

Univerzita Karlova v Praze  
Matematicko-fyzikální fakulta

## DIPLOMOVÁ PRÁCE



Barbora Vaňková

### Vývoj dynamického modelu pro odhad radonové zátěže budov

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Ing. Marek Brabec, Ph.D., Státní zdravotní ústav  
Studijní program: Matematika, matematická statistika

Ráda bych zde poděkovala panu Ing. Brabcovi, Ph.D. za trpělivý přístup a poskytnutí dat, panu doc. Mgr. Michalu Kulichovi, Ph.D. za velkou pomoc při přepracování práce a dále svojí rodině za podporu, které se mi od ní dostalo.

Prohlašuji, že jsem svou diplomovou práci napsala samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 4.8.2011

Barbora Vaňková

# Obsah

<b>1</b>	<b>Úvod</b>	<b>5</b>
<b>2</b>	<b>Data</b>	<b>7</b>
2.1	Základní charakteristiky dat . . . . .	8
<b>3</b>	<b>Odhad spojité funkce na základě diskrétních pozorování</b>	<b>11</b>
3.1	Volba vhodného modelu . . . . .	12
3.2	Odhad parametrů modelu . . . . .	13
3.3	Výběr báze . . . . .	18
3.3.1	Fourierova báze . . . . .	18
3.3.2	B-spline báze . . . . .	18
<b>4</b>	<b>Concurrent model</b>	<b>24</b>
4.1	Popis modelu . . . . .	24
4.2	Výběr vhodného modelu . . . . .	25
4.3	Odhad funkcí $\beta_1, \dots, \beta_P$ . . . . .	26
<b>5</b>	<b>Simulační studie</b>	<b>29</b>
5.1	Simulační studie I - Porovnání Fourierovy a B-spline báze . . . . .	29
5.2	Simulační studie II . . . . .	33
<b>6</b>	<b>Model závislosti OAR</b>	<b>37</b>
6.1	Funkcionální pozorování . . . . .	37
6.1.1	Model pro odhad funkcionálního pozorování . . . . .	37
6.1.2	Simulační studie IV. . . . .	39
6.1.3	Odhad funkcionálního pozorování . . . . .	42
6.2	Concurrent model . . . . .	46
6.2.1	Přehled modelů . . . . .	46
<b>7</b>	<b>Závěr</b>	<b>59</b>

Název práce: Vývoj dynamického modelu pro odhad radonové zátěže budov  
Autor: Barbora Vaňková  
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky  
Vedoucí diplomové práce: Ing. Marek Brabec, Ph.D.  
e-mail vedoucího: mbrabec@cs.cas.szu

Abstrakt: V předložené práci je popsána metoda odhadu funkcionálních dat na základě diskrétních pozorování a základní postupy funkcionální datové analýzy. Konkrétně se zaměřuje na vliv porušení předpokladů zobecněné cross-validační metody a na konstrukci concurrent modelu pro funkcionální data. Postupy jsou aplikovány na data, která popisují průběh objemové aktivity radonu.

Klíčová slova: Fourierova báze, B-spline, concurrent model

Title: Dynamic model for estimation of radon concentration in buildings  
Author: Barbora Vaňková  
Department: Department of probability and mathematical statistics  
Supervisor: Ing. Marek Brabec, Ph.D.  
Supervisor's e-mail address: mbrabec@cs.cas.cz

Abstract: In the present work there is described the method for estimation of functional data from discrete values and basic methods of functional data analysis.

Keywords: Fourier basis, B-spline, concurrent model

# Kapitola 1

## Úvod

Tato práce se zabývá aplikací funkcionální datové analýzy na data popisující průběh objemové aktivity radonu (dále OAR) měřené v testovacím objektu. Cílem je najít model popisující vztah mezi OAR, která byla naměřena ve dvou různých místnostech, popřípadě ukázat, že průběhy OAR spolu nesouvisí. Dále se budu zabývat testováním citlivosti metod funkcionální datové analýzy na nesplnění předpokladů apod.. K dispozici mám hodnoty OAR naměřené Státním ústavem radiální ochrany během osmnácti dnů v roce 2008. Práce je rozdělena do tří myšlenkových celků.

V první části je uveden popis dat, aby bylo zřejmé, na jakou situaci má být popsána teorie aplikovaná. V další kapitole jsou uvedené obecné definice základních pojmů. Není však cílem čtenáře seznámit s podrobnou teorií funkcionální datové analýzy (dále FDA), ale pouze zadefinovat značení a základní pojmy používané v dalším textu. Podrobně se této problematice věnuje kniha [6], na kterou navazuje kniha [5]. Druhá zmíněná publikace obsahuje popis praktické aplikace FDA v programovacím jazyku R (resp. S+) včetně konkrétních příkladů. Obě tyto knihy jsou napsány velmi srozumitelným a přehledným způsobem.

Druhá část práce se zabývá tím, jak převést diskrétní časovou řadu na funkcionální pozorování. Nezabývám se dopodrobna všemi možnostmi provedení, ale podrobně analyzuji a srovnávám dvě možnosti provedení. Odhad funkcionálního pozorování konstruuji jako lineární kombinaci spojitých diferencovatelných funkcí. Pro srovnání jsem vybrala dva typy báze s odlišnými vlastnostmi a sleduji vliv volby báze na výsledný model. Výběr bází je proveden tak, aby měly odlišné vlastnosti a jejich srovnání bylo zajímavé. Možností je celá řada, ale vybrala jsem dvě s co nejširším uplatněním. Data nevykazují nějaký speciální charakter (např. monotonií), proto jsem se držela co možná nejobecnějšího řešení.

Třetí část obsahuje simulační studie, které testují porušení předpokladů použitých metod apod.

Čtvrtá část je pro tuto práci klíčová. Snažím se v ní najít vhodný model pro průběh OAR a zároveň si vyzkoušet aplikaci funkcionální datové analýzy. Data,

která jsem měla k dispozici nejsou zcela typická pro aplikaci FDA, proto se v některých situacích jedná spíše o aproximaci. Průběh OAR během jednoho dne (k dispozici je více pozorování) považuji za funkcionální pozorování. Výstupem této části je model pro závislost OAR mezi dvěma sledovanými pokoji.

Výpočty byly provedeny za pomoci softwaru R 2.12.0 a jejich podrobné kódy jsou k dispozici na přiloženém CD.

# Kapitola 2

## Data

Data byla naměřena ve dnech 3.10.2008 až 20.10.2008 v rodinném domě Lažný v rámci studie vlivu užívání stavby na výsledné hodnoty objemové aktivity radonu. Měřila se OAR ( $Rn \cdot m^{-3}$ ) a teplota ( $^{\circ}C$ ). Měření probíhalo v pěti různých místnostech v daném objektu, ale jednotlivé časové řady se liší v časech měření. Pro dětský pokoj, kuchyň a chodbu je interval měření 30 minut, ale počátek měření je pro každou místnost jiný. Navíc došlo v jednom okamžiku k prodloužení intervalu měření, proto není možné předpokládat, že je dělení časového intervalu měření ekvidistantní. Ve zbylých dvou místnostech je interval měření 2 minuty a 60 minut. Přehled časové struktury měření je shrnut v tabulce (2.1).

Pro vývoj modelu jsem použila pouze data naměřená v dětském pokoji (dále značím  $p1$ ) a na chodbě (dále značím  $p2$ ), proto se v dalších odstavcích budu věnovat popisu pouze těchto dvou datových řad. Z fyzikálního pohledu hraje důležitou roli relativní vlhkost, která může hodnoty OAR významně ovlivnit. Tuto informaci bohužel nemám k dispozici, proto ji nemohu zahrnout do modelu. Vybrala jsem dětský pokoj a chodbu, protože měření probíhala v podobných časových okamžicích. Interval měření je v obou případech 30 minut. Navíc prodloužení časového intervalu nastalo mezi odpovídajícími si měřeními (tato podmínka není pro konstrukci modelu nutná).

Místo měření	Počátek měření	Konec měření	Interval měření
Dětský pokoj - p1	3.10.2008 16:10	20.10.2008 13:56	30 min
Chodba - p2	3.10.2008 16:04	20.10.2008 13:55	30 min
Kuchyň - p3	3.10.2008 15:58	20.10.2008 13:52	30 min
Kuchyň 1 - p4	3.10.2008 16:24	20.10.2008 15:00	2 min
Obývací pokoj - p5	3.10.2008 17:00	20.10.2008 15:00	60 min

Tabulka 2.1: Časová struktura měření

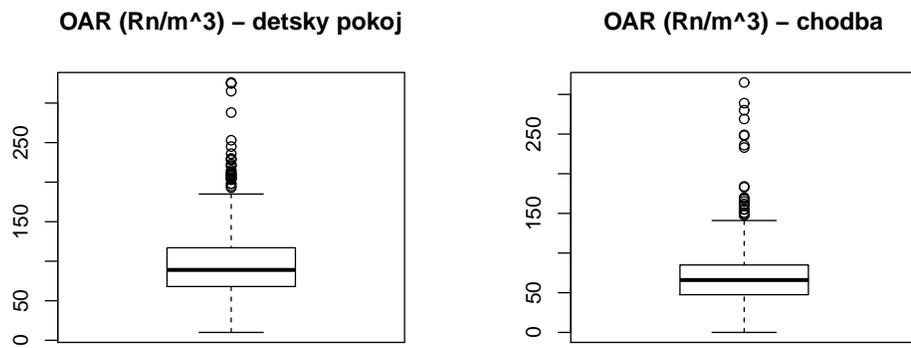
## 2.1 Základní charakteristiky dat

Jak jsem již zmínila, měření probíhalo od 3.10.2008 do 20.10.2008, tedy v průběhu 18 dní. Celkově bylo provedeno 812 měření v každé místnosti. Pro 16 dní je k dispozici měření popisující vývoj OAR v průběhu celého dne (48 měření pro každý den). První a poslední den je neúplný (první den proběhlo 16 měření a poslední den 28 měření). Tato data jsem pro odvození modelu nepoužila (a to ani v případě, že neúplnost dnů nebyla podstatná), protože jsem chtěla získat vzájemně porovnatelné výsledky. Z celkových 812 pozorování pro jednotlivé místnosti použiji tedy pouze pozorování 16 až 784. V obou místnostech došlo v jednom okamžiku k prodloužení intervalu mezi měřeními. V dětském pokoji i na chodbě nastalo toto prodloužení intervalu mezi 529. a 530. měřením. Pro dětský pokoj se prodloužil na 46 minut a pro chodbu na 51 minut.

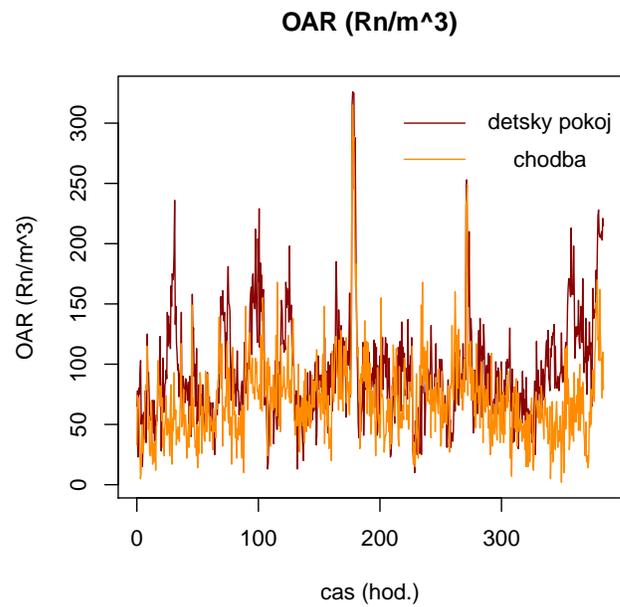
Na data budu pohlížet dvěma odlišnými způsoby. Nejprve je budu považovat za jednu datovou řadu (v tomto případě by nebylo nutné vyřazovat pozorování z prvního a posledního dne). Pro každý pokoj mám tedy k dispozici právě jedno funkcionální pozorování. Druhá možnost je data rozdělit na jednotlivé dny a pohlížet na ně jako na opakované pozorování průběhu OAR během jednoho dne. Tento pohled s sebou nese velkou míru zjednodušení. Podrobněji se tím zabývá kapitola (6).

Na obrázku (2.1) jsou pomocí krabicového grafu znázorněny základní statistické vlastnosti dat (bez rozlišení dnu měření) naměřených v dětském pokoji a na chodbě. Je vidět, že  $p1$  má vyšší průměr i směrodatnou odchylku než  $p2$ . Průběh OAR (naměřené ve zmíněných místnostech) v čase je vykreslen na obrázku (2.2). OAR naměřená v dětském pokoji je znázorněna červeně a OAR naměřená na chodbě oranžově. Průběh OAR v obou místnostech je na první pohled značně podobný. Ale při podrobnějším prozkoumání je vidět, že příčinou vyššího průměru a směrodatné odchylky u  $p1$  je několik výkyvů (kladným směrem) v první polovině měření a celkově vyšší naměřené hodnoty v závěru měření.

Na obrázku (2.3) je krabicový graf OAR naměřené v jednotlivých místnostech rozdělených podle dne měření (okraje vyznačeného obdélníku odpovídají prvnímu a třetímu kvartilu). Z obrázku (2.3) je patrné, že data vykazují podobné vlastnosti. Soubor  $p1$  vykazuje o něco vyšší hodnoty průměrů a směrodatných odchylek, ale nejedná se o výrazné rozdíly. V obou případech vykazují měření z 8. a 12. dne výrazně vyšší směrodatnou odchylku i průměr než data odpovídající zbylým dnům. Odlišnost 8. a 12. dne je dobře patrná z obrázku (2.4), na kterém je vykreslen vývoj OAR během jednotlivých dní. Výkyv se bohužel nedá vysvětlit např. náhlou změnou teploty apod. Pravděpodobně je způsoben změnou vnějších podmínek, která v datech není zachycena (např. změna relativní vlhkosti).

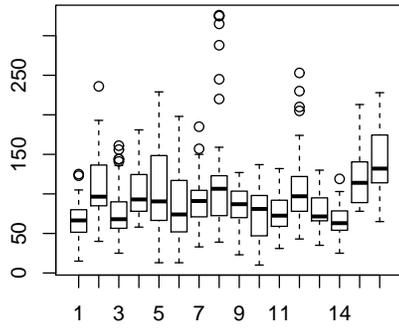


Obrázek 2.1: Krabicový graf pro data naměřená v dětském pokoji a na chodbě

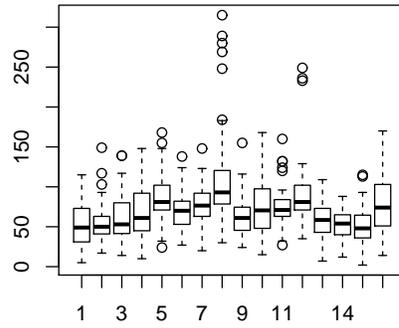


Obrázek 2.2: Průběh OAR v dětském pokoji ( $p1$ ) a na chodbě ( $p2$ ) bez rozlišení dnů měření

OAR (Rn/m<sup>3</sup>) po dnech – det. pokoj

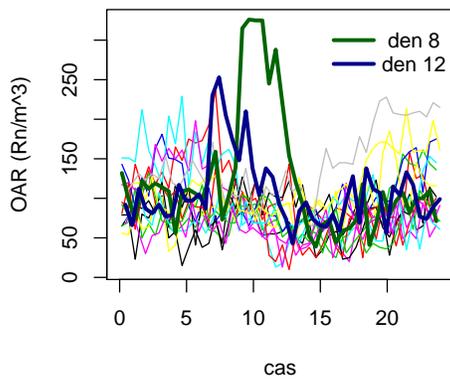


OAR (Rn/m<sup>3</sup>) po dnech – chodba

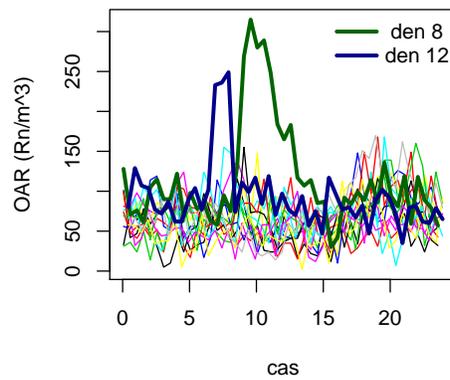


Obrázek 2.3: Krabicový graf (podle pokojů a dnů měření)

OAR po dnech – det. pokoj



OAR po dnech – chodba



Obrázek 2.4: Průběh OAR naměřený v dětském pokoji a na chodbě - zobrazeno po dnech měření. Zvýrazněn 8. a 12. den pro jejich odlišné vlastnosti.

# Kapitola 3

## Odhad spojité funkce na základě diskretních pozorování

Cílem této kapitoly je popsat postup odhadnutí spojité funkce na základě diskretních pozorování. Odhady budeme konstruovat jako lineární kombinace spojitých diferencovatelných funkcí. První část kapitoly se zabývá výběrem vhodného modelu a odhadem koeficientů lineární kombinace. Druhá část se zabývá volbou vhodného typu báze. Kapitulu jsem doplnila o příklad, na kterém demonstruji základní rozdíly mezi použitými bázemi.

Nechť  $y(t_j)$  jsou diskretní pozorování odpovídající modelu

$$y(t_j) = f(t_j) + \epsilon(t_j), \quad (3.1)$$

kde  $t_j \in \langle 0, T \rangle$  pro  $j = 1, \dots, J$  a  $\epsilon(t_j)$  jsou nezávislé, náhodné chyby s nulovou střední hodnotou a konečným konstantním rozptylem. Za pozorování  $y(t_j)$  budeme považovat hodnoty OAR naměřené v dětském pokoji (resp. na chodbě). Časový interval  $\langle 0, T \rangle$  odpovídá buď jednomu dni vyjádřenému v hodinách (pak  $T = 24$ ), nebo celkové době měření (pak  $T = 384$ ).

Předpokládejme, že funkce  $f$  je prvkem prostoru spojitých diferencovatelných funkcí, proto je možné ji vyjádřit jako lineární kombinaci funkcí báze tohoto prostoru, kterou budu značit  $\{\phi_l\}_{l=1}^{\infty}$ . Odhad  $\hat{f}$  budeme konstruovat jako prvek podprostoru prostoru spojitých diferencovatelných funkcí.

$$\hat{f}(t_j) = \sum_{l=1}^L \phi_l(t_j) c_l, \quad (3.2)$$

kde  $L$  je počet funkcí báze a  $\mathbf{c} = (c_1, \dots, c_L)$  je vektor lineárních koeficientů. V případě, že jsou funkce báze lineárně nezávislé, je počet parametrů modelu  $L$ . Dále pro lineárně nezávislou bázi platí, že pokud  $L = J$ , jedná se o saturovaný model a platí  $\hat{f}(t_j) = y(t_j)$ . Takový model je obvykle nepoužitelný. Zpravidla obsahuje

vysoký počet parametrů a nevykazuje dobré vlastnosti. Velmi často však slouží jako výchozí model při hledání optimálního modelu.

