UNIVERZA V LJUBLJANI

Fakulteta za strojništvo

Identifikacija dinamičnih sistemov z globokimi Gaussovimi procesi

Magistrsko delo Magistrskega študijskega programa II. stopnje
STROJNIŠTVO

Mitja Jančič

Ljubljana, november 2017

UNIVERZA V LJUBLJANI

Fakulteta za strojništvo

Identifikacija dinamičnih sistemov z globokimi Gaussovimi procesi

Magistrsko delo Magistrskega študijskega programa II. stopnje
STROJNIŠTVO

Mitja Jančič

Mentor: prof. dr. Edvard Govekar
Somentor: prof. dr. Juš Kocijan

Ljubljana, november 2017

[Prostor za podpisano temo zaključnega dela]

Zahvala

Za uspešno zaključeno magistrsko delo se zahvaljujem mentorjema prof. dr. Edvardu Govekarju in prof. dr. Jušu Kocijanu. Zahvaljujem se za odlično vodenje, konstruktivne pripombe predvsem pa za nesebično deljenje znanja in izkušenj s tehničnega področja ter področja pisanja strokovnih del. Zahvala tudi Martinu Stepančiču z odseka za vodenje sistemov na Institutu Jožef Stefan, za pomoč pri razumevanju latentnih spremenljivk in njihovega pomena nasploh.

Zahvaljujem se avtorju uporabljene programske opreme, dr. Andreasu Damianou, za pomoč pri razumevanju kode in nasploh za vodenje pri uporabi programskega paketa.

Posebna zahvala tudi podjetju *MEIS d.o.o.* za meritve meteoroloških spremenljivk okoli jedrske elektrarne Krško, s katerimi smo lahko metodo ovrednotili na praktičnem primeru. Iskreno se zahvaljujem tudi Andreju Kolarju in Mihu Vrhovniku iz podjetja *Naviter d.o.o.,* ki sta omogočila brezplačen dostop do precej zmogljivejše strojne opreme kot je moj prenosni računalnik. Mnoge analize tekom magistrskega dela so bile za moj prenosni računalnik računsko preveč kompleksne.

Na koncu bi se rad zahvalil tudi prijateljem in sošolcem, ki so mi pomagali na poti do magistrske izobrazbe. Posebej izpostavljam podporo in pomoč Pie, Doris in Tilna. Zahvala tudi mami in bratu za neskončno moralno in finančno podporo pri študiju, ter očetu, ki je bil v času študija glavna motivacija.

Izjava

1. Spodaj podpisani Mitja Jančič, rojen 23.06.1992 v Celju, študent Fakultete za strojništvo Univerze v Ljubljani, izjavljam, da sem magistrsko delo z naslovom Identifikacija dinamičnih sistemov z metodo globokih Gaussovih procesov, izdelal samostojno v sodelovanju z mentorjem prof. dr. Edvardom Govekarjem in somentorjem prof. dr. Jušem Kocijanom.

2. Izjavljam, da je magistrsko delo, ki sem ga oddal v elektronski obliki, identična tiskani verziji.

3. Izrecno izjavljam, da v skladu z določili Zakona o avtorski in sorodnih pravicah (Ur. l. RS, št. 21/1995 s spremembami) dovolim objavo magistrskega dela na spletnih straneh Fakultete in Univerze v Ljubljani.

4. S podpisom se strinjam z javno objavo svoje magistrskega dela na straneh na svetovnem spletu preko Repozitorija Univerze v Ljubljani.

S svojim podpisom zagotavljam, da:

- je predloženo besedilo rezultat izključno mojega lastnega raziskovalnega dela;
- je predloženo besedilo jezikovno korektno in tehnično pripravljeno v skladu z Navodili za izdelavo zaključnih del, kar pomeni, da sem:
  - poskrbel, da so dela in mnenja drugih avtorjev oziroma avtoric, ki jih uporabljam v magistrskem delu, citirana oziroma navedena v skladu z Navodili za izdelavo zaključnih del, in
  - pridobil vsa dovoljenja za uporabo uporabljenih podatkov in avtorskih del, ki so v celoti (v pisni ali grafični obliki) uporabljena v tekstu, in sem to v besedilu tudi jasno zapisal;
- se zavedam, da je plagiatorstvo – predstavljanje tujih del (v pisni ali grafični obliki) kot mojih lastnih – kaznivo po Kazenskem zakoniku (Ur. l. RS, št. 55/2008 s spremembami);
- se zavedam posledic, ki bi jih na osnovi predložene magistrskega dela dokazano plagiatorstvo lahko predstavljalo za moj status na Fakulteti za strojništvo Univerze v Ljubljani v skladu z relevantnim pravilnikom.

V Ljubljani, dne 15.07.2017 Podpis avtorja: _____

Izvleček

UDK 517.938:519.218.7(043.2)

Tek. štev.: MAG II/450

Identifikacija dinamičnih sistemov z globokimi Gaussovimi procesi

Mitja Jančič

Ključne besede: identifikacija dinamičnih sistemov
globoki Gaussovi procesi
autoregresivni model
enokoračna napoved
neparametričen model
nelinearni sistem

Zaradi naraščajoče kompleksnosti obravnavanih sistemov in posledično zahtevnega matematičnega modeliranja, v praksi pogosto uporabimo empirične modele ali modele črne škatle, s katerimi modeliramo le povezave med vhodno-izhodnimi vrednostmi, ne pa tudi fizikalnih zakonitosti, ki se jim sistem podreja. Za modeliranje oziroma identifikacijo zveze med vhodnimi in izhodnimi vrednostmi sistema se uporabljajo tudi globoki Gaussovi procesi. Ti za opis kompleksnejših procesov uporabljajo gnezdenje in hierarhično strukturo. Z identificirano zvezo med vhodno-izhodnimi vrednostmi z uporabo Gaussovih procesov lahko za dane vhodne vrednosti napovemo vrednost izhoda in pripadajočo negotovost, kar lahko s pridom uporabimo. V okviru magistrskega dela predstavimo teoretične osnove modeliranja z globokimi Gaussovimi procesi in njihove prednosti. V ta namen v ilustrativnem primeru uspešno identificiramo dinamični sistem nelinearnega nihanja mase, v bolj praktičnem primeru pa obravnavamo bistveno kompleksnejši sistem napovedovanja temperature v prizemni plasti atmosfere.

Abstract

UDC 517.938:519.218.7(043.2)

No.: MAG II/450

Identification of dynamic systems using deep Gaussian Processes

Mitja Jančič

Key words: identification of dynamic systems
deep Gaussian Processes
autoregressive models
one-step ahead prediction
nonparametric model
nonlinear systems

Mathemathical and physical modelling only provide approximate description of the true nature of a dynamic system. The higher the precision of the model the more likely it becomes analytically intractable and, therefore, empirical models or black box models are used. When dynamic systems are considered as black box models, almost no prior knowledge about the system is considered. Deep Gaussian Processes, which use hierarchical structure to provide adequate identification of very complex systems, can be used to identify the mapping between the system input and output values. With the given mapping function we can then provide a one-step ahead prediction of the system output values, together with its uncertainty, which can be advantageously used. In this paper we use deep Gaussian Processes to identify a dynamic system and present its advantages by studying two cases. In the first illustrative case we successfully identify the dynamic properties of a nonlinear oscillating mass, while in the second, more realistic and complex case, we study one-step ahead prediction of air temperature in the atmospheric surface layer.

Kazalo

Kazalo slik

Kazalo preglednic

Seznam uporabljenih simbolov

Oznaka	Enota	Pomen
$\mathcal{D}$	/	učna množica
e	/	napaka
f	/	zveza med vhodnimi in izhodnimi vrednostmi sistema
$\mathbf{F}$	/	matrika latentnih vrednosti preslikave
$\mathcal{F}$	/	funkcional
$\mathcal{GP}$	/	Gaussov proces
$\mathbf{h}$	/	vektor spremenljivk skritega sloja
$k(x, x')$	/	kovariančna funkcija
$\mathbf{K}$	/	kovariančna matrika
KL	/	Kullback-Leiblerjeva divergenca
L	/	število skritih slojev
$\mathcal{L}$	/	logaritem verjetnostne porazdelitve
m	/	število induciranih spremenljivk
n	/	število podatkov
$\mathcal{N}$	/	Gaussova verjetnostna porazdelitev
p	mbar	zračni tlak
$p(x)$	/	verjetnostna porazdelitev spremenljivke x
P	$\mathrm{W/m^2}$	globalno sončno sevanje
$\mathcal{PG}$	/	Pasquill-Girardovi indeksi stabilnosti atmosfere
q	/	dimenzija latentnega prostora
$q(x)$	/	variacijska verjetnostna porazdelitev
T	°C	temperatura zraka
$\mathbf{U}$	/	matrika induciranih spremenljivk
v	m/s	hitrost vetra
x	/	vhodna vrednost
$\mathbf{X}$	/	matrika latentnih vrednosti vhodov
$\mathbf{X}_u$	/	matrika latentnih vrednosti psevdovhodov
y	/	izhodna vrednost
$\mathbf{Y}$	/	matrika izhodnih vrednosti
$\mathbf{Z}$	/	matrika izmerjenih vhodnih vrednosti
β^{-1}	/	varianca gaussovsko porazdeljenega šuma
$\boldsymbol{\epsilon}$	/	vektor gaussovsko porazdeljenega šuma
$\boldsymbol{\Theta}$	/	hiperparametri kovariančne funkcije
$\boldsymbol{\mu}$	/	vektor srednjih vrednosti napovedi
ξ	/	statistika
σ	/	standardna deviacija
φ	%	relativna vlažnost zraka
$\boldsymbol{\Phi}$	/	statistika
$\boldsymbol{\psi}$	/	statistika

l	skriti sloj
NRMSE	normalizirana vrednost korena srednje vrednost kvadratične napake
u	psevdovhod
$*$	testna vrednost

Seznam uporabljenih okrajšav

Okrajšava	Pomen
ARD	avtomatsko določanje ustreznosti (angl. *Automatic Relevance Determination*)
GP	Gaussov proces
JEK	jedrska elektrarna Krško
LVM	model latentnih spremenljivk (angl. *Latent Variable Model*)
MAP	maksimalno aposteriorno verjetje (angl. *Maximum Aposteriori Probability*)
NARX	angl. *Nonlinear autoregressive model with exogenous inputs*
NRMSE	normalizirana vrednost korena srednje vrednosti kvadratične napake (angl. *Normalized Root Mean Square Error*)

1 Uvod

Modeliranje dinamičnega sistema je pogosto naloga raziskovalca. Človek je radoveden in rad postavlja matematične in fizikalne modele naravnih pojavov in sistemov. S tem dobimo bistveno globlji vpogled v razumevanje sistema, predvsem pa omogočimo napoved njegovega obnašanja.

V dejanskem svetu se pogosto srečamo s sistemi, katerih fizikalno in matematično ozadje ni enolično določeno ali iz različnih razlogov ni dosegljivo. Tedaj postavimo eksperimentalni model in sistem ali proces obravnavamo kot model črne škatle. Ti pa zahtevajo nekoliko drugačen pristop. V praksi se pogosto uporabijo linearne aproksimacije procesov, nelinearne popravke pa bodisi zanemarimo bodisi jih poiščemo z drugimi metodami, npr. z nevronskimi mrežami, kot v delih [1] in [2], kjer avtorji napovejo porabo zemeljskega plina za en dan vnaprej. Podobno v članku [3] opišejo enodnevno napoved obremenitve ogrevalnega sistema.

V sklopu tega magistrskega dela bomo model črne škatle reševali z modelom globokih Gaussovih procesov (globokih GP) [4, 5]. GP-model je neparametričen model, zato je število parametrov potrebnih optimizacije precej manjše v primerjavi z nevronskimi mrežami.

Namen magistrskega dela je seznanitev in vrednotenje metode identifikacije dinamičnih sistemov z globokimi GP. Globoki GP so zaradi podobnosti s principom umetne inteligence in globokim učenjem primernejši za identifikacijo kompleksnejših sistemov, predvsem pa so zaradi svoje strukture in majhnega števila parametrov primernejši za obdelavo ogromne količine podatkov. Model globokih GP bomo uporabili in ovrednotili na dveh praktičnih primerih.

1.1 Struktura dela

Magistrsko delo se začne s splošno predstavitvijo GP-modela. V poglavju 2 pojasnimo kaj GP sploh je. Predstavimo le glavne značilnosti GP-modela in poudarimo pomen kovariančne funkcije ter matematični proces učenja modela.

S splošnim znanjem o GP-modelih nato v poglavju 3 predstavimo poseben GP-model, prvotno predstavljen v namen nenadzorovanega učenja z latentnimi spremenljivkami, to je GP-model latentnih spremenljivk (GP-LVM) [4,6]. V uvodnih podpoglavjih pojasnimo pomen latentnih vhodov in njihove posledice na učenje matematičnega modela.

Izkaže se, da postane robna verjetnost napovedi z vpeljavo latentnih spremenljivk analitično neizračunljiva. S tem razlogom v poglavju 3.3 dodatno vpeljemo še inducirane spremenljivke in več variacijskih verjetnostnih porazdelitev kot aproksimacije pravim verjetnostnim porazdelitvam. Računska kompleksnost modela je opisana v podpoglavju 3.3.3, v naslednjih dveh pa tudi princip enokoračnega napovedovanja vrednosti.

Variacijski model GP-LVM lahko gnezdimo in s tem tvorimo hierarhično strukturo globokih GP v poglavju 4. Izpeljava je narejena za primer nenadzorovanega učenja. Razširitev na problem nadzorovanega učenja je preprosta in zahteva le razmeroma preprost dodatek k modelu.

Predstavljena modela globokih GP in model latentnih spremenljivk uporabimo na konkretnem ilustrativnem primeru v poglavju 5, kjer identificiramo sistem z nelinearnim nihanjem mase. Sistem najprej matematično opišemo, a ga v nadaljevanju obravnavamo kot model črne škatle. Več pozornosti namenimo premišljeni inicializaciji hiperparametrov, latentnih in induciranih spremenljivk v poglavju 5.3.2.

V poglavju 6 modela (globoki GP in GP-LVM) uporabimo tudi na kompleksnejšem realnem problemu napovedi temperature v okolici jedrske elektrarne Krško (JEK). V podpoglavju 6.2 najprej poskrbimo za predobdelavo meritev in jih hkrati razdelimo na učno in testno množico. Sledi podpoglavje 6.3 kjer poiščemo optimalno obliko regresorskega vektorja ter opravimo vrsto eksperimentov z naključno in premišljeno inicializacijo globokega GP-modela.

1.2 Pregled literature

V uvodni predstavitvi GP se v veliki meri opiramo na knjigo Rasmussena in Williamsa [7], diplomski deli [8, 9] ter uvodna poglavja knjige [10], kjer so povzete glavne značilnosti standardnih GP-modelov.

Posebna izpeljanka Gaussovih procesov, model latentnih spremenljivk, je dobro opisana v delih Lawrenca [6, 11]. A model GP-LVM ni analitično izračunljiv in posledično tvorjenje globokih mrež ni mogoče. Zato je v doktorski disertaciji [4] predstavljen variacijski model GP-LVM z izračunljivo spodnjo mejo verjetnosti. Poleg doktorske disertacije se pri izpeljavi variacijskega GP-LVM opiramo tudi na dela [12–14].

Z vpeljanim variacijskim GP-LVM-modelom lahko končno tvorimo tudi model globokih GP, kjer se opiramo na doktorsko disertacijo [4] in članek [5].

Modela globokih GP in GP-LVM preizkusimo na sintetičnih podatkih iz [15, 16] kjer za definicijo regresorskega vektorja privzamemo rezultate članka [17].

Nazadnje se v sklopu magistrskega dela posvetimo realnemu problemu napovedi temperature v okolici JEK. Za opis sistema in razumevanje meteoroloških pojavov v prizemni plasti atmosfere se zanašamo na deli [18] in [19].

2 Gaussovi procesi

Modeliranje z GP uvrščamo na področje empiričnega modeliranja. V grobem ločimo *nadzorovano* in *nenadzorovano* učenje modela, kjer v prvem primeru poznamo vrednosti izhodov in vhodov obravnavanega sistema, medtem ko v drugem primeru poznamo le njegove izhodne vrednosti [7]. Modeliranje z GP je empirično modeliranje. To pomeni, da želimo za vhodne vrednosti x_i in izhodne vrednosti y_i iz množice učnih podatkov $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ določiti matematično strukturo, ki bo kar najbolje opisovala obravnavani sistem - da bo izhod modela pri vhodnih podatkih x_i čim bolj podoben izhodu sistema y_i [8]. Model na podlagi GP lahko opredelimo kot model *črne škatle* (angl. *black-box model*) [7], kjer podrobne informacije o fizikalni in matematični naravi sistema niso poznane ali iz nekih razlogov celo niso dosegljive in so nam na voljo zgolj vhodno-izhodne vrednosti $\{(x_i, y_i)\}_{i=1}^{n}$ sistema.

Številni problemi regresije, klasifikacije in zmanjšanje dimenzionalnosti problema obsegajo učenje neznane preslikave f med vhodnimi in izhodnimi vrednostmi sistema. Znani postopki identifikacije preslikave f običajno omogočajo, da dobimo dobre informacije o preslikavi šele kadar upoštevamo nekaj splošnih predpostavk o njeni naravi, njenem obnašanju [4,12]. Predpostavke (npr. gladkost preslikave, periodičnost sistema, idr.) matematično vpeljemo v sistem z ustrezno definicijo apriornega verjetja Gaussovih procesov. Ti pa skupaj z optimizacijskim postopkom in upoštevanjem meritev vodijo do neparametrične oblike preslikave f, ki se kar najbolje prilega učnim podatkom iz množice $\mathcal{D}$.

2.1 Od Gaussovih procesov do regresijskega modela

Komponente naključnega vektorja so naključne spremenljivke. Naključni proces (angl. *stohastic process*) je večkratna realizacija naključne spremenljivke v odvisnosti od nekih neodvisnih spremenljivk, npr. časa [9]. Fizikalno gledano so naključni procesi tisti procesi, ki niso ponovljivi in pri katerih se opazovane veličine obnašajo naključno.

Kadar so vrednosti naključne spremenljivke normalno porazdeljene, pravimo takemu procesu GP. GP torej razumemo kot porazdelitev *funkcijskega prostora*. Kar je Gaussova porazdelitev za spremenljivko, predstavlja GP za celotno funkcijo [8]. Še drugače razumemo GP kot posplošitev večdimenzionalne Gaussove porazdelitve na neskončno število dimenzij [4], t.j. naključnih spremenljivk. Vzorec iz GP je funkcija f.

Podobno kot je Gaussova porazdelitev popolnoma določena s končnim vektorjem povprečja $\boldsymbol{\mu}$ in kovariančno matriko $\mathbf{K}$, je tudi GP popolnoma definiran s povprečno funkcijo $\mu(x)$ in kovariančno funkcijo $k_y(x, x')$, obe definirani za katerikoli $x, x' \in \mathbb{R}$. Tu sta x in x' dve vhodni vrednosti. Izbrana kovariančna funkcija definira lastnosti kot sta recimo gladkost in stacionarnost preslikave f, ne pa tudi njene parametrične oblike [4, 7].

Vzemimo vektor meritev $\mathbf{y} = (y_1, \ldots, y_n) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$ izhodnih vrednosti opravljenih na mestih x_i. Naključna spremenljivka y_i lahko predstavlja temperaturo zraka v odvisnosti od nadmorske višine x_i. Koreliranje meritev pomeni, da mora biti temperatura na podobni višini x_{i-1} in x_{i+1} podobna, kar se mora odražati tudi v kovariančni matriki na mestih $\mathbf{K}_{i,i-1}$ in $\mathbf{K}_{i,i+1}$. Kovariančna funkcija tako omogoča interpolacijo meritev glede na njihovo bližino [4]. Toda v dejanskem svetu bi namesto vnaprej izbranih diskretnih točk $\mathbf{x}$ raje upoštevali celotno domeno vhodov, za kar pa uporabimo GP.

Pomembna lastnost Gaussove porazdelitve omogoča, da obravnavamo le končen set meritev $\mathbf{y} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))$. Privzemimo, da smo meritev temperature $\mathbf{y}_A$ in $\mathbf{y}_B$ opravili na dveh mestih $\mathcal{X}_A, \mathcal{X}_B \subseteq \mathcal{X}$. Tedaj je verjetnost

$$p(\mathcal{X}) = p(\mathcal{X}_A, \mathcal{X}_B) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}) \tag{2.1}$$

za dana

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{bmatrix} \text{ in}$$
$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{AA} & \mathbf{K}_{AB} \\ \mathbf{K}_{BA} & \mathbf{K}_{BB} \end{bmatrix}.$$

Lastnost Gaussove porazdelitve pravi, da lahko z meritvami temperature iz množice $\mathcal{X}_A$ sklepamo na temperaturo na lokaciji $\mathcal{X}_B$ saj sta robni verjetnosti neodvisni od meritev iz druge množice

$$p(\mathbf{y}_A) = \int_{\mathbf{y}_B} p(\mathbf{y}_A, \mathbf{y}_B) \mathrm{d}\mathbf{y}_B = \mathcal{N}(\mathbf{y}_A | \boldsymbol{\mu}_A, \mathbf{K}_{AA}), \tag{2.2}$$

$$p(\mathbf{y}_B) = \int_{\mathbf{y}_A} p(\mathbf{y}_A, \mathbf{y}_B) \mathrm{d}\mathbf{y}_A = \mathcal{N}(\mathbf{y}_B | \boldsymbol{\mu}_B, \mathbf{K}_{BB}). \tag{2.3}$$

Definicija direktno implicira, da prek neznanih vrednosti funkcije na neopazovanih mestih (takih je lahko neskončno mnogo) integriramo, kar pa pogosto ni mogoče [4]. Toda v primeru večdimenzionalne Gaussove porazdelitve vemo, da je podmnožica Gaussove množice normalno porazdeljena s povprečjem in kovariančno matriko podmnožice.

Pri uporabi GP pogosto uporabljamo matematični zapis iz [4]

$$\begin{aligned} f &\sim \mathcal{GP}(\mu(\mathbf{x}), k_y(\mathbf{x}, \mathbf{x}')), \\ \mu(\mathbf{x}) &= \mathbb{E}\left[f(\mathbf{x})\right], \\ k_y(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\left[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))\right], \end{aligned}$$

pri čemer s standardizacijo vhodno-izhodnih vrednosti sistema poskrbimo, da je povprečje $\mu(\mathbf{x}) = 0$ in standardna deviacija $\sigma_f = 1$. Takšna izbira je značilna za statično modeliranje, kjer je proces popolnoma opisljiv s statističnim momentom drugega reda, t.j. kovariančno funkcijo [4, 9].

V praksi torej zadostuje, da ovrednotimo kovariančno matriko na končnem vzorcu podatkov, s čimer določimo Gaussovo porazdelitev učne množice. Verjetnostno porazdelitev izhodov tedaj zapišemo

$$p(\mathbf{y}|x) = \mathcal{N}(\mathbf{y}|0, \mathbf{K_{yy}}) = (2\pi)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\mathbf{y}^\top \mathbf{K_{yy}^{-1}} \mathbf{y}\Big), \tag{2.4}$$

kjer matriko $\mathbf{K_{yy}} = k_y(\mathbf{x}, \mathbf{x}')$ izračunamo na vseh n učnih podatkih. V splošnem imajo kovariančne funkcije parametre $\boldsymbol{\Theta_f}$, ki smo jih v zapisu $p(\mathbf{y}|x, \boldsymbol{\Theta_f})$ izpustili. Več o kovariančnih funkcijah bomo napisali v poglavju 2.3.

Zavedajmo se, da natančne vrednosti $f(\mathbf{x})$ z meritvijo enostavno ni moč določiti, pravzaprav niti teoretično ni dostopna. V matematičnem modelu zato za bolj realističen model pogosto dodamo šum

$$\mathbf{y} = f(\mathbf{x}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I}), \tag{2.5}$$

zaradi česar vrednosti funkcije $f(\mathbf{x})$ označimo za *latentne* ali *skrite* spremenljivke $\mathbf{f}$ [4].

Največja prednost GP v praksi se pokaže v analitično izračunljivi robni verjetnosti izhodnih vrednosti glede na opazovane vhodne vrednosti

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{x}) \mathrm{d}\mathbf{f} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K_{yy}} + \beta^{-1}\mathbf{I}). \tag{2.6}$$

Zgornji izračuni vodijo do GP-modela, ki hkrati upošteva celo družino funkcij.

2.2 Bayesovo sklepanje

Matematična oblika Bayesovega izreka je [8]:

$$p(f(\mathbf{x})|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|f(\mathbf{x}), \mathbf{X}) p(f(\mathbf{x}))}{p(\mathbf{y}|\mathbf{X})}, \tag{2.7}$$

pri čemer so

- $p(f(\mathbf{x})|\mathbf{y}, \mathbf{X})$ verjetnost učnih izhodov glede na funkcijo $f(\mathbf{x})$,
- $p(f(\mathbf{x}))$ predstavlja apriorno verjetnostno porazdelitev v prostoru funkcij,
- $p(\mathbf{y}|\mathbf{X})$ robna verjetnost podatkov.

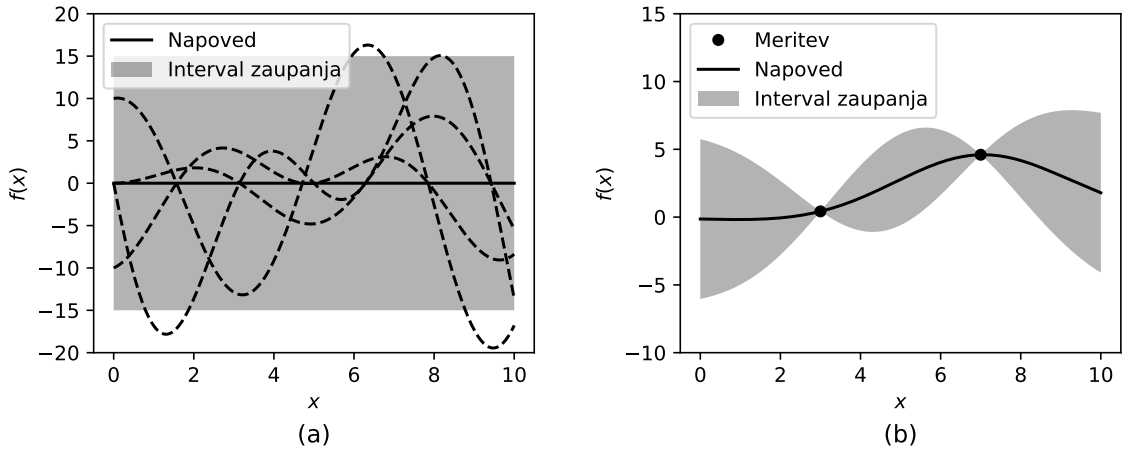
GP vidimo tudi kot neparametrično varianto bayesovske parametrične regresije [8]. Standardni parametrični model definira:

$$\mathbf{y} = \boldsymbol{\phi}(x)^\top \mathbf{w} + \boldsymbol{\epsilon}, \tag{2.8}$$

kjer je $\boldsymbol{\phi}(x)$ vektor baznih funkcij, ki vhodni podatek x nelinearno preslikajo v prostor stanj, npr. $\boldsymbol{\phi}(x) = (1, x, x^2, \ldots, x^k)^\top$ preslika vhodne podatke v krivuljo k-tega reda, medtem ko $\phi(x) = x$ predstavlja linearno regresijo. Vektor $\mathbf{w}$ je set uteži, ki jih v postopku optimizacije določimo tako, da je prileganje meritvam kar se da dobro. V bayesovski metodologiji namesto iskanja maksimalne vrednosti posteriorne verjetnosti

(angl. *Maximum Aposteriori Probability, MAP*) rajši upoštevamo predhodno znanje o sistemu in ga uporabimo na utežeh $\mathbf{w}$. Efektivno s tem utežimo vse možne kombinacije parametrov [4, 8].

Po podrobnejši analizi, npr. v [7], ugotovimo, da je tak pristop povsem ekvivalenten GP-modelu. Razlika je le, da bayesovski pristop upošteva parametrično obliko $\boldsymbol{\phi}(x)^\top \mathbf{w}$ in z apriornim znanjem direktno vplivamo na parametre $\mathbf{w}$, medtem ko pri GP z apriornim znanjem o sistemu vplivamo direktno na preslikavo f. Rezultat je bogatejši in bolj fleksibilen sistem, saj z apriornim znanjem omejimo le določene lastnosti modeliranih funkcij, medtem ko se njihova oblika formulira samodejno na osnovi vhodnih podatkov in predpostavk, ki so vgrajene v kovariančni funkciji.



