Jindřich Duník

IDENTIFIKACE SYSTÉMŮ A FILTRACE

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

Jindřich Duník

IDENTIFIKACE SYSTÉMŮ A FILTRACE

FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

Identifikace systémů a filtrace

Jindřich Duník

Skripta jsou přepracovaným a rozšířeným vydáním původních skript Identikace systémů a filtrace napsaných v roce 1994 prof. Ing. Miroslavem Šimandlem, CSc.

Grafický návrh obálky:
Tereza Saitzová

Vydala:
Západočeská univerzita v Plzni
Univerzitní 2732/8, 301 00 Plzeň

2. přepracované a rozšířené vydání
Plzeň 2018

ISBN 978-80-261-0775-0
ISBN 978-80-261-0940-2 (tištěné vydání)

https://doi.org/10.24132/ZCU.2018.07750

Předmluva k druhému vydání

Pozorování a poznávání prostředí, které nás obklopuje, a tvorba popisu tohoto prostředí je nedílnou součástí lidského snažení. Již od prvopočátku se lidstvo snaží najít popis okolního prostředí za účelem předpovědi jeho budoucího vývoje, která následně umožní efektivněji plánovat lidskou činnost. Jako příklad výše zmíněného může být uvedeno počasí, které významným způsobem ovlivňuje kvalitu lidského života. Proto se lidé již odnepaměti snaží na základě pozorování popsat vývoj počasí formou různých pranostik[1] a na jejich základě vývoj počasí předvídat.

S rozvojem přírodních i technických věd však slovní popis okolního prostředí přestal dostačovat a pozornost se začala upírat na popis pozorovaných jevů a procesů formou různých matematických modelů, které byly založeny nejen na různých fyzikálních, chemických a matematických principech, ale i na zpracování dostupných pozorování. Tento moment lze tak považovat za počátek rozvoje technik a metod *identifikace a estimace*, které se zaměřují na nalezení struktury (nebo-li formy) a parametrů matematických modelů okolních jevů, procesů a systémů na základě měřených dat.

Cílem těchto skript je seznámit čtenáře se základními myšlenkami, koncepty a metodami identifikace a estimace. Předkládaná skripta je možné chápat jako druhé vydání skript "Identifikace systémů a filtrace" napsané v roce 1994 prof. Ing. Miroslavem Šimandlem, CSc. Text skript byl nejen revidován, ale i výrazněji rozšířen tak, aby reflektoval vývoj vědní disciplíny i současnou náplň přednášek stejnojmenného předmětu vyučovaného na Katedře kybernetiky, Fakulty aplikovaných věd, Západočeské univerzity v Plzni, v rámci magisterského studia. Skripta jsou zpracovány podrobněji než odpovídá přednášené látce, proto některé části mohou být užitečné i posluchačům doktorského studia.

Tvorba skript by se neobešla bez mnoha cenných rad, podnětů a postřehů jak od kolegů, tak i studentů předmětu, za což jim patří velké díky. Poděkování samozřejmě patří i rodině za trpělivost a podporu v přípravě textu. I přes mnohá čtení těchto skript, stále se jistě najdou mnohé překlepy, stylistické i věcné chyby apod. Proto, jakékoliv podněty či komentáře k obsahu skript jsou vítány na e-mailové adrese "dunikj@kky.zcu.cz".

V Plzni, 2020 Jindřich Duník

[1] Slovo *pranostika* je odvozeno z řeckého slova *prognósis* znamenající *předpověď*.

Předmluva k prvnímu vydání

Objevy sedmnáctého a osmnáctého století vedly k přesvědčení, že přírodní jevy lze popisovat pomocí jednoduchých matematických předpisů. Avšak v tomto století zřetelně vyšlo najevo, že pro řešení problémů spojených s novými technologiemi především v komunikacích a řízení je nutné explicitní modelování neurčitosti. Byly vyřešeny důležité inženýrské problémy jako kódování, filtrace, predikce, automatické řízení. Nové výsledky v těchto oblastech umožnily pokrok nejenom v pozemských aktivitách lidstva, ale též při průzkumu vesmíru (např. satelitní komunikace). Ukázalo se, že revoluční techniky modelování založené na koncepci nacházení matematických modelů přímo z měřených dat (identifikace systémů a filtrace) přebírají význam role fyzikálně motivovaných postupů při tvorbě matematických modelů (matematické modelování).

V této souvislosti se ukazuje, že by měla být věnována mnohem větší pozornost realitě, která je vytvářena mutací různorodých procesů v měnícím se prostředí. V takovém případě je samozřejmě nezbytné vhodné modelování neurčitosti respektující složitost popisovaných situací. Rozumnost modelování, nacházení a ladění modelů v reálném čase na základě měřených dat je pak evidentní. Často se nabízí i analogie s funkcí mozku a procesem učení se z experimentu.

Cílem těchto skript je poskytnout úvod do kybernetické teorie zabývající se stavbou modelů a zpracováním signálů, která tvoří základnu pro algoritmy filtrace, predikce, automatického řízení, adaptace i učení.

Skriptum se skládá ze dvou dílů. První díl se zabývá identifikací systémů a druhý díl filtrací. Oba díly jsou zpracovány tak, že je lze studovat nezávisle na sobě. Obsah skript tvoří přednášky pro studenty 4. ročníku oboru Kybernetika a řídicí techniky, které proběhly v minulých letech v rámci předmětů Identifikace systémů a Nelineární filtrace a od akademického roku 1991/92 v předmětu Identifikace systémů a filtrace na Západočeské univerzitě, Fakultě aplikovaných věd. Některé pasáže byly součástí přednášek pro studenty v rámci doktorandského (kandidátského) studia vědeckého oboru Kybernetika.

Na závěr bych rád vyjádřil poděkování paní Haně Němečkové za včasné a trpělivé přepisování rukopisu a realizaci nesčetných oprav při vědomí termínu odevzdání práce.

Listopad, 1994 Miroslav Šimandl

Historie změn v obsahu skript

1994: První vydání skript (listopad, 1994).

2018: Druhé vydání skript (únor, 2018). Korekce jazyka a vzorců. Hlavní změny:

- První díl: Rozšíření popisu metody nejmenších čtverců o uvažování zpětné vazby, rovnostní a nerovnostní omezení a vlastnosti odhadu. Rozšíření ilustrace metody chyby predikce. Rozšíření diskuze k numerickému ošetření rekurzivních identifikačních algoritmů.
- Druhý díl: Přidáno alternativní odvození Kalmanova filtru. Přidána diskuze k vlastnostem, konvergenci a konzistenci odhadu Kalmanova filtru. Přidán popis moderních lokálních estimátorů (diferenční a unscentovaný filtr). Přidán popis metody bodových mas. Přidán popis korelační metody pro odhad (identifikaci) vlastností šumů stavového modelu.

2020: Korigované druhé vydání skript (duben, 2020). Korekce jazyka a vzorců. Hlavní změny:

- První díl: Přidány definice týkající se popisu stochastického procesu. Přidán popis metody podprostorů pro přímou identifikaci parametrů lineárních stavových modelů. Přidán popis metod identifikace nelineárních vstupně-výstupních modelů.
- Druhý díl: Rozšíření ilustrace Kalmanova filtru. Rozšíření popisu diferenčního filtru. Rozšíření popisu o využití estimačních technik v úloze identifikace systémů. Přidán popis aplikací estimačních metod v oblasti navigace.

DÍL PRVNÍ
IDENTIFIKACE SYSTÉMŮ

Obsah

Poznámka: Tato publikace obsahuje i druhý díl, který následuje za poslední stranou dílu prvního.

Kapitola 1

Úvod

Identifikace systémů se zabývá hledáním matematických modelů reálných systémů z experimentálních dat. Dosahuje širokého uplatnění v nejrůznějších sférách lidské činnosti. V oblasti automatického řízení se metody identifikace systémů používají k získání vhodných modelů pro syntézu regulátorů, návrh algoritmů predikce nebo pro simulaci. V oblasti zpracování signálů (např. v komunikacích, geofyzikálním inženýrství a mechanice) jsou modely získané identifikací používány pro spektrální analýzu, detekci poruch, rozpoznávání obrazů, filtraci či adaptivní filtraci atd. Identifikace systémů se rovněž výrazně prosazuje i v netechnických disciplínách jako např. v biologii, ekonometrii či ekologii.

1.1 Typy modelů

Modely dynamických systémů mohou být velmi rozmanité. Mezi základní typy modelů patří:

- Mentální, intuitivní či slovní modely. Tento typ modelu používáme např. při řízení auta („stlačením brzdy snižujeme rychlost“, „otáčením volantu měníme směr jízdy“ atd.)
- Tabulky nebo grafy. Typickým příkladem grafické reprezentace modelů je logaritmická frekvenční charakteristika. Obdobně zápis vztahu ceny a spotřeby jistého zboží v tabulce může být též chápán jako model.
- Matematické modely. Ačkoliv tabulky a grafy mohou být vnímány též jako „matematické“ modely, zde omezíme třídu matematických modelů na diferenciální a především diferenční rovnice. Takové modely jsou vhodné jak pro analýzu a predikci chování dynamických systémů, tak i pro návrh regulátorů a filtrů. Tento typ modelu bude převážně využíván ve skriptu. Poznamenejme, že v souvislosti s matematickými modely můžeme dále mluvit o lineárních a nelineárních modelech, deterministických a stochastických modelech, časově invariantních a časově variantních (nebo-li t-invariantní a t-variantních) modelech, modelech se soustředěnými a rozloženými parametry atd.