### 3.1 Volba vhodného modelu

Kvalitu modelu budeme posuzovat podle střední čtvercové chyby (chceme, aby byla co nejmenší)

$$MSE(\hat{\mathbf{f}}) = \frac{1}{J} \sum_{j=1}^J \left( y(t_j) - \hat{f}(t_j) \right)^2, \quad (3.3)$$

jejíž střední hodnotu je možné vyjádřit následujícím způsobem:

$$E \left\{ MSE(\hat{\mathbf{f}}) \right\} = \frac{1}{J} \sum_{j=1}^J E \left( y(t_j) - \hat{f}(t_j) \right)^2,$$

kde

$$\begin{aligned} E \left( y(t_j) - \hat{f}(t_j) \right)^2 &= E \left( y(t_j) - Ey(t_j) + Ey(t_j) - \hat{f}(t_j) \right)^2 = \\ &= E \left( y(t_j) - Ey(t_j) \right)^2 + E \left( Ey(t_j) - \hat{f}(t_j) \right)^2 + \\ &+ \left( Ey(t_j) \right)^2 - E \left( y(t_j) \hat{f}(t_j) \right) - \left( Ey(t_j) \right)^2 + Ey(t_j) E\hat{f}(t_j). \end{aligned}$$

Dále platí, že

$$\begin{aligned} E \left( y(t_j) \hat{f}(t_j) \right) &= E \left( \hat{f}(t_j) (f(t_j) + \epsilon(t_j)) \right) = \\ &= E \left( \hat{f}(t_j) f(t_j) \right) + E \left( \hat{f}(t_j) \epsilon(t_j) \right) = \\ &= E \left( \hat{f}(t_j) f(t_j) \right) + 0 = Ey(t_j) E\hat{f}(t_j), \end{aligned}$$

$$\begin{aligned} E \left( Ey(t_j) - \hat{f}(t_j) \right)^2 &= E \left( f(t_j) - \hat{f}(t_j) \right)^2 \\ &= E \left( f(t_j) - E\hat{f}(t_j) + E\hat{f}(t_j) - \hat{f}(t_j) \right)^2 = \\ &= E \left( f(t_j) - E\hat{f}(t_j) \right)^2 + E \left( E\hat{f}(t_j) - \hat{f}(t_j) \right)^2 = \\ &= bias^2 \left( \hat{f}(t_j) \right) + var \left( \hat{f}(t_j) \right). \end{aligned} \quad (3.4)$$

Střední hodnotu střední čtvercové chyby je možné zapsat jako:

$$\begin{aligned} E \left\{ MSE \left( \hat{\mathbf{f}} \right) \right\} &= \frac{1}{J} \sum_{j=1}^J \left( var \left( y \left( t_j \right) \right) + bias^2 \left( \hat{f} \left( t_j \right) \right) + var \left( \hat{f} \left( t_j \right) \right) \right) \\ &= var \left( \epsilon \left( t_j \right) \right) + bias^2 \left( \hat{f} \left( t_j \right) \right) + var \left( \hat{f} \left( t_j \right) \right). \end{aligned} \quad (3.5)$$

Jedná se o součet rozptylu náhodné chyby, vychýlení odhadu  $\hat{f}$  a rozptylu tohoto odhadu. Odhad  $\hat{f}$  je konstruován na základě pozorovaného průběhu OAR v dané místnosti. Pro různé průběhy získáme různé hodnoty odhadu  $\hat{f}$ . Vysoká hodnota vychýlení říká, že střední hodnota výsledného odhadu se značně liší od odhadované funkce  $f$ , ale že pro různá pozorování průběhu OAR bude dávat stabilní výsledky (to odpovídá nízké hodnotě rozptylu odhadu  $\hat{f}$ ). V opačném případě (vychýlení je nízké, ale rozptyl vysoký) bude střední hodnota odhadu blízká odhadované funkci  $f$ , ale jednotlivé realizace odhadu (pro různé průběhy OAR) budou velmi nestabilní. Dobrý model budeme tedy hledat tak, že budeme chtít co nejmenší střední čtvercovou chybu, ale zároveň se budeme snažit o dosažení kompromisu mezi rozptylem a vychýlením odhadu  $\hat{f}$ .

V případě, že je odhadovaná funkce  $f$  známá, budeme k ohodnocení modelu používat kritérium

$$MSE^* \left( \hat{\mathbf{f}} \right) = \frac{1}{J} \sum_{j=1}^J \left( f \left( t_j \right) - \hat{f} \left( t_j \right) \right)^2, \quad (3.6)$$

kteřé je možné vyjádřit jako součet druhé mocniny vychýlení odhadu  $\hat{f}$  a jeho rozptylu.

## 3.2 Odhad parametrů modelu

V této kapitole se budeme věnovat odhadu vektoru parametrů  $\mathbf{c}$  modelu (3.1). Jednou z možností, jak odhadnout vektor  $\mathbf{c}$  je minimalizovat součet čtverců ( $SSE$ ),

$$SSE(\mathbf{c}) = \sum_{j=1}^J \left[ y(t_j) - \sum_{l=1}^L \phi_l(t_j) c_l \right]^2 \quad (3.7)$$

přes všechna možná  $\mathbf{c}$ . Dále budeme používat označení  $\mathbf{y} = (y(t_1), \dots, y(t_J))'$ ,  $\mathbf{f} = (f(t_1), \dots, f(t_J))'$ ,  $\hat{\mathbf{f}} = (\hat{f}(t_1), \dots, \hat{f}(t_J))'$  a  $\Phi$  nechť je matice, jejíž sloupce tvoří funkce  $\phi_l$  v bodech  $t_1, \dots, t_J$  pro  $l = 1, \dots, L$ . Výraz (3.7) je možné zapsat

jako

$$\begin{aligned} SSE(\mathbf{c}) &= (\mathbf{y} - \Phi\mathbf{c})'(\mathbf{y} - \Phi\mathbf{c}) = \\ &= \mathbf{y}'\mathbf{y} - \mathbf{y}'\Phi\mathbf{c} - \mathbf{c}'\Phi'\mathbf{y} + \mathbf{c}'\Phi'\Phi\mathbf{c}. \end{aligned} \quad (3.8)$$

Tento výraz je minimální pro  $\hat{\mathbf{c}} = (\Phi'\Phi)^{-1}\Phi'\mathbf{y}$ , pokud existuje inverzní matice. Při použití bohaté báze bude střední čtvercová chyba vysoká díky vysokému rozptylu odhadu, což způsobí jeho nestabilitu. Aby k tomuto efektu nedocházelo a zároveň jsme nemuseli příliš omezit prostor funkcí nad kterým odhad konstruujeme, použijeme tzv. penalizovaný součet čtverců

$$\begin{aligned} PSSE_\lambda(\mathbf{c}) &= \sum_{j=1}^J \left[ y(t_j) - \sum_{l=1}^L \phi_l(t_j) c_l \right]^2 + \lambda PEN = \\ &= (\mathbf{y} - \Phi\mathbf{c})'(\mathbf{y} - \Phi\mathbf{c}) + \lambda PEN = \\ &= SSE(\mathbf{c}) + \lambda PEN, \end{aligned} \quad (3.9)$$

kde parametr  $\lambda$  nabývá hodnot z intervalu  $\langle 0, \infty \rangle$  a určuje míru vlivu penalizačního členu  $PEN$ :

$$\begin{aligned} PEN &= \int_T \left[ D^2 \hat{f}(t) \right]^2 dt = \\ &= \int_T (D^2 \Phi^*(t) \mathbf{c})^2 dt = \\ &= \int_T (D^2 \mathbf{c}' \Phi^{*'}(t)) (D^2 \Phi^*(t) \mathbf{c}) dt = \\ &= \mathbf{c}' \left[ \int_T (D^2 \Phi^{*'}(t)) (D^2 \Phi^*(t)) dt \right] \mathbf{c} = \\ &\stackrel{ozn.}{=} \mathbf{c}' \mathbf{R} \mathbf{c}, \end{aligned} \quad (3.10)$$

kde  $D^2$  značí druhou derivaci a  $\Phi^* = (\phi_1(t), \dots, \phi_L(t))$ . Matice  $\mathbf{R}$  má dimenzi  $L \times L$  a její prvky jsou tvaru  $R_{ij} = \int_T D^2 \phi_i(t) D^2 \phi_j(t) dt$ , kde  $i, j = 1, \dots, L$ . Prostřednictvím parametru  $\lambda$  je možné určit vliv penalizačního členu  $PEN$  na penalizovaný součet čtverců  $PSSE_\lambda(\mathbf{c})$ , a tím regulovat křivost výsledného odhadu  $\hat{f}$  (resp. jeho stabilitu).

- $\lambda = 0$ : Efekt penalizačního členu je nulový a pro dostatečně bohatou bázi platí, že  $\hat{f}(t_j) = y(t_j)$ .
- $\lambda \rightarrow \infty$ : Výsledný odhad je přímka.

Dále platí

$$\begin{aligned} PSSE_\lambda(\mathbf{c}) &= (\mathbf{y} - \Phi\mathbf{c})'(\mathbf{y} - \Phi\mathbf{c}) + \lambda\mathbf{c}'\mathbf{R}\mathbf{c} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{c}'\Phi'\mathbf{y} + \mathbf{c}'\Phi'\Phi\mathbf{c} + \lambda\mathbf{c}'\mathbf{R}\mathbf{c} \end{aligned} \quad (3.11)$$

$$D(PSSE_\lambda(\mathbf{c})) = -2\Phi'\mathbf{y} + 2\Phi'\Phi\mathbf{c} + 2\lambda\mathbf{R}\mathbf{c} \quad (3.12)$$

Označme  $\hat{\mathbf{c}} = \arg \min (PSSE_\lambda(\mathbf{c}))$ , potom platí  $D(PSSE_\lambda(\hat{\mathbf{c}})) = 0$  a  $\hat{\mathbf{c}}$  vyjádříme z rovnice

$$\begin{aligned} -2\Phi'\mathbf{y} + 2\Phi'\Phi\hat{\mathbf{c}} + 2\lambda\mathbf{R}\hat{\mathbf{c}} &= 0, \\ \Phi'\mathbf{y} &= (\Phi'\Phi + \lambda\mathbf{R})\hat{\mathbf{c}}, \\ (\Phi'\Phi + \lambda\mathbf{R})^{-1}\Phi'\mathbf{y} &= \hat{\mathbf{c}}. \end{aligned} \quad (3.13)$$

Odhad  $\hat{\mathbf{f}}$  lze zapsat jako

$$\hat{\mathbf{f}} = \Phi(\Phi'\Phi + \lambda\mathbf{R})^{-1}\Phi'\mathbf{y} \stackrel{\text{ozn.}}{=} \mathbf{S}_{\lambda\Phi}\mathbf{y}. \quad (3.14)$$

Hodnotu penalizačního parametru je možné stanovit dle vlastního úsudku na základě velmi dobré znalosti dat nebo pomocí automatických metod.

Za předpokladu, že pozorování  $y(t_j)$  jsou vzájemně nezávislá, lze použít metodu *cross-validace* nebo *zobecněné cross-validace*. První metoda je zde popsána na základě informací z knihy [3] a popis druhé je čerpán z knihy [6].

### Cross-validace (CV):

- Principem této metody je hledání odhadu, který bude co nejméně citlivý na vynechání jednotlivých pozorování při jeho konstrukci.
- Na základě datových bodů  $\mathbf{y}_{-j} = (y(t_1), \dots, y(t_{j-1}), y(t_{j+1}), \dots, y(t_J))'$  určíme pro  $j = 1, \dots, J$  odhady  $\hat{\mathbf{f}}_{-j} = (\hat{f}_{-j}(t_1), \dots, \hat{f}_{-j}(t_J))'$ . Označme symbolem  $\Phi_{-j}$  matici  $\Phi$  s vynechaným  $j$ -tým řádkem, pak

$$\hat{\mathbf{f}}_{-j} = \Phi(\Phi'_{-j}\Phi_{-j} + \lambda\mathbf{R})^{-1}\Phi'_{-j}\mathbf{y}_{-j}$$

- Definujme

$$\begin{aligned} CV(\lambda) &= \frac{1}{J} \sum_{j=1}^J [y(t_j) - \hat{f}_{-j}(t_j)]^2 \\ &= \frac{1}{J} \sum_{j=1}^J [y(t_j) - \Phi_{(j,\cdot)}(\Phi'_{-j}\Phi_{-j} + \lambda\mathbf{R})^{-1}\Phi'_{-j}\mathbf{y}_{-j}]^2, \end{aligned} \quad (3.15)$$

kde  $\Phi_{(j,\cdot)}$  značí  $j$ -tý řádek matice  $\Phi$ .

- Optimální  $\lambda$  získáme minimalizací  $CV(\lambda)$ .

### Zobecněná cross-validace (GCV):

- Definujme

$$GCV(\lambda) = \left[ \frac{n}{n - \text{tr}(\mathbf{S}_{\lambda\Phi})} \right] \left[ \frac{SSE}{n - \text{tr}(\mathbf{S}_{\lambda\Phi})} \right], \quad (3.16)$$

kde  $\text{tr}(\cdot)$  značí stopu matice.

- Optimální  $\lambda$  získáme minimalizací výrazu  $GCV(\lambda)$ .

Metoda GCV byla poprvé popsána v článku [2] a původně sloužila jako aproximace metody CV.

### Vztah mezi CV a GCV:

V předchozím textu (viz výraz 3.14) jsme odvodili vyjádření odhadu  $\hat{\mathbf{f}} = (\hat{f}(t_1), \dots, (t_J))'$ :

$$\hat{\mathbf{f}} = \mathbf{S}_{\lambda\Phi} \mathbf{y}.$$

Cross-validační kritérium pro hledání optimální hodnoty parametru  $\lambda$  je tvaru

$$CV(\lambda) = \frac{1}{J} \sum_{j=1}^J \left[ y(t_j) - \hat{f}_{-j}(t_j) \right]^2.$$

Zobecněné cross-validační kritérium jsme definovali

$$GCV(\lambda) = \left[ \frac{J}{J - \text{tr}(\mathbf{S}_{\lambda\Phi})} \right] \left[ \frac{SSE}{J - \text{tr}(\mathbf{S}_{\lambda\Phi})} \right].$$

V obou případech je dosaženo optimální hodnoty parametru  $\lambda$  právě když kritérium dosahuje minimální hodnoty.

Pokud nahradíme  $j$ -té pozorování vektoru  $\mathbf{y}$  hodnotou odhadu  $\hat{f}_{-j}(t_j)$ , označíme výsledný vektor jako

$$\tilde{\mathbf{y}}_j = \left( y(t_1), \dots, y(t_{j-1}), \hat{f}_{-j}(t_j), y(t_{j+1}), \dots, y(t_J) \right).$$

Odhad spojitě funkce zkonstruovaný na základě vektoru  $\tilde{\mathbf{y}}_j$  budeme značit  $\tilde{\mathbf{f}}_{-j}$ .

Dále budeme předpokládat, že platí

$$\tilde{f}_{-j}(t_j) = \hat{f}_{-j}(t_j)$$

pro všechna  $j = 1, \dots, J$ .

Pokud na pravé straně rovnice  $\hat{\mathbf{f}} = \mathbf{S}_{\lambda\Phi} \mathbf{y}$  nahradíme  $\mathbf{y}$  za  $\tilde{\mathbf{y}}_j$  a  $\mathbf{S}_{\lambda\Phi}$  budeme dále značit jako  $\mathbf{S}$ , platí

$$\left( \tilde{f}_{-j}(t_1), \dots, \tilde{f}_{-j}(t_J) \right)' = \mathbf{S} \tilde{\mathbf{y}}_j.$$

Dále platí

$$\hat{f}(t_j) = \sum_{i=1}^J S_{ji} y(t_i) = \sum_{i \neq j} S_{ji} y(t_i) + S_{jj} y(t_j),$$

$$\hat{f}_{-j}(t_j) = \tilde{f}_{-j}(t_j) = \sum_{i=1}^J S_{ji} \tilde{y}_j(t_i) = \sum_{i \neq j} S_{ji} \tilde{y}_j(t_i) + S_{jj} \hat{f}_{-j}(t_j) = \sum_{i \neq j} S_{ji} y(t_i) + S_{jj} \hat{f}_{-j}(t_j).$$

Odečtením těchto dvou rovnic získáme rovnici

$$\begin{aligned} \hat{f}(t_j) - \hat{f}_{-j}(t_j) &= S_{jj} \left( y(t_j) - \hat{f}_{-j}(t_j) \right), \\ \hat{f}(t_j) - y(t_j) &= S_{jj} \left( y(t_j) - \hat{f}_{-j}(t_j) \right) + \hat{f}_{-j}(t_j) - y(t_j), \\ \hat{f}(t_j) - y(t_j) &= (1 - S_{jj}) \left( \hat{f}_{-j}(t_j) - y(t_j) \right), \\ \frac{y(t_j) - \hat{f}(t_j)}{1 - S_{jj}} &= y(t_j) - \hat{f}_{-j}(t_j). \end{aligned}$$

Potom je možné cross-validační kritérium zapsat jako:

$$CV(\lambda) = \frac{1}{J} \sum_{j=1}^J \left( \frac{y(t_j) - \hat{f}(t_j)}{1 - S_{jj}} \right)^2$$

Pokud diagonální člen  $S_{jj}$  nahradíme průměrným diagonálním členem  $\frac{\text{tr}(\mathbf{S})}{J}$  získáme zobecněné cross-validační kritérium

$$GCV(\lambda) = \left[ \frac{J}{J - \text{tr}(\mathbf{S})} \right] \left[ \frac{\sum_{j=1}^J \left( y(t_j) - \hat{f}(t_j) \right)^2}{J - \text{tr}(\mathbf{S})} \right].$$

Dále je zřejmé, že pokud označíme  $\left( \frac{y(t_j) - \hat{f}(t_j)}{1 - S_{jj}} \right)^2$  jako  $CV_j(\lambda)$ , platí

$$CV(\lambda) = \frac{1}{J} \sum_{j=1}^J CV_j(\lambda)$$

a

$$GCV(\lambda) = \frac{1}{J} \sum_{j=1}^J CV_j(\lambda) \frac{J^2 (1 - S_{jj})^2}{(J - \text{tr}(\mathbf{S}))^2}.$$

### 3.3 Výběr báze

V předchozí části jsme popsali konstrukci odhadu funkce  $f$  na základě diskrétních pozorování  $y(t_j)$ ,  $j = 1, \dots, J$ . Připomeňme, že

$$\hat{f}(t_j) = \sum_{l=1}^L \phi_l(t_j) c_l, \quad (3.17)$$

kde  $L$  je počet funkcí dané báze a  $\mathbf{c} = (c_1, \dots, c_L)$  je vektor lineárních koeficientů. V této kapitole uvedeme dva typy báze, které se dále používají při konstrukci odhadu průběhu OAR. Jak již bylo řečeno dříve, předpokládáme, že funkce  $f$  je prvkem prostoru spojitých, diferencovatelných funkcí. Jednotlivé bazické funkce  $\phi_l$  jsou tedy také spojitě a diferencovatelné. Pro konstrukci odhadu průběhu OAR budeme používat Fourierovu bázi a B-spline. Kromě nich by bylo možné použít například jádrové funkce, bázi vhodnou speciálně pro monotónní odhady apod..

Fourierova báze se skládá z periodických funkcí. Není tedy příliš vhodná pro neperiodická data a nemusí vykazovat dobré vlastnosti zejména v okrajových bodech definičního oboru odhadované funkce. B-spline má polynomiální charakter a je vhodný pro neperiodická i periodická data.

#### 3.3.1 Fourierova báze

Tato báze je odvozena od Fourierových řad. V knize [6] je definována následovně:

**Definice 3.3.1** *Fourierova báze je definována jako  $\{\phi_l(t)\}_{l=1}^L$ . Jednotlivé funkce  $\phi_l(t)$  mají tvar*

$$\begin{aligned} \phi_1(t) &= 1, \\ \phi_{2r}(t) &= \sin r\omega t, \\ \phi_{2r+1}(t) &= \cos r\omega t, \end{aligned}$$

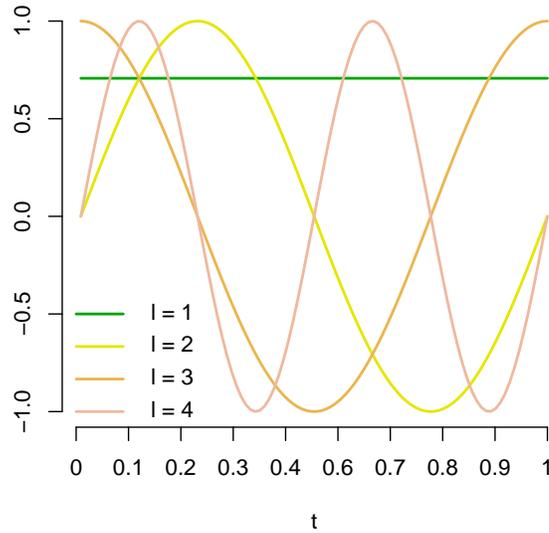
kde  $r$  je celé číslo a  $\omega$  parametr, který určuje délku periody  $2\pi/\omega$ .

Funkce  $\phi_l$  pro  $l = 1, 2, 3$  jsou zobrazeny na obrázku (3.1). Tato báze je periodická a je vhodná hlavně pro data, která jsou stabilní v čase, popřípadě mají periodický charakter.

#### 3.3.2 B-spline báze

Spline definujeme jako spojitou po částech polynomiální reálnou funkci. Existuje velké množství různých typů splinů. Jak již bylo řečeno, budeme se zabývat B-splinem.

### Fourierova baze



Obrázek 3.1: První tři funkce Fourierovy báze.

Uvažujme interval  $\langle t_1, t_J \rangle$  (v našem případě se jedná o interval  $\langle 0, T \rangle$ ) s vnitřními body (uzly)  $t_2, \dots, t_{J-1}$ . Obecně se polynomiální spline konstruuje tak, že na sub-intervalech mezi libovolnými dvěma sousedními uzly je definován polynom stupně  $S$  a v každém uzlu mají sousední polynomy stejnou hodnotu (výsledná funkce je spojitá). Pokud chceme hladkou (nejen spojitou)  $p$ -tou derivaci, musí být stupeň polynomu alespoň  $p+2$ . Nejběžněji používaný je kubický spline, tj.  $S = 3$ . Podrobný popis teorie splinů je možné nalézt v knize [1], ze které pochází také následující definice.

