MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Bakalářská práce

BRNO 2021 MAGDALÉNA ZEDNÍČKOVÁ

MASARYKOVA
UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
ÚSTAV MATEMATIKY A STATISTIKY

Gaussovské procesy v regresi

Bakalářská práce

Magdaléna Zedníčková

Vedoucí práce: doc. Mgr. Jan Koláček, Ph.D. **Brno 2021**

Bibliografický záznam

Autor:	Magdaléna Zedníčková Přírodovědecká fakulta, Masarykova univerzita Ústav matematiky a statistiky
Název práce:	Gaussovské procesy v regresi
Studijní program:	Matematika
Studijní obor:	Statistika a analýza dat
Vedoucí práce:	doc. Mgr. Jan Koláček, Ph.D.
Akademický rok:	2020/2021
Počet stran:	ix + 36
Klíčová slova:	gaussovský proces v regresi; hyperparametr; kovarianční funkce; predikce; simulace

Bibliographic Entry

Author:	Magdaléna Zedníčková Faculty of Science, Masaryk University Department of Mathematics and Statistics

Title of Thesis:	Gaussian Processes for Regression
Degree Programme:	Mathematics
Field of Study:	Statistics and Data Analysis
Supervisor:	doc. Mgr. Jan Koláček, Ph.D.
Academic Year:	2020/2021
Number of Pages:	ix + 36
Keywords:	Gaussian process for regression; hyperparameter; covariance function; prediction; simulation

Abstrakt

Bakalářská práce se věnuje gaussovským procesům, což je jeden z moderních přístupů k regresní analýze. Teoretická část obsahuje definici gaussovského procesu a jeho vlastnosti, uvádí stručný přehled kovariančních funkcí, popisuje metodu pro užití gaussovských procesů v regresi a podrobně znázorňuje vliv hyperparametrů na odhad regresní křivky. Cílem praktické části práce je použití této teorie v praxi, proto je aplikována na simulovaná a reálná data za pomocí softwaru R. Na těchto datech je vytvořeno několik různých modelů a jsou vyhodnoceny ty nejlepší.

Abstract

The bachelor thesis deals with Gaussian processes, which is one of the modern approaches to regression analysis. The theoretical part contains the definition of the Gaussian process and its properties and a brief summary of covariance functions. It also includes a description of method for using Gaussian processes for regression and it represents the influence of hyperparameters on the estimation of regression curve. The aim of the practical part is usage of the theory in practice, therefore it is applied to simulated and real data using R software. Several different models are made using this data and the best models are evaluated.

MASARYKOVA UNIVERZITA
PŘÍRODOVĚDECKÁ FAKULTA
KOTLÁŘSKÁ 2, 611 37 BRNO
IČ: 00216224
DIČ: CZ00216224

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Akademický rok: 2020/2021

Ústav:	Ústav matematiky a statistiky
Studentka:	Magdaléna Zedníčková
Program:	Matematika
Obor:	Statistika a analýza dat

Ředitel *ústavu* PřF MU Vám ve smyslu Studijního a zkušebního řádu MU určuje bakalářskou práci s názvem:

Název práce:	Gaussovské procesy v regresi
Název práce anglicky:	Gaussian Processes for Regression
Jazyk závěrečné práce:	čeština

Oficiální zadání:
Teorie gaussovských procesů je jedním z moderních přístupů k regresní analýze. Cílem práce je uvést základní popis této metodologie a pomocí simulací ukázat vliv hyperparametrů na celkový odhad regresní křivky. V závěru tento přístup bude aplikován na reálných datech.

Literatura:
RASMUSSEN, Carl Edward a Christopher K. I. WILLIAMS. *Gaussian processes for machine learning*. Cambridge, Mass.: MIT Press, 2006. xviii, 248. ISBN 026218253X.

Vedoucí práce:	doc. Mgr. Jan Koláček, Ph.D.
Datum zadání práce:	19. 4. 2020
V Brně dne:	16. 5. 2021

Zadání bylo schváleno prostřednictvím IS MU.

Magdaléna Zedníčková, 2. 11. 2020
doc. Mgr. Jan Koláček, Ph.D., 9. 11. 2020
RNDr. Jan Vondra, Ph.D., 18. 11. 2020

Poděkování

Na tomto místě bych chtěla poděkovat vedoucímu práce doc. Mgr. Janu Koláčkovi, Ph.D. za vedení, podněty a připomínky v průběhu vypracování práce.

Prohlášení

Prohlašuji, že jsem svoji bakalářskou práci vypracovala samostatně pod vedením vedoucího práce s využitím informačních zdrojů, které jsou v práci citovány.

Brno 20. května 2021

. . . . . .
Magdaléna Zedníčková

Obsah

Přehled použitého značení

Pro snažší orientaci v textu zde čtenáři předkládáme přehled základního značení, které se v celé práci vyskytuje.

$Y_1, \ldots, Y_n$	vysvětlované proměnné
$x_{1,1}, \ldots, x_{n,k}$	kovariáty neboli vysvětlující proměnné
$\beta_0, \ldots, \beta_k$	regresní koeficienty
$\varepsilon_1, \ldots, \varepsilon_n$	náhodné chyby
$Cov(\varepsilon_i, \varepsilon_j)$	kovariance náhodných chyb
$Cov(Y_i, Y_j)$	kovariance vysvětlovaných proměnných
$Var\varepsilon_i$	rozptyl i-té náhodné chyby
$VarY_i$	rozptyl i-té vysvětlované proměnné
$E\varepsilon_i$	střední hodnota i-té náhodné chyby
EY_i	střední hodnota i-té vysvětlované proměnné
$h(\mathbf{X})$	hodnost matice plánu
ρ	korelační koeficient
$\sim$	rozděleno jako; např. $\mathcal{N}(\mu, \sigma^2)$
$\overset{\text{iid}}{\sim}$	nezávislé stejně rozdělené náhodné veličiny
$\propto$	přímá úměrnost
$\{X_t;\ t \in T\}$	stochastický proces se spojitým časem
$t_1, \ldots, t_k$	indexy z T
T	indexová množina
$\mathcal{GP}$	gaussovský proces
$\mathcal{D}$	datová sada, tzv. trénovací množina
X	náhodná veličina v kapitole 2, vstupní trénovací hodnoty v kapitole 3 a 4
$\mathbf{X}$	náhodný vektor
X_*	vstupní testovací hodnoty
$\mathcal{X}$	abstraktní prostor
f_*	výstupní testovací hodnoty
f	výstupní trénovací hodnoty
σ^2	rozptyl

$\mathbf{\Sigma}$	kovarianční matice
$k(\mathbf{x}, \mathbf{x}')$	kovarianční funkce
$cov(\mathbf{f}_*)$	aposteriorní kovarianční matice GPR
$K(X,X)$	kovarianční matice mezi trénovacími vstupy
$K(X_*,X_*)$	kovarianční matice mezi testovacími vstupy
$K(X,X_*)$	kovarianční matice mezi trénovacími a testovacími vstupy
μ	střední hodnota
$\boldsymbol{\mu}$	vektor středních hodnot
$m(\mathbf{x})$	funkce středních hodnot
$\bar{\mathbf{f}}_*$	aposteriorní vektor středních hodnot GPR
σ_f^2	celkový rozptyl
σ_n^2	šum
l	měřítko délky
p	perioda
ν	stupeň hladkosti
K_ν	modifikovaná Besselova funkce
δ_{pq}	Kroneckerovo delta
I	jednotková matice
$\boldsymbol{\theta}$	vektor hyperparametrů; např. $\boldsymbol{\theta} = (\sigma_f, l, \sigma_n)$
$\hat{\boldsymbol{\theta}}$	odhad vektoru hyperparametrů $\boldsymbol{\theta}$
$L(\boldsymbol{\theta})$	věrohodnostní funkce
$\ell(\boldsymbol{\theta})$	logaritmická věrohodnostní funkce
$\mathbb{R}$	množina všech reálných čísel
$\mathbb{R}^n$	n-rozměrná množina všech reálných čísel

Úvod

Teorie gaussovských procesů je jeden z moderních přístupů k regresní analýze. Jedná se o neparametrický přístup k regresi, který nese jméno po známém matematikovi Carlu Friedrichu Gaussovi, jelikož je založen na pojmu normálního (Gaussova) rozdělení.

Cílem práce je popsat teorii a postup užití tohoto přístupu a pomocí simulací ukázat vliv hyperparametrů kovarianční funkce na celkový odhad regresní křivky. Na základě teoretických poznatků jsou tyto znalosti aplikovány na simulovaná a reálná data. Praktická část je vytvořena za pomocí softwaru R.

V první kapitole je stručně popsána jedna ze základních metod regresní analýzy, a to lineární regrese. Tato metoda je v praktické části srovnána s užitím gaussovských procesů v regresi. Druhá kapitola nejprve krátce uvádí bayesovskou statistiku a vícerozměrné normální rozdělení. Pojmy a vztahy z těchto dvou podkapitol jsou velmi důležité při následném definování gaussovských procesů. Pro ty je zásadním pojmem kovarianční funkce, proto následuje výčet několika nejznámějších. Třetí kapitola se zabývá postupem užití gaussovských procesů v regresi a znázorňuje, jak volba hyperparametrů kovarianční funkce ovlivňuje chování odhadované regresní křivky.

Poslední kapitola je stěžejní, jelikož popisuje chování gaussovských procesů na simulovaných a reálných datech. Podkapitola se simulovanými daty uvádí srovnání křivky získané pomocí gaussovských procesů se vzhledem známé funkce, poté je použit volně dostupný balíček z R ke srovnání jeho automaticky vypočítané predikce a predikce získané pomocí vzorců zmíněných v kapitole 3. Dále je odhadnutá křivka získaná pomocí gaussovských procesů srovnána s křivkou lineárního modelu. Na závěr jsou zpracovány dvě sady reálných dat. Nejprve je sledován růst dugonga indického v závislosti na jeho věku, druhá datová sada zaznamenává počet úmrtí způsobených nehodou v USA mezi lety 1973 až 1978.

Kapitola 1

Lineární regrese

Lineární regrese se ve statistice využívá k popisu vztahu mezi jednou náhodnou veličinou (vysvětlovanou proměnnou) a několika jinými veličinami (vysvětlujícími proměnnými nebo kovariátami). Lineární regrese požaduje, aby byl model lineární v regresních koeficientech. Vysvětlovaná proměnná je tedy lineární funkcí kovariát.

Obsahem kapitoly je popis lineárního modelu, který je čerpán z [1] a [2].

1.1 Lineární model

Lineární model lze zapsat ve tvaru

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i, \quad i = 1,2,\ldots,n, \tag{1.1}$$

který popisuje závislost mezi vysvětlovanou proměnnou Y_i a kovariátami $x_{i,1}, x_{i,2}, \ldots, x_{i,k}$. V rovnici (1.1) se $\beta_0, \ldots, \beta_k$ nazývají regresní koeficienty a ε_i náhodné chyby. Kovariáty a regresní koeficienty jsou nenáhodné veličiny, naopak vysvětlovaná proměnná a chyby jsou veličiny náhodné.

Aby se jednalo o lineární model, musí být splněno několik předpokladů:

1. $E\varepsilon_i = 0$ pro $i = 1,2,\ldots,n$ neboli $EY_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k}$,
2. $Var\varepsilon_i = \sigma^2$ pro $i = 1,2,\ldots,n$ neboli $VarY_i = \sigma^2$,
3. $Cov(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j,\ i,j = 1,2,\ldots,n$ neboli $Cov(Y_i, Y_j) = 0$.

Předpokládá se tedy, že náhodné chyby mají nulovou střední hodnotu, stejný rozptyl, stejné rozdělení a jsou nekorelované. Tyto předpoklady lze zapsat ve tvaru $\varepsilon_i \sim (0, \sigma^2)$. V některých případech může být požadováno, aby byly normálně rozdělené. Jelikož v normálním rozdělení nekorelovanost implikuje nezávislost, jsou náhodné chyby i nezávislé. Pak lze tyto předpoklady zapsat ve tvaru $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

Pro jednoduchost se někdy využívá maticového zápisu

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.2}$$

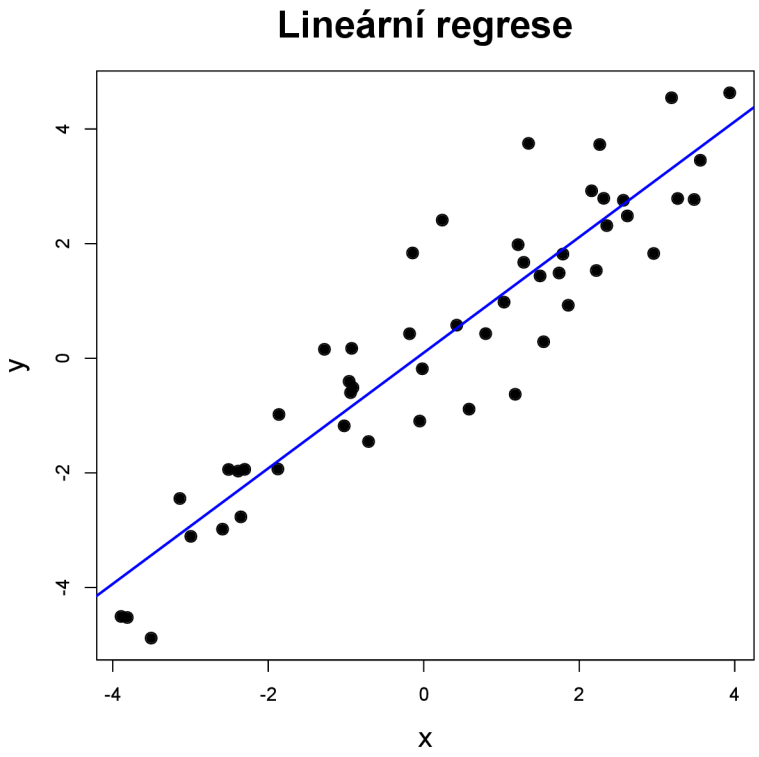
kde

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,k} \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

V (1.2) se matice $\mathbf{X}$ nazývá matice plánu, její dimenze je $n \times p$, kde $p = k+1$. Předpokládá se, že $n > p$ a $h(\mathbf{X}) = p$, tedy matice $\mathbf{X}$ má plnou hodnost. Předchozí předpoklady lze pro (1.2) zapsat takto:

1. $E\boldsymbol{\varepsilon} = \mathbf{0}$, tedy $E\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}$,
2. $Var\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$, tedy $Var\mathbf{Y} = \sigma^2\mathbf{I}$.

Jejich zápis je pak ve tvaru $\boldsymbol{\varepsilon} \sim (0, \sigma^2\mathbf{I})$ nebo $\boldsymbol{\varepsilon} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2\mathbf{I})$ pro normální model. Na obrázku 1.1 je znázorněna lineární regrese pro náhodně vygenerované body.



Obr. 1.1: Lineární regrese náhodně vygenerovaných bodů

Kapitola 2

Gaussovské procesy

Gaussovské procesy (GP) umožňují vytvářet predikce o datech. Definují apriorní rozdělení nad funkcemi, které může být po zpozorování dat převedeno na aposteriorní rozdělení nad funkcemi. Tento přístup popisuje bayesovská statistika. Gaussovské procesy jsou založeny na vícerozměrném normálním rozdělení, které rozšiřují do nekonečné dimenze. Na základě spojitosti nebo diskrétnosti dat se gaussovské procesy využívají k regresi nebo klasifikaci.