1.2 Matematické modelování a identifikace systémů

Jak již bylo řečeno, matematické modely dynamických systémů jsou užitečné z mnoha důvodů. V podstatě jsou známy dva základní přístupy ke konstrukci matematických modelů:

- Matematické modelování. Charakteristickým rysem matematického modelování je využívání fyzikálních, chemických, ekonomických a jiných známých zákonů k popisu dynamického chování zkoumaných systémů s cílem vytvoření matematického modelu bez nutnosti využití měřených veličin. Jedná se tedy o analytický přístup.
- Identifikace systémů. Charakteristickým rysem identifikace systémů je využívání různorodých experimentů prováděných na sledovaném systému, získávání reálných měřených dat a na tomto základě vytváření matematických modelů vyhovujících co nejlépe naměřeným veličinám. Jedná se tedy o experimentální přístup.

V mnoha případech jsou sledované systémy a procesy tak složité, že není možné zkonstruovat rozumný model pouhým použitím matematického modelování (např. použitím kinematických rovnic, zákona o zachování energie apod.). Často tak model založený na matematickém modelování obsahuje jistý počet neznámých parametrů, které nelze určit bez analýzy měřených dat (např. síla větru v dané lokaci a daném období) a je nutné provést jejich odhad pomocí vhodné identifikační metody. Oblasti matematického modelování a identifikace systémů jsou tak v mnoha případech v realitě těsně spjaty.

Modely získané identifikací mají na rozdíl od modelů získaných výhradně matematickým modelováním (využitím např. fyzikálního pohledu) následující vlastnosti:

- Je relativně snadné je navrhnout a využívat.
- Jejich platnost je limitována (jsou platné pro určitý pracovní bod, určitý typ vstupu atd.).
- Nabízí menší vysvětlující charakter o chování systému, protože získané parametry modelu jsou mnohdy značně komplikovanou (a *neznámou*) funkcí skutečných parametrů systému, které pro nás mají zřejmý fyzikální, ekonomický, či chemický.

Identifikace systémů není snadno ovládnutelná a plně dokazatelná metodologie a nelze ji často použít bez spolupráce s odborníkem na daný problém. Uveďme několik důvodů pro toto tvrzení:

- Musí být nalezena vhodná struktura modelu. To může být těžký problém především v situacích, kdy dynamika systému je nelineární.
- Vlastnosti sledovaného procesu či systému se mohou měnit v čase a pak vznikají problémy při popisu založeném na t-invariantním modelu.
- V reálném světě nejsou „dokonalá“ data. Je třeba vzít v úvahu, že naměřená data jsou jistě pod vlivem různých poruch či šumu.
- Může se stát, že není možné měřit proměnné veličiny či signály, které mají stěžejní důležitost pro zdárnou identifikaci systému.

1.3 Jak postupovat při identifikaci systémů

Věnujme se nyní hlavním krokům, které jsou prováděny při identifikaci systému. Nejdříve je provedeno vybuzení systému užitím nějakého vstupního signálu jako jsou např. jednotkový skok, náhodný signál či sinusový signál. Vstupní a výstupní signály systému jsou pak (v určitém určitém intervalu) sledovány a zaznamenány v paměti počítače pro následné informační zpracování. Zpracování spočívá v nalezení vhodného modelu sledovaného procesu, který co nejvíce

vyhovuje zaznamenané vstupní a výstupní sekvenci dat. To vyžaduje stanovit vhodnou formu modelu (typický příklad je lineární diferenční rovnice určitého řádu) a poté použít vhodnou metodu k odhadu neznámých parametrů modelu (reprezentovaných koeficienty diferenční rovnice). Výběr struktury a odhad parametrů modelu se v praxi často provádí iterativně. To znamená, že je vybrána prozatímní struktura modelu a jsou odhadnuty odpovídající parametry. Takto získaný model je pak testován, aby bylo možné rozhodnout, zda se jedná o vhodnou reprezentaci systému. V případě, že se nejedná o vhodnou reprezentaci, je třeba uvažovat alternativní strukturu modelu (např. složitější), provést odhad parametrů, ověřování výsledku atd. Jedná se tedy o iterativní proces hledání přijatelného modelu. Na závěr poznamenejme, že zpracování experimentálních dat může být provedeno jednorázově po naměření všech dat (off-line) nebo okamžitě při příchodu nové informace, nového jednotlivého měření, potom mluvíme o rekurzivní identifikaci (nebo též on-line). Rekurzivní identifikace je zvláště vhodná tam, kde dochází k nějakým změnám v popisu systému (tj. neznámý t-variantní systém). Rekurzivní algoritmy proto tvoří jádro mnoha adaptivních systémů.

1.4 Shrnutí

Tato kapitola byla věnována přiblížení předmětu identifikace a modelování. Bylo rovněž naznačeno, kdy a jak identifikaci systémů využívat. Literatura zabývající se identifikací systémů je velmi bohatá. Z česky psaných publikací se identifikací systémů zcela nebo částečně zabývají [1]-[12], [35]-[36], [57]-[59]. Ze zahraniční literatury pak alespoň uveďme [13]-[21], [60]-[63]. Matematické modelování reprezentuje např. [22],[23]. Velký význam pro rozvoj identifikace na mezinárodním poli mají sympozia "Symposium on System Identification (SYSID)", pořádané mezinárodní federací automatického řízení (International Federation of Automatic Control (IFAC)), které se konají každé tři roky. První bylo uspořádáno v Praze v roce 1967 a zatím poslední v roce 2018 ve Stockholmu. Ze špičkových časopisů, které se věnují identifikaci jmenujme alespoň časopisy IFAC Automatica a IEEE Transactions on Automatic Control.

Kapitola 2

Základní pojmy a úvodní příklady

2.1 Koncepce S,M,I,X

V této kapitole zavedeme základní pojmy, které budou důležité při popisu a analýze identifikačních metod. Důležitost těchto pojmů budeme ilustrovat na jednoduchých příkladech. Výsledek identifikace je ovlivňován přinejmenším následujícími čtyřmi faktory, které budeme diskutovat v této i dalších kapitolách.

- Systém S. V identifikaci pod pojmem systém často rozumíme neznámou fyzikální realitu, kterou chceme poznat a ze které získáváme experimentální (měřená) data. Označujeme ji též pojmem proces, soustava, ale i objekt či reálný systém. Abychom však mohli provést teoretickou analýzu výsledků identifikace, je nutné zavést předpoklady na data. V takovém případě budeme zde užívat pojem systém pro označení nám známého, úplného matematického popisu generátoru dat. V praxi, když pracujeme s reálnými daty, je systém neznámý, chceme jej poznat, identifikovat. Generování dat např. počítačem je naopak založeno na dokonalé znalosti systému. Toto pojetí budeme používat při zkoumání chování různých identifikačních metod v různých situacích.
- Struktura modelu M. Identifikační metody jsou často děleny na neparametrické a parametrické podle toho, zda poskytují neparametrický nebo parametrický model. Neparametrické modely jsou představovány tabulkou, funkcí či křivkou. Jako příklad neparametrického modelu uveďme odezvu na jednotkový skok. Je to křivka, která přináší informaci o charakteristických vlastnostech systému. Jiný příklad neparametrického modelu je frekvenční charakteristika. Nicméně v mnoha případech je výhodné a důležité se zabývat spíše parametrickými modely. Takové modely jsou charakterizovány vektorem parametrů, který budeme označovat Θ. Jestliže Θ může nabývat hodnoty z nějaké množiny přípustných hodnot, dostáváme množinu modelů nebo strukturu modelu $M(\Theta)$. Poznamenejme však, že toto dělení identifikačních metod na parametrické a neparametrické má spíše historické důvody, protože přísně vzato neparametrický model můžeme též parametrizovat.
- Identifikační metoda I. Doposud, jak víme z rozsáhlé literatury věnované identifikaci, bylo navrženo množství identifikačních metod. Nejdůležitější z nich budou uvedeny a diskutovány v těchto skriptech. Poznamenejme, že některé metody mohou být z dnešního pohledu chápány jako stejné, ale jejich původní návrhy byly provedeny pro odlišné struktury modelů, takže mohou být známy pod různými názvy.
- Experimentální podmínky X. Pod symbolem X budeme chápat na obecné úrovni způsob provedení identifikačního experimentu. To jest výběr a generování vstupního signálu,

možný výskyt zpětných vazeb, vzorkovací periodu, předfiltraci dat atd.

Předtím než přejdeme k příkladům, poznamenejme, že ze čtyř pojmů S,M,I,X musíme jako daný a fixovaný chápat systém S. Získávání dat ze systému mohou experimentální podmínky X často do jisté míry ovlivnit. A naopak nezřídka je nutné vzít na vědomí různá omezení znemožňující volný výběr experimentálních podmínek jako např. bezpečnostní požadavky, výrobní podmínky atd. Po získání dat je třeba vybrat identifikační metodu I a strukturu modelu M. Na stejnou množinu dat mohou být použity různé výběry I a M, dokud není dosaženo uspokojivého modelu systému.

2.2 Generátory dat

V této části budeme definovat dva systémy, které budou sloužit v celé 2. kapitole jako generátory dat. Cílem kapitoly bude ukázat použití různých identifikačních metod na tyto systémy formou příkladů.
Předpokládejme, že data jsou generována systémem prvního řádu popsaným diferenční rovnicí

$$y(t) + a_0 y(t-1) = b_0 u(t-1) + e(t) + c_0 e(t-1), \tag{2.2.1}$$

kde $\{e(t)\}$ je posloupnost nezávislých náhodných proměnných identicky distribuovaných[1] (tj. náhodný proces). Střední hodnota náhodných proměnných nechť je nula a variance λ^2. Poznamenejme, že takový typ náhodného procesu se označuje jako *bílý šum*. Dále $u(t)$ je vstup a $y(t)$ výstup systému v čase t.

Dále předpokládejme dvě různé množiny hodnot parametrů. Pro $a_0 = -0,8 \;\; b_0 = 1,0$
$c_0 = 0,0 \;\; \lambda = 1,0$ dostáváme systém S_1

$$S_1 : \qquad y(t) - 0,8y(t-1) = 1,0u(t-1) + e(t) \tag{2.2.2}$$

a pro $a_0 = -0,8 \;\; b_0 = 1,0 \;\; c_0 = -0,8 \;\; \lambda = 1,0$ dostáváme následující systém S_2

$$S_2 : \qquad y(t) - 0,8y(t-1) = 1,0u(t-1) + e(t) - 0,8e(t-1) \tag{2.2.3}$$

který lze alternativně vyjádřit takto:

$$S_2 : \qquad x(t) - 0,8x(t-1) = 1,0u(t-1) \tag{2.2.4}$$

$$y(t) = x(t) + e(t) \tag{2.2.5}$$

Všimněme si, že bílý šum má rozdílné postavení v uvažovaných systémech. V systému S_1 působí jako „chyba rovnice“, zatímco v systému S_2 aditivně ovlivňuje signál $x(t)$, který lze interpretovat jako deterministický výstup. Model (2.2.3) je tak v anglicky psané literatuře často označován jako „output-error model“, což lze přeložit jako model s chybou výstupu.

Poznámka . Lineární vstupně-výstupní model ve formě (2.2.3) je označován jako ARMAX model prvního řádu. V této poznámce vysvětlíme, co zkratka ARMAX v oblasti identifikace systémů znamená, detailnější diskuzi lze najít v kapitole 5. Uvažujme model systému

$$y(t)+a_1y(t-1)+a_2y(t-2)+\ldots = b_1u(t-1)+b_2u(t-2)+\ldots+e(t)+c_1e(t-1)+c_2e(t-2)+\ldots \tag{2.2.6}$$

[1]Pojem identicky distribuované náhodné veličiny značí náhodné veličiny popsané stejnou hustotou pravděpodobnosti.

kde $y(t)$ značí známý výstup systému, $u(t)$ známý vstup systému a $e(t)$ neznámou poruchu ovlivňující systém. Proměnné $a_1, a_2, \ldots, b_1, b_2, \ldots, c_1, c_2, \ldots$ představují parametry modelu. Pak mohou nastat následující speciální případy (modely)

- *autoregresní model* (označovaný zkratkou AR z anglického výrazu „autoregressive model“) ve struktuře
$$y(t) = -a_1 y(t-1) - a_2 y(t-2) - \ldots + e(t) \tag{2.2.7}$$
kde výstup $y(t)$ je dán váženým součtem předchozích *výstupů* (tj. předchozích hodnot *stejného* procesu). AR model tak lze chápat jako filtr s nekonečnou impulsní odezvou (IIR, z anglického „infinite impluse response“).

- *klouzavý průměr* (označovaný zkratkou MA z anglického výrazu „moving average model“) ve struktuře
$$y(t) = e(t) + c_1 e(t-1) + c_2 e(t-2) + \ldots \tag{2.2.8}$$
kde výstup $y(t)$ je dán váženým součtem předchozích *vstupů* (tj. hodnot *jiného* procesu). MA model tak lze chápat jako filtr s konečnou impulsní odezvou (FIR, z anglického „finite impluse response“).

- ARMA model je daný kombinací předchozích modelů a vede na
$$y(t) + a_1 y(t-1) + a_2 y(t-2) + \ldots = e(t) + c_1 e(t-1) + c_2 e(t-2) + \ldots \tag{2.2.9}$$

- ARMAX model (2.2.6) vznikne doplněním předchozího ARMA modelu o (filtrovaný) vstupní signál $u(t)$, kdy písmeno X pochází a z anglického výrazu „eXternal/eXogenous input“.

2.3 Ukázka použití neparametrických metod

V této části využijeme dvě neparametrické metody k identifikaci systému S_1.

Příklad 2.3.1 (Přechodová analýza)
Typický příklad přechodové analýzy je zaznamenání odezvy reálného systému na vstupní signál ve tvaru jednotkového skoku. Obecně, odezva systému obsahuje některé důležité charakteristické vlastnosti systému jako je statické zesílení a časová konstanta. Při malé amplitudě vstupního signálu systému S_1, bude velmi těžké díky vysoké úrovni šumu dedukovat z grafického vyjádření odezvy na jednotkový skok cokoliv o dynamických vlastnostech S_1. Také je vhodné poznamenat, že odezva systému bude odlišná pro různé realizace experimentu, což dále komplikuje možnost identifikace.

Příklad 2.3.2 (Korelační analýza)
Předpokládejme, že model výstupu systému S_1 je
$$y(t) = \sum_{k=0}^{\infty} h(k) u(t-k) + v(t) \tag{2.3.1}$$
kde $\{h(k)\}$ je váhová funkce (váhová sekvence) a $v(t)$ reprezentuje poruchu. Nechť $\{u(t)\}$ je bílý šum se střední hodnotou nula a variancí σ^2, nezávislý na poruchách $v(t)$. Vynásobením

(2.3.1) $u(t-\tau)$ ($\tau > 0$) a zavedením operátoru střední hodnoty $E[\cdot]$ dostaneme

$$r_{yu}(\tau) \stackrel{\triangle}{=} E[y(t)u(t-\tau)] = \sum_{k=0}^{\infty} h(k)E[u(t-k)u(t-\tau)] = \sigma^2 h(\tau) \tag{2.3.2}$$

kde byla využita následující vlastnost bílého šumu

$$\begin{aligned} E[u(t-k)u(t-\tau)] &= 0, \text{ pokud } k \neq \tau \\ &= \sigma^2, \text{ pokud } k = \tau \\ Ev(t)u(t-\tau) &= (E[v(t)])(E[u(t-\tau)]) = 0 \end{aligned}$$

Na základě tohoto vztahu lze koeficienty váhové funkce $\{h(k)\}$ odhadnout podle následujícího výrazu

$$\hat{h}(\tau) = \frac{\frac{1}{N-\tau}\sum_{t=\tau+1}^{N} y(t)u(t-\tau)}{\frac{1}{N}\sum_{t=1}^{N} u^2(t)} \tag{2.3.3}$$

kde N označuje počet dat. Snadno můžeme simulovat systém S_1 s výše definovaným vstupem a graficky znázornit odhadnutou váhovou funkci podle (2.3.3).
Jak lze rychle ověřit, skutečná váhová sekvence systému S_1 je

$$h(k) = 0,8^{k-1} \quad k \geq 1 \quad h(0) = 0$$

Z graficky znázorněné váhové funkce bychom opět velmi těžko zjišťovali parametr pravděpodobně exponenciálního poklesu odhadovaných $\{h(k)\}$.

2.4 Ukázka použití parametrické metody

V následující části budeme identifikovat systémy S_1 a S_2 pomocí jedné z parametrických metod, a to metody nejmenších čtverců. Parametrické metody můžeme obecně charakterizovat jako prostředek pro hledání zobrazení měřených veličin na odhadovaný vektor parametrů.
Uvažujme strukturu modelu M, která je dána diferenční rovnicí

$$M: \qquad y(t) + ay(t-1) = bu(t-1) + \epsilon(t) \tag{2.4.1}$$

Struktura modelu M je stanovena lineární diferenční rovnicí prvního řádu. Vektor parametrů je pak

$$\Theta = \begin{bmatrix} a \\ b \end{bmatrix} \tag{2.4.2}$$

V (2.4.1) je $y(t)$ výstupní signál v čase t, $u(t)$ vstupní signál a $\epsilon(t)$ představuje chybu rovnice, označovanou často jako residuum. Proměnou $\epsilon(t)$ v rovnici (2.4.1) lze chápat jako chybu modelu, protože sotva můžeme doufat, že (2.4.1) s $\epsilon(t) \stackrel{\triangle}{=} 0$ může přesně vyhovovat sekvenci měřených dat. Tudíž $\epsilon(t)$ bude popisovat odchylku v datech od dokonalého (deterministického) lineárního systému prvního řádu a pro danou množinu dat $\{u(1), y(1), u(2), y(2), ..., u(N), y(N)\}$ je $\{\epsilon(t)\}$ funkcí vektoru parametrů Θ. To můžeme jednoduše ukázat přepsáním (2.4.1) na

$$\epsilon(t) = y(t) + ay(t-1) - bu(t-1) = y(t) - [-y(t-1), u(t-1)]\Theta \tag{2.4.3}$$

V následujících kapitolách budeme zavádět různá zobecnění jednoduché struktury modelu (2.4.1). Povšimněme si, že ji lze jednoduše zobecnit na lineární model n-tého řádu pouhým přidáním členů $a_i y(t-i),\ b_i u(t-i)$ pro $i = 1, 2, ..., n$.

Nyní specifikujme identifikační metodu I. V této kapitole se omezíme na metodu nejmenších čtverců. Vektor parametrů je potom určen minimalizací kvadrátů chyby rovnice (residuí $\{\epsilon(t)\}$). To znamená, že dostaneme odhad

$$\hat{\Theta} = \arg\min_{\Theta} V(\Theta) \tag{2.4.4}$$

kde ztrátová funkce $V(\Theta)$ je dána

$$V(\Theta) = \sum_{t=1}^{N} \epsilon^2(t) \tag{2.4.5}$$

Jak je vyjádřeno v (2.4.3), residua jsou funkcí Θ, a tudíž $V(\Theta)$ je definováno pro každou hodnotu Θ.

Pro jednoduchou strukturu modelu (2.4.1), můžeme snadno zapsat explicitní vyjádření závislosti $V(\Theta)$ na Θ. Pro zkrácení zápisu označme $\sum_{t=1}^{N}$ jako $\sum$, pak dostaneme

$$\begin{aligned} V(\Theta) &= \sum [y(t) + ay(t-1) - bu(t-1)]^2 \\ &= [a^2 \sum y^2(t-1) + b^2 \sum u^2(t-1) - 2ab \sum y(t-1)u(t-1)] \\ &+ [2a \sum y(t)y(t-1) - 2b \sum y(t)u(t-1)] + [\sum y^2(t)] \end{aligned} \tag{2.4.6}$$

Odhad Θ je pak získán podle (2.4.4) minimalizací (2.4.6). Bod, pro který funkce nabývá minimální hodnoty můžeme nalézt položením gradientu $V(\Theta)$ rovno nule. To jest

$$\begin{aligned} 0 = \frac{\partial V(\Theta)}{\partial a} &= 2[\hat{a} \sum y^2(t-1) - \hat{b} \sum y(t-1)u(t-1) + \sum y(t)y(t-1)] \\ 0 = \frac{\partial V(\Theta)}{\partial b} &= 2[\hat{b} \sum u^2(t-1) - \hat{a} \sum y(t-1)u(t-1) - \sum y(t)u(t-1)] \end{aligned} \tag{2.4.7}$$

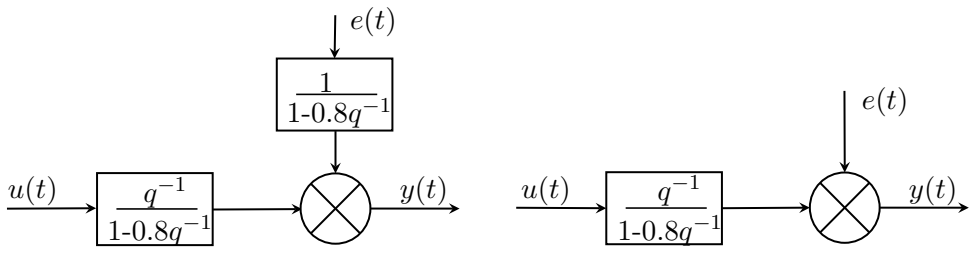
nebo v maticové formě

$$\begin{bmatrix} \sum y^2(t-1) & -\sum y(t-1)u(t-1) \\ -\sum y(t-1)u(t-1) & \sum u^2(t-1) \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} -\sum y(t)y(t-1) \\ \sum y(t)u(t-1) \end{bmatrix} \tag{2.4.8}$$

Poznamenejme, že (2.4.8) je systém lineárních rovnic se dvěma neznámými $\hat{a}$ a $\hat{b}$.

V následující části se budeme zabývat odhadem parametrů počítaným podle (2.4.8) pro různé případy. Budeme používat simulovaná data generovaná počítačem a jako důležitý doplněk provedeme i teoretickou analýzu. Při analýze budeme předpokládat velký počet dat N, stacionaritu a ergodicitu procesů, a proto může být zavedena následující aproximace

$$\frac{1}{N} \sum_{t=1}^{N} y^2(t-1) \approx E y^2(t-1) \tag{2.4.9}$$



Systém S_1 Systém S_2

Obrázek 2.1: Grafické znázornění systémů S_1 a S_2.

a obdobně i pro další součty. Může být ukázáno, že pro všechny zde uvažované případy bude levá strana (2.4.9) konvergovat k pravé straně, když N se blíží k nekonečnu. Výhoda střední hodnoty oproti součtu je v tom, že analýzu lze provádět v deterministickém rámci, přesněji problém nezávisí na konkrétní realizaci dat. Pro deterministický signál bude mít operátor E význam

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N}$$

Vraťme se nyní již k dříve uvažovaným systémům S_1 a S_2. Poznamenejme, že pro S_2 signál $x(t)$ může být chápán jako deterministický výstup bez přítomnosti šumu. To je více zřejmé z obr. 2.1, kde jsou v grafické podobě znázorněny systémy S_1 a S_2. V obrázku použitý operátor q^{-1} značí jednokrokové zpoždění, tj. $u(t-1) = q^{-1}u(t)$ a $y(t-1) = q^{-1}y(t)$.

Definice 2.4.1. Uvažujme skalární spojitou náhodnou veličinu x s hustotou pravděpodobnosti $p(x)$. Pak střední hodnota, tj. první necentrální moment, náhodné veličiny je definována jako

$$E[x] = \int_{-\infty}^{\infty} xp(x)dx$$

a variance, tj. druhý centrální moment, jako

$$\begin{aligned} var[x] &= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx \\ &= E[x^2] - \mu^2 \end{aligned}$$

Poznámka . Uvažujme dvě nezávislé skalární náhodné veličiny x a v se známými středními hodnotami a variancemi. Uvažujme dále náhodnou veličinu $z = Ax + v$, kde A je známá konstanta. Pak první dva momenty veličiny z jsou

$$\begin{aligned} E[z] &= AE[x] + E[v] \\ var[z] &= A^2 var[x] + var[v] \end{aligned}$$

Definice 2.4.2. Stacionaritou rozumíme stacionaritu v širším smyslu, která je definována následujícím způsobem [68]. Předpokládejme náhodný proces (sekvenci) $x(t) \in \mathbb{R}$. Proces je stacionární v širším smyslu pokud střední hodnota je konstantní, tj. nezávislá na čase t, a autokovarianční funkce závisí jen na rozdílu časových okamžiků a ne na konkrétním okamžiku t.

Tedy, pokud pro střední hodnotu platí

$$E[x(t)] = \mu_x, \forall t$$

a pro autokovarianční funkci definovanou

$$C_{xx}(t_1, t_2) = E\left[(x(t_1) - E[x(t_1)])(x(t_2) - E[x(t_2)])\right]$$

platí

$$C_{xx}(t_1, t_2) = C_{xx}(\tau), \forall \tau$$

kde rozdíl $\tau = t_2 - t_1$.

Definice 2.4.3. Ergodicitou rozumíme ergodicitu ve střední hodnotě, která je definována následujícím způsobem [68]. Předpokládejme náhodný proces $x(t) \in \mathbb{R}$ s konstantní střední hodnotou μ_x. Proces je ergodický ve střední hodnotě pokud

$$\lim_{T\to\infty} \frac{1}{2T+1} \sum_{t=-T}^{T} x(t) = \mu_x$$

Všimněme si konstanty $\frac{1}{2T+1}$ zajišťující nestranný odhad střední hodnoty μ_x. Odpovídající podmínky pro autokovarianční funkci $C_{xx}(\tau)$ pak jsou

$$\lim_{T\to\infty} \frac{1}{2T+1} \sum_{\tau=-2T}^{2T} \left(1 - \frac{|\tau|}{2T+1}\right) C_{xx}(\tau) = 0$$

$$\sum_{\tau=-\infty}^{\infty} |C_{xx}(\tau)| < \infty$$

Příklad 2.4.1. Simulujme systém S_1 a S_2 pro 1000 časových kroků. Vstupní signál nechť je PRBS (z anglického „pseudo random binary sequence“). Tento signál nabývá pouze dvě úrovně a to takovým způsobem, že jeho momenty prvního a druhého řádu jsou dosti podobné bílému šumu se střední hodnotou nula a variancí σ^2. Při simulaci veličiny $u(t)$ nechť je $\sigma = 1$. Pro odhad parametrů použijeme metodu nejmenších čtverců, rovnice (2.4.8). Výsledky jsou shrnuty do tabulky 2.4.1.

parametr	skutečná hodnota	odhadnutá hodnota	
		systém S_1	systém S_2
a	-0,8	-0,795	-0,580
b	1,0	0,941	0,959

Tabulka 2.4.1 Odhady parametrů pro příklad 2.4.1

Z tabulky 2.4.1 je vidět, že získaný model dává dobré výsledky pro systém S_1, zatímco systém S_2 je modelován dosti špatně. Vysvětlíme nyní tento výsledek teoretickou analýzou. Vydělíme všechny členy ve (2.4.8) počtem uvažovaných dat N a použijeme aproximaci (2.4.9) (pracujeme v oblasti stacionárních a ergodických procesů). Pak dostaneme následující rovnici pro odhady

$$\begin{bmatrix} E[y^2(t)] & -E[y(t)u(t)] \\ -E[y(t)u(t)] & E[u^2(t)] \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} -E[y(t)y(t-1)] \\ E[y(t)u(t-1)] \end{bmatrix} \tag{2.4.10}$$

Dále předpokládejme, že $u(t)$ je bílý šum s nulovou střední hodnotou a variancí σ^2. Takže PRBS je přesná aproximace bílého šumu prvním a druhým momentem. Při analýze je vhodné použít i stacionárních vlastností, to jest $E[y^2(t)] = E[y^2(t-1)]$. Pak pro systém (2.2.1) po jednoduchých výpočtech dostaneme:

$$E[y^2(t)] = \frac{b_0^2\sigma^2 + (1 + c_0^2 - 2a_0c_0)\lambda^2}{1 - a_0^2}$$

$$E[y(t)u(t)] = 0$$

$$E[u^2(t)] = \sigma^2$$

$$E[y(t)y(t-1)] = \frac{-a_0b_0^2\sigma^2 + (c_0 - a_0)(1 - a_0c_0)\lambda^2}{1 - a_0^2}$$

$$E[y(t)u(t-1)] = b_0\sigma^2$$

Aplikací těchto výsledků v (2.4.10) dostaneme následující výrazy pro parametrické odhady

$$\hat{a} = a_0 + \frac{-c_0(1 - a_0^2)\lambda^2}{b_0^2\sigma^2 + (1 + c_0^2 - 2a_0c_0)\lambda^2} \tag{2.4.11}$$

$$\hat{b} = b_0$$

Takže pro systém S_1, kde $c_0 = 0$, jsou odhadované parametry pro nekonečný počet dat

$$\hat{a} = -0,8 \quad \hat{b} = 1,0 \tag{2.4.12}$$

To znamená, že pro velké hodnoty N, tj asymptoticky, můžeme očekávat, že odhad parametrů bude blízko skutečným hodnotám parametrů a_0, b_0. Tento závěr je v souladu s dosaženými výsledky při simulacích.
Pro systém S_2 dostaneme asymptotické odhady parametrů

$$\hat{a} = \frac{-0,8\sigma^2}{\sigma^2 + 0,36\lambda^2} \approx -0,588 \quad \hat{b} = 1,0 \tag{2.4.13}$$

Pro tento případ zjišťujeme odchylku u odhadu $\hat{a}$ od skutečné hodnoty. Výsledek potvrzuje i simulace (viz tabulka 2.4.1). Teoretická analýza nám dokazuje, že výsledek získaný simulací nebyl ovlivněn ani malým počtem dat nebo nedostatkem štěstí v experimentu. Nehledě na počet dat, i když N bude blízké nekonečnu, odhad parametru a bude podle (2.4.13) obsahovat systematickou odchylku.

2.5 Strannost, konsistence a aproximace modelu

Po příkladu 2.4.1 z předchozí kapitoly je vhodné zavést další pojmy, které se vztahují ke kvalitě odhadu parametrů. Jmenovitě se budeme věnovat pojmům strannost, asymptotická strannost a konzistence.

Řekneme, že odhad $\hat{\Theta}$ je stranný, jestliže se jeho střední hodnota odchyluje od skutečné hodnoty, to jest

$$E[\hat{\Theta}] \neq \Theta_0 \tag{2.5.1}$$

Rozdíl $E[\hat{\Theta}]-\Theta_0$ je strannost. Jestliže v (2.5.1) nastává rovnost, říkáme, že $\hat{\Theta}$ je nestranný odhad.

Vysvětlení pojmu strannost. Předpokládejme zobrazení $Z : \mathbb{R}^{n_z} \to \mathbb{R}^{n_\Theta}$, které představuje estimátor neznámých parametrů $\Theta_0 \in \mathbb{R}^{n_\Theta}$. To znamená, že pro daná měření soustředěná do vektoru $z \in \mathbb{R}^{n_z}$ představuje $Z(z)$ odhad parametrů Θ_0. Označme ho $\hat{\Theta}$. Jedná se tedy o transformaci z prostoru měření do prostoru parametrů. Je rozumné tedy požadovat, aby $Z(z)$ bylo definováno pro všechna možná data z. Aby Z mohlo být chápáno jako dobrý estimátor, mělo by mít určité vlastnosti. Všechny úvahy vztažené ke kvalitě odhadu jsou založeny na chybě

$$\tilde{\Theta} = \Theta_0 - Z(z)$$

která při odhadu vzniká. Ideálně bychom si přáli, aby byla chyba nulová, popř. aby odhad $Z(z)$ byl roven skutečné hodnotě parametru s pravděpodobností jedna

$$P[Z(z) = \Theta_0] = 1$$

Obecně však, pro konečný počet dat, je to nerealizovatelný požadavek, a tudíž je třeba jej oslabit. Je vhodné požadovat, aby průměrná hodnota chyby byla nulová

$$E[Z(z) - \Theta_0] = 0$$

nebo ekvivalentně, aby očekávaná hodnota odhadu se rovnala očekávané hodnotě parametru

$$E[Z(z)] = E[\Theta_0] = \Theta_0$$

Estimátorům, které vykazují tuto vlastnost říkáme nestranné.

Všimněme si, že strannost estimátoru není obecně funkcí počtu dostupných dat. V některých případech, jakými je například v předchozí části uvažovaný problém identifikace parametrů dynamického systému, však může nastat situace, kdy odhad parametrů pro konečný počet dat je stranný, zatímco pro nekonečný počet dat je nestranný. Tato vlastnost je obvykle označována jako asymptotická nestrannost.

Ilustrujme pojem asymptotická strannost odhadu za pomoci příkladu 2.4.1. a systému S_1. Z rovnice (2.4.8) plyne následující maticový vztah pro odhad neznámých parametrů ve smyslu nejmenších čtverců

$$\hat{\Theta} = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = A^{-1} \begin{bmatrix} -\sum y(t)y(t-1) \\ \sum y(t)u(t-1) \end{bmatrix}$$

kde $A = \begin{bmatrix} \sum y^2(t-1) & -\sum y(t-1)u(t-1) \\ -\sum y(t-1)u(t-1) & \sum u^2(t-1) \end{bmatrix}$, který může být, dosazením za $y(t)$ z popisu systému

(2.2.2), dále upraven

$$\begin{aligned}\hat{\Theta} &= A^{-1}\begin{bmatrix} -\sum(-a_0y(t-1)+b_0u(t-1)+e(t))y(t-1) \\ \sum(-a_0y(t-1)+b_0u(t-1)+e(t))u(t-1) \end{bmatrix} \\ &= A^{-1}\underbrace{\begin{bmatrix} \sum y(t-1)^2 & -\sum u(t-1)y(t-1) \\ \sum -y(t-1)u(t-1) & u(t-1)^2 \end{bmatrix}}_{A}\begin{bmatrix} a_0 \\ b_0 \end{bmatrix} + A^{-1}\begin{bmatrix} -\sum e(t)y(t-1) \\ \sum e(t)u(t-1) \end{bmatrix} \\ &= \Theta_0 + A^{-1}\begin{bmatrix} -\sum e(t)y(t-1) \\ \sum e(t)u(t-1) \end{bmatrix}\end{aligned}$$

K vyhodnocení strannosti či nestrannosti odhadu $\hat{\Theta}$ je nutné najít střední hodnotu odhadu, tj. vypočítat

$$E[\hat{\Theta}] = \Theta_0 + E\left[\begin{bmatrix} \sum y^2(t-1) & -\sum y(t-1)u(t-1) \\ -\sum y(t-1)u(t-1) & \sum u^2(t-1) \end{bmatrix}^{-1}\begin{bmatrix} -\sum e(t)y(t-1) \\ \sum e(t)u(t-1) \end{bmatrix}\right]$$

Z předchozího vztahu je patrné, že odhad $\hat{\Theta}$ není nestranný, protože střední hodnota členu na pravé straně nebude nulová ($y(t)$ závisí na všech předchozích vstupech a výstupech). Avšak, pokud připustíme nekonečné množství dat, tj. $N \to \infty$, lze psát

$$E[\hat{\Theta}] = \Theta_0 + E\left[\begin{bmatrix} E[y^2(t-1)] & -E[y(t-1)u(t-1)] \\ -E[y(t-1)u(t-1)] & E[u^2(t-1)] \end{bmatrix}^{-1}\begin{bmatrix} -E[e(t)y(t-1)] \\ E[e(t)u(t-1)] \end{bmatrix}\right]$$

a dále, díky předpokladu bělosti šumu $e(t)$,

$$E[\hat{\Theta}] = \Theta_0, \quad \text{když} \quad N \to \infty$$

Tudíž získaný odhad pro systém S_1 je asymptoticky nestranný. Poznamenejme, že v literatuře se můžeme setkat i s alternativním označením asymptotické nestrannosti říkající, že odhad $\hat{\Theta}$ *konverguje ve střední hodnotě* ke skutečnému vektoru parametrů Θ_0. Není těžké ověřit, že odhad parametrů systému S_2 již nelze považovat za asymptoticky nestranný.

Asymptotická nestrannost má v jistém smyslu blízko k pojmu konzistence. Říkáme, že odhad $\hat{\Theta}$ je konsistentní, jestliže

$$\hat{\Theta} \to \Theta_0 \quad \text{když} \quad N \to \infty \tag{2.5.2}$$

Protože $\hat{\Theta}$ je stochastická proměnná, musíme definovat v jakém smyslu budeme brát limitu v (2.5.2). Jednou z možností je „limita s pravděpodobností 1“, která je definována $\forall \epsilon > 0$ jako

$$\lim_{N\to\infty} P[\mid \hat{\Theta} - \Theta \mid < \epsilon] = 1$$

Tuto definici konzistence odhadu budeme obyčejně používat i v následujících kapitolách.

Provedená analýza v příkladu 2.4.1 ukazuje, že $\hat{\Theta}$ je konsistentní pro systém S_1, ale není konsistentní pro systém S_2.

Nyní si všimneme pojmu identifikovatelnost systému. Zhruba řečeno, říkáme, že systém je identifikovatelný, jestliže odhady parametrů jsou konsistentní. Poznamenejme, že identifikovatelnost daného systému S závisí na struktuře modelu M, identifikační metodě I

a experimentálních podmínkách X.

V následujícím příkladu ukážeme, jak experimentální podmínky mohou ovlivnit výsledek identifikace.

Příklad 2.5.1 Nechť systémy S_1 a S_2 jsou simulovány a je použito 1000 měření. Vstupem je jednotkový skok. Po vypočtení odhadu ve smyslu nejmenších čtverců dostaneme výsledky, které jsou zachyceny v tabulce 2.5.1.

parametr	skutečná hodnota	odhadnutá hodnota	
		systém S_1	systém S_2
a	-0,8	-0,788	-0,058
b	1,0	1,059	4,693

Tabulka 2.5.1 Odhady parametrů pro příklad 2.5.1

Vidíme, že dostáváme dobrý model pro systém S_1. Pro systém S_2 dostáváme značnou odchylku od skutečných parametrů. Odhad je také dosti odlišný od výsledků, které jsme dostali v příkladu 2.4.1. Např. zde je také značná odchylka v odhadu $\hat{b}$.

Teoretická analýza těchto pozorování vyžaduje řešit rovnici (2.4.10). Nejprve vypočteme jednotlivé kovariance. Nechť $u(t)$ je jednotkový skok velikosti σ. Označme statický zisk systému S $(S = b_0/(1 + a_0))$. Pak dostaneme

$$E[y^2(t)] = S^2\sigma^2 + \frac{(1 + c_0^2 - 2a_0c_0)\lambda^2}{1 - a_0^2}$$

$$E[y(t)u(t)] = S\sigma^2$$

$$E[u^2(t)] = \sigma^2$$

$$E[y(t)y(t-1)] = S^2\sigma^2 + \frac{(c_0 - a_0)(1 - a_0c_0)\lambda^2}{1 - a_0^2}$$

$$E[y(t)u(t-1)] = S\sigma^2$$

Dosazením těchto výsledků do (2.4.10) odvodíme následující výrazy pro odhady parametrů

$$\hat{a} = a_0 - \frac{c_0(1 - a_0^2)}{1 + c_0^2 - 2a_0c_0}$$

$$\hat{b} = b_0 - b_0c_0\frac{1 - a_0}{1 + c_0^2 - 2a_0c_0} \tag{2.5.3}$$

Povšimněme si, že nyní se oba odhady parametrů obecně odlišují od skutečných hodnot parametrů. Navíc odchylka je nezávislá na velikosti vstupního skoku σ. Odchylka je nulová, jestliže $c_0 = 0$. Pro uvažovaný systém S_1 dostaneme

$$\hat{a} = -0,8 \quad \hat{b} = 1,0 \tag{2.5.4}$$

(jako v příkladu 2.4.1), zatímco pro systém S_2 vypočteme následující výsledek

$$\hat{a} = 0,0 \quad \hat{b} = \frac{b_0}{1 + a_0} = 5,0 \tag{2.5.5}$$

Je zřejmé, že je velice odlišný od skutečných hodnot. Nicméně všimněme si, že statický zisk je odhadnut správně, protože

$$\frac{\hat{b}}{1 + \hat{a}} = \frac{b_0}{1 + a_0}$$

Teoretické výsledky (2.5.4) a (2.5.5) jsou velmi blízké simulačním výsledkům, které jsou uvedeny v tabulce 2.5.1. Měli bychom poznamenat, že v případě absolutní nepřítomnosti šumu (tj. když $\lambda^2 = 0$), nastanou problémy, protože v (2.4.10) vznikne singulární matice

$$\sigma^2 \begin{bmatrix} S^2 & -S \\ -S & 1 \end{bmatrix}$$

Pak řešení (2.4.10) může být charakterizováno

$$\frac{\hat{b}}{1 + \hat{a}} = S$$

Na uvedených příkladech bylo vidět, že odhady jsou pro S_1 konzistentní, zatímco odhady pro systém S_2 vykazují systematickou chybu. Získané modely pro S_2 mohou být chápány jako aproximace skutečného systému. Aproximace je zřejmě závislá na užitých experimentálních podmínkách. V následujícím příkladu ukážeme detailní výpočty pro různé experimentální podmínky.

Příklad 2.5.2 Použijme model (2.4.1) jako základ pro výpočet predikce. Nejrozumnější predikce hodnoty $y(t)$ na základě dat až do času $t - 1$ je, bez znalosti rozdělení $\epsilon(t)$, dána vztahem

$$\hat{y}(t) = -ay(t-1) + bu(t-1) \tag{2.5.6}$$

Chyba predikce bude podle (2.2.1) splňovat

$$\tilde{y}(t) = y(t) - \hat{y}(t) = (a - a_0)y(t-1) + (b_0 - b)u(t-1) + e(t) + c_0 e(t-1) \tag{2.5.7}$$

Vypočtěme varianci chyby predikce $W = E[\tilde{y}^2(t)]$ pro několik případů. Pro systém S_2 stále předpokládáme $c_0 = a_0$.

Nejprve vypočtěme varianci chyby predikce pro situaci, kdy jsou pro výpočet predikce (2.5.6) použity skutečné hodnoty parametrů, to jest $a = a_0, b = b_0$. Pak chyba predikce je

$$\tilde{y}(t) = e(t) + c_0 e(t-1) = e(t) + a_0 e(t-1)$$

a variance chyby predikce

$$W_1 = \lambda^2(1 + a_0^2) \tag{2.5.8}$$

bude nezávislá na experimentálních podmínkách. Dále spočítejme varianci chyby predikce pro situaci, kdy predikce (2.5.6) je založena na odhadech parametrů získaných při buzení systému jednotkovým skokem o velikosti σ. Využitím odhadů (2.5.5) dostaneme v ustáleném stavu pro systém S_2

$$\begin{aligned}
\tilde{y}(t) &= -a_0 y(t-1) + (b_0 - \frac{b_0}{1+a_0})u(t-1) + e(t) + a_0 e(t-1) \\
&= -a_0[\frac{b_0}{1+a_0}\sigma + e(t-1)] + \frac{a_0 b_0}{1+a_0}\sigma + e(t) + a_0 e(t-1) \\
&= e(t)
\end{aligned}$$

a tak

$$W_2 = \lambda^2 < W_1 \tag{2.5.9}$$

Povšimněme si, že jsme dostali lepší výsledek (nižší varianci chyby predikce) než v případě, když jsme použili skutečných hodnot a_0 a b_0. Můžeme tedy říci, že identifikační metoda používá a a b jako *prostředku* k získání dobré predikce. Poznamenejme, že v předchozích výpočtech je stěžejní, abychom při identifikaci použili stejné experimentální podmínky ($u(t)$ skok o velikosti σ) jako při výpočtu predikce. Dokumentujme toto tvrzení. Předpokládejme, že odhady parametrů jsou určeny z experimentu, kde $u(t)$ je bílý šum s variancí $\bar{\sigma}^2$ a s nulovou střední hodnotou. Pak odhady parametrů jsou dány (2.4.11), (2.4.13). Použitím těchto výrazů pro $c_0 = a_0$ dostaneme odhad parametrů. Nyní předpokládejme, že odhadnutý model použijeme pro predikci, ale za vstup budeme považovat skok o velikosti σ. Pak

$$\begin{aligned}
\tilde{y}(t) &= (a - a_0)y(t-1) + e(t) + a_0 e(t-1) \\
&= (a - a_0)[\frac{b_0}{(1+a_0)}\sigma + e(t-1)] + e(t) + a_0 e(t-1) \\
&= (a - a_0)\frac{b_0}{(1+a_0)}\sigma + e(t) + ae(t-1)
\end{aligned}$$

Nechť z označuje $b_0^2\bar{\sigma}^2/[(1-a_0^2)\lambda^2]$. Střední hodnota $\tilde{y}^2(t)$ pak bude

$$\begin{aligned}
W_3 &= \lambda^2(1 + a^2) + (a - a_0)^2\frac{b_0^2}{(1+a_0)^2}\sigma^2 \\
&= \lambda^2[1 + (\frac{a_0 z}{z+1})^2] + (\frac{a_0}{z+1})^2(\frac{b_0}{1+a_0})^2\sigma^2
\end{aligned}$$

Je zřejmé, že vždy bude $W_3 > W_2$.

V následující části omezíme náš rozbor pouze na systém S_1. Budeme především analyzovat chování matice vznikající ve (2.4.8). Předpokládejme, že tato matice je "dobře podmíněná".

Pak existuje jediné řešení (2.4.8). Toto řešení je pro systém S_1 dáno asymptoticky ($N \to \infty$). Využitím skutečných parametrů a_0, b_0 a dosazením do pravé strany v (2.4.8) za $y(t)$ dostaneme

$$\begin{aligned} & \frac{1}{N}\begin{bmatrix} \sum y^2(t-1) & -\sum y(t-1)u(t-1) \\ -\sum y(t-1)u(t-1) & \sum u^2(t-1) \end{bmatrix}\begin{bmatrix} a_0 \\ b_0 \end{bmatrix} - \frac{1}{N}\begin{bmatrix} -\sum y(t)y(t-1) \\ \sum y(t)u(t-1) \end{bmatrix} \\ = & \frac{1}{N}\begin{bmatrix} \sum y(t-1)e(t) \\ -\sum u(t-1)e(t) \end{bmatrix} \to E\begin{bmatrix} y(t-1)e(t) \\ -u(t-1)e(t) \end{bmatrix} = 0 \end{aligned} \tag{2.5.10}$$

Poslední rovnost je splněna, protože $\{e(t)\}$ je bílý šum, a tudíž je nezávislý na všech minulých datech.

Bohužel, nelze ve všech případech zaručit „dobrou podmíněnost“ matice v rovnici (2.4.8), popř. v (2.5.10), a tím i možnost její inverze. Invertovatelnost matice je do značné míry dána aktuálními experimentálními podmínkami, které jsou ovlivněny jak vstupním signálem $u(t)$, tak i tím, zda systém obsahuje zpětnou vazbou či nikoliv. Proto, v následujících subkapitolách se budeme v příkladech věnovat situacím, kdy čtvercová matice vyskytující se v (2.5.10) není, z důvodu experimentálních podmínek, dobře podmíněná.

2.6 Trvale vybuzený systém

Příklad 2.6.1. Simulujme systém S_1 a uvažujme 1000 simulačních kroků. Vstup bude jednotkový impuls v čase $t = 1$. Vypočtěme odhady parametrů ve smyslu nejmenších čtverců. Numerické výsledky jsou ukázány v tabulce 2.6.1.

Parametr	Skutečná hodnota	Odhadnutá hodnota
a	-0,8	-0,796
b	1,0	2,950

Tabulka 2.6.1 Odhady parametrů pro příklad 2.6.1

Z tabulky je zřejmé, že odhad parametru a je velmi dobrý, zatímco odhad parametru b je špatný. To je přirozené, protože vstup málo ovlivňuje výstup. Informaci o b_0 můžeme dostat pouze přes $y(t)$, které je ovlivněno vstupem. Na druhé straně, parametr a_0 bude také popisovat účinek šumu na výstup. Protože šum je obsažen ve všech datech je velmi přirozené, že a_0 je odhadnuto mnohem přesněji než b_0.

Proveďme teoretickou analýzu. Uvažujme (2.4.8), kde $u(t)$ je impuls velikosti σ v čase $t = 1$. Označme

$$R_0 = \frac{1}{N}\sum y^2(t-1) \qquad R_1 = \frac{1}{N}\sum y(t)y(t-1)$$

Odhad parametrů lze snadno získat z (2.4.8)

$$\begin{aligned}\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix} &= \begin{bmatrix} NR_0 & -y(1)\sigma \\ -y(1)\sigma & \sigma^2 \end{bmatrix}^{-1} \begin{bmatrix} -NR_1 \\ y(2)\sigma \end{bmatrix} \\ &= \frac{1}{R_0 - y^2(1)/N} \begin{bmatrix} -R_1 + y(1)y(2)/N \\ (-y(1)R_1 + y(2)R_0)/\sigma \end{bmatrix} \end{aligned} \tag{2.6.1}$$

Pro velké N pak zjistíme, že

$$R_0 \to \frac{\lambda^2}{1-a_0^2}, \quad R_1 \to \frac{-a_0\lambda^2}{1-a_0^2} = -a_0 R_0$$

Pro limitní případ (tedy $N \to \infty$) dostaneme dosazením za R_0 a R_1 do (2.6.1)

$$\begin{aligned}\hat{a} &= a_0 \\ \hat{b} &= (a_0 y(1) + y(2))/\sigma = b_0 + e(2)/\sigma \end{aligned} \tag{2.6.2}$$

Je zřejmé, že $\hat{b}$ obsahuje člen, který způsobuje odchylku odhadu od skutečné hodnoty. Tato odchylka závisí na konkrétní realizaci náhodného procesu $\{e(t)\}$ a velikosti impulsu σ. V uvažované simulaci bylo $e(2) = 1,957$ a $\sigma = 1$, což podle (2.6.2) by mělo dát $\hat{b} = 2,957$. Tato hodnota je v souladu s výsledky uvedenými v tabulce 2.6.1.
Výše pozorované chování odhadů ve smyslu nejmenších čtverců lze vysvětlit. V tomto příkladě analyzovaná situace je zvláštní ve dvou aspektech. První je, že matice v (2.6.1) vynásobená $1/N$ se blíží k singulární matici, když $N \to \infty$. Přesto ale odhad ve smyslu nejmenších čtverců existuje a může být vypočítán pro každé N. Důležitější je však druhý aspekt. Neplatí totiž vztah (2.5.10), protože součty obsahující vstupní signál nesměřují k očekávaným hodnotám.
Pro systém S_1, jak bylo vidět v příkladech 2.4.1, 2.5.1 a 2.6.1 jsme dostali konsistentní odhady parametrů za předpokladu, že vstup je bílý šum nebo skoková funkce (ve druhém případě navíc musíme předpokládat existenci šumu působícího na systém, tedy $\lambda^2 > 0$). Jestliže $u(t)$ je impuls, metodou nejmenších čtverců nezískáme konsistentní odhady. Zhruba řečeno, důvodem je skutečnost, že impulsní funkce je „příliš často“ rovna nule. Abychom garantovali konsistenci, potřebujeme použít vstup, který dostatečně ovlivňuje proces. Na tomto základě definujme pojem trvalé buzení.

Definice 2.6.1. Říkáme, že signál $u(t)$ je trvale budicí - p.e. (z anglického termínu „persistently exciting“) řádu n, jestliže

i) existuje limita

$$r_u(\tau) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} [u(t+\tau)][u^T(t)] \tag{2.6.3}$$

ii) a následující matice je pozitivně definitní

$$R_u(n) = \begin{bmatrix} r_u(0) & r_u(1) & \dots & r_u(n-1) \\ r_u(-1) & r_u(0) & & \\ \vdots & \ddots & & \\ r_u(1-n) & & & r_u(0) \end{bmatrix} \tag{2.6.4}$$

Poznámka 1. Většina stacionárních stochastických procesů je ergodická. To znamená, že v (2.6.3) můžeme $\lim_{N\to\infty}$ nahradit operátorem střední hodnoty E. Pak matice $R_u(n)$ je obyčejná kovarianční matice signálu $u(t)$ (s prvky dané autokovarianční funkcí signálu).
Ilustrujme zavedený pojem na vstupy použité v této kapitole.

Příklad 2.6.2. Nechť $u(t)$ je bílý šum s nulovou střední hodnotou a variancí σ^2. Pak dostaneme $r_u(\tau) = \sigma^2$ pro $\tau = 0$ a 0 pro $\tau \neq 0$, a matice $R_u(n) = \sigma^2 I_n$ je tedy pozitivně definitní pro libovolné n. Tudíž bílý šum je trvale budicí signál všech řádů.

Dále uvažujme $u(t)$ jako skok o velikosti σ. Pak dostaneme $r_u(\tau) = \sigma^2$ pro všechna τ a pak bude $R_u(1)$ pozitivně definitní, zatímco $R_u(n)$ pro $n = 2, 3, ...$ bude singulární. Tudíž signál typu skoková funkce je p.e. řádu jedna.

Konečně $u(t)$ nechť je impuls. To dává $r_u(\tau) = 0, R_u(n) = 0$. Tento signál není p.e. žádného řádu. To vysvětluje, proč jsme nedostali konsistentní odhad parametrů, když vstup byl impuls.

Poznámka 2. V pracích zabývajících se adaptivním řízením se používají alternativní definice trvale budicího signálu.

Poznámka 3. Pojem trvale budicí signál, tak jak byl zaveden, je motivován úlohou určení odhadu koeficientů „useknuté“ váhové funkce [20]. S touto úlohou se setkáme ve 3. kapitole při výkladu neparametrických metod, konkrétně u korelační analýzy. Nutná podmínka pro konsistentní odhad lineárního systému n-tého řádu je, aby vstupní signál byl p.e. řádu $2n$. V některých případech při použití metody nejmenších čtverců stačí, aby signál byl p.e. řádu n.

Poznámka 4. Tvrzení v předchozí poznámce jsou použitelná při hledání konsistentních odhadů systémů se šumem. Pro systémy bez šumu není nutné, aby vstup byl p.e. Uvažujme např. deterministický lineární systém n-tého řádu s nulovými počátečními podmínkami. Jako vstup použijme impuls a zaznamenáme impulsní odezvu. Z $2n$ nenulových hodnot impulsní odezvy je možné najít parametry systému, i když vstup není p.e. Důvodem je, že systém bez šumu může být identifikován z konečného počtu dat ($N < \infty$), zatímco trvalé vybuzení se týká vlastností vstupního signálu při použití nekonečného počtu dat ($N \to \infty$), které je uvažováno při analýze konsistence odhadu parametrů v systémech se šumem).

Poznámka 5. Původní práce týkající se analýzy trvale budicích signálů i některé současné analýzy jsou prováděny ve frekvenční oblasti. Protože je však celá práce zaměřena na popis signálů v časové oblasti, nebudeme se analýzami ve frekvenční oblasti rozsáhleji zabývat.

2.7 Vliv zpětné vazby

Viděli jsme, že musí být zavedeno určité omezení na vstupní signál, abychom garantovali, že matice vyskytující se v (2.4.8) je dobře podmíněná. V této kapitole budeme sledovat situaci, kdy vstup je determinován výstupní zpětnou vazbou. Při provádění reálného identifikačního experimentu se mnohdy nelze obejít bez takovéto zpětné vazby. Identifikovaný systém může být v otevřené smyčce nestabilní, takže bez stabilizující zpětné vazby může být nemožné získat nějakou, třeba i nevelkou, množinu dat. Také bezpečnost nebo požadavek na normální pracovní režim může být dostatečný důvod pro použití zpětné vazby během identifikačního experimentu.

Rovněž není jisté, že přímá vazba je lepší pro identifikaci než využití zpětné vazby.

Příklad 2.7.1. Uvažujme systém S_1 (2.2.2). Předpokládejme, že vstup je určen proporcionální zpětnou vazbou

$$u(t) = -ky(t) \tag{2.7.1}$$

Pak matice v (2.4.8)

$$\sum y^2(t-1) \begin{bmatrix} 1 & k \\ k & k^2 \end{bmatrix}$$

je zřejmě singulární. Zároveň je zřejmé, že systém (2.2.2) se vstupem generovaným regulátorem (2.7.1) nelze identifikovat. Je vidět, že pouze $\{y(t)\}$ přináší informaci o dynamice systému S_1 v tom smyslu, že $\{u(t)\}$ nám nemůže přinést nic nového. Kombinací modelu (2.4.1) a regulátoru (2.7.1) dostaneme

$$\epsilon(t) = y(t) + (a + bk)y(t-1) \tag{2.7.2}$$

Tento výraz ukazuje, že z dat může být pouze odhadnuta lineární kombinace $a + bk$. Všechny hodnoty a a b, které dávají stejnou hodnotu $a + bk$ budou dávat stejná residua $\{\epsilon(t)\}$ a stejné hodnoty ztrátové funkce. Neexistuje tedy jediné minimum ztrátové funkce, ale ta je minimalizována množinou bodů. Pro asymptotický případ ($N \to \infty$) je tato množina definována

$$\{\Theta \mid a + bk = a_0 + b_0 k\}$$

Protože neexistuje jediné minimum, matice druhých derivací kritéria $V(\Theta)$ musí být singulární. Vraťme se proto k matici, která se objevuje v (2.4.8). Ta nás vede zpět k počátečnímu zjištění, že parametry a a b nemůžeme identifikovat vstupem (2.7.1). Příklad 2.7.1 nám ukazuje, že použití zpětné vazby (2.7.1) v průběhu identifikačního experimentu znemožňuje zajistit konsistentní odhad. Naštěstí situace není tak zlá. O tom nás přesvědčí následující příklady.

Příklad 2.7.2. Simulujme systém S_1 a proveďme 1000 iterací. Vstup nechť je definován jako zpětná vazba, na kterou aditivně působí časově variantní referenční signál

$$u(t) = -ky(t) + r(t) \tag{2.7.3}$$

Referenční signál $r(t)$ uvažujme jako PRBS velikosti 0,5 a zpětnovazební zisk vybereme $k = 0,5$. Vypočtěme odhady ve smyslu nejmenších čtverců dle vztahu (2.4.8). Tabulka 2.7.1 shrnuje dosažené výsledky.

parametr	skutečná hodnota	odhadnutá hodnota
a	-0,8	-0,754
b	1,0	0,885

Tabulka 2.7.1 Odhady parametrů pro příklad 2.7.2.

Jak je zřejmé z tabulky v tomto případě jsme získali rozumné odhady, ačkoli data byla generována za přítomnosti zpětné vazby. Je vhodné také zdůraznit, že pro vlastní identifikační proces, tak jak je doposud uvažován, není zapotřebí znalost zisku k ani referenčního signálu

$r(t)$. Odhad parametrů je počítán pouze na základě měřených vstupních dat $u(t)$ a výstupních dat $y(t)$.

Analyzujme tuto situaci. Nejdříve si všimněme, že (2.5.10) stále platí. Je tudíž postačující ukázat, že matice v (2.4.8) je regulární a dobře podmíněná pro velké N. Za tím účelem předpokládejme, že $r(t)$ je bílý šum s nulovou střední hodnotou a variancí σ^2. Pak dostaneme pro (2.2.1), kde $c_0 = 0$ a (2.7.3)

$$y(t) + (a + bk)y(t-1) = br(t-1) + e(t) \tag{2.7.4}$$
$$u(t) + (a + bk)u(t-1) = r(t) + ar(t-1) - ke(t) \tag{2.7.5}$$

Rovnici (2.7.4) lze interpretovat jako popis uzavřeného systému (tj. systém S_1 (2.2.2) + regulátor (2.7.1)). Po určitých výpočtech dostaneme matici

$$\begin{bmatrix} E[y^2(t)] & -E[y(t)u(t)] \\ -E[y(t)u(t)] & E[u^2(t)] \end{bmatrix} =$$
$$= \frac{1}{1-(a+bk)^2} \begin{bmatrix} b^2\sigma^2 + \lambda^2 & -k(b^2\sigma^2 + \lambda^2) \\ -k(b^2\sigma^2 + \lambda^2) & k^2(b^2\sigma^2 + \lambda^2) + [1-(a+bk)^2]\sigma^2 \end{bmatrix}$$

která je pozitivně definitní. Poznamenejme jen, že jsme předpokládali, že uzavřený systém je asymptoticky stabilní ($\mid a + bk \mid$) < 1.

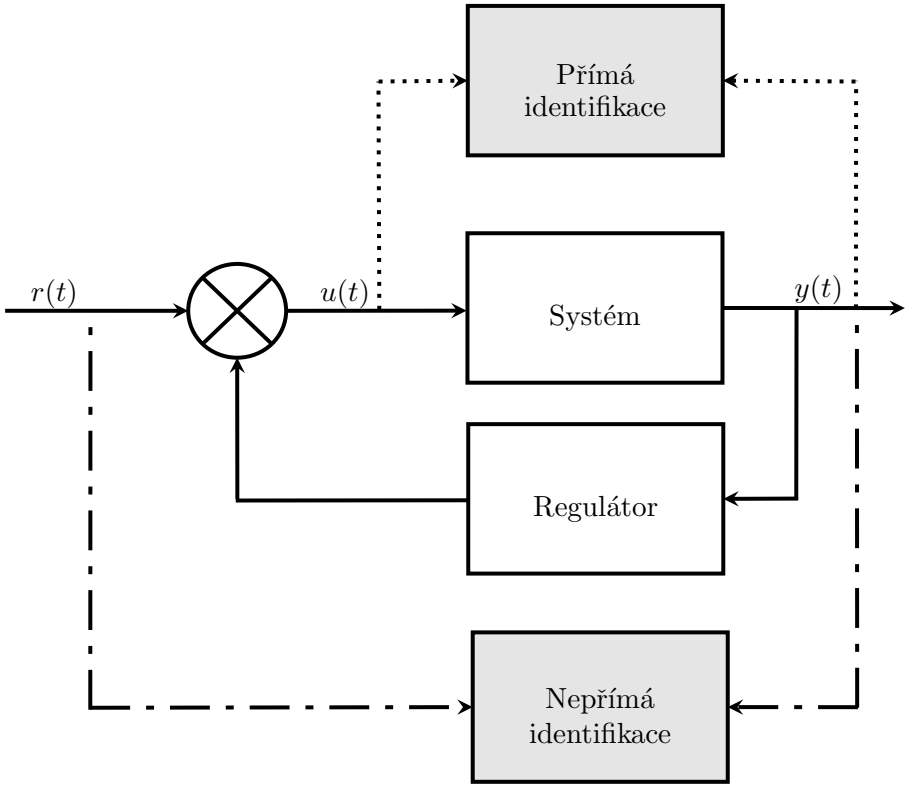
Výše představený postup identifikace parametrů systému je v literatuře označován jako přímá identifikace, kdy existence zpětné vazby je identifikační metodou v podstatě ignorována, tj. pro určení parametrů systému nepotřebujeme znát zisk regulátoru k ani referenční signál $r(t)$. Předpokládáme však, že referenční signál je přítomen a je dostatečně bohatý. Naproti tomu existuje i přístup označovaný jako nepřímá identifikace, kde je nejprve identifikován model uzavřeného systému a pak na jeho základě a známého popisu regulátoru je určen model vlastního systému. Nepřímá identifikace tak předpokládá známý popis regulátoru, avšak umožňuje najít parametry modelu systému i v situaci, kdy není referenční signál dostatečně bohatý (tj. není p.e. dostatečného řádu). Principiální rozdíl mezi přímou a nepřímou identifikací je znázorněn na obrázku 2.2.
Dva přístupy k nepřímé identifikaci systému S_1 jsou ilustrovány v následujících příkladech.

Příklad 2.7.3. Předpokládejme známý zisk k a referenční signál $r(t)$. Pak lze, analogicky k (2.4.8), sestavit následující soustavu dvou rovnic na základě (2.7.4) zapsanou v maticové formě

$$\begin{bmatrix} \sum y^2(t-1) & -\sum y(t-1)r(t-1) \\ -\sum y(t-1)r(t-1) & \sum r^2(t-1) \end{bmatrix} \begin{bmatrix} \hat{\tilde{a}} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} -\sum y(t)y(t-1) \\ \sum y(t)r(t-1) \end{bmatrix} \tag{2.7.6}$$

kde $\tilde{a} = a + bk$. Řešení soustavy rovnic (2.7.6) umožní nalézt odhady parametrů popisu uzavřeného systému $\hat{\tilde{a}}$ a $\hat{b}$ na jejichž základě lze snadno vypočítat odhad zbývajícího neznámého parametru a systému S_1 dle vztahu $\hat{a} = \hat{\tilde{a}} - \hat{b}k$.

Ačkoliv představený přístup k nepřímé identifikaci je relativně přímočarý, je závislý na předpokladu dostatečně bohatého referenčního signálu $r(t)$, který dostatečně vybudí systém (podobně jako přímá identifikace v příkladu 2.7.2). Tento předpoklad může být v některých situacích limitující (ne vždy je možné zvolit libovolný referenční signál $r(t)$ s ohledem na chování výstupu systému $y(t)$).



Obrázek 2.2: Ilustrace přímé a nepřímé identifikace systému se zpětnou vazbou.

Pokud tedy není možné použít referenční signál, který je trvale budicí dostatečného řádu, můžeme použít přístup založený na specifikaci regulátoru vyššího řádu. Tento přístup je ilustrován následujícím příkladem.

Příklad 2.7.4. Předpokládejme známý regulátor prvního řádu popsaný následujícím vztahem

$$u(t) = -k_1 y(t) - k_2 y(t-1)$$

Pak popis uzavřeného systému je dán následující rovnicí

$$y(t) + (a + bk_1)y(t-1) + bk_2 y(t-2) = e(t) \tag{2.7.7}$$

a soustava rovnic umožňující odhad neznámých parametrů uzavřeného systému $\tilde{a} = a + bk_1$ a $\tilde{b} = bk_2$ je

$$\begin{bmatrix} \sum y^2(t-1) & \sum y(t-1)y(t-2) \\ \sum y(t-1)r(t-1) & \sum y^2(t-2) \end{bmatrix} \begin{bmatrix} \hat{\tilde{a}} \\ \hat{\tilde{b}} \end{bmatrix} = \begin{bmatrix} -\sum y(t)y(t-1) \\ -\sum y(t)y(t-2) \end{bmatrix} \tag{2.7.8}$$

Jakmile je proveden odhad parametrů uzavřeného systému $\tilde{a}$ a $\tilde{b}$, můžeme, vzhledem ke známému popisu regulátoru (2.7.7), spočítat odhady parametrů původního systému S_1 dle vztahů

$$\hat{b} = \hat{\tilde{b}}/k_2$$
$$\hat{a} = \hat{\tilde{a}} - \hat{b}k_1$$

a to bez ohledu na přítomnost či vlastnosti případného referenčního signálu.

Poznamenejme, že představené metody pro identifikaci systému se zpětnou vazbou jsou založeny na předpokladu stabilního zpětnovazebního systému.
Závěrem tak dodejme, že konzistentní odhady parametrů systému ovlivněného zpětnou vazbou lze získat buďto za předpokladu dostatečně bohatého referenčního signálu z „externího“ zdroje, či použít regulátor stejného či vyššího řádu než je řád systému. Jinou možností, která však nebyla zde diskutována, je použití regulátoru s časově proměnnými parametry.

2.8 Shrnutí a zhodnocení výsledků

Shrňme zkