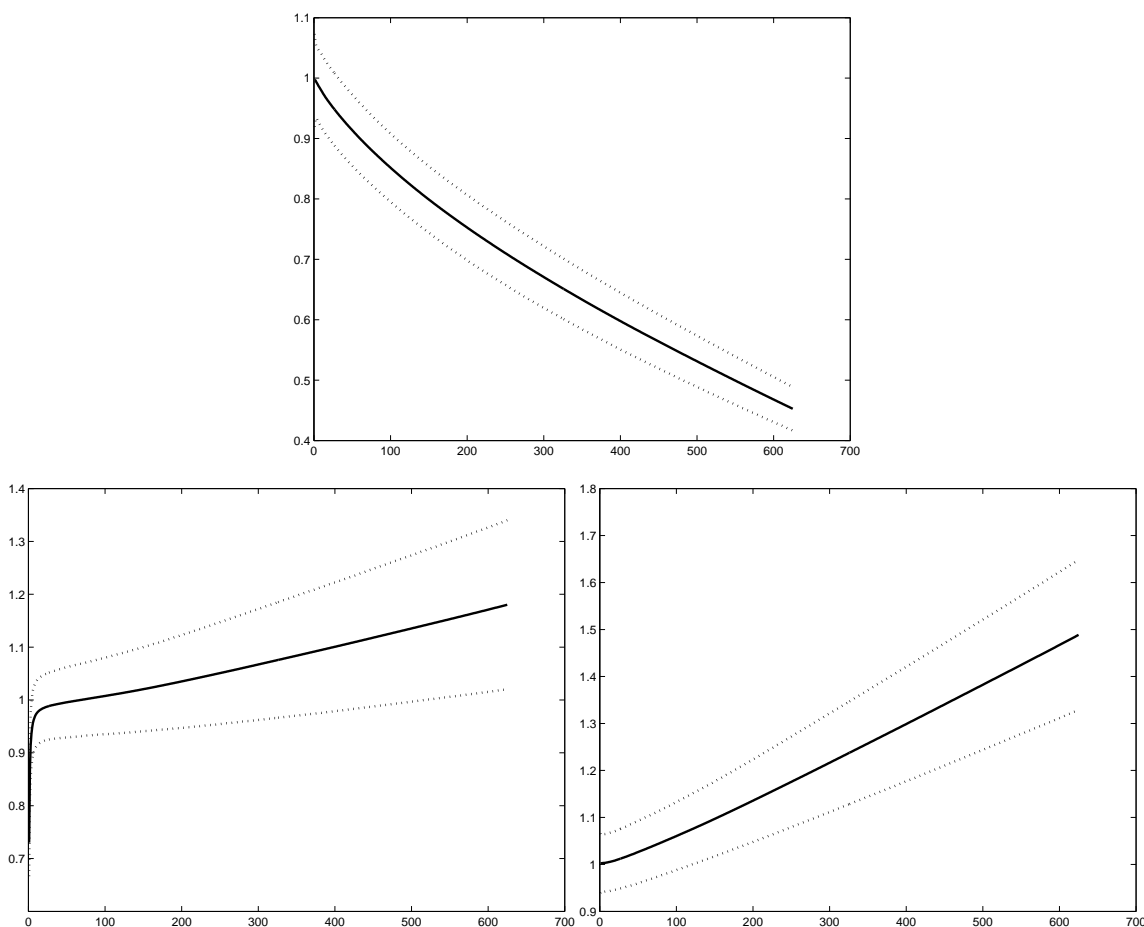
Obr. 5.18: Průběh odhadu entropie v závislosti na parametru m pro rozdělení $G(7;2)$, 250 pozorování. Vykreslen je průměrný odhad ze 100 realizací (plná čára) a kolem něj pás z výběrové směrodatné odchylky (tečkovaně). Teoretická entropie je rovna $-0,3687$. Nahoře odhad založený na rozestupu vzorků (5.5) (optimální $m = 93$, entropie $-0,368$), vlevo dole Vasickuv odhad (5.6) (optimální $m = 26$, entropie $-0,4124$) a vpravo dole vylepšený Vasickuv odhad (5.7) (optimální $m = 7$, entropie $-0,3687$).

nejbližšího souseda je potom

$$H_n = \frac{1}{n} \sum_{i=1}^n \ln(n\rho_{n,i}) + \ln 2 + C_E, \quad (5.8)$$

kde C_E je Eulerova konstanta $C_E = -\int_0^\infty e^{-t} \ln t dt$ ($\approx 0,577215$). Podobný odhad je popsán v práci Kozachenko a Leonenko (1987).

Tento odhad vykazuje v průměru dobré výsledky. S vyšším počtem pozorování se odhad zpřesňuje. Jeho výhodou je oproti ostatním odhadům nezávislost na jakékoli uživatelem volené hodnotě.



Obr. 5.19: Průběh odhadu entropie v závislosti na parametru m pro Rozdělení $E(1)$, 1250 pozorování. Vykreslen je průměrný odhad ze 100 realizací (plná čára) a kolem něj pás z výběrové směrodatné odchylky (tečkovaně). Teoretická entropie je rovna 1. Nahoře odhad založený na rozestupu vzorků (5.5) (optimální $m = 1$, entropie 0,997), vlevo dole Vasickuv odhad (5.6) (optimální $m = 68$, entropie 1) a vpravo dole vylepšený Vasickuv odhad (5.7) (optimální $m = 1$, entropie 1,0024).

5.4.5 Přímý odhad entropie na základě momentových charakteristik

Vzhledem k dobrým výsledkům při odhadu entropie z momentového odhadu hustoty jsme odvodili pro odhad entropie explicitní vzorec. Přímo ze vzorců (3.13) a (3.14) se entropie pravděpodobně vypočítat nedá. Výpočet lze provést v případě, kdy ve vzorci (3.11) nerozvinujeme exponenciálu do řady, ale pracujeme s jeho přesným tvarem. Nebude se pak jednat o vzorce odvozené z (3.13) a (3.14), ale protože v nich budou vystupovat členy k nim analogické, budeme je tak nazývat. Odhady jsou

konstruovány pro hustoty na kladné polopřímce. Pro vzorec odhadu entropie na základě Gram-Charlierova rozvoje označíme

$$A_1 = \frac{1}{4}(1 + \ln(2\pi\sigma^2)) \left(1 + \operatorname{erf} \left(\frac{\mu}{\sqrt{2}\sigma} \right) \right),$$

$$A_2 = \frac{-\gamma_1(8\sigma^5 + 4\mu^4\sigma + 4\mu^2\sigma^3) + \gamma_2(\mu^5 - \mu^3\sigma^2) + 24\mu\sigma^4}{\sigma^5} + \frac{\ln(2\pi\sigma^2)(\gamma_1 4(\sigma^3 - \mu^2\sigma) + \gamma_2(\mu^3 - 3\mu\sigma^2))}{\sigma^3},$$

$$A_3 = \frac{\gamma_1 \mu^2 - \sigma^2}{12 \sigma^2} - \frac{\gamma_1^2 \mu \mu^4 - 10\mu^2\sigma^2 + 15\sigma^4}{72 \sigma^5} - \frac{\gamma_2 \mu - 3\sigma^2 + \mu^2}{48 \sigma^3} + \frac{\gamma_1 \gamma_2 45\sigma^4 \mu^2 + \mu^6 - 15\sigma^2 \mu^4 - 15\sigma^6}{144 \sigma^6} - \frac{\gamma_2^2 \mu 105\sigma^4 \mu^2 - 105\sigma^6 + \mu^6 - 21\sigma^2 \mu^4}{1152 \sigma^7},$$

$$A_4 = \exp \left(-\frac{\mu^2}{2\sigma^2} \right) \sqrt{\frac{2}{\pi}},$$

kde

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Odhad analogický k Gram-Charlierovu rozvoji pak vyjádříme jako

$$E_{GC} = A_1 - A_4 \left(\frac{A_2}{96} + A_3 \right). \quad (5.9)$$

Podobně ve vzorci pro odhad entropie na základě Edgeworthova rozvoje označíme

$$A_1 = \frac{1}{4}(1 + \ln(2\pi\sigma^2)) \left(1 + \operatorname{erf} \left(\frac{\mu}{\sqrt{2}\sigma} \right) \right),$$

$$\begin{aligned}
A_2 = & \frac{\gamma_1 \mu^2 - \sigma^2}{12 \sigma^2} - \frac{\gamma_1^2 \mu \mu^4 - 10 \mu^2 \sigma^2 + 15 \sigma^4}{72 \sigma^5} - \frac{\gamma_2 \mu - 3 \sigma^2 + \mu^2}{48 \sigma^3} \\
& + \frac{\gamma_1 \gamma_2 (45 \sigma^4 \mu^2 + \mu^6 - 15 \sigma^2 \mu^4 - 15 \sigma^6)}{144 \sigma^6} \\
& - \frac{\gamma_2^2 \mu (105 \sigma^4 \mu^2 - 105 \sigma^6 + \mu^6 - 21 \sigma^2 \mu^4)}{1152 \sigma^7} \\
& + \frac{\gamma_1^3 (105 \sigma^8 - 28 \mu^6 \sigma^2 + 210 \mu^4 \sigma^4 - 420 \mu^2 \sigma^6)}{864 \sigma^8} \\
& - \frac{\gamma_1^2 \gamma_2 (-36 \mu^6 \sigma^2 + 378 \mu^4 \sigma^4 - 1260 \mu^2 \sigma^6 + 945 \sigma^8 + \mu^8)}{3456 \sigma^9},
\end{aligned}$$

$$A_3 = \frac{\ln(2\pi\sigma^2)(\gamma_1^2(10\mu^3\sigma^2 - \mu^5 - 15\mu\sigma^4) + \gamma_2(-3\mu^3\sigma^2 + 9\mu\sigma^4) + 12\gamma_1(\mu^2\sigma^3 - \sigma^5))}{\sigma^5},$$

$$\begin{aligned}
A_4 = & \frac{72\sigma^6\mu + \gamma_1^2(5\mu^3\sigma^4 - 8\mu^5\sigma^2 + \mu^7) + 3\gamma_2(\mu^5\sigma^2 - \mu^3\sigma^4)}{\sigma^7} \\
& - \frac{12\gamma_1(\mu^2\sigma^5 + 2\sigma^7 + \mu^4\sigma^3)}{\sigma^7},
\end{aligned}$$

$$A_5 = \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \sqrt{\frac{2}{\pi}}.$$

Odhad analogický k Edgeworthovu rozvoji pak vyjádříme jako

$$E_{Edg} = A_1 + A_5 \left(\frac{1}{288} (A_3 - A_4) - A_2 \right). \quad (5.10)$$

Vzorce (5.9) a (5.10) jsou sice dlouhé, nicméně výpočetně rychlé. Na základě simulací můžeme říci, že odhady entropie nejsou nestranné, ale vychýlení je prakticky nulové. Oba vzorce jsme testovali na široké škále rozdělení. Obecně jsme dostali velmi dobré výsledky pro rozdělení s koeficientem variace $CV < 0,8$ s výjimkou inverzního Gaussova rozdělení (7). Pro toho rozdělení výsledky nebyly příliš dobré pro $CV < 0,05$ a $0,55 < CV < 0,8$, ovšem pro $0,05 < CV < 0,55$ jsme získali velmi dobré odhady. Nyní budeme diskutovat výsledky pro vybraná rozdělení podrobněji.