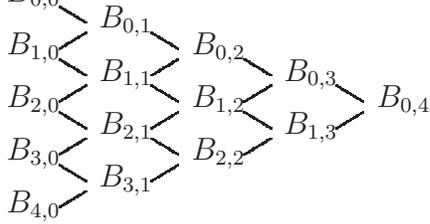
**Definice 3.3.2** *Nechť je  $\mathbf{s} = (s_0, \dots, s_{M-1})$  neklesající posloupnost bodů z intervalu  $\langle s_0, s_{M-1} \rangle$ . Pak pomocí rekurzivního vzorce definujeme funkce  $B_{m,k}$ , které tvoří B-spline bázi stupně  $k = 0, \dots, M - 2$ .*

$$B_{m,0}(t) := \begin{cases} 1 & \text{pro } s_m \leq s < s_{m+1}, \quad m = 0, \dots, M - 2 \\ 0 & \text{jinak,} \end{cases}$$

$$B_{m,k}(t) := \frac{s - s_m}{s_{m+k} - s_m} B_{m-1,k-1}(s) + \frac{s_{m+k+1} - s}{s_{m+k+1} - s_{m+1}} B_{m,k-1}(s), \quad m = 0, \dots, M - k - 2.$$

V případě, že se ve sčítanci vyskytuje výraz  $\frac{0}{0}$ , položíme ho roven 0 (tato situace nastává v případě vícenásobných uzlů).

Následující schéma popisuje rekurzivní závislost  $B_{m,k}$  na  $B_{m,k-1}$  a  $B_{m+1,k-1}$  (pro  $m = 0, \dots, 4$  a  $k = 0, \dots, 4$ ). Je zřejmé, že  $B_{m,0}(s)$  jsou nenulová pro  $s \in \langle s_m, s_{m+1} \rangle$  a  $B_{m,k}(s)$  pro  $s \in \langle s_m, s_{m+k} \rangle$ . Tato vlastnost je pro B-spline charakteristická.



V našem případě budeme předpokládat, že počet uzlů je  $J + 2k$ ,  $s_0 = \dots = s_k = t_1$ ,  $s_{k+j-1} = t_j$  pro  $j = 2, \dots, J - 1$  a  $s_{k+J-1} = \dots = s_{2k+J-1} = t_J$ . Znamená to, že množinu uzlů konstruujeme z  $(t_1, \dots, t_J)$  tak, že krajní uzel se v ní vyskytuje  $(k + 1)$ -krát.

Na obrázku (3.2) jsou vykresleny B-spliny stupně 0, 1 a 2. Dále platí, že počet funkcí báze = stupeň splinu + počet vnitřních uzlů + 1. V následující větě jsou shrnuty základní vlastnosti B-splinu.

**Věta 3.3.1** *Nechť  $B_{m,k}(s)$  je B-spline z definice (3.3.2) definovaný na intervalu  $\langle s_0, s_{M-1} \rangle$ , potom pro všechna  $k = 0, \dots, M - 2$  platí následující vlastnosti.*

- B-spline je složen z nezáporných funkcí, tj.

$$B_{m,k}(s) = 0, \quad s \notin \langle s_m, s_{m+k+1} \rangle \quad \text{a zároveň} \quad B_{m,k}(s) > 0, \quad s \in (s_m, s_{m+k+1}). \quad (3.18)$$

- Součet všech funkcí báze je v každém bodě intervalu  $\langle s_0, s_{M-1} \rangle$  roven 1, tj.

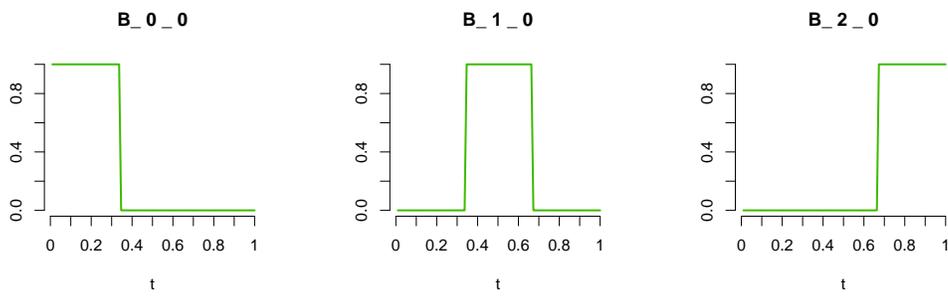
$$\sum_m B_{m,k}(s) = 1, \quad s \in \langle s_0, s_{M-1} \rangle. \quad (3.19)$$

*Důkaz:*

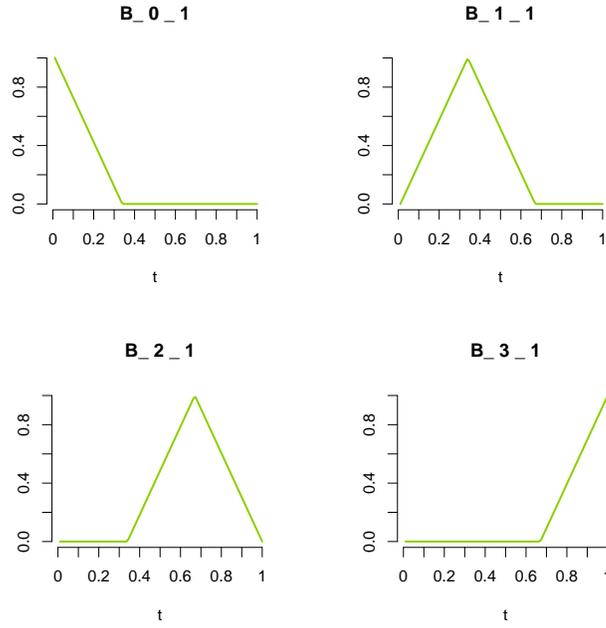
Nechť  $k = 0$ :

$$B_{m,0}(s) = \begin{cases} 1 & \text{pro } s_m \leq s < s_{m+1}, \\ 0 & \text{jinak,} \end{cases}$$

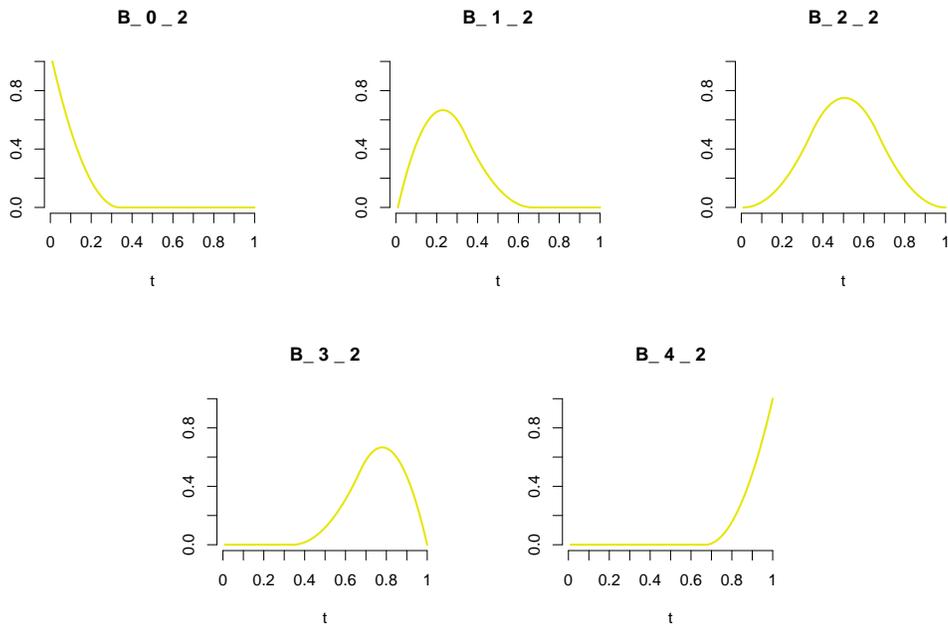
pro všechna  $m = 0, \dots, M - 2$ . Je zřejmé, že platí  $B_{m,0}(s) = 0$ ,  $s \notin \langle s_m, s_{m+1} \rangle$  a zároveň  $B_{m,0}(s) > 0$ ,  $s \in (s_m, s_{m+1})$ .



(a) B-spline báze stupně 0



(c) B-spline báze stupně 1



(e) B-spline báze stupně 2

Obrázek 3.2: Jednotlivé B-spliny stupně 0, 1 a 2

V dalším textu budeme používat značení  $B_{m,0}(s) = \mathbb{I}_{(s_m, s_{m+1})}(s)$ , kde  $s \in \langle s_0, s_{M-1} \rangle$ .

Nechť  $k = 1$ :

$$\begin{aligned} B_{m,1}(s) &= \frac{s-s_m}{s_{m+1}-s_m} B_{m,0}(s) + \frac{s_{m+2}-s}{s_{m+2}-s_{m+1}} B_{m+1,0} \\ &= \underbrace{\frac{s-s_m}{s_{m+1}-s_m} \mathbb{I}_{(s_m, s_{m+1})}(s)}_{\geq 0} + \underbrace{\frac{s_{m+2}-s}{s_{m+2}-s_{m+1}} \mathbb{I}_{(s_{m+1}, s_{m+2})}(s)}_{\geq 0}. \end{aligned}$$

První sčítanec je nenulový pouze na intervalu  $(s_m, s_{m+1})$  (na intervalu  $\langle s_m, s_{m+1} \rangle$  je to zaručeno funkcí  $\mathbb{I}_{(s_m, s_{m+1})}(s)$ ). Čítec i jmenovatel zlomku  $\frac{s-s_m}{s_{m+1}-s_m}$  nabývají na intervalu  $(s_m, s_{m+1})$  hodnot  $> 0$  (v krajním bodě  $s_m$  je čítec roven 0). Obdobný vztah platí i pro druhý sčítanec. Tím je dokázáno, že platí  $B_{m,1}(s) = 0$ ,  $s \notin \langle s_m, s_{m+2} \rangle$  a zároveň  $B_{m,0}(s) > 0$ ,  $s \in (s_m, s_{m+2})$ .

Předpokládejme, že věta platí pro  $k = i, i \geq 2$ :

Ukážeme, že platí i pro  $k = i + 1, i \geq 2$ :

$$B_{m,i+1}(s) = \underbrace{\frac{s-s_m}{s_{m+i+1}-s_m} B_{m,i}(s)}_{\geq 0} + \underbrace{\frac{s_{m+i+2}-s}{s_{m+i+2}-s_{m+1}} B_{m+1,i}}_{\geq 0}$$

Prvek  $B_{m,i}(s)$  je větší než 0 na intervalu  $\langle s_m, s_{m+i+1} \rangle$ . Na tomto intervalu nabývá čítec i jmenovatel zlomku  $\frac{s-s_m}{s_{m+i+1}-s_m}$  hodnot  $\geq 0$  (přičemž pro  $s = s_m$  je čítec roven 0). Tím je tvrzení věty dokázáno.

Při dokazování druhé části věty je důležité si uvědomit, že s rostoucím řádem splinu roste i počet multiplikativních kořenů použitých ke konstrukci báze (viz věta 3.3.1).

Nechť  $k = 0$ :

*Počet funkcí báze:*  $J - 1$

*Počet uzlů:*  $J$

*Uzly:*  $s_0 \leq \dots \leq s_{J-1}$

$$\sum_{m=0}^{J-1} B_{m,0}(s) = \sum_{m=0}^{J-1} \mathbb{I}_{(s_m, s_{m+1})}(s) = \mathbb{I}_{\langle s_0, s_{J-1} \rangle}(s)$$

Nechť  $k = 1$ :

Počet funkcí báze:  $J$

Počet uzlů:  $J + 2$

Uzly:  $s_0 = s_1 \leq \dots \leq s_J = s_{J+1}$

$$\begin{aligned}
\sum_{m=0}^{J-1} B_{m,1}(s) &= \sum_{m=0}^{J-1} \left\{ \frac{s-s_m}{s_{m+1}-s_m} B_{m,0}(s) + \frac{s_{m+2}-s}{s_{m+2}-s_{m+1}} B_{m+1,0}(s) \right\} \\
&= \sum_{m=0}^{J-1} \left\{ \frac{s-s_m}{s_{m+1}-s_m} \mathbb{I}_{\langle s_m, s_{m+1} \rangle}(s) + \frac{s_{m+2}-s}{s_{m+2}-s_{m+1}} \mathbb{I}_{\langle s_{m+1}, s_{m+2} \rangle}(s) \right\} \\
&= \underbrace{\frac{s-s_0}{s_1-s_0} \mathbb{I}_{\langle s_0, s_1 \rangle}(s)}_{=0} + \sum_{m=0}^{J-2} \left\{ \frac{s_{m+2}-s}{s_{m+2}-s_{m+1}} \mathbb{I}_{\langle s_{m+1}, s_{m+2} \rangle}(s) \right. \\
&\quad \left. + \frac{s-s_{m+1}}{s_{m+2}-s_{m+1}} \mathbb{I}_{\langle s_{m+1}, s_{m+2} \rangle}(s) \right\} + \underbrace{\frac{s-s_J}{s_{J+1}-s_J} \mathbb{I}_{\langle s_J, s_{J+1} \rangle}(s)}_{=0} \\
&= \sum_{m=0}^{J-2} \mathbb{I}_{\langle s_{m+1}, s_{m+2} \rangle}(s) = \mathbb{I}_{[s_1, s_J]}(s) = \mathbb{I}_{[s_0, s_{J+1}]}(s)
\end{aligned}$$

Nechť  $k = i$ :

Počet funkcí báze:  $J + i - 1$

Počet uzlů:  $J + 2i$

Uzly:  $s_0 = \dots = s_i \leq \dots \leq s_{J+i-1} = \dots = s_{J+2i-1}$

Předpokládejme, že

$$\sum_{m=0}^{J+i-2} B_{m,i}(s) = \mathbb{I}_{\langle s_0, s_{J+2i-1} \rangle}(s)$$

Nechť  $k = i + 1$ :

Počet funkcí báze:  $J + i$

Počet uzlů:  $J + 2(i + 1)$

Uzly:  $s_0 = \dots = s_{i+1} \leq \dots \leq s_{J+i} = \dots = s_{J+2i+1}$

$$\begin{aligned}
\sum_{m=0}^{J+i-1} B_{m,i+1}(s) &= \sum_{m=0}^{J+i-1} \left\{ \frac{s-s_m}{s_{m+i+1}-s_m} B_{m,i}(s) + \frac{s_{m+i+2}-s}{s_{m+i+2}-s_{m+1}} B_{m+1,i}(s) \right\} \\
&= \underbrace{\frac{s-s_0}{s_{i+1}-s_0} B_{0,i}(s)}_{=0} + \sum_{m=0}^{J+i-2} \left\{ \frac{s_{m+i+2}-s}{s_{m+i+2}-s_{m+1}} B_{m+1,i}(s) \right. \\
&\quad \left. + \frac{s-s_{m+1}}{s_{m+i+2}-s_m} B_{m+1,i}(s) \right\} + \underbrace{\frac{s_{J+2i+1}-s}{s_{J+2i+1}-s_{J+i}} B_{J+i,i}(s)}_{=0} \\
&= \sum_{m=0}^{J+i-2} B_{m+1,i}(s) = \mathbb{I}_{\langle s_0, s_{J+2i+2} \rangle}(s)
\end{aligned}$$

Poslední rovnost plyne z předpokladu pro  $k = i$  v případě, že přechísľujeme uzly splinu  $s_0, \dots, s_{J+2i} \rightarrow s_1, \dots, s_{J+2i+1}$ . Tím je i druhé tvrzení věty dokázáno.

# Kapitola 4

## Concurrent model

### 4.1 Popis modelu

Tento model popisuje vztah dvou (popřípadě více) funkcionálních náhodných veličin. V podstatě se jedná o rozšíření klasického regresního modelu. Namísto diskretních pozorovaných hodnot však stojí funkce se shodným definičním oborem. Označení *concurrent* je možné do češtiny přeložit jako *souběžný* nebo *paralelní*. Běžně se však tento výraz nepoužívá, proto je v dalším textu používána anglická varianta. Označení je odvozeno od toho, že pomocí vysvětlujících proměnných v čase  $t$  je modelovaná závislá proměnná rovněž v čase  $t$ . Formálně model zapíšeme takto

$$g_i(t) = \sum_{p=1}^P h_{ip}(t) \beta_p(t) + \psi_i(t), \quad (4.1)$$

kde  $\psi_i(t)$  jsou nezávislé, normálně rozdělené chyby s nulovou střední hodnotou a konečným, konstantním rozptylem. Realizace vysvětlované funkcionální náhodné veličiny značíme  $g_i$ . Jednotlivé realizace vysvětlujících proměnných značíme  $h_{ip}$ , přičemž  $p = 0, \dots, P$  je index vysvětlující proměnné a  $i = 1, \dots, I$  značí jednotlivá pozorování. Všechny funkce jsou definované na reálném intervalu  $T$ . Pokud platí, že  $h_{i1} = 1$ , jedná se o analogii regresního modelu s absolutním členem. Maticový zápis modelu je následující

$$\mathbf{g}(t) = \mathbf{H}(t) \boldsymbol{\beta}(t) + \boldsymbol{\psi}(t), \quad (4.2)$$

kde  $\mathbf{g}(t) = (g_1(t), \dots, g_I(t))'$ ,  $\boldsymbol{\beta}(t) = (\beta_1(t), \dots, \beta_P(t))'$ ,  $\boldsymbol{\psi}(t) = (\psi_1(t), \dots, \psi_I(t))'$  a  $\mathbf{H}(t) = (h_{ip}(t))_{i=1, p=1}^I$ .

Nyní se budeme věnovat souvislostem mezi modelem 4.1 a sledovanými průběhy OAR. V předchozí kapitole jsme popsali konstrukci odhadu spojité funkce na základě diskretních pozorování. Ve stručnosti (podrobný popis modelu je možné najít v

kapitole 3). Připomeňme, že výsledný odhad byl konstruovaný penalizovanou minimalizací nejmenších čtverců za předpokladu platnosti modelu

$$y(t_j) = f(t_j) + \epsilon(t_j), \quad (4.3)$$

kde  $y(t_j)$  jsou hodnoty OAR naměřené v časech  $t_j$ , kde  $j = 1, \dots, J$  a  $\epsilon(t_j)$  jsou nezávislé, náhodné chyby s nulovou střední hodnotou a konečným konstantním rozptylem. Nyní budeme navíc předpokládat, že chyby jsou normálně rozdělené. Veškeré proměnné týkající se OAR naměřené v dětském pokoji budeme dále značit pomocí horního indexu <sup>1</sup> a proměnné týkající se průběhu OAR na chodbě pomocí horního indexu <sup>2</sup>. Nejjednodušší model popisující vztah mezi průběhem OAR v dětském pokoji (dále  $f^1(t)$ ) a na chodbě (dále  $f^2(t)$ ) je následujícího tvaru

$$f_i^1(t) = f_i^2(t) \beta_0(t) + \psi_i(t), \quad (4.4)$$

kde  $\psi_i(t)$  jsou nezávislé, normálně rozdělené chyby s nulovou střední hodnotou a konečným, konstantním rozptylem. Index  $i$  označuje jednotlivé napozorované průběhy a  $t \in T = \langle t_1, t_J \rangle$ .

## 4.2 Výběr vhodného modelu

To jak je model dobrý, budeme posuzovat podle střední čtvercové chyby. Definujeme ji jako rozšíření jednorozměrné střední čtvercové chyby pro funkcionální pozorování. Nejprve však definujeme průměr a rozptyl funkcionálních pozorování.

**Definice 4.2.1** *Nechť  $f_1(t), \dots, f_K(t)$  je  $K$  realizací funkcionální náhodné veličiny  $f(t)$ , kde  $t \in \mathbf{T}$ . Potom její průměr a  $(f(t))$  definujeme jako*

$$a(f(t)) = \frac{1}{K} \sum_{k=1}^K f_k(t)$$

a výběrový rozptyl  $v(f(t))$  je definová jako

$$v(f(t)) = \frac{1}{K} \sum_{k=1}^K \left( f_k(t) - \frac{1}{K} \sum_{l=1}^K f_l(t) \right)^2.$$

Nyní již můžeme přistoupit k vlastní definici střední čtvercové chyby  $MSE_f$ :

$$\begin{aligned} MSE_f(\hat{\mathbf{g}}(t)) &= \frac{1}{I} \sum_{i=1}^I (g_i(t) - \hat{g}_i(t))^2 \\ E\{MSE_f(\hat{\mathbf{g}}(t))\} &= \frac{1}{I} \sum_{i=1}^I \{var(g_i(t)) + var(\hat{g}_i(t)) + bias^2(\hat{g}_i(t))\} \quad (4.5) \end{aligned}$$

Důkaz tohoto rozkladu je analogický s jednorozměrným případem, proto ho nebudeme uvádět. Dále definujeme střední čtvercovou chybu v případě, kdy jsou funkce  $\boldsymbol{\beta}(t)$  známé. Budeme ji značit  $MSE^*$ .

$$\begin{aligned} MSE_f^*(\hat{\mathbf{g}}(t)) &= \frac{1}{I} \sum_{i=1}^I \left( \mathbf{H}(t) \boldsymbol{\beta}(t) - \mathbf{H}(t) \hat{\boldsymbol{\beta}} \right)^2 = \\ &= \frac{1}{I} \sum_{i=1}^I \{var(\hat{g}_i(t)) + bias^2(\hat{g}_i(t))\} \end{aligned} \quad (4.6)$$

Z definic je zřejmé, že všechna kritéria jsou funkce s definičním oborem  $\mathbf{T}$ . Jejich porovnáním můžeme získat v různých bodech definičního oboru různé výsledky. Proto zavedeme ještě integrované střední čtvercové chyby  $IMSE_f$  a  $IMSE_f^*$ .

$$IMSE_f(\hat{\mathbf{g}}(t)) = \int_{\mathbf{T}} MSE(\hat{\mathbf{g}}(t)) dt \quad (4.7)$$

$$IMSE_f^*(\hat{\mathbf{g}}(t)) = \int_{\mathbf{T}} MSE_f^*(\hat{\mathbf{g}}(t)) dt \quad (4.8)$$

Na závěr této kapitoly zdefinujeme funkcionální verzi Akaikého kritéria  $AIC_f$ :

$$AIC_f(\hat{g}_i(t)) = \int_{\mathbf{T}} \left( C + \ln \left( \sum_{i=1}^I (\hat{g}_i(t) - g_i(t))^2 \right) + 2(k+1) \right) dt, \quad (4.9)$$

kde  $k$  je počet parametrů a  $C$  je konstanta, která závisí pouze na  $\mathbf{g}(t)$ . Pro vzájemné porovnání modelů jí tedy není nutné explicitně vyjádřit. Stejně jako v předchozím případě prosté rozšíření AIC integruji, abych získala jeho jednoznačnou porovnatelnost. Aby bylo možné toto kritérium použít, musí být splněna normalita a homoskedasticita náhodných chyb modelu.

### 4.3 Odhad funkcí $\beta_1, \dots, \beta_P$

Optimální vektor funkcí  $\boldsymbol{\beta}$  budeme hledat obdobně jako v klasickém regresním modelu pomocí minimalizace součtu čtverců reziduí. Situace je však komplikovanější, protože jednotlivé  $\beta_j$  nejsou konstanty, ale funkce definované na intervalu  $\langle t_1, t_J \rangle$ . Minimalizační kritérium se odvodí jako rozšíření  $SSE$ . Stejně jako v jednorozměrné regresi hledáme optimální  $\boldsymbol{\beta}$  tak, abych minimalizovala součet čtvercových chyb. Prostým rozšířením by výsledným kritériem byla funkce s definičním oborem  $\langle t_1, t_J \rangle$  a nebylo by možné jednoznačně určit, zda je minimální. Z tohoto důvodu ho integrujeme podle času  $t$ .

$$SSE_f(\boldsymbol{\beta}) = \int_{\mathbf{T}} (\mathbf{g}(t) - \mathbf{H}(t) \boldsymbol{\beta}(t))' (\mathbf{g}(t) - \mathbf{H}(t) \boldsymbol{\beta}(t)) dt \quad (4.10)$$

Při použití přímého rozšíření součtu čtverců reziduí klasického regresního modelu (budeme ho značit jako  $SSE_f$ ) může být výsledný odhad  $\hat{\boldsymbol{\beta}}(t)$  málo stabilní. V takovém případě je dobré použít rozšíření penalizovaného součtu čtverců (budeme ho značit jako  $PSSE_f$ ).

$$PSSE_f(\boldsymbol{\beta}) = \int_T (\mathbf{g}(t) - \mathbf{H}(t)\boldsymbol{\beta}(t))' (\mathbf{g}(t) - \mathbf{H}(t)\boldsymbol{\beta}(t)) dt + \sum_{p=1}^P \lambda_p \int_T [L_p \beta_p(t)]^2 dt, \quad (4.11)$$

kde  $L$  je označení diferenciálního operátoru a  $\boldsymbol{\lambda}$  je vektor penalizačních parametrů. Toto kritérium uvádí J. O. Ramsay v knize [6].

Funkce  $\beta_p(t)$  je možné vyjádřit jako lineární kombinace funkcí báze  $\theta_{kp}(t)$  prostoru hladkých diferencovatelných funkcí, nad kterým odhad konstruujeme.

$$\beta_p(t) = \sum_{k=1}^{K_p} b_{kp} \theta_{kp}(t) = \boldsymbol{\theta}_p(t)' \mathbf{b}_p, \quad (4.12)$$

kde  $K_p$  je počet funkcí báze  $\{\theta_{kp}\}$ . Dále zavedeme značení

$$\mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_P \end{pmatrix}.$$

Tento vektor má  $K_\beta = \sum_{p=1}^P K_p$  řádků. Potom  $\boldsymbol{\Theta}(t)$  je definována jako

$$\boldsymbol{\Theta}(t) = \begin{bmatrix} \boldsymbol{\theta}'_1(t) & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\theta}'_2(t) & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\theta}'_P(t) \end{bmatrix}.$$

Dále  $\boldsymbol{\beta}(t) = \boldsymbol{\Theta}(t)\mathbf{b}$  a výraz (4.2) je možné zapsat jako

$$\mathbf{g}(t) = \mathbf{H}(t)\boldsymbol{\Theta}(t)\mathbf{b} + \boldsymbol{\psi}(t).$$

Definujme blokovou diagonální matici  $\mathbf{R}$  s  $P$  bloky

$$\mathbf{R}_p = \lambda_p \int_T [L\boldsymbol{\theta}_p(t)]' [L\boldsymbol{\theta}_p(t)] dt.$$

Minimalizační kritérium  $PSSE_f$  (4.11) se dá přepsat jako

$$\begin{aligned} PSSE_f(\mathbf{b}) &= \int_T (\mathbf{g}(t) - \mathbf{H}(t) \Theta(t) \mathbf{b})' (\mathbf{g}(t) - \mathbf{H}(t) \Theta(t) \mathbf{b}) dt + \sum_{p=1}^P \mathbf{b}'_p \mathbf{R}_p \mathbf{b}_p = \\ &= \int_T (\mathbf{g}'(t) \mathbf{g}(t) - \mathbf{g}'(t) \mathbf{H}(t) \Theta(t) \mathbf{b} - \mathbf{b}' \Theta'(t) \mathbf{H}'(t) \mathbf{g}(t) + \\ &+ \mathbf{b}' \Theta'(t) \mathbf{H}'(t) \mathbf{H}(t) \Theta(t) \mathbf{b}) dt + \mathbf{b}' \mathbf{R} \mathbf{b}. \end{aligned}$$

Odhad parametru  $\mathbf{b}$  je řešením soustavy normálních rovnic.

Platí tedy  $\hat{\mathbf{b}} = \arg \min (PSSE_f(\mathbf{b}))$  a

$$\begin{aligned} \int_T \left( -\Theta'(t) \mathbf{H}'(t) \mathbf{g}(t) + \Theta'(t) \mathbf{H}'(t) \mathbf{H}(t) \Theta(t) \hat{\mathbf{b}} \right) dt + \mathbf{R} \hat{\mathbf{b}} &= 0, \\ \int_T \left( \Theta'(t) \mathbf{H}'(t) \mathbf{H}(t) \Theta(t) \hat{\mathbf{b}} \right) dt + \mathbf{R} \hat{\mathbf{b}} &= \int_T (\Theta'(t) \mathbf{H}'(t) \mathbf{g}(t)) dt. \end{aligned}$$

Pokud označíme

$$\mathbf{A} = \int_T (\Theta'(t) \mathbf{H}'(t) \mathbf{H}(t) \Theta(t)) dt + \mathbf{R}, \quad (4.13)$$

$$\mathbf{d} = \int_T (\Theta'(t) \mathbf{H}'(t) \mathbf{g}(t)) dt, \quad (4.14)$$

pak soustavu normálních rovnic je možné zapsat jako

$$\mathbf{A} \hat{\mathbf{b}} = \mathbf{d}. \quad (4.15)$$

V některých případech je možné řešení soustavy (4.15) vyjádřit explicitně, ale obecně je vhodné tuto soustavu řešit numerickými metodami integrace. Odhad jednotlivých realizací funkcionální náhodné veličiny  $\mathbf{g}(t)$  je možné tedy vyjádřit jako

$$\hat{\mathbf{g}}(t) = \mathbf{H}(t) \Theta(t) \hat{\mathbf{b}}. \quad (4.16)$$

# Kapitola 5

## Simulační studie

V této kapitole uvádím simulační studie. První studie se zaměřuje na nevhodnost Fourierovy báze pro modelování neperiodických dat. Druhá řeší chování zobecněné cross-validační metody v případě porušení předpokladu nezávislosti dat.

### 5.1 Simulační studie I - Porovnání Fourierovy a B-spline báze

Tento příklad ukazuje rozdílné chování Fourierovy báze a B-spline báze pro periodická a neperiodická data.

1. Uvažuji periodickou funkci

$$f_1(t) = \sin(t) + \cos(\pi t), \quad t \in (0, 2\pi).$$

2. Uvažuji neperiodickou funkci

$$f_2(s) = e^{s^3}, \quad s \in (0, 1).$$

Vygenerovala jsem 1000 replikací dat podle každého z následujících modelů:

1.  $y_1(t_j) = f_1(t_j) + \epsilon_1(t_j)$ , kde  $t_j = \frac{2\pi j}{100}$ ,  $j = 0, \dots, 100$  a  $\epsilon_1 \sim N\left(0; \frac{1}{25}\right)$ .
2.  $y_2(s_j) = f_2(s_j) + \epsilon_2(s_j)$ , kde  $s_j = \frac{j}{100}$ ,  $j = 0, \dots, 100$  a  $\epsilon_2 \sim N\left(0; \frac{1}{25}\right)$ .

Odhadla jsem funkce  $f_1$  a  $f_2$  pomocí Fourierovy báze a pomocí B-splinu. Výsledné odhady odpovídající jednotlivým simulacím značím  $\hat{f}_{1r}(t_j)$  (resp.  $\hat{f}_{2r}(s_j)$ ), kde  $r$  je index simulace. K porovnání výsledků jsem použila rozdělení  $MSE^*$ .

Typ báze		$f_1$ (periodická)	$f_2$ (neperiodická)
Fourierova báze	$\frac{1}{J} \sum_{j=1}^J \text{var}(\hat{f}(t_j))$	0,00734	0,00758
	$\frac{1}{J} \sum_{j=1}^J \text{bias}^2 \hat{f}(t_j)$	0,00230	0,02512
	$E(MSE(\hat{\mathbf{f}}))$	0,00964	0,03269
B-spline	$\frac{1}{J} \sum_{j=1}^J \text{var}(\hat{f}(t_j))$	0,00825	0,00273
	$\frac{1}{J} \sum_{j=1}^J \text{bias}^2 \hat{f}(t_j)$	0,00102	0,00054
	$E(MSE(\hat{\mathbf{f}}))$	0,00926	0,00326

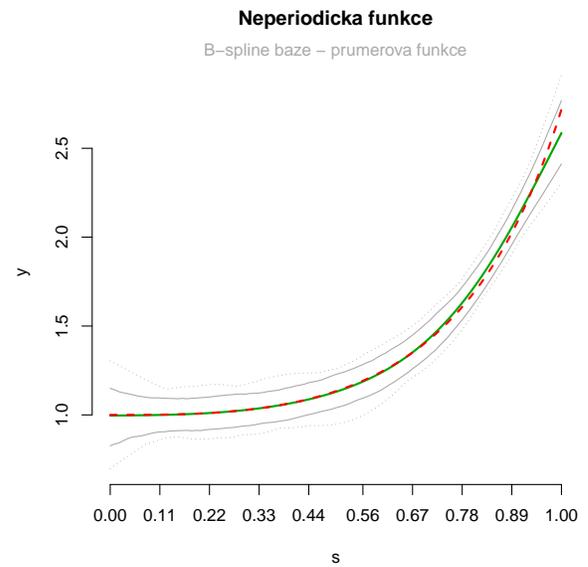
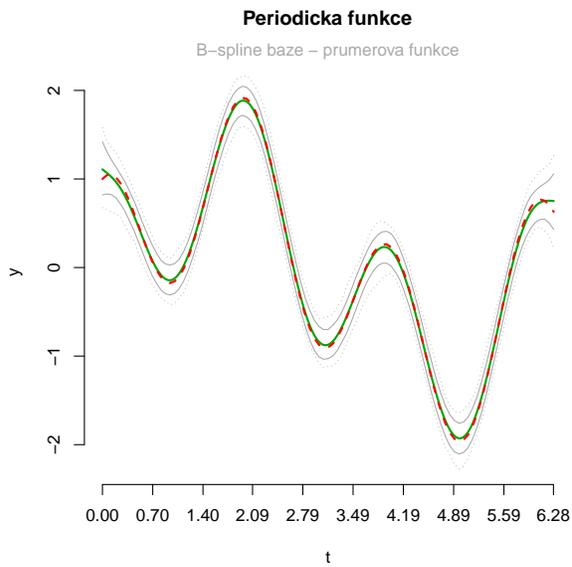
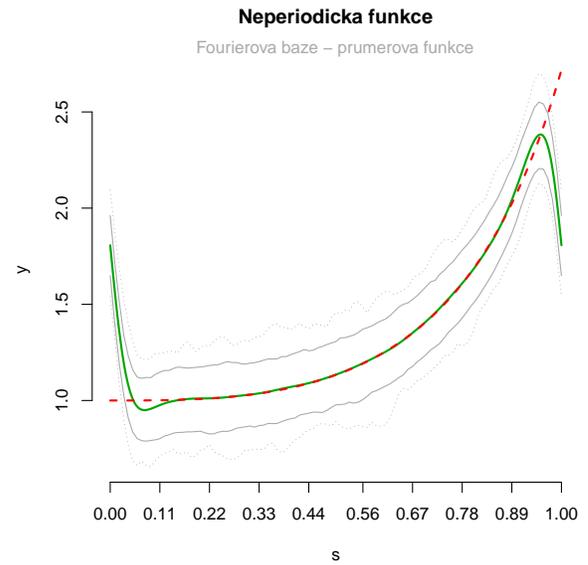
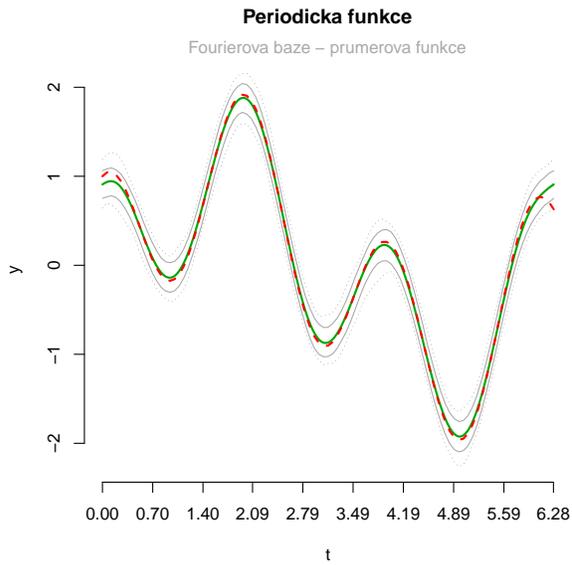
Tabulka 5.1: Odhad střední hodnoty a rozptylu RMSE

Fourierova báze se skládá z 101 funkcí. Pro určení penalizačního parametru  $\lambda$  jsem použila metodu GCV a jako penalizační člen jsem zvolila  $PEN$  (viz (3.10)).

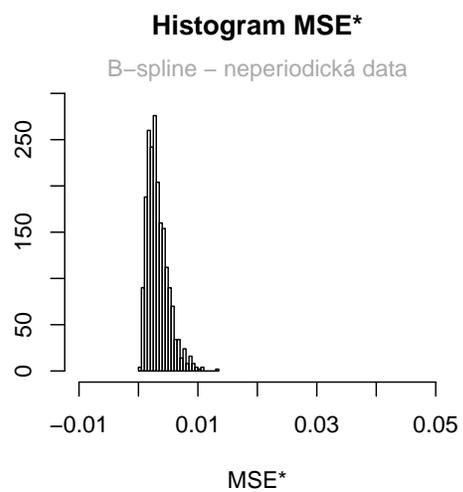
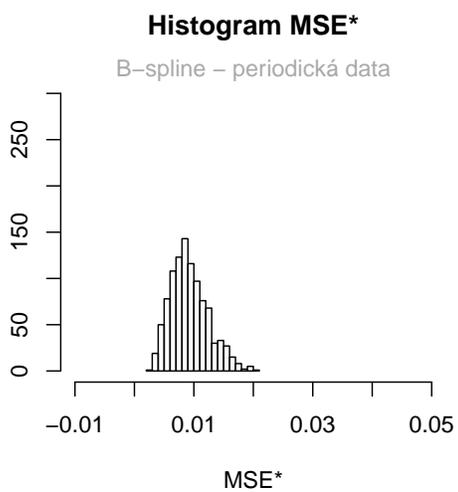
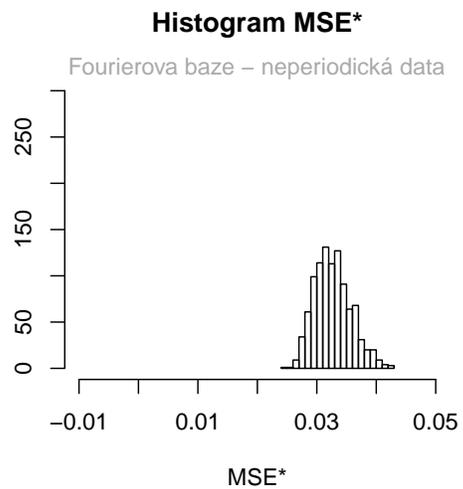
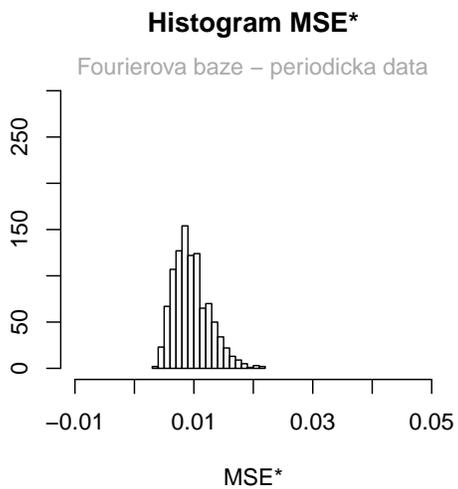
Kubický B-spline má uzly v bodech  $t_j$  (resp.  $s_j$ ), kde  $j = 0, \dots, 100$ . Pro penalizaci jsem, stejně jako u Fourierovy báze, použila  $PEN$  a hodnotu parametru  $\lambda$  jsem stanovila na základě GCV metody.

Na obrázku (5.2) jsou graficky znázorněny výsledné odhady. Zeleně je vyznačen odhad střední hodnoty  $\hat{f}_1(t_j)$  (resp.  $\hat{f}_2(s_j)$ ) a červeně přerušovaně funkce  $f_1$  (resp.  $f_2$ ). Dále je v grafu vyznačena obálka simulovaných odhadů, která je definovaná jako nejmenší konvexní množina obsahující hodnoty  $\hat{f}_{1r}(t_j)$  (resp.  $\hat{f}_{2r}(s_j)$ ), a množiny obsahující 95% dat, které jsou zkonstruované následovně. Dolní mez pro každé  $t_j$  (resp.  $s_j$ ) je 2,5%tní výběrový kvantil z  $\{f_{1r}(t_j)\}_{r=1}^R$  (resp.  $\{f_{2r}(s_j)\}_{r=1}^R$ ) a horní mez je 97,5%tní výběrový kvantil. Dále uvádím histogramy charakterizující rozdělení  $MSE^*$ .

Pro periodická data jsou výsledky sice srovnatelné, ale B-spline báze dává odhad s nižší průměrnou  $MSE^*$  a menším vychýlením na úkor vyššího rozptylu odhadu (viz tabulka (5.1)). U neperiodických dat se potvrdilo to, že funkční hodnoty odpovídající krajním bodům definičního intervalu mohou být při použití Fourierovy báze velmi vychýlené. Průměrná střední čtvercová chyba je v tomto případě daleko vyšší než při použití B-splinu. Rozptyl odhadu je srovnatelný s rozptylem odhadu periodické funkce, ale vychýlení je daleko vyšší. Odhad zkonstruovaný pomocí B-spline báze vykazuje pro neperiodická data výrazně lepší vlastnosti. Podrobné výsledky simulace jsou shrnuty v tabulce (5.1), která obsahuje rozklad střední čtvercové chyby pro periodická i neperiodická data a oba typy báze.



Obrázek 5.1: Odhad  $E(\hat{f}_1)$  a  $E(\hat{f}_2)$  (zeleně),  $f_1(t_j)$  a  $f_2(s_j)$  (červeně čárkovaně), množina obsahující 95% dat (šedě) a obálka simulací (šedě čárkovaně).



Obrázek 5.2: Histogramy  $MSE^*$

## 5.2 Simulační studie II

Tato studie má za úkol zjistit, zda je lepší určit parametr  $\lambda$  v penalizovaném součtu čtverců pomocí metody *GCV* nebo na základě vlastního úsudku v případě, kdy je porušen předpoklad nezávislosti vstupních dat a předpoklad normality náhodných chyb. Vycházela jsem z měření provedených v dětském pokoji. Na základě těchto dat jsem zkonstruovala odhad funkce  $f^1$  (viz kapitola (3)), který jsem použila jako střední hodnotu simulovaných dat.

$$y_s(t_j) = \widehat{f}(t_j) + \widehat{\epsilon}(t_j). \quad (5.1)$$

Dále jsem vygenerovala  $R$  replikací (každá obsahuje  $J$  prvků) náhodných chyb, které jsem uspořádala do matice  $\mathbf{E}$ .

$$\mathbf{E} = \begin{pmatrix} \widehat{\epsilon}_{11} & \dots & \widehat{\epsilon}_{1R} \\ \vdots & & \vdots \\ \widehat{\epsilon}_{J1} & \dots & \widehat{\epsilon}_{JR} \end{pmatrix}.$$

Simulovaná data jsem zkonstruovala jako součet generovaných náhodných chyb a hodnot funkce  $\widehat{f}^1$  v časech měření

$$\mathbf{Y}_s = \begin{pmatrix} y_{s1}(t_1) & \dots & y_{sR}(t_1) \\ \vdots & & \vdots \\ y_{s1}(t_J) & \dots & y_{sR}(t_J) \end{pmatrix} = \begin{pmatrix} \widehat{f}^1(t_1) & \dots & \widehat{f}^1(t_1) \\ \vdots & & \vdots \\ \widehat{f}^1(t_J) & \dots & \widehat{f}^1(t_J) \end{pmatrix} + \mathbf{E}.$$

Na základě pseudodat zkonstruuji odhady funkce  $\widehat{f}^1$  (budu je značit indexem replikací).

$$\mathbf{Y}_s = \left( \widehat{f}_1^1(\mathbf{t}), \dots, \widehat{f}_R^1(\mathbf{t}) \right) + \widehat{\mathbf{E}},$$

kde  $\mathbf{t} = (t_1, \dots, t_J)$ ,  $\widehat{f}_r^1$  jsou odhady funkce  $\widehat{f}^1$  odpovídající jednotlivým simulacím a  $\widehat{\mathbf{E}}$  je odhad matice  $\mathbf{E}$ .

Ke konstrukci odhadu  $\widehat{\mathbf{f}}^1$  jsem použila Fourierovu bázi, která se skládá z 301 funkcí. Optimální lineární kombinaci funkcí báze jsem spočítala na základě minimalizace penalizovaného součtu čtverců, a to pro dvě hodnoty parametru  $\lambda$ :

1.  $\lambda_1 = 0$ : výsledný odhad (v dalším textu ho značím  $\widehat{\mathbf{f}}^1_{\lambda_1} = \left( \widehat{f}^1_{\lambda_1}(t_1), \dots, \widehat{f}^1_{\lambda_1}(t_J) \right)'$ ) je poměrně variabilní (dá se očekávat, že odhad nebude příliš stabilní). Z analýzy reziduí modelu za předpokladu vyloučení první a poslední hodnoty (jedná se o odlehlá pozorování) je patrné, že nelze zamítnout hypotézu normality reziduí (p-hodnota Shapirova-Wilkova testu normality je 0.1). Rezidua modelu jsou korelovaná;

2.  $\lambda_2 = 2$  : výsledný odhad (v dalším textu ho značím  $\widehat{\mathbf{f}}^1_{\lambda_2} = \left( \widehat{f}^1_{\lambda_2}(t_1), \dots, \widehat{f}^1_{\lambda_2}(t_J) \right)'$ ) se zdá stabilnější než v prvním případě, rezidua jsou méně korelovaná, ale jejich rozdělení má ve srovnání s normálním rozdělením těžké chvosty.

Na obrázku (5.3) jsou zeleně zobrazeny výsledné odhady  $\widehat{\mathbf{f}}^1_{\lambda_1}$  a  $\widehat{\mathbf{f}}^1_{\lambda_2}$  a na obrázku (5.4) jsou histogramy reziduí modelu 5.1, jejich Q-Q grafy a empirické autokovarianční funkce. Obrázky potvrzují vlastnosti reziduí popsané v předchozím odstavci (nižší korelace reziduí pro  $\lambda_2$  a těžké chvosty jejich rozdělení).

Počet simulací  $R$  je 600 (ověřovala jsem dostatečnost počtu simulací pomocí stability rozptylu odhadů). Náhodné chyby jsem generovala třemi způsoby:

1.  $\widehat{\epsilon}_{jr} \sim N(0, 17^2)$ , pro všechna  $j = 1, \dots, J$  a  $r = 1, \dots, R$ .
2. Vektory  $\mathbf{E}_{(:,r)}$  (tímto značením mám na mysli sloupce matice  $\mathbf{E}$ ) konstruuji pro všechna  $r = 1, \dots, R$  jako permutace vektoru reziduí  $\widehat{\epsilon}_{\lambda_2}$  (tento vektor jsem vybrala, protože jeho složky jsou méně korelované).
3. Vektory  $\mathbf{E}_{(:,r)}$  konstruuji pomocí tzv. *blokového bootstrapu* (viz [4]). Výchozí množina pro bootstrap je  $\widehat{\epsilon}_{\lambda_1}$ . Používám klouzavé bloky a délka bloku je 12 pozorování (což odpovídá 6 hodinám). Naprogramovaná procedura je součástí přiloženého CD. Tuto metodu jsem zvolila proto, abych byla schopná modelovat korelovaná data a přitom se co nejvíce přiblížila korelační struktuře originálních dat.

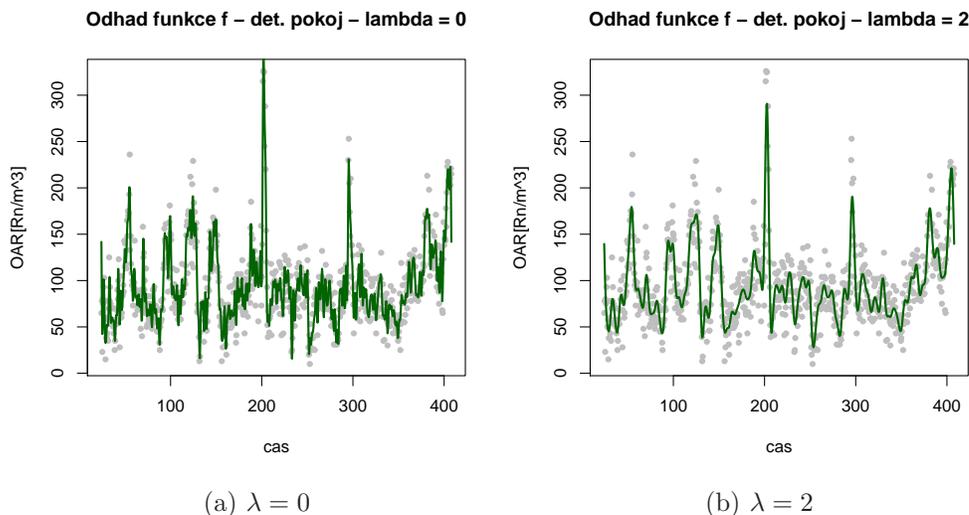
V prvním případě očekávám, že metoda GCV bude pracovat pro simulovaná data dobře (jsou splněny předpoklady modelu). V druhém případě budu porovnávat, jaký vliv má použití rozdělení s těžkými chvosty (ve srovnání s normálním rozdělením). V obou těchto případech se jedná o nezávislá pozorování. V posledním případě simulovaná data částečně zachovávají původní korelační strukturu modelu. Jelikož se mi jedná hlavně o testování GCV pro korelovaná data, mohu tento zjednodušený přístup použít. V tabulce (5.2) je přehled všech způsobů generování pseudodat  $\mathbf{Y}_s$ .

Dále jsem pro jednotlivé modely spočítala odhady funkcí  $\widehat{\mathbf{f}}^1_{\lambda_1}$ ,  $\widehat{\mathbf{f}}^1_{\lambda_2}$  a porovnála jejich vlastnosti. Funkce  $\widehat{\mathbf{f}}^1_{\lambda_1}$  a  $\widehat{\mathbf{f}}^1_{\lambda_2}$  jsou spolu s Q-Q grafy a histogramy reziduí  $\widehat{\epsilon}_{\lambda_1}$  a  $\widehat{\epsilon}_{\lambda_2}$  graficky znázorněny na obrázku (5.3).

Z grafů je patrné, že vyšší hodnota penalizačního parametru dává vyšší *hladkost* výsledných odhadů. Histogramy ukazují, že rezidua modelů jsou symetrická, střední hodnota je nulová a rozptyl je vyšší pro vyšší hodnotu penalizačního parametru. Z Q-Q grafu je vidět, že rozdělení reziduí se nejvíce podobá normálnímu rozdělení pro nepenalizovanou variantu (s vyloučením odlehlých pozorování). Testovala jsem

$r = 1, \dots, 600$	Střední hodnota $\mathbf{y}_{sr}(t_j)$	
Způsob generování matice $\mathbf{E}$	$\widehat{f}_{\lambda_1}^1(t_j)$	$\widehat{f}_{\lambda_2}^1(t_j)$
$N(0, 17^2)$	$\mathbf{Y}_{s(\cdot,r)}^{n1} = \widehat{f}_{\lambda_1}^1 + \mathbf{E}^{n1}_{(\cdot,r)}$	$\mathbf{Y}_{s(\cdot,r)}^{n2} = \widehat{f}_{\lambda_2}^1 + \mathbf{E}^{n2}_{(\cdot,r)}$
Permutace $\widehat{\epsilon}_{\lambda_2}$	$\mathbf{Y}_{s(\cdot,r)}^{p1} = \widehat{f}_{\lambda_1}^1 + \mathbf{E}^{p1}_{(\cdot,r)}$	$\mathbf{Y}_{s(\cdot,r)}^{p2} = \widehat{f}_{\lambda_2}^1 + \mathbf{E}^{p2}_{(\cdot,r)}$
Blokový bootstrap $\widehat{\epsilon}_{\lambda_1}$	$\mathbf{Y}_{s(\cdot,r)}^{b1} = \widehat{f}_{\lambda_1}^1 + \mathbf{E}^{b1}_{(\cdot,r)}$	$\mathbf{Y}_{s(\cdot,r)}^{b2} = \widehat{f}_{\lambda_2}^1 + \mathbf{E}^{b2}_{(\cdot,r)}$

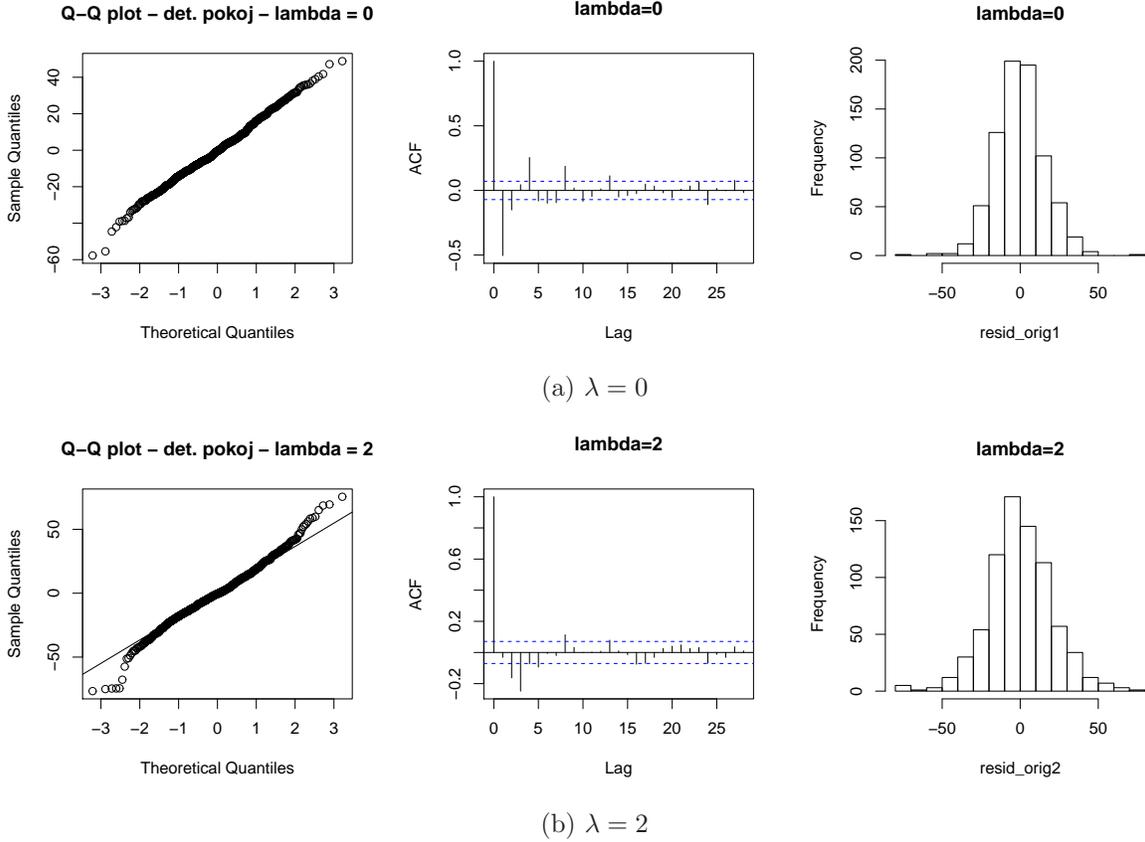
Tabulka 5.2: Přehled značení simulací pro různé hodnoty parametru  $\lambda$  a různé způsoby generování náhodných chyb



Obrázek 5.3: Zobrazení funkce  $\widehat{f}(t)$ ,  $t \in \mathbf{T}$ , pro různé hodnoty parametru  $\lambda$

různé kombinace počtu funkcí báze a výše penalizačního parametru a obecně se dá říci, že modely konstruované za pomoci penalizace mají rezidua s těžkými chvosty (ve srovnání s normálním rozdělením). Vyzkoušené transformace dat tento efekt zvýrazní, popřípadě způsobí zešikmení rozdělení reziduí modelu.

Z výsledků je patrné, že metoda GCV pro rozdělení s těžkými chvosty dává vychýlené odhady, ale celková střední čtvercová chyba je srovnatelná s hodnotou pro model s normálním rozdělením (pro  $\lambda = 2$  dosahuje dokonce nižší hodnoty). Z porovnání modelů s korelovanými a nekorelovanými náhodnými chybami je vidět, že porušení předpokladu nezávislosti dává vychýlené odhady, ale rozdíl oproti variantě se splněnými předpoklady není tak výrazný (především v případě, kdy data odpovídají hladšímu modelu), abychom nemohli metodu aplikovat i na korelovaná data.



Obrázek 5.4: Základní charakteristiky reziduí modelu (5.1)

Model	$MSE_f$	$var$	$bias^2$
$\mathbf{Y}_{s(\cdot,r)}^{n1} = \widehat{\mathbf{f}}_{\lambda_1}^1 + \mathbf{E}_{(\cdot,r)}^{n1}$	114,39	102,96	11,60
$\mathbf{Y}_{s(\cdot,r)}^{n2} = \widehat{\mathbf{f}}_{\lambda_2}^1 + \mathbf{E}_{(\cdot,r)}^{n2}$	42,14	42,13	0,08
$\mathbf{Y}_{s(\cdot,r)}^{p1} = \widehat{\mathbf{f}}_{\lambda_1}^1 + \mathbf{E}_{(\cdot,r)}^{p1}$	118,85	92,98	26,02
$\mathbf{Y}_{s(\cdot,r)}^{p2} = \widehat{\mathbf{f}}_{\lambda_2}^1 + \mathbf{E}_{(\cdot,r)}^{p2}$	40,71	39,89	0,89
$\mathbf{Y}_{s(\cdot,r)}^{b1} = \widehat{\mathbf{f}}_{\lambda_1}^1 + \mathbf{E}_{(\cdot,r)}^{b1}$	155,90	71,53	84,48
$\mathbf{Y}_{s(\cdot,r)}^{b2} = \widehat{\mathbf{f}}_{\lambda_2}^1 + \mathbf{E}_{(\cdot,r)}^{b2}$	40,51	37,47	3,10

Tabulka 5.3: Srovnání střední čtvercové chyby (a rozklad na jednotlivé komponenty) odhadů pro různé hodnoty penalizačního parametru ( $\lambda_1$  a  $\lambda_2$ ) a různé způsoby generování náhodných chyb.

# Kapitola 6

## Model závislosti OAR

Předpokládejme, že průběhy OAR během jednotlivých dnů jsou nezávislá funkcionální pozorování. Uvědomuji si, že to není zcela korektní předpoklad, protože data jsou vzájemně korelovaná, a proto se nedá vyloučit ani korelace mezi jednotlivými dny. Z tohoto důvodu je výsledný model spíše aproximativní. Předpokládám, že OAR se mění v průběhu dne, ale toto chování je pro jednotlivé dny podobné. Data rozdělíme podle dnů měření a pozorování z jednotlivých dnů budu považovat za nezávislé realizace funkcionální náhodné veličiny popisující průběh OAR v jednom dni. V rozporu s tímto předpokladem je skutečnost, že pozorování s blízkým časem měření jsou korelovaná. Tento fakt v tomto případě zanedbáme (korelace je prokazatelná pouze pro několik blízkých pozorování). Předpoklad nezávislosti se projeví ve vlastnostech odhadů denního průběhu OAR, které nebudou splňovat podmínku návaznosti mezi jednotlivými dny. Pro jednoduchost vyřadíme první a poslední den, protože měření neprobíhalo v průběhu celého dne.

### 6.1 Funkcionální pozorování

#### 6.1.1 Model pro odhad funkcionálního pozorování

Tato část popisuje několik modelů vhodných k popisu denního průběhu OAR. Budu předpokládat, že denní průběh OAR je možné popsat funkcemi  $f_i(t)$  ( $i$  je index jednotlivých dnů) a že tyto funkce jsou vzájemně nezávislé. Díky tomuto předpokladu nebudu požadovat, aby na sebe odhady  $\hat{f}_i$  navazovaly.

Obecný model je možné zapsat následovně

$$y_i(t_{ij}) = f_i(t_{ij}) + \epsilon(t_{ij}), \quad (6.1)$$

kde  $\epsilon_i$  jsou nezávislé, normálně rozdělené chyby se střední hodnotou 0 a konečným, konstantním rozptylem. Dále  $t_{ij} \in T = (0, 24)$ ,  $j = 1, \dots, J$  a  $i = 1, \dots, 16$ .

$t_{1j}$	0:04	0:34	...	16:04	16:34	17:04	...	23:34
$t_{11j}$	0:04	0:34	...	16:04	16:55	17:25	...	23:55
$t_{12j}$	0:25	0:55	...	16:25	16:55	17:25	...	23:55

Tabulka 6.1: Časy měření

Předpokládáme, že časy měření v průběhu jednotlivých dnů mohou být různé (nemusí být ani ekvidistantní).

Odhady funkcí  $f_i$  můžeme vyjádřit jako lineární kombinaci funkcí báze podprostoru prostoru spojitých, diferencovatelných funkcí:

$$\widehat{f}_i(t_{ij}) = \sum_{l=1}^L \phi_l(t_{ij}) c_{il}, \quad (6.2)$$

kde lineární koeficienty  $c_{il}$  spočítáme pomocí minimalizace penalizovaného součtu čtverců (3.9). Tyto odhady není možné pomocí standardních funkcí knihovny *fda* spočítat, protože nelze zadat různé argumenty  $t_{ij}$  pro různá  $i$ . Snažila jsem se najít alternativní metodu. Použiji následující rozklad do podmodelů

$$\begin{aligned} y_i(t_{1j}) &= f_i(t_{1j}) + \epsilon_i(t_{1j}), \quad i = 1, \dots, 10, \\ y_i(t_{11j}) &= f_i(t_{11j}) + \epsilon_i(t_{11j}), \quad i = 11, \\ y_i(t_{12j}) &= f_i(t_{12j}) + \epsilon_i(t_{12j}), \quad i = 12, \dots, 16, \end{aligned} \quad (6.3)$$

kde  $t_{1j} \in T$ ,  $t_{11j} \in T$ ,  $t_{12j} \in T$ ,  $j = 1, \dots, 48$ . Hodnoty  $t_{1j}$ ,  $t_{11j}$  a  $t_{12j}$  jsou shrnuty v tabulce (6.1) (pro modelování jsou tyto hodnoty převedeny na hodiny). Odhady funkcí  $f_i$  můžeme zapsat jako

$$\widehat{f}_i(t_{1j}) = \sum_{l=1}^L \phi_l(t_{1j}) c'_{il}, \quad i = 1, \dots, 10, \quad (6.4)$$

$$\widehat{f}_i(t_{11j}) = \sum_{l=1}^L \phi_l(t_{11j}) c'_{il}, \quad i = 11, \quad (6.5)$$

$$\widehat{f}_i(t_{12j}) = \sum_{l=1}^L \phi_l(t_{12j}) c'_{il}, \quad i = 12, \dots, 16. \quad (6.6)$$

Při použití standardních metod knihovny *fda* však budu mít pro každý podmodel zvláštní proměnnou a nebude možné tato funkcionální pozorování použít jako realizace jedné náhodné veličiny při konstrukci concurrent modelu.

Tento problém řeším následujícím způsobem:

- Uvažujme množinu  $(t_1, \dots, t_K)$ , která je tvořena uspořádaným sjednocením  $t_{ij}$  pro všechna  $i$  a  $j$ . Zároveň vektory  $y_i(t_k)$  rozšířím o nedefinované hodnoty (dále značeno jako NA) pro všechna  $t_k$ , pro která nejsou definovány.
- Standardním postupem knihovny *fda* nelze spočítat koeficienty  $c_{il}$  pro replikace obsahující hodnoty NA (v tomto případě se to týká všech replikací), proto koeficienty  $c_{il}$  nahradíme koeficienty  $c'_{il}$ . Je nutné ověřit vlastnosti reziduí takto zkonstruovaného modelu. Aby bylo možné tuto aproximaci použít, musejí být rezidua stejně rozdělená.

Odhad funkce  $f_i$  bude mít tvar

$$\hat{f}_i(t_k) = \sum_{l=1}^L \phi_l(t_k) c'_{il}. \quad (6.7)$$

Poslední model, se kterým budu pracovat, předpokládá shodné časy měření pro všechny replikace. Je pouze aproximativní. Nahradíme  $t_{ij}$ ,  $j = 1, \dots, J$  pro  $i > 1$  hodnotami  $t_{1j}$ . Jedná se o značné zjednodušení skutečných dat a v dalších odstavcích budu porovnávat, jak velký vliv toto zjednodušení má na celkové výsledky. Model má následující tvar

$$y_i(t_{1j}) = f_i(t_{1j}) + \epsilon_i(t_{1j}), \quad (6.8)$$

kde  $\epsilon_1$  jsou nezávislé, normálně rozdělené chyby s nulovou střední hodnotou a konečným rozptylem. Dále  $t_{1j} \in T$ ,  $j = 1, \dots, 48$ ,  $i = 1, \dots, 16$ . Díky tomu, že posunutí není příliš velké, lze očekávat, že výsledky budou pouze posunuté.

### 6.1.2 Simulační studie IV.

V této simulační studii se budu zabývat tím, nakolik se mohou změnit výsledné odhady, pokud se v datech vyskytují nedefinované hodnoty. Budu testovat změnu odhadu v závislosti na počtu nedefinovaných hodnot a na jejich poloze vůči ostatním pozorováním

Jako výchozí funkci pro testování budu používat odhad průběhu OAR během prvního dne měření. Pseudodata pro testování vygeneruji podle následujícího předpisu

$$y_r^*(t_k) = f(t_k) + \psi_r(t_k) = \sum_{l=1}^L \phi_l(t_k) c'_{1l} + \psi_r(t_k), \quad (6.9)$$

kde pro přehlednost písmenem  $f$  značím odhad  $\hat{f}_1$  (odhad průběhu OAR během prvního dne), přičemž koeficienty  $c'_{1l}$  spočítám pomocí minimalizace penalizovaného

součtu čtverců (hodnota penalizačního parametru je stanovena metodou GCV). Náhodné chyby generuji jako bootstrapový výběr z množiny odhadů  $(\widehat{\epsilon}_1(t_{11}), \dots, \widehat{\epsilon}_1(t_{1J}))$ . Pomocí indexu  $r$  čísluji jednotlivé simulace. Index nabývá hodnot  $1, \dots, R$ , kde  $R = 800$ . Definiční obor  $(t_1, \dots, t_K) \subset T = (0, 24)$ . Výpočty provedu pro dva typy báze  $\{\phi_l\}_{l=1}^L$ : Fourierova báze o 47 funkcích a b-spline s uzly v bodech  $t_{1j}$ .

Nyní budu testovat citlivost odhadu

$$\widehat{f}_r(t_k) = \sum_{l=1}^L \phi_l(t_k) d_{rl}, \quad (6.10)$$

zkonstruovaného na základě pseudodat  $\mathbf{y}_r^*$ , na chybějící pozorování. Označení  $d_{rl}$  jsem zavedla pro odhady koeficientů  $c'_{1l}$ . Pro hodnocení jednotlivých efektů budu používat opět střední čtvercovou chybu  $MSE^*$ .

Nejprve budu volit  $t_k$  ekvidistantně. Testovaným faktorem v tomto případě bude počet nedefinovaných hodnot a jejich poloha. Další možností je neekvidistantní rozložení  $t_k$ . V tomto případě mě zajímá závislost chyby na poloze chybějící hodnoty v rámci intervalu definovaného sousedními hodnotami. Budu testovat následující situace:

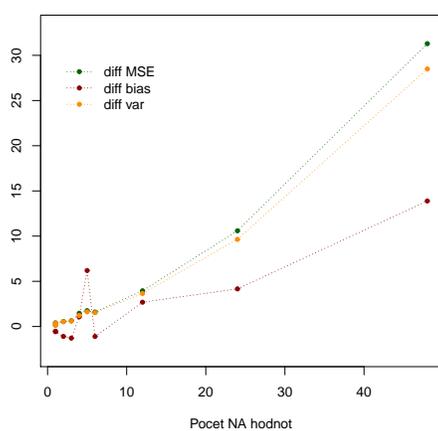
1. Definiční obor  $\{t_k\}_{k=1}^{96}$  definuji jako ekvidistantní dělení intervalu  $(t_{11}, t_{1J})$ . Počet nedefinovaných hodnot je postupně 1, 2, 3, 4, 5, 6, 12, 24 a 48 a jsou rovnoměrně rozmístěné v definičním oboru. Rovnoměrným rozmístěním mám na mysli, že se nedefinovaná hodnota nachází na pozicích s pořadím *celá část*  $\left(\frac{\text{počet nedefinovaných hodnot}}{96}\right)$ . V případě, že je nedefinovaná právě 1 hodnota, uvažuji dvě možnosti umístění. První je standardní dle výše uvedeného vzorce (pořadí 96, dále značím jako 1a) a druhá možnost je pozice 48 (dále značím jako 1b).
2. V tomto případě při konstrukci hodnot  $t_k$  vycházím  $\{t_{1j}\}_{j=1}^J$ . Mezi  $t_{1p}$  a  $t_{1(p+1)}$  ( $p = 24$ ) dodefinuji nový bod definičního oboru tak, že vzdálenost mezi novým prvkem a  $t_{1p}$  bude postupně  $\frac{1}{10}, \frac{2}{10}, \dots, \frac{9}{10}$  délky tohoto intervalu. Právě tyto vložené časy budou odpovídat nedefinovaným hodnotám.

Výsledky analýzy předchozích dvou situací obsahují tabulky (6.2) a (6.3). V tabulce (6.2) srovnávám rozdíl průměrné střední čtvercové chyby  $MSE^*$  odhadu, který byl zkonstruován na základě všech dat (značení  $\widehat{\mathbf{f}}_r^c$ , kde  $r$  je index simulace), a odhadu s chybějícími pozorováními (značení  $\widehat{\mathbf{f}}_r^t$ , kde  $r$  je index simulace). Dále sleduji rozdíl v rozptylu a vychýlení těchto odhadů.

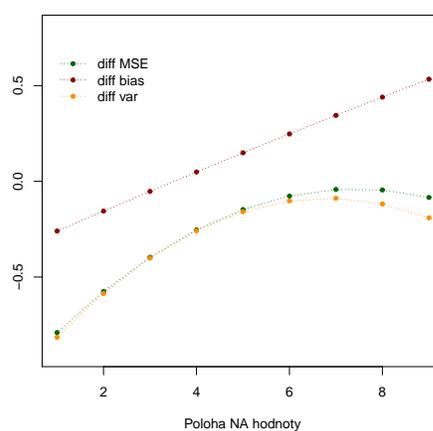
Na obrázku (6.1a) je zobrazena závislost rozdílu  $MSE^*$ , rozptylu a vychýlení na počtu vynechaných hodnot.

Fourier	1a	1b	2	3	4	5	6	12	24	48
$EMSE^* \widehat{f}^c$	38,42									
$bias \widehat{f}^c$	-0,02									
$var \widehat{f}^c$	33,95									
$EMSE^* \widehat{f}^t$	38,57	38,82	38,97	39,05	39,89	40,17	40,02	42,37	49,01	69,71
$bias \widehat{f}^t$	-0,02	-0,02	-0,03	-0,03	-0,01	0,04	-0,03	0,01	0,02	0,12
$var \widehat{f}^t$	34,10	34,33	34,48	34,56	35,18	35,58	35,48	37,61	43,59	62,44

Tabulka 6.2: Vliv vynechaných pozorování



(a) Závislost na počtu NA



(b) Závislost na poloze NA

Obrázek 6.1: Závislost rozdílu  $MSE^*$ , rozptylu a vychýlení na počtu vynechaných hodnot (vychýlení je zobrazeno jako 100 násobek původních hodnot) a na poloze vynechaného pozorování (vychýlení je zobrazeno jako 10 násobek původních hodnot).

Fourier	1 : 9	2 : 8	3 : 7	4 : 6	5 : 5	6 : 4	7 : 3	8 : 2	9 : 1
$EMSE^* \widehat{\mathbf{f}}^c$	76,76								
$bias \widehat{\mathbf{f}}^c$	-0,27								
$var \widehat{\mathbf{f}}^c$	72,37								
$EMSE^* \widehat{\mathbf{f}}^t$	76,43	76,39	76,37	76,36	76,37	76,39	76,42	76,47	76,53
$bias \widehat{\mathbf{f}}^t$	-0,31	-0,30	-0,28	-0,27	-0,26	-0,25	-0,24	-0,23	-0,21
$var \widehat{\mathbf{f}}^t$	72,28	72,26	72,25	72,23	72,22	72,20	72,20	72,18	72,17

Tabulka 6.3: Vliv polohy vynechaného pozorování

V případě použití Fourierovy báze byla ve všech výpočtech použita maximální báze (47 funkcí). Pro všechny výpočty byl použit penalizační parametr  $\lambda = 0,32$  (optimální hodnota stanovená metodou GCV pro základní model bez vynechaných pozorování). Výsledky ukazují, že střední čtvercová chyba odhadu roste s množstvím nedefinovaných pozorování rychleji než lineárně. Rozdíl vychýlení odhadů je ve srovnání s rozdílem středních čtvercových chyb velmi malý. Simulace jsem provedla i pro b-spline bázi, ale rozdíl mezi výsledky byl zanedbatelný, proto je zde neuvádím.

Druhým bodem zkoumání je vliv polohy vyloučeného měření vůči sousedním hodnotám. Tato analýza byla provedena pouze pro Fourierovu bázi (se shodnými vlastnostmi jako v předchozím případě). Na obrázku (6.1b) je zobrazena závislost změny  $MSE^*$  (resp. rozptylu a vychýlení) v důsledku vynechání jednoho pozorování na poloze tohoto pozorování (odpovídající hodnoty je možné nalézt v tabulce (6.3), přičemž sloupce odpovídají dělicímu poměru intervalu, který je dán umístěním vynechaného pozorování). Konkávní charakter říká, že s klesající vzdáleností NA hodnoty od okraje sledovaného intervalu klesá i celková střední čtvercová chyba. Ale nedá se říci, že by to nezáviselo na výběru sledovaného intervalu (o tom svědčí rozdílné okrajové hodnoty a nesymetrický tvar funkce).

### 6.1.3 Odhad funkcionálního pozorování

V této kapitole popíšu postup výpočtu odhadů denního průběhu OAR. Zvolila jsem dva různé odhady. První je odhad (6.6) (dále ho budu značit jako odhad 1) a druhý vychází z modelu (6.8), který zanedbává časový posun pozdějších měření (dále odhad 2). Parametry modelu jsem určila na základě minimalizace penalizovaného součtu čtverců. Hodnotu penalizačního parametru jsem stanovila za pomoci metody GCV, přičemž jsem se rozhodovala na základě součtu hodnot GCV kritéria jednotlivých křivek.

Při použití metody GCV je výsledná hodnota parametru  $\lambda$  silně ovlivněna dny

osm a dvanáct, které prokazují výrazně odlišné vlastnosti od zbytku souboru. Z toho důvodu jsem tyto dny do výpočtu nezahrnula.

Pro  $d = 11$  (den, ve kterém nastal posun v časech měření) jsem hodnotu parametru  $\lambda$ , kterou dává metoda GCV navýšila. A to proto, že pro odhad penalizačního parametru byla k dispozici pouze jedna replikace. V takovém případě metoda nedává srovnatelné výsledky se situací, kdy je k dispozici více replikací. Je to patrné již při vizuální kontrole, protože výsledný odhad dne 11 se zdá více variabilní než odhady pro ostatní dny. Navíc jsou rezidua modelu silně korelovaná. Potvrzuje to i fakt, že při použití odhadu 2, je odhad jedenáctého dne daleko méně variabilní ( $\lambda$  je v tomto případě odhadována na základě všech 16 replikací). Statistický test nezamítá na 5% hladině významnosti hypotézu normality reziduí.

### **Odhad 1 - Fourierova báze versus B-spline báze**

Rozdělení reziduí výsledných modelů vesměs vykazují těžší chvosty (ve srovnání s normálním rozdělením), ale na základě Shapirova-Wilkova testu normality (podrobný popis testu je popsán v [7]) není možné hypotézu normality na 5% hladině významnosti zamítnout. Navíc jsou rezidua daleko méně korelovaná než v případě, kdy je  $\lambda$  určeno na základě všech dat.

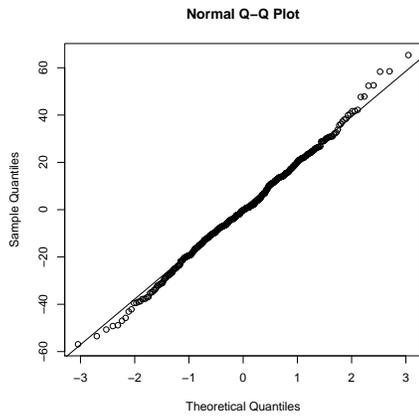
Výsledné Q-Q grafy jsou na obrázcích (6.2) (Fourierova báze) a (6.3) (b-spline báze). Přesné p-hodnoty Shapirova-Wilkova testu pro jednotlivé části modelu (6.3) jsou v tabulce (6.4). Shodnost rozdělení reziduí jednotlivých částí modelu (6.3) jsem dále testovala pomocí dvouvýběrového Kolmogorova-Smirnova testu. Nulovou hypotézu na 5% hladině významnosti nezamítám pro žádné dvě kombinace testovaných dat. Mohu tedy předpokládat, že  $\epsilon_i(t_j)$  jsou stejně rozdělená. Totéž platí pro data naměřená na chodbě.

### **Odhad 2 - Fourierova báze versus B-spline báze**

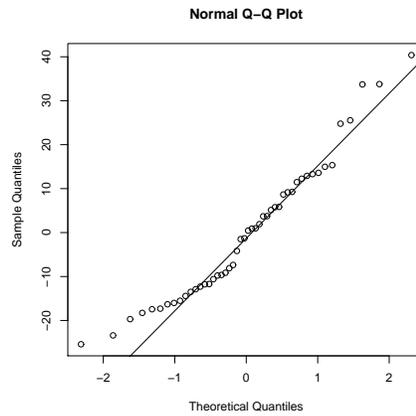
Postupovala jsem obdobně jako v předchozím případě, ale vynecháním odlehlých dnů 8 a 12 výrazné zlepšení výsledků nenastalo (viz tabulka (6.4)) a hypotézu normality na 5% hladině významnosti zamítám.

Q-Q grafy reziduí modelu (6.8) mají ve srovnání s normálním rozdělením těžké chvosty. Výsledky formálního testu jsou shrnuty v tabulce (6.4).

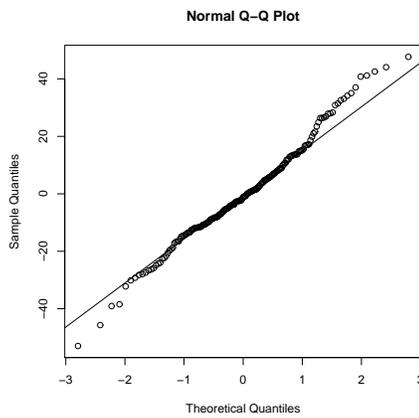
Vlastnosti odhadů jsou lepší v případě vyloučení odlehlých pozorování (dnů). K dispozici je však velmi málo dní, proto budu dále pracovat i s těmito daty. Do výpočtu penalizačního parametru  $\lambda$  však tato pozorování nevstupují. Je to kvůli tomu, abych získala model vhodný pro většinu dní, i za cenu, že pro modelování dnů 8 a 12 je nevhodný. Je však nutné brát v úvahu, že další výsledky mohou být zahrnutím těchto dnů deformovány.



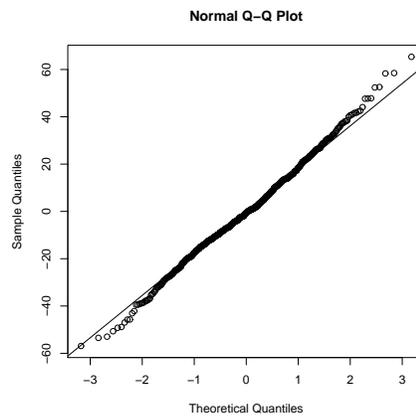
(a) Odhad (6.5) ( $d \neq 8$ )



(b) Odhad (6.6)

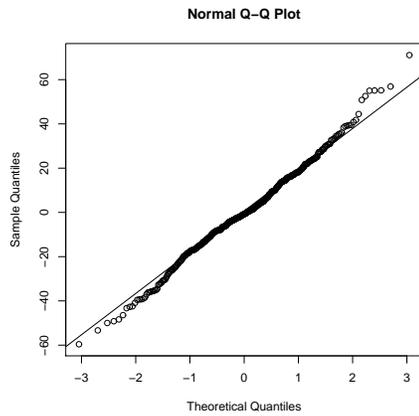


(c) Odhad (6.6) ( $d \neq 12$ )

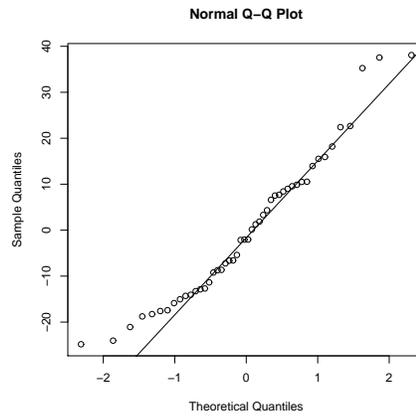


(d) Celkový odhad ( $d \neq 8$ ) a ( $d \neq 12$ )

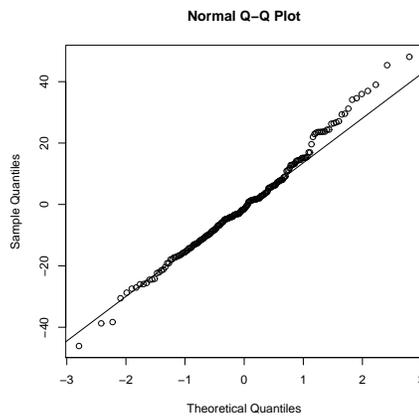
Obrázek 6.2: Q-Q grafy reziduí modelu (6.3) pro průběh OAR v dětském pokoji (Fourierova báze) - nejprve jsem testovala normalitu pro každý podmodel separátně, poslední graf ukazuje charakter celkového rozdělení reziduí.



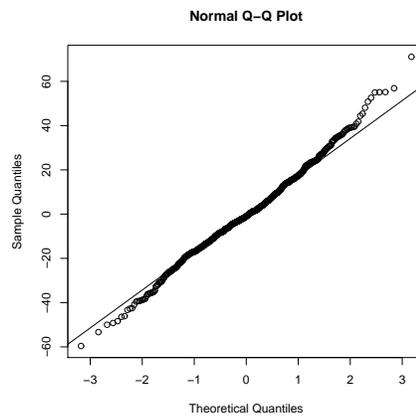
(a) Odhad (6.5) ( $d \neq 8$ )



(b) Odhad (6.6)



(c) Odhad (6.6) ( $d \neq 12$ )



(d) Celkový odhad ( $d \neq 8$ ) a ( $d \neq 12$ )

Obrázek 6.3: Q-Q grafy reziduí modelu (6.3) průběh OAR na chodbě (b-spline báze) - nejprve jsem testovala normalitu pro každý podmodel separátně, poslední graf ukazuje charakter celkového rozdělení reziduí.

Typ báze	Podmodel	Pokoj 1		Pokoj 2	
		Odhad 1	Odhad 2	Odhad 1	Odhad 2
Fourier	Odhad (6.5) ( $d \neq 8$ )	0,586	—	0,002	—
	Odhad (6.6)	0,051	—	0,056	—
	Odhad (6.6) ( $d \neq 12$ )	0,087	—	0,900	—
	Celkem	0,109	0,119	0,004	$1 \cdot 10^{-5}$
B-spline	Odhad (6.5) ( $d \neq 8$ )	0,063	—	$2 \cdot 10^{-5}$	—
	Odhad (6.6)	0,052	—	0,371	—
	Odhad (6.6) ( $d \neq 12$ )	0,194	—	0,621	—
	Celkem ( $d \neq 8$ ) a ( $d \neq 12$ )	0,007	—	$2 \cdot 10^{-5}$	—

Tabulka 6.4: Výsledné p-hodnoty Shapirova-Wilkova testu normality pro jednotlivé modely

Konkrétní předpoklady pro konstrukci odhadů jsou shrnuty v tabulce (6.5). Na obrázku (6.4) je zobrazen průměr čtverců rozdílu odhadu 1 při použití Fourierovy báze a b-splinu ( $\frac{1}{16} \sum_{i=1}^{16} \left( \widehat{f}_i^B(t_j) - \widehat{f}_i^F(t_j) \right)^2$ , kde  $\widehat{f}_i^F$  je odhad průběhu OAR během  $d$ -tého dne zkonstruovaný pomocí Fourierovy báze a  $\widehat{f}_i^B$  pomocí b-splinu). Rozdíl je patrný hlavně v okrajových bodech definičního oboru. Nejvyšších rozdílů dosahují odhady dnů 8 a 12.

Dále jsem srovnávala odhady 1 a 2. Rozdíl mezi nimi je velmi malý. Jelikož zjednodušení spočívá v použití shodných časů měření pro všechna data, ačkoli v některých dnech došlo k časovému posunu, jsou i výsledné odhady vzájemně posunuté. Jelikož zjednodušený odhad vykazuje horší statistické vlastnosti ( $MSE$ , výrazná nenormalita reziduí pro pokoj 2), nebudu ho dále používat.

## 6.2 Concurrent model

### 6.2.1 Přehled modelů

V této kapitole budu modelovat závislost OAR v dětském pokoji na průběhu OAR na chodbě pomocí concurrent modelu (4.2). Budu uvažovat následující modely (přehled viz tabulka (6.7)):

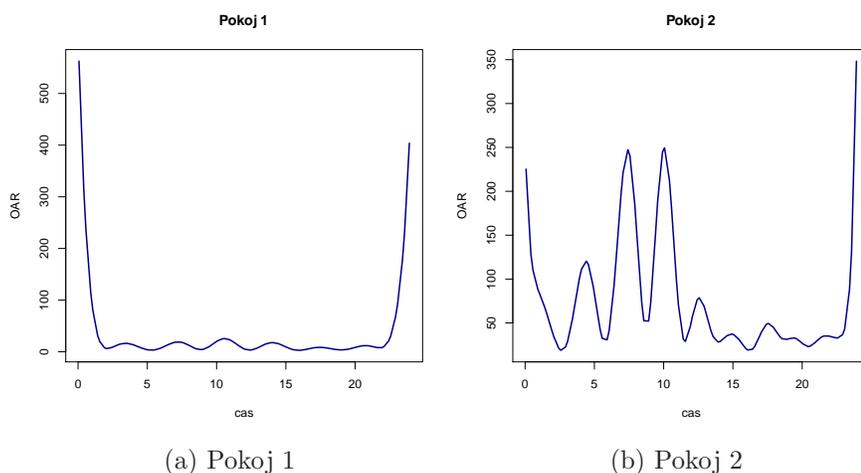
**Model 1: Bez konstantního členu + odhad 1 + Fourierova báze**

$$p_{1i}^1(t) = p_{2i}^1(t) \beta_1^1(t) + \psi_i^1(t)$$

Pro odhad funkce  $\beta_1^1$  byla použita Fourierova báze o 47 funkcích, penalizační pa-

Typ báze	Pokoj 1		Pokoj 2	
	Fourier	B-spline	Fourier	B-spline
Počet prvků báze	47	100	47	100
Definiční obor báze	[0,1667;23,9333]		[0,1667;23,9333]	
Penalizační člen	$\int_T [D^4 \hat{f}(t)]^2 dt$		$\int_T [D^4 \hat{f}(t)]^2 dt$	
$\lambda$ - odhad 1 ( $d = 1, \dots, 10$ )	2,57	9,55	0,09	9,55
$\lambda$ - odhad 1 ( $d = 11$ )	0,80	6,00	22,91	9,55
$\lambda$ - odhad 1 ( $d = 12, \dots, 17$ )	4,47	12,02	15,49	1,58
$MSE$ - odhad 1	420,58	426,44	408,99	467,11
$\lambda$ - odhad 2	3,72	—	5,25	—
$MSE$ - odhad 2	428,71	—	482,34	—

Tabulka 6.5: Výsledný odhad denního průběhu OAR. Označením  $MSE$  mám v tuto chvíli na mysli průměrnou hodnotu  $MSE$  pro jednotlivé dny.



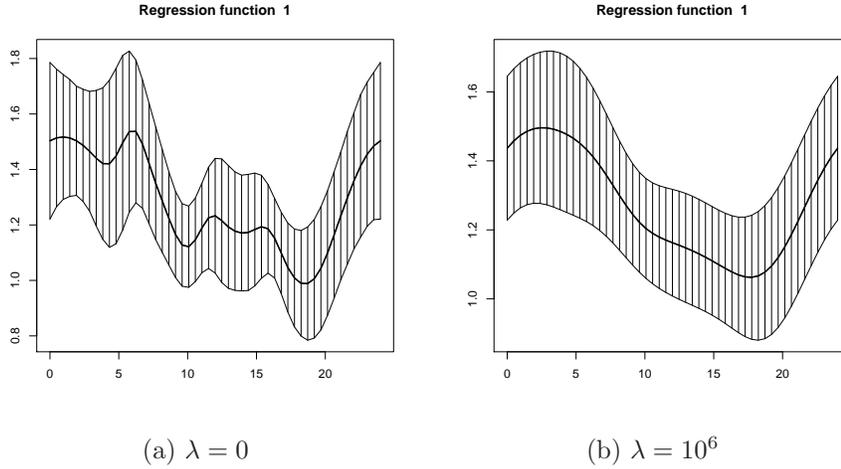
Obrázek 6.4:  $\frac{1}{16} \sum_{i=1}^{16} \left( \widehat{f}_i^B(t_j) - \widehat{f}_i^F(t_j) \right)^2$ , kde  $\widehat{f}_i^F$  je odhad průběhu OAR během  $i$ -tého dne zkonstruovaný pomocí Fourierovy báze a  $\widehat{f}_i^B$  pomocí b-splinu

	Pokoj 1		Pokoj 2	
	Fourier	B-spline	Fourier	B-spline
den 1	<b>396,39</b>	404,48	<b>178,02</b>	274,17
den 2	<b>467,61</b>	490,22	<b>356,23</b>	399,32
den 3	234,26	<b>180,52</b>	<b>291,05</b>	360,65
den 4	387,60	<b>370,25</b>	<b>424,74</b>	540,92
den 5	<b>529,37</b>	552,77	<b>538,18</b>	624,99
den 6	286,60	<b>265,48</b>	<b>263,90</b>	382,43
den 7	516,45	<b>486,98</b>	<b>376,14</b>	426,24
den 8	<b>978,33</b>	1235,05	<b>511,39</b>	1057,99
den 9	386,31	<b>384,51</b>	<b>397,71</b>	467,79
den 10	<b>398,02</b>	408,80	<b>343,25</b>	501,73
den 11	<b>241,82</b>	245,04	415,90	<b>380,09</b>
den 12	<b>708,48</b>	743,61	1097,84	<b>861,16</b>
den 13	<b>237,43</b>	240,54	391,14	<b>330,13</b>
den 14	262,33	<b>254,18</b>	258,20	<b>232,80</b>
den 15	234,71	<b>230,30</b>	403,45	<b>368,00</b>
den 16	463,62	<b>330,34</b>	296,66	<b>265,31</b>

Tabulka 6.6:  $MSE$  odhadu 1 a 2 pro oba typy báze.

Model	Báze (data)	Báze (funkce $\beta$ )		
Model 1	Fourier	Fourier	Odhad 1	$p_{1i}^1(t) = p_{2i}^1(t) \beta_1^1(t) + \psi_i^1(t)$
Model 2				$p_{1i}^2(t) = \beta_0^2(t) + p_{2i}^2(t) \beta_1^2(t) + \psi_i^2(t)$
Model 3			$p_{1i}^3(t) = \beta_0^3(t) + p_{2i}^3(t) \beta_1^3(t) + \psi_i^3(t)$	
Model 5	B-spline	B-spline	Odhad 1	$p_{1i}^5(t) = \beta_0^5(t) + p_{2i}^5(t) \beta_1^5(t) + \psi_i^5(t)$
Model 4			Odhad 1	$p_{1i}^4(t) = \beta_0^4(t) + p_{2i}^4(t) \beta_1^4(t) + \psi_i^4(t)$

Tabulka 6.7: Přehled modelů



Obrázek 6.5: Model 1 - funkce  $\beta_1^1(t)$

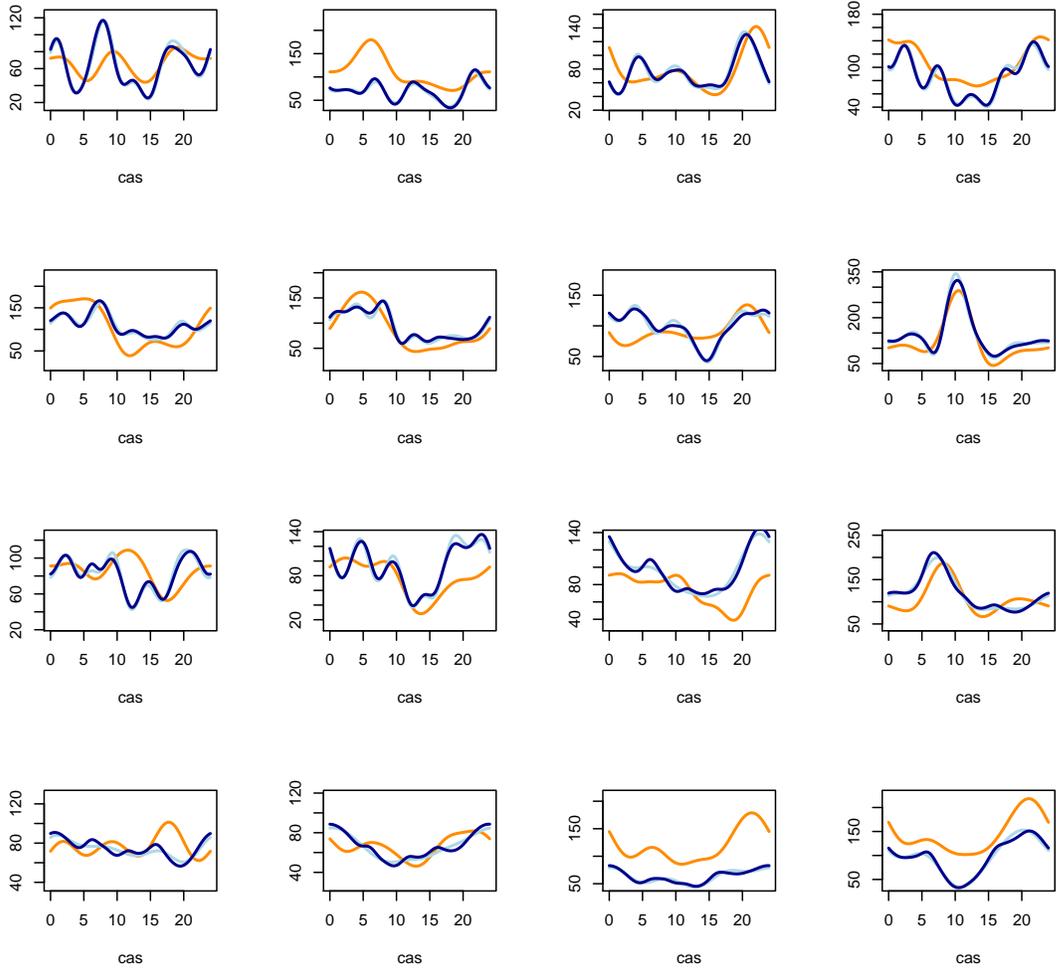
parametr  $\lambda = 0$  a  $\lambda = 10^6$ . Výsledná funkce  $\widehat{\beta}_1^1$  je na obrázku (6.5). Dle očekávání má funkce  $\widehat{\beta}_1^1$  hladší průběh pro vyšší hodnotu parametru  $\lambda_\beta$ . Šířka intervalu je pro obě varianty velmi podobná. Na obrázku 6.6 je srovnání fitů obou variant modelu. Je vidět, že model nefituje příliš dobře. Přehled hodnot  $MSE^*$  a  $AIC$  je v tabulce (6.8).

### Model 2: S konstantním členem + odhad 1 + Fourierova báze

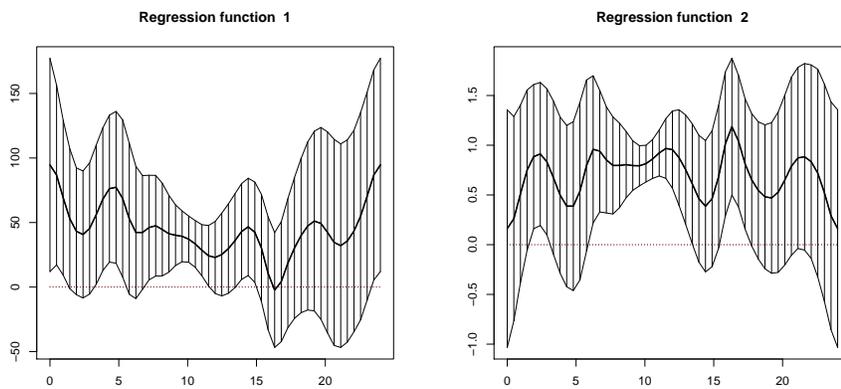
$$p_{1i}^2(t) = \beta_0^2(t) + p_{2i}^2(t) \beta_1^2(t) + \psi_i^2(t)$$

Pro odhad funkcí  $\beta_0^2$  a  $\beta_1^2$  byla použita Fourierova báze o 47 funkcích, penalizační parametry  $\lambda_\beta = (0, 0)$ ,  $\lambda_\beta = (10^3, 10^3)$  a  $\lambda_\beta = (10^6, 10^6)$ . Výsledné funkce  $\widehat{\beta}_0^2$  a  $\widehat{\beta}_1^2$  jsou na obrázku (6.7). Podobně jako v předchozím případě platí, že pro vyšší hodnoty  $\lambda_\beta$  mají funkce  $\widehat{\beta}_0^2$  a  $\widehat{\beta}_1^2$  hladší průběh. V tomto případě je značný rozdíl v šířce intervalu spolehlivosti. V prvním případě je interval spolehlivosti oproti druhým dvěma možnostem velmi široký. Pro druhé dvě varianty je šířka intervalu spolehlivosti srovnatelná. Na obrázku (6.9b) je zobrazen průběh  $MSE$  pro jednotlivé hodnoty parametru  $\lambda_\beta$ .

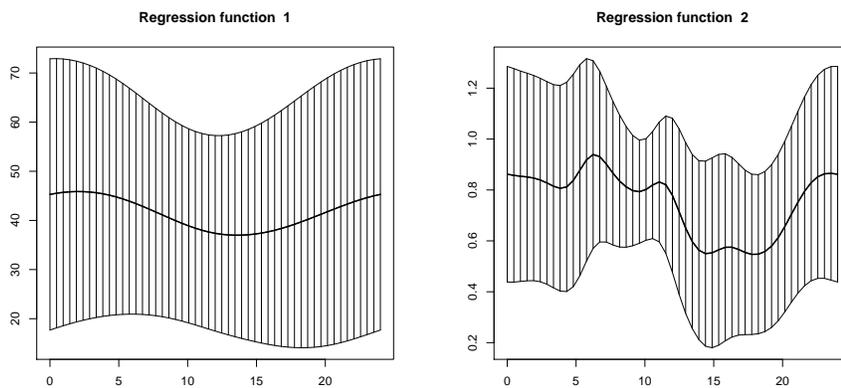
Na obrázku (6.8) je porovnání modelu s konstantním členem a bez něho. Byl použit parametr  $\lambda_\beta$ . Překvapivě dobré výsledky dávají oba odhady pro den 8, ačkoli OAR nabývala extrémních hodnot. Ani jeden z modelů nedokázal postihnout průběh OAR během 2. dne měření a oba podhodnocují množství OAR naměřené během posledních dvou dní. Porovnání  $MSE(\widehat{p}_{1i}^1)$  a  $MSE(\widehat{p}_{2i}^1)$  (viz obrázek (6.9a)) ukazuje, že nejvyšší rozdíl mezi výslednými odhady nastává kolem páté a desáté



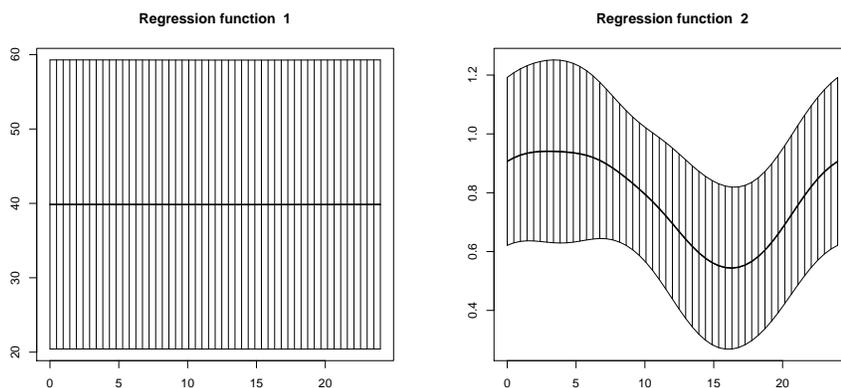
Obrázek 6.6: Model 1 -  $g_i(t)$  (oranžově),  $\hat{g}_i(t)$ :  $\lambda = 0$  (tmavěmodře);  $\lambda = 1000$  (světlemodře)



(a)  $\lambda = 0$



(b)  $\lambda = 10^3$



(c)  $\lambda = 10^6$

Obrázek 6.7: Model 2 - funkce  $\beta_0^2(t)$  a  $\beta_1^2(t)$

hodiny dopolední a naopak téměř stejné odhady získáme mezi čtvrtou a pátou hodinou odpoledne. Na základě porovnání  $AIC$  je lepší model 1. Znamená to, že zlepšení vlastností modelu přidáním konstantního členu je nedostatečné.

### Model 3: S konstantním členem + odhad 2 + Fourierova báze

$$p_{1i}^3(t) = \beta_0^3(t) + p_{2i}^3(t) \beta_1^3(t) + \psi_i^3(t)$$

Pro odhad funkcí  $\beta_0^3$  a  $\beta_1^3$  byla použita Fourierova báze o 47 funkcích, penalizační parametr  $\lambda_\beta = (0, 0)$ . Výsledné funkce  $\widehat{\beta}_0^3$  a  $\widehat{\beta}_1^3$  je na obrázku (6.10). Hodnoty střední čtvercové chyby a  $AIC$  jsou v tabulce (6.8). Vstupní data pro tento model jsou tvořena odhady průběhu OAR typu 2 (došlo k zanedbání časového posunu, ke kterému došlo 11. den, ale odhady byly konstruovány na základě všech replikací a nedošlo k rozdělení do tří podmodelů). Je vidět, že výsledky tohoto modelu jsou lepší než výsledky modelu 2. Je to pravděpodobně způsobeno vyšší stabilitou odhadů průběhu OAR.

### Model 4: S konstantním členem + odhad 1 + B-spline báze

$$p_{1i}^4(t) = \beta_0^4(t) + p_{2i}^4(t) \beta_1^4(t) + \psi_i^4(t)$$

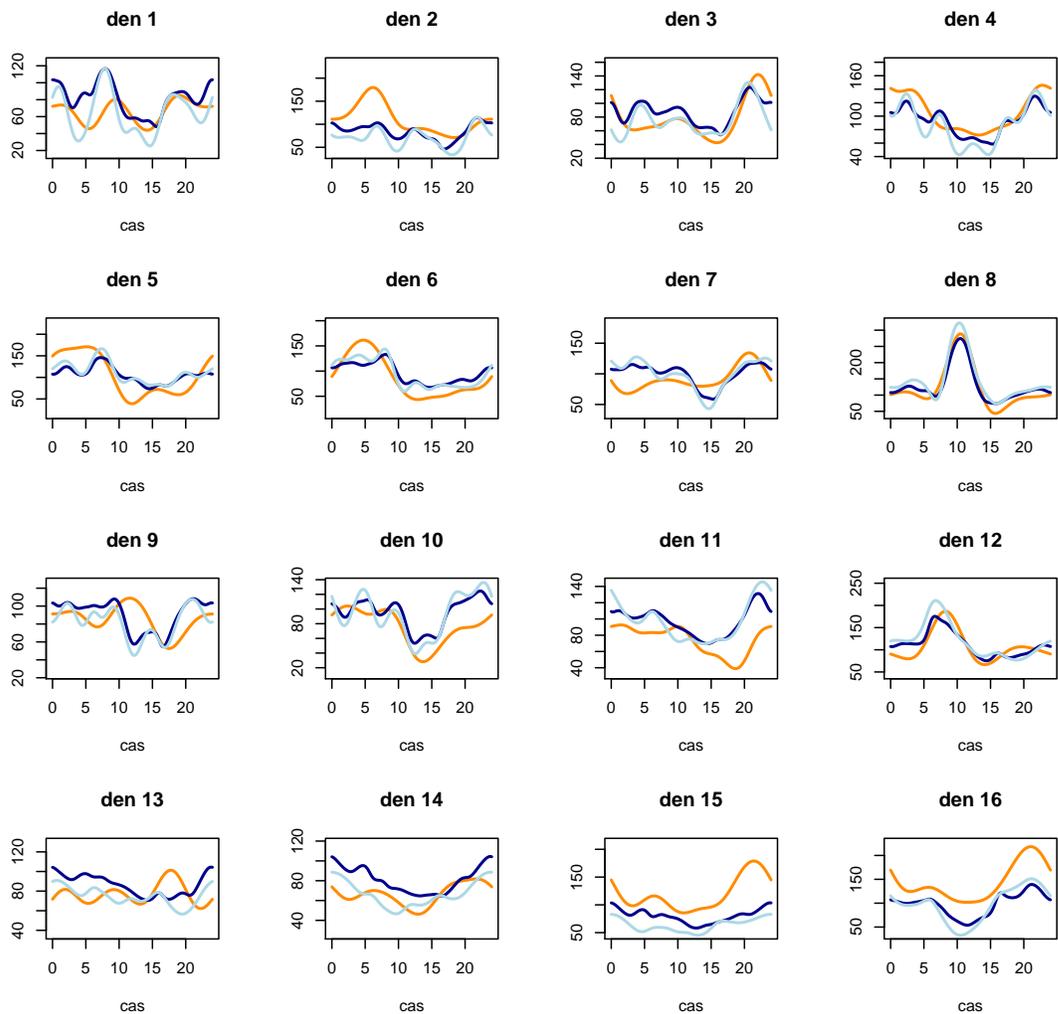
Pro odhad funkcí  $\beta_0^4$  a  $\beta_1^4$  byla použita b-spline báze s uzly v časech měření, penalizační parametr  $\lambda_\beta = (0, 0)$ . Výsledné funkce  $\widehat{\beta}_0^4$  a  $\widehat{\beta}_1^4$  je na obrázku (6.11). Vyznačují se užšími intervaly spolehlivosti než odhady ostatních modelů. Tento model ve srovnání s modelem 2 lépe fituje (nižší střední čtvercová chyba). Na obrázku (6.13) je srovnání průběhů střední čtvercové chyby modelu 2 a modelu 4. Pro většinu definovaných hodnot dává b-spline báze (model 4) lepší výsledky. Funkcionální data však v tomto případě byla konstruována pomocí b-spline báze. Předchozí analýza ukázala, že Fourierova báze je v tomto případě pro konstrukci odhadu funkcionálních pozorování vhodnější, proto uvažují následující model.

### Model 5: S konstantním členem + odhad 1 + B-spline báze pro odhad regresní funkce + Fourierova báze pro odhad dat

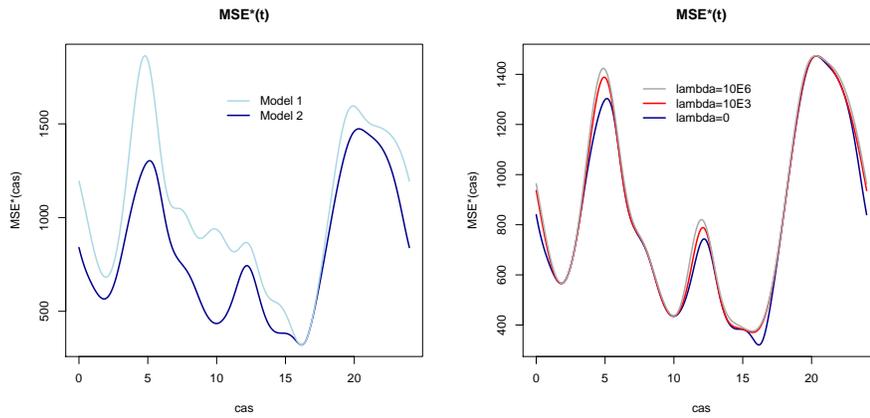
$$p_{1i}^5(t) = \beta_0^5(t) + p_{2i}^5(t) \beta_1^5(t) + \psi_i^5(t)$$

Pro odhad funkcí  $\beta_0^5$  a  $\beta_1^5$  byla použita b-spline báze s uzly v časech měření, penalizační parametr  $\lambda_\beta = (0, 0)$ . Výsledné funkce  $\widehat{\beta}_0^5$  a  $\widehat{\beta}_1^5$  je na obrázku (6.12) a je patrné, že model selhává v okrajových bodech definičního oboru.

Protože předchozí výsledky nejsou příliš dobré, snažila jsem se zvýšit kvalitu

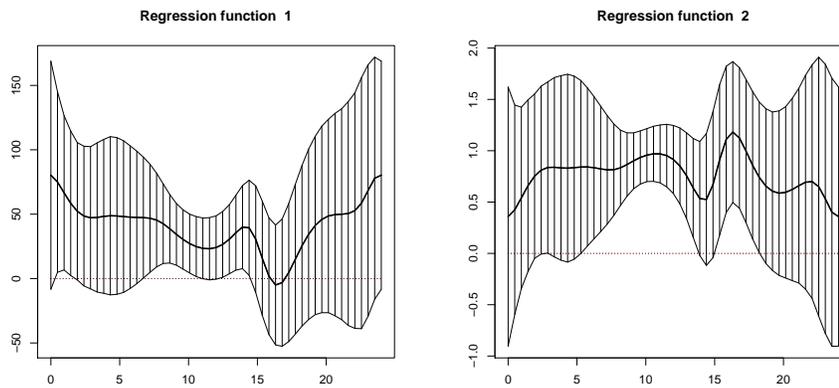


Obrázek 6.8: Srovnání fitů modelu s konstantním členem (tmavě modře) a modelu bez konstantního členu (světle modře) s původními pozorováními (oranžově).

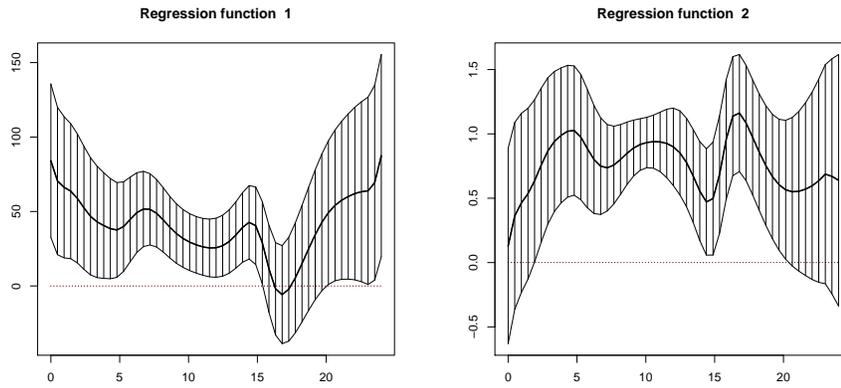


(a)  $MSE^*(\hat{p}_{1i}^1)$  a  $MSE^*(\hat{p}_{1i}^2)$  (b) Srovnání  $MSE^*(\hat{p}_{1i}^2)$  pro  $\lambda = 0$ ,  $\lambda = 10^3$  a  $\lambda = 10^6$

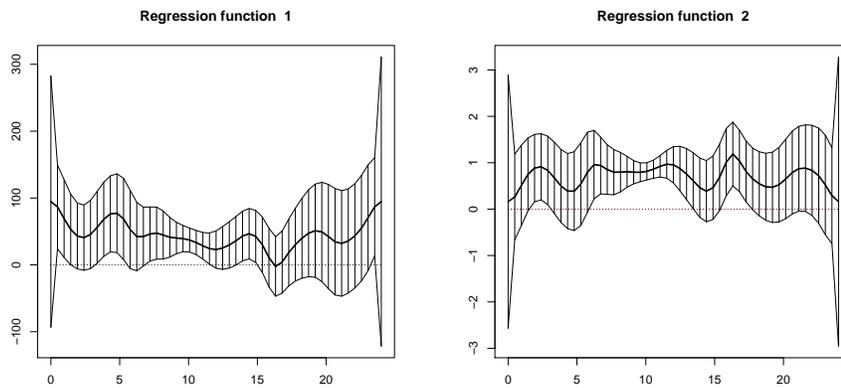
Obrázek 6.9: Průběh  $MSE^*$



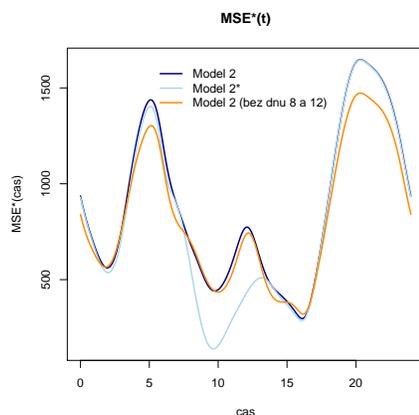
Obrázek 6.10: Model 3 - funkce  $\beta_0^3(t)$  a  $\beta_1^3(t)$  pro  $\lambda = 0$



Obrázek 6.11: Model 4 - funkce  $\beta_0^4(t)$  a  $\beta_1^4(t)$  pro  $\lambda = 0$



Obrázek 6.12: Model 5 - funkce  $\beta_0^5(t)$  a  $\beta_1^5(t)$  pro  $\lambda = 0$



Obrázek 6.13: Srovnání  $MSE^*(\hat{p}_{1i}^2)$  a  $MSE(\hat{p}_{1i}^4)$  pro  $\lambda = 0$

modelů zahrnutím další vysvětlující proměnné. V tomto případě to byla teplota (vhodnější by byla vlkost, ale bohužel její hodnoty jsem neměla k dispozici). Ukázalo se, že kvalita modelu se nezlepšila.

Další cestou, kterou jsem vyzkoušela bylo vyloučení kritických pozorování ze dne 8 a 12. Domnívala jsem se, že tato dvě pozorování příliš ovlivňují celkový charakter modelu a mohou způsobovat to, že model nefunguje dobře pro ostatní data. Na obrázku (6.13) je průběh střední čtvercové chyby modelu s vyloučenými pozorováními (dále model 2\*), modelu 2 a střední čtvercové chyby modelu 2, při jejímž výpočtu byly vynechány dny 8 a 12. Z porovnání vychází, že model 2\* lépe fituje kolem desáté hodiny dopolední. V tuto dobu dosahoval průběh OAR ve dnech 8 a 12 extrémních hodnot, a tudíž ovlivnily i výsledný model. Když však porovnam průběh střední čtvercové chyby modelu 2 (varianta počítaná ze všech data versus varianta počítaná bez dnů 8 a 12), jsou výsledky překvapivě podobné. Vyloučením kritických dnů dokonce docílím zhoršení střední čtvercové chyby (model 2 kritické dny fituje velmi dobře). Dá se říci, že vliv kritických dnů je pouze lokální a jejich vyloučením nedojde ke zlepšení celkového chování modelu.

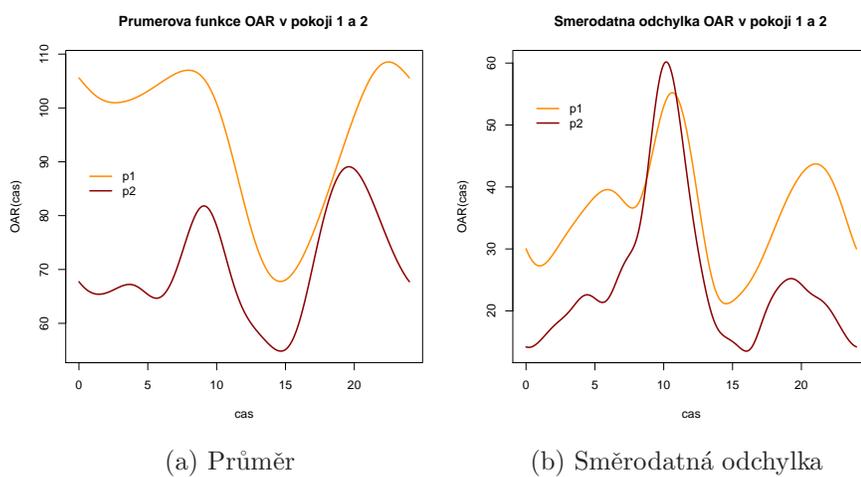
Z výsledků shrnutých v tabulce (6.8) a z poznatků získaných testováním obměn původních modelů, které nepřinesly žádný užitek, jsem dospěla k závěru, že část variability průběhu OAR naměřené v prvním pokoji je vysvětlitelná hodnotami z druhého pokoje. Toto množství je ale malé a bylo by dobré zařadit vhodnější regresor. Navíc residua žádného z modelů nevykazují normální charakter. Zkoušela jsem konstruovat model, který by popisoval vztah mezi průběhy OAR v různých časech (se zpožděním), ale ani toto rozšíření původního modelu nedává dobré smysluplné výsledky.

Průběh střední hodnoty obou souborů dat je velmi podobný (viz obr. (6.14a)) stejně jako směrodatná odchylka (viz obr. (6.14b)), která však dosahuje vysokých

	<i>MSE</i>	<i>AIC</i>
Model 1 ( $\lambda = 0$ )	24869,3	1383,2
Model 1 ( $\lambda = 10^6$ )	25417,4	1383,8
Model 2 ( $\lambda = 0$ )	19766,5	2529,4
Model 2 ( $\lambda = 10^3$ )	20276,1	2530,0
Model 2 ( $\lambda = 10^6$ )	20670,0	2530,5
Model 3 ( $\lambda = 0$ )	19454,2	2529,2
Model 4 ( $\lambda = 0$ )	19515,9	2529,3
Model 5 ( $\lambda = 0$ )	19766,5	2529,4

Tabulka 6.8: Porovnání modelů: Do výpočtu *AIC* jsem nezahrnula člen *C*, který závisí pouze na vstupních datech. Z tohoto důvodu nejsou všechny modely porovnatelné.

hodnot.



Obrázek 6.14: Průměr a směrodatná odchylka průběhu OAR v pokoji 1 a 2.

# Kapitola 7

## Závěr

Funkcionální datová analýza je rychle se rozvíjející směr a její možnosti jsou velmi široké. Ve svojí práci jsem se zabývala aplikací concurrent modelu na reálná data a testováním citlivosti metod na nesplnění předpokladů. Vhodnost výběru concurrent modelu jsem si ověřila pomocí klasických regresních modelů. Za náhodné veličiny jsem považovala měření provedená v jednom časovém okamžiku a modelovala jsem závislost jejich různých kombinací. Na základě vyhodnocení koeficientů determinace jsem zjistila, že nejvíce variability modelované proměnné vysvětlím pomocí hodnot vysvětlující proměnné naměřených ve stejném okamžiku. Je nutné poznamenat, že závislost mezi sledovanými veličinami nebyla prokazatelná ve všech časech měření a nebyla příliš silná. Tato skutečnost se později odrazila i ve vlastnostech výsledného concurrent modelu.

Kapitola (3) se zabývá obecným postupem vyhlazení řady diskrétních hodnot a vytvořením funkcionálního pozorování. Odhady konstruuji jako lineární kombinaci funkcí, které tvoří bázi podprostoru prostoru spojitých diferencovatelných funkcí. Na vlastním příkladu jsem ukázala důležitost správného výběru báze. Porovnála jsem chování Fourierovy báze a b-splinu pro periodická a neperiodická data. Závěrem je, že b-spline je možné použít pro oba typy dat, ale periodická Fourierova báze pro neperiodická data selhává v okrajových hodnotách definičního oboru. Dále v této kapitole zmiňuji možnosti odhadu koeficientů lineární kombinace. Výsledky čerpané z literatury jsem doplnila o postupy jejich odvození (například odvození minimalizačních kritérií (3.14), důkaz věty shrnující základní vlastnosti b-splinu, vztah mezi CV a GCV metodou).

V kapitole (4) uvádím rozšíření lineárně regresního modelu pro funkcionální náhodné veličiny. Základní výsledky jsem převzala z literatury, ale doplnila jsem je o postupy odvození (viz (4.13)) a o kritéria pro vzájemné porovnávání modelů (čtvercové chyby, AIC). Ta jsem získala jako vhodné rozšíření jejich jednozměrných variant (bylo nutné získat jejich jednoznačnou porovnatelnost).

V dalších kapitolách se věnuji aplikaci dříve popsané teorie na konkrétní data. Ka-

pitola (2) shrnuje vlastnosti použitých dat. Prostřednictvím simulačních studií jsem porovnávala chování modelu s normálně rozdělenými náhodnými složkami proti rozdělení s těžšími chvosty. Ukázalo se, že metoda není příliš citlivá na porušení normality náhodných složek modelu. Dále mě zajímalo, zda je metodu možné použít i pro korelovaná data. V tomto případě jsou výsledné odhady sice vychýlené, ale ne tolik, abych mohla prohlásit, že metoda selhala. V další simulační studii porovnávala chování concurrent modelu pro různé hodnoty parametru  $\lambda$ .

Pro účely hledání vhodného modelu popisující vztah mezi průběhem OAR v dětském pokoji a na chodbě považuji data za realizaci náhodné veličiny, která popisuje průběh OAR během jednoho dne. Tento přístup není zcela korektní, protože data jsou ve skutečnosti korelovaná. Proto se jedná spíše o aproximaci. Nezávislost se projeví v tom, že průběhy OAR v jednotlivých dnech na sebe nenavazují. V této kapitole jsem mimo jiné testovala jaký vliv mohou na vyhlazení mít nedefinované hodnoty. Přistupovala jsem k tomu dvěma způsoby. První je kvantitativní a ukázal, že s lineárně rostoucím počtem nedefinovaných hodnot roste střední čtvercová chyba rychleji než lineárně a že je tvořena převážně rozptylem odhadu (vychýlení je malé). Druhý přístup byl kvalitativní a jeho výsledky ukazují, že závisí nejen na tom, jak daleko se nedefinovaná hodnota nachází od sousedního pozorování, ale také na konkrétních hodnotách sousedních pozorování.

Ve zbylé části této kapitoly jsem se snažila zkonstruovat model vhodný k popisu vztahu mezi sledovanými náhodnými veličinami. Porovnávala několik variant (testovala jsem přidání dalšího regresoru, odlehlá pozorování, různé varianty konstrukce odhadů funkcionálních dat apod.) a jako nejlepší se mi jeví nejjednodušší model bez konstantního členu (jeho přidáním nedošlo k dostatečnému zlepšení modelu). Nedá se však prohlásit, že by pro modelování závislé proměnné byl vhodný, protože vysvětluje příliš malé množství variability. Potvrdila se tedy silná souvislost s klasickou lineární regresí.

# Literatura

- [1] de Boor, C.: *A Practical Guide to Splines*. Springer, 2001.
- [2] Craven, P.; Wahba, G.: Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-validation. *Numerische Mathematik*, ročník 31, 1998: s. 377–403.
- [3] Efron, B.; Tibshirani, R. J.: *An Introduction to the Bootstrap*. CRC Press LLC, 1998.
- [4] Petrásek, J.: Metody bootstrap pro závislá pozorování. *Diplomová práce*, 2008.
- [5] Ramsay, J. O.; Hooker, G.; Graves, S.: *Functional data analysis with R and MATLAB*. Springer, 2009, 202 s.
- [6] Ramsay, J. O.; Silverman, B. W.: *Functional Data Analysis*. Springer, 1997.
- [7] Royston, P.: Algorithm AS 181: The W test for Normality. *Applied Statistics*, ročník 31, 1982: s. 176–180.