Slika 2.1: Slika (a) prikazuje štiri naključno izbrane funkcije iz apriorne verjetnosti. Slika (b) prikazuje primer, ko v obravnavo dodamo še dve učni točki. Srednja vrednost napovedi je prikazana s polno črto, osenčeno območje prikazuje standardni odklon od povprečja.

Obravnavajmo preprost enodimenzionalni regresijski problem, s preslikavo vhodnega podatka x v izhod $f(x)$. Na Sliki 2.1(a) so prikazane štiri naključno izbrane funkcije iz funkcijskega prostora. Funkcije so gladke in se podrejajo apriornemu verjetju, s katerim izražamo naša pričakovanja kako naj bi funkcija f izgledala, še preden v obravnavo dodamo dejanske meritve. Sivo obarvano območje označuje področje standardnega odklona od povprečne vrednosti, ki ga interpretiramo kot mero zaupanja. V konkretnem primeru je uporabljen GP, pri katerem apriorna varianca ni odvisna od vrednosti x.

V naslednjem koraku dodamo točki iz učne množice $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$. Radi bi obravnavali le še tiste funkcije, ki gredo skozenj (v obravnavo bi lahko vzeli tudi funkcije, ki se točkama zgolj dovolj približajo). Primer je prikazan na Sliki 2.1(b). Polna črta prikazuje srednjo vrednost funkcijskega prostora. Opazimo, da se z dodajanjem meritev negotovost modela v njihovi okolici močno zmanjša. Zveza med apriornim verjetjem in merjenimi podatki vodi do posteriornega verjetja nad prostorom funkcij.

Če bi dodali še več točk, bi slej ali prej dosegli, da bi povprečje funkcij prečkalo vse dane točke in da bi se negotovost močno zmanjšala. Poudarimo, da GP-model ni parametričen model in da je skrb, ali lahko model podatke dobro popiše (kot bi bilo v primeru kadar bi skušali z linearnim modelom opisati močno nelinearne podatke),

odveč [7]. Četudi dodamo ogromno število točk, po navadi še vedno ostane nekaj fleksibilnosti za funkcije iz funkcijskega prostora. Za še lažji prikaz kako se z večanjem števila podatkov manjša fleksibilnost funkcij si zamislimo, da iz apriorne verjetnosti na Sliki 2.1(a) naključno izbiramo funkcije in zavračamo vse tiste, ki se ne ujemajo z meritvami iz množice $\mathcal{D}$ na Sliki 2.1(b).

2.3 Kovariančna funkcija

Vrednost kovariančne funkcije $k(\mathbf{x}_i, \mathbf{x}_j)$ izraža korelacijo med izhodoma $f(\mathbf{x}_i)$ in $f(\mathbf{x}_j)$ modela [9]. Bližje kot je vrednost kovariančne funkcije -1 ali 1, bolj sta njena argumenta korelirana. Kadar je vrednost kovariančne funkcije 0, med argumentoma ni korelacije.

V splošnem je lahko kovarinačna funkcija poljubna funkcija, ki tvori pozitivno definitno kovariančno matriko $\mathbf{K}$ za poljuben nabor vhodnih vektorjev [9]. Najpreprostejša oblika kovariančne funkcije ima konstantno vrednost

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2. \tag{2.9}$$

S fizikalnega vidika je primernejša izbira kovariančne funkcije, ki bolj korelira izhodne točke, ki so si v vhodnem prostoru blizu [8, 9]. Eksponentna kovariančna funkcija

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\Big(-(\frac{r}{l})^d \Big) \quad \text{za} \quad 0 < d \leq 2, \tag{2.10}$$

je že bližje fizikalni naravi sistema. Z razdaljo $r = |\mathbf{x}_i - \mathbf{x}_j|$ upoštevamo razdaljo v vhodnem prostoru, σ_f^2, l in d pa so parametri kovariančne funkcije.

Najbolj običajno je, da v primeru ko o sistemu ne vemo veliko, npr. model črne škatle, izberemo Gaussovo kovariančno funkcijo,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\Big(-\frac{1}{2} \sum_{d=1}^{D} w_d (x_i^d - x_j^d)^2 \Big), \tag{2.11}$$

kjer z D označujemo dimenzijo vhodnih vektorjev $\mathbf{x}_i$ in $\mathbf{x}_j$, σ_f in $\mathbf{w} = (w_1, \ldots, w_D)$ pa so parametri kovariančne funkcije. Parametre kovariančne funkcije imenujemo *hiperparametri* ($\boldsymbol{\Theta}$). S tem poudarimo, da so to parametri sicer neparametričnega modela, ki določajo verjetnostno porazdelitev nad funkcijskim prostorom [9].

S kovariančno funkcijo v matematični model vgradimo svoje znanje o sistemu. Konkretno v primeru (2.11) predpostavimo, da funkcije, ki jih imamo za bolj verjetne, izražajo gladkost (t.j. če se vhod malo spremeni, se tudi izhod malo spremeni) in stacionarnost (kovarianca je odvisna le od medsebojne razdalje med vhodnima vektorjema, ne pa tudi od njune medsebojne lege v prostoru) [9].

Poleg Gaussove kovariančne funkcije uporabljamo tudi eksponentno, ki je prav tako neskončnokrat odvedljiva, a manj prilagodljiva in vodi do izhoda GP-modela, ki ni tako gladek. Omenimo še linearno, polinomsko, periodično in Matérnovo kovariančno funkcijo. K temu dodajmo, da lahko kovariančne funkcije med seboj seštevamo in s tem popišemo bistveno bolj kompleksno strukturo [9, 10]. Opis oscilirajočega sistema čigar odziv se približno linearno veča s časom bi zelo dobro povzeli z vsoto linearne in periodične kovariančne funkcije.

2.3.1 Učenje

Učenje GP-modela v matematičnem smislu pomeni iskanje numeričnih vrednosti hiperparametrov modela. Pogosto v ta namen uporabimo optimizacijo po metodi največjega verjetja, po kateri za hiperparametre privzamemo tiste numerične vrednosti, ki so najbolj verjetne [8]. Določimo jih prek posteriorne verjetnostne porazdelitve:

$$p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\boldsymbol{\Theta}, \mathbf{X})p(\boldsymbol{\Theta})}{p(\mathbf{y}, \mathbf{X})}. \tag{2.12}$$

Optimalne vrednosti poiščemo z iskanjem največje vrednosti logaritma porazdelitve $p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X})$, torej

$$\log(p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X})) = \frac{1}{2}\log(|\mathbf{K}|) - \frac{1}{2}\mathbf{y}_n^\top \mathbf{K}^{-1}\mathbf{y}_n - \frac{n}{2}\log(2\pi). \tag{2.13}$$

Za določitev hiperparametrov modela torej z neko optimizacijsko metodo poiščemo maksimum logaritma (2.13). Običajno se uporabi negativno predznačena vrednost in poišče minimalna vrednost [8]. V praksi se zelo pogosto uporabljajo gradientne metode, ki pa zahtevajo tudi izračun odvoda po hiperparametrih

$$\frac{\partial}{\partial \Theta_i}\log(p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{X})) = -\frac{1}{2}\mathrm{Tr}(\mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \Theta_i}) + \frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\frac{\partial \mathbf{K}}{\partial \Theta_i}\mathbf{K}^{-1}\mathbf{y}^\top. \tag{2.14}$$

Pri tem omenimo, da je rang kovariančne matrike $\mathbf{K}$ enak številu podatkov učne množice. Več je učnih podatkov, večja je matrika $\mathbf{K}$, bolj zapleten in dolgotrajen je izračun njenega inverza. Računska kompleksnost takega modela narašča s tretjo potenco števila učnih podatkov n [4, 7].

3 GP-model latentnih spremenljivk

Kadar govorimo o GP imamo na voljo več različnih modelov [11]. V naslednjih poglavjih bomo predstavili značilnosti GP-modela latentnih spremenljivk (angl. *Gaussian Process Latent Variable Model, GP-LVM*), ki velja za zelo močno in robustno orodje za zmanjšanje dimenzij regresijskega problema. Zgodovinsko gledano je bil GP-LVM-model razvit za potrebe nenadzorovanega učenja [4], a je za prehod na nadzorovano učenje potreben le en dodaten razmislek.

Za identifikacijo dinamičnih sistemov je pogosto potrebna predobdelava opravljenih meritev. Bogatejša kot je dinamika in kompleksnejši kot je sistem, več meritev potrebujemo za izbrano natančnost identifikacije [6]. K sreči se v praksi pogosto izkaže, da lahko že z bistveno manjšo množico podatkov povsem zadovoljivo popišemo lastnosti opazovanega sistema. Možen razlog je, da imajo podatki, čeprav navidezno večdimenzionalni, nizkodimenzionalno ovojnico s katero opišemo vso poglavitno dinamiko sistema [7]. Ali drugače povedano, da lahko večdimenzionalne probleme iz narave zadovoljivo opišemo s podatki manjše dimenzije, kar je v resnici pogosta praksa aproksimiranja kompleksnih sistemov in struktur.

3.1 Latentni vhodi

Primarna naloga modela GP-LVM je, da poleg zmanjšanja dimenzionalnosti problema poišče tudi zvezo med matriko vhodnih vrednosti $\mathbf{X} \in \mathbb{R}^{n \times q}$ in latentnimi vrednostmi preslikave $\mathbf{Y} \in \mathbb{R}^{n \times p}$, oziroma $\mathbf{F} \in \mathbb{R}^{n \times p}$ s prištetim šumom po enačbi (2.5).

Glavni izziv pri tem predstavlja matrika neznanih vhodnih vrednosti $\mathbf{X}$. Najbolj elegantna rešitev iz [4,14] pravi, da matriko $\mathbf{X}$ obravnavamo kot latentne spremenljivke in hkrati uporabimo p neodvisnih GP kot apriorno verjetje latentnih vrednosti preslikave $\mathbf{f}(\mathbf{X}) = (f_1(\mathbf{X}), \ldots, f_p(\mathbf{X}))$ tako, da velja

$$f_i(\mathbf{X}) = \mathcal{GP}(0, k_f(\mathbf{X}, \mathbf{X}')), \quad i = 1, \ldots, p. \tag{3.1}$$

Z izbiro nelinearne kovariančne funkcije omogočimo nelinearno redukcijo dimenzije regresijskega problema.

Porazdelitev izhodne vrednosti pri danih vhodnih vrednostih modela zapišemo

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}) &= \int p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X})\mathrm{d}\mathbf{F} \\ &= \int \prod_{j=1}^{p}\prod_{i=1}^{n} p(y_{i,j}|f_{i,j}) \prod_{j=1}^{p} p(\mathbf{f}_j|\mathbf{X})\mathrm{d}\mathbf{F} \\ &= \prod_{j=1}^{p} \mathcal{N}(\mathbf{y}_j|\mathbf{0}, \mathbf{K}_{ff} + \beta^{-1}\mathbf{I}), \end{aligned} \tag{3.2}$$

za neodvisne izhode

$$y_{i,j} = f_j(\mathbf{x}_i) + \epsilon_{i,j}, \quad \epsilon_{i,j} \sim \mathcal{N}(0, \beta^{-1}). \tag{3.3}$$

Kovariančno matriko v enačbi (3.2) dobimo z izračunom kovariančne funkcije $k_f(\mathbf{X}, \mathbf{X})$ na vseh n učnih podatkih, toda njeni argumenti so v primeru modela GP-LVM neznane latentne spremenljivke. Opomnimo, da so podrejene apriornemu verjetju latentnega prostora $p(\mathbf{X}) \triangleq p(\mathbf{X}|\mathbf{\Theta}_x)$ s hiperparametri $\mathbf{\Theta}_x$. Struktura apriornega verjetja zavisi od dane uporabe modela, npr. ali so meritve neodvisne in enakomerno porazdeljene.

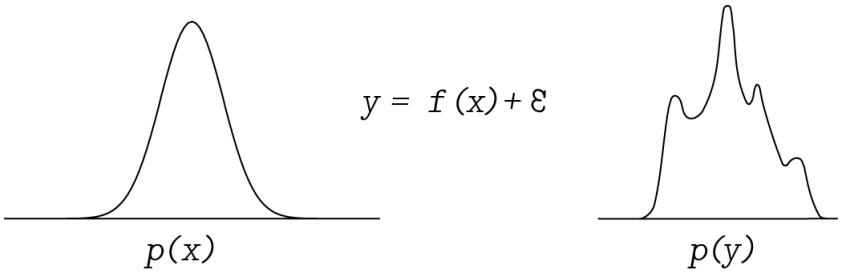
Zaradi recipročne povezanosti $\mathbf{F}$ in $\mathbf{X}$ sta sklepanje na vrednosti izhodov $\mathbf{Y}$ in analitična definicija modela zelo zahtevna [12]. Kljub temu lahko za trenutek matriko $\mathbf{X}$ obravnavamo kot matriko konstant in integriramo po prostoru latentnih vrednosti preslikave $\mathbf{F}$. Tako dobimo robno verjetnost $p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})$, ki pa je analitično izračunljiva. Slednji izraz odpira možnosti učenja modela s postopkom MAP kjer so latentni vhodi $\mathbf{X}$ izbrani tako, da velja

$$\mathbf{X}_{\mathrm{MAP}} = \arg\max_{\mathbf{X}} p(\mathbf{Y}|\mathbf{X})p(\mathbf{X}). \tag{3.4}$$

Za normalno porazdeljeno naključno spremenljivko s povprečjem μ in varianco σ^2 interpretiramo območje $\mu \pm 2\sigma$ kot interval 95 % zaupanja. S tem izražamo negotovost naključne spremenljivke. Širjenje negotovosti povezanih z latentnimi spremenljivkami skozi nelinearno preslikavo pri učenju nelinearnega modela z učnim postopkom MAP je brez aproksimacij nemogoče [4]. Preslikava vhodne verjetnostne porazdelitve z nelinearnim sistemom namreč rezultira v zelo splošno obliko, ki jo je po navadi težko normalizirati. Na Sliki 3.1 je prikazan primer, ko Gaussovo porazdelitev za vrednost x širimo skozi nelinearno preslikavo f. Na desni je kot rezultat preslikave skiciran le en možen primer iz bogatega funkcijskega prostora.

Verjetnostna porazdelitev latentnih spremenljivk $\mathbf{X}$ se v postopku MAP učenja sesede v Diracovo delta - δ funkcijo v trenutku, ko nanj deluje nelinearna preslikava f. Slednje dejstvo močno uteži nadgradnjo ali v splošnem uporabo modela v kompleksnejših modelih [4, 6]. Več o slabostih učenja MAP bomo napisali v poglavju 3.2.

Alternativo MAP učnemu postopku predstavlja integracija po latentnih spremenljivkah $\mathbf{X}$. Postopek je podrobneje opisan v poglavju 3.3.



Slika 3.1: Razširjanje Gaussove porazdelitve skozi nelinearno preslikavo. Slika po viru [4].

3.2 Slabosti MAP učenja modelov

V literaturi pogosto zasledimo, da optimizacija vrednosti vhodnih latentnih spremenljivk in hiperparametrov modela GP-LVM temelji na učnem postopku MAP, a ima tak pristop vrsto pomanjkljivosti. Kot prvo, je rezultat odvisen od vrednosti latentnih spremenljivk. Posledično je model zelo občutljiv na prekomerno prilagajanje podatkom (angl. *overfitting*). In kot drugo, postopek MAP ne omogoča določevanja optimalne dimenzije latentnega prostora [6, 14]. Kar je tudi razlog, zakaj večina algoritmov in literature navaja zahtevo, da dimenzijo latentnega prostora določi uporabnik.

V definiranem modelu GP-LVM iz poglavja 3.1 uporabljamo le preprostejše oblike kovariančnih funkcij - funkcije s konstantno vrednostjo nekaterih hiperparametrov, ne glede na dimenzijo latentnih vhodov. Gaussova kovariančna funkcija iz poglavja 2.3, z večjim številom hiperparametrov, omogoča avtomatsko določanje ustreznosti (angl. *Automatic Relevance Determination, ARD*), kjer so dimenzijam dodeljene uteži w_k, npr. vsaki nepomembni dimenziji je dodeljena utež z vrednostjo blizu 0. Če bi takšno kovariančno funkcijo uporabili v fazi učenja s postopkom MAP, bi model pri majhnem številu podatkov lahko rezultiral v prekomernem prilagajanju. Zato so standardni modeli GP-LVM s postopkom učenja MAP zelo občutljivi na dimenzijo latentnega prostora q in se izogibajo kovariančnim funkcijam z lastnostjo ARD.

Po drugi strani bayesovski pristop omogoča implementacijo *mehkega* modela [4], kjer dimenzijo latentnega prostora q razumemo kot neko »začetno konzervativno oceno«. Dodatna optimizacijska stopnja v modelu nato poišče dimenzije, ki jih lahko zanemarimo in jim pripiše utež blizu ničle. Za napoved izhoda model sicer še vedno upošteva vse dimenzije, a pri tem ustrezno uteži prispevek vsake posebej. V tem primeru začetna vrednost q torej niti ne igra tako pomembne vloge, le da je dovolj velika, da z njo zajamemo vso poglavitno dinamiko opazovanega sistema.

Zgoraj opisani problemi so v največji meri posledice latentnih spremenljivk, ki jih v standardnem GP-modelu ni. Zato tudi učenje MAP standardnega GP-modela iz poglavja 2 v splošnem ni problematično. Pomembna posledica težav povezanih s postopkom MAP učenja modela GP-LVM je, da tak algoritem močno omejuje in upočasnjuje razvoj kompleksnejših modelov, npr. *globokih GP* [4], ki jih bomo opisali v naslednjem poglavju.

Na tem mestu je jasno, da je izpeljava bayesovskega pristopa k reševanju in učenju modela GP-LVM nujna. S tem postane model zanesljivejši in odpira možnosti hierarhičnih

struktur [4,6]. Učenje modela GP-LVM z *variacijsko metodo*, pojasnjeno v naslednjem podpoglavju, se je izkazalo za zelo dobro alternativo učnemu postopku MAP. Variacijska metoda v splošnem ni izpostavljena problemu prekomernega prilagajanja in hkrati omogoča samodejno identifikacijo dimenzije latentnega prostora.

3.3 Variacijski GP-LVM

V učnem postopku MAP iščemo maksimalne vrednosti $p(\mathbf{Y}, \boldsymbol{\Theta})$ glede na hiperparametre $\boldsymbol{\Theta}$, medtem ko bayesovske metode temeljijo na izračunu logaritma robne verjetnosti

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} = \log \int p(\mathbf{Y}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta}. \tag{3.5}$$

Slednji pristop je z matematičnega vidika ugodnejši, saj gre v resnici za povprečenje prek vseh možnih kombinacij parametrov $\boldsymbol{\Theta}$. Izkaže pa se, da za večino praktičnih primerov zgornji integral ni izračunljiv, zato uporabimo enega izmed aproksimacijskih pristopov [4,12]. Le-ti po navadi temeljijo na iskanju spodnje meje funkcionala $\mathcal{F}(q(\boldsymbol{\Theta}))$ v odvisnosti od variacijske porazdelitve $q(\boldsymbol{\Theta})$. Spodnjo mejo poiščemo z uporabo Jensenove neenačbe [13]

$$\log \int f(x) \mathrm{d}x \geq \int \log f(x) \mathrm{d}x, \tag{3.6}$$

za poljubno pozitivno funkcijo $f(x)$. Tedaj je

$$\begin{aligned} \log p(\mathbf{Y}) &= \log \int p(\mathbf{Y}, \boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} \\ &= \log \int q(\boldsymbol{\Theta}) \frac{p(\mathbf{Y}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} \mathrm{d}\boldsymbol{\Theta} \\ &\geq \int q(\boldsymbol{\Theta}) \log \frac{p(\mathbf{Y}, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} \mathrm{d}\boldsymbol{\Theta} \\ &= \mathcal{F}(q(\boldsymbol{\Theta})). \end{aligned} \tag{3.7}$$

Neenačaj spremenimo v enačaj za $q(\boldsymbol{\Theta}) = p(\mathbf{Y}|\boldsymbol{\Theta})$. V primeru, da je posteriorna verjetnost kompleksna (glej Sliko 3.1 desno), vpeljemo dodatne omejitve variacijske porazdelitve $q(\boldsymbol{\Theta})$ in posledično zgolj aproksimiramo posteriorno verjetnostno porazdelitev. Cilj optimizacijske metode je določiti takšen $q(\boldsymbol{\Theta})$, da bo ujemanje s pravo posteriorno verjetnostjo $p(\boldsymbol{\Theta}|\mathbf{Y})$ čim boljše, t.j. $\mathcal{F}(q(\boldsymbol{\Theta})) \to \log p(\mathbf{Y})$. Kako kvaliteten približek za spodnjo mejo robne verjetnostni predstavlja $q(\boldsymbol{\Theta})$ objektivno ovrednotimo s *Kullback-Leiblerjevo* divergenco [4]. Dodatna pojasnila o Kullback-Leiblerjevi divergenci in njeno izpeljavo najdemo v prilogi A.

Standardni bayesovski pristop k izračunu logaritmirane robne verjetnosti (3.5) je težaven [4]. Morali bi namreč integrirati po obeh množicah spremenljivk: vrednostih preslikave $\mathbf{F}$ in latentnih spremenljivkah $\mathbf{X}$

$$\log p(\mathbf{Y}) = \log \int p(\mathbf{Y}, \mathbf{F}, \mathbf{X}) \mathrm{d}\mathbf{X} \mathrm{d}\mathbf{F} = \log \int p(\mathbf{Y}|\mathbf{F}) \Big(p(\mathbf{F}|\mathbf{X}) p(\mathbf{X}) \mathrm{d}\mathbf{X} \Big) \mathrm{d}\mathbf{F}, \tag{3.8}$$

kar pa ni izračunljivo. Poglavitni problem bayesovske metodologije predstavlja širjenje apriorne verjetnostne porazdelitve $p(\mathbf{X})$ skozi nelinearno preslikavo f. Vsak člen verjetnosti $p(\mathbf{F}|\mathbf{X})$ iz integrala (3.8) je namreč proporcionalen

$$|\mathbf{K}_{ff}|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\mathbf{f}_i^\top \mathbf{K}_{ff}^{-1}\mathbf{f}_i\Big) \tag{3.9}$$

kjer se $\mathbf{X}$ v matriki $\mathbf{K}_{ff}$ pojavi nelinearno. Od tod je jasno, da so latentni vhodi $\mathbf{X}$ v integralu (3.8) vključeni na zelo kompleksen način in da integracija po celotni domeni $\mathbf{X}$ skoraj zagotovo ne bo mogoča [12].

V ta namen uporabimo *standardni variacijski pristop* in z njim aproksimiramo robno verjetnost $p(\mathbf{Y})$ [12]. Kar v konkretnem primeru pomeni, da vpeljemo variacijsko porazdelitev naključnih spremenljivk:

$$q(\mathbf{F}, \mathbf{X}) = q(\mathbf{F})q(\mathbf{X}), \tag{3.10}$$

s čimer želimo aproksimirati pravo posteriorno verjetje $p(\mathbf{F}|\mathbf{Y}, \mathbf{X})p(\mathbf{X}|\mathbf{Y})$. Z Jensenovo neenačbo izračunamo oceno za logaritem robne verjetnosti

$$\log p(\mathbf{Y}) \geq \int q(\mathbf{F})q(\mathbf{X}) \log \frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{F})q(\mathbf{X})} \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{F}, \tag{3.11}$$

a je integral še vedno analitično neizračunljiv kot je vidno iz

$$\begin{aligned}\log p(\mathbf{Y}) \geq &\int q(\mathbf{F})q(\mathbf{X}) \log p(\mathbf{F}|\mathbf{X}) \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{F} + \\ &+ \int q(\mathbf{F})q(\mathbf{X}) \log \frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{X})}{q(\mathbf{F})q(\mathbf{X})} \mathrm{d}\mathbf{X}\mathrm{d}\mathbf{F},\end{aligned} \tag{3.12}$$

kjer v prvem integralu integracija po $\mathbf{X}$ še vedno poteka po zapletenih nelinearnih oblikah $\mathbf{X}$ v $\mathbf{K}_{ff}^{-1}$ in $\log|\mathbf{K}_{ff}|$.

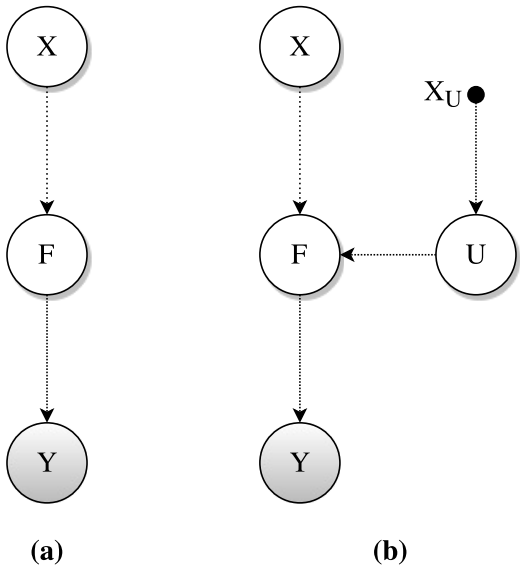
3.3.1 Izračunljiva spodnja meja robne verjetnosti

Da bo spodnja meja $\log p(\mathbf{Y})$ analitično izračunljiva vpeljemo poljubne inducirane spremenljivke [4]. Z njimi razširimo verjetnosti prostor in jih v modelu intepretiramo kot dodatne parametre [20, 21]. Običajno se inducirane spremenljivke uporabijo za dimenzijsko aproksimacijo kovariančne matrike in posledično računsko pohitritev, mi pa jih bomo uporabili za izračun integrala (3.11) [4].

Natančneje, skupno verjetnost napovedi razširimo z m induciranimi spremenljivkami $\mathbf{u}_i \in \mathbb{R}^p$ latentne preslikave $f(\mathbf{x})$. Inducirane točke zberemo v matriki $\mathbf{U} \in \mathbb{R}^{m\times p}$, njihove neodvisne vrednosti pa v matriki psevdovhodov $\mathbf{X}_u \in \mathbb{R}^{m\times q}$. Slika 3.2(b) grafično prikazuje vpliv induciranih točk. Navidezna verjetnostna gostota tedaj privzame obliko

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{U}, \mathbf{X}) = p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{U}, \mathbf{X})p(\mathbf{U})p(\mathbf{X}). \tag{3.13}$$

Za izračun variacijske spodnje meje v navideznem prostoru induciranih spremenljivk,



Slika 3.2: Grafični prikaz modela GP-LVM. Slika (a) prikazuje standardni GP-LVM-model (b) njegovo nadgradnjo v variacijski GP-LVM-model. Slika po viru [4].

vpeljemo še variacijsko porazdelitev $q(\mathbf{U})$. Ker želimo integrirati tudi po latentnem prostoru dodatno vpeljemo še variacijsko porazdelitev $q(\mathbf{X})$. Zaenkrat privzemimo, da je $q(\mathbf{X})$ Gaussova porazdelitev. Za oceno logaritma robne verjetnosti lahko tedaj pišemo

$$\begin{aligned}
\log p(\mathbf{Y}) &= \log \int p(\mathbf{Y}|\mathbf{U},\mathbf{X})p(\mathbf{U})p(\mathbf{X})\mathrm{d}\mathbf{X}\mathrm{d}\mathbf{U} \\
&\geq \int q(\mathbf{X})q(\mathbf{U}) \log \frac{p(\mathbf{Y}|\mathbf{U},\mathbf{X})p(\mathbf{U})p(\mathbf{X})}{q(\mathbf{X})q(\mathbf{U})}\mathrm{d}\mathbf{X}\mathrm{d}\mathbf{U} \\
&= \int q(\mathbf{X})q(\mathbf{U}) \log p(\mathbf{Y}|\mathbf{U},\mathbf{X})\mathrm{d}\mathbf{X}\mathrm{d}\mathbf{U} - \mathrm{KL}(q(\mathbf{U})||p(\mathbf{U})) - \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X})) \\
&\geq \langle \mathcal{L} \rangle_{q(\mathbf{X})q(\mathbf{U})} - \mathrm{KL}(q(\mathbf{U})||p(\mathbf{U})) - \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X})) \\
&= \hat{\mathcal{F}}(q(\mathbf{X}), q(\mathbf{U})) - \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X})) \\
&= \mathcal{F}(q(\mathbf{X}), q(\mathbf{U})),
\end{aligned} \tag{3.14}$$

kjer z $\langle \cdot \rangle_{q(\mathbf{X})q(\mathbf{U})}$ označujemo povprečenje po $q(\mathbf{X})q(\mathbf{U})$.