Obsahem kapitoly je úvod do bayesovské statistiky, shrnutí vícerozměrného normálního rozdělení, popis gaussovských procesů a jejich kovarianční funkce.

2.1 Úvod do bayesovské statistiky

Podkapitola se krátce zmiňuje o bayesovské statistice, jelikož je využívána při práci s gaussovskými procesy. Základem bayesovské statistiky je Bayesův vzorec. Existuje mnoho jeho modifikací, zde je uvedena ta, která je využívána při výběru vhodného modelu a je zavedena v [3].

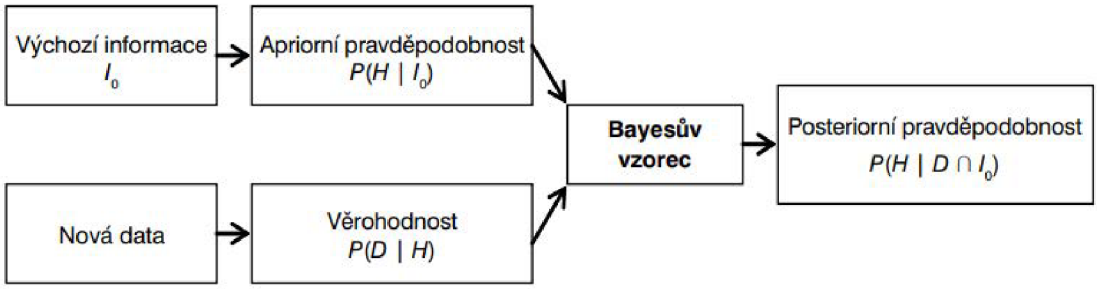
Věta 2.1 (Bayesův vzorec)**.** *Nechť θ je parametr modelu a D jsou pozorovaná data. Pak má Bayesův vzorec tvar*

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}, \text{ kde } p(D) = \int p(D|\theta)p(\theta)\,\mathrm{d}\theta. \tag{2.1}$$

Jednotlivé části vzorce (2.1) jsou důležité pojmy pro bayesovskou statistiku:

- $p(\theta)$ - apriorní pravděpodobnost; vyjadřuje prvotní znalosti o parametru θ předtím, než jsou k dispozici data D
- $p(\theta|D)$ - aposteriorní pravděpodobnost; oproti apriorní pravděpodobnosti už bere v úvahu pozorovaná data D
- $p(D|\theta)$ - věrohodnostní funkce; pravděpodobnost, že budou zpozorována data D za podmínky parametru θ
- $p(D)$ - evidence

Jelikož evidence nezávisí na parametru θ, může být aposteriorní pravděpodobnost psána jako $p(\theta|D) \propto p(D|\theta)p(\theta)$, kde $\propto$ značí přímou úměrnost.



Obr. 2.1: Proces získání aposteriorní pravděpodobnosti (Zdroj:[4])

2.2 Vícerozměrné normální rozdělení

Vícerozměrné normální rozdělení (MVN) je zobecněním normálního rozdělení $\mathcal{N}(\mu, \sigma^2)$ pro n-rozměrnou náhodnou veličinu $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$. Každá náhodná veličina je normálně rozdělena a i jejich sdružené rozdělení je normální. Jde o nejčastěji používané rozdělení u spojitých náhodných veličin. Následující definice vychází z [5] a [6].

Definice 1. Řekneme, že náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ má vícerozměrné normální (Gaussovo) rozdělení s parametry $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T \in \mathbb{R}^n$ a maticí $\boldsymbol{\Sigma} > 0$ rozměru $n \times n$, pokud jeho hustota má tvar

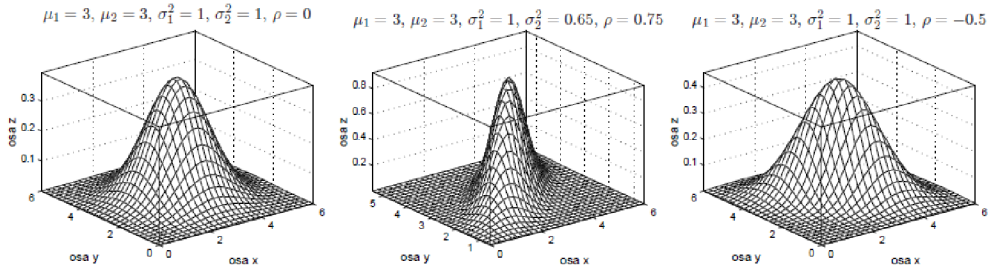
$$f(\mathbf{x}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}, \tag{2.2}$$

a budeme psát $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
$\boldsymbol{\Sigma} > 0$ znamená, že matice je pozitivně definitní a tedy i regulární. Této matici se říká kovarianční matice. Parametr $\boldsymbol{\mu}$ se nazývá vektor středních hodnot.

Například pro $n = 2$ má hustota tvar

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1^2\sigma_2^2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}$$

a značí se $\mathbf{X} = (X_1, X_2)^T \sim \mathcal{N}_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.



Obr. 2.2: Ukázky hustot vícerozměrného normálního rozdělení pro $n = 2$ (Zdroj:[5])

Definice 2. Nechť $\boldsymbol{\mu} \in \mathbb{R}^n$ a $\boldsymbol{\Sigma} > 0$ je matice $n \times n$. Náhodný vektor $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ má vícerozměrné normální rozdělení $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ právě tehdy, když $\mathbf{a}^T\mathbf{X} \sim \mathcal{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a})$ pro každé $\mathbf{a} \in \mathbb{R}^n$.

Obě definice jsou ekvivalentní. Následující věty jsou získány z [6] a [7].

Věta 2.2 (Vlastnosti)**.** *Nechť $\boldsymbol{\mu} \in \mathbb{R}^n$ a $\boldsymbol{\Sigma} > 0$ je $n \times n$ symetrická pozitivně definitní matice. Pak platí:*

1. *Nechť $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Pak $E\mathbf{X} = \boldsymbol{\mu}$ a $Var\mathbf{X} = \boldsymbol{\Sigma}$.*
2. *Nechť $Z_1, \ldots, Z_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ a $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$. Pak $\mathbf{Z} \sim \mathcal{N}_n(0, \mathbf{I})$.*
3. *Nechť $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ a nechť $\mathbf{A}$ je $m \times n$ reálná matice a $\mathbf{b} \in \mathbb{R}^m$ je vektor konstant. Potom $\mathbf{AX} + \mathbf{b} \sim \mathcal{N}_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$.*

Věta 2.3 (Marginalizace a podmíněnost)**.** *Nechť $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T)^T$, $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$ a dimenze jednotlivých složek jsou $k \times 1$ pro $\mathbf{X}_1$ a $\boldsymbol{\mu}_1$ a $k \times k$ pro $\boldsymbol{\Sigma}_{11}$. Pak platí:*

1. *$\mathbf{X}_1 \sim \mathcal{N}_k(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$.*
2. *Jestliže $\boldsymbol{\Sigma}_{12} = 0$, pak $\mathbf{X}_1$ a $\mathbf{X}_2$ jsou nezávislé.*
3. *Je-li $\boldsymbol{\Sigma}_{22}$ regulární, pak podmíněné rozdělení $\mathbf{X}_1$, je-li dáno $\mathbf{X}_2 = \mathbf{x}_2$, je k-rozměrné normální se střední hodnotou*

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$$

a rozptylem

$$\boldsymbol{\Sigma}_{11|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

Tedy $\mathbf{X}_1|\mathbf{X}_2 \sim \mathcal{N}_k(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{11|2})$.
(Analogicky by platilo i pro $\mathbf{X}_2|\mathbf{X}_1 \sim \mathcal{N}_k(\boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{22|1})$).

Z věty vyplývá, že marginální i podmíněné rozdělení je opět normální.

2.3 Gaussovské procesy

Pro další úvahy bude potřeba následující definice, která je uvedena v [8].

Definice 3. Stochastický proces se spojitým časem $\{X_t;\ t \in T\}$ je gaussovský právě tehdy, když pro každou konečnou množinu indexů $t_1, \ldots, t_k$ z indexové množiny T je

$$X_{t_1,\ldots,t_k} = (X_{t_1}, \ldots, X_{t_k})$$

vícerozměrná normální náhodná veličina.

Gaussovský proces je podle [9] definovaný funkcí středních hodnot $m(\mathbf{x})$ a kovarianční funkcí $k(\mathbf{x},\mathbf{x}')$, které se definují jako

$$m(\mathbf{x}) = E[f(\mathbf{x})], \tag{2.3}$$

$$k(\mathbf{x},\mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], \tag{2.4}$$

a zápis gaussovského procesu má tvar

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x},\mathbf{x}')). \tag{2.5}$$

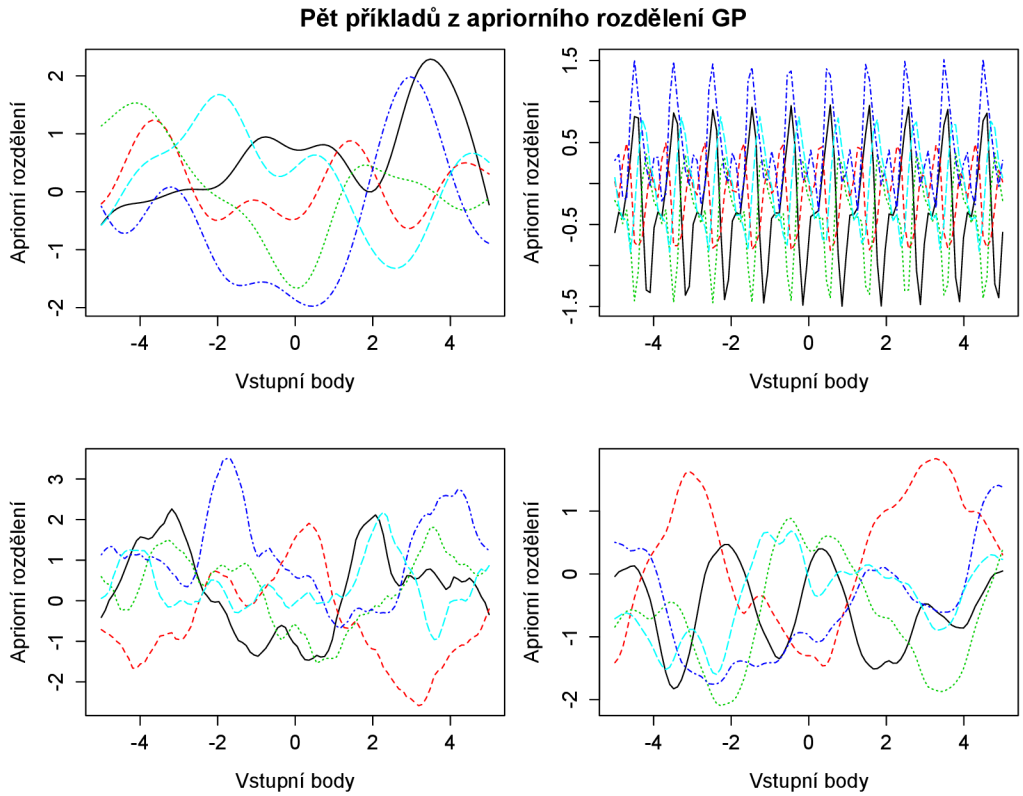
Zápis znamená, že funkce $f(\mathbf{x})$ je rozdělena stejně jako gaussovský proces s funkcí středních hodnot $m(\mathbf{x})$ a kovarianční funkcí $k(\mathbf{x},\mathbf{x}')$. V tomto případě hodnota funkce $f(\mathbf{x})$ představuje náhodnou veličinu v $\mathbf{x}$. Pro jednoduchost se uvažuje funkce středních hodnot $m(\mathbf{x}) = 0$, z čehož vychází i tato bakalářská práce.

Jakmile je určena funkce středních hodnot a kovarianční funkce, lze pomocí gaussovských procesů vykreslit hodnoty apriorního rozdělení. Rozdělení může být získáno v libovolných vstupních bodech X_*. Apriorní rozdělení představuje očekávané výstupní hodnoty f_* vstupů X_*, aniž jsou k dispozici data, a definuje se dle [9] jako

$$\mathbf{f}_* \sim \mathcal{N}(\mathbf{0}, K(X_*, X_*)), \tag{2.6}$$

kde $K(X_*,X_*)$ je kovarianční matice, ve které je na každý prvek aplikována vybraná kovarianční funkce $k(x_*,x_*)$.

Na obrázku 2.3 lze vidět několik apriorních rozdělení pro různé kovarianční funkce, které jsou popsány v následující podkapitole.



Obr. 2.3: Apriorní rozdělení gaussovských procesů různých kovariančních funkcí

Cílem gaussovských procesů většinou není pozorování apriorního rozdělení, ale získání nějaké informace o výsledné funkci na základě pozorovaných bodů. Tedy jde o získání aposteriorního rozdělení, což popisuje kapitola 3.

2.4 Kovarianční funkce

Kovarianční funkce též kernel je důležitým nástrojem při použití gaussovských procesů. Podle [9] a [10] je definovaná jako reálná funkce dvou argumentů, tj. $k(\mathbf{x},\mathbf{x}') \in \mathbb{R}$ pro $\mathbf{x},\mathbf{x}' \in \mathcal{X}$, kde $\mathcal{X}$ je nějaký abstraktní prostor. Musí být:

1. symetrická, tj. $k(\mathbf{x},\mathbf{x}') = k(\mathbf{x}',\mathbf{x})$,
2. nezáporná, tj. $k(\mathbf{x},\mathbf{x}') \geq 0$ a
3. pozitivně semidefinitní, tj. $\int k(\mathbf{x},\mathbf{x}')f(\mathbf{x})f(\mathbf{x}')\,\mathrm{d}\mu(\mathbf{x})\,\mathrm{d}\mu(\mathbf{x}') \geq 0 \quad \forall f \in L_2(\mathcal{X},\mu)$.

Z pohledu Gaussovských procesů definuje míru podobnosti vstupů $\mathbf{x}$ a $\mathbf{x}'$, to znamená, že pokud vstupní hodnoty leží blízko sebe, očekává se, že jejich výstupní hodnoty budou velmi podobné, ne-li stejné. Kovarianční funkce ovlivňuje chování hledané funkce f a určuje kovarianci mezi dvěma náhodnými veličinami. Čím blíž tyto veličiny jsou, tím větší je jejich kovariance. Každá kovarianční funkce má volné parametry, tzv. hyperparametry, které upřesňují její tvar.[9]

Existují různé kovarianční funkce, které jsou zmíněny v následující podkapitole.