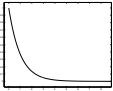
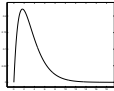
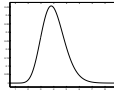
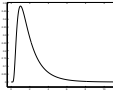
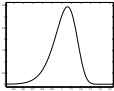
Pro gamma rozdělení (2) se vyhovující odhad entropie získá pro kombinace parametrů s koeficientem variace nižším než 0,82, dobrý s CV nižším než 0,77. To odpovídá parametru $\vartheta > 1,5$ resp. $\vartheta > 1,7$ Pro $\vartheta > 8,2$ je lepší E_{Edg} vzorec (5.10). U Weibullova rozdělení (5) dostaneme vyhovující odhad entropie pro kombinace parametrů s CV nižším než 0,84 ($b > 1,2$). Pro velmi nízké hodnoty parametru a (menší než 0,01) s CV nižším než 0,68 ($b > 1,4$).

Pro inverzní Gaussovo rozdělení (7) získáme dobrý odhad entropie pro kombinaci parametrů s CV od 0,05 do 0,55, přijatelný s koeficientem variace od 0,02 do 0,77. Lepší je E_{GC} vzorec (5.9). Exponenciální rozdělení (1) jako zvláštní případ Weibullova a Gamma rozdělení nevyhovuje námi stanovené podmínce ($CV = 1$). Teoreticky by byl vhodným odhadem průměr z odhadů E_{Edg} a E_{GC} , pro malé λ pak odhad E_{Edg} . Prakticky, pro málo pozorování, bude lepší odhad E_{GC} , protože 3. a 4. moment je podhodnocený. Pro malé λ bude opět lepší E_{Edg} .

5.5 Srovnání všech uvedených odhadů entropie

Nyní srovnáme 12 dříve uvedených metod odhadu entropie v kombinaci s odhady posunutí (4.1) a (4.7). Metody založené na odhadu hustoty rozdělení využívají odhad hustoty renormalizovaný na adekvátní interval. Při zkoumání vlivu posunutí na odhad entropie používáme posun $0,3 \times$ střední hodnota. Pokaždé jsme nasimulovali 10 opakování 50 nebo 250 pozorování. Jako srovnávací kritérium používáme střední kvadratickou chybu odhadu. Simulace byly generovány z rozdělení uvedených v tab. 5.4. Naším záměrem bylo testovat širokou škálu rozdělení o různé šikmosti.

Tab. 5.4: Rozdělení použita pro generování simulací při srovnávání metod odhadu entropie. Zahrnuli jsme exponenciální, gamma, inverzní Gaussovo a Weibullovo rozdělení.

Rozdělení	exp(1)	gam.(0,6;2)	gam.(5;20)	inv. Gauss(2;4)	Weib.(0,5;10)
Graf hustoty					
CV	1	0,7071	0,2236	0,7071	0,1203
Entropie	1	2,0880	1,2905	1,4556	-0,7138

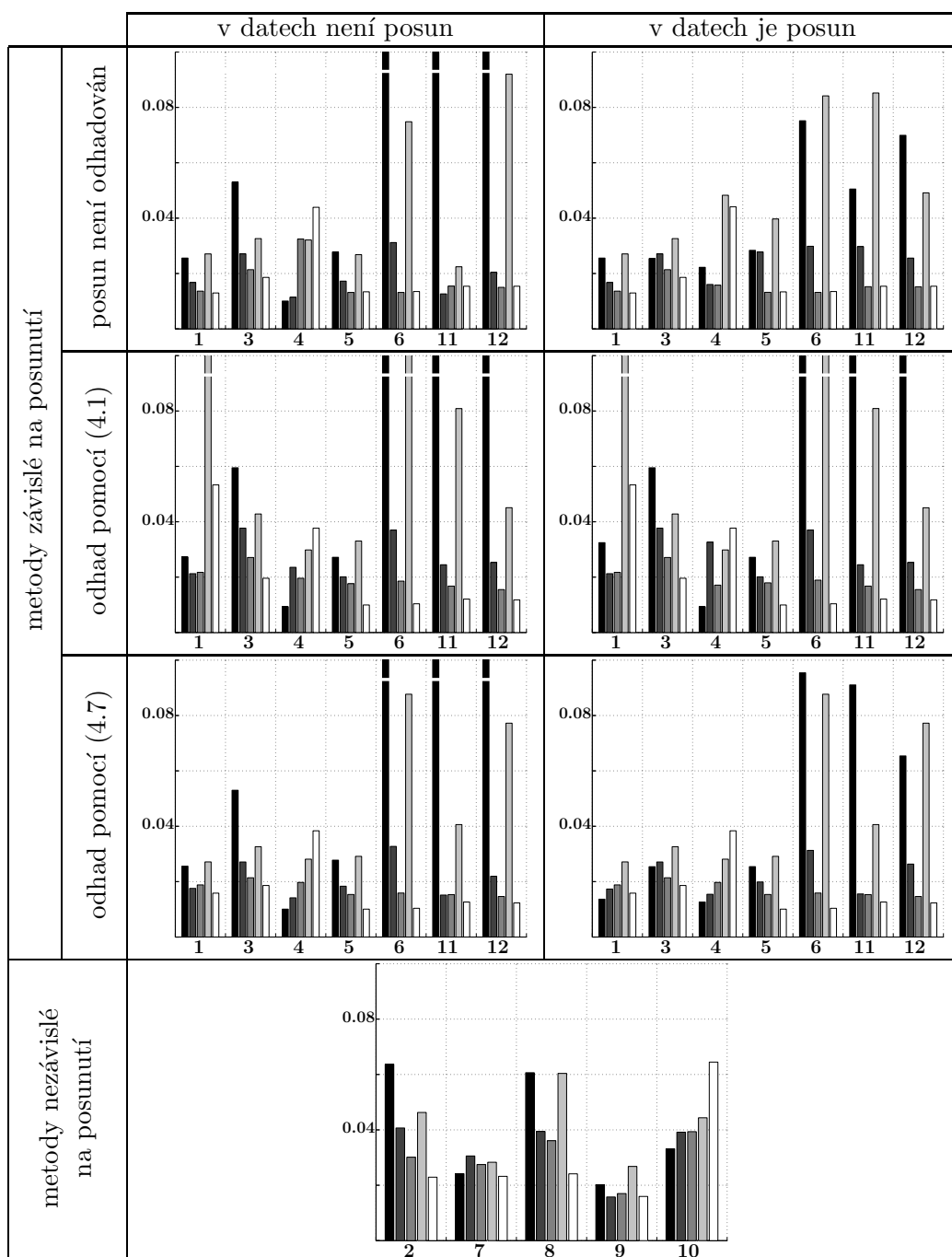
Pro srovnání jsme vybrali následující odhady entropie:

1. Parametrický odhad, kdy užitím metody maximální věrohodnosti spočteme parametry rozdělení a entropii odhadneme pomocí (5.1). Předpokládáme, že rozdělení je až na parametry známé.
2. Plug-in odhad pomocí histogramu. Použijeme histogram s neekvidistantními dělicími intervaly (3.8) konstruovaný dle pravidla (5.2) na intervalu daném jako $[t_{(1)}, t_{(n)}]$.

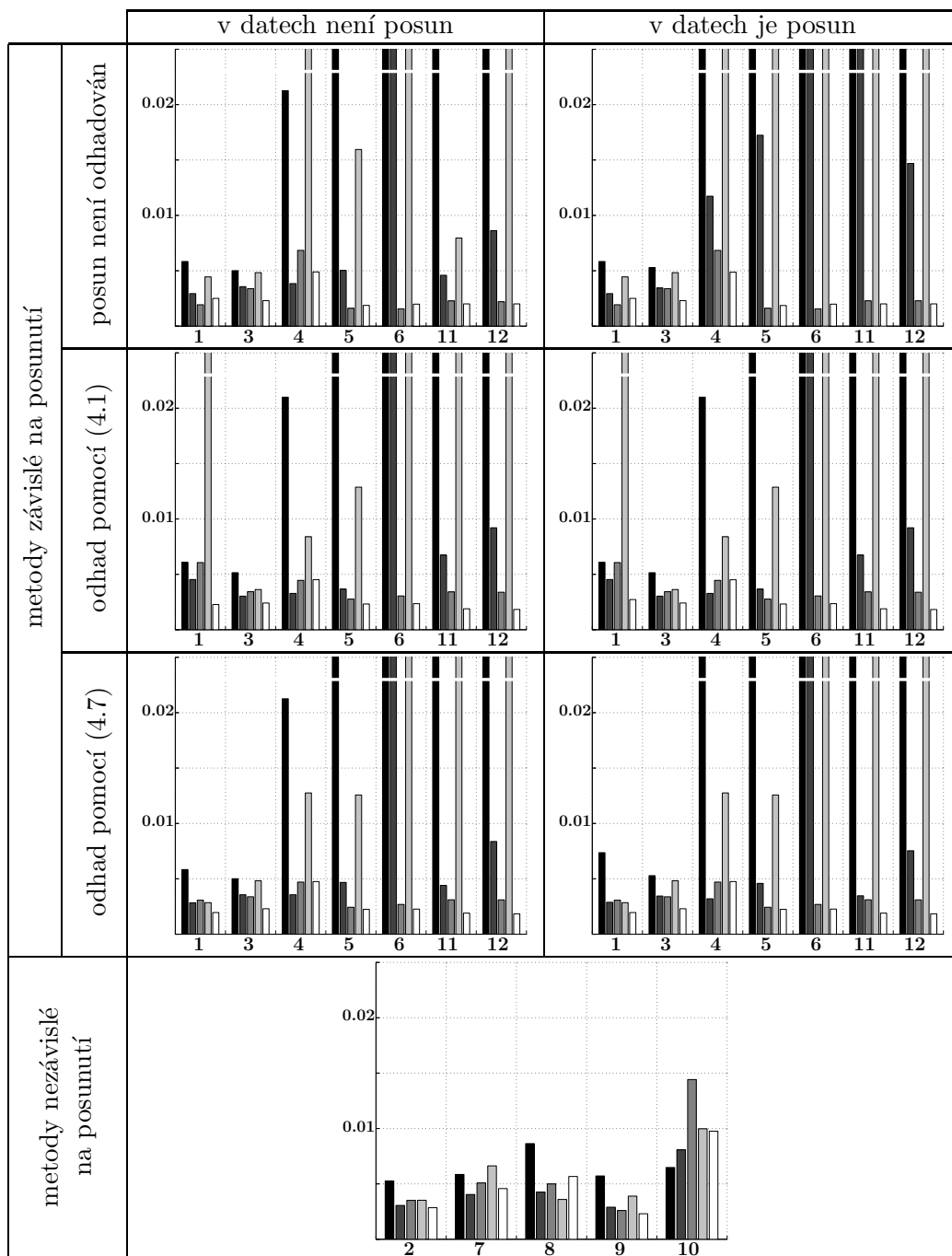
3. Plug-in odhad pomocí histogramu. Viz předchozí metoda, ale na intervalu $\left[\max\{t_{(1)} - \frac{st}{\sqrt{n}}, 0\}, t_{(n)} + \frac{st}{\sqrt{n}}\right]$.
4. Plug-in jádrový odhad (3.9). Aplikujeme často používané Epanečnikovo jádro $\left(\frac{3}{4}(1 - x^2)\right)$ pro $x \in [-1, 1]$ a 0 jinak), vyhlazovací parametr je určen pomocí metody křížové validace.
5. Plug-in odhad založený na Edgeworthově odhadu hustoty (3.14).
6. Plug-in odhad založený na Gramm–Charlierově odhadu hustoty (3.13).
7. Přímý odhad založený na „rozestupu vzorků“ (5.5). Volíme $m = 1$.
8. Přímý Vasickuv odhad (5.6). Používáme $m = 6$ pro 50 pozorování a $m = 13$ pro 250 pozorování.
9. Přímý „vylepšený“ Vasickuv odhad (5.7). Volíme $m = 4$.
10. Přímý odhad založený na „vzdálenosti od nejbližšího souseda“ (5.8).
11. Přímý momentový odhad založený na Gramm–Charlierově konceptu (5.9).
12. Přímý momentový odhad založený na Edgeworthově konceptu (5.10).

V naší práci můžeme vzhledem ke vlivu posunutí rozlišit mezi šesti různými situacemi: V datech není posun, a my jej neodhadujeme, nebo jej odhadneme pomocí minima (4.1), strana 52, anebo pomocí korigovaného minima (4.7), strana 54. Další možností je, že v datech je posun, a my jej opět buď neodhadujeme, nebo jej odhadneme pomocí minima (4.1) anebo pomocí korigovaného minima (4.7). Část metod odhadu entropie je na posunutí i na jeho případném odhadu nezávislá.

Na obr. 5.20 a 5.21 jsou znázorněny výsledky pro 50 a 250 simulovaných hodnot v jedné sérii. Na jejich základě můžeme vyslovit následující závěry. Přímé momentové odhady založené na Gram-Charlierově konceptu dávají výsledky srovnatelné s parametrickými a dalšími prověřenými odhady, ovšem pouze pro rozdělení s koeficienty variace v rozsahu daném v podkapitole 5.4.5. Dobrý výsledek byl dosažen pro oba rozsahy výběru. Ukazuje se, že tento odhad je silně závislý na odhadu posunutí. Analogicky bychom mohli hodnotit odhad založený na Edgeworthově konceptu, který se ale obecně jeví mírně horší než přímý odhad založený na Gram-Charlierově konceptu.



Obr. 5.20: Srovnání středních kvadratických chyb jednotlivých metod odhadu entropie. Použili jsme simulace o rozsahu 50 pozorování rozdělení uvedených v tab. 5.4. Černé sloupčky jsou použity pro exponenciální rozdělení, tmavě šedé pro rozdělení $\text{gamma}(0,6;2)$, šedé pro $\text{gamma}(5;20)$, světle šedé pro inverzní Gaussovo rozdělení a bílé pro Weibullovo rozdělení. Čísla pod sloupčky odkazují na výčet metod popsaných výše.



Obr. 5.21: Srovnání středních kvadratických chyb jednotlivých metod odhadu entropie. Použili jsme simulace o rozsahu 250 pozorování rozdělení uvedených v tab. 5.4. Černé sloupčky jsou použity pro exponenciální rozdělení, tmavě šedé pro rozdělení gamma(0,6;2), šedé pro gamma(5;20), světle šedé pro inverzní Gaussovo rozdělení a bílé pro Weibullovo rozdělení. Čísla pod sloupčky odkazují na výčet metod popsaných výše.

Histogram konstruovaný ve shodě s pravidlem (5.2) vskutku asymptoticky dává dobré výsledky odhadu entropie. Není vidět významná diference mezi metodami označenými 2. a 3., tento odhad tedy nebude příliš citlivý na odhad posunutí. Celkově nejlepším odhadem entropie se jeví vylepšený Vasickuv odhad (5.7).

Odhad posunutí minimem se v naší situaci zdá být příliš hrubý (reálný posun je nadhodnocen) a může způsobit významné znehodnocení odhadu entropie. Odhad (4.7) toto nadhodnocení snižuje a s výjimkou exponenciálního rozdělení a některých kombinací parametrů inverzního Gaussova rozdělení může kvalitu odhadu entropie výrazně zlepšit. Pokud v datech není žádný posun, ale my ho odhadujeme jako minimální hodnotu, pro menší počet pozorování dojde ke zhoršení odhadu entropie oproti stavu, kdy bychom posun neodhadovali. Pro větší počet pozorování se rozdíly stírají. Pokud v datech posun je, tak obecně můžeme říci, že pro menší rozsahy výběru použití odhadů posunutí (4.1) a (4.7) k vylepšení odhadu entropie nevede, ovšem pro větší rozsah pozorování je již zlepšení znatelné.

Kapitola 6

Odhad entropie z intervalově cenzorovaných dat

K odhadu entropie z intervalově cenzorovaných dat můžeme přikročit několika strategiemi.

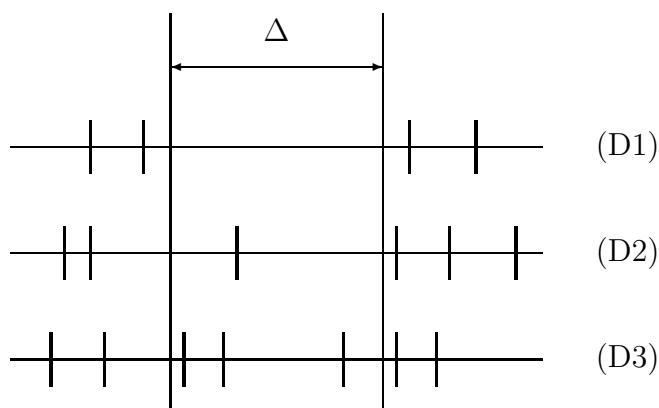
1. Můžeme pracovat s intervalově cenzorovanými daty, jako by šlo o kompletní pozorování. Budeme moci využít všechny již uvedené metody odhadu entropie, nicméně se vědomě dopouštíme chyby, když považujeme neúplná data za úplná.
2. Dále můžeme vyloučit z dat neúplná pozorování, a entropii odhadovat pouze na základě úplných. Podle situace nám zbude relativně málo pozorování, nutno bude použít metody vhodné pro malé vzorky. Zbavujeme se části informace, naše výpočty budou méně přesné než při další uvažované strategii.
3. Zohledníme charakter dat, a použijeme adekvátní, byť složitou, metodu. Využitelné budou jen plug-in odhady, do kterých dosadíme vhodný odhad hustoty.

Posledním případem se budeme zabývat v následujícím textu, na závěr uvedené přístupy srovnáme.

6.1 Intervalově cenzorovaná data

Předpokládejme, že proces obnovy je zaznamenáván v n intervalech („datových oknech“ či jen „oknech“) $[\tau_i, \tau_{i+1}]$, $i = 1, \dots, n$ relativně malé šířky $\Delta = \tau_{i+1} - \tau_i$ pro všechna $i = 1, \dots, n$. Obecně můžeme získat tři typy pozorování (viz obr. 6.1):

- (D1) V datovém okně se nenachází žádná událost.
- (D2) V datovém okně se nachází právě jedna událost.
- (D3) V datovém okně se nachází alespoň dvě události.



Obr. 6.1: Typy pozorování procesu obnovy v datovém okně.

Pouze typ (D3) obsahuje úplnou realizaci, zatímco v případě (D1) je interval mezi událostmi delší než datové okno a v případě (D2) máme o intervalu mezi událostmi pouze částečnou informaci.

6.2 Metoda založená na Vardiho algoritmu

V práci Vardi (1989) je uvedena metoda odhadu distribuční funkce z intervalově cenzorovaných dat pomocí neparametrické metody maximální věrohodnosti. K výpočtům je možno použít všech tří typů diskutovaných dat (viz obr. 6.1 a text výše). Na druhou stranu, tato metoda nerozlišuje mezi dopředným a zpětným rekurenčním časem (viz definici 3.12).

Nejdříve uvedeme zde použité označení dat. Nechť X_1, \dots, X_m a Z_1, \dots, Z_n jsou nezávislé stejně rozdělené náhodné veličiny s distribuční funkcí F a U_1, \dots, U_n jsou nezávislé stejně rozdělené náhodné veličiny z rovnoměrného spojitého rozdělení na intervalu $(0, 1)$. Neúplná pozorování Y_1, \dots, Y_n jsou definována jako

$$Y_i = Z_i U_i, \quad i = 1, \dots, n. \quad (6.1)$$

Nakonec tedy máme realizace úplných pozorování x_1, \dots, x_m a realizace neúplných pozorování y_1, \dots, y_n , které použijeme pro odhad distribuční funkce F .

Nechť t_1, \dots, t_h označuje navzájem různé hodnoty realizací $x_1, \dots, x_m, y_1, \dots, y_n$ seřazené vzestupně. V případě, že nemáme žádná data typu (D1), s vysokou pravděpodobností bude pro počet navzájem různých hodnot h platit vztah $h = m + n$. V opačném případě dostaneme h jako $h = m + n + 1 - n_{lr}$, kde n_{lr} je počet dat typu (D1). Obecně definujeme ξ_j a η_j jako četnosti hodnot x_1, \dots, x_m a y_1, \dots, y_n pro $t_j, j = 1, \dots, h$. Nyní můžeme pomocí nově zavedených veličin zapsat věrohodnostní funkci ve formě

$$L_I = \prod_{j=1}^h p_j^{\xi_j} \left(\sum_{k=j}^h \frac{p_k}{t_k} \right)^{\eta_j} \quad (6.2)$$

s parametry $p_j \geq 0$ ($j = 1, \dots, h$), $\sum_{j=1}^h p_j = 1$. Pro maximalizaci této věrohodnostní funkce použijeme EM algoritmus uvedený v práci Dempster et al. (1977). Počáteční odhady parametrů p_j označíme p_j^{old} (pro $j = 1, \dots, h$). Stanovíme je libovolně tak, aby platily podmínky $p_j^{old} > 0$ a $\sum_{j=1}^h p_j^{old} = 1$. Dále spočteme

$$p_j^{new} = (m + n)^{-1} \left[\xi_j + t_j^{-1} p_j^{old} \sum_{k=1}^j \eta_k \left(\sum_{i=k}^h t_i^{-1} p_i^{old} \right)^{-1} \right] \quad (6.3)$$

a iterujeme tak dlouho, až dosáhneme konvergence (ve smyslu nějakého vhodného pravidla); hodnoty $p_1^{new}, \dots, p_h^{new}$ jsou poté hledaným odhadem rozdělení intervalů mezi událostmi. Přesněji řečeno, tyto hodnoty dávají pouze částečný neparametrický odhad distribuční funkce na intervalu $(0, \Delta)$. Tento odhad lze zapsat jako

$$\hat{F}(v) = \begin{cases} 0 & v < t_1 \\ \sum_{i=1}^c p_i^{new} & t_c \leq v < t_{c+1} \text{ pro } c = 1, \dots, h-1. \end{cases} \quad (6.4)$$

Vzhledem k tomu, že platí $\sum_{j=1}^h p_j^{new} = 1$, je také možné zapsat formálně správný úplný odhad distribuční funkce

$$\hat{F}_1(v) = \begin{cases} 0 & v < t_1 \\ \sum_{i=1}^c p_i^{new} & t_c \leq v < t_{c+1} \text{ pro } c = 1, \dots, h-1 \\ \sum_{i=1}^h p_i^{new} = 1 & t_h \leq v. \end{cases} \quad (6.5)$$

Zde je ovšem (v závislosti na šířce datového okna) silně nadhodnocen odhad distribuční funkce pro $v \geq t_h$. Přesné využití odhadů $p_1^{new}, \dots, p_h^{new}$ pro náš účel je uvedeno v podkapitole 6.4.

6.3 Metoda založená na algoritmu McCleanové a Devineho

V práci McClean a Devine (1995) je použita jiná konstrukce odhadu distribuční funkce neparametrickou metodou maximální věrohodnosti. Tato metoda je určena pro data typu (D2) a (D3), narozdíl od přístupu z práce Vardi (1989) nelze využít typ dat (D1) (přehled typů dat viz obr. 6.1 a text výše). Předností této metody je, že jsou rozlišeny dopředné a zpětné rekurenční časy (viz definici 3.12).

Označme (t_i, x_i) , $i = 1, \dots, h$ uspořádané hodnoty úplných mezimpulsních intervalů s jejich případnými multiplicitami. Data typu (D1) neuvažujeme, zabýváme se spojitými daty. Pravděpodobnost takovýchto multiplicit je tedy zanedbatelná, ale v praxi k nim může dojít např. díky zaokrouhlování, snížené přesnosti přístroje a podobně. Symbolem $N_x = \sum_{i=1}^h x_i$ označíme počet navzájem rozdílných úplných pozorování. Trojicemi (t_{ij}, y_{ij}, z_{ij}) , $j = 1, \dots, c_i$, označíme hodnoty dopředných a zpětných rekurenčních časů (t_{ij}) v intervalu $[t_{i-1}, t_i)$, $i = 1, \dots$, s jejich případnými multiplicitami (y_{ij}) pro dopředný rekurenční čas a (z_{ij}) pro zpětný rekurenční čas. Poznamenejme, že $t_0 = 0$ a c_i je počet navzájem různých dopředných a zpětných rekurenčních mezimpulsních intervalů v intervalu $[t_{i-1}, t_i)$. Hodnoty $t_{h+1,j}$, $y_{h+1,j}$ a $z_{h+1,j}$, $j = 1, \dots, c_{h+1}$ analogicky definujeme na intervalu $[t_h, \infty)$. Nakonec definujeme celkový počet dopředných resp. zpětných rekurenčních mezimpulsních intervalů v intervalu $[t_{i-1}, t_i)$, $i = 1, \dots, h + 1$ jako $Y_i = \sum_{j=1}^{c_i} y_{ij}$ resp. $Z_i = \sum_{j=1}^{c_i} z_{ij}$ a celkový počet dopředných resp. zpětných rekurenčních mezimpulsních intervalů v celých datech jako $N_y = \sum_{i=1}^{h+1} Y_i$ resp. $N_z = \sum_{i=1}^{h+1} Z_i$.

Budeme rozlišovat mezi následujícími dvěma situacemi:

- (A) Největší pozorování je úplné. V tomto případě dostaneme věrohodnostní funkci (založenou stejně jako u předchozí metody na neúplných pozorováních) ve tvaru

$$L_I = (\mu^*)^{-N_y} \prod_{i=1}^h p_i^{x_i} \left(\sum_{k=i+1}^h p_k \right)^{Y_i + Z_i}, \quad (6.6)$$

kde $\mu^* = \sum_{i=1}^h p_i t_i$, kterou maximalizujeme vzhledem k parametrům $p_i > 0$, $i = 1, \dots, h$ a $\sum_{i=1}^h p_i = 1$.

- (B) Nejméně jedno neúplné pozorování je větší než největší úplný interval mezi událostmi. V tomto případě je nutno definovat $t_{h+1} > t_{hc_h}$ ale zároveň nejmenší možné. Formálně můžeme psát

$$t_{h+1} = t_{hc_h} + \varepsilon,$$

kde ε je neměřitelně malý časový interval. Poznamenejme, že pro korektní proveditelnost následujícího algoritmu je nutno položit $x_{h+1} = 1$. Věrohodnostní funkce je zde dána ve formě

$$L_I = (\mu^*)^{-N_y} \prod_{i=1}^h p_i^{x_i} \left(\sum_{k=i+1}^{h+1} p_k \right)^{Y_i + Z_i}, \quad (6.7)$$

kde

$$\mu^* = \begin{cases} \sum_{i=1}^{h+1} p_i t_i & \text{pro } t_{h+1} \leq \Delta, \\ \Delta p_{h+1} + \sum_{i=1}^h p_i t_i & \text{pro } t_{h+1} > \Delta. \end{cases}$$

Tuto věrohodnostní funkci maximalizujeme vzhledem k parametrům $p_i > 0$, $i = 1, \dots, h+1$ a $\sum_{i=1}^{h+1} p_i = 1$. Zápis této situace může být jednoduše redukován pro případ (A).

Pro maximalizaci stanovené věrohodnostní funkce je opět aplikován EM algoritmus (Dempster et al., 1977). Celá maximalizace je popsána v následující pětikrokové proceduře.

1. Najdeme startovací odhady pro p_1, \dots, p_{h+1} . My používáme $p_i = \frac{1}{h+1}$, $i = 1, \dots, h+1$.
2. Označíme stávající hodnoty p_1, \dots, p_{h+1} jako $p_1^{old}, \dots, p_{h+1}^{old}$ a určíme

$$\xi_i = x_i + \sum_{j=1}^i Z_j \pi_i^j \quad i = 1, \dots, h+1 \quad (6.8)$$

a

$$\eta_i = \sum_{j=1}^i Y_j \pi_i^j \quad i = 1, \dots, h+1 \quad (6.9)$$

kde

$$\pi_i^j = \begin{cases} \frac{p_i^{old}}{\sum_{k=j}^{h+1} p_k^{old}} & \text{pro } j \leq i, \\ 0 & \text{jinak.} \end{cases}$$

3. Řešení rovnice

$$\sum_{i=1}^{h+1} \frac{(\xi_i + \eta_i)t_i}{N_Y t_i + (N_x + N_z)\mu} = 1 \quad (6.10)$$

vzhledem k μ označíme $\hat{\mu}$. K získání řešení se nám ukázala jako vhodná metoda půlení intervalu.

4. Spočteme upravené odhady pro p_1, \dots, p_{h+1} pomocí vzorce

$$\hat{p}_i = \frac{(\xi_i + \eta_i)\hat{\mu}}{N_Y t_i + (N_x + N_z)\hat{\mu}} \quad i = 1, \dots, h + 1 \quad (6.11)$$

a označíme je $p_1^{new}, \dots, p_{h+1}^{new}$.

5. Použijeme vhodné kritérium konvergence, například když

$$\sum_{i=1}^{h+1} (p_i^{old} - p_i^{new})^2 < \epsilon \quad (6.12)$$

pro předem danou malou konstantu ϵ výpočet ukončíme. Pokud nerovnost (6.12) neplatí, tak položíme $p_i^{old} = p_i^{new}$, $i = 1, \dots, h + 1$ a přejdeme ke kroku 2. Opakujeme kroky 2. až 5. dokud nedosáhneme konvergence. Hodnoty $p_1^{new}, \dots, p_{h+1}^{new}$ potom považujeme za neparametrický maximálně věrohodný odhad rozdělení intervalů mezi událostmi.

Analogicky jako u předchozí metody můžeme definovat částečný a úplný odhad distribuční funkce rozdělení mezispulsních intervalů. Pro plug-in odhady je ovšem třeba stanovit odhad hustoty, což provedeme v následující podkapitole.

6.4 Odhad hustoty a entropie na základě výstupů předchozích metod

Z odhadů $p_1^{new}, \dots, p_h^{new}$ zformujeme histogram s r dělicími intervaly. Poznamenejme, že tento histogram je definován pouze na intervalu $(0, \Delta)$ a integrál z tohoto histogramu není roven jedné ale hodnotě $p_{hist} < 1$. Chybějící chvost odhadované hustoty

odhadneme pomocí hustoty exponenciálního rozdělení (1) s parametrem $a = 0$ a parametrem λ vhodně stanoveným. Tento parametr určíme jako řešení rovnice

$$\int_{\Delta}^{\infty} \frac{1}{\lambda} e^{-\lambda x} dx = 1 - p_{hist} \quad (6.13)$$

vzhledem k parametru λ . Po integraci dostaneme rovnici

$$\frac{1}{\lambda^2} e^{-\lambda \Delta} = 1 - p_{hist}. \quad (6.14)$$

Její řešení, opět např. pomocí metody půlení intervalu, označíme λ_{hist} . Nakonec dostaneme odhad hustoty pro celý definiční obor a ten použijeme pro odhad diferenciální entropie. Entropie odhadnutá na základě histogramu se spočte jako

$$E_{hist} = \sum_{i \in S} -(r_i - l_i) f_i \log(f_i), \quad (6.15)$$

kde S je indexová množina dělicích intervalů, v nichž je histogram nenulový, l_i resp. r_i je levá resp. pravá hranice i -tého intervalu a f_i je hodnota histogramu v i -tém dělicím intervalu. Část odhadu entropie vypočtené na základě exponenciálního chvostu je dána jako

$$E_{tail} = e^{-\lambda_{hist} \Delta} \left(\ln \left(\frac{1}{\lambda_{hist}} e^{-\lambda_{hist} \Delta} \right) - 1 \right) \quad (6.16)$$

a celková entropie je odhadnuta jako $E = E_{hist} + E_{tail}$.

6.5 Srovnání uvedených metod

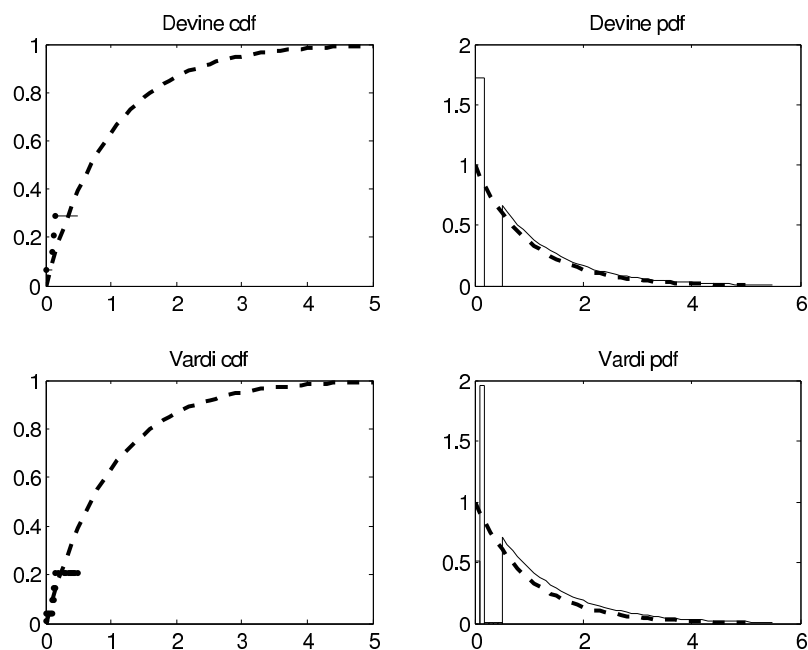
V této podkapitole bychom chtěli srovnat tři metody odhadu entropie na základě intervalově cenzorovaných dat. První dvě metody, tj. metoda založená na Vardiho algoritmu a metoda založená na algoritmu McCleanové a Devineho, se řadí mezi plug-in odhady a respektují v různé míře vlastnosti dat. Jako třetí metodu pro srovnání jsme zvolili přímý vylepšený Vasickuv odhad (5.7) určený k výpočtům s kompletními daty. V minulé kapitole jsme ukázali, že tento odhad je ve srovnání s ostatními metodami určenými pro odhad entropie z úplných dat velmi dobrý. Nyní chceme pomocí tohoto odhadu ilustrovat ztrátu kvality odhadu entropie v případě, že nerespektujeme podstatu dat. Vasickuv odhad použijeme k odhadu entropie ve dvou situacích:

1. Budeme uvažovat pouze úplné intervaly mezi událostmi. Nasazení Vasickova odhadu zde bude adekvátní, ovšem ztratíme velké množství informace. Pro malé šířky datového okna nemusíme získat žádná úplná pozorování.
2. Budeme uvažovat všechna pozorování a odhadovat entropii Vasickovym odhadem na základě úplných i cenzorovaných pozorování. Cenzorovaná pozorování budeme chybně považovat za úplná.

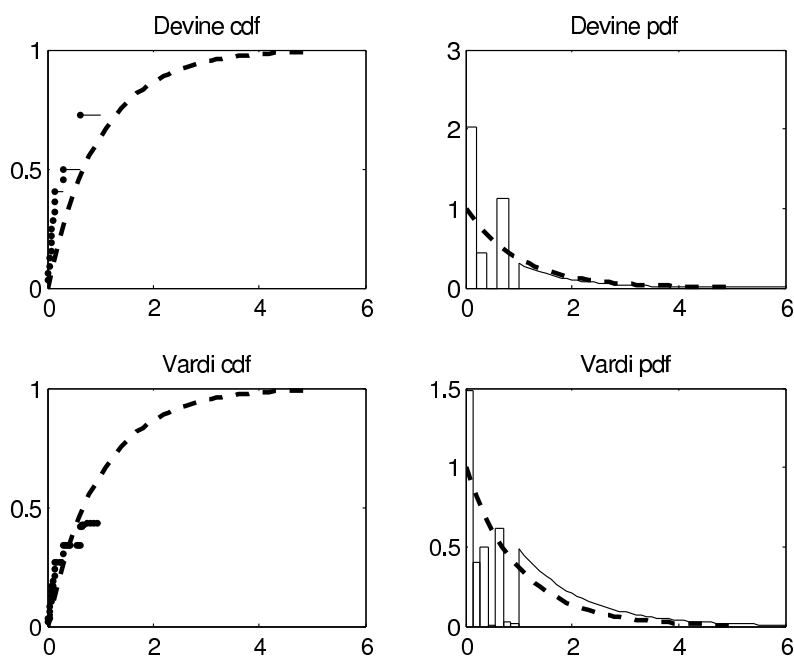
Uvedené metody budeme srovnávat pomocí odhadů ze simulovaných dat. Technika simulace je následující: Pro získání n vzorků generujeme n -krát 1000 realizací, které nám představují délky intervalů mezi událostmi. Tyto uspořádáme do posloupnosti, a pro každou tuto posloupnost náhodně určíme polohu datového okna o délce Δ (sice náhodně, ale tak, aby celé datové okno bylo mezi první a poslední událostí posloupnosti). Tímto způsobem získáme

- n_c pozorování úplných intervalů,
- $n_l \leq n$ zleva cenzorovaných intervalů,
- $n_r \leq n$ zprava cenzorovaných intervalů a
- $n_{lr} \leq n$ oboustranně cenzorovaných intervalů.

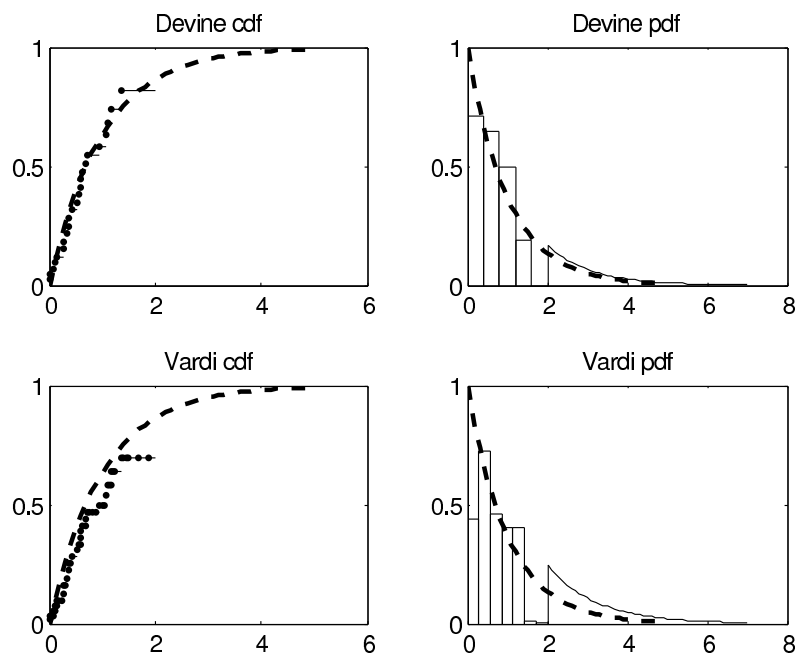
Nejdříve zhodnotíme kvalitu odhadu distribuční funkce a hustoty. Nejnázornější možností se jeví grafická prezentace dosažených odhadů. Na obr. 6.2 až 6.9 jsou vykresleny odhady pro distribuční funkci a hustotu exponenciálního rozdělení (1) s parametrem $\lambda = 1$. Analyzovány jsou situace s šířkou datového okna Δ 0,5; 1; 2 a 4 v kombinaci s počtem datových oken (25 a 100). Není třeba zkoumat odhady distribuční funkce a hustoty pro jiný parametr λ exponenciálního rozdělení, změny šířky datového okna jsou tomu ekvivalentní. Zkratkou „Devine“ označujeme metodu založenou na algoritmu McCleanové a Devineho, zkratkou „Vardi“ metodu založenou na Vardiho algoritmu.



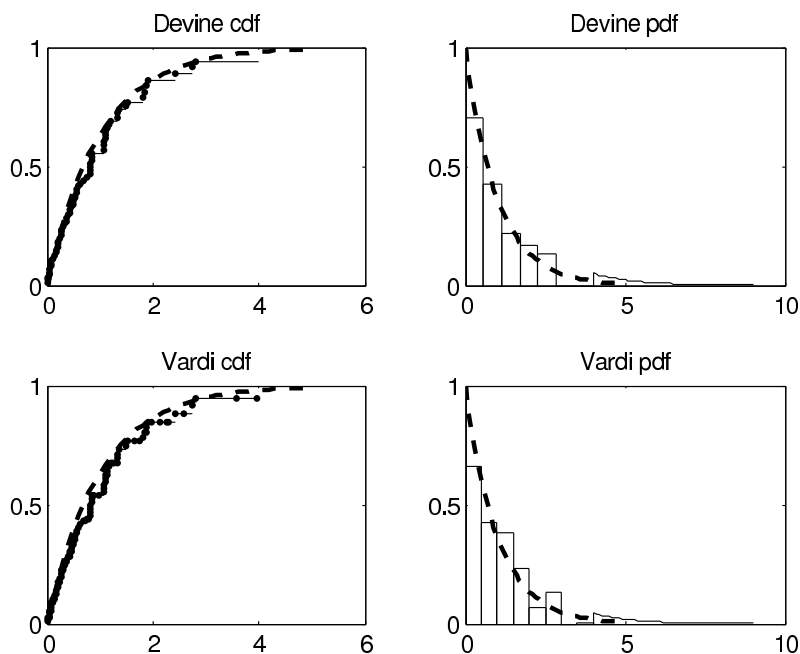
Obr. 6.2: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 25 datových oken o šířce $\Delta = 0,5$. Teoretické charakteristiky jsou vykresleny čárkovaně, odhady plnou čarou.



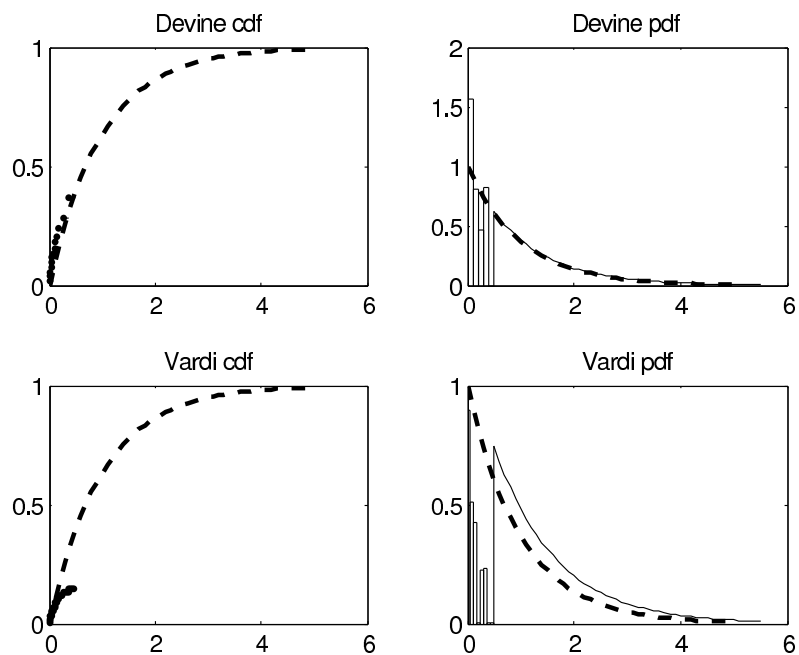
Obr. 6.3: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 25 datových oken o šířce $\Delta = 1$. Teoretické charakteristiky jsou vykresleny čárkovaně, odhady plnou čarou.



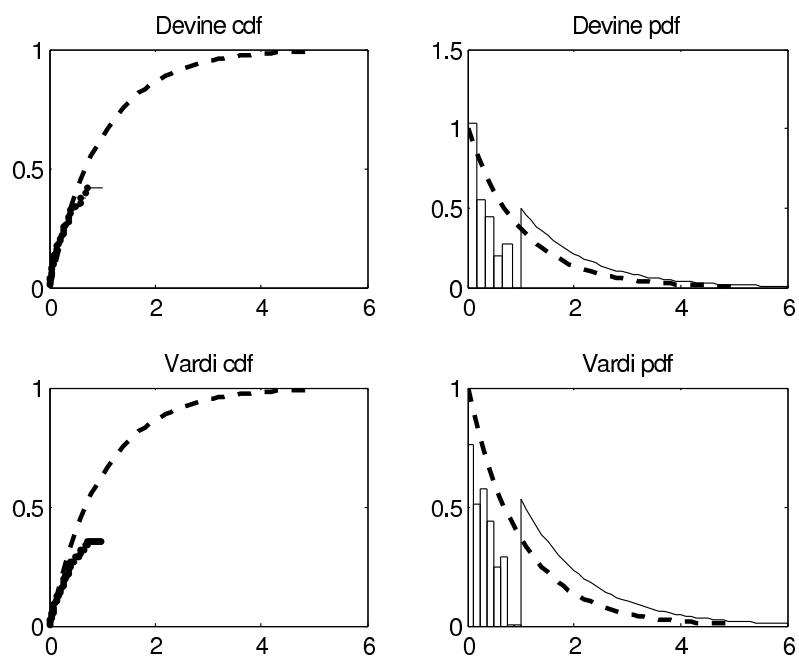
Obr. 6.4: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 25 datových oken o šířce $\Delta = 2$. Teoretické charakteristiky jsou vykresleny čárkovane, odhady plnou čarou.



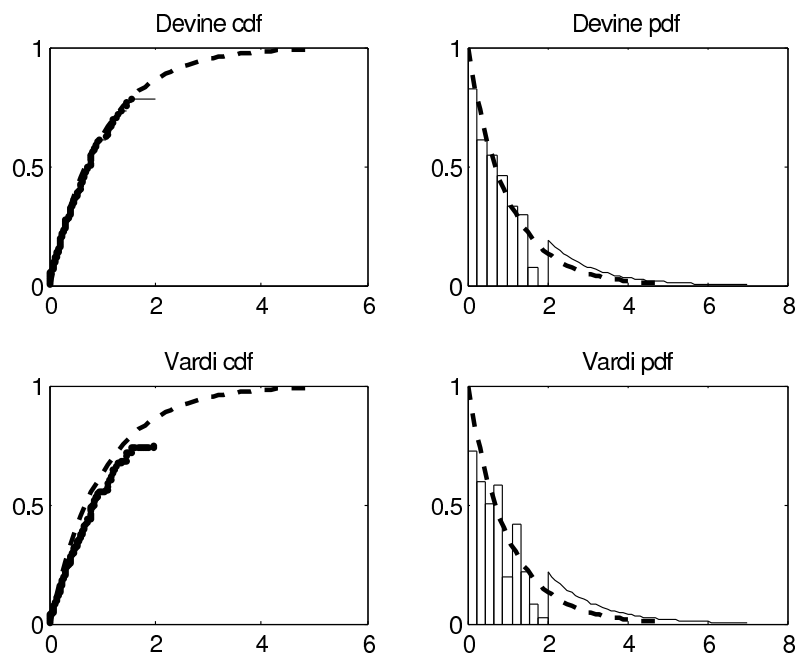
Obr. 6.5: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 25 datových oken o šířce $\Delta = 4$. Teoretické charakteristiky jsou vykresleny čárkovane, odhady plnou čarou.



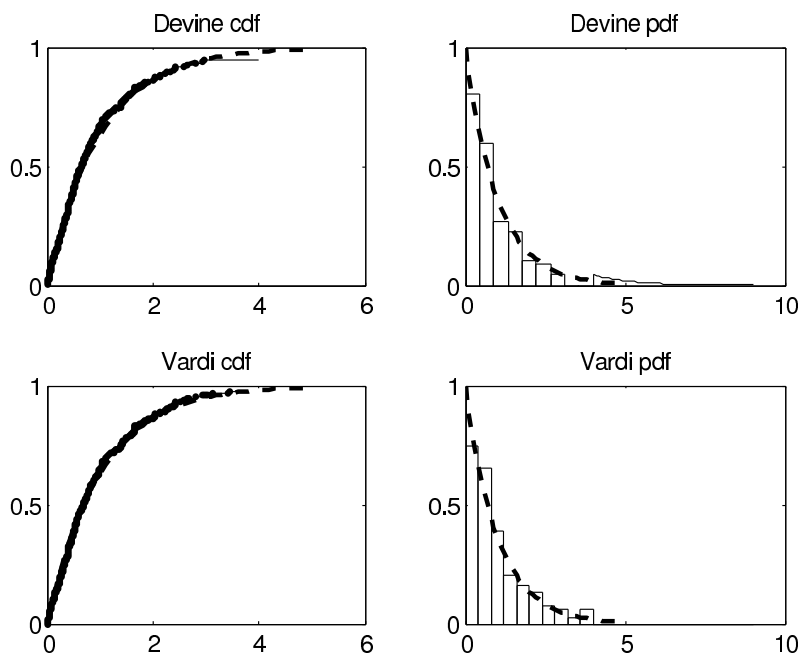
Obr. 6.6: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 100 datových oken o šířce $\Delta = 0,5$. Teoretické charakteristiky jsou vykresleny čárkovaně, odhady plnou čarou.



Obr. 6.7: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 100 datových oken o šířce $\Delta = 1$. Teoretické charakteristiky jsou vykresleny čárkovaně, odhady plnou čarou.



Obr. 6.8: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 100 datových oken o šířce $\Delta = 2$. Teoretické charakteristiky jsou vykresleny čárkovane, odhady plnou čarou.



Obr. 6.9: Odhad distribuční funkce (cdf) a hustoty (pdf) z intervalově cenzorovaných exponenciálních dat. Simulováno 100 datových oken o šířce $\Delta = 4$. Teoretické charakteristiky jsou vykresleny čárkovane, odhady plnou čarou.

Případ odhadu distribuční funkce a hustoty, kdy pracujeme s nejvíce omezenou informací (v rámci naší studie), je znázorněn na obr. 6.2. Šířka datového okna 0,5 je menší než střední hodnota generovaných simulací (ta je rovna jedné). Odhad se jeví jako velmi dobrý, což je ovšem dáno použitým rozdělením spíše než vlastnostmi odhadu. Při stejně širokém datovém okně, ale pro 100 pozorování (obr. 6.6) již vidíme rozdíly mezi metodami - metoda založená na algoritmu McCleanové a Devineho dává viditelně lepší odhad pro distribuční funkci i hustotu. Při posuzování odhadů pro šířku datového okna rovnu 1 se však výstupy obou metod jeví stejně kvalitně. Posouzení založené na výsledku pro šířku datového okna 0,5 tedy není možné zobecnit. Pro šířku datového okna 2 je již ale přínos metody založené na algoritmu McCleanové a Devineho systematický. Při šířce datového okna 4 dostáváme relativně velké množství úplných realizací mezimpulsních intervalů a rozdíly v odhadu použitými metodami se stírají.

Celkově se dá říci, že obě metody dávají odhady v předpokládané kvalitě. Pro relativně malou šířku datového okna dostaneme špatný, ne však nepoužitelný odhad. Při větším počtu datových oken kvalita odhadu roste, stejně jako při větších šířkách datového okna. Právě šířka datového okna je pro odhad distribuční funkce a hustoty zásadní. Můžeme například vidět, že odhad při 25 oknech o šířce 4 (obr. 6.5) je mnohem lepší než při 100 oknech o šířce 1 (obr. 6.7), ačkoliv celková data mají stejnou „délku“. Objektivnější hodnocení obou metod nám poskytne analýza odhadů entropie.

Výsledky samotného odhadu entropie (průměry a směrodatné odchylky) stokrát opakovaných simulací exponenciálního rozdělení (1) s parametrem $\lambda = 1$ jsme zpracovali do tab. 6.1 až 6.4. Entropii jsme odhadovali pro kombinace šířky datového okna Δ 0,5; 1; 2; 4 a počtu datových oken (25, 50, 75 a 100). Připomeňme, že teoretická diferenciální entropie je v případě Poissonova procesu s parametrem $\lambda = 1$ rovna 1. Kromě již zavedeného značení označíme zkratkou „Vasicek A“ Vasickuv vylepšený odhad pouze s úplnými daty a zkratkou „Vasicek B“ Vasickuv odhad se všemi typy dat.

Tab. 6.1: Výsledky odhadu entropie z intervalově cenzorovaných dat pro šířku datového okna $\Delta = 0,5$. Symbolem NaN (z anglického Not A Number) označujeme situaci, kdy výpočet není možno provést.

Počet oken		Devine	Vardi	Vasicek A	Vasicek B
25	průměr	0,8170	1,1305	$-\infty$	$-\infty$
	směr. odchylka	0,4153	0,0771	NaN	NaN
50	průměr	0,9006	1,1799	$-\infty$	$-\infty$
	směr. odchylka	0,2911	0,0537	NaN	NaN
75	průměr	0,9135	1,2057	-0,9636	$-\infty$
	směr. odchylka	0,2729	0,0420	0,2997	NaN
100	průměr	0,8577	1,2143	-0,9858	$-\infty$
	směr. odchylka	0,3087	0,0340	0,2691	NaN

Tab. 6.2: Výsledky odhadu entropie z intervalově cenzorovaných dat pro šířku datového okna $\Delta = 1$.

Počet oken		Devine	Vardi	Vasicek A	Vasicek B
25	průměr	0,9137	1,1719	-0,4016	0,0859
	směr. odchylka	0,2866	0,1315	0,4951	0,0393
50	průměr	0,9191	1,2045	-0,3155	0,0683
	směr. odchylka	0,2441	0,0974	0,1761	0,0275
75	průměr	0,8719	1,2210	-0,2573	0,0663
	směr. odchylka	0,2119	0,0893	0,1255	0,0269
100	průměr	0,8953	1,2382	-0,2729	0,0632
	směr. odchylka	0,1894	0,0829	0,0987	0,0217

Tab. 6.3: Výsledky odhadu entropie z intervalově cenzorovaných dat pro šířku datového okna $\Delta = 2$.

Počet oken		Devine	Vardi	Vasicek A	Vasicek B
25	průměr	0,9262	1,0706	0,3070	0,6202
	směr. odchylka	0,2183	0,1977	0,1461	0,0519
50	průměr	0,9415	1,1137	0,3474	0,6076
	směr. odchylka	0,1310	0,1026	0,0960	0,0305
75	průměr	0,9276	1,1314	0,3672	0,6076
	směr. odchylka	0,1424	0,0769	0,0611	0,0216
100	průměr	0,9673	1,1389	0,3720	0,6026
	směr. odchylka	0,1178	0,0770	0,0627	0,0203

Tab. 6.4: Výsledky odhadu entropie z intervalově cenzorovaných dat pro šířku datového okna $\Delta = 4$.

Počet oken		Devine	Vardi	Vasicek A	Vasicek B
25	průměr	0,9808	0,9788	0,8314	1,0358
	směr. odchylka	0,1298	0,1124	0,1114	0,0685
50	průměr	0,9872	1,0430	0,8393	1,0639
	směr. odchylka	0,0962	0,0821	0,0874	0,0307
75	průměr	0,9743	1,0317	0,8761	1,0530
	směr. odchylka	0,0774	0,0765	0,0694	0,0325
100	průměr	0,9886	1,0384	0,8983	1,0536
	směr. odchylka	0,0598	0,0685	0,0575	0,0295

Zásadní informací při hodnocení odhadnuté entropie je fakt, že použití uvedených metod je ve srovnání se situací, kdy neuvažujeme charakter dat, přínosné. Pro šířku datového okna 0,5 není dostupný Vasickuv odhad se všemi typy dat v žádném případě a Vasickuv odhad pouze s úplnými realizacemi až pro větší počet datových oken. Poznamenejme, že metoda založená na algoritmu McCleanové a Devineho stejně jako Vasickuv odhad entropie s úplnými realizacemi vyžadují alespoň jeden úplný mezimpulsní interval. V případě $\Delta = 0,5$ tato podmínka nebyla splněna asi v 1 % simulací, takové série dat byly z analýzy vyloučeny a nahrazeny vyhovujícími daty. Obě posuzované metody jsou v průměru vychýlené, metoda založená na algoritmu McCleanové a Devineho při šířce datového okna 0,5 systematicky entropii podhodnocuje, metoda založená na Vardiho algoritmu ji systematicky nadhodnocuje. Obecně, pro všechny zkoumané šířky, to ale neplatí. Odchylka výsledků metody založené na algoritmu McCleanové a Devineho je řádově vyšší než pro metodu založenou na Vardiho algoritmu. To je dáno zejména tím, že data typu (D1), v případě délky datového okna 0,5 relativně časté, využívá pouze metoda založená na Vardiho algoritmu.

Odchylka výsledků metody založené na algoritmu McCleanové a Devineho je vyšší, i když ne tak výrazně, i v případě datového okna o šířce 1. Zde je ale již patrné, že tato metoda dává výsledky průměrně bližší teoretické hodnotě entropie. Vasickuv odhad vždy vyjde velmi špatně. Pro šířku datového okna 2 a 4 je patrné přibližování odchylky výsledků metody založené na algoritmu McCleanové a Devineho a metody založené na Vardiho algoritmu. Tato hodnota je zde srovnatelná i s odchylkou Vasickova odhadu pouze s úplnými daty. Průměrný odhad posuzovaných metod je srovnatelný, pro šířku datového okna 2 je možné říci, že metoda založená na algoritmu McCleanové a Devineho dává lepší výsledky pro větší počet datových

oken, zatímco metoda založená na Vardiho algoritmu je úspěšná zejména pro menší počet datových oken.

Hlavním faktorem určujícím průměrnou kvalitu odhadu entropie je i zde šířka datového okna. Pro šířku datového okna 4 dostaneme srovnatelné průměrné výsledky i pro Vasickuv odhad se všemi typy dat, což je dáno velkou četností úplných realizací. Výsledky posuzovaných metod jsou ale vždy průměrně lepší.

Závěr

V předložené dizertační práci jsme nejdříve krátce prezentovali neurofyziologické základy, poté jsme je ve druhé kapitole zasadili do rámce matematických modelů samostatného neuronu. Zásadním východiskem pro další práci se stal fakt, že posloupnosti akčních potenciálů lze často považovat za realizace procesu obnovy. Tento je plně popsán hustotou mezipulsních intervalů, na které jsme se v převážné části práce zaměřili. Uvedli jsme přehled statistických definic, postupů a metod odhadu použitých při zkoumání odhadu posunutí a diferenciální entropie.

U odhadu posunutí jsme shrnuli dostupné parametrické i neparametrické metody a ověřili uplatnění nově navrženého odhadu. Můžeme konstatovat, že odhad pomocí minimální hodnoty, obvykle používaný pro posunutí, není nejlepší volbou. Porovnali jsme několik metod odhadu posunutí v kombinaci se třemi modely generování mezipulsních intervalů. Srovnali jsme jejich přesnost a asymptotické vlastnosti na simulovaných datech. Z výsledků plyne, že nejstabilnější výsledky dává aplikace neparametrických metod, konkrétně Cookova odhadu. Žádná z metod ale nedosahuje zřetelně lepších výsledků než ostatní metody. Námi navržený odhad dosáhl dobrých výsledků pro gamma model, ne však pro exponenciální model (jak jsme předpokládali). Analýza odhadů pro experimentální data poukázala na omezené možnosti užití srovnávaných metod.

Při odhadu entropie byly zkoumány plug-in odhady a přímé odhady. U prvně jmenovaných jsme se zaměřili zejména na volbu vnitřních parametrů dávající optimální odhad entropie. S tímto cílem jsme stanovili vztah pro počet dělicích intervalů v neekvidistantním histogramu. Ověřili jsme použitelnost momentových odhadů hustoty (a následně entropie). U přímých odhadů závislých na parametru jsme zjišťovali jeho optimální volbu. Zkonstruovali jsme dva přímé momentové odhady entropie a diskutovali jejich použitelnost. Všechny uvedené odhady jsme srovnali, a to v kombinaci se dvěma odhady posunutí. Námi navržené postupy dávají dobré výsledky, ovšem pouze pro vybrané kombinace rozdělení a koeficientu variace.

Při intervalově cenzorovaném záznamu dat v časovém okně získáme i neúplné realizace. Modifikovali jsme dva postupy původně určené pro odhad distribuční funkce z takovýchto dat, a zkonstruovali jsme postup pro výpočet odhadu entropie na základě jejich výstupů. Tyto odhady jsme diskutovali z hlediska délky časového okna a počtu záznamů. Srovnali jsme je s odhadem entropie vhodným pro úplné realizace. Námi prezentovaný způsob výpočtu entropie dává pro malou šířku časového okna vždy lepší odhady než neadekvátně použitý standardní přímý odhad entropie.

Prezentace

- Posterové sdělení na European Conference on Mathematical and Theoretical Biology, Dresden 18.–22. 7. 2005;
- Přednáška „Refractory period and its role in information transfer“, 1st International Summer School on Computational Biology, Brno 4.–5. 9. 2005;
- Posterové sdělení na Fifth Conference of the Czech Neuroscience Society, Praha 19.–21. 11. 2005;
- Přednáška „Estimation of the shift parameter in neurophysiological models“, Seminary of Department of Mathematical and Statistical Methods, AU Poznan, 27. 4. 2007;
- Přednáška „Odhad některých charakteristik neurofyziologických dat“, Výjezdní zasedání Centra Jaroslava Hájka pro teoretickou a aplikovanou statistiku, Telč, 25.–27. září 2008;
- Přednáška „Odhady posunutí rozdělení definovaných na polopřímce s aplikací v neurofyziologii“, seminář Statistika, Optimalizace a Funkcionální modelování, FSI VUT Brno, 9. 4. 2009.

Publikace

- Hampel D., 2005. Refractory period and its role in information transfer. Sborník 1st International Summer School on Computational Biology, Brno, s. 75–83.
- Hampel D., 2007. Estimation of Differential Entropy for Positive Random Variables and Its Application in Computational Neuroscience. In: Mathematical Modeling of Biological Systems, Volume II. Birkhäuser, Boston, s. 223–234.
- Hampel D., Lansky P., 2008. On the estimation of refractory period. Journal of Neuroscience Methods, 171, s. 288–295.

Literatura

- Abramowitz, M., Stegun, I. A., 1965. Handbook of mathematical functions, with formulas, graphs, and mathematical tables. Vol. 55. Superintendent of Documents, U.S. Government Printing Office, Washington, D.C.
- Albano, G., Giorno, V., Nobile, A. G., Ricciardi, L. M., 2007. A wiener-type neuronal model in the presence of exponential refractoriness. *BioSystems* 88, 202–215.
- Anděl, J., 1976. *Statistická analýza časových řad*. SNTL, Praha.
- Anděl, J., 1978. *Matematická statistika*. SNTL, Praha.
- Baillo, A., Cuevas, A., Justel, A., 2000. Set estimation and nonparametric detection. *Can. J. Stat.* 28, 765–782.
- Barbieri, R., Quirk, M. C., Frank, L. M., Wilson, M. A., Brown, E. N., 2001. Construction and analysis of non-poisson stimulus response models of neural spike train activity. *J Neurosci. Methods* 105 (1), 25–37.
- Beirlant, J., Dudewicz, E. J., Györfi, L., van der Meulen, E. C., 1997. Nonparametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.* 6 (1), 17–39.
- Berry II, M. J., Meister, M., 1998. Refractoriness and neural precision. *J. Neurosci.* 18 (6), 2200–2211.
- Bowers, N. L., 1966. Expansion of probability density functions as a sum of gamma densities with applications in risk theory. *Trans. Soc. Actuaries* 18, 125–137.
- Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., Frank, L. M., 2002. The time-rescaling theorem and its application to neural spike train data analysis. *Neural Comput.* 14, 325–346.

- Buonocore, A., Esposito, G., Giorno, V., Valerio, C., 2003. Towards dead time inclusion in neuronal modeling. *Sci. Math. Jpn.* 58 (2), 323–334.
- Buonocore, A., Giorno, V., Nobile, A. G., Ricciardi, L. M., 2002. A neuronal modeling paradigm in the presence of refractoriness. *BioSystems* 67, 35–43.
- Cooke, P., 1979. Statistical inference for bounds of random variables. *Biometrika* 66 (2), 367–374.
- Cover, T. M., Thomas, J. A., 1991. *Elements of information theory*. John Wiley & Sons Inc., New York.
- Cox, D. R., 1962. *Renewal Theory*. Methuen and Co. Ltd, London.
- Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* 39 (1), 1–38, with discussion.
- Doane, D. P., 1976. Aesthetic frequency classifications. *J. Amer. Statist.* 30, 181–183.
- Fenton, A. A., Csizmadia, G., Muller, R. U., 2000. Conjoint control of hippocampal place cell firing by two visual stimuli: I. the effects of moving the stimuli on firing field positions. *J. Gen. Physiol.* 116, 191–210.
- Fetz, E. E., Gerstein, G. L., 1963. An RC model for spontaneous activity of single neurons. *MIT Quarterly Report of the Res. Lab. of Electronics* 71, 249–256.
- Fiore, L., Lorenzetti, W., Ratti, G., 2005. Comparing neuronal spike trains with inhomogeneous poisson distribution: Evaluation procedure and experimental application in cases of cyclic activity. *J. Neurosci. Methods* 149, 7–14.
- Freedman, D., Diaconis, P., 1981. On the histogram as a density estimator: L_2 theory. *Z. Wahrsch. Verw. Gebiete* 57 (4), 453–476.
- Gerstner, W., Kistler, W. M., 2002. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, Cambridge.
- Guillory, K., Shoham, S., Normann, R., 2006. Discrete stimulus estimation from neural responses in the turtle retina. *Vis. Res.* 46, 1876–1885.

- Holden, A. V., 1976. *Models of the Stochastic Activity of Neurones*. Springer-Verlag, New York.
- Hátle, J., Likeš, J., 1974. *Základy počtu pravděpodobnosti a matematické statistiky*. SNTL - Alfa, Praha.
- Kagan, A. M., Linnik, Y. V., Rao, C. R., 1973. *Characterization problems in mathematical statistics*. John Wiley & Sons, New York-London-Sydney, translated from the Russian by B. Ramachandran, Wiley Series in Probability and Mathematical Statistics.
- Katz, B., 1966. *Nerve, Muscle and Synapse*. McGraw-Hill Book Company, New York.
- Koláček, J., Karunamuni, R. J., 2009. On boundary correction in kernel estimation of ROC curves. *Austrian Journal of Statistics* 38 (1), 17–32.
- Kostal, L., Lansky, P., 2006. Similarity of interspike interval distributions and information gain in a stationary neuronal firing. *Biol. Cybernet.* 94 (2), 157–167.
- Kostal, L., Lansky, P., 2007. Variability and randomness in stationary neuronal activity. *BioSystems* 89, 44–49.
- Kozachenko, L. F., Leonenko, L., 1987. Sample estimate of the entropy of a random vector. *Problems of Information Transmission* 23, 95–101.
- Lansky, P., Sanda, P., He, J., 2006. The parameters of the stochastic leaky integrate-and-fire neuronal model. *J. Comput. Neurosci.* 21 (2), 211–223.
- Lansky, P., Sato, S., 1999. The stochastic diffusion models of nerve membrane depolarization and interspike interval generation. *J Periph Nervous Syst* 4, 27–42.
- Lehky, S. R., 2004. Bayesian estimation of stimulus responses in poisson spike trains. *Neural Comput.* 16, 1325–1343.
- Lehmann, E. L., Casella, G., 1998. *Theory of Point Estimation*, 2nd Edition. Springer-Verlag, New York.
- Levick, W. R., Williams, W. O., 1964. Maintained activity of lateral geniculate neurones in darkness. *J Physiol Lond* 170, 582–597.
- Machek, J., 1980. *Teorie odhadu*. SPN, Praha.

- McClellan, S., Devine, C., 1995. A nonparametric maximum likelihood estimator for incomplete renewal data. *Biometrika* 82 (4), 791–803.
- Meunier, N., Marion-Poll, F., Lansky, P., Rospars, J. P., 2003. Estimation of the individual firing frequencies of two neurons recorded with a single electrode. *Chem. Senses* 28, 671–679.
- Ochs, S., 1965. *Elements of Neurophysiology*, 1st Edition. John Wiley & Sons, New York.
- Paninsky, L., 2003. Estimation of entropy and mutual information. *Neural Comput.* 15 (6), 1191–1253.
- Poggio, G. F., Vienerstein, L. J., 1964. Time series analysis of impulse sequences of thalamic somatic sensory neurons. *J Neurophysiol* 27, 517–545.
- Prášková, Z., Lachout, P., 2001. *Základy náhodných procesů*. Karolinum, Praha.
- Reeke, G. N., Coop, A. D., 2004. Estimating the temporal interval entropy of neuronal discharge. *Neural Comput.* 16, 941–970.
- Ricciardi, L. M., Esposito, F., 1966. On some distribution functions for non-linear switching elements with finite dead time. *Kybernetik* 3, 148–152.
- Rousseeuw, P. J., Croux, C., 1993. Alternatives to the median absolute deviation. *J. Amer. Statist. Assoc.* 88 (424), 1273–1283.
- Sampath, G., Srinivasan, S. K., 1977. *Stochastic Models for Spike Trains of Single Neurons*. Springer-Verlag, New York.
- Schaette, R., Gollisch, T., Herz, A., 2005. Spike-train variability of auditory neurons in vivo: Dynamic responses follow predictions from constant stimuli. *J. Neurophysiol.* 93, 3270–3281.
- Scott, D. W., 1979. On optimal and data-based histograms. *Biometrika* 66 (3), 605–610.
- Skvaril, J., Radil-Weiss, T., Bohdanecky, Z., Syka, J., 1971. Spontaneous discharge patterns of mesencephalic neurons: interval histogram and mean interval relationship. *Kybernetik* 9 (1), 11–15.

- Sturgers, H. A., 1926. Choice of a class interval. *J. Amer. Statist. Assoc.* 21, 65–66.
- Teich, M. C., Matin, L., Cantor, B. I., 1978. Refractoriness in the maintained discharge of the cat's retinal ganglion cell. *J. Opt. Soc. Am.* 68 (3), 386–402.
- Terrell, G. R., Scott, D. W., 1992. Variable kernel density estimation. *Ann. Statist.* 20 (3), 1236–1265.
- Tuckwell, H. C., 1988. Elementary applications of probability theory. Chapman & Hall, London.
- Vardi, Y., 1989. Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika* 76 (4), 751–761.
- Vasicek, O., 1976. A test for normality based on sample entropy. *J. Roy. Statist. Soc. Ser. B* 38 (1), 54–59.

Příloha A

Popis programových implementací

V následujícím přehledu uvádíme vybrané implementace prezentovaných modelů, metod a postupů ve výpočetním systému Matlab. Jejich funkčnost je ověřena pro verzi R2008b, vyžadován je Statistics toolbox. Kromě funkcí uvedených v následujícím výčtu jsou na přiloženém CD i další, pomocné funkce. Tyto zde sice nekomenujeme, ale jejich popis a vazba na další funkce jsou dostupné příkazem `help`.

Distribuční funkce alfa rozdělení (3).

$$F_x = \text{alfacdf}(\text{lam}, a, x)$$

Výstup:

`Fx` ... hodnota distribuční funkce

Vstupy:

`lam` ... parametr λ

`a` ... parametr a

`x` ... argument distribuční funkce

Inverzní funkce k distribuční funkci alfa rozdělení (3).

$$F_x = \text{alfaicdf}(\text{lam}, a, x)$$

Výstup:

F_x ... hodnota funkce inverzní k distribuční funkci

Vstupy:

lam ... parametr λ

a ... parametr a

x ... argument funkce

Hustota alfa rozdělení (3).

$$f_x = \text{alfapdf}(\text{lam}, a, x)$$

Výstup:

f_x ... hodnota hustoty

Vstupy:

lam ... parametr λ

a ... parametr a

x ... argument hustoty

Simulace realizací alfa rozdělení (3).

$$x = \text{alfasim}(\text{lambda}, a, n)$$

Výstup:

x ... vektor simulovaných hodnot

Vstupy:

lam ... parametr λ

a ... parametr a

n ... požadovaný počet realizací

Odhad posunutí dle Bailla.

$$[\text{ha1}, \text{ha2}] = \text{baillo}(\mathbf{X})$$

Výstupy:

ha1 ... odhad pro spojitý nosič (4.5)

ha2 ... odhad pro téměř spojitý nosič (4.6)

Vstup:

X ... data

Cookeův odhad posunutí (4.2).

$$\text{hata} = \text{cooke}(\mathbf{X})$$

Výstup:

hata ... odhad posunutí

Vstup:

X ... data

Odhad entropie z intervalově cenzorovaných dat na základě algoritmu McCleanové a Devineho.

$$[\text{Entropie}] = \text{devine}(\mathbf{Xc}, \mathbf{Xl}, \mathbf{Xr}, \mathbf{Xlr}, \text{delta})$$

Výstup:

Entropie ... odhadnutá entropie

Vstupy:

Xc ... vektor úplných pozorování

Xl ... vektor zleva cenzorovaných pozorování

Xr ... vektor zprava cenzorovaných pozorování

Xlr ... vektor oboustranně cenzorovaných pozorování

delta ... šířka datového okna

Konstrukce histogramu jako odhadu hustoty, vypočtení entropie z histogramu (3.8).
Dělicí intervaly jsou tvořeny tak, aby v každém byl stejný počet pozorování až na poslední.

$$[\text{Entrhist}, \text{odhpst}, \mathbf{s}, N, \text{edges}] = \text{histe}(\mathbf{X}, \text{del}, \mathbf{a}, \mathbf{b})$$

Výstupy:

`Entrhist` ... odhad entropie

`odhpst` ... vektor odhadu hustoty po intervalech

`s` ... středy dělicích intervalů

`N` ... počty pozorování po intervalech

`edges` ... hranice intervalů

Vstupy:

`X` ... řádkový vektor dat

`del` ... počet dělicích intervalů

`a` ... levý okraj histogramu

`b` ... pravý okraj histogramu

Teoretická entropie inverzního Gaussova rozdělení (7) počítaná numericky.

$$E = \text{igentr}(\text{xmax}, \mathbf{k}, \mu, \lambda)$$

Výstup:

`E` ... entropie

Vstupy:

`xmax` ... bod useknutí hustoty IG rozdělení

`k` ... jemnost vzorkování

`mu` ... parametr μ

`lambda` ... parametr λ

Generování simulací z inverzního Gaussova rozdělení (7).

$$\mathbf{x} = \text{igsim}(\mu, \lambda, n)$$

Výstup:

\mathbf{x} ... vektor simulovaných realizací

Vstupy:

μ ... parametr μ

λ ... parametr λ

n ... požadovaný počet realizací

Jádrový odhad hustoty (3.9). Tato funkce se využívá v rámci funkce `optjodh`, ale lze ji použít i samostatně.

$$[\text{jodhf}, \text{Entrj}] = \text{jodh}(a, b, k, X, h, \text{jadro})$$

Výstupy:

`jodhf` ... jádrový odhad hustoty

`Entrj` ... „jádrový“ odhad entropie

Vstupy:

a, b ... interval, na němž konstruujeme jádrový odhad hustoty

k ... dělicí krok pro interval (a, b) (pro výpočet entropie)

X ... data, jejichž hustotu odhadujeme

h ... parametr vyhlazovacího jádra

`jadro` ... v uvozovkách zadaný typ jádra, implementovány (jako funkce) jsou následující

typy jader: `biwj`, `cauj`, `epj`, `gauj`, `kosj`, `lapj`, `obj`, `triwj`, `trojj` (viz tabulka 3.1

a `help` pro jednotlivé funkce)

Simulace realizací Ornstein-Uhlenbeckova modelu.

```
t = LIFsim(tau, I, C, delta, sigma, Vth, limit, n)
```

Výstup:

t ... vektor realizací

Vstupy:

tau ... parametr τ

I ... proud I

C ... kapacitance C , $\mu = I/C$

delta ... parametr δ

sigma ... parametr σ

Vth ... prahové napětí

limit ... maximální počet kroků k dosažení prahu

n ... požadovaný počet simulací

Příklad neurofyziologicky korektního vstupu:

```
tau = 0.007
```

```
I = 0.5 * 10-7
```

```
C = 6 * 10-7
```

```
delta = 0.001
```

```
sigma = 0.07
```

```
Vth = 0.0007
```


Přímé momentové odhady entropie na intervalu $[a, \infty)$. Poznamenejme, že v práci uvádíme a diskutujeme pouze odhady na intervalu $[0, \infty)$ (5.9) a (5.10), zde je dostaneme volbou $a = 0$.

$$[\text{Egc}, \text{Eedg}] = \text{momev}(\text{sigma}, \text{mu}, \text{g1}, \text{g2}, \text{a})$$

Výstupy:

Egc ... entropie na základě Gram-Charlierova rozvoje

Eedg ... entropie na základě Edgeworthova rozvoje

Vstupy:

sigma ... odhad směrodatné odchylky

mu ... odhad střední hodnoty

g1 ... odhad šikmosti

g2 ... odhad špičatosti

a ... odhad posunutí

Odhad entropie založený na vzdálenosti od nejbližšího souseda (5.8).

$$\text{Hnearn} = \text{nearneighf}(\text{X})$$

Výstup:

Hnearn ... odhadnutá entropie

Vstup:

X ... data

Jádrový odhad hustoty (3.9) s optimálním h dle křížového ověření (3.10).

$$[\text{jodhf}, \text{Entrj}, \text{opth}] = \text{optjodh}(a, b, \text{krok}, X, \text{param}, k, \text{jadro})$$

Výstupy:

`jodhf` ... jádrový odhad hustoty

`Entrj` ... „jádrový“ odhad entropie

`opth` ... h dané křížovou validací

Vstupy:

`a`, `b` ... interval, na němž konstruujeme jádrový odhad hustoty

`krok` ... dělicí krok pro interval (a, b) (pro výpočet entropie)

`X` ... data, jejichž hustotu odhadujeme

`param` ... počáteční volba h

`k` ... maximální počet iterací optimalizačního algoritmu

`jadro` ... viz funkce `jodh`

Odhad entropie založený na rozestupu vzorků (5.5).

$$[\text{Hsam}] = \text{samsfaf}(X, m)$$

Výstup:

`Hsam` ... odhad entropie

Vstupy:

`X` ... data

`m` ... parametr m

Edgeworthův odhad hustoty (3.14).

```
tildef = tilfE(y, gamma1, gamma2)
```

Výstup:

tildef ... navzorkovaný odhad hustoty

Vstupy:

y ... standardizovaná data

gamma1 ... odhad šikmosti

gamma2 ... odhad špičatosti

Gram-Charlierův odhad hustoty (3.13).

```
tildef = tilfGC(y, gamma1, gamma2)
```

Výstup:

tildef ... navzorkovaný odhad hustoty

Vstupy:

y ... standardizovaná data

gamma1 ... odhad šikmosti

gamma2 ... odhad špičatosti

Odhad entropie z intervalově cenzorovaných dat na základě Vardiho algoritmu.

$$[\text{Entropie}] = \text{vardi}(\text{Xc}, \text{Xl}, \text{Xr}, \text{Xlr}, \text{delta}, \text{nc}, \text{nl}, \text{nr}, \text{nlr})$$

Výstup:

Entropie ... odhadnutá entropie

Vstupy:

Xc ... vektor úplných pozorování

Xl ... vektor zleva cenzorovaných pozorování

Xr ... vektor zprava cenzorovaných pozorování

Xlr ... vektor oboustranně cenzorovaných pozorování

delta ... šířka datového okna

nc ... počet úplných pozorování

nl ... počet zleva cenzorovaných pozorování

nr ... počet zprava cenzorovaných pozorování

nlr ... počet oboustranně cenzorovaných pozorování

Vasickovy odhady entropie (5.6) a (5.7).

$$[\text{Hvas}, \text{Hvasv}, \text{H}, \text{Hv}] = \text{vasodh}(\text{X}, \text{mn}, \text{mnv})$$

Výstupy:

Hvas ... Vasickuv odhad entropie

Hvasv ... vylepšený Vasickův odhad entropie

Vstupy:

X ... data

mn ... parametr pro Vasickův odhad

vmn ... parametr pro vylepšený Vasickuv odhad