Definiramo variacijsko porazdelitev

$$q(\mathbf{F},\mathbf{U},\mathbf{X}) = q(\mathbf{X}) \prod_{i=1}^{p} p(\mathbf{f}_i|\mathbf{u}_i,\mathbf{X})q(\mathbf{u}_j), \tag{3.15}$$

ki po uporabi Jensenove neenačbe vodi do

$$\log p(\mathbf{Y}) \geq \int q(\mathbf{F},\mathbf{U},\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U},\mathbf{X})}{q(\mathbf{F},\mathbf{U},\mathbf{X})}\mathrm{d}\mathbf{F}\mathrm{d}\mathbf{U}\mathrm{d}\mathbf{X}, \tag{3.16}$$

pri čemer se z uporabo (3.15) členi $p(\mathbf{f}_i|\mathbf{u}_i,\mathbf{X})$ v enačbi pokrajšajo. Jasno je, da v matematični model vgrajujemo predpostavke o variacijski distribuciji katere namen je

aproksimacija pravega posteriornega verjetja

$$p(\mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Y}) = p(\mathbf{F}|\mathbf{U}, \mathbf{Y}, \mathbf{X})p(\mathbf{U}|\mathbf{Y}, \mathbf{X})p(\mathbf{X}|\mathbf{Y}). \tag{3.17}$$

Izračun logaritma $\langle\mathcal{L}\rangle_{q(\mathbf{X})q(\mathbf{U})}$ iz funkcionala $\mathcal{F}(q(\mathbf{X}), q(\mathbf{U}))$ v (3.14) je sedaj analitičen, a še vedno odvisen od izbrane kovariančne funkcije [4].

Pričakovana vrednost logaritma glede na porazdelitev $q(\mathbf{X})$ je

$$\langle\mathcal{L}\rangle_{q(\mathbf{X})} = \sum_{i=1}^{n}\sum_{j=1}^{p}\langle\mathcal{L}_{i,j}\rangle_{q(\mathbf{X})} \tag{3.18}$$

pri čemer

$$\begin{aligned}\langle\mathcal{L}_{i,j}\rangle_{q(\mathbf{X})} = &-\frac{1}{2}\log(2\pi\beta^{-1}) - \frac{\beta}{2}y_{i,j}^2 + \beta\mathrm{Tr}(y_{i,j}\mathbf{\Psi}_i\mathbf{K}_{uu}^{-1}\mathbf{u}_j) - \\ &-\frac{\beta}{2}\mathrm{Tr}(\mathbf{K}_{uu}^{-1}\mathbf{u}_j\mathbf{u}_j^\top\mathbf{K}_{uu}^{-1}\hat{\mathbf{\Phi}}_i) - \frac{\beta}{2}(\hat{\xi}_i - \mathrm{Tr}(\hat{\mathbf{\Phi}}_i\mathbf{K}_{uu}^{-1})),\end{aligned} \tag{3.19}$$

kjer smo vpeljali statistike

$$\xi = \langle\mathrm{Tr}(\mathbf{K}_{ff})\rangle_{q(\mathbf{X})} = \sum_{i=1}^{n}\hat{\xi}_i, \tag{3.20}$$

$$\mathbf{\Psi} = \langle\mathbf{K}_{fu}\rangle_{q(\mathbf{X})} = \{\mathbf{\Psi}_i\}_{i=1}^{n} \text{ in} \tag{3.21}$$

$$\mathbf{\Phi} = \langle\mathbf{K}_{uf}\mathbf{K}_{fu}\rangle_{q(\mathbf{X})} = \sum_{i=1}^{n}\hat{\mathbf{\Phi}}_i. \tag{3.22}$$

Njihovo povprečje po $q(\mathbf{X})$ je možno izračunati neodvisno in ločeno za vsako robno porazdelitev $q(\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_i, \mathbf{S}_i)$ iz $q(\mathbf{X})$ [4]. Potemtakem velja

$$\hat{\xi}_i = \int k_f(\mathbf{x}_i, \mathbf{x}_i)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_i, \mathbf{S}_i)\mathrm{d}\mathbf{x}_i, \tag{3.23}$$

$$\mathbf{\Psi}_{i,k} = \int k_f(\mathbf{x}_i, (\mathbf{X}_u)_k)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_i, \mathbf{S}_i)\mathrm{d}\mathbf{x}_i \text{ in} \tag{3.24}$$

$$(\hat{\mathbf{\Phi}}_i)_{k,k'} = \int k_f(\mathbf{x}_i, (\mathbf{X}_u)_k)k_f((\mathbf{X}_u)_{k'}, \mathbf{x}_i)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_i, \mathbf{S}_i)\mathrm{d}\mathbf{x}_i. \tag{3.25}$$

Opazimo, da so statistike $\{\mathbf{\Phi}, \mathbf{\Psi}, \xi\}$ v resnici konvolucije kovariančne funkcije k_f z Gaussovo porazdelitvijo in so analitično izračunljive za večino standardnih kovariančnih funkcij, npr. eksponentna kvadratična ali linearna kovariančna funkcija. Z uporabo zgornjih integralskih definicij lahko vrednost logaritma (3.19) prepišemo v

$$\begin{aligned}\langle\mathcal{L}_{i,j}\rangle_{q(\mathbf{X})} = &\,\mathcal{N}(y_{i,j}|\mathbf{\Psi}_i, \mathbf{K}_{uu}^{-1}\mathbf{u}_j, \beta^{-1}) - \\ &-\frac{\beta}{2}\mathrm{Tr}\Big(\mathbf{K}_{uu}^{-1}\mathbf{u}_j\mathbf{u}_j^\top\mathbf{K}_{uu}^{-1}(\hat{\mathbf{\Phi}}_i - \mathbf{\Psi}_i^\top\mathbf{\Psi}_i)\Big) - \frac{\beta}{2}\Big(\hat{\xi}_i - \mathrm{Tr}\big(\hat{\mathbf{\Phi}}_i\mathbf{K}_{uu}^{-1}\big)\Big).\end{aligned} \tag{3.26}$$

K temu dodamo še pričakovano vrednost glede na porazdelitev $q(\mathbf{U})$

$$\begin{aligned}\Big\langle\langle\mathcal{L}_{i,j}\rangle_{q(\mathbf{X})}\Big\rangle_{q(\mathbf{U})} = &-\frac{1}{2}\log(2\pi\beta^{-1}) - \frac{\beta}{2}y_{i,j}^2 + \beta\mathrm{Tr}(y_{i,j}\mathbf{\Psi}_i\mathbf{K}_{uu}^{-1}(\boldsymbol{\mu})_j) - \\ &-\frac{\beta}{2}\mathrm{Tr}\Big(\mathbf{K}_{uu}^{-1}\big((\boldsymbol{\mu})_j(\boldsymbol{\mu})_j^\top + (\mathbf{\Sigma}_u)_j\big)\mathbf{K}_{uu}^{-1}\hat{\mathbf{\Phi}}_i\Big) - \\ &-\frac{\beta}{2}\Big(\hat{\xi}_i - \mathrm{Tr}\big(\hat{\mathbf{\Phi}}_i\mathbf{K}_{uu}^{-1}\big)\Big).\end{aligned} \tag{3.27}$$

S tem lahko končno izračunamo funkcional $\mathcal{F}(q(\mathbf{X}), q(\mathbf{U}))$ iz enačbe (3.14) tako, da zgornji izraz seštejmo po indeksih i in j.

Optimizacija formulacije (3.27) vključuje tako parametre modela $\mathbf{\Theta}$, inducirane točke $\mathbf{X}_u$, kot tudi parametre porazdelitev $q(\mathbf{X})$ in $q(\mathbf{U})$. Veliko število parametrov in njihova medsebojna povezanost se za gradientne metode mnogokrat izkaže za zahtevno nalogo [4].

Efektivno z variacijskim pristopom v modelu GP-LVM dosežemo, da je spodnja meja robne verjetnosti $p(\mathbf{Y})$, oziroma njen logaritem $\log p(\mathbf{Y})$, izračunljiv. S tem je možna tudi propagacija vseh negotovosti povezanih z vhodnimi vrednostmi. Integriranje po $\mathbf{X}$ ni več problematično [4], saj smo namesto $\mathbf{K}_{ff}$, $\mathbf{K}_{fu}$ in $\mathbf{K}_{fu}\mathbf{K}_{uf}$ vpeljali povprečja $\{\mathbf{\Phi}, \mathbf{\Psi}, \xi\}$. Poudarimo, da je variacijski GP-LVM-model omejen na uporabo kovariančnih funkcij, za katere so $\{\mathbf{\Phi}, \mathbf{\Psi}, \xi\}$ statistike izračunljive.

3.3.2 Različna apriorna verjetja variacijskega GP-LVM-modela

S spreminjanjem strukture apriornega verjetja $q(\mathbf{X})$ nad latentnim prostorom obravnavamo številne različice variacijskega GP-LVM-modela. Pri tem je pomembno, da se $p(\mathbf{X})$ v variacijski spodnji meji iz enačbe (3.14) pojavi le v KL členu

$$\log p(\mathbf{Y}) = \hat{\mathcal{F}}(q(\mathbf{X}), q(\mathbf{U})) - \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X})), \tag{3.28}$$

ki je analitično izračunljiv kadar je $p(\mathbf{X})$ Gaussova porazdelitev.

Popolnoma neodvisni podatki

V najpreprostejšem primeru privzamemo, da so spremenljivke latentnega prostora med seboj popolnoma neodvisne. Takšna izbira še zdaleč ni optimalna, saj lahko z apriornim verjetjem v matematični model vnesemo bistveno kompleksnejše informacije o naravi sistema, je pa z računskega stališča preprosta in v večini fizikalnih sistemov smiselna.

Variacijska porazdelitev $q(\mathbf{X})$ za popolnoma neodvisne točke je

$$q(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \mathbf{S}_i), \tag{3.29}$$

za polno matriko $\mathbf{S}_i$. Kullback-Leiblerjeva divergenca je tedaj

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X})) = \frac{1}{2}\sum_{i=1}^{n} \mathrm{Tr}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top + \mathbf{S}_i - \log \mathbf{S}_i) - \frac{nq}{2}, \tag{3.30}$$

kjer oznaka $\log \mathbf{S}_i$ pomeni matriko $\mathbf{S}_i$ z logaritmiranimi elementi.

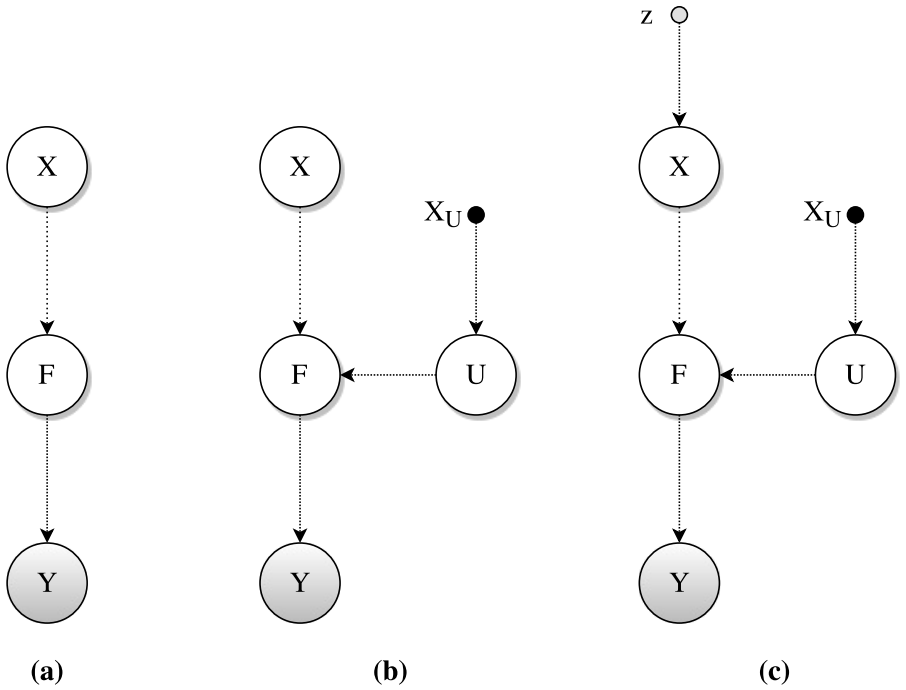
Variacijski GP-LVM-model za nadzorovano učenje

V primeru funkcijskega sistema za nadzorovano učenje obravnavamo množico učnih vrednosti $\mathcal{D} = \{\mathbf{z}_i, y_i\}_{i=1}^n$. Spremenljivke $\mathbf{z}_i$ naj v tem primeru predstavljajo poljubne vhodne vrednosti, odvisno od obravnavanega sistema.

Splošen model GP-LVM vpeljemo preko definicije začasnih latentnih funkcij $\mathbf{x}(\mathbf{z}) = (x_1(\mathbf{z}), \ldots, x_q(\mathbf{z}))$, kjer vsaka komponenta predstavlja neodvisen GP

$$x_j(\mathbf{z}) \sim \mathcal{GP}(0, k_x(\mathbf{z}, \mathbf{z}')), \quad k = 1, \ldots, q, \tag{3.31}$$

za kovariančno funkcijo $k_x(\mathbf{z}, \mathbf{z}')$. Točko y_i dobimo s preslikavo latentnega vektorja $\mathbf{x}_i$, kot je na Sliki 3.3(c) tudi grafično prikazano.



Slika 3.3: Grafični prikaz modela GP-LVM za nadzorovano učenje. Slika (a) prikazuje standardni GP-LVM-model (b) njegovo nadgradnjo v variacijski GP-LVM-model in (c) izpeljanko modela z nadzorovanim učenjem. Zgornji nivo na primeru (c) je v splošnem poljuben, odvisno od obravnavanega sistema. Slika po viru [4].

Vektorje latentnih vrednosti $\mathbf{x}_i$ shranimo v matriko $\mathbf{X}$ in lahko zapišemo

$$p(\mathbf{X}|\mathbf{z}) = \prod_{i=1}^{q} p(\mathbf{x}_i|\mathbf{z}) = \prod_{i=1}^{q} \mathcal{N}(\mathbf{x}_i|\mathbf{0}, \mathbf{K}_x), \tag{3.32}$$

pri čemer kovariančno matriko izračunamo z vrednotenjem kovariančne funkcije na poljubnih vrednostih vhodov $\mathbf{z}$. Kovariančna funkcija k_x ima parametre $\boldsymbol{\Theta}_x$. Le-ti določajo lastnosti vsake začasne funkcije $x_k(\mathbf{z})$, npr. eksponentna kvadratična funkcija vodi v zelo gladek izhod modela [9, 12].

Za vpeljavo variacijskega GP-LVM-modela znova privzamemo neodvisne meritve sistema

$$q(\mathbf{X}) = \prod_{j=1}^{q} \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_j, \mathbf{S}_j), \tag{3.33}$$

za polno $n \times n$ matriko $\mathbf{S}_j$. Spomnimo se oblike pogojne verjetnosti $p(\mathbf{X}|\mathbf{z})$ v enačbi (3.32) in izračunajmo pripadajoč KL člen iz enačbe (3.14)

$$\mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{z})) = \frac{1}{2}\sum_{j=1}^{q}\Big[\mathrm{Tr}(\mathbf{K}_x^{-1}\mathbf{S}_j + \mathbf{K}_x^{-1}\boldsymbol{\mu}_j\boldsymbol{\mu}_j^{\top}) + \log|\mathbf{K}_x| - \log|S_j|\Big] - \frac{nq}{2}. \tag{3.34}$$

3.3.3 Časovna kompleksnost

Velika omejitev standardnih GP-modelov je, da je za končni model potreben izračun inverza kovariančne matrike $\mathbf{K}_{ff}$ [10]. Računska zahtevnost za izračun inverza poljubne $n \times n$ matrike narašča sorazmerno s tretjo potenco števila podatkov n [21]. S tem je uporaba standardnih GP-modelov omejena za modeliranje največ nekaj tisoč podatkov.

Z vpeljavo m induciranih točk efektivno aproksimiramo kovariančno matriko $\mathbf{K}_{ff}$ z matriko manjših dimenzij. Kar za standardne modele GP-LVM pomeni, da zmanjšamo računsko kompleksnost iz n^3 na nm^2 v primeru variacijskega GP-LVM-modela. Po navadi je $m \ll n$ kar po logičnem sosledju pomeni, da lahko z modelom GP-LVM obdelujemo tudi relativno velike množice učnih točk [6].

Dinamični GP-LVM-model pa v enačbi (3.34) še vedno zahteva inverz matrike $\mathbf{K}_x$ velikosti $n \times n$ za izračun robne verjetnosti. Računska kompleksnost n^3 torej ostaja, s čimer je uporaba na ogromnem številu podatkov še vedno otežena.

3.3.4 Napovedovanje

Z variacijskim GP-LVM-modelom bi radi napovedali vrednosti izhoda modela. Z zvezdico označujemo testne količine, npr. matrika testnih podatkov $\mathbf{Y}_* \in \mathbb{R}^{n_* \times p}$. Da napovemo izhod iz GP-modela moramo določiti funkcijo srednje srednje vrednosti in pripadajočo varianco za določitev negotovosti [8]. Vhod v GP-model so posamezne vrednosti neodvisnih spremenljivk, medtem ko je izhod iz GP-modela verjetnostna porazdelitev izhodne vrednosti.

Za uspešno eno koračno napoved najprej pojasnimo kako aproksimirati robno verjetnost $p(\mathbf{Y}_*|\mathbf{Y})$. Z vpeljavo latentnih spremenljivk $\mathbf{X}$, pripadajočimi učnimi vrednostmi izhodov $\mathbf{Y}$ ter testnimi latentnimi spremenljivkami $\mathbf{X}_* \in \mathbb{R}^{n_* \times q}$ pišemo [4]

$$p(\mathbf{Y}_*|\mathbf{Y}) = \frac{p(\mathbf{Y}_*, \mathbf{Y})}{p(\mathbf{Y})} = \frac{\int p(\mathbf{Y}_*, \mathbf{Y}|\mathbf{X}, \mathbf{X}_*)p(\mathbf{X}, \mathbf{X}_*)\mathrm{d}\mathbf{X}\mathrm{d}\mathbf{X}_*}{\int p(\mathbf{Y}|\mathbf{X})p(\mathbf{X})\mathrm{d}\mathbf{X}}. \tag{3.35}$$

Integral v imenovalcu aproksimiramo s spodnjo mejo $e^{\mathcal{F}(q(\mathbf{X}))}$, pri čemer je $\mathcal{F}(q(\mathbf{X}))$ izračunljiva variacijska spodnja meja kot smo jo izpeljali v poglavju 3.3.1. Maksimizacija spodnje meje določa variacijsko porazdelitev $q(\mathbf{X})$ glede na latentne spremenljivke

učnih točk $\mathbf{X}$. Nato z $e^{\mathcal{F}(q(\mathbf{X},\mathbf{X}_*))}$ aproksimiramo še spodnjo mejo števca iz enačbe (3.35).

Uporabimo torej aproksimacijo

$$p(\mathbf{Y}_*|\mathbf{Y}) \approx e^{\mathcal{F}(q(\mathbf{X},\mathbf{X}_*))-\mathcal{F}(q(\mathbf{X}))}. \tag{3.36}$$

Obravnavajmo primer, ko imamo delno opazovane testne podakte $\mathbf{Y}_* = (\mathbf{Y}_*^u, \mathbf{Y}_*^o)$ in bi radi rekonstruirali manjkajoči del $\mathbf{Y}_*^u$. Verjetnostna porazdelitev napovedi je tedaj $p(\mathbf{Y}_*^u|\mathbf{Y}_*^o, \mathbf{Y})$. Da jo izračunamo, najprej vpeljemo vrednosti latentnih funkcij $\mathbf{F}_*^u$ (torej vrednosti $\mathbf{Y}_*^u$ brez dodanega šuma) in latentne spremenljivke $\mathbf{X}_*$, ter zapišemo

$$p(\mathbf{Y}_*^u|\mathbf{Y}_*^o, \mathbf{Y}) = \int p(\mathbf{Y}_*^u|\mathbf{F}_*^u)p(\mathbf{F}_*^u|\mathbf{X}_*, \mathbf{Y}_*^o, \mathbf{Y})p(\mathbf{X}_*|\mathbf{Y}_*^o, \mathbf{Y})\mathrm{d}\mathbf{F}_*^u\mathrm{d}\mathbf{X}_*. \tag{3.37}$$

Uporabimo aproksimacijo variacijske porazdelitve

$$p(\mathbf{Y}_*^u|\mathbf{Y}_*^o, \mathbf{Y}) \approx q(\mathbf{Y}_*^u|\mathbf{Y}_*^o, \mathbf{Y}) = \int p(\mathbf{Y}_*^u|\mathbf{F}_*^u)q(\mathbf{F}_*^u, \mathbf{X}_*)q(\mathbf{X}_*)\mathrm{d}\mathbf{F}_*^u\mathrm{d}\mathbf{X}_*, \tag{3.38}$$

na osnovi česar želimo napovedati $\mathbf{Y}_*^u$ tako, da ocenimo povprečno vrednost $\mathbb{E}(\mathbf{Y}_*^u)$ in pripadajočo kovarianco. Prvi člen zgornjega integrala prihaja iz Gaussove verjetnostne porazdelitve, ostala dva člena določimo z uporabo variacijske metodologije in optimiziramo variacijsko spodnjo mejo za $\log p(\mathbf{Y}_*^o, \mathbf{Y})$

$$\begin{aligned}\log p(\mathbf{Y}_*^o, \mathbf{Y}) &= \log \int p(\mathbf{Y}_*^o, \mathbf{Y}|\mathbf{X}_*, \mathbf{X})p(\mathbf{X}_*, \mathbf{X})\mathrm{d}\mathbf{X}_*\mathrm{d}\mathbf{X} \\ &= \log \int p(\mathbf{Y}^u|\mathbf{X})p(\mathbf{Y}_*^o, \mathbf{Y}^o|\mathbf{X}_*, \mathbf{X})p(\mathbf{X}_*, \mathbf{X})\mathrm{d}\mathbf{X}_*\mathrm{d}\mathbf{X},\end{aligned} \tag{3.39}$$

kar z znanjem iz poglavja 3.3 brez težav izračunamo prek funkcionala $\mathcal{F}(q(\mathbf{X}, \mathbf{X}_*))$.

Z integracijo po latentnih spremenljivkah dobimo kot rezultat večdimenzionalno porazdelitev. K sreči, lahko kljub temu statistične momente $\{\mathbf{\Phi}, \mathbf{\Psi}, \xi\}$ izračunamo analitično, če je le kovariančna funkcija primerno izbrana [4]. Tako je

$$\mathbb{E}(\mathbf{f}_{i,*}^u) = \mathbf{B}^\top \boldsymbol{\psi}_* \text{ in} \tag{3.40}$$

$$\mathrm{Cov}(\mathbf{f}_{i,*}^u) = \mathbf{B}^\top(\mathbf{\Phi}_* - \boldsymbol{\psi}_*\boldsymbol{\psi}_*^\top)\mathbf{B} + \xi_*\mathbf{I} - \mathrm{Tr}\big((\mathbf{K}_{uu}^{-1} - (\mathbf{K}_{uu} + \beta\mathbf{\Phi})^{-1})\mathbf{\Phi}_*\big)\mathbf{I}, \tag{3.41}$$

kjer smo definirali statistike

$$\begin{aligned}\xi_* &= \langle k_f(\mathbf{x}_*, \mathbf{x}_*)\rangle\,, \\ \boldsymbol{\psi}_* &= \langle \mathbf{K}_{u*}\rangle \text{ in} \\ \mathbf{\Phi}_* &= \big\langle \mathbf{K}_{u*}\mathbf{K}_{u*}^\top\big\rangle\end{aligned}$$

in matriko $\mathbf{B} = \beta(\mathbf{K}_{uu} + \beta\mathbf{\Phi})^{-1}\boldsymbol{\psi}\mathbf{y}_j$.

Navedimo še, da po vzoru enačbe (2.5) velja

$$\mathbb{E}(\mathbf{y}_{i,*}^u) = \mathbb{E}(\mathbf{f}_{i,*}^u) \text{ in} \tag{3.42}$$

$$\mathrm{Cov}(\mathbf{y}_{i,*}^u) = \mathrm{Cov}(\mathbf{f}_{i,*}^u) + \beta^{-1}\mathbf{I} \tag{3.43}$$

saj je $\mathbf{F}$ le $\mathbf{Y}$ brez upoštevanega šuma.

Napoved dinamičnih sistemov

Variacijska aproksimacija dinamičnih sistemov prav tako temelji na izračunu funkcionala $\mathcal{F}(q(\mathbf{X}, \mathbf{X}_*))$. Toda v dinamičnih sistemih kjer so vhodi $(\mathbf{X}, \mathbf{X}_*)$ apriori korelirani, faktorizacija variacijske porazdelitve $q(\mathbf{X}, \mathbf{X}_*)$ ni mogoča. S tem je optimizacija dinamičnih sistemov računsko bistveno bolj zapletena saj moramo optimizirati kar $2(n + n_*)q$ parametrov [4]. Problem je sicer podoben kot v primeru (3.37), le da je množica opazovanih meritev prazna. Sledi torej

$$p(\mathbf{Y}_*|\mathbf{Y}) \approx \int p(\mathbf{Y}_*|\mathbf{F}_*)q(\mathbf{F}_*|\mathbf{X}_*)q(\mathbf{X}_*)\mathrm{d}\mathbf{X}_*\mathrm{d}\mathbf{F}_*. \tag{3.44}$$

Postopek dalje je enak, le da za izračun $q(\mathbf{X}_*)$

$$q(\mathbf{X}_*) = \int p(\mathbf{X}_*|\mathbf{X})q(\mathbf{X})\mathrm{d}\mathbf{X}, \tag{3.45}$$

ni potrebna optimizacija, saj so členi $p(\mathbf{x}_{*,i}|\mathbf{X}_j)$ določeni iz apriornega verjetja [12]. Ker je $q(\mathbf{X})$ gaussovsko porazdeljen, je tudi zgornji izraz gaussovsko porazdeljen s povprečjem in varianco

$$\mu_{x_*,i} = \mathbf{K}_{*f}\bar{\mu}_j \text{ in} \tag{3.46}$$
$$\sigma^2_{x_*,i} = \mathbf{K}_{**} - \mathbf{K}_{*f}(\mathbf{K}_x + \lambda_j^{-1})^{-1}\mathbf{K}_{f*}, \tag{3.47}$$

za $\mathbf{K}_{*f} = k_x(\mathbf{t}_*, \mathbf{t})$, $\mathbf{K}_{*f} = \mathbf{K}_{*f}^\top$ in $\mathbf{K}_{**} = k_x(\mathbf{t}_*, \mathbf{t}_*)$.

3.3.5 Nadzorovano učenje

Za nadzorovano učenje modela potrebujemo dve orodji: variacijski pristop iz poglavja 3.3.1 in *variacijsko omejitev*, ki delno omeji porazdelitev vhodnega prostora glede na dane izhodne vrednosti.

V večini fizikalnih uporabah GP-modelov imamo na voljo vhodno-izhodne vrednosti $\mathbf{Z}$ in $\mathbf{Y}$. Iščemo vrednosti izhodov $\mathbf{Y}_*$ za neke nove vhodne vrednosti iz matrike $\mathbf{Z}_*$. Upoštevamo še negotovost vhodnih vrednosti in dodamo šum

$$y_{i,j} = \mathbf{f}_j(\mathbf{x}_i) + (\epsilon_f)_{i,j}, \quad (\epsilon_f)_{i,j} \sim \mathcal{N}(0, \beta^{-1}) \tag{3.48}$$
$$\mathbf{x}_i = \mathbf{z}_i + (\boldsymbol{\epsilon}_x)_i, \quad (\boldsymbol{\epsilon}_x)_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_x), \tag{3.49}$$

kjer so vektorji $\{\mathbf{z}_i\}_{i=1}^n$ vhodi iz matrike $\mathbf{Z} \in \mathbb{R}^{n\times q}$, medtem ko so $\{\mathbf{x}_i\}_{i=1}^n$ latentne spremenljivke. Negotovost vhodnih podatkov lahko upoštevamo na vsaj dva načina [4].

Lahko direktno definiramo apriorno porazdelitev latentnih vhodov

$$p(\mathbf{X}|\mathbf{Z}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\mathbf{z}_i, \boldsymbol{\Sigma}_x), \tag{3.50}$$

ki jo upoštevamo tako, da v enačbi (3.14) popravimo funkcional

$$\mathcal{F} = \langle \log p(\mathbf{Y}|\mathbf{X})\rangle_{q(\mathbf{X})} - \mathrm{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Z})), \tag{3.51}$$

pri čemer KL-člen ostaja analitično izračunljiv. Problem pa je, da je potrebno s takim pristopom oceniti kovariančno matriko šuma $\mathbf{\Sigma}_x$. V model torej vpeljemo še dodane parametre in kompleksnost, zato se raje odločimo za alternativen pristop.