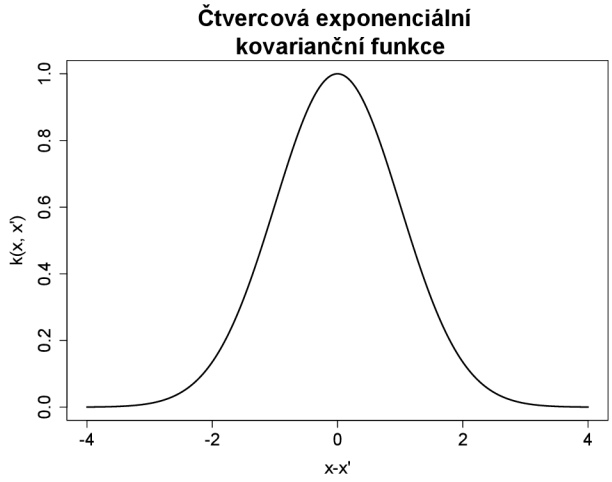
2.4.1 Nejčastěji používané kovarianční funkce

Následující kovarianční funkce, které jsou popsány v [11], mohou být rozděleny na stacionární a nestacionární. Do stacionárních patří čtvercová exponenciální, racionální kvadratická, periodická kovarianční funkce a třída Matérnových kovariančních funkcí. Hodnota těchto funkcí dle [9] záleží na rozdílu $\mathbf{x}-\mathbf{x}'$. Z níže zmíněných je nestacionární pouze lineární kovarianční funkce. V praktické části této práce se využívají pouze stacionární kovarianční funkce.

Čtvercová exponenciální (SE) kovarianční funkce
V gaussovských procesech je nejvyužívanější kovarianční funkcí, jelikož implikuje předpoklady jako hladkost a stacionaritu. Má tvar:

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp\left[\frac{-(\mathbf{x}-\mathbf{x}')^2}{2l^2}\right], \tag{2.7}$$

kde $\sigma_f^2 > 0$ je celkový rozptyl a $l > 0$ značí měřítko délky a udává, jak široká a hladká daná funkce je.

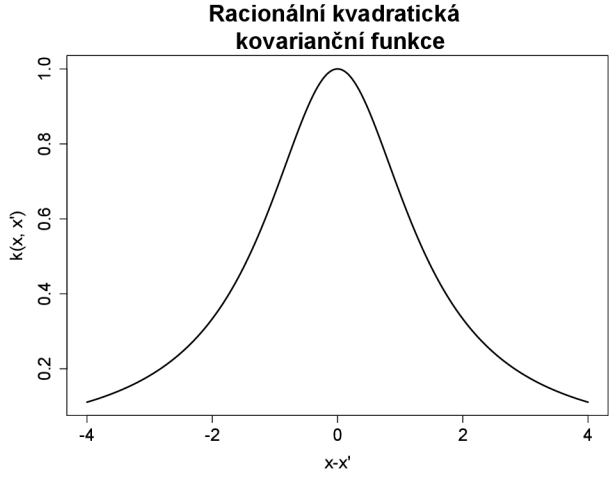


Obr. 2.4: Graf čtvercové exponenciální kovarianční funkce

Racionální kvadratická (RQ) kovarianční funkce

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \left[1 + \frac{(\mathbf{x}-\mathbf{x}')^2}{2\alpha l^2}\right]^{-\alpha}, \tag{2.8}$$

kde $\sigma_f^2 > 0$ je celkový rozptyl a parametry $\alpha, l > 0$ jsou v roli nekonečné sumy čtvercových exponenciálních kovariančních funkcí s různými měřítky délky l.

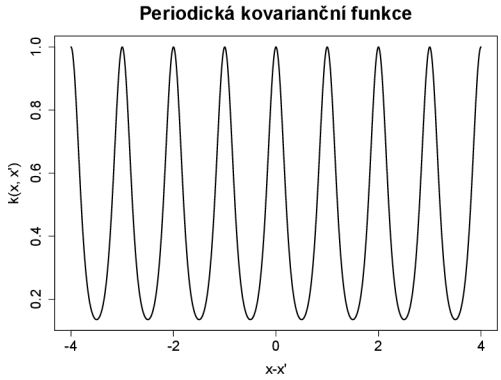


Obr. 2.5: Graf racionální kvadratické kovarianční funkce

Periodická kovarianční funkce

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp\left[-\frac{2}{l^2}\sin^2\left(\pi\frac{|\mathbf{x}-\mathbf{x}'|}{p}\right)\right], \tag{2.9}$$

kde $\sigma_f^2 > 0$ je celkový rozptyl, l značí měřítko délky a p je perioda. Tato kovarianční funkce je vhodná pro modelování funkcí, které jsou periodické.



Obr. 2.6: Graf periodické kovarianční funkce

Třída Matérnových kovariančních funkcí

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}(\mathbf{x}-\mathbf{x}')}{l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}(\mathbf{x}-\mathbf{x}')}{l} \right), \tag{2.10}$$

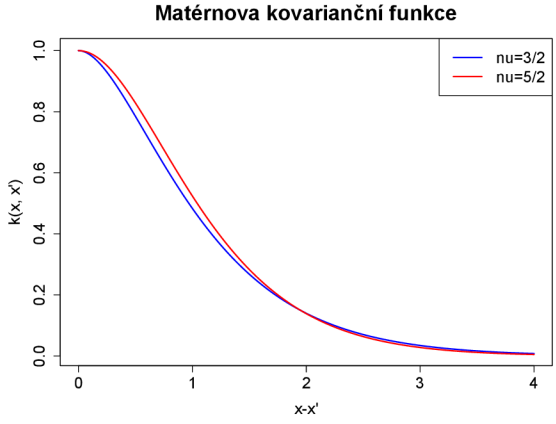
kde $\sigma_f^2 > 0$ je celkový rozptyl, $\nu > 0$ značí stupeň hladkosti, $l > 0$ měřítko délky a K_ν je modifikovaná Besselova funkce, což je řešení Besselovy diferenciální rovnice $z^2 \frac{\mathrm{d}^2 w(z)}{\mathrm{d}z^2} + + \frac{\mathrm{d}w(z)}{\mathrm{d}z} + (z^2 - \nu^2) w(z) = 0$.

Čím větší je ν, tím hladší je funkce. Když $\nu \to \infty$, kovarianční funkce konverguje k čtvercové exponenciální kovarianční funkci.

Nejčastější volby ν pro regresi jsou $\nu = 3/2$ a $\nu = 5/2$. Pak má Matérnova funkce následující tvary:

$$k_{\nu=3/2}(\mathbf{x},\mathbf{x}') = \sigma_f^2 \left(1 + \frac{\sqrt{3}(\mathbf{x}-\mathbf{x}')}{l} \right) \exp\left(-\frac{\sqrt{3}(\mathbf{x}-\mathbf{x}')}{l} \right), \tag{2.11}$$

$$k_{\nu=5/2}(\mathbf{x},\mathbf{x}') = \sigma_f^2 \left(1 + \frac{\sqrt{5}(\mathbf{x}-\mathbf{x}')}{l} + \frac{5(\mathbf{x}-\mathbf{x}')^2}{3l^2} \right) \exp\left(-\frac{\sqrt{5}(\mathbf{x}-\mathbf{x}')}{l} \right). \tag{2.12}$$

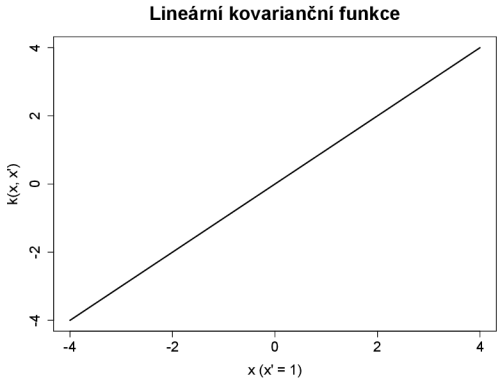


Obr. 2.7: Graf Matérnovy kovarianční funkce při $\nu = 3/2$ a $\nu = 5/2$

Lineární kovarianční funkce

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \mathbf{x}^T \mathbf{x}', \tag{2.13}$$

kde $\sigma_f^2 > 0$ je celkový rozptyl. Pokud je v gaussovských procesech použita tato kovarianční funkce, pak jde o lineární regresi, viz kapitola 1.



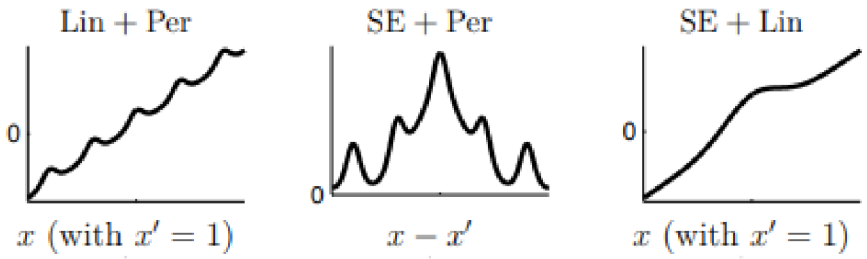
Obr. 2.8: Graf lineární kovarianční funkce

2.4.2 Tvorba nové kovarianční funkce

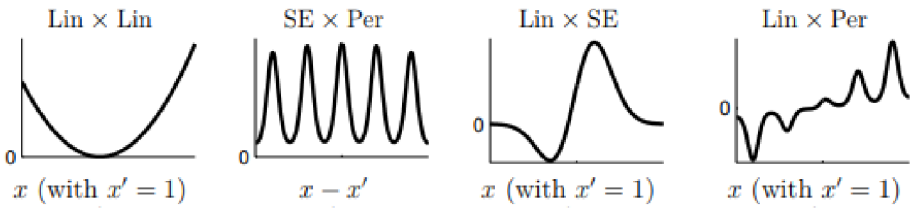
Při modelování nemusí být použity pouze funkce z části 2.4.1, podle [9] lze vytvořit novou kovarianční funkci, a to buď kombinací nebo modifikací již zmíněných kovariančních funkcí.

Věta 2.4. *Součet nebo násobení dvou kovariančních funkcí je opět kovarianční funkce, tj.* $k(\mathbf{x},\mathbf{x}') = k_1(\mathbf{x},\mathbf{x}') + k_2(\mathbf{x},\mathbf{x}')$ *a* $k(\mathbf{x},\mathbf{x}') = k_1(\mathbf{x},\mathbf{x}') \times k_2(\mathbf{x},\mathbf{x}')$.

Na obrázcích 2.9 a 2.10 lze vidět grafy několika nově vytvořených kovariančních funkcí.



Obr. 2.9: Ukázky grafů nově vzniklých kovariančních funkcí získaných jejich součtem. Značení: Lin - lineární kernel, Per - periodický kernel, SE - čtvercový exponenciální kernel (Zdroj:[11])



Obr. 2.10: Ukázky grafů nově vzniklých kovariančních funkcí získaných jejich násobením. Značení stejné jako u obr. 2.9 (Zdroj:[11])

Kapitola 3

Gaussovské procesy v regresi

Gaussovské procesy v regresi (GPR) jsou využívány k predikci spojitých náhodných veličin. Jedná se o proložení dat nějakou funkcí $f(x)$. Jde o neparametrický přístup k regresi, jehož cílem je nalezení rozdělení nad možnými funkcemi $f(x)$, které odpovídají pozorovaným datům. Neparametrický přístup znamená, že má nekonečně mnoho parametrů, resp. počet parametrů roste s počtem pozorovaných dat.

Aby byla regrese spolehlivá, musí být správně zvolena kovarianční funkce, tj. její hyperparametry. Důsledkem špatně zvolených hyperparametrů je nesmyslná regrese. Pro odhad těchto hyperparametrů se často používá metoda maximální věrohodnosti.

Hlavními výhodami gaussovských procesů je dobrá funkčnost na malých datových sadách a schopnost predikovat nejistá měření.

3.1 Predikce pozorování

Je dána množina vstupů x a výstupů y, tzv. trénovací množina $\mathcal{D} = \{(\mathbf{x}_i, y_i)|i = 1,\ldots,n\}$, která má n trénovacích (pozorovaných) bodů. Každý vstup $\mathbf{x}_i$ je d-dimenzionální vektor čísel. Každý výstup y_i je reálný skalár.

Cílem je predikovat funkce na základě pozorovaných dat. Uvedené formule v této podkapitole jsou čerpány z [9] a [10].

3.1.1 Predikce pozorování bez šumu

Je dána trénovací množina $\mathcal{D} = \{(\mathbf{x}_i, f_i)|i = 1,\ldots,n\}$, kde $f_i = f(\mathbf{x}_i)$. Předpokládá se, že data z trénovací množiny jsou bez šumu. To znamená, že predikované funkce budou procházet každým bodem z trénovací množiny $\mathcal{D}$. Dále je dána množina testovací vstupů X_* velikosti $n_* \times d$. Cílem je predikovat funkci výstupů f_* této testovací množiny. Toho se dosáhne pomocí aposteriorního rozdělení gaussovských procesů.

Sdružené rozdělení trénovacích (pozorovaných) hodnot f a testovacích (predikovaných) hodnot f_* je podle apriorní pravděpodobnosti dáno vztahem:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right), \tag{3.1}$$

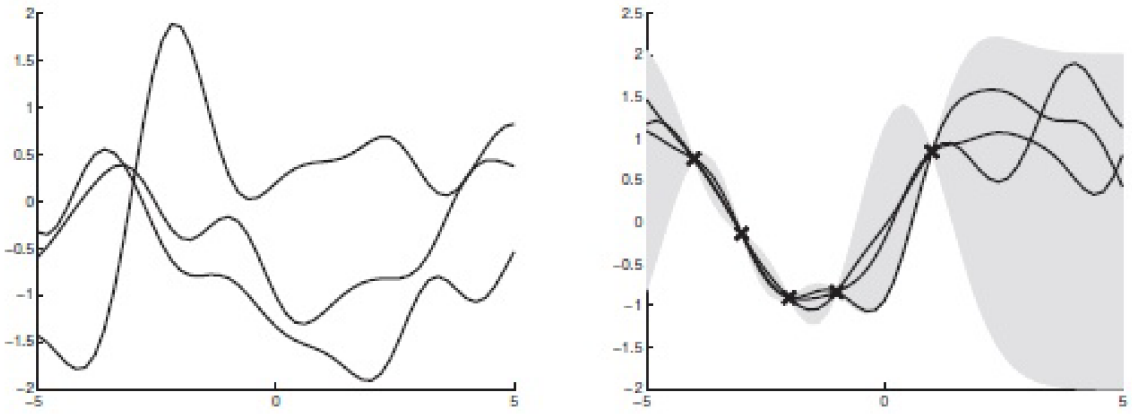
kde X je množina trénovacích vstupů délky n a kovarianční matice $K(X,X_*)$ velikosti $n \times n_*$ počítá kovarianci mezi všemi dvojicemi trénovacích a testovacích bodů. Podobně se dají spočítat i kovarianční matice $K(X,X)$, $K(X_*,X_*)$ a $K(X_*,X)$.

K získání aposteriorního rozdělení nad funkcemi stačí spočítat vektor středních hodnot a kovarianční matici pravidla podmíněnosti z věty 2.3. Výsledné vztahy jsou:

$$\mathbf{f}_*|X_*,X,\mathbf{f} \sim \mathcal{N}(\bar{\mathbf{f}}_*, cov(\mathbf{f}_*)), \text{ kde} \tag{3.2}$$

$$\bar{\mathbf{f}}_* = K(X_*,X)K(X,X)^{-1}\mathbf{f}, \tag{3.3}$$

$$cov(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*). \tag{3.4}$$



Obr. 3.1: Ukázka převodu apriorního rozdělení na aposteriorní rozdělení po zpozorování dat (Zdroj:[10])

3.1.2 Predikce pozorování se šumem

Pro modelovaní reálných situací je typické, že pozorování mají určitý šum. Není podmínkou, aby predikované funkce procházely pozorovanými daty, ale měly by být blízko. Je tedy dána trénovací množina $\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1,\ldots,n\}$, kde pro funkční hodnoty platí: $y = f(\mathbf{x}) + \varepsilon$, kde $\varepsilon \sim \mathcal{N}(0,\sigma_n^2)$. Potom apriorní rozdělení trénovacích výstupů se šumem má tvar:

$$cov(y_p,y_q) = k(\mathbf{x}_p,\mathbf{x}_q) + \sigma_n^2\delta_{pq} \text{ nebo } cov(\mathbf{y}) = K(X,X) + \sigma_n^2 I, \tag{3.5}$$

kde δ_{pq} je Kroneckerovo delta, pro kterou platí $\delta_{pq} = \begin{cases} 1 & \text{pro } p = q \\ 0 & \text{jinak} \end{cases}$.

Sdružené rozdělení pozorovaných hodnot y a predikovaných hodnot f_* po zavedení šumu σ_n^2 má následující tvar:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(X,X) + \sigma_n^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right). \tag{3.6}$$

Pak pro aposteriorní rozdělení platí vzorce:

$$\mathbf{f}_*|X_*,\mathbf{y},X \sim \mathcal{N}(\bar{\mathbf{f}}_*, cov(\mathbf{f}_*)), \text{ kde} \tag{3.7}$$

$$\bar{\mathbf{f}}_* = K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}\mathbf{y}, \tag{3.8}$$

$$cov(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)[K(X,X) + \sigma_n^2 I]^{-1}K(X,X_*). \tag{3.9}$$

3.2 Vliv a odhad hyperparametrů

Z podkapitoly 2.4 vychází, že kovarianční funkce má několik hyperparametrů, které upřesňují její tvar a chování. Tato podkapitola ukazuje vliv hyperparametrů na predikci při jejich změně a následně je uveden postup, jak získat jejich optimální hodnoty.

3.2.1 Vliv hyperparametrů

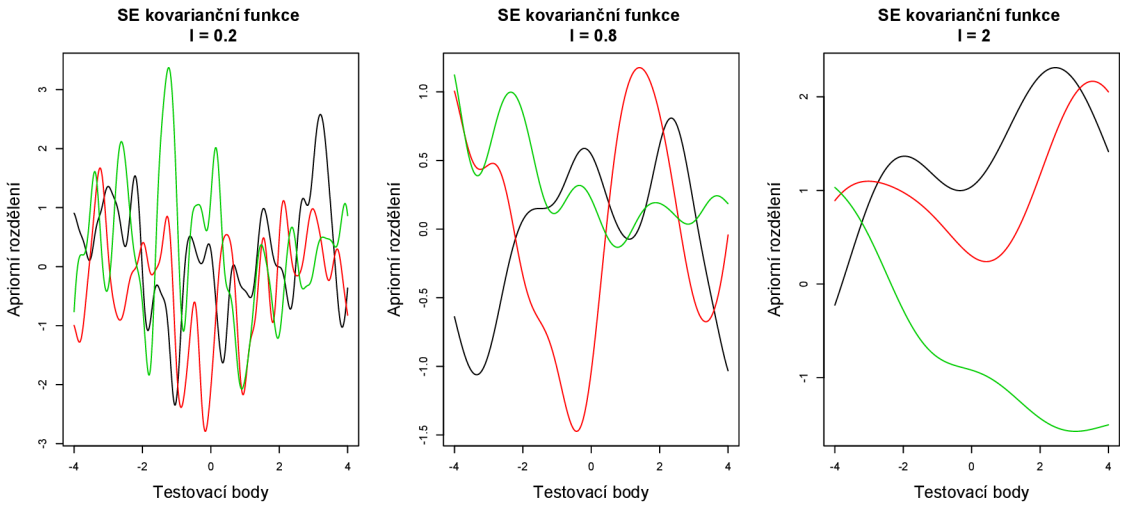
Podkapitola čerpá z [9], [10], [12] a [13].

Je uvažována čtvercová exponenciální (SE) kovarianční funkce, která je dána vzorcem:

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp\left[\frac{-(\mathbf{x}_p - \mathbf{x}_q)^2}{2l^2}\right] + \sigma_n^2 \delta_{pq}, \tag{3.10}$$

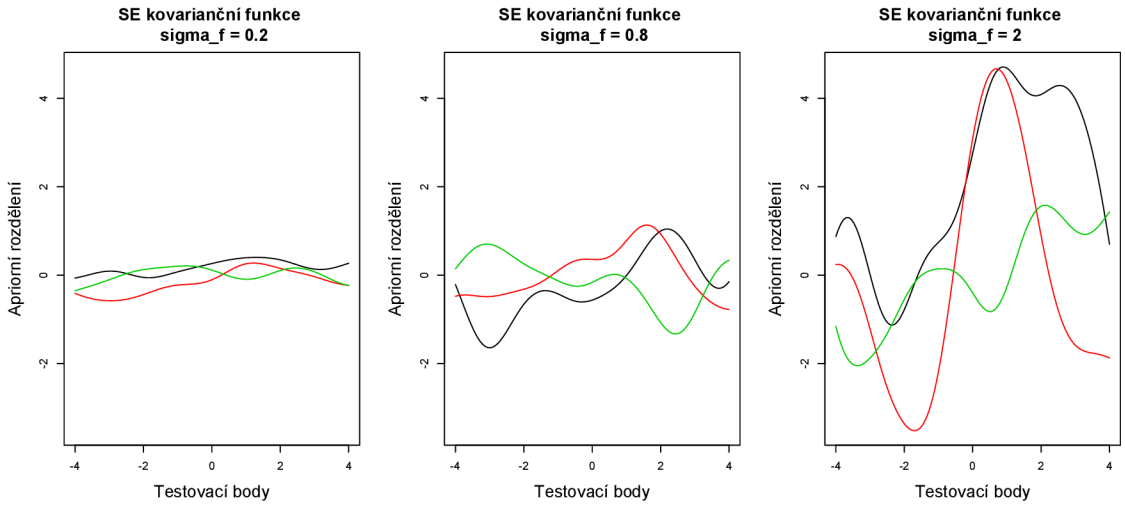
kde hyperparametr l značí měřítko délky, σ_f^2 celkový rozptyl a σ_n^2 šum.

Nejprve se ukáže vliv hyperparametrů na apriorní rozdělení. Ve všech případech bude $\sigma_n = 0$. Hyperparametr $\sigma_f = 1$ je zafixován a l se mění. Na obrázku 3.2 lze pozorovat, že čím vyšší je hodnota l, tím hladší je apriorní rozdělení.



Obr. 3.2: Apriorní rozdělení za použití fixních hyperparametrů $(\sigma_f, \sigma_n) = (1, 0)$ a měnícího se hyperparametru l

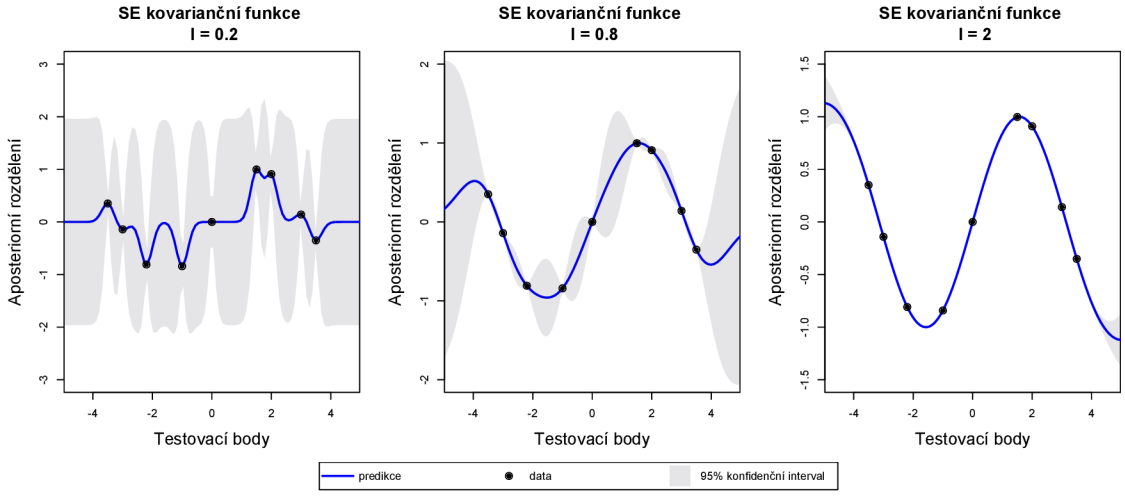
Nyní je zafixován hyperparametr $l = 1$ a σ_f se bude lišit. Tento hyperparametr, jak je vidět na obrázku 3.3, mění pouze vertikální rozptyl. Se zvětšujícím se σ_f, se zvyšují hodnoty na ose y.



Obr. 3.3: Apriorní rozdělení za použití fixních hyperparametrů $(l, \sigma_n) = (1, 0)$ a měnícího se hyperparametru σ_f

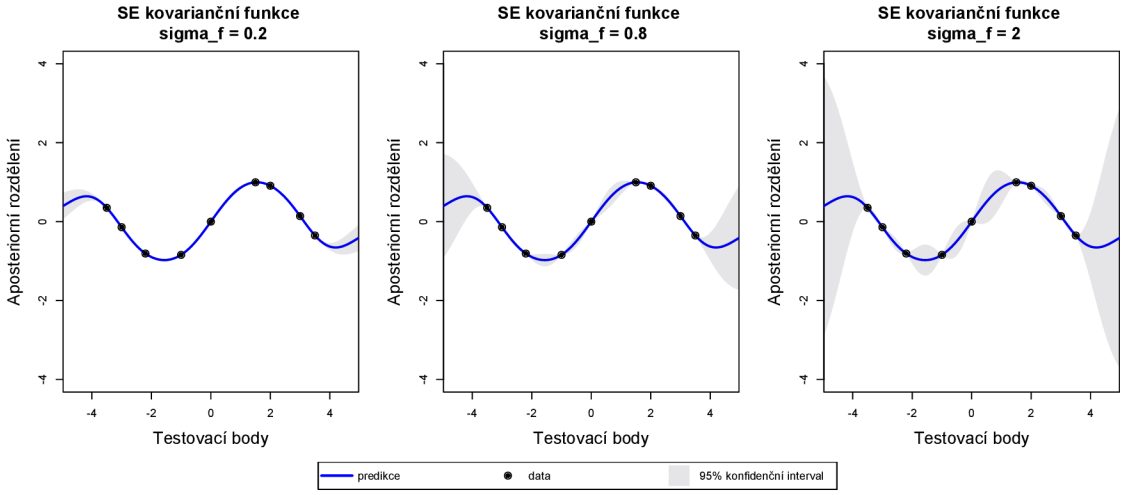
Stejným postupem jako u apriorního rozdělení je znázorněn vliv hyperparametrů na aposteriorní rozdělení, dva hyperparametry jsou zafixovány a jeden se mění. Černé body na obrázcích představují pozorovaná data.

Zafixují se hyperparametry $\sigma_f = 1, \sigma_n = 0$ a l se bude měnit. Stejně jako u apriorního rozdělení je na obrázku 3.4 vidět, že čím vyšší je hodnota l, tím hladší je predikce, tedy hodnoty predikované funkce se mění velmi pomalu a nevyskytují se zde téměř žádné oblasti nejistoty (šedé plochy). Naopak, čím menší je hodnota l, tím je predikce „vlnitější“, tím rychleji se mění hodnoty predikované funkce a mezi trénovacími body jsou větší oblasti nejistoty.



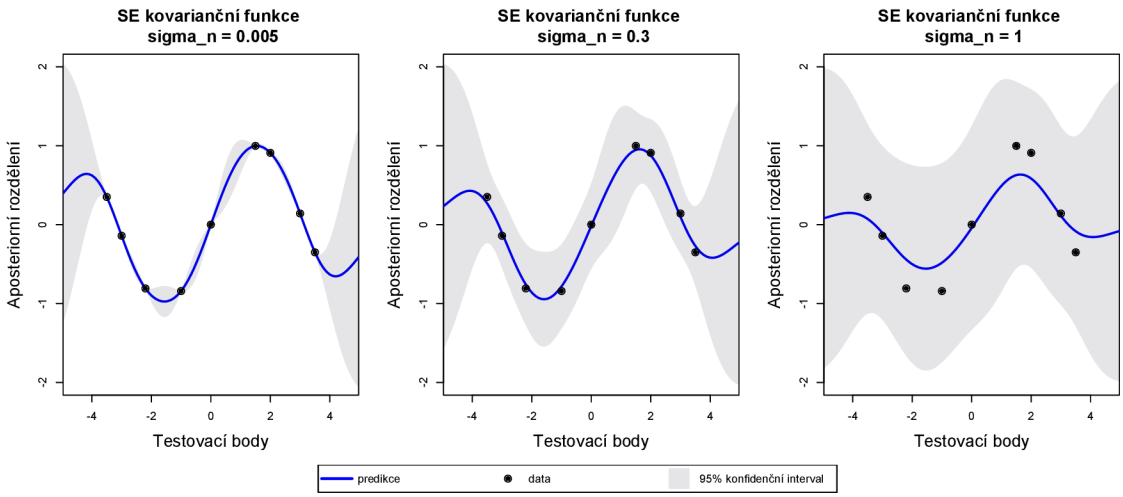
Obr. 3.4: Aposteriorní rozdělení za použití fixních hyperparametrů $(\sigma_f, \sigma_n) = (1, 0)$ a měnícího se hyperparametru l

V druhém případě na obrázku 3.5 jsou zafixovány hyperparametry $l = 1, \sigma_n = 0$ a σ_f se bude měnit. Opět jako u apriorního rozdělení má tento hyperparametr vliv zejména na vertikální rozptyl, přesněji udává odchylku od predikce. Malá hodnota σ_f znamená, že funkce se nachází blízko predikci. Velká hodnota σ_f zvětšuje odchylku, tedy oblasti nejistoty.



Obr. 3.5: Aposteriorní rozdělení za použití fixních hyperparametrů $(l, \sigma_n) = (1, 0)$ a měnícího se hyperparametru σ_f

Obrázek 3.6 ukazuje, jak je ovlivněno aposteriorní rozdělení při zafixovaných hyperparametrech $l = 1, \sigma_f = 1$ a měnícím se σ_n. Šum σ_n^2 není formálně součástí kovarianční funkce, ale je také považován za hyperparametr, pomocí kterého je specifikováno, jak velký šum je v datech očekáván. Při velmi malém šumu prochází predikce trénovacími body. Naopak při velkém šumu je predikce méně přesná, jelikož neprochází všemi daty.



Obr. 3.6: Aposteriorní rozdělení za použití fixních hyperparametrů $(l, \sigma_f) = (1, 1)$ a měnícího se hyperparametru σ_n

3.2.2 Odhad hyperparametrů

Optimální hodnoty hyperparametrů lze dle [14] a [15] získat metodou maximální věrohodnosti (MLE). Ta je založena na výpočtu věrohodnostní funkce $L(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})$, kde $\mathbf{y}$ je vektor výstupů trénovacích bodů a $\boldsymbol{\theta}$ (např. $\boldsymbol{\theta} = (\sigma_f, l, \sigma_n)$) je vektor hyperparametrů, tj.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}). \tag{3.11}$$

Jelikož platí $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, K(X,X) + \sigma_n^2 I)$, věrohodnostní funkce je rovna hustotě vícerozměrného normálního rozdělení (viz (2.2)), tj.

$$L(\boldsymbol{\theta}) = (2\pi)^{-\frac{n}{2}} |K(X,X) + \sigma_n^2 I|^{-\frac{1}{2}} e^{-\frac{1}{2}\mathbf{y}^T (K(X,X)+\sigma_n^2 I^{-1})\mathbf{y}}. \tag{3.12}$$

V praxi je výhodnější používat logaritmickou věrohodnostní funkci $\ell(\boldsymbol{\theta})$, která je ve tvaru

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|K(X,X) + \sigma_n^2 I| - \frac{1}{2}\mathbf{y}^T (K(X,X) + \sigma_n^2 I)^{-1}\mathbf{y}. \tag{3.13}$$

První část vzorce 3.13 je normalizační konstanta, druhá část značí složitost modelu a poslední část určuje datový fit, tedy jak dobře současná parametrizace vysvětluje závislou proměnnou.

Optimální vektor hyperparametrů $\hat{\boldsymbol{\theta}}$ je získán minimalizováním negativní logaritmické věrohodnostní funkce

$$-\log L(\boldsymbol{\theta}) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log|K(X,X) + \sigma_n^2 I| + \frac{1}{2}\mathbf{y}^T (K(X,X) + \sigma_n^2 I)^{-1}\mathbf{y}, \text{ tedy} \tag{3.14}$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}(-\log L(\boldsymbol{\theta})) \tag{3.15}$$

Na závěr lze použít jakýkoli algoritmus pro optimalizaci více proměnných a jsou vypočítány hodnoty odhadovaných hyperparametrů $\boldsymbol{\theta}$.

Kapitola 4

Užití Gaussovských procesů v regresi na datech

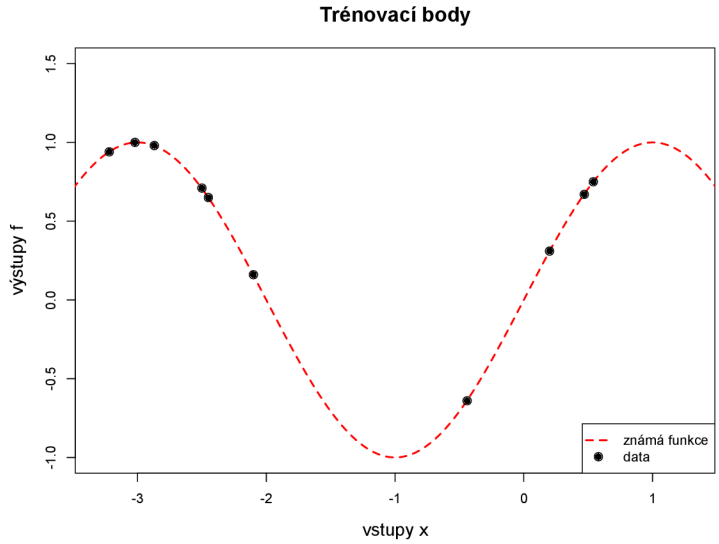
Kapitola představuje užití gaussovských procesů v regresi na simulovaných a reálných datech. Všechny grafy a výpočty jsou získány za použití softwaru `R`.

4.1 Simulovaná data

Na simulovaných datech lze ukázat ideální chování gaussovských procesů v regresi. V části 4.1.1 je porovnána predikce pomocí gaussovských procesů se známou zvolenou funkcí. V 4.1.2 je provedena predikce většího množství náhodně vygenerovaných dat, která je srovnána s predikcí pomocí balíčku **laGP** softwaru `R`. V části 4.1.3 je chování gaussovských procesů porovnáno s lineárním modelem. Následující postupy vychází z [9], [12], [16] a [17].

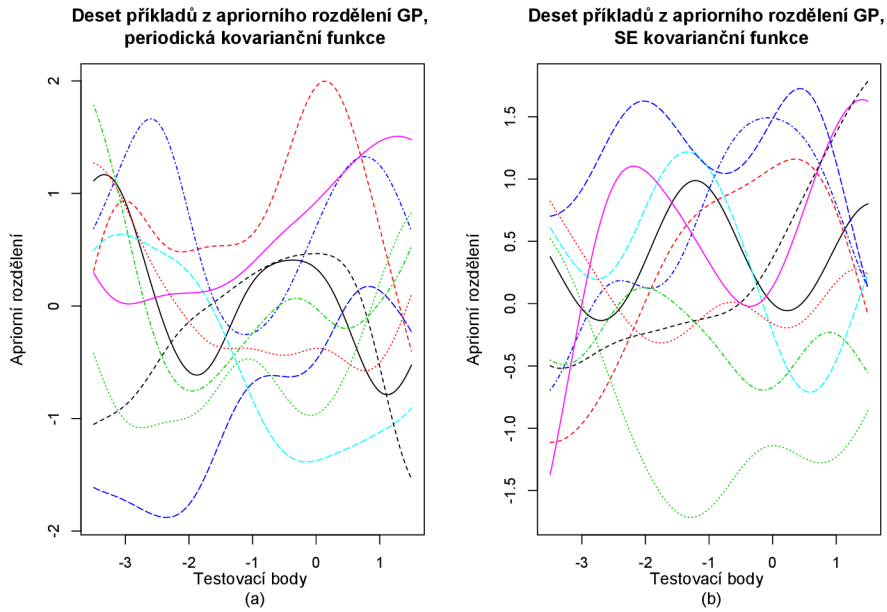
4.1.1 Srovnání predikce pomocí GP se známou funkcí

Na základě poznatků z podkapitoly 3.1.1 je provedeno srovnání predikce se známou zvolenou funkcí bez šumu. Pro predikci je potřeba nejprve zvolit známou funkci, trénovací body a testovací body, ve kterých bude uskutečněna predikce. Jako známá funkce je zvolena $sin(0,5\pi x)$. Náhodně vygenerované hodnoty z intervalu $(-3,5;1)$ slouží jako vstupy x_i. Pomocí známé funkce se získají výstupy trénovacích bodů f_i. Je predikováno 200 stejně vzdálených testovacích vstupů X_* z intervalu $(-3,5;1,5)$. Na obrázku 4.1 lze vidět zvolené trénovací body a známou funkci.



Obr. 4.1: Trénovací body společně se známou funkcí

Pro použití gaussovských procesů je třeba zvolit kovarianční funkci. Známá funkce $sin(0,5\pi x)$ je hladká a periodická. Nejpodobnější kovarianční funkce je periodická nebo čtvercová exponenciální (SE), což je vidět na apriorním rozdělení na obrázku 4.2.

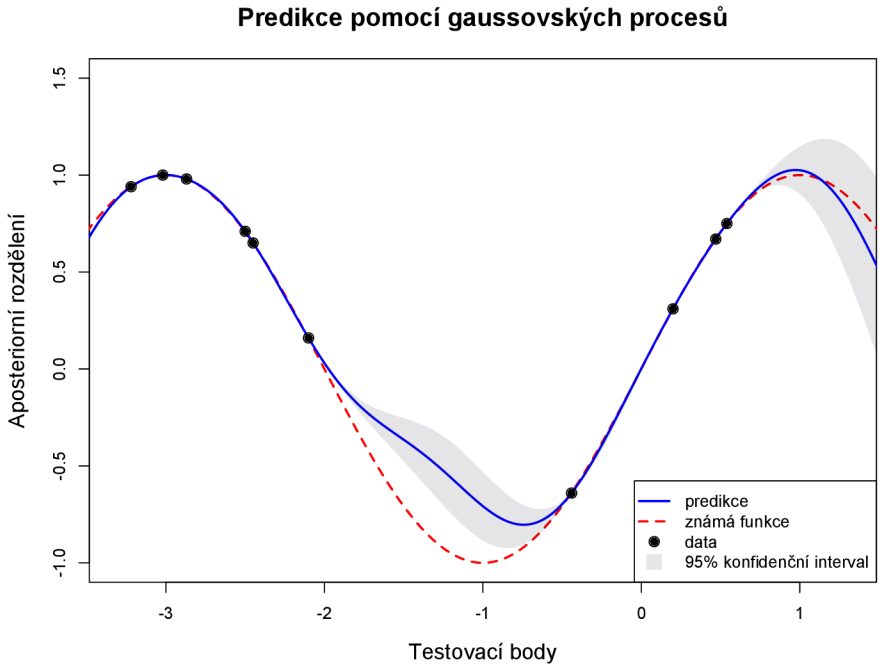


Obr. 4.2: Apriorní rozdělení s hyperparametry (a) $(l, \sigma_f, p) = (1, 1, 2\pi)$, (b) $(l, \sigma_f) = (1, 1)$

Periodická kovarianční funkce má hyperparametry $(l, \sigma_f, p) = (1, 1, 2\pi)$, u SE jsou $(l, \sigma_f) = (1, 1)$. Jelikož černá křivka v grafu 4.2 (b) připomíná sinusoidu, i SE kovarianční funkce by mohla být vhodná. Známá funkce $sin(0,5\pi x)$ je periodická, a proto je nejprve predikováno pomocí periodické kovarianční funkce.

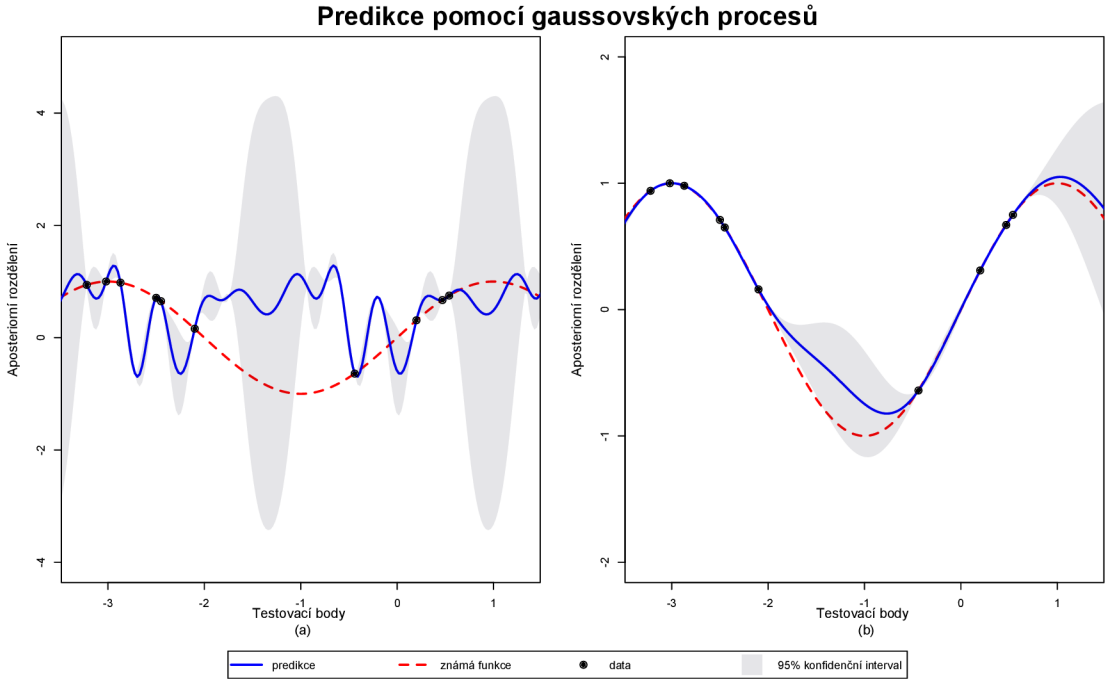
Periodická kovarianční funkce

Nejdříve jsou zvoleny hyperparametry $(l, \sigma_f, p) = (1, 1, 2\pi)$. Pro výpočet aposteriorního rozdělení je spočítána střední hodnota a kovarianční matice pomocí vzorců (3.3) a (3.4). Na obrázku 4.3 lze sledovat, že predikce je podobná známé funkci. Prochází trénovacími daty, jelikož data jsou bez šumu. Největší odchylka od známé funkce se nachází v místech s malým množstvím dat.



Obr. 4.3: Aposteriorní rozdělení za užití periodické kovarianční funkce a hyperparametrů $(l, \sigma_f, p) = (1, 1, 2\pi)$

Další predikce je s optimalizovanými hyperparametry. V softwaru `R` je použita funkce `optim`, které je na vstupu zadán vektor hodnot počátečních hyperparametrů `par`, přes které bude provedena optimalizace, a funkce `fn`, která má být minimalizována. Jsou vytvořeny dvě funkce `optim` pro různé vektory `par`. Nejprve pro `par`$= c(1, 1, \pi)$ (obrázek 4.4 (a)), poté pro `par`$= c(1, 1, 2\pi)$ (obrázek 4.4 (b)). V obou případech je minimalizována negativní logaritmická věrohodnostní funkce (3.14). Následně je postup stejný jako u předchozího výpočtu aposteriorního rozdělení. Je vypočítána střední hodnota a kovarianční matice. Na obrázku 4.4 (a) je graf s hyperparametry $(l, \sigma_f, p) = (0,5065; 2,0223; 2,2792)$, graf (b) má hyperparametry $(l, \sigma_f, p) = (0,6360; 0,6693; 6,9962)$.



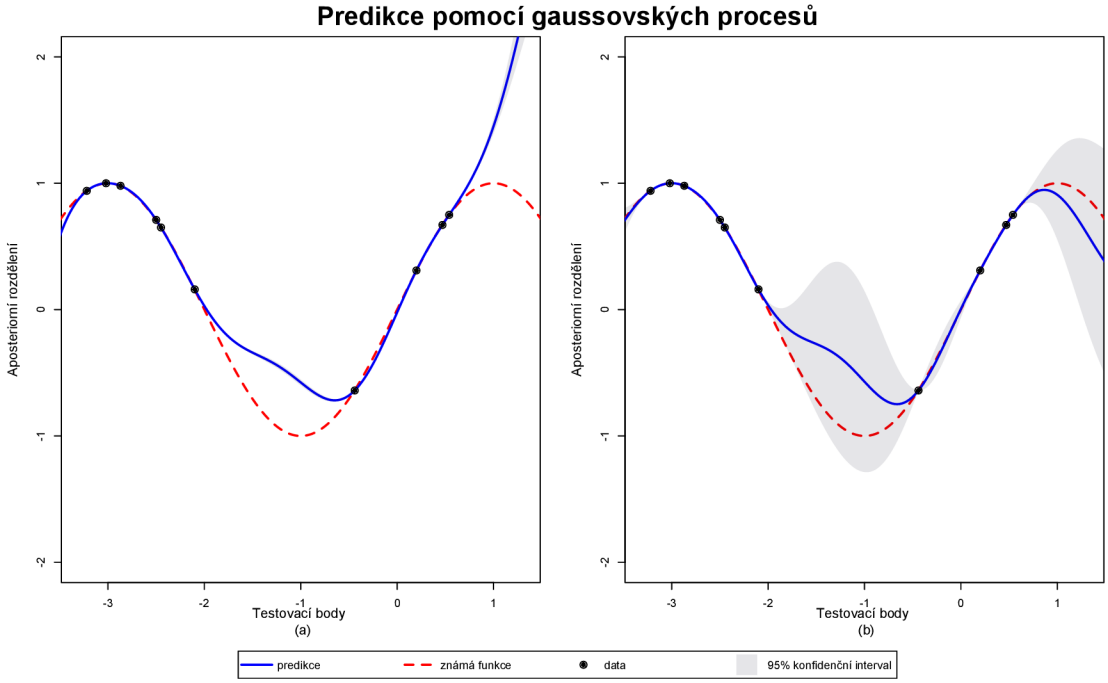
Obr. 4.4: Aposteriorní rozdělení za užití periodické kovarianční funkce a optimalizovaných hyperparametrů (a) $(l, \sigma_f, p) = (0,5065; 2,0223; 2,2792)$, (b) $(l, \sigma_f, p) = (0,6360; 0,6693; 6,9962)$

Z obrázku 4.4 (a) vyplývá, že predikce je špatná. Vznikají velké oblasti nejistoty, predikce je „vlnitá“ a vzhledem vzdálená od známé funkce. V grafu 4.4 (b) je predikce téměř totožná se známou funkcí. Hyperparametr periody je velmi blízko skutečné periodě 6. Oblast nejistoty je větší pouze v oblasti bez bodů.

I přes dobrý výsledek v 4.4 (b) je v další části popsán postup s použitím SE kovarianční funkce.

SE kovarianční funkce

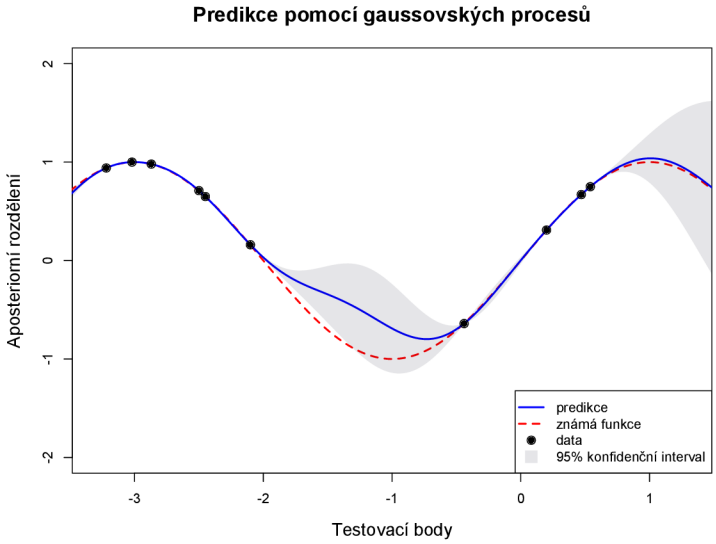
Postup je stejný jako u periodické kovarianční funkce. Nejprve jsou zvoleny hyperparametry $(l, \sigma_f) = (1,1)$ (obrázek 4.5 (a)), následně jsou použity menší hyperparametry $(l, \sigma_f) = (0,5; 0,5)$ (obrázek 4.5 (b)).



Obr. 4.5: Aposteriorní rozdělení za užití SE kovarianční funkce a zvolených hyperparametrů (a) $(l, \sigma_f) = (1,1)$, (b) $(l, \sigma_f) = (0,5;0,5)$

Funkce na obrázku 4.5 (a) neobsahuje žádné oblasti nejistoty. Zároveň predikovaná funkce od hodnoty $0,5$ na ose x roste až do nekonečna a liší se tak od známé funkce. Z těchto poznatků je zřejmé, že hodnoty hyperparametrů jsou příliš vysoké. Predikce na grafu 4.5 (b) je velmi podobná predikci v grafu 4.4 (b). Jde o první predikci, která má v oblastech nejistoty kompletně zachycenou známou funkci. Její hyperparametry by mohly být podobné těm optimálním.

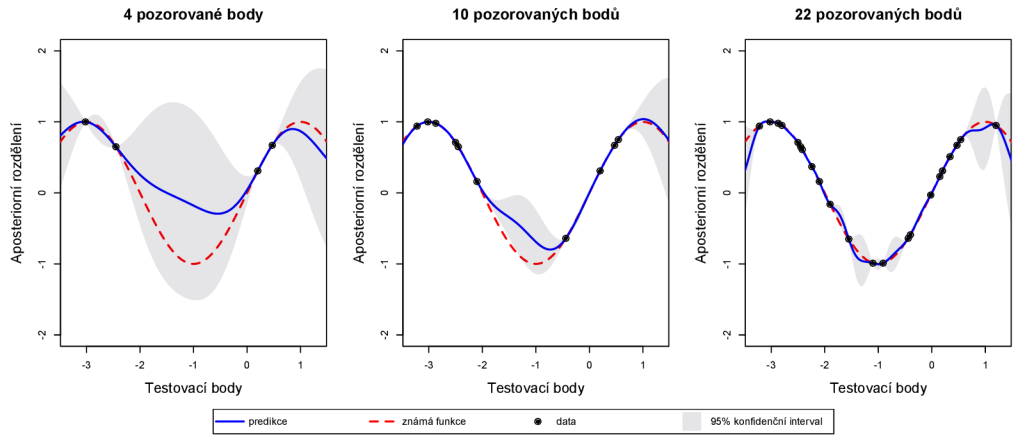
Na závěr je vytvořena predikce s optimalizovanými hyperparametry, kde ve funkci `optim` je zvolen `par`$= c(1,1)$ a ve `fn` je zadána negativní logaritmická věrohodnostní funkce (3.14). Optimální hyperparametry jsou $(l, \sigma_f) = (0,6267;0,6333)$. Na obrázku 4.6 lze vidět, že oblasti nejistoty jsou menší a predikce se blíží známé funkci.



Obr. 4.6: Aposteriorní rozdělení za užití SE kovarianční funkce a optimalizovaných hyperparametrů $(l, \sigma_f) = (0,6267; 0,6333)$

Nejlepších výsledků bylo dosaženo s použitím optimalizovaných hyperparametrů. U periodické funkce s vektorem počátečních hodnot $\texttt{par} = c(1,1,2\pi)$, u SE kovarianční funkce s $\texttt{par} = c(1,1)$. Vliv `par` na predikci s SE kovarianční funkcí je výrazně nižší než u predikce s periodickou funkcí. Lepší predikce lze tedy dosáhnout pomocí SE kovarianční funkce, jelikož je o jeden hyperparametr méně náročnější na optimalizaci.

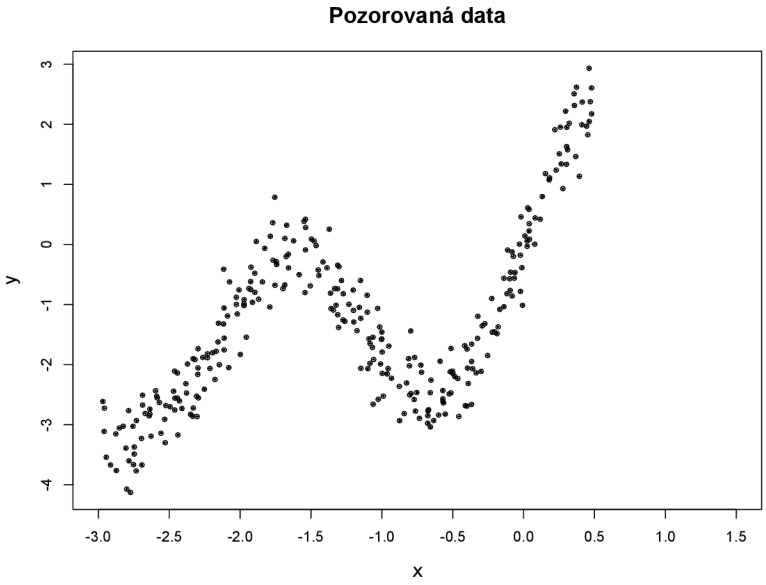
Závěrem je důležité zdůraznit, že gaussovské procesy modelují nejistotu v závislosti na pozorovaných datech. To znamená, že čím více trénovacích bodů je k dispozici, tím menší jsou oblasti nejistoty. Obrázek 4.7 popisuje situaci použití 4, 10 a 22 trénovacích bodů. V trénovacích bodech bez šumu je nejistota vždy nulová, mimo pozorované body jsou velikosti oblastí nejistoty závislé na počtu trénovacích bodů. I pro predikovanou funkci platí, že čím více bodů je známo, tím lepší a přesnější je.



Obr. 4.7: Aposteriorní rozdělení za užití SE kovarianční funkce při různém počtu trénovacích bodů

4.1.2 Srovnání predikce dat s predikcí pomocí balíčku z R

Pro zpracování jsou v podkapitole použita data zobrazená na obrázku 4.8. Predikce má být provedena v 500 stejně vzdálených testovacích vstupech X_* z intervalu $(-3;1,5)$. Na základě struktury dat lze soudit, že predikovaná funkce bude hladká. Použití SE kovarianční funkce je tedy vhodné.



Obr. 4.8: Pozorovaná data

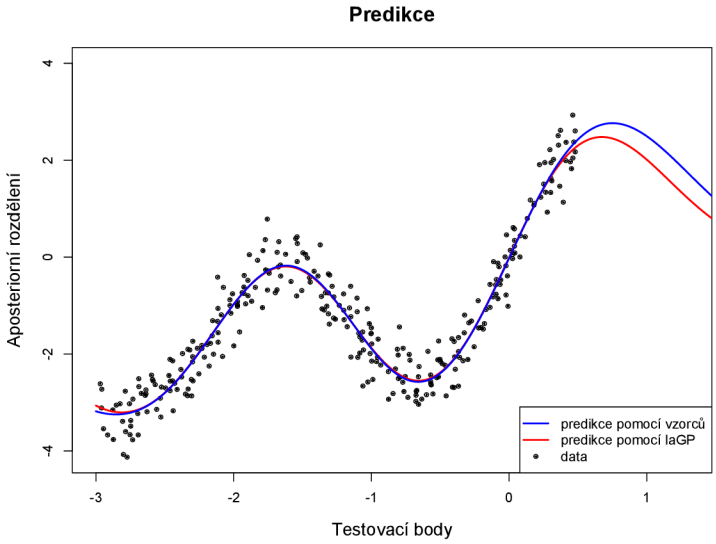
Základem zpracování dat je určení optimálních hodnot pro hyperparametry l a σ_f. Data obsahují šum s neznámou hodnotou, takže i hyperparametr σ_n^2 je třeba optimalizovat. K tomu je stejně jako v podkapitole 4.1.1 použita funkce `optim`, které je na vstupu zadán `par`$= c(1,1,0.5)$ odpovídající (l,σ_f,σ_n), a jako `fn` je použita negativní logaritmická věrohodnostní funkce (3.14). Optimalizací jsou získány hyperparametry $(l,\sigma_f,\sigma_n) = (0,7802;2,4146;0,3946)$. Následně jsou pomocí vzorce (3.8) vypočítány predikované hodnoty testovacích výstupů.

Získaná predikce je srovnána s predikcí stejných dat pomocí balíčku z R. K tomu je použitý volně dostupný balíček **laGP**, který je podrobněji popsán v [18] a slouží k predikci pomocí gaussovských procesů. Balíček je vhodný pro zpracování větších datových sad a je výpočetně velmi rychlý. Proto je zvolen i pro účely této podkapitoly. Dalšími příklady balíčků mohou být například **GauPro**, **GPfit**, **kernlab** a další.

Užití balíčku laGP v R

Nejprve je třeba vytvořit nový objekt gaussovského procesu `gpi` pomocí funkce `newGP`. Vstupem této funkce je matice nebo datová tabulka trénovacích vstupů `X`, vektor odpovídajících výstupů `Z` a hodnoty hyperparametrů `d` a `g`. Parametr `d` značí hodnotu měřítka délky a `g` je šum. K získání predikovaných hodnot je třeba použít funkci `predGP`. Té je na vstupu zadán předchozí funkcí vytvořený objekt `gpi` a matice nebo datová tabulka vstupů `XX`, ve kterých má být provedena predikce.

Získané predikce pomocí obou způsobů jsou zobrazeny na obrázku 4.9.



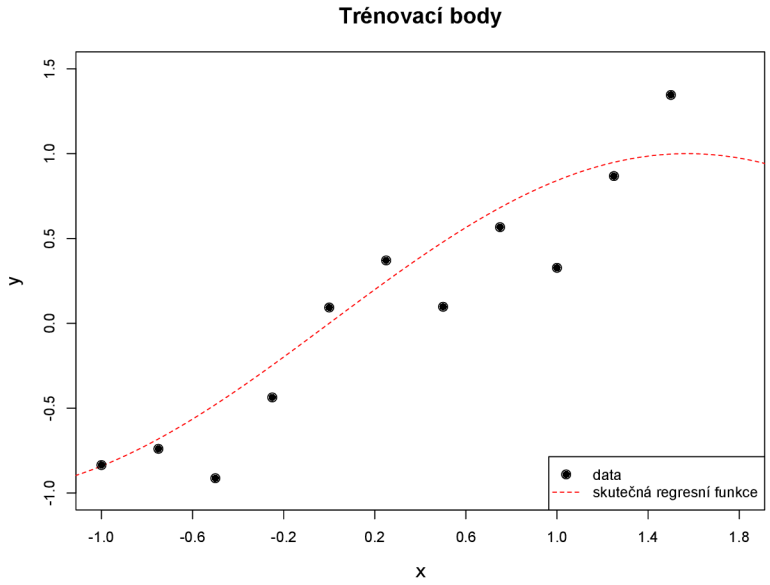
Obr. 4.9: Srovnání predikce pozorovaných dat s predikcí pomocí balíčku **laGP**

Predikce pomocí vzorců a optimalizovaných hyperparametrů je srovnatelná s predikcí pomocí balíčku **laGP**.

4.1.3 Srovnání GPR s lineárním modelem

Chování gaussovských procesů v regresi se dá porovnat s mnoha dalšími statistickými metodami. V této podkapitole je srovnání s lineárním modelem.

Je zvoleno 11 náhodných trénovacích bodů (x_i, y_i), $i = 1, \ldots, 11$, regresní funkce $sin(x)$, u kterých je známo, že mají šum $\sigma_n^2 = 0,1$, viz obrázek 4.10. Data jsou v `R` zapsána do datové tabulky s názvem `data`.



Obr. 4.10: Trénovací body

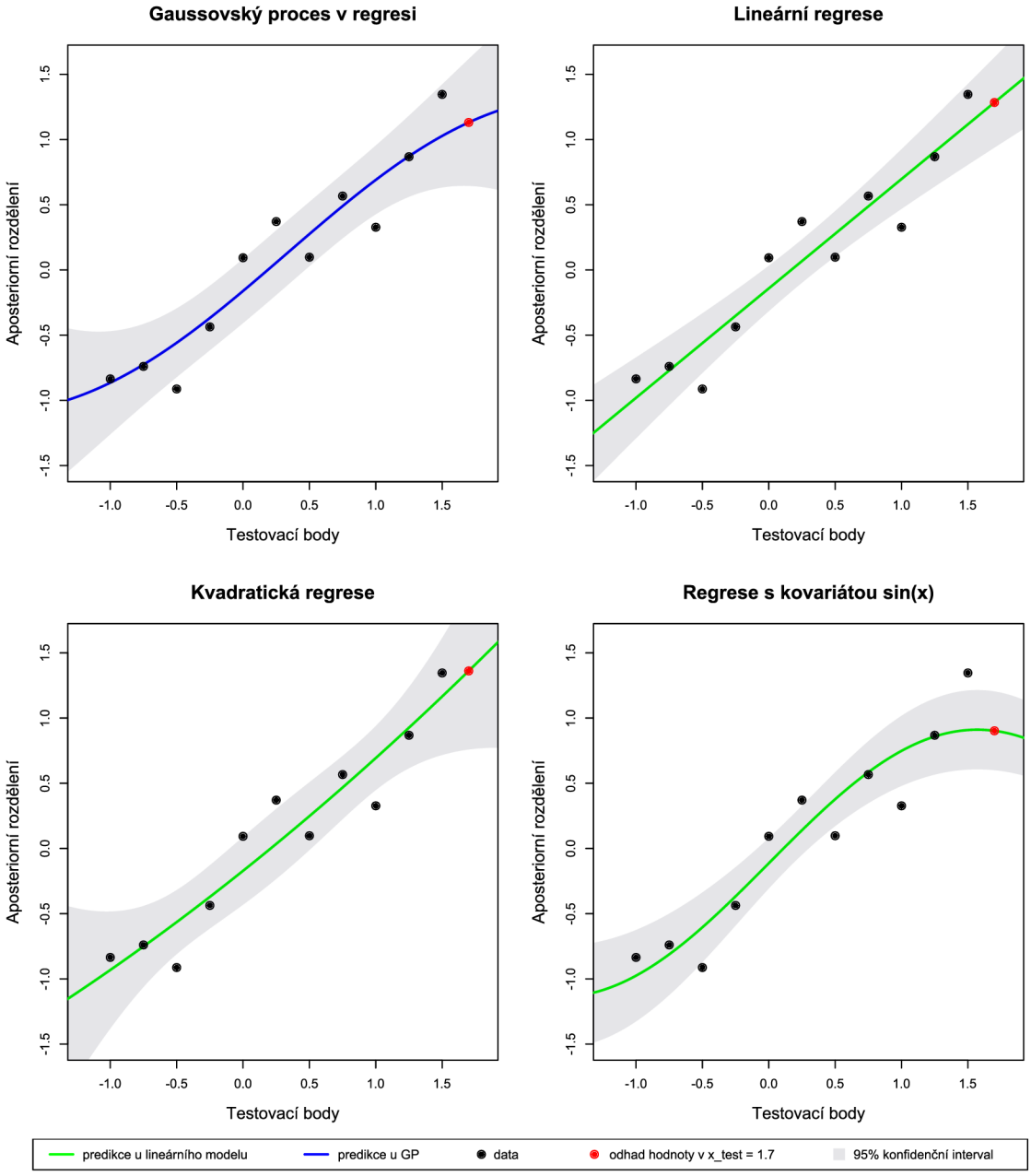
Nejprve je odhadnuta funkční hodnota v jednom testovacím bodě $x_* = 1,7$. Následně je predikováno 200 stejně vzdálených vstupů X_* z intervalu $(-1,2; 1,8)$. Predikce je provedena více způsoby - pomocí gaussovských procesů a pomocí různých lineárních modelů.

Predikce pomocí gaussovských procesů je provedena prostřednictvím vzorců z části 3.1.2, jelikož data obsahují šum. Použije se SE kovarianční funkce, pro kterou jsou použity optimalizované hyperparametry $(l, \sigma_f) = (2,1315; 1,2146)$. Užití SE kovarianční funkce implikuje hladkost predikované funkce.

Predikce lineárního modelu je popsána v kapitole 1. V softwaru `R` je k odhadu využita funkce $\texttt{lm}(\texttt{y} \sim \texttt{x},\ \texttt{data} = \texttt{data})$, což odpovídá rovnici ve tvaru $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, 11$. Regresní koeficienty jsou $(\beta_0, \beta_1) = (-0,1422; 0,8389)$.

Pro kvadratický model je tvar funkce $\texttt{lm}(\texttt{y} \sim \texttt{x} + \texttt{I}(\texttt{x}^2),\ \texttt{data} = \texttt{data})$, čemuž odpovídá rovnice $Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$, $i = 1, \ldots, 11$. Vypočítané regresní koeficienty jsou $(\beta_0, \beta_1, \beta_2) = (-0,1716; 0,8127; 0,0524)$.

Jelikož u predikce pomocí gaussovských procesů je použita SE kovarianční funkce, která je hladká, je vytvořen i odhad pomocí lineárního modelu, který má v matici plánu $\mathbf{X}$ kovariátu $sin(x)$. Rovnici $Y_i = \beta_0 + \beta_1 sin(x_i) + \varepsilon_i$, $i = 1, \ldots, 11$ je možné v `R` zapsat jako $\texttt{lm}(\texttt{y} \sim \texttt{sin}(\texttt{x}),\ \texttt{data} = \texttt{data})$. Výsledné regresní koeficienty jsou $(\beta_0, \beta_1) = (-0,1137; 1,0244)$. Výsledky těchto predikcí jsou na obrázku 4.11.



Obr. 4.11: Porovnání predikce pomocí gaussovských procesů s různými lineárními modely na náhodně vygenerovaných bodech

Z obrázku vyplývá, že predikce pomocí gaussovských procesů je lepší než lineární i kvadratická regrese. Srovnatelných hodnot je dosaženo v případě užití kovariáty $sin(x)$ v matici plánu. Na závěr je srovnána odhadnutá hodnota jednotlivých predikcí v bodě $x_* = 1,7$ se skutečnou hodnotou, což napoví, která metoda je přesnější.

	GPR	lineární regrese	kvadratická regrese	lineární regrese s kovariátou $sin(x)$	skutečná hodnota
hodnota v $x_* = 1,7$	1,1306	1,2840	1,3614	0,9021	**0,9917**

Tabulka 4.1: Srovnání odhadu funkční hodnoty v $x_* = 1,7$ pomocí jednotlivých metod se skutečnou hodnotou regresní funkce $sin(x)$

Skutečné hodnotě regresní funkce $sin(x)$ se nejvíce blíží odhad pomocí regrese s kovariátou $sin(x)$ a poté pomocí gaussovských procesů, což potvrzují i výsledky z obrázku 4.11. Je ale třeba uvažovat, že data měla šum $\sigma_n^2 = 0,1$. Tedy odhadovaná hodnota se pohybuje v intervalu $(0,6754; 1,3079)$.

Závěrem lze říci, že odhad pomocí gaussovských procesů je v porovnání s lineárním modelem velmi dobrý, v některých případech může dosahovat i lepších výsledků.

4.2 Reálná data

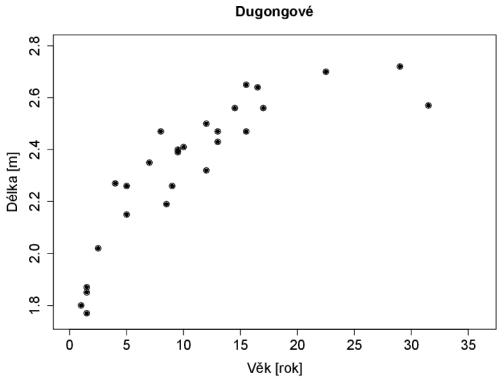
Tato podkapitola ukazuje užití gaussovských procesů v regresi na reálných datech. V části 4.2.1 jde o data s názvem *Dugongové* dostupná z [19]. Podle [20] je dugong druh vodního savce žijícího u pobřeží východní Afriky, Asie a Austrálie. V podkapitole 4.2.2 se jedná o data *Úmrtí způsobená nehodou v USA (Accidental Deaths in the US)*, viz [21]. Jsou použity stejné postupy a výpočty jako v podkapitole 4.1 se simulovanými daty.

4.2.1 Dugongové

V této části je zpracován datový soubor s názvem *Dugongové*, který obsahuje 27 pozorování. Data udávají věk a délku dugongů indických žijících nedaleko Townsville v severním Queenslandu v Austrálii. Tito savci se dožívají 50–60 let, rekordem je věk 73, a dosahují délky 2,4–3,2 metrů. Na obrázku 4.12 je ukázka datového souboru a pozorovaných dat znázorněných v grafu.

	Age	Length
1	1.0	1.80
2	1.5	1.85
3	1.5	1.87
4	1.5	1.77
5	2.5	2.02
6	4.0	2.27
7	5.0	2.15
8	5.0	2.26
9	7.0	2.35
10	8.0	2.47

(a) Ukázka datového souboru



(b) Pozorovaná data

Obr. 4.12: Ukázka datové tabulky a pozorovaných dat

K získání predikce je třeba vypočítat střední hodnotu a kovarianční matici podle vzorců 3.8 a 3.9. Jako kovarianční funkce se použije SE kovarianční funkce, jelikož je hladká, což je podle rozložení dat v obrázku 4.12 (b) nejspíše i hledaná funkce. SE kovarianční funkce je použita s optimalizovanými hyperparametry $\boldsymbol{\theta} = (l, \sigma_f, \sigma_n)$, které jsou získány pomocí funkce `optim`. Celkem jsou zjištěny tři různé trojice optimalizovaných hyperparametrů, které závisí na volbě vektoru jejich vstupních hodnot `par`. V tabulce 4.2 jsou popsány různé `par` a jim odpovídající výsledné optimální hyperparametry.

$\boldsymbol{\theta}$	odhad $\boldsymbol{\theta}$
l	15,9573
σ_f	1,7877
σ_n	0,0849

(a) `par`$= c(1,1,0.3)$

$\boldsymbol{\theta}$	odhad $\boldsymbol{\theta}$
l	39,3312
σ_f	1,9077
σ_n	0,0952

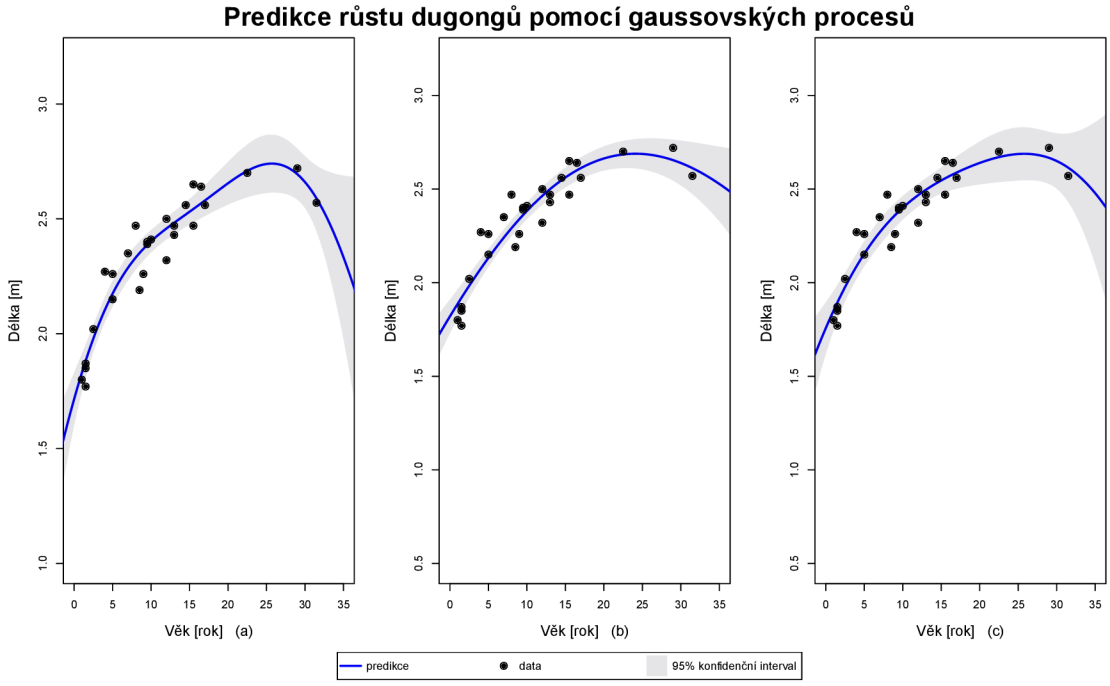
(b) `par`$= c(0.4,1,0.6)$

$\boldsymbol{\theta}$	odhad $\boldsymbol{\theta}$
l	19,1186
σ_f	2,1394
σ_n	0,1132

(c) `par`$= c(2,1,0.3)$

Tabulka 4.2: Optimalizované hodnoty hyperparametrů SE kovarianční funkce pro tři různé `par` ve funkci `optim`

Následně jsou zjištěné hyperparametry dosazeny do výpočtů. Výsledné predikce jsou zobrazeny na obrázku 4.13.

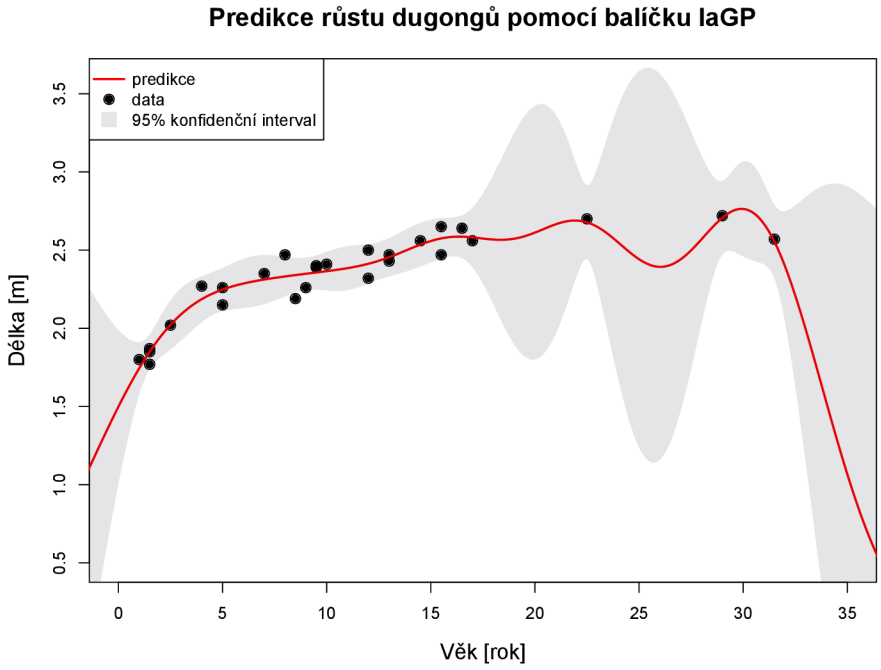


Obr. 4.13: Predikce reálných dat *Dugongové* odpovídající hyperparametrům z tabulky 4.2

Na základě vizuálního srovnání se jako nejlepší predikce jeví 4.13 (b), jelikož podle pásu nejistoty je zřejmé, že délka s věkem bude klesat, což u pásů nejistoty ostatních grafů nemusí platit. U 4.13 (c) by zřejmě dugong mohl stále růst, což není moc pravděpodobné, jelikož podle naměřených dat délka s věkem vyšším než 30 let spíše klesá. U 4.13 (a) i

(c) jsou pásy nejistoty širší než u 4.13 (b), a proto je těžší odhadovat další délku dugonga. Navíc kolem 25. roku je v délce zaznamenán určitý výkyv, což neodpovídá skutečnosti. Proto je vhodné jako nejlepší predikci vybrat 4.13 (b).

Predikci lze provést i pomocí balíčku laGP. Zde je vcelku složité nalézt ve funkci `newGP` vhodné vstupní hyperparametry `d` a `g`. Proto je v tomto případě lepší pro výpočet predikce užití vzorců. Na obrázku 4.14 lze pozorovat jednu z možných predikcí pomocí balíčku.



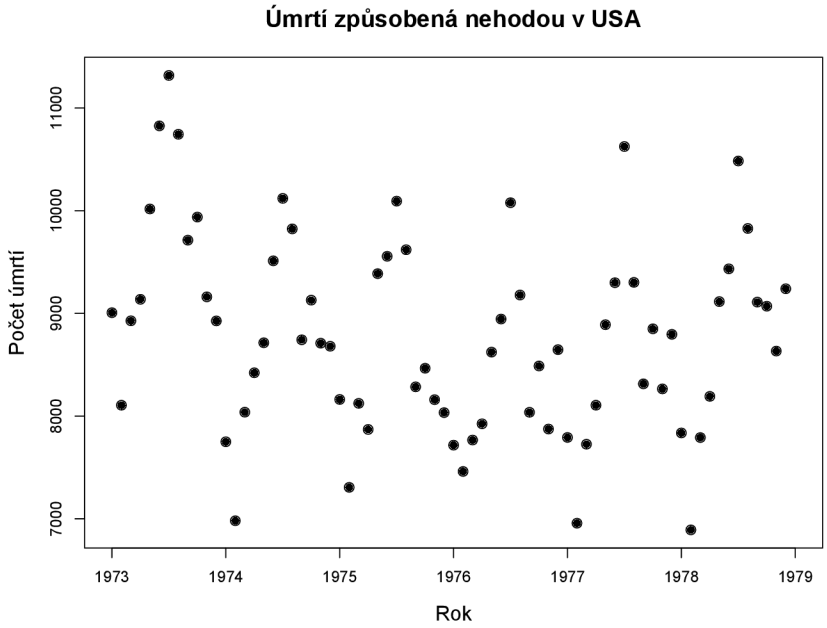
Obr. 4.14: Predikce reálných dat *Dugongové* pomocí balíčku **laGP**

4.2.2 Úmrtí způsobená nehodou v USA

Druhý datový soubor tvoří *Úmrtí způsobená nehodou v USA* během let 1973 až 1978. Jedná se o časovou řadu udávající měsíční součty úmrtí. Datová sada je volně dostupná v softwaru `R`. Na obrázku 4.15 je zobrazena datová tabulka, obrázek 4.16 znázorňuje tato data v grafu.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1973	9007	8106	8928	9137	10017	10826	11317	10744	9713	9938	9161	8927
1974	7750	6981	8038	8422	8714	9512	10120	9823	8743	9129	8710	8680
1975	8162	7306	8124	7870	9387	9556	10093	9620	8285	8466	8160	8034
1976	7717	7461	7767	7925	8623	8945	10078	9179	8037	8488	7874	8647
1977	7792	6957	7726	8106	8890	9299	10625	9302	8314	8850	8265	8796
1978	7836	6892	7791	8192	9115	9434	10484	9827	9110	9070	8633	9240

Obr. 4.15: Pozorovaná data



Obr. 4.16: Pozorovaná data

Podle obrázku 4.16 lze soudit, že hledaná křivka je periodická a hladká, a proto je využita periodická a SE kovarianční funkce. Jsou vytvořeny dvě různé predikce, nejprve s kovarianční funkcí získanou součinem periodické a SE, poté s kovarianční funkcí získanou součinem periodické a dvěma různými SE.

Kovarianční funkce periodická×SE
Pro optimalizaci je použita funkce `optim` s `par`$= c(1,1,0.5,1,1,0.2)$ odpovídající hyperparametrům (l,σ_f,p) pro periodickou kovarianční funkci, (l,σ_f) pro SE kovarianční funkci a σ_n pro obě kovarianční funkce společně. Tabulka 4.3 zobrazuje zjištěné hodnoty hyperparametrů.

hyperparametr	**odhad hyperparametru**
periodické l	2,9677
periodická σ_f	10,3021
perioda p	28,3388
SE l	0,1118
SE σ_f	21,1898
σ_n	7,1618

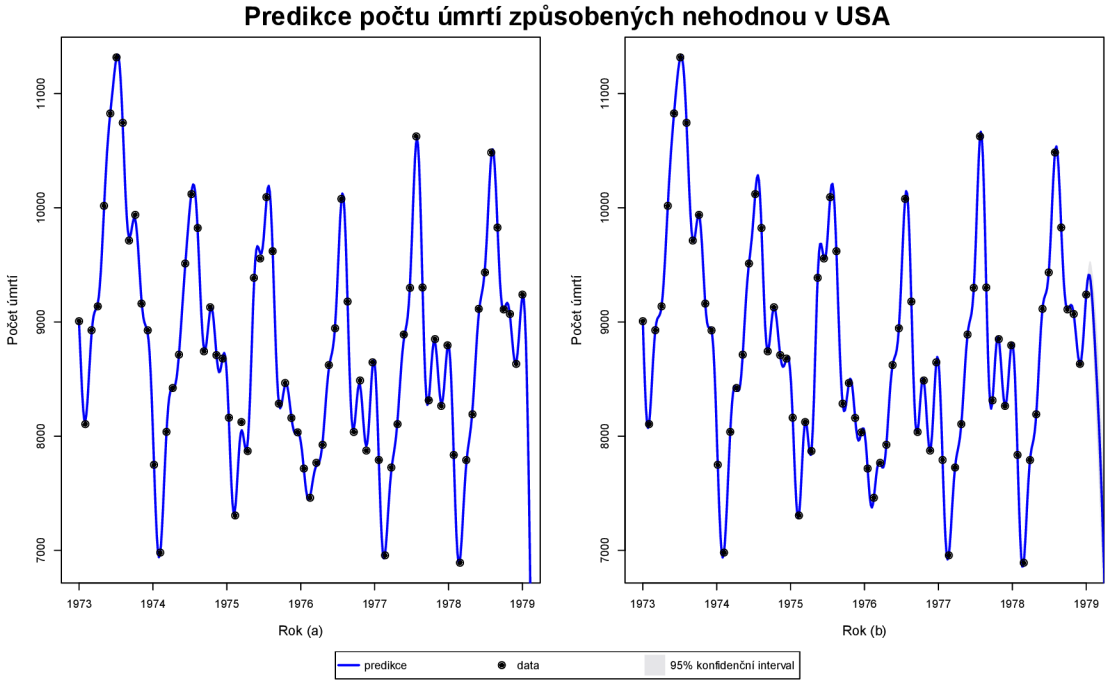
Tabulka 4.3: Optimalizované hodnoty hyperparametrů kovarianční funkce periodická×SE

Kovarianční funkce periodická×SE×SE
Opět jsou vypočítány optimální hyperparametry, a to (l,σ_f,p) pro periodickou kovarianční funkci, (l,σ_f) pro SE kovarianční funkce, pro každou rozdílné hodnoty, a σ_n pro všechny tři kovarianční funkce společně. Ve funkci `optim` je jako `par` zadán vektor $c(1,1,1,1,1,1,1,0.2)$. V tabulce 4.4 jsou vypsány výsledné hyperparametry.

hyperparametr	**odhad hyperparametru**
periodické l	1,4427
periodická σ_f	10,5298
perioda p	0,6471
SE1 l	0,5419
SE1 σ_f	2,7413
SE2 l	1,5141
SE2 σ_f	13,5570
σ_n	0,4092

Tabulka 4.4: Optimalizované hodnoty hyperparametrů kovarianční funkce periodická×SE×SE

V obou případech jsou následně pomocí vzorců z podkapitoly 3.1.2 vypočítány střední hodnoty a kovarianční matice, čímž jsou získány predikované hodnoty. Odhadnuté křivky jsou na obrázku 4.17.



Obr. 4.17: Predikce reálných dat *Úmrtí způsobená nehodou v USA* (a) Periodická×SE kovarianční funkce, (b) Periodická×SE×SE kovarianční funkce

Obě predikce jsou podobné, 4.17 (b) je v některých místech přesnější než 4.17 (a). Oblasti nejistoty ani v jednom grafu nejsou výrazné, proto jde o velmi přesná proložení dat křivkou. Lepší predikce se zdá být na obrázku 4.17 (b), jelikož protíná všechny body, nemá v závěru tak strmý pokles a celkově působí realističtěji. Je ale těžké rozhodnout o lepší predikci, jelikož obě vychází velmi dobře.

Závěr

Hlavním cílem bakalářské práce s názvem Gaussovské procesy v regresi bylo uvést základní popis této metodologie, pomocí simulací ukázat vliv hyperparametrů na celkový odhad regresní křivky a následně aplikovat tento přístup na simulovaná a reálná data. K tvorbě grafů a predikcí byl využit software `R`.

Při práci s gaussovskými procesy bylo zjištěno, že velmi záleží na volbě kovarianční funkce, jelikož se od ní odvíjí její hyperparametry. Jejich volba je také stěžejní, jelikož ovlivňují chování predikované křivky, hlavně její hladkost a rozptyl. Často se tyto hyperparametry odhadují pomocí metody maximální věrohodnosti, k čemuž v softwaru `R` slouží například funkce `optim`. U této funkce nastává problém při volbě vstupního parametru `par`, což je vektor počátečních hodnot hyperparametrů, které mají být odhadnuty. Je třeba vyzkoušet predikce s různými odhady a rozhodnout, se kterými by predikce mohla být vhodná. Při vhodné volbě všech těchto částí, lze jednoduše zjistit odhadovanou křivku.

V praktické části bylo ukázáno, že výsledky toho přístupu jsou velmi přesné. Při porovnání se známou funkcí je zřejmé, že predikce se velmi blíží známé křivce. Predikované hodnoty se také dají srovnat s výsledky lineárního modelu.

Nakonec byly gaussovské procesy v regresi aplikovány na reálná data. U první datové sady *Dugongové* bylo za cíl zjistit další růst dugongů během stárnutí. Zde nastal problém s volbou optimálních hyperparametrů, jelikož několik různých vstupních hodnot `par` dávalo rozdílné hypeparametry. Byly vybrány tři různé trojice hyperparemtrů a bylo diskutováno, se kterými vyšla lepší predikce. Druhá datová sada *Úmrtní způsobená nehodou v USA* měla složitější data, zde bylo cílem najít vhodnou kovarianční funkci. Podle vzhledu dat byla zvolena periodická funkce, která byla násobena jednou nebo dvěma SE kovariančními funkcemi. Obě predikce byly ve výsledku velmi přesné.

Z výsledků bakalářské práce vyplývá, že gaussovské procesy jsou za předpokladu existence normality vhodnou metodou regresní analýzy. Při vhodně zvolené kovarianční funkci a funkci středních hodnot jde o účinný nástroj pro tvorbu predikcí.

Seznam použité literatury

[1] YAN, Xin a Xiao Gang SU. *Linear Regression Analysis: Theory and Computing* [online]. World Scientific Publishing Co. Pte., 2009 [cit. 2021-02-03]. ISBN 978-981-4470-08-7. Dostupné z: http://www.manalhelal.com/Books/geo/LinearRegressionAnalysisTheoryandComputing.pdf

[2] RENCHER, Alvin C. a G. Bruce SCHAALJE. *Linear Models in Statistics* [online]. Second Edition. John Wiley & Sons, 2008 [cit. 2021-02-03]. ISBN 9780470192603. Dostupné z: http://www.utstat.toronto.edu/brunner/books/LinearModelsInStatistics.pdf

[3] FALTÝNKOVÁ, Jana. *Základy bayesovské analýzy dat* [online]. Brno, 2012 [cit. 2021-04-09]. Dostupné z: https://is.muni.cz/th/ut7y9/. Bakalářská práce. Masarykova univerzita, Přírodovědecká fakulta. Vedoucí práce Martin KOLÁŘ.

[4] HEBÁK, Petr. A Comparison of Classical and Bayesian Probability and Statitics (1). *Acta Oeconomica Pragensia* [online]. 2012, February 1, 2012, **20**(1), 69-87 [cit. 2021-4-9]. ISSN 05723043. Dostupné z: http://aop.vse.cz/doi/10.18267/j.aop.359.html

[5] FORBELSKÁ, Marie a Jan KOLÁČEK. *Pravděpodobnost a statistika II*.

[6] KRAUS, Andrea. *Linear Models in Statistics I: Lecture 3: Multivariate normal distribution*. Brno, 2020.

[7] KULICH, Michal. *Souhrn teorie pravděpodobnosti: Pro obor Finanční matematika* [online]. Karlín, 2013 [cit. 2021-04-10]. Dostupné z: https://www2.karlin.mff.cuni.cz/~pesta/NMFM301/pravdepodobnost_fm.pdf

[8] Gaussian process. *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation [cit. 2021-5-8]. Dostupné z: https://en.wikipedia.org/wiki/Gaussian_process

[9] RASMUSSEN, Carl Edward a Christopher K.I. WILLIAMS. *Gaussian Processes for Machine Learning* [online]. Cambridge: The MIT Press, 2006 [cit. 2021-02-09]. ISBN 026218253X. Dostupné z: http://www.gaussianprocess.org/gpml/chapters/RW.pdf

[10] MURPHY, Kevin P. *Machine learning: A probabilistic perspective* [online]. The MIT Press, 2012 [cit. 2021-02-09]. ISBN 9780262304320. Dostupné z: http://noiselab.ucsd.edu/ECE228/Murphy_Machine_Learning.pdf

[11] DUVENAUD, David Kristjanson. *Automatic Model Construction with Gaussian Processes* [online]. Pembroke College, 2014 [cit. 2021-02-09]. Dostupné z: https://www.cs.toronto.edu/~duvenaud/thesis.pdf. Dissertation. University of Cambridge.

[12] KRASSER, Martin. *Gaussian processes* [online]. March 19, 2018 [cit. 2021-04-02]. Dostupné z: http://krasserm.github.io/2018/03/19/gaussian-processes/

[13] WEI, Yi. Understanding Gaussian Process, the Socratic Way. *Towards Data Science* [online]. Dec 1, 2019 [cit. 2021-04-10]. Dostupné z: https://towardsdatascience.com/understanding-gaussian-process-the-socratic-way-ba02369d804

[14] ROELANTS, Peter. Fitting a Gaussian process kernel. In: *Notes on machine learning* [online]. [cit. 2021-04-06]. Dostupné z: https://peterroelants.github.io/posts/gaussian-process-kernel-fitting/

[15] COUFAL, Martin. *Gaussian Processes Based Hyper-Optimization of Neural Networks* [online]. Brno, 2020 [cit. 2021-04-12]. Dostupné z: https://www.fit.vut.cz/study/thesis/22368/.en. Master's Thesis. Brno University of Technology, Faculty of Information Technology. Vedoucí práce Karel Beneš.

[16] EBDEN, Mark. *Gaussian Processes: A Quick Introduction* [online]. In: . August 2008, s. 1-6 [cit. 2021-04-06]. Dostupné z: https://arxiv.org/pdf/1505.02965.pdf

[17] GUNDERSEN, Gregory. *Gaussian Process Regression with Code Snippets* [online]. 27 June 2019 [cit. 2021-04-02]. Dostupné z: http://gregorygundersen.com/blog/2019/06/27/gp-regression/

[18] GRAMACY, Robert B. a Furong SUN. *Package -laGP-* [online]. September 7, 2019 [cit. 2021-5-8]. Dostupné z: https://cran.r-project.org/web/packages/laGP/laGP.pdf

[19] Age and Length of Dugongs near Townsville. *StatSci.org* [online]. [cit. 2021-5-8]. Dostupné z: http://www.statsci.org/data/oz/dugongs.html

[20] Dugong indický. *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation [cit. 2021-5-8]. Dostupné z: https://cs.wikipedia.org/wiki/Dugong_indick%C3%BD

[21] Accidental Deaths in the US 1973–1978. *The R Datasets Package* [online]. [cit. 2021-5-8]. Dostupné z: https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/USAccDeaths.html