Namesto vgraditve negotovosti v model preko latentnega prostora, definiramo variacijsko omejitev. Z njo negotovosti upoštevamo direktno na aproksimaciji posteriornega verjetja $q(\mathbf{X}|\mathbf{Z})$. Tako v model ne vpeljemo dodatnih parametrov. Tipično je

$$q(\mathbf{X}|\mathbf{Z}) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\mathbf{z}_i, \mathbf{S}_i), \tag{3.52}$$

funkcional $\mathcal{F}$, ki omogoča uporabo modela variacijskega GP-LVM pa

$$\mathcal{F} = \langle \log p(\mathbf{Y}|\mathbf{X}) \rangle_{q(\mathbf{X}|\mathbf{Z})} - \mathrm{KL}(q(\mathbf{X}|\mathbf{Z})||p(\mathbf{X})). \tag{3.53}$$

Od tod dalje je optimizacija modela enaka kot v primeru nenadzorovanega učenja.

4 Globoko učenje

Izumitelji so vrsto let sanjali o napravi, ki bi posnemala in nadgrajevala človeško mišljenje, inteligenco. S prvim računalnikom smo naredili tudi prvi resen korak k *umetni inteligenci* [22]. Razvoj umetne inteligence je bil od samega začetka zelo hiter. Njena uporaba na praktičnih primerih v industriji je hitro pokazala svoje prednosti pred tehnologijo tistega časa. Za človeka navidezno zapleteni problemi so z razvojem umetne inteligence postajali vse lažje rešljivi.

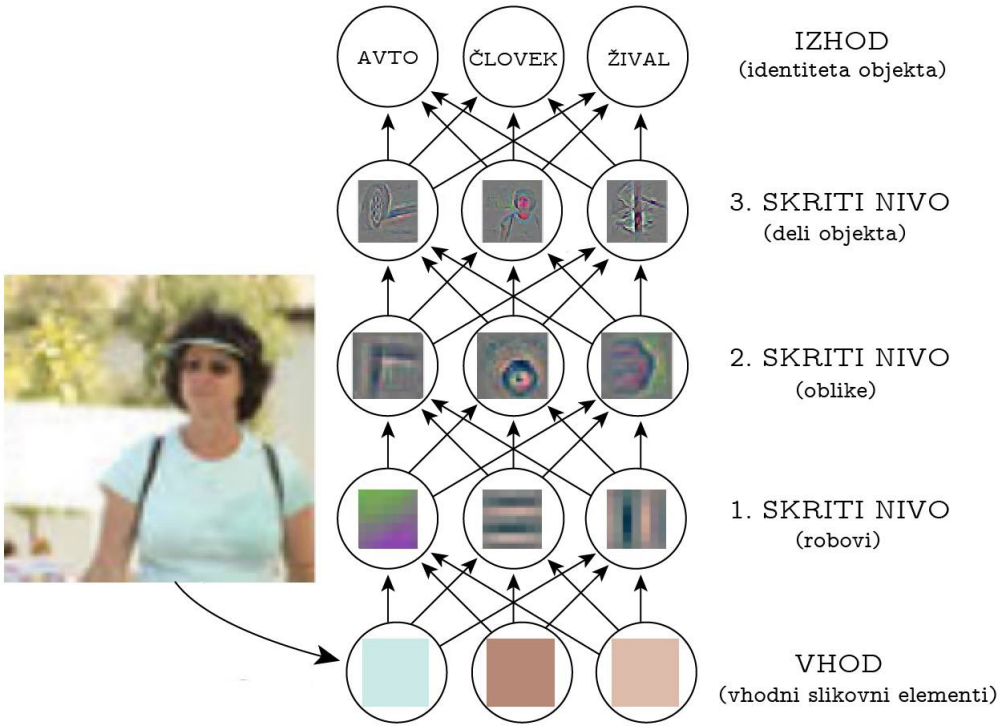
Uporaba umetne inteligence je poskrbela za velik napredek na področju reševanja problemov, ki jih človek zlahka reši, ne zna pa jih pojasniti, npr. intuicija, prepoznavanje besed, obrazov, slik, itd. Z umetno inteligenco smo računalnikom omogočili, da se učijo iz lastnih izkušenj in razumejo svet kot hierarhično strukturo preprostejših, osnovnih, konceptov. Z učenjem iz lastnih izkušenj se izognemo mnogim težavam, kot na primer vprašanju kako formalno definirati človeško intuicijo ali inteligenco in ju prenesti na računalniški čip. Koncept hierarhije računalniku poenostavi učenje in sestavljanje kompleksnih modelov iz preprostejših, osnovnih, struktur. Grafično si hierarhični koncept predstavljamo kot številne *skrite sloje* naložene drug nad drugim, kot prikazuje Slika 4.1. Pravimo, da je tak graf *globok*. Od tod tudi posebna veja umetne inteligence *globoko učenje* [22].

Globoki GP so relativno nov koncept pristopa k umetni inteligenci. Prvo formulacijo iz leta 2013 najdemo v [5] in dve leti kasneje rigorozno formulacijo v doktorski disertaciji [4]. Pristop se osredotoča na variacijski izračun spodnje meje robnega verjetja. S tem postane model sicer aproksimativen, a analitično izračunljiv. Do danes so se razvile že druge izpeljanke variacijske aproksimacije. Na primer v doktorski disertaciji [23] je namesto variacijskega pristopa uporabljen algoritem vzorčenja. Spet v članku [24] za nadzorovano učenje uporabljajo drugačen tip aproksimacije spodnje vrednosti verjetja, ki naj bi bil za primer regresivskega modela nekoliko robustnejši. V delu [25] so uporabili tudi računsko zahtevnejši model globokih GP v kombinaciji z MAP učnim postopkom modela. V nadaljevanju se bomo dotaknili le variacijske izpeljave globokih GP pri čemer se v veliki meri opiramo na disertacijo [4] in članek [25].

Oznake

Ker želimo v nadaljevanju magistrskega dela vse spremenljivke med seboj ločiti glede na pripadajoči sloj v hierarhični strukturi, uvedemo nov matematični zapis. Privzemimo matematični zapis iz [4] in uvedimo pravilo, da spodnji indeksi identificirajo sloj hierarhične strukture, npr. $\mathbf{h}_l$ je vektor skritih spremenljivk l-tega sloja. Ostale

elemente znotraj istega nivoja označujemo enako kot prej, le da so njihovi indeksi v oklepaju zgoraj, npr. $h_l^{(i)}$ označuje i-ti element vektorja $\mathbf{h}_l \in \mathbb{R}^{n_l}$.



Slika 4.1: Struktura globokega učenja na primeru razpoznavanja vsebine slike. Slika po viru [22].

4.1 Globoki Gaussovi procesi

V naslednji poglavjih obravnavamo *gnezdenje GP-modelov* in s tem zaradi podobnosti z globokim učenjem uvedemo pojem globokih GP (angl. *Deep Gaussian Processes*).

V globokih GP opazujemo izhodne vrednosti GP-modela, ki za vhodne vrednosti prejme izhode drugega GP-modela. Celotnega procesa v splošnem ne interpretiramo več kot model GP. Z rekurzijo take strukture tvorimo poljubno število nivojev globokega modela.

Čeprav je bayesovsko sklepanje standardnih GP-modelov iz poglavja 2 analitično izračunljivo, pa to v primeru globokih GP v splošnem več ne velja [25]. Razen v primeru, ko vhodne vrednosti vsakega nivoja hierarhične strukture interpretiramo kot latentne spremenljivke. Tedaj lahko uporabimo znanje prejšnjih poglavij (glej poglavje 3.3) in v poljubno globokih strukturah preprosto integriramo prek latentnih spremenljivk. Na tak način formuliramo globok neparametričen GP-model za nenadzorovano učenje. Za prehod na nadzorovano učenje le dodatno omejimo hierarhično strukturo v skladu z opazovanimi vrednostmi vhodov. Te v matematični model vgradimo na enak način kot v poglavju 3.3.

4.1.1 Definicija

Za izpeljavo modela globokih GP izhajamo iz splošne oblike gnezdenja nevronskih mrež [4,22]. Vsaka preslikava modela je neodvisen GP kjer funkcije obravnavamo z ustreznim verjetnostnim modelom, t.j. po funkcijskem prostoru integriramo in ne poskušamo optimizirati določene oblike preslikave f. Posledica integracije po funkcijskem prostoru je, da je kljub večjemu številu nivojev, število parametrov še vedno bistveno manjše kot na primer v modelih nevronskih mrež [4].

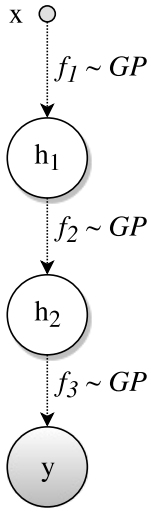
Posplošitev GP-modela na domeno globokih modelov dosežemo s kompozitumom poljubnega števila funkcij

$$\begin{aligned}\mathbf{y} &= \mathbf{f}_{1:L} + \boldsymbol{\epsilon} \\ &= \mathbf{f}_L(\mathbf{f}_{L-1}(\ldots \mathbf{f}_1(\mathbf{x}))) + \boldsymbol{\epsilon}.\end{aligned} \tag{4.1}$$

V enačbi je $\boldsymbol{\epsilon}$ Gaussov šum, ki ga prištejemo v vsakem sloju hierarhije. Tedaj zašumljene vrednosti funkcije zberemo v vektor spremenljivk nivoja $\mathbf{h}_l$ in zapišemo rekurzivno zvezo

$$\mathbf{h}_l = \mathbf{f}_l(\mathbf{h}_{l-1}) + \boldsymbol{\epsilon}_l. \tag{4.2}$$

Z dodatno vpeljavo $\mathbf{y} \triangleq \mathbf{h}_{L+1}$ in $\mathbf{x} \triangleq \mathbf{h}_0$ je rekurzivna definicija globokih GP popolna. Na Sliki 4.2 je grafičen prikaz modela globokih GP z dvema skritima nivojema.



Slika 4.2: Globoki GP z dvema skritima nivojema. Slika po viru [4].

Sklepanje v modelu globokih GP z L prostori vhodnih spremenljivk $\{\mathbf{h}_l\}_{l=1}^{L}$ je v splošnem težavno. Vrednosti spremenljivk vhodnega prostora obravnavamo kot latentne spremenljivke in bi po njih radi integrirali. Žal integracija ni analitično izračunljiva. V ta namen je bil predstavljen hierarhični GP-LVM-model [26].

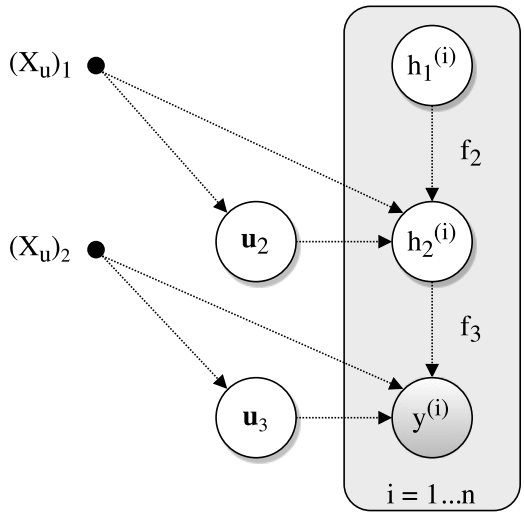
Večina globokih algoritmov potrebuje ogromno količino podatkov za učenje [22]. Prilagajanje zapletenih, globokih modelov na malem številu podatkov se zdi povsem odvečna komplikacija, če v zakup vzamemo računsko kompleksnost in vrsto potrebnih aproksimacij, da je model analitično sploh izračunljiv.

Kadar uporabljamo večslojno strukturo se pojavi vprašanje, ali v splošnem večje število slojev vedno vodi do boljšega modela. Če za trenutek pozabimo problem prekomernega prilagajanja in eksplozijo parametrov se izkaže, da kljub temu zelo globokih modelov v praksi ne uporabljamo, saj običajno vodijo do iskanja izjem v procesu in ne splošne preslikave procesa [4]. Ali drugače povedano, izkaže se, da zelo globoki modeli svojo reprezentacijsko moč osredotočijo na zelo majhnem vzorcu vhodnih podatkov. K sreči nam bayesovski pristop na nek način tudi onemogoča, da bi model iskal prezapletene strukture [25].

Računska kompleksnost nLm^2 globokih GP narašča linearno s številom nivojev in skritih dimenzij in je skoraj neodvisna od dimenzije izhodnih vrednosti p [24]. S tem postane model globokih GP zelo privlačen za obdelavo ogromnega števila podatkov [25]. Odvisnost od števila vhodnih podatkov je enaka kot v primeru modela GP-LVM. To pomeni n^3 v primeru nadzorovanega učenja in nm^2 za nenadzorovano učenje (glej poglavje 3.3.3).

4.1.1.1 Nenadzorovano učenje

Obravnavajmo primer modela globokih GP. Model naj ima L skritih slojev z latentnimi spremenljivkami $\{\mathbf{h}_l\}_{l=1}^{L}$ in opazovanimi izhodi $\mathbf{Y}$. Vmesno vozlišče l-tega nivoja, $\mathbf{h}_l \in \mathbb{R}^n$, je sestavljeno iz izhodov tega nivoja in deluje kot vhod naslednjemu skritemu nivoju, $l+1$. Sloji so medsebojno povezani s preslikavami kot prikazuje Slika 4.3. Vsaka preslikava predstavlja neodvisen GP, s kovariančno funkcijo k_l in hiperparametri $\mathbf{\Theta}_l$. Vsak set latentnih spremenljivk $\mathbf{h}_l$ dobimo kot izhod predhodnega nivoja z dodatkom Gaussovega šuma $\boldsymbol{\epsilon}_j \sim \mathcal{N}(\mathbf{0}, \beta_l^{-1}\mathbf{I})$ po enačbi (4.2). Posamezen nivo hierarhične zgradbe modela globokih Gaussovih procesov torej predstavlja model GP-LVM.



Slika 4.3: Grafični prikaz globokih GP za primer nenadzorovanega učenja. Model ima $L = 2$ skrita nivoja. Vodoravno poravnana vozlišča so del istega nivoja. Slika povzeta po viru [4]

Skupno verjetnostno porazdelitev modela globokih GP z L skritimi nivoji zapišemo [4]

$$p(\mathbf{y}, \{\mathbf{h}_l\}_{l=1}^{L}) = p(\mathbf{y}|\mathbf{h}_L)p(\mathbf{h}_L|\mathbf{h}_{L-1}) \cdots p(\mathbf{h}_2|\mathbf{h}_1)p(\mathbf{h}_1), \tag{4.3}$$

za $p(\mathbf{h}_1) = \mathcal{N}(\mathbf{h}_1|\mathbf{0}, \mathbf{I})$ in pogojne verjetnosti

$$p(\mathbf{h}_l|\mathbf{h}_{l-1}) = \int p(\mathbf{h}_l|\mathbf{f}_l)p(\mathbf{f}_l|\mathbf{h}_{l-1})\mathrm{d}\mathbf{f}_l. \tag{4.4}$$

Opazovane vrednosti izhodov upoštevamo z rekurzivno zvezo $\mathbf{h}_{L+1} \triangleq \mathbf{y}$. Velja torej [25]

$$\begin{aligned} p(\mathbf{h}_1|\mathbf{x}) &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{h_1h_1} + \sigma_1^2\mathbf{I}), \\ p(\mathbf{h}_i|\mathbf{h}_{i-1}) &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{h_ih_i} + \sigma_i^2\mathbf{I}) \text{ in} \\ p(\mathbf{y}|\mathbf{h}_{l-1}) &\sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{f_lf_l} + \sigma_l^2\mathbf{I}). \end{aligned}$$

Pravi izziv predstavlja učenje modela globokih GP, še posebej če za to uporabljamo postopek učenja MAP [26]. Že intuitivno se zdi, da bo postopek učenja MAP problematičen saj bo vsa skrita vozlišča $\{\mathbf{h}_l\}_{l=1}^{L}$ interpretiral kot dodatne parametre modela. V nadaljevanju se zato raje poslužujemo bayesovskega pristopa in integriramo po vseh latentnih spremenljivkah. S tem se drastično zmanjša število parametrov saj vsak nov sloj doprinese le variacijske parametre in ne tudi parametrov modela [4]. Dodatno lahko z uporabo ARD lastnosti kovariančne funkcije avtomatsko sklepamo na dimenzionalnost strukture dinamičnega sistema.

Kljub bayesovskem pristopu so nelinearne kovariančne funkcije še vedno problematične. Za večjo nazornost težave vzemimo model globokih GP z dvema skritima nivojema. Robna verjetnost takšnega modela je

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{h}_2)\Big(\int p(\mathbf{h}_2|\mathbf{h}_1)p(\mathbf{h}_1)\mathrm{d}\mathbf{h}_1\Big)\mathrm{d}\mathbf{h}_2. \tag{4.5}$$

Integral po prostoru spremenljivk $\mathbf{h}_1$ lahko z definicijo pogojnih verjetnosti (4.4) zapišemo v obliki

$$p(\mathbf{h}_2) = \int p(\mathbf{h}_2|\mathbf{f}_2)p(\mathbf{f}_2|\mathbf{h}_1)p(\mathbf{h}_1)\mathrm{d}\mathbf{h}_1\mathbf{f}_2 = \langle p(\mathbf{h}_2|\mathbf{h}_1)\rangle_{p(\mathbf{h}_1)}, \tag{4.6}$$

kar ni izračunljivo, saj zaradi nelinearne kovariančne funkcije $k_1(\mathbf{h}_1, \mathbf{h}_1)$, verjetnostne gostote $p(\mathbf{h}_1)$ ne moremo razširiti skozi funkcijo $p(\mathbf{f}_2|\mathbf{h}_1)$. Do podobne situacije pridemo na vsakem nivoju globokih GP [4]. Problem zapišemo z rekurzivno zvezo

$$\widetilde{p}(\mathbf{h}_l) = \langle p(\mathbf{h}_l|\mathbf{h}_{l-1})\rangle_{\widetilde{p}(\mathbf{h}_{l-1})}. \tag{4.7}$$

Z zgornjo težavo analitične neizračunljivosti se, podobno kot v poglavju 3.3, spopademo z variacijskim pristopom.

4.1.2 Variacijsko sklepanje znotraj nivoja globokih GP

Obravnavajmo neizračunljiv integral (4.6). Podobno kot v modelu variacijskega GP-LVM iz poglavja 3.3 vpeljemo navidezen verjetnostni prostor za vsak nivo. To pomeni, da v model dodamo poljubne spremenljivke $(\mathbf{x}_u)_l$ in pripadajoče vrednosti $\mathbf{u}_l$.

Vpeljava novih spremenljivk modela razširi vsak člen $p(\mathbf{h}_l|\mathbf{h}_{l-1})$ iz robne verjetnosti (4.3) v $p(\mathbf{h}_l|\mathbf{h}_{l-1}, \mathbf{u}_l)$. Le-ti so sedaj dodatno odvisni od vrednosti spremenljivk $\mathbf{u}_l = (u_l^{(1)}, \ldots, u_l^{(m_l)})$ in njihovih neodvisnih psevdo-vhodov $(\mathbf{x}_u)_{l-1} = ((x_u)_{l-1}^{(1)}, \ldots, (x_u)_{l-1}^{(m_1)})$, za $l = 2, \ldots, L+1$. Slika 4.3 grafično prikazuje zgradbo takega modela. V nadaljevanju obravnave poenostavimo, da je število induciranih točk m_l na vseh nivojih enako.

Pogojno verjetnost (4.4) sedaj zapišemo

$$p(\mathbf{h}_l|\mathbf{u}_l, \mathbf{h}_{l-1}) = \int p(\mathbf{h}_l|\mathbf{f}_l)p(\mathbf{f}_l|\mathbf{u}_l, \mathbf{h}_{l-1})\mathrm{d}\mathbf{f}_l, \tag{4.8}$$

kjer je $p(\mathbf{h}_l|\mathbf{f}_l) = \mathcal{N}(\mathbf{h}_l|\mathbf{f}_l, \beta_l^{-1}\mathbf{I})$. Od tod

$$p(\mathbf{h}_l|\mathbf{u}_l, \mathbf{h}_{l-1}) = \mathcal{N}(\mathbf{f}_l|\mathbf{a}_l, \widetilde{\mathbf{K}}_l), \tag{4.9}$$

za definirana

$$\mathbf{a}_l = \mathbf{K}_{f_{l-1}u_{l-1}}\mathbf{K}_{u_{l-1}u_{l-1}}^{-1}\mathbf{u}_l \text{ in} \tag{4.10}$$

$$\widetilde{\mathbf{K}}_l = \mathbf{K}_{f_{l-1}f_{l-1}} - \mathbf{K}_{f_{l-1}u_{l-1}}\mathbf{K}_{u_{l-1}u_{l-1}}^{-1}\mathbf{K}_{u_{l-1}f_{l-1}}. \tag{4.11}$$

Toda verjetnost iz enačbe (4.9) je še vedno problematična: širjenje verjetnostne gostote $\mathbf{h}_{l-1}$ iz nivoja prej je še vedno nemogoče [4]. Na tem mestu uporabimo enak trik kot pri izpeljavi variacijskega GP-LVM-modela: Robno verjetnost (4.3) logaritmiramo in uporabimo Jensenovo neenačbo

$$\begin{aligned}\log p(\mathbf{y}, \{\mathbf{h}_l\}_{l=1}^L \,|\, \{\mathbf{u}_l\}_{l=2}^{L+1}) &= \log p(\mathbf{y}|\mathbf{h}_L, \mathbf{u}_{L+1}) + \sum_{l=2}^{L} \log p(\mathbf{h}_l|\mathbf{h}_{l-1}, \mathbf{u}_l) + \log p(\mathbf{h}_1) \\ &\geq \Big(\log p(\mathbf{h}_1) + \sum_{l=2}^{L+1} \mathcal{L}_l\Big) = \mathcal{L},\end{aligned} \tag{4.12}$$

pri čemer $\mathcal{L}_{L+1}$ navzdol omejuje $\log p(\mathbf{y}|\mathbf{h}_L, \mathbf{u}_{L+1})$, medtem ko za ostale člene velja $\mathcal{L}_l \leq \log p(\mathbf{h}_l|\mathbf{u}_l, \mathbf{h}_{l-1})$. Tu je

$$\mathcal{L}_l = \log \mathcal{N}(\mathbf{h}_l|\mathbf{a}_l, \beta_l^{-1}\mathbf{I}) - \frac{\beta_l}{2}\mathrm{Tr}(\widetilde{\mathbf{K}}_l). \tag{4.13}$$

Zgornja formulacija omogoča, da v nadaljevanju enačbo (4.12) integriramo po latentnih prostorih $\{\mathbf{h}_l\}_{l=1}^L$ in induciranih izhodnih vrednostih $\{\mathbf{u}_l\}_{l=2}^{L+1}$ ter se s tem znebimo ogromnega števila spremenljivk [4].

4.1.3 Variacijsko sklepanje med nivoji globokega GP

V prejšnjem poglavju smo ocenili vrednost logaritma robne verjetnosti (4.12). V tem poglavju se bomo z integracijo izraza znebili še vseh latentnih spremenljivk $\{\mathbf{h}_l\}_{l=1}^L$ in induciranih vrednosti $\{\mathbf{u}_l\}_{l=2}^{L+1}$. Zopet uporabimo variacijski pristop iz poglavja 3.3 in upoštevamo, da so globoki GP le v slojih zloženi variacijski GP-LVM-modeli.

V matematičnem smislu bi radi aproksimirali logaritem robne verjetnosti

$$\log p(\mathbf{y}) = \log \int p(\mathbf{y}, \{\mathbf{h}_l\} \,|\, \{\mathbf{u}_l\}) \prod_{l=2}^{L+1} p(\mathbf{u}_l) \mathrm{d}\,\{\mathbf{h}_l\}\,\mathrm{d}\,\{\mathbf{u}_l\}\,, \tag{4.14}$$

kjer s krajšim zapisom označujemo $\{\mathbf{h}_l\} = \{\mathbf{h}_l\}_{l=1}^{L}$ in podobno $\{\mathbf{u}_l\} = \{\mathbf{u}_l\}_{l=2}^{L+1}$. Za izračun zgornjega integrala vpeljemo variacijsko porazdelitev [4]

$$\mathcal{Q} = \prod_{l=1}^{L} q(\mathbf{u}_{l+1}) q(\mathbf{h}_l) \tag{4.15}$$

in uporabimo Jensenovo neenakost, da dobimo spodnjo oceno $\mathcal{F} \leq \log p(\mathbf{y})$ (izpeljava v [4])

$$\begin{aligned}\mathcal{F} &= \int \mathcal{Q} \log \frac{p(\mathbf{y}, \{\mathbf{h}_l\} \,|\, \{\mathbf{u}_l\}) \prod_{l=2}^{L+1} p(\mathbf{u}_l)}{\mathcal{Q}} \mathrm{d}\,\{\mathbf{h}_l\}\,\mathrm{d}\,\{\mathbf{u}_l\} \\ &= \underbrace{\langle \log p(\mathbf{y}, \{\mathbf{h}_l\} \,|\, \{\mathbf{u}_l\}) \rangle_{\mathcal{Q}}}_{\langle \mathcal{L} \rangle_{\mathcal{Q}}} - \sum_{l=2}^{L+1} \mathrm{KL}(q(\mathbf{u}_l) || p(\mathbf{u}_l)) + \sum_{l=1}^{L} \mathcal{H}(q(\mathbf{h}_l)).\end{aligned} \tag{4.16}$$

Tu smo z $\mathcal{H}(\cdot)$ označili *Shannonovo informacijsko entropijo* [27], ki meri količino informacije. Bistvo naključnega procesa je njegova negotovost. Izidi so nepredvidljivi, a vsaka nova meritev zmanjša negotovost - nosi nekaj informacije o izidu. Negotovost gotovega dogodka je nič, $\mathcal{H}(p = 1) = 0$. Tudi negotovost nemogočega dogodka je nič, saj v takem dogodku ni nič nejasnega. Vrednost entropije je odvisna izključno od verjetnostne porazdelitve $q(\mathbf{h}_l)$ in je maksimalna, ko smo maksimalno negotovi, t.j. ko so vsi izidi enako verjetni. Več o informacijski entropiji najdemo v delu [27].

Ker so vse verjetnostne porazdelitve $p(\cdot)$ v zgornjem funkcionalu (4.16) normalno porazdeljene, se tudi v primeru variacijskih porazdelitev $q(\mathbf{h}_l)$ in $q(\mathbf{u}_l)$ omejimo na družino Gaussovih funkcij. S tem postanejo vsi členi v zgornji enačbi izračunljivi [4, 25].

Za Gaussovo porazdeljena $q(\mathbf{h}_l)$ in $q(\mathbf{u}_l)$ se variacijska porazdelitev $\mathcal{Q}$ iz enačbe (4.15) prepiše v

$$\begin{aligned}\mathcal{Q} &= q(\{\mathbf{h}_l\}) q(\{\mathbf{u}_l\}) \\ &= \prod_{l=1}^{L} \Big[\mathcal{N}(\mathbf{h}_l | \mathbf{m}_l, \mathbf{S}_l) \mathcal{N}(\mathbf{u}_{l+1} | \boldsymbol{\mu}_{l+1}, \boldsymbol{\Sigma}_{l+1}) \Big] \\ &= \prod_{l=1}^{L} \Big[\mathcal{N}(\mathbf{u}_{l+1} | \boldsymbol{\mu}_{l+1}, \boldsymbol{\Sigma}_{l+1}) \prod_{i=1}^{n} \mathcal{N}(h_l^{(i)} | m_l^{(i)}, s_l^{(i)}) \Big].\end{aligned}$$

za posamezni nivo. Za serijo L nivojev [4] pa je

$$\mathcal{Q} = q(\{\mathbf{H}_l\}) q(\{\mathbf{U}_l\}) = \prod_{l=1}^{L} \Big[\prod_{j=1}^{q_{l+1}} \mathcal{N}(\mathbf{u}_{l+1}^{(j)} | \boldsymbol{\mu}_{l+1}^{(j)}, \boldsymbol{\Sigma}_{l+1}^{(j)}) \prod_{i=1}^{n} \mathcal{N}(\mathbf{h}_l^{(i)} | \mathbf{m}_l^{(i)}, \mathbf{S}_l^{(i)}) \Big], \tag{4.17}$$

pri čemer je $\mathbf{S}_l^{(i)}$ diagonalna $q_l \times q_l$ matrika in $\boldsymbol{\Sigma}_l^{(j)}$ polna $m_l \times m_l$ matrika.

Končno izpeljemo oceno prvega člena iz spodnje meje verjetnosti funkcionala (4.16)

$$\langle \mathcal{L} \rangle_{\mathcal{Q}} = \langle \log p(\mathbf{h}_1) \rangle_{q(\mathbf{h}_1)} + \sum_{l=2}^{L+1} \langle \mathcal{L}_l \rangle_{q(\mathbf{h}_{l-1})q(\mathbf{h}_l)q(\mathbf{u}_l)}. \tag{4.18}$$

Končno oblika funkcionala (4.16) za aproksimacijo spodnje meje strnemo v zapisu [4]

$$\begin{aligned}\mathcal{F} = \sum_{l=2}^{L+1} \Big(\langle \log \mathcal{N}(\mathbf{h}_l | \mathbf{a}_l, \beta_l^{-1}\mathbf{I}) \rangle_{q(\mathbf{h}_{l-1})q(\mathbf{h}_l)q(\mathbf{u}_l)} - \frac{\beta_l}{2} \langle \mathrm{Tr}(\widetilde{\mathbf{K}}_l) \rangle_{q(\mathbf{h}_{l-1})} \Big) - \\ - \mathrm{KL}(q(\mathbf{h}_1)||p(\mathbf{h}_1)) - \sum_{l=2}^{L+1} \mathrm{KL}(q(\mathbf{u}_l)||p(\mathbf{u}_l)) + \sum_{l=2}^{L} \mathcal{H}(q(\mathbf{h}_l)),\end{aligned} \tag{4.19}$$

kjer opomnimo, da je $\mathbf{h}_{L+1} \triangleq \mathbf{y}$. Vsi členi v zgornji aproksimaciji so izračunljivi. Še posebej Kullback-Leiblerjeve divergence in entropija predstavljajo zaradi Gaussovih porazdelitev sila preprost preračun [4]. Konkretno

$$\mathcal{H}(q(\mathbf{h}_l)) = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{g_l} \Big(\log(2\pi) + 1 + \log(\mathbf{S}_l^{(i)})^{(j,j)} \Big), \tag{4.20}$$

če z $(\mathbf{S}_l^{(i)})^{(j,j)}$ označimo j-ti element diagonalne matrike $\mathbf{S}_l^{(i)}$. Še najbolj zahteven je prvi člen, ki pa ga razširimo z upoštevanjem vseh Gaussovih porazdelitev. Tedaj opazimo, da lahko vse pričakovane vrednosti glede na variacijsko porazdelitev $q(\mathbf{h}_l)$ namesto kovariančnih matrik zapišemo s statistikami $\{\mathbf{\Phi}, \mathbf{\Psi}, \xi\}$ za vsak nivo posebej.

$$\begin{aligned}\xi_l &= \langle \mathrm{Tr}(\mathbf{K}_{f_l f_l}) \rangle_{q(\mathbf{H}_l)}, \\ \mathbf{\Psi}_l &= \langle \mathbf{K}_{f_l u_l} \rangle_{q(\mathbf{H}_l)} \text{ in} \\ \mathbf{\Phi}_l &= \langle \mathbf{K}_{f_l u_l} \mathbf{K}_{u_l f_l} \rangle_{q(\mathbf{H}_l)}.\end{aligned}$$

Brez izpeljave (narejena v [4]) navedimo končno obliko funkcionala (4.19), ki upošteva vse nivoje in vpeljane statistike

$$\begin{aligned}\mathcal{F} = \sum_{i=1}^{n} \Big[\sum_{j=1}^{p} \langle \mathcal{L}_{L+1}^{(i,j)} \rangle_{\mathcal{Q}} + \sum_{l=2}^{L} \sum_{j=1}^{g_l} \langle \mathcal{L}_l^{(i,j)} \rangle_{\mathcal{Q}} - \mathrm{KL}(q(\mathbf{h}_1^{(i)})||p(\mathbf{h}_1^{(i)})) + \\ + \frac{1}{2} \sum_{l=1}^{L} \sum_{j=1}^{q_l} \Big(\log(2\pi) + 1 + \log(\mathbf{S}_l^{(i)})^{(j,j)} \Big) \Big] - \sum_{l=1}^{L+1} \sum_{j=1}^{q_l} \mathrm{KL}(q(\mathbf{u}_l^{(j)})||p(\mathbf{u}_l^{(j)})),\end{aligned} \tag{4.21}$$

kjer je

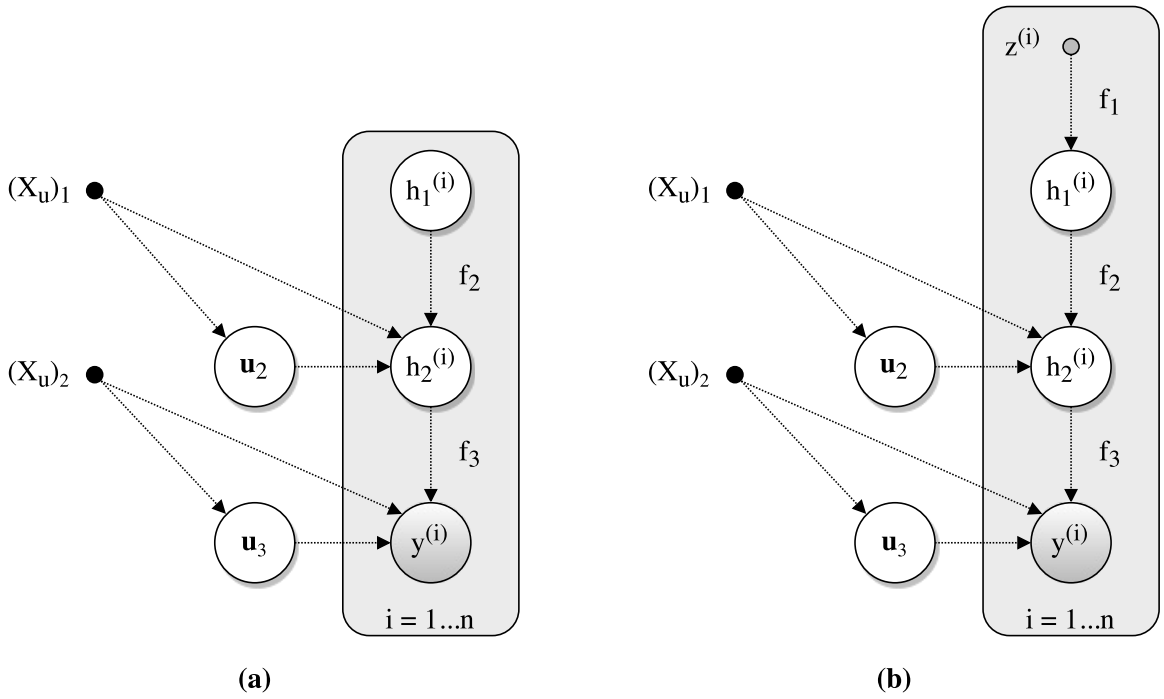
$$\begin{aligned}\langle \mathcal{L}_{l+1}^{(i,j)} \rangle_{\mathcal{Q}} = &- \frac{1}{2} \log(2\pi\beta_{l+1}^{-1}) + \frac{\beta_{l+1}}{2} \mathrm{Tr}\Big((m_{l+1}^{(i,j)})^2 + (\mathbf{S}_{l+1}^{(i)})^{(j,j)} \Big) \\ &- \beta_{l+1} \mathrm{Tr}\Big(m_{l+1}^{(i,j)} \mathbf{\Psi}_l^{(i)} \mathbf{K}_{u_l u_l}^{-1} \boldsymbol{\mu}_{l+1}^{(j)} \Big) - \frac{\beta_{l+1}}{2} \Big(\hat{\xi}_l^{(i)} - \mathrm{Tr}(\mathbf{\Phi}_l^{(i)} \mathbf{K}_{u_l u_l}^{-1}) \Big) \\ &- \frac{\beta_{l+1}}{2} \mathrm{Tr}\Big(\mathbf{K}_{u_l u_l}^{-1} \Big(\boldsymbol{\mu}_{l+1}^{(j)} (\boldsymbol{\mu}_{l+1}^{(j)})^\top + \mathbf{\Sigma}_{l+1} \Big) \mathbf{K}_{u_l u_l}^{-1} \mathbf{\Phi}_l^{(i)} \Big).\end{aligned} \tag{4.22}$$

Analitična izračunljivost spodnje meje (4.21) temelji na enakih pogojih kot izračunljivost variacijskega GP-LVM-modela. Kovariančna funkcija mora biti taka, da je konvolucija z Gaussovo porazdelitvijo $q(\mathbf{h})$ izračunljiva. Za optimizacijo se po navadi uporabljajo gradientne metode [4]. Pri tem določimo

– hiperparametre modela: $\{\beta_l, \mathbf{\Theta}_{f,l}\}_{l=2}^{L+1}$ in
– variacijske parametre $\left\{(\mathbf{x}_u)_l, \mathbf{m}_l, \left\{\text{diag}(\mathbf{S}_l^{(i)}), \boldsymbol{\mu}_{l+1}, \boldsymbol{\Sigma}_{l+1}\right\}_{i=1}^{n}\right\}_{l=1}^{L}$.

4.1.4 Nadzorovano učenje

Možnost, da lahko zapišemo skupno verjetnostno porazdelitev za globoke GP ima veliko prednost. S tem je prehod iz nenadzorovanega učenja v nadzorovano zelo preprost. Grafično je primer nadzorovanega učenja prikazan na Sliki 4.4(b).



Slika 4.4: Grafični prikaz globokih GP. (a) primer nenadzorovanega in (b) primer nadzorovanega učenja z induciranimi spremenljivkami in $L = 2$ skritima nivojema. Vodoravno poravnana vozlišča so del istega nivoja. Slika povzeta po viru [4]

Za primer nadzorovanega učenja le dodamo pogoj na vrhu modela [4, 24]. Odvisnost od izmerjenih vrednosti se nato pojavi v vseh pogojnih verjetnostih, ki vključujejo prvi nivo $\mathbf{h}_1$, npr. $p(\mathbf{h}_1)$ pišemo kot $p(\mathbf{h}_1|\mathbf{x})$. K sreči so vrednosti $\mathbf{x}$ deterministične in njihova propagacija z nelinearno preslikavo f ni problematična. Potemtakem tudi izračun Gaussove porazdelitve $p(\mathbf{h}_1|\mathbf{x})$ ne predstavlja več težave. Če pa je temu tako, so vsi triki navideznega prostora odveč, kot tudi inducirane točke najvišjega sloja.

Čeprav nadzorovan model učenja vpelje dodaten nivo v hierarhični strukturi, pa tega nivoja ne štejemo med skrite sloje, saj je njegova vloga in tudi pomen povsem drugačen. Od tod sledi, da imata oba modela na Sliki 4.4 enako število skritih slojev.

Pri razvoju nadzorovanega učenja je potrebno posebno pozornost nameniti le pravilni definiciji odvisnosti verjetnosti od $\mathbf{x}$ tam, kjer je to potrebno. Ali drugače povedano, za nadzorovano učenje je variacijska distribucija $\mathcal{Q}$ povsem enaka kot v enačbi (4.17),

spremeni se faktorizacija latentnega prostora

$$q(\{\mathbf{H}_l\}) = \prod_{j=1}^{q_1} \mathcal{N}(\mathbf{h}_1^{(j)} | \mathbf{m}_1^{(j)}, \mathbf{S}_1^{(j)}) \prod_{l=2}^{L} \prod_{i=1}^{n} \mathcal{N}(\mathbf{h}_l^{(i)} | \mathbf{m}_l^{(i)}, \mathbf{S}_l^{(i)}). \tag{4.23}$$

Ostali postopek izračune spodnje meje robnega verjetja in napovedi modela iz prejšnjih poglavij ostaja nespremenjen.

5 Ilustrativni primer identifikacije nelinearnega dinamičnega sistema

Teoretično znanje iz prejšnjih poglavij bomo prikazali na primeru identifikacije dinamičnega sistema. Na sintetičnih podatkih bomo preizkusili delovanje modela globokih GP. Rezultate bomo primerjali z modelom GP-LVM iz poglavja 3.

Cilj sintetičnega eksperimenta je na praktičnem primeru prikazati uporabnost regresijskega modela globokih GP. Radi bi identificirali zvezo med vhodnimi in izhodnimi vrednostmi podatkov. Pri tem želimo pravilno napovedati vrednost izhoda en korak vzorčenja podatkov vnaprej, t.j. da je razlika med meritvijo in napovedjo modela pri novih vhodnih podatkih čim manjša. Manjša kot je negotovost povezana z napovedjo modela, bolj smo z rezultati zadovoljni.

Rezultate ovrednotimo na dva načina. Če ni drugače navedeno, grafično prikazujemo ujemanje napovedi modela z meritvami iz testne množice. V ta namen na grafih izrisujemo vrednosti napovedi $\mu(k)$ ter napako $e(k)$ definirano kot absolutno vrednost razlike med napovedanimi in testnimi vrednostmi, npr. Slika 5.4. Pri obeh grafih s sivo pobarvanim območjem označujemo interval 95 % zaupanja, t.j. $\pm 2\sigma$.

Naj poudarimo, da v splošnem opazujemo večdimenzionalni regresijski problem. V veliki meri se za primerjavo rezultatov obeh modelov oziramo na dvodimenzionalne grafe, npr. Slika 5.4. Za objektivno merilo kakovosti prilagajanja napovedanih in testnih vrednosti pa bomo uporabili normalizirano vrednost korena srednje vrednosti kvadratične napake (angl. *Normalized Root Mean Square Error - NRMSE*). Njena definicija je [28]:

$$e_{\text{NRMSE}} = 1 - \frac{||\mathbf{y}_{ref} - \boldsymbol{\mu}||^2}{||\mathbf{y}_{ref} - \langle \mathbf{y}_{ref} \rangle||^2}, \tag{5.1}$$

kjer je $\mathbf{y}_{ref}$ vektor testnih vrednosti, $\boldsymbol{\mu}$ vektor napovedanih vrednosti izhoda in z $\langle \cdot \rangle$ označujemo povprečno vrednost komponent vektorja. Odstopek e_{NRMSE} je definiran za vrednosti na intervalu $[1, -\infty)$ pri čemer 1 označuje popolno ujemanje in $-\infty$ zelo slabo ujemanje napovedi z referenčnimi vrednostmi. Kadar je $e_{\text{NRMSE}} = 0$, ni prileganje nič boljše od konstantne srednje vrednosti vektorja $\mathbf{y}_{ref}$. Zavedajmo se, da z NRMSE napako večdimenzionalni problem prikazujemo v eni dimenziji.

Če ni drugače navedeno, uporabljamo sledečo konfiguracijo:

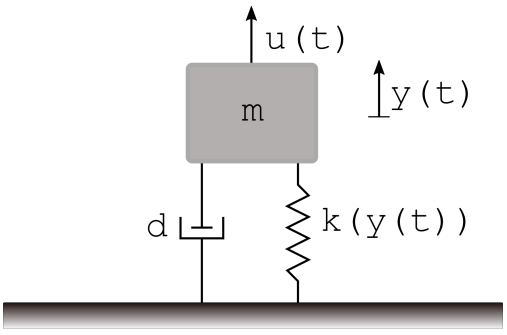
– v obeh modelih (GP-LVM in globokih GP) uporabljamo Gaussovo kovariančno funkcijo iz poglavja 2.3 s čimer omogočimo ARD in
– 100 induciranih točk za model GP-LVM in 10 induciranih točk za model globokih GP. S takim številom induciranih točk smo dosegli optimalen kompromis med napako e_{NRMSE} in časom izračuna modela.

V sklopu magistrskega dela uporabljamo že napisano programsko opremo. Avtor programske opreme, Andreas C. Damianou, omogoča prost dostop do programske implementacije globokih GP na spletni strani [29] za programsko okolje Python. Na voljo je tudi programski paket v okolju MATLAB [30]. V konkretnem primeru smo v prvi fazi laboratorijskega dela uporabljali okolje MATLAB, a kasneje upoštevali avtorjev nasvet in vso analizo opravili v programskem okolju Python. Določene datoteke programskega paketa smo prilagodili zahtevam konkretnega primera, sicer pa v matematično jedro programa nismo posegali. V veliko pomoč pri razumevanju programskega paketa je doktorska disertacija [4] in kratka predstavitev dostopna na spletnem naslovu [31].

Pri vrednotenju ilustrativnega primera uporabljamo prenosni računalnik z 64-bitnim operacijskim sistemom Windows 10, i7-6500U 2,5 GHz dvojedrnim procesorjem in 8 GB delovnega spomina.

5.1 Opis sistema

Kot del sintetičnega eksperimenta želimo modelirati dinamiko nelinearnega nihanja. Sistem mase m, dušilke z viskoznim faktorjem dušenja d in nelinearno vzmetjo $k(y(t))$ je simulirano z elektronskim vezjem [15, 16]. Mehanski ekvivalent dinamičnega sistema je skiciran na Sliki 5.1.



Slika 5.1: Skica mehanskega sistema, ki ga simuliramo z električnim vezjem.

Odmik mase $y(t)$ (izhodna vrednost) iz ravnovesne lege je povezana s silo vzbujanja $u(t)$ (vhodna vrednost) sistema. Gibalna enačba sistema je nelinearna

$$m\frac{\mathrm{d}^2y(t)}{\mathrm{d}t^2} + d\frac{\mathrm{d}y(t)}{\mathrm{d}t} + k(y(t))y(t) = u(t), \tag{5.2}$$

saj je koeficient vzmeti odvisen od odmika iz ravnovesne lege

$$k(y(t)) = a + by(t)^2. \tag{5.3}$$

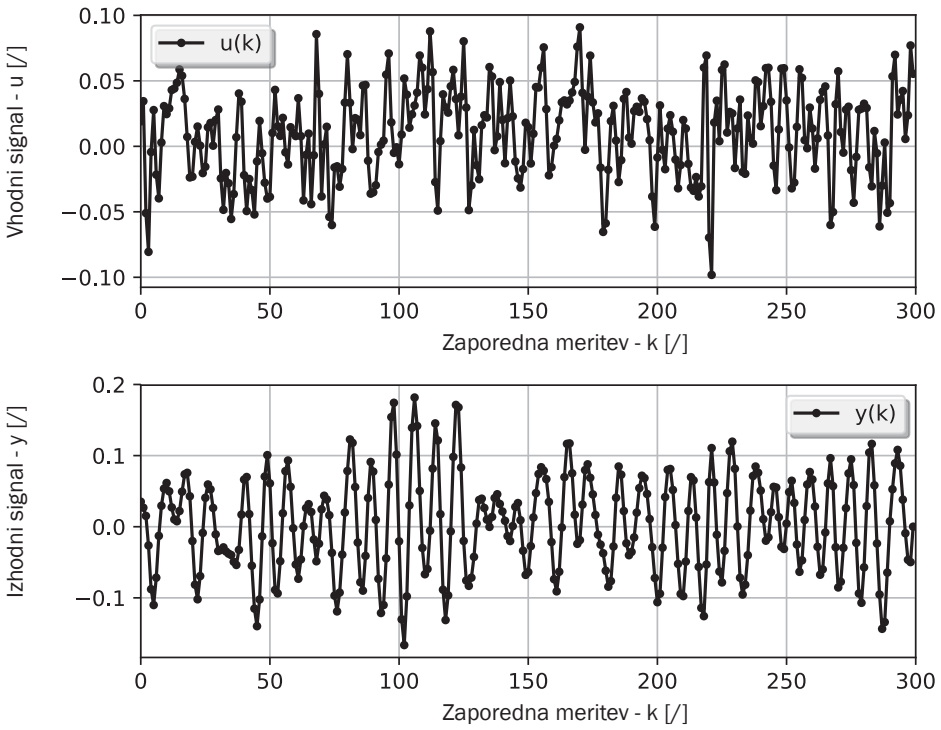
Za majhne odmike $y(t)$ po navadi zanemarimo nelinearni del koeficienta vzmeti. Tokrat se osredotočamo prav na nelinearni del, saj naj bi tako model GP-LVM kot tudi model globokih GP uspešno identificirala dinamiko nelinearnih sistemov. Celoten model odmika bomo identificirali kot črno škatlo in ne bomo upoštevali strukture iz enačb (5.2) in (5.3).

5.2 Podatki

Izmerjene vrednosti $\{u(t), y(t)\}$ električnega vezja ločimo v *učno* in *testno* množico. Z učno množico model naučimo zveze med vhodnimi in izhodnimi vrednostmi, s testnimi podatki pa preverjamo kako uspešno je bilo učenje modela. Pravimo jima tudi množici podatkov za *identifikacijo* in *vrednotenje*. Po navadi v učno množico pospravimo vse podatke, ki so dobri, t.j. vsebujejo nove informacije o sistemu, so popolni, nimajo preveč šuma, itd. V testno množico pa shranimo vse ostalo, kar pač ostane.

V konkretnem primeru imamo podatke (glej [15]) že ločene. Na voljo imamo 10 000 učnih in 4000 testnih vrednosti vhodov $u(t)$ in izhodov $y(t)$ zajetih s frekvenco 610.35 Hz. Preden jih prepustimo optimizacijskim algoritmom programske opreme, je potrebno še nekaj predobdelave.

Na Sliki 5.2 je izrisanih prvih 300 vhodno-izhodnih vrednosti.



Slika 5.2: Prvih 300 vrednosti vhodnih in izhodnih vrednosti sintetičnih podatkov. Na absciso nanašamo zaporedno število meritve k.

5.2.1 Predobdelava podatkov

Izkaže se, da sta oba modela v fazi učenja bolj učinkovita, kadar je povprečje podatkov enako 0 in njihova varianca enaka 1 [4]. V praksi to pomeni, da meritve sistema $u(t)$ in $y(t)$ iz testne in učne množice standardiziramo. V nadaljevanju prikazujemo postopek za normalizacijo vhodnih vrednosti $u(t)$. Povsem enaki koraki so potrebni za standardizacijo učnih in testnih vrednosti izhodov $\mathbf{y}(t)$.

Ker so meritve opravljene ob ekvidistantnih diskretnih časih, se hkrati znebimo zapisa $u(t) \to u(k)$, pri čemer je $k \in \mathbb{Z}_+$ zaporedno število meritve. Proces standardizacije zahteva operaciji

1. odštevanja povprečne vrednosti $\widetilde{u}(k) = u(k) - \langle \mathbf{u} \rangle$ in nato še
2. normalizacijo standardne deviacije $\hat{u}(k) = \frac{\widetilde{u}(k)}{\sigma_{\widetilde{\mathbf{u}}}}$,

pri čemer je $u(k)$ k-ta komponenta vektorja $\mathbf{u}$ v katerem so zbrane vrednosti iz učne in testne množice. S simbolom $\langle \cdot \rangle$ označujemo povprečenje po komponentah in z $\sigma_{\widetilde{\mathbf{u}}}$ standardno deviacijo komponent. Proces ponovimo za vsako komponento u iz vektorja $\mathbf{u}$. Vektor $\hat{\mathbf{u}}(k)$ ima tedaj dve lastnosti:

$$\langle \hat{\mathbf{u}} \rangle = 0 \text{ in} \tag{5.4}$$

$$\mathrm{Var}(\hat{\mathbf{u}}) = 1. \tag{5.5}$$

V nadaljevanju pozabimo na oznako s strešico in privzamemo, da imajo vsi numerični podatki od tod dalje zgornji dve lastnosti (5.4) in (5.5).

5.2.2 Regresorski vektor

Dinamični sistem je vsak sistem, katerega trenutno stanje je odvisno od prejšnjih stanj [21]. Dinamični sistem v zveznem prostoru opišemo z diferencialnimi enačbami. Diskretiziran sistem pa je predstavljen z diferenčnimi enačbami. V zveznem prostoru odvisnost od prejšnjih stanj prepoznamo kot odvode po času, v diskretnem prostoru pa kot zakasnitve vrednosti različnih fizikalnih spremenljivk. Obravnavani sistem iz poglavja 5.1 je tako dinamični sistem.

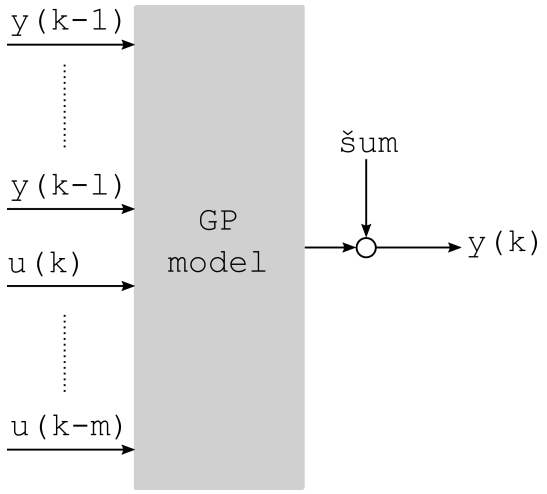
Iščemo preslikavo iz večdimenzionalnega prostora vhodnih vrednosti $\mathbf{Z} \in \mathbb{R}^{n \times q}$ v vektor izhodnih vrednosti sistema $\mathbf{y} \in \mathbb{R}^n$.

$$y_i(k) = f(\mathbf{z}_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \beta^{-1}\mathbf{I}), \tag{5.6}$$

kjer je ϵ_i Gaussov šum in $\mathbf{z}_i$ *regresorski vektor*, v katerem so zbrane vse vhodne vrednosti modela.

Regresorski vektor $\mathbf{z}_i$ predstavlja vrstico matrike $\mathbf{Z}$ in je v dinamičnih sistemih sestavljen iz zakasnjenih vrednosti vhodov in izhodov

$$\mathbf{z}_i = (y(k-1), y(k-2), \ldots, y(k-l), u(k-1), u(k-2), \ldots, u(k-m)), \tag{5.7}$$



Slika 5.3: Grafični prikaz modela NARX. Slika po viru [10].

pri čemer z l označujemo največjo zakasnitev izhodnih vrednosti in z m največjo zakasnitev vhodnih vrednosti. Takšna sestava regresorja je znana tudi pod imenom NARX (angl. *nonlinear autoregressive model with exogenous input*) [7, 10], ki je grafično prikazan na Sliki 5.3. Red dinamičnega sistema določa število zakasnitev izhoda l.

V splošnem je potrebno za uspešno identifikacijo sistema ugotoviti strukturo regresorskega vektorja. Slednje je lahko precej težavno in zamudno delo, še posebej v primeru modelov črne škatle, kjer ne poznamo fizikalne narave sistema in njegovih zakonitosti. Ker ne poznamo reda dinamičnega sistema, lahko v regresorskem vektorju upoštevamo neoptimalno število časovnih zakasnitev. Preveč zakasnitev in s tem regresorjev je z računskega stališča preveč obremenjujoče ter posledično model prezapleten. S premalo zakasnitvami pa poglavitno dinamiko sistema praviloma preslabo opišemo. Nekatere metode za določitev regresorjev na podlagi identifikacije dinamičnega reda sistema z Gaussovimi procesi najdemo v [10] in [8]. Lahko pa uberemo tudi bolj empiričen pristop in z ARD lastnostjo kovariančnih funkcij ter poskušanjem različnih zakasnitev poiščemo optimalno obliko regresorskega vektorja, kar pa je lahko zelo zamudno.

V konkretnem primeru iz poglavja 5.1 z iskanjem oblike regresorskega vektorja nimamo težav, saj so avtorji članka [17] že ugotovili, da regresor oblike

$$\mathbf{z}_i = (y(k-4), y(k-2), y(k-1), u(k-3), u(k-2), u(k-1)), \tag{5.8}$$

dobro opiše dinamiko sistema.

5.3 Identifikacija in ugotovitve

S podatki in programsko opremo smo opravili več serij eksperimentov. V nadaljevanju prikazujemo le najpomembnejše ugotovitve in rezultate. Največjo pozornost smo namenili različnim inicializacijam hiperparametrov, latentnega prostora ter induciranih spremenljivk. Preverili smo tudi kako na rezultate modela globokih GP vpliva večje števila skritih slojev, uporaba apriornega verjetja in različno število induciranih točk.

5.3.1 Naključna inicializacija

Analizo sintetičnih podatkov opravimo z regresorskim vektorjem kot ga definirajo avtorji dela [17]. Le-ti trdijo, da je obravnavani sistem četrtega reda in da vso poglavitno dinamiko opišemo z regresorskim vektorjem v enačbi (5.8).

Inicializacija vseh hiperparametrov, latentnih in induciranih spremenljivk naj bo zaenkrat naključna. Zaenkrat se omejimo na le en skriti sloj v modelu globokih GP. Rezultati obeh modelov so na Sliki 5.4, kjer prikazujemo le majhen del iz domene testnih podatkov, t.j. le 50 testnih meritev na intervalu $k \in [2000, 2050]$. V nasprotnem primeru so namreč grafi povsem neberljivi in bi jih le stežka komentirali. Črtkana črta na grafu prikazuje izmerjen odziv sistema, polna črta napoved modela in s sivo je obarvan interval 95 % zaupanja.

S primerjavo modelov na Sliki 5.4 in rezultatov v Preglednici 5.1 ugotovimo, da se z naključno inicializacijo pri danih sintetičnih podatkih oba modela približno enako dobro naučita preslikave med vhodnimi in izhodnimi vrednostmi. Napovedi modela globokih GP se kljub manjšemu številu induciranih točk ujemajo z realnimi meritvami enako dobro kot napovedi modela GP-LVM. Rezultati analize so v splošnem boljši v korist modela globokih GP.

Model	Št. induciranih točk	Čas računanja [s]	e_{NRMSE}	$\langle 2\sigma \rangle$
GP-LVM	100	135	0.972	0.354
Globoki GP	10	8	0.970	0.148

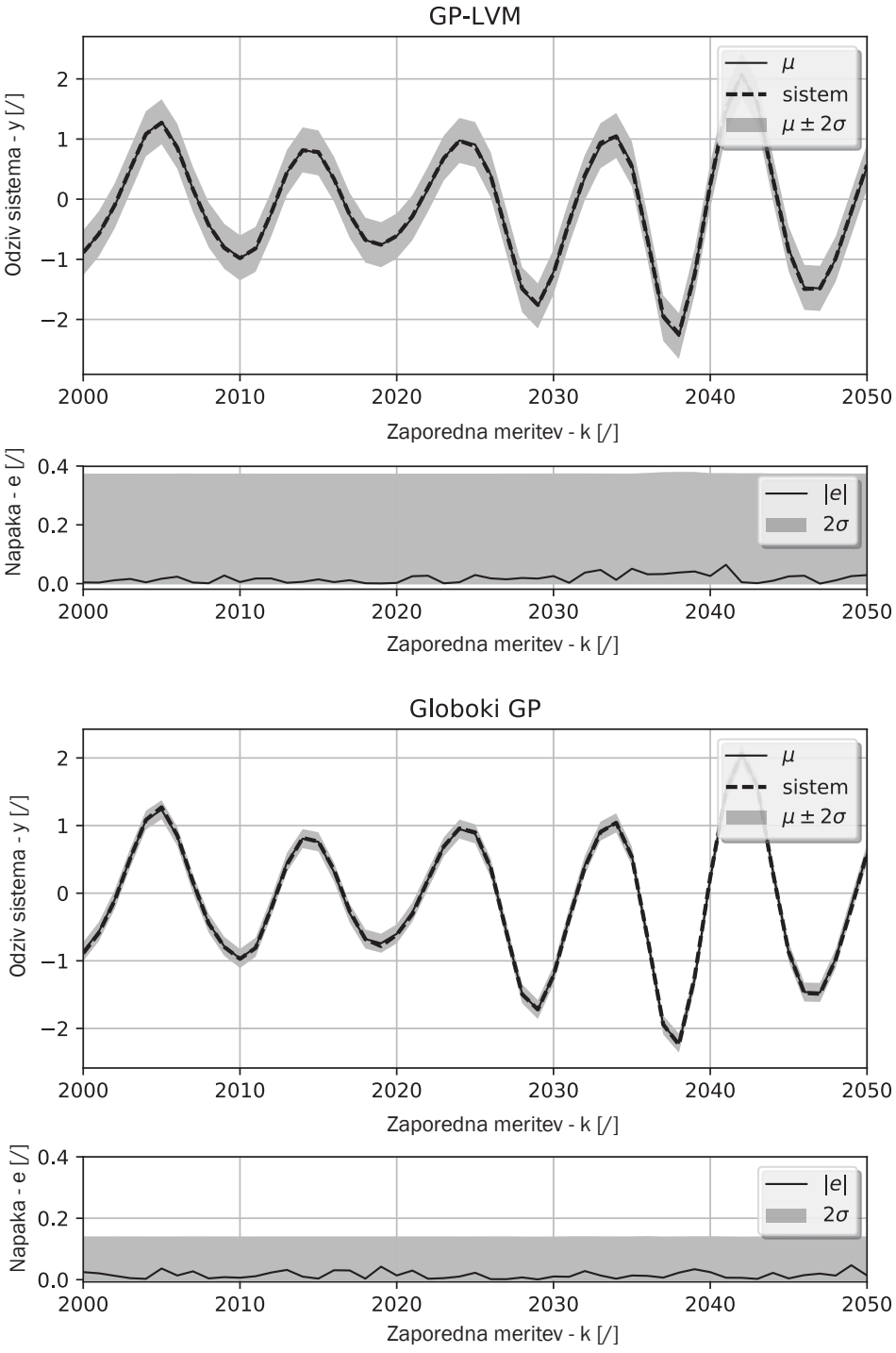
Legenda: $\langle 2\sigma \rangle$ povprečna negotovost napovedi modela

Preglednica 5.1: Glavne karakteristike obeh regresijskih modelov z naključno inicializacijo.

Najpomembneje je, da smo z modelom globokih GP našli boljši lokalni minimum. Slednjo informacijo iz Slike 5.4 razberemo kot bistveno večjo negotovost napovedi modela GP-LVM. Povprečna vrednost negotovosti je podana v Preglednici 5.1. Oba modela se zelo dobro naučita preslikave f (za objektivno merilo glej e_{NRMSE} v Preglednici 5.1). Razliko dveh tisočink NRMSE napake lahko zanemarimo. Opazimo tudi, da je model globokih GP s časovnega stališča bistveno manj zahteven.

Izračun modela GP-LVM smo tisočkrat ponovili, a pri tem z naključno inicializacijo nismo našli boljšega ali vsaj ekvivalentno dobrega lokalnega minimuma, kot ga najde model globokih GP. Izkaže pa se, da brez težav najdemo še boljši lokalni minimum modela globokih GP (glej naslednje poglavje 5.3.2).

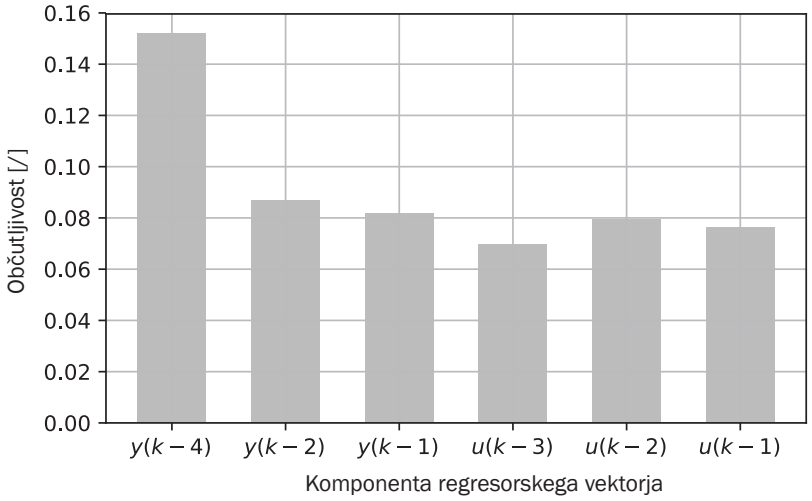
Na Sliki 5.5 so izrisane vrednosti hiperparametrov kovariančne funkcije. Posamezno vrednost hiperparametra interpretiramo kot utež pripadajoče prostostne stopnje regresorskega vektorja - občutljivost modela na prostostno stopnjo. Opazimo, da ima največjo utež komponenta $y(k-4)$, medtem ko so uteži ostalih zakasnitev približno enake.



Slika 5.4: Rezultati naključne inicializacije obeh modelov za osnovno obliko regresorja in $k \in [2000, 2050]$.

5.3.2 Premišljena inicializacija modela globokih GP

Hiperparametre modela ter latentne in inducirane spremenljivke lahko inicializiramo naključno ali *premišljeno*. Premišljena inicializacija pomeni, da nekatere spremenljivke inicializiramo z določenimi vrednostmi in preverimo, kako se s tem spremenijo rezul-



Slika 5.5: Uteži posameznih prostostnih stopenj regresorskega vektorja.

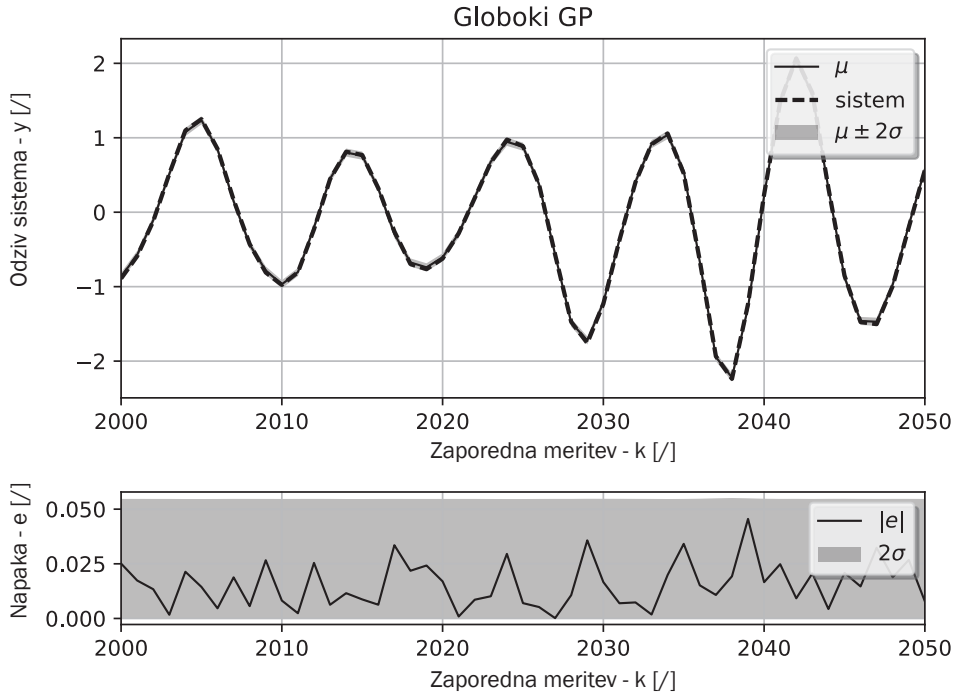
tati modela. Premišljena zato, ker po navadi najprej poiščemo tiste hiperparametre ali spremenljivke, ki močno vplivajo na numerično stabilnost modela. Tem določimo vrednosti in šele potem spreminjamo vrednosti ostalih hiperparametrov in spremenljivk. S premišljeno inicializacijo hiperparametrov ter induciranega in latentnega prostora, bi morali v principu poiskati boljši lokalni minimum ali še boljše, globalnega. S tem se močno zmanjša negotovost napovedanih vrednosti modela.

V predhodnem poglavju smo ugotovili, da oba modela v primeru naključne inicializacije hiperparametrov, latentnega in induciranega prostora kljub večkratnim ponovitvam vztrajno zaideta v relativno slab lokalni minimum. Drugačna inicializacija modela GP-LVM ni vodila do boljših rezultatov. Po tisoč ponovitvah so na Sliki 5.4 prikazani najboljšega lokalnega minimuma, kot ga najdemo z naključno inicializacijo. S premišljeno inicializacijo nismo uspeli dobiti boljših rezultatov. V nadaljevanju zato obravnavamo le različne inicializacije modela globokih GP, medtem ko inicializacijo modela GP-LVM ohranimo naključno.

V splošnem ni pravila kako inicializirati parametre, da bo program zašel v najboljši lokalni minimum. Izkazalo se je, da je konvergenca optimizacijskega postopka zelo pogojena z inicializacijo latentnega prostora. Z nepravilno izbiro smo kaj kmalu podvojili čas izračuna in pri tem dočakali silno slabe rezultate - do 10-krat večjo negotovost napovedi in NRMSE napako približno 0.1-0.4. Na Sliki 5.6 so rezultati modela globokih GP, pri katerem smo za inicializacijo latentnega prostora vzeli 9996 (število učnih podatkov) enakomerno porazdeljenih števil na intervalu $[0, 1]$. To je že zadostovalo za precej boljše rezultate v primerjavi z naključno inicializacijo modela.

Odvisnost modela od ostalih parametrov, npr. hiperparametri kovariančne funkcije, varianca latentnih spremenljivk, inducirane spremenljivke ipd., smo pustili naključno, saj smo s spreminjanjem njihovih vrednosti rezultate le poslabšali. Eksperiment smo s privzeto inicializacijo latentnega prostora ponovili tisočkrat, a pri tem nismo našli boljšega lokalnega minimuma.

S primerjavo rezultatov naključne inicializacije na Sliki 5.4 in rezultatov premišljene inicializacije na Sliki 5.6 opazimo, da smo s stališča negotovosti še dodatno izboljšali



Slika 5.6: Rezultati modela globokih GP s premišljeno inicializacijo parametrov.

model globokih GP. S tem je bil dosežen primarni cilj premišljene inicializacije. Z njo smo sicer nekoliko podaljšali čas izračuna, a dosegli več kot pol manjšo negotovost meritev (glej Preglednici 5.2 in 5.1). Obenem se izboljša NRMSE napaka, ki sedaj znaša 0.976 v primeru s prejšnjih 0.970 iz naključne inicializacije. Izboljšanje NRMSE napake je sicer dobrodošlo, a sprememba 6 tisočink ni ključnega pomena. Pomembneje je to, da smo s premišljeno inicializacijo dosegli skoraj trikrat manjšo negotovost napovedi modela.

Model	Čas računanja [s]	e_{NRMSE}	$\langle 2\sigma \rangle$	Premišljena inicializacija
Globoki GP	8	0.970	0.148	×
	31	0.976	0.054	✓

Legenda: $\langle 2\sigma \rangle$ povprečna negotovost napovedi modela

Preglednica 5.2: Primerjava modela globokih GP z naključno in premišljeno inicializacijo.

5.3.3 Dodatni eksperimenti

Rezultati modela globokih GP z več kot enim skritim slojem so bili venomer slabši in časovno zahtevnejši. V Preglednici 5.3 so zbrani podatki eksperimentov z več skritimi sloji. V vseh poskusih z večjim številom skritih slojev je bila negotovost napovedi večja, čas izračuna daljši, vrednost NRSME napake pa manj ugodna. Zaključimo lahko, da večje število slojev v konkretnem primeru ne vodi do boljšega regresijskega modela.

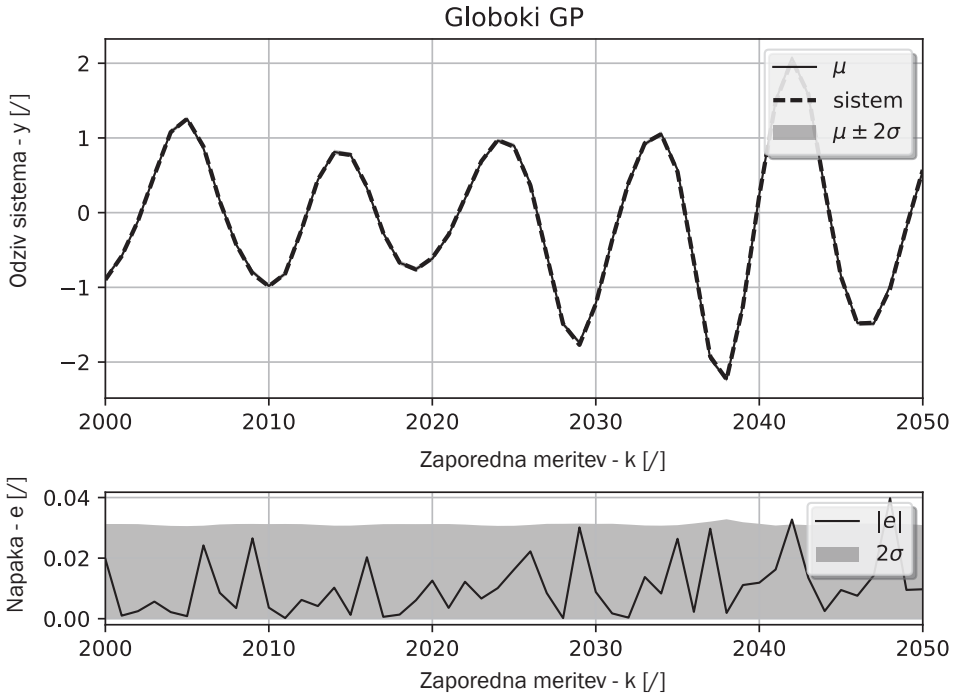
Št. skritih slojev	Čas računanja [s]	e_{NRMSE}	$\langle 2\sigma \rangle$
1	8	0.974	0.142
2	49	0.971	0.155
3	73	0.965	0.229
4	90	0.963	0.273

Legenda: $\langle 2\sigma \rangle$ povprečna negotovost napovedi modela

Preglednica 5.3: Vpliv števila slojev na model globokih GP.

V serijah eksperimentov smo skušali boljši rezultat doseči tudi z definicijo apriornega verjetja nad parametri. Rezultati sicer niso bili slabi, a se nam kljub temu ni uspelo približati rezultatom na Sliki 5.6.

Še manjšo negotovost napovedi modela globokih GP dosežemo na račun večjega števila induciranih točk (glej Sliko 5.7 in Preglednico 5.4). Če število induciranih točk z 10 povečamo na 100, se negotovost in tudi NRMSE napaka močno izboljšata. Se pa v zakup temu za faktor 60 podaljša čas izračuna. Večje število induciranih točk modela GP-LVM ni vodilo do boljših rezultatov.



Slika 5.7: Rezultati modela globokih GP s premišljeno inicializacijo parametrov in 100 induciranimi točkami.

Medtem ko je model z 10 induciranimi točkami najbolj občutljiv na inicializacijo latentnega prostora, je občutljivost modela globokih GP s 100 induciranimi točkami na Sliki 5.7 močno odvisna od večjega števila incializiranih parametrov. Inicializacije, ki smo jih uporabili so:

– varianca kovariančne matrike je 20,

Model	Št. induciranih točk	Čas računanja [s]	e_{NRMSE}	$\langle 2\sigma \rangle$
Globoki GP	10	31	0.976	0.054
	100	1902	0.984	0.031

Legenda: $\langle 2\sigma \rangle$ povprečna negotovost napovedi modela

Preglednica 5.4: Odvisnost modela globokih GP od števila induciranih točk.

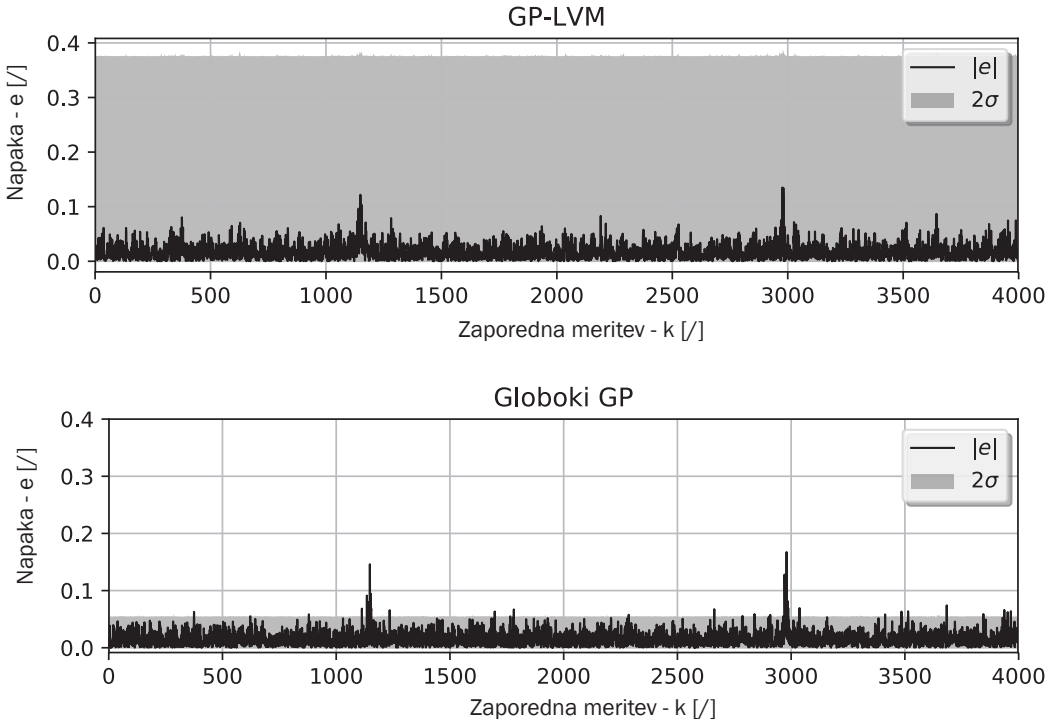
– horizontalni skalirni faktor kovariančne matrike je 10,
– latentni prostor inicializiramo z 9996 vrednostmi med -1 in 1,
– inducirane točke skritega sloja inicializiramo enako kot latentni prostor,
– varianca kovariančne matrike skritega sloja je 80 in
– horizontalni skalirni faktor kovariančne matrike skritega sloja je 15.

5.3.4 Zaključki

Oba modela se dobro naučita zveze med vhodnim prostorom $\mathbf{Z}$ in izhodnimi vrednostmi $\mathbf{y}$. Rezultati naključne inicializacije modela globokih GP sicer niso slabi, a lahko z nekoliko več truda dobimo bistveno boljše. Boljšega lokalnega minimuma nam z modelom GP-LVM in naključno inicializacijo po 1000 ponovitvah ni uspelo najti. S še večjim številom ponovitev eksperimenta bi to morda uspelo. Z nekoliko več truda smo uspešno izboljšali rezultate modela globokih GP. S premišljeno inicializacijo modela globokih GP smo poiskali boljši lokalni minimum, kot ga najde naključno inicializiran model.

V splošnem lahko ocenimo, da je v konkretnem primeru sintetičnih podatkov model globokih GP boljši in hitrejši v primerjavi z GP-LVM modelom. Izkaže se, da je obravnavani sistem ravno dovolj kompleksen, da se pokažejo prednosti modela globokih GP.

V sklopu ilustrativnega eksperimenta smo dosegli zadani cilj - z modelom globokih GP smo identificirali nelinearni dinamični sistem in pokazali lastnosti modela globokih GP. Vsa teoretična dognanja iz predhodnih poglavij smo pokazali na praktičnem primeru. V Preglednici 5.4 in na Sliki 5.7 prikazujemo uspešnost modela globokih GP pri identifikaciji nelinearnega dinamičnega sistema. Model se je uspešno naučil preslikave med vhodnimi in izhodnimi vrednostmi. Z negotovostjo napovedanih vrednosti modela globokih GP na Sliki 5.8 smo zadovoljni, saj vidimo, da smo s premišljeno inicializacijo napako precej zmanjšali negotovost napovedanih vrednosti.



Slika 5.8: Napaka in negotovost napovedi obeh modelov za vse testne podatke.

6 Napoved temperature

V prejšnjem primeru smo pokazali, da lahko z globokimi GP uspešno modeliramo nelinearne dinamične sisteme. V nadaljevanju metodo modeliranja z globokimi GP uporabimo tudi na realnem problemu: napoved temperature v okolici JEK za pol ure vnaprej.

Cilj tega poglavja je določiti zvezo med vhodno-izhodnimi vrednostmi izmerjenih meteoroloških spremenljivk. Model želimo naučiti tako dobro, da bodo napovedane vrednosti čim bližje dejanski meritvi in negotovost povezana z napovedjo čim manjša.

Napovedane vrednosti modela globokih GP tudi tokrat primerjamo s preprostejšim modelom GP-LVM, ki je v predhodnem poglavju vselej uspešno poiskal dokaj dober lokalni minimum. Podobno kot v prejšnjem poglavju, rezultate obeh modelov primerjamo na osnovi dvodimenzionalnih grafov. Na posamezni sliki prikazujemo napovedane vrednosti modela $\mu(k)$ in testne podatke ter napako $e(k)$ definirano kot absolutno vrednost razlike med njima, npr. Slika 6.6. Pri obeh grafih s sivo pobarvanim območjem označujemo interval 95 % zaupanja. Za objektivno vrednotenje ujemanja napovedanih in testnih vrednosti se zanašamo na NRMSE-napako iz enačbe (5.1).

Če v poglavju ni drugače navedeno, uporabljamo sledečo konfiguracijo:

– v obeh modelih (GP-LVM in globokih GP) uporabljamo Gaussovo kovariančno funkcijo iz poglavja 2.3 s čimer omogočimo ARD in
– 100 induciranih točk za model GP-LVM in 10 induciranih točk za model globokih GP. S takim številom induciranih točk dosežemo optimalen kompromis med e_{NRMSE} napako in časom izračuna modela.

Pri analizi večjega števila podatkov z modelom GP-LVM se izkaže, da prenosni računalnik z 8 GB delovnega spomina več ne zadostuje. V namen raziskovalnega dela smo na tem mestu uporabljali strežnik z i5-6400 2,7 GHz štirijedrnim procesorjem in kar 24 GB delovnega spomina. Slednja konfiguracija je zadoščala za vso nadaljnjo analizo.

6.1 Opis sistema

Osrednji problem obsega napoved mikroskopske klime v okolici JEK, t.j. v radiju dveh kilometrov in manj. V primeru nesreče in odpovedi varnostnih sistemov v JEK se od pristojnih organov namreč pričakuje, da bodo ustrezno ukrepali. Primerna reakcija,

poleg vseh predpisanih varnostnih protokolov, obsega tudi morebitno evakuacijo prebivalcev iz okoliških vasi. Pri tem je v veliko pomoč, če lahko z uporabo meteoroloških modelov zanesljivo predvidimo vsaj nekatere meteorološke parametre. Dolgoročni cilj je čim bolj natančna napoved gibanja zračne mase in s tem morebitnega širjenja oblaka radioaktivnih delcev.

V sklopu tega poglavja skušamo iz meritev različnih meteoroloških spremenljivk in uporabo modela globokih GP napovedati temperaturo zraka 2 m nad tlemi.

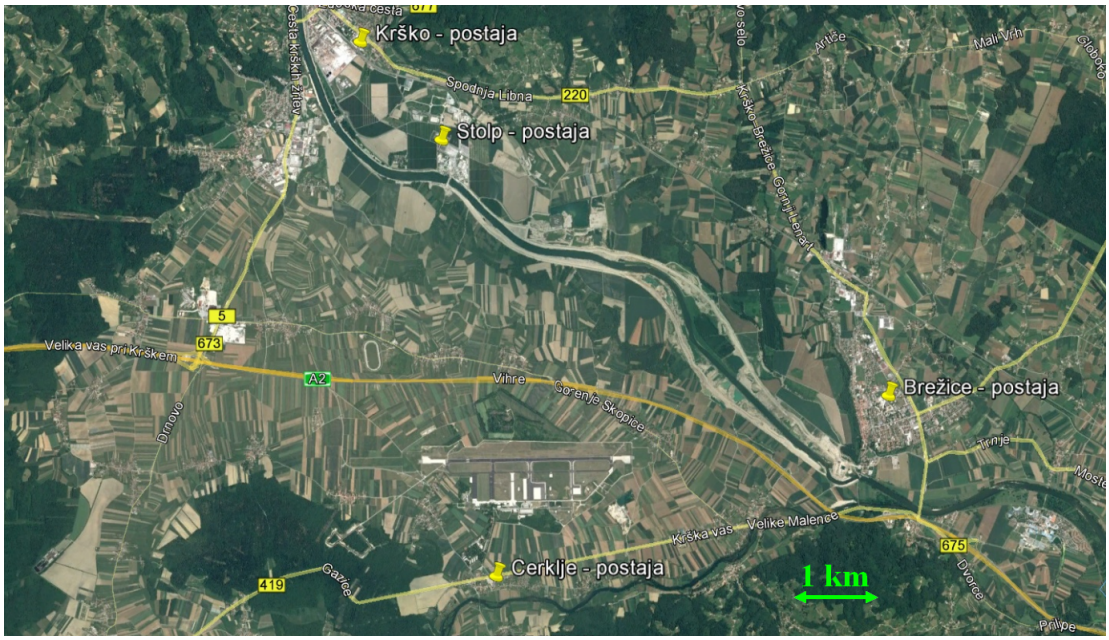
Temperatura zraka ima zelo bogate dinamične lastnosti (glej Sliko 6.2) [19]. Poleg dnevnih nihanj imamo letne čase, ohladitve po padavinah, vročinske vale, obdobja polarne zime in druge vremenske pojave pri katerih občutimo tudi spremembo temperature. Naravna sprememba temperature je počasen proces, po navadi povezan s spremembo drugih parametrov, npr. tik pred fronto ali nevihto, ki zrak ohladi, se zračni tlak zmanjša. Podobna opažanja danes upoštevajo številne kratkoročne meteorološke napovedi [18]. Tudi pooblačitev, zaradi katere kratkovalovno sončno sevanje več ne greje tal, zaustavi oddajanje toplote v okoliško zračno maso. Podobno oblačna noč preprečuje uhajanje dolgovalovnega sevanja segretih tal iz atmosfere - takrat se zbudimo v toplo jutro.

Razumeti moramo, da lahko informacije o trendu temperature zraka dobimo iz zelo širokega nabora meteoroloških spremenljivk. Glavni cilj vsake identifikacije dinamičnega sistema je, da iz meteoroloških meritev, ki so nam na voljo, zaobjamemo vse naštete in tudi nenaštete pojave.

Atmosfera je zelo kompleksen sistem [19]. Radi bi napovedali temperaturo 2 m nad tlemi, t.j. temperaturo v prizemni plasti atmosfere kjer je ogromno vplivnih veličin - merljivih in tudi nemerljivih. Jasno je, da so v prizemni plasti prisotne močne turbulence zračnih mas in vertikalno mešanje zraka [18]. Poleg geografskih značilnosti okoliškega terena (glej Sliko 6.1) imajo velik vpliv na gibanje zračnih mas v prizemni plasti tudi gozdovi, reke in objekti postavljeni v okolici. Znano je, da rjavo pšenično polje bistveno bolje zadržuje toploto kot nizka zelena trava. Zaradi rjave barve je absorpcijski koeficient pšenice precej večji kot absorpcijski koeficient zelene trave. Topel zrak se nato ujame v visoki pšenici in počasi ohlaja. Podobno so gozdovi običajno hladnejši od okolice: absorpcijski koeficient zelenih listov je majhen, poleg tega sončni žarki le redko sežejo do tal, da bi se lahko okoliški zrak segrel.

JEK se sicer ne nahaja sredi pšeničnega polja ali gostega zelenega gozdu, je pa locirana na relativno razgibanem geografskem področju. S splošnim šibkim vetrom se toplota iz pšeničnega polja z vetrom pomakne proti JEK, ki lahko deluje kot sprožilec termičnega stebra in radioaktivni oblak ponese visoko od tal.

Kompleksnost obravnavnega sistema je torej velika in med drugim obsega tudi opazovane količine, ki jih težko izmerimo ali objektivno predstavimo. Mnoge aktualne napovedi še vedno temeljijo na meritvah, ki jih meteorologi opravljajo ročno ali po občutku [19], npr. vidljivost ter tip in višina oblakov. Oba opazovana parametra brez dvoma vplivata na stabilnost atmosfere, ki je pomembna meteorološka spremenljivka kadar govorimo o napovedi temperature.



Slika 6.1: Geografske značilnosti okoliškega terena in merilne postaje.

6.2 Podatki

V skladu s fizikalnim razumevanjem sistema iz prejšnjega poglavja si želimo opazovati zelo širok spekter meteoroloških spremenljivk. V nadaljevanju se omejimo le na tiste spremenljivke, ki so merljive, predvsem pa na tiste, katerih meritve so na voljo. Četudi se nabor opazovanih spremenljivk morda ne zdi idealen, skušamo iz njih dobiti čim več informacij o dinamičnih lastnostih sistema.

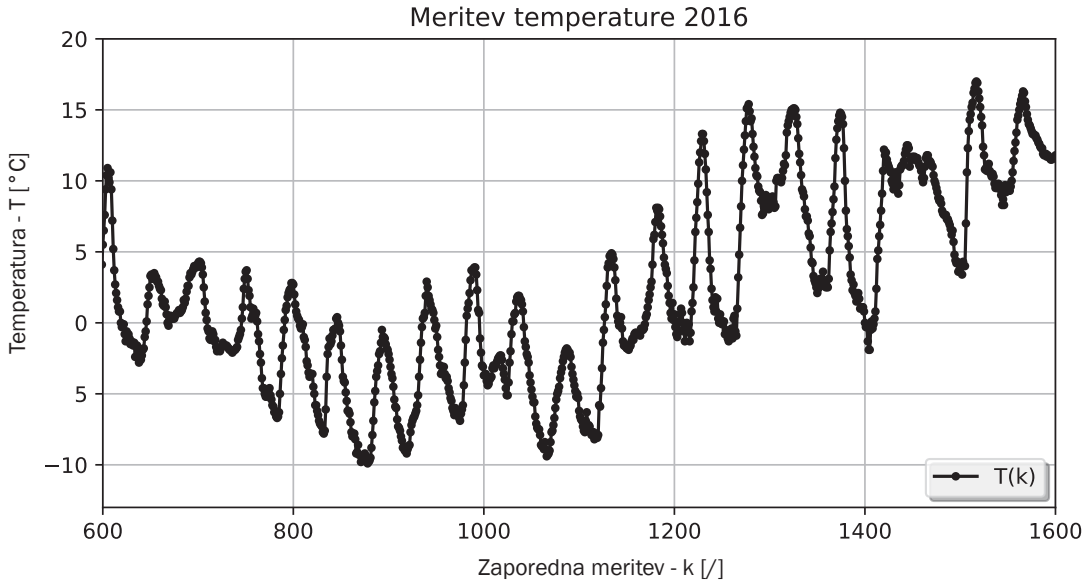
Podjetje *MEIS d.o.o.* že vrsto let opravlja meteorološke meritve v JEK in njeni okolici. Njim gre zahvala za meritve iz stolpa na Sliki 6.1:

– temperature (T) 2 m nad tlemi,
– relativne vlažnosti zraka (φ) 2 m nad tlemi,
– stabilnosti atmosfere ($\mathcal{PG}$),
– zračnega tlaka pri tleh (p),
– globalnega sončnega sevanja (P) in
– hitrosti vetra (v) 10 m nad tlemi

za obdobje štirih let (2013-2016). Tu je stabilnost atmosfere definirana po *Pasquill-Girardovih* stabilnostnih razredih A-G, kjer A pomeni zelo nestabilno in G zelo stabilno ozračje [32]. Stabilnostni razredi so zelo opisno definirani, npr. na sončen dan brez oblačka in z manj kot 2 m/s splošnega vetra ocenimo stabilnost atmosfere s črko A.

Vse spremenljivke so izmerjene v 30 minutnih intervalih, kar znese približno 17 500 meritev na leto. Skupno torej približno 70 000 meritev vsake opazovane veličine - bistveno več v primerjavi s 14 000 sintetičnih podatkov v prejšnjem poglavju ilustrativnega eksperimenta.

Na Sliki 6.2 je prikazan le del meritev temperature iz leta 2016.



Slika 6.2: Izmerjena temperatura v letu 2016. Prikazanih je le zaporedje 1000 meritev $\approx$ 21 dni z naključno izbranim začetkom. S k označimo zaporedno število meritve.

6.2.1 Predobdelava podatkov

Preden podatke prepustimo optimizacijskemu postopku programa jih še ustrezno normaliziramo. Postopek normalizacije je povsem enak kot pri normalizaciji sintetičnih podatkov v poglavju 5.2.1. Edina razlika je, da tokrat obravnavamo 6 spremenljivk, prej pa le 2. Postopek torej ponovimo za vsako opazovano spremenljivko. Glavno je, da ima na koncu matrika izmerjenih vrednosti po stolpcih povprečje nič in standardno deviacijo enako 1.

Delitev na testno in učno množico

Na voljo imamo meritve iz leta 2013 do vključno leta 2016. Podatke moramo smiselno razdeliti na učno in testno množico. Želimo, da se model iz testne množice čim bolje nauči zveze med vhodom $\mathbf{z}_i$ in izhodom T sistema. V učno množico zato zberemo čim več podatkov, ki nosijo nove informacije o sistemu, v testno po shranimo vse kar ostane.

Z ozirom na fizikalno vsebino opazovanega sistema in tudi napoved modelov, se zdi smiselno, da meritve iz leta 2013, 2014 in 2015 uporabimo kot učne podatke. Na podatkih iz leta 2016 pa preverimo, kako dobro smo model naučili. Kar v grobem pomeni približno 52 500 učnih podatkov in 17 500 testnih vrednosti, s katerimi model ovrednotimo.

Definicija regresorskega vektorja

Medtem ko z iskanjem pravih komponent regresorskega vektorja v sintetičnem eksperimentu iz poglavja 5.2.2 nismo imeli težav, v tem primeru to ne drži. Sestavo

regresorskega vektorja moramo določiti sami. Na voljo imamo meritve šestih različnih meteoroloških spremenljivk. Uporabili bomo NARX-model s Slike 5.3 pri čemer se moramo odločiti koliko in katere zakasnitve bomo v regresorskem vektorju upoštevali.

Za definicijo regresorskega vektorja se opiramo na fizikalno intuicijo iz poglavja 6.1 in razumevanje termodinamičnih procesov nasploh [18]. Pri tem izpostavimo globalno sončno sevanje, čigar vpliv na temperaturo zagotovo ni zanemarljiv. Ne pozabimo, da je najnižja dnevna temperatura po navadi izmerjena šele eno ali dve uri po sončnem vzhodu. Sodeč po takšnih opažanjih, bi bilo v regresorskem vektorju smiselno upoštevati več zakasnitev globalnega sončnega sevanja. Od ostalih meteoroloških parametrov pričakujemo, da bosta dve zakasnitvi zadostovali za identifikacijo sistema. Zdi se tudi, da so vsi parametri pomembni in vsi vplivajo na napoved temperature. Težko bi vnaprej z zagotovostjo trdili, da je kakšna spremenljivka povsem odveč. Analizo torej začnemo z regresorjem oblike

$$\begin{aligned}\mathbf{z}_i = (&T(k-2), T(k-1),\\ &\varphi(k-2), \varphi(k-1),\\ &\mathcal{PG}(k-2), \mathcal{PG}(k-1),\\ &P(k-4), P(k-3), P(k-2), P(k-1),\\ &p(k-2), p(k-1),\\ &v(k-2), v(k-1)),\end{aligned} \tag{6.1}$$

kjer je k zaporedno število meritve. Zaenkrat predpostavimo, da bo sledeča definicija regresorskega vektorja zadostovala. Po potrebi bomo v nadaljevanju regresorski vektor razširili ali z ARD lastnostjo kovariančnih funkcij odstranili nepomembne prostostne stopnje. Poudarimo, da v regresorju iz enačbe (6.1) upoštevamo le dve zakasnitvi izhoda, t.j. temperature. Vse ostale dimenzije predstavljajo zakasnjene vrednosti vhodov.

Manjkajoči podatki

Merilna tehnika je močno napredovala. Tehnologija je v zadnjih letih poskrbela za avtomatizacijo različnih procesov, med drugim tudi meritev. Žal popolni merilni sistemi ne obstajajo. Vsak realni sistem je tudi pokvarljiv.

Njihova nepopolnost se kaže v večih pogledih. Med njimi se dotaknimo nezanesljivega delovanja oziroma manjkajočih meritev. Spomnimo se, da opazujemo 6 meteoroloških spremenljivk. Kar v praksi pomeni, da lahko na primer zaradi okvare anemometra manjkajo meritve o skalarni hitrosti vetra. Zaradi izpada električne energije lahko manjkajo meritve vseh spremenljivk. Tu so še mehanske napake ali splošne okvare zaradi dotrajanosti opreme. V takem primeru so podatki *nepopolni*.

Manjkajoče vrednosti v regresorskem vektorju $\mathbf{z}_i$ iz matrike $\mathbf{Z}$ je potrebno ustrezno obravnavati. Postopamo lahko na več načinov. Morda pomislimo na interpolacijo med zadnjo znano meritvijo ter prvo vrednostjo po vnovični vzpostavitvi merilnega sistema. Ideja se zdi smiselna, a se pri tem zavedajmo, da z interpolacijo umetno vnašamo informacije v sistem, ki morda sploh niso pravilne. V konkretnem primeru zato k težavi pristopimo drugače.

Na voljo imamo ogromno število merilnih podatkov. Če kakšno vrstico, recimo tisto z manjkajočo vrednostjo, izmed 70 000 izbrišemo, pri tem najverjetneje ne izgubimo ogromno informacij o sistemu. Hkrati ne model globokih GP ne GP-LVM-model ne potrebuje in nima informacije o tem ali so meritve opravljene vsakih 30 minut ali pa enkrat na teden. Modela se naučita preslikave na osnovi vhodno-izhodne karakteristike sistema. Pri tem oba upoštevata, da bosta imela dva regresorja, ki sta si v vhodnem prostoru blizu, zelo verjetno podobno tudi izhodno vrednost. Zato iz matrike $\mathbf{Z}$ in iz vektorja izhodov $\mathbf{y}$ v enačbi

$$\mathbf{y} = f(\mathbf{Z}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \beta^{-1}, \mathbf{I}) \tag{6.2}$$

pobrišemo vse vrstice z manjkajočimi vrednostmi. Pri tem je nujno skrbno upoštevati tudi zakasnitve iz regresorja v enačbi (6.1). Če namreč manjka i-ti element poljubne prostostne stopnje regresorskega vektorja je potrebno iz matrike izbrisati tudi vse tiste vrstice kjer i-ta dimenzija nastopa kot zakasnitev.

Primer: Za večjo nazornost vzemimo matriko vhodnih regresorjev $\mathbf{A}$ in vektor izhodnih vrednosti $\mathbf{b}$. Vrednosti c in d naj predstavljata poljubno veličino.

$$\mathbf{A} = \begin{bmatrix} b(k-2) & b(k-1) & c(k-1) & c(k) & d(k) \\ b(k-1) & b(k) & c(k) & c(k+1) & d(k+1) \\ b(k) & b(k+1) & c(k+1) & c(k+2) & d(k+2) \\ b(k+1) & \boldsymbol{b(k+2)} & c(k+2) & c(k+3) & d(k+3) \\ \boldsymbol{b(k+2)} & b(k+3) & c(k+3) & c(k+4) & d(k+4) \\ b(k+3) & b(k+4) & c(k+4) & c(k+5) & d(k+5) \end{bmatrix} \quad \text{in} \quad \mathbf{b} = \begin{bmatrix} b(k) \\ b(k+1) \\ \boldsymbol{b(k+2)} \\ b(k+3) \\ b(k+4) \\ b(k+5) \end{bmatrix}$$

Predpostavimo, da meritev izhoda $b(k+2)$ iz nekih razlogov manjka. Vrednosti $b(k+2)$ so zgoraj označene s krepko pisavo. V takem primeru zavržemo 3., 4. in 5. vrstico iz matrike $\mathbf{A}$ in vektorja $\mathbf{b}$, saj vse vsebujejo vsaj eno neznano vrednost. Tedaj pišemo

$$\mathbf{A} = \begin{bmatrix} b(k-2) & b(k-1) & c(k-1) & c(k) & d(k) \\ b(k-1) & b(k) & c(k) & c(k+1) & d(k+1) \\ b(k+3) & b(k+4) & c(k+4) & c(k+5) & d(k+5) \end{bmatrix} \quad \text{in} \quad \mathbf{b} = \begin{bmatrix} b(k) \\ b(k+1) \\ b(k+5) \end{bmatrix}$$

6.3 Identifikacija in ugotovitve

S podatki in programsko opremo smo opravili več serij eksperimentov. V nadaljevanju prikazujemo le najpomembnejše ugotovitve in rezultate. Največjo pozornost smo namenili določitvi komponent regresorskega vektorja in različnim inicializacijam hiperparametrov, latentnega prostora ter induciranih spremenljivk. Preverili smo tudi kako na rezultate modela globokih GP vpliva večje število skritih slojev, uporaba apriornega verjetja in različno število induciranih točk.

6.3.1 Regresorski vektor

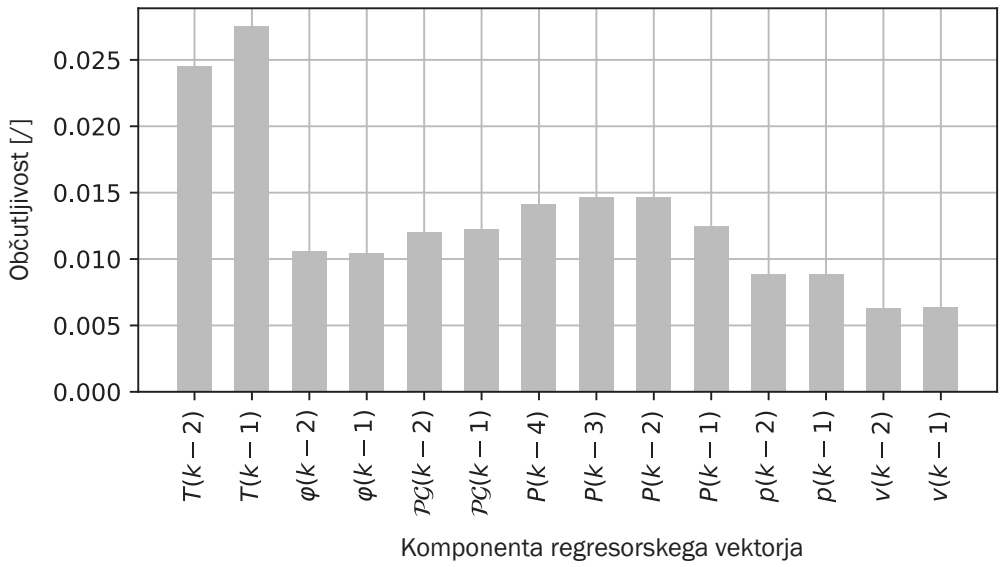
Iskanje optimalne strukture regresorskega vektorja začnemo z enačbo (6.1). V naslednjem koraku regresorju dodamo ali odvzamemo prostostne stopnje in rezultate

lastnosti ARD primerjamo. Postopek ponavljamo dokler nismo zadovoljni z napako NRMSE, časom izračuna, predvsem pa z negotovostjo napovedi. Na koncu izberemo »optimalno« obliko regresorskega vektorja. V tej fazi laboratorijskega dela uporabljamo le model GP-LVM in lastnost ARD kovariančne funkcije. Na prejšnjem ilustrativnem primeru smo namreč ugotovili, da numerična stabilnost modela GP-LVM ni močno odvisna od inicializacije modela. Kadar najdemo optimalen regresorski vektor, je dober za oba modela.

Poudarimo, da ni nujno, da model z rezultati odraža tudi naravne fizikalne lastnosti. Če na podlagi fizikalnega znanja, osebnih izkušenj ali intuicije pričakujemo, da bo napoved temperature odvisna od vseh izmerjenih meteoroloških parametrov, še ne pomeni, da je to nujno res. Oba modela sta namreč numerična in niti ne poznata niti ne razumeta naravnih zakonitosti. Modela pač iščeta preslikavo med vhodnimi in izhodnimi vrednostmi sistema. V nadaljevanju nas zato ne bo motilo, če se izkaže, da je regresorski vektor neodvisen od hitrosti vetra na višini 10 m ipd.

Osnovna oblika regresorja

Prvi korak v optimizaciji regresorskega vektorja je analiza občutljivosti modela GP-LVM na komponente iz enačbe (6.1). Občutljivost modela je dana z vrednostjo hiperparametrov kovariančne funkcije. Rezultati ARD lastnosti kovariančne funkcije so prikazani na Sliki 6.3. Pomembnejše računske lastnosti modela najdemo v Preglednici 6.1.



Slika 6.3: Občutljivost modela GP-LVM na komponente regresorskega vektorja za $\dim(\mathbf{z}_i) = 14$.

Opazimo, da imata obe zakasnitvi hitrosti vetra v regresorskem vektorju najmanjši uteži. Opazimo tudi, da je prva zakasnitev globalnega sončnega sevanja $P(k-1)$ nekoliko slabše utežena v primerjavi z ostalimi zakasnitvami iste spremenljivke. To

delno potrjuje razmišljanje iz poglavja 6.2.1, kjer govorimo, da naj bi na temperaturo najbolj vplivala jakost globalnega sončnega sevanja eno ali dve uri nazaj [18]. Iz slike je jasno tudi, da imata zakasnitvi izhodne vrednosti $T(k-2)$ in $T(k-1)$ največji uteži.

Relativno velika občutljivost modela se pokaže tudi na edini meteorološki parameter, ki ni povsem objektivno definiran, t.j. stabilnost atmosfere po Pasquill-Girardovi lestvici.

Z analizo Slike 6.3 se zdi, da bi bilo smiselno upoštevati še več zakasnitev. V sklopu eksperimentalnega dela smo zato v regresorskem vektorju upoštevali kar 5 zakasnitev vsake spremenljivke ($\dim(\mathbf{z}_i) = 30$). Ugotovimo, da so vse dodatne zakasnitve slabše utežene in posledično ne prispevajo k boljšemu modelu. Zato za izhodišče privzamemo obliko regresorskega vektorja iz enačbe (6.1).

Optimizacija regresorja

Iz Slike 6.3 je jasno, da smo bili že v prvo kar uspešni z določevanjem prostostnih stopenj regresorskega vektorja. Nobena komponenta namreč nima uteži bistveno manjše v primerjavi z ostalimi. Model se s takšnim regresorskim vektorjem dobro nauči preslikave kar potrjuje napaka NRMSE v Preglednici 6.1. Kljub temu preverimo ali lahko brez velikega poslabšanja napake NRSME in časa izračuna kakšno komponento regresorja (6.1) zanemarimo.

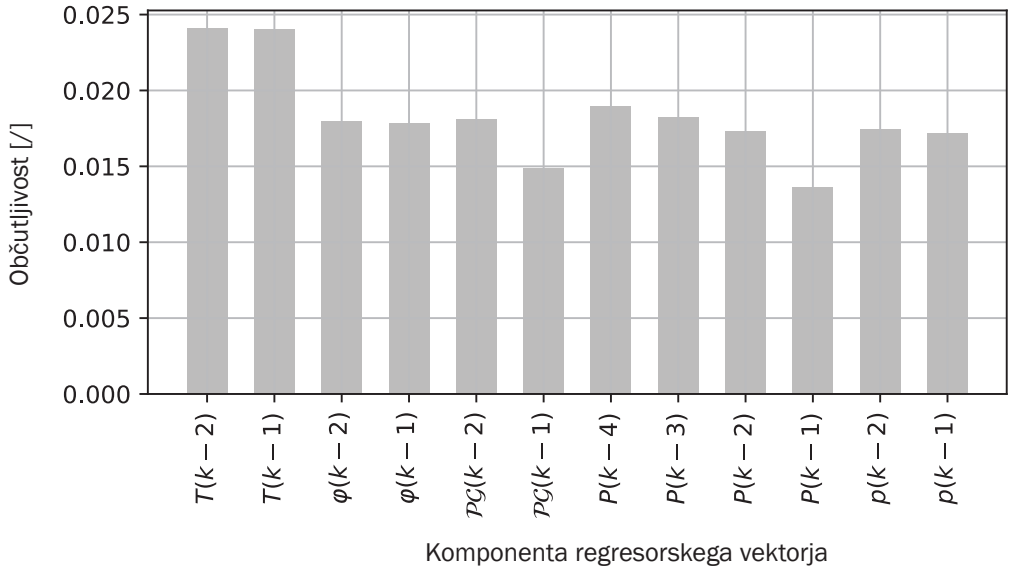
Omenili smo, da imata obe zakasnitvi hitrosti vetra majhno utež v primerjavi z ostalimi prostostnimi stopnjami v enačbi (6.1). Regresorski vektor zato skrajšamo tako, da v celoti zanemarimo meritev hitrosti. Nov vektor vhodnih podatkov ima tedaj 12 komponent

$$\begin{aligned}\mathbf{z}_i = (&T(k-2), T(k-1),\\ &\varphi(k-2), \varphi(k-1),\\ &\mathcal{PG}(k-2), \mathcal{PG}(k-1),\\ &P(k-4), P(k-3), P(k-2), P(k-1),\\ &p(k-2), p(k-1)).\end{aligned} \tag{6.3}$$

Vrednosti hiperparametrov z novim regresorskim vektorjem so prikazani na Sliki 6.4, pomembnejše karakteristike modela najdemo v Preglednici 6.1.

Opazimo, da občutljivost modela na obe zakasniti izhodne vrednosti $T(k-2)$ in $T(k-1)$ še vedno nekoliko izstopa. Občutljivost modela na ostale komponente je približno enaka za vse. Tudi z novim regresorskim vektorjem, čeprav za dve prostostni stopnji krajši, se model dobro nauči preslikave iz vhodnih v izhodne vrednosti sistema (glej Preglednico 6.1). Napaka NRMSE se pri krajšem regresorskem vektorju spremeni zgolj za 2 tisočinki, kar lahko zanemarimo, medtem ko skoraj 100 sekund krajši čas izračuna ni zanemarljiv.

Pomembno je, da v primerjavi s Sliko 6.3 opazimo ponavljajoče se trende. S tem mislimo na nezanemarljive občutljivosti modela od relativne vlažnosti, tlaka ipd. Če bi se pri spremembi regresorja iz enačbe (6.1) v (6.3) bistveno spremenila občutljivost modela na kakšno izmed komponent regresorskega vektorja, potem je to pomembno opozorilo. V konkretnem primeru k sreči ne opazimo izstopajočih sprememb, razen



Slika 6.4: Občutljivost modela GP-LVM na komponente regresorskega vektorja za $\dim(\mathbf{z}_i) = 12$.

morda enkratne zakasnitve globalnega sončnega sevanja $P(k-1)$, ki ima po novem najmanjšo utež. Ista komponenta je bila v primerjavi z zakasnitvami $P(k-2)$,$P(k-3)$ in $P(k-4)$ že na Sliki 6.3 najmanj utežena. To sicer še ne pomeni, da je zanemarljiva, lahko pa trditev preprosto preverimo.

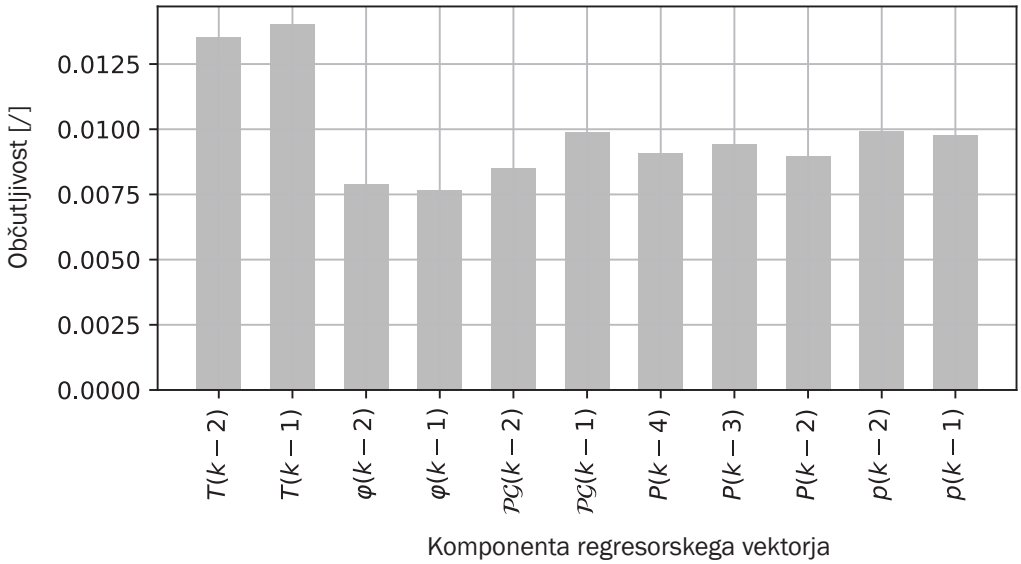
V naslednjem koraku zato z novim regresorskim vektorjem

$$\begin{aligned}\mathbf{z}_i = (&T(k-2), T(k-1),\\ &\varphi(k-2), \varphi(k-1),\\ &\mathcal{PG}(k-2), \mathcal{PG}(k-1),\\ &P(k-4), P(k-3), P(k-2),\\ &p(k-2), p(k-1)),\end{aligned} \tag{6.4}$$

preverimo, ali lahko zanemarimo $P(k-1)$. Rezultati ARD-lastnosti so prikazani na Sliki 6.5 in v Preglednici 6.1.

Čeprav smo iz regresorja (6.3) odstranili le komponento $P(k-1)$ se GP-LVM model precej slabše nauči preslikave (glej NRMSE napako v Preglednici 6.1). Vrednost NRMSE napake se za stotinko poslabša in pri tem čas izračuna kar precej poveča. Čas izračuna je celo daljši kot čas izračuna v primeru začetnega regresorja iz enačbe (6.1) s tremi prostostnimi stopnjami več.

Za optimalno obliko regresorskega vektorja tako privzamemo strukturo v enačbi (6.3). S takim regresorjem je čas izračuna relativno kratek in vse komponente na Sliki 6.4 se zdijo pomembne. Tudi napaka NRMSE iz Preglednice 6.1 je ugodna. Opazimo, da smo pri tem iz prvotnega regresorja v enačbi (6.1) ohranili predpostavko, da je obravnavani dinamični sistem drugega reda. Vsi rezultati v nadaljevanju so torej osnovani na regresorskem vektorju iz enačbe (6.3) z 12 prostostnimi stopnjami.



Slika 6.5: Občutljivost modela GP-LVM na komponente regresorskega vektorja za $\dim(\mathbf{z}_i) = 11$.

$\dim(\mathbf{z}_i)$	Čas računanja [s]	e_{NRMSE}
14	546	0.925
12	447	0.923
11	593	0.911

Preglednica 6.1: Glavne karakteristike modela GP-LVM za različne definicije regresorskih vektorjev.

6.3.2 Število skritih slojev

Podobno kot v sintetičnem eksperimentu v poglavju 5.3.3 se izkaže, da tudi na podatkih realnega eksperimenta z večjim številom slojev ne dobimo boljšega regresijskega modela. Rezultati naključno inicializiranega modela globokih GP z 10 induciranimi točkami in različnim številom skritih slojev so zbrani v Preglednici 6.2.

Št. skritih slojev	Čas računanja [s]	e_{NRMSE}	$\langle 2\sigma \rangle$
1	37	0.941	0.138
2	151	0.937	0.218
3	223	0.930	0.472
4	99	0.917	0.350

Legenda: $\langle 2\sigma \rangle$ povprečna negotovost napovedi modela

Preglednica 6.2: Vpliv števila slojev na model globokih GP.

Opazimo trend, kako se z večanjem števila slojev veča napaka NRMSE, čas izračuna in negotovost napovedi. Izjema je model s štirimi skritimi sloji. Izračun je v tem primeru sicer nepričakovano hiter, a vrednost napake NRMSE v primerjavi z ostalimi slabša.

Ugotavljamo, da večje število skritih slojev v konkretnem primeru ne vodi do boljšega regresijskega modela. Dinamični sistem je najbrž dovolj enostaven, da ga že z enim skritim slojem povsem zadovoljivo identificiramo. Vsi rezultati modela globokih GP v nadaljevanju so zato rezultati modela z enim skritim slojem.

6.3.3 Naključna inicializacija modela globokih GP

Analizo izmerjenih meteoroloških spremenljivk opravimo z regresorskim vektorjem v enačbi (6.4). Inicializacija vseh hiperparametrov, latentnih in induciranih spremenljivk naj bo zaenkrat naključna. Omejimo se na le en skriti sloj v modelu globokih GP. Rezultati obeh modelov (model globokih GP in model GP-LVM) so na Sliki 6.6, kjer prikazujemo le majhen del iz domene testnih podatkov, t.j. le 200 testnih meritev na intervalu $k \in [8000, 8200]$. V nasprotnem primeru so namreč grafi povsem neberljivi in bi jih le stežka komentirali. Črtkana črta na grafu prikazuje izmerjen odziv sistema, polna črta napoved modela in s sivo je obarvan interval 95 % zaupanja.

S primerjavo modelov na Sliki 6.6 in rezultatov v Preglednici 6.3, ugotovimo, da se z naključno inicializacijo model globokih GP precej bolje nauči preslikave med vhodnimi in izhodnimi vrednostmi. Napaka NRMSE se razlikuje za kar dve desetinki. Tudi s časovnega stališča je ugodnejši model globokih GP, kar je razumljivo, saj uporabljamo manj induciranih točk kot v modelu GP-LVM. A najpomembneje je, da je negotovost napovedi modela globokih GP za skoraj tretjino manjša. Rezultati analize so v splošnem bolj v korist modela globokih GP.

Model	Št. induciranih točk	Čas računanja [s]	e_{NRMSE}	$\langle 2\sigma \rangle$
GP-LVM	100	400	0.923	0.326
Globoki GP	10	172	0.947	0.113

Legenda: $\langle 2\sigma \rangle$ povprečna negotovost napovedi modela

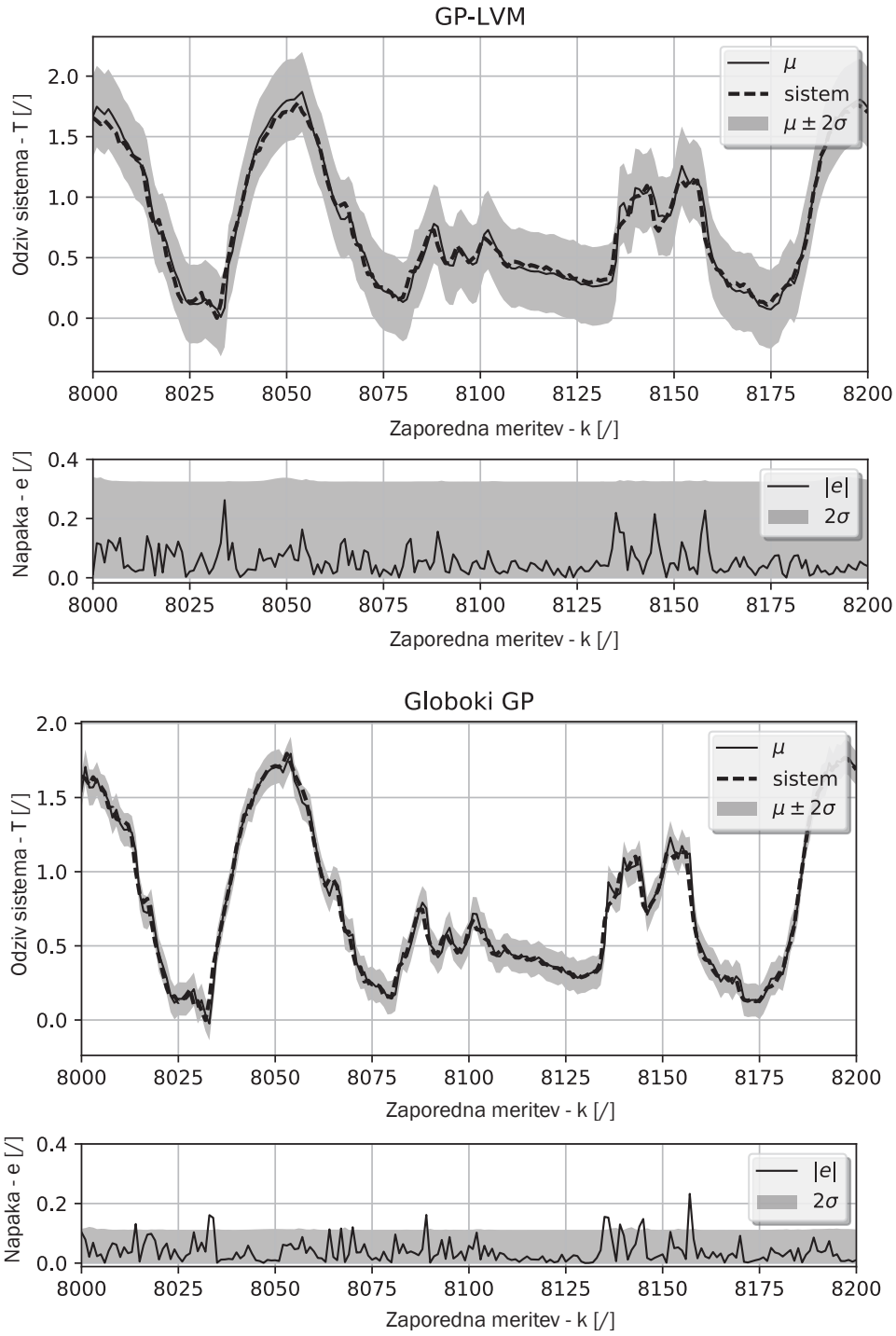
Preglednica 6.3: Glavne karakteristike obeh regresijskih modelov z naključno inicializacijo.

Izračun modela GP-LVM z naključno inicializacijo smo 100 krat ponovili. Da bi našli boljši lokalni minimum smo uporabili tudi premišljeno inicializacij. Identificirali smo le vrsto ekvivalentnih lokalnih minimumov, ne pa bistveno boljših. Model globokih GP z naključno inicializacijo smo ponovili tisočkrat in prikazujemo najboljše rezultate.

Več pozornosti v nadaljevanju namenimo premišljeni inicializaciji modela globokih GP. Spomnimo, da v sintetičnem eksperimentu iz poglavja 5.3 nismo uspeli poiskati boljšega minimuma modela GP-LVM, medtem ko smo bili z modelom globokih GP pri tem uspešni. Še pred tem opozorimo, da smo sodeč po napaki e globokega GP-modela s Slike 6.6 poiskali zadovoljivo dober minimum. Negotovost napovedi je relativno majhna zato ne pričakujemo bistveno boljšega regresijskega modela.

6.3.4 Premišljena inicializacija modela globokih GP

S Slike 6.6 lahko sklepamo, da numeričen model globokih GP ni zašel v globalni minimum. Sodeč po izrisu iz intervala za $k =\in [8000, 8200]$ dobimo občutek, da bi lahko z



Slika 6.6: Rezultati naključne inicializacije obeh modelov za $k \in [8000, 8200]$.

boljšo inicializacijo še vsaj malo zmanjšali negotovost napovedi modela globokih GP.

V sklopu premišljene inicializacije smo preizkusili inicializacijo vseh možnih parametrov. To pomeni različne vrednosti inicializacije horizontalnih skalirnih faktorjev in variance kovariančne matrike na vsakem nivoju. Tudi različne inicializacije induciranih in latentnih spremenljivk. S tem mislimo na enakomerne porazdelitve v različnih intervalih, npr. $[-1, 1]$, $[0, 2]$, ipd.

Neodvisno od parametra, ki smo ga inicializirali, nismo uspeli najti boljšega minimuma, kot ga najdemo že z naključno optimizacijo na Sliki 6.6. Identificirali smo več kombinacij inicializacij parametrov s katerimi se približamo tem rezultatom, a nobena izmed njih ni opazno boljša.

6.3.5 Dodatni eksperimenti

V serijah dodatnih eksperimentov smo skušali boljši rezultat doseči tudi z definicijo apriornega verjetja nad hiperparametri modela globokih GP. Rezultati so bili znova zgolj primerljivo dobri rezultatom na Sliki 6.6.

Tudi večje število induciranih točk ni imelo opaznega vpliva na kvaliteto regresijskega modela. V Preglednici 6.4 so zbrani rezultati modela globokih GP z 10 in 100 induciranimi točkami. Pri tem se čas izračuna podaljša za faktor 15, a je sprememba napake NRMSE in negotovosti napovedi zanemarljiva.

Model	Št. induciranih točk	Čas računanja [s]	e_{NRMSE}	$\langle 2\sigma \rangle$
Globoki GP	10	172	0.947	0.113
	100	2596	0.949	0.109

Legenda: $\langle 2\sigma \rangle$ povprečna negotovost napovedi modela

Preglednica 6.4: Odvisnost modela globokih GP od števila induciranih točk.

Čeprav smo v poglavju 6.3.2 ugotovili, da večje število skritih slojev ne pomeni nujno boljšega regresijskega modela, pa smo preverili, če večje število parametrov večslojnega modela morda vodi do boljšega lokalnega minimuma. Premišljeno smo inicializirali model globokih GP z dvema skritima slojema, a prav tako nismo uspeli dobiti boljšega lokalnega minimuma.

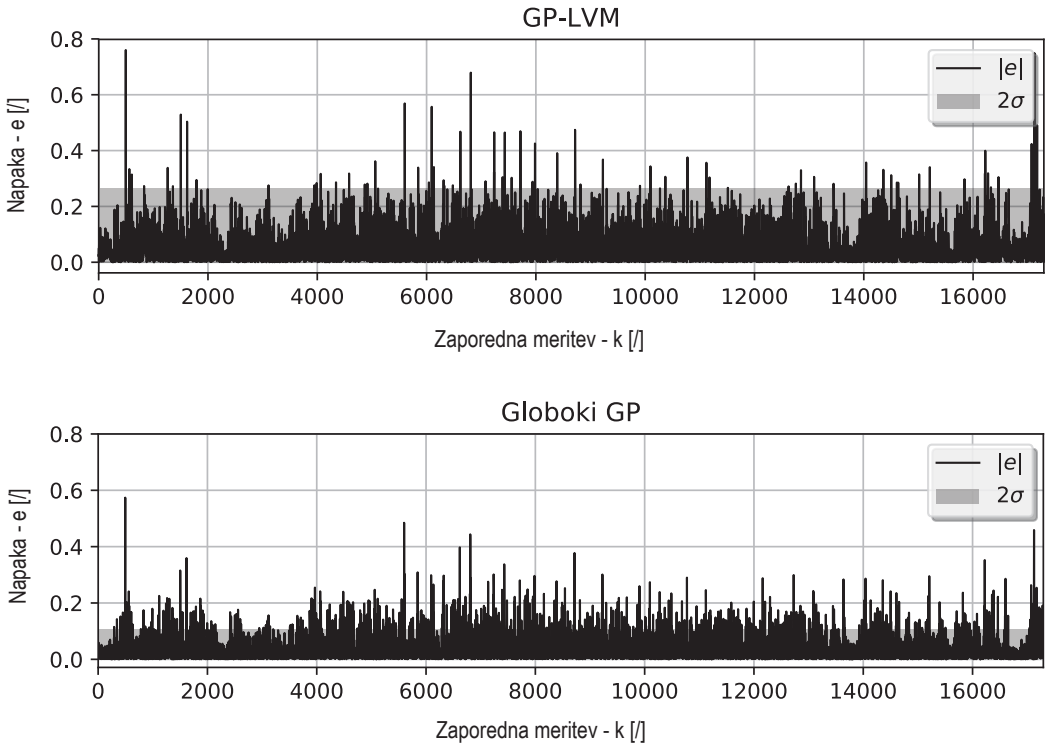
6.4 Zaključki

Oba modela, model globokih GP in model GP-LVM, se dobro naučita zveze med vhodnim prostorom $\mathbf{Z}$ in izhodnimi vrednostmi $\mathbf{y}$. Rezultati naključne inicializacije modela globokih GP so dobri, a smo po zaključkih sintetičnega eksperimenta iz poglavja 5.3.4 napačno predvidevali, da bomo s premišljeno inicializacijo dosegli še bolšje. Boljšega lokalnega minimuma z modelom GP-LVM in naključno inicializacijo tudi po 100 ponovitvah nismo našli.

Nekoliko več truda smo vložili v inicializacijo modela globokih GP. Efektivno smo s premišljeno inicializacijo globokega GP skušali poiskati boljši lokalni minimum, kot ga najde naključno inicializiran model. Kljub različnim kombinacijam premišljene inicializacije nismo uspeli zmanjšati negotovosti napovedi.

V splošnem je v konkretnem primeru model globokih GP boljši in hitrejši v primerjavi z modelom GP-LVM. V obeh primerih se izkaže, da že naključno inicializirana modela poiščeta zelo dober lokalni minimum, če ne globalnega. Tudi večje število induciranih točk ni vodilo do boljšega regresijskega modela.

V sklopu eksperimenta smo dosegli zadani cilj - z modelom globokih GP smo identificirali nelinearni dinamični sistem in uspešno napovedali temperaturo za en korak vnaprej. Vsa teoretična dognanja iz predhodnih poglavij smo pokazali na praktičnem primeru. Na Sliki 6.6 prikazujemo uspešnost modela globokih GP pri identifikaciji nelinearnega dinamičnega sistema. Model se je uspešno naučil preslikave med vhodnimi in izhodnimi vrednostmi. Z negotovostjo napovedanih vrednosti modela globokih GP na Sliki 6.7 smo zadovoljni, saj vidimo, da je napaka napovedi e znotraj intervala 95 % zaupanja. Vidimo tudi precej manjšo negotovost napovedanih vrednosti modela globokih GP kot modela GP-LVM.



Slika 6.7: Napaka in negotovost napovedi obeh modelov za vse testne podatke.

7 Zaključki

Predstavili smo modele na podlagi GP in nato še posebno izpeljanko modela latentnih spremenljivk. Ugotovili smo, da GP-model latentnih spremenljivk ni analitično izračunljiv v okviru učnega postopka MAP, zato smo dodatno k temu vpeljali model variacijskega GP-LVM. S tem smo dobili ne le analitično izračunljiv model, ampak tudi zmanjšali računsko kompleksnost modela. Nazadnje smo z gnezdenjem modelov GP-LVM tvorili še model globokih GP.

Uporabnost modela globokih GP smo preverili na ilustrativnem primeru ter rezultate primerjali z modelom GP-LVM. Ugotovili smo, da se oba modela dobro naučita zveze med vhodnimi in izhodnimi vrednosti opazovanega sistema. Model globokih GP je v splošnem hitrejši in negotovost napovedi manjša v primerjavi z rezultati modela GP-LVM. V zameno je potrebnega nekoliko več truda pri premišljeni inicializaciji hiperparametrov, induciranih in latentnih spremenljivk. Ugotovili smo tudi, da v splošnem večje število skritih slojev modela globokih GP ne pomeni nujno boljšega lokalnega minimuma numeričnega modela. Z večjim številom induciranih točk smo še dodatno izboljšali model globokih GP, a pri tem močno povečali računsko kompleksnost modela in posledično čas izračuna.

Tudi rezultati analize meteoroloških spremenljivk na primeru napovedi temperature v okolici JEK so v splošnem bolj v korist modelu globokih GP. Z večjim številom podatkov in večimi prostostnimi stopnjami regresorskega vektorja smo to tudi pričakovali. Čeprav je sistem bolj kompleksen od ilustrativnega primera, se izkaže, da večje število skritih slojev tudi tokrat ne vodi do boljših rezultatov. Ugotovili smo, da premišljena inicializacija in večje število induciranih točk ne vodita nujno do boljšega lokalnega minimuma optimizacijskega postopka numeričnega modela. Z rezultati globokega GP modela smo povsem zadovoljni že z naključno inicializacijo.

Z uspešno uporabo modela globokih GP na ilustrativnem in dejanskem primeru smo dosegli cilj magistrskega dela. Demonstrirali smo uporabo modela globokih GP in pri tem pokazali, da se model globokih GP hitreje in bolje nauči preslikave med vhodnimi in izhodnimi vrednostmi sistema. Napovedane vrednosti imajo manjšo negotovost.

V sklopu magistrskega dela se osredotočamo le na napoved vrednosti za en korak vnaprej. V nadaljevanju dela bi radi model globokih GP uporabili tudi za dolgoročne napovedi. Poleg temperature pri tleh, bi radi napovedali tudi temperaturo na različnih višinah in s tem napovedali temperature za celoten profil v območju interesa. Seveda samo temperatura ne zadostuje za dobro napoved gibanja radioaktivnega oblaka okoli

JEK. Za detajlno poznavanje mikroklime potrebujemo tudi vrednosti drugih meteoroloških parametrov. Dolgoročno bomo model globokih GP uporabili tudi za napoved drugih meteoroloških spremenljivk.

8 Literatura

[1] P. Potočnik, B. Soldo, G. Šimunović, T. Šarić, A. Jeromen, E. Govekar: *Comparison of static and adaptive models for short-term residential natural gas forecasting in Croatia*. Applied Energy **129**(2014) str. 94 – 103.

[2] B. Soldo, P. Potočnik, G. Šimunović, T. Šarić, E. Govekar: *Improving the residential natural gas consumption forecasting models by using solar radiation*. Energy and Buildings **69**(2014) str. 498 – 506.

[3] P. Potočnik, E. Strmčnik, E. Govekar: *Linear and Neural Network-based Models for Short-Term Heat Load Forecasting*. Strojniški vestnik - Journal of Mechanical Engineering **61**:9 (2015) str. 543–550.

[4] A. C. Damianou: *Deep Gaussian Processes and Variational Propagation of Uncertainty*: *Doktorska disertacija*, University of Sheffield, 2015.

[5] A. Damianou, N. Lawrence: *Deep Gaussian Processes*. V: C. Carvalho, P. Ravikumar (ur.): *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*. AISTATS '13 JMLR W&CP 31, 2013, str. 207–215.

[6] N. D. Lawrence: *The Gaussian Process Latent Variable Model*. `ftp://ftp.dcs.shef.ac.uk/home/neil/gplvmTutorial.pdf` 2006.

[7] C. Rasmussen, C. Williams: *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning MIT Press, Cambridge, MA, USA, 2006.

[8] T. Šuštar: *Določanje reda dinamičnega modela na podlagi Gaussovih procesov*: Diplomsko delo, Univerza v Ljubljani, 2013.

[9] V. Tanko: *Kovariančne funkcije v modelih na podlagi Gaussovih procesov*: Diplomsko delo, Univerza v Ljubljani, 2015.

[10] J. Kocijan: *Modelling and control of dynamic systems using Gaussian process models*. Advances in industrial control. Springer, Cham, 2016.

[11] N. Lawrence: *Probabilistic Non-linear Principal Component Analysis with Gaussian Process Latent Variable Models*. J. Mach. Learn. Res. **6**(2005) str. 1783–1816.

[12] A. C. Damianou, M. K. Titsias, N. D. Lawrence: *Variational Inference for Latent Variables and Uncertain Inputs in Gaussian Processes*. Journal of Machine Learning Research **17**:42 (2016) str. 1–62.

[13] N. D. Lawrence: *Variational Inference Guide*. `ftp://ftp.dcs.shef.ac.uk/home/neil/variationalInference.pdf` 2002, Ogled: 01.05.2017.

[14] Y. Gal, M. van der Wilk, C. Rasmussen: *Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models* V: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (ur.): *Advances in Neural Information Processing Systems 27* Curran Associates, Inc. 2014 str. 3257–3265.

[15] T. Wigren, J. Schoukens: *Three free data sets for development and benchmarking in nonlinear system identification*. V: *2013 European Control Conference (ECC)*, 2013, str. 2933–2938.

[16] R. Pintelon, J. Schoukens: *Frequency Response Function Measurements in the Presence of Nonlinear Distortions*. V: *System Identification*. John Wiley & Sons, Inc., 2005, str. 69–113.

[17] D. Aleksovski, D. Dovžan, S. Džeroski, J. Kocijan: *A comparison of fuzzy identification methods on benchmark datasets*. IFAC-PapersOnLine **49**:5 (2016) str. 31 – 36, 4th IFAC Conference on Intelligent Control and Automation SciencesICONS 2016.

[18] J. Holton: *An Introduction to Dynamic Meteorology*. zvezek 4 od *An Introduction to Dynamic Meteorology* Elsevier Academic Press, 2004.

[19] J. Rakovec, T. Vrhovec: *Osnove meterologije za naravoslovce in tehnike*. tretja izdaja DMFA, 2007.

[20] E. Snelson: *Variable noise and dimensionality reduction for sparse Gaussian processes*. V: *Proc. of UAI-06*. AUAI Press, 2006, .

[21] D. Petelin: *Aproksimacijske metode pri modeliranju dinamičnih sistemov z Gaussovimi procesi*: *Doktorska disertacija*, Mednarodna podiplomska šola Jožefa Stefana, 2014.

[22] I. Goodfellow, Y. Bengio, A. Courville: *Deep Learning*. MIT Press, 2016, `http://www.deeplearningbook.org`.

[23] K. Vafa: *Training and Inference for Deep Gaussian Processes*: *Doktorska disertacija*, Harvard College, 2016.

[24] T. D. Bui, J. M. Hernández-Lobato, D. Hernández-Lobato, Y. Li, R. E. Turner: *Deep Gaussian Processes for Regression Using Approximate Expectation Propagation*. V: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16 JMLR.org, 2016, str. 1472–1481.

[25] J. Hensman, N. D. Lawrence: *Nested Variational Compression in Deep Gaussian Processes*. arXiv preprint arXiv:1412.1370 (2014).

[26] N. D. Lawrence, A. J. Moore: *Hierarchical Gaussian Process Latent Variable Models*. V: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07 ACM, New York, NY, USA, 2007, str. 481–488.

[27] S. Širca: *Verjetnost v fiziki*. DMFA Ljubljana, 2016.

[28] MATLAB: *MATLAB Documentation - System Identification Toolbox*. The MathWorks Inc., 2017.

[29] Z. Dai, A. Damianou, J. Gonzalez: *PyDeepGP*. URL `https://github.com/SheffieldML/PyDeepGP`, Ogled: 01.05.2017.

[30] A. Damianou, N. Lawrence: *deepGP v.1.0*. URL `https://github.com/SheffieldML/deepGP`, Ogled: 01.05.2017.

[31] A. Damianou, N. Lawrence: *Deep Gaussian Processes (deep GPs)*. URL `http://git.io/A51g`, Ogled: 01.06.2017.

[32] *Pasquill Stability Classes*. URL `https://www.ready.noaa.gov/READYpgclass.php`, Ogled: 01.06.2017.

9 Priloga

A Izpeljava KL divergence

Kako dober približek spodnje meje v variacijskem modelu GP-LVM zagotavlja funkcija $q(\mathbf{\Theta})$ pove *Kullback-Leibler* divergenca [13].

Pravilo produkta dveh verjetnosti pravi

$$p(\mathbf{Y}, \mathbf{\Theta}) = p(\mathbf{Y}|\mathbf{\Theta})p(\mathbf{\Theta}),$$
$$p(\mathbf{Y}, \mathbf{\Theta}) = p(\mathbf{\Theta}|\mathbf{Y})p(\mathbf{Y})$$

od koder očitno velja

$$p(\mathbf{\Theta}|\mathbf{Y})p(\mathbf{Y}) = p(\mathbf{Y}|\mathbf{\Theta})p(\mathbf{\Theta}). \tag{A.1}$$

Spodnja meja robne verjetnosti je z uporabo Jensenove neenačbe

$$\begin{aligned}
\log p(\mathbf{Y}) &= \log \int p(\mathbf{Y}, \mathbf{\Theta}) \mathrm{d}\mathbf{\Theta} \\
&= \log \int q(\mathbf{\Theta}) \frac{p(\mathbf{Y}, \mathbf{\Theta})}{q(\mathbf{\Theta})} \mathrm{d}\mathbf{\Theta} \\
&\geq \int q(\mathbf{\Theta}) \log \frac{p(\mathbf{Y}, \mathbf{\Theta})}{q(\mathbf{\Theta})} \mathrm{d}\mathbf{\Theta} \\
&= \int q(\mathbf{\Theta}) \log \frac{p(\mathbf{Y}|\mathbf{\Theta})p(\mathbf{\Theta})}{q(\mathbf{\Theta})} \mathrm{d}\mathbf{\Theta}.
\end{aligned} \tag{A.2}$$

V enačbi (A.2) uporabimo identito (A.1) in dobimo

$$\log p(\mathbf{Y}) \geq \int q(\mathbf{\Theta}) \log p(\mathbf{Y}) \mathrm{d}\mathbf{\Theta} + \int q(\mathbf{\Theta}) \log p(\mathbf{\Theta}|\mathbf{Y}) \mathrm{d}\mathbf{\Theta} - \int q(\mathbf{\Theta}) \log q(\mathbf{\Theta}) \mathrm{d}\mathbf{\Theta}. \tag{A.3}$$

V prvem členu prepoznamo pričakovano vrednost $\log p(\mathbf{Y})$ pri porazdelitvi $q(\mathbf{\Theta})$ in je torej kar enak $\log p(\mathbf{Y})$

$$\log p(\mathbf{Y}) \geq \log p(\mathbf{Y}) + \int q(\mathbf{\Theta}) \log p(\mathbf{\Theta}|\mathbf{Y}) \mathrm{d}\mathbf{\Theta} - \int q(\mathbf{\Theta}) \log q(\mathbf{\Theta}) \mathrm{d}\mathbf{\Theta}. \tag{A.4}$$

Od tod je razlika med pravo vrednostjo spodnje meje verjetnosti in njeno aproksimacijo dobljeno aproksimacijo $q(\boldsymbol{\Theta})$ ter uporabo Jensenove neenačbe

$$\mathrm{KL}(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\mathbf{Y})) = \int q(\boldsymbol{\Theta}) \log q(\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{\Theta} - \int q(\boldsymbol{\Theta}) \log p(\boldsymbol{\Theta}|\mathbf{Y}) \mathrm{d}\boldsymbol{\Theta}, \tag{A.5}$$

Kullback-Leiblerjeva divergenca dveh porazdelitev. Člen KL je vedno pozitiven, razen kadar je $q(\boldsymbol{\Theta}) = p(\boldsymbol{\Theta}|\mathbf{Y})$, tedaj je njegova vrednost 0 in se neenakost spremeni v enakost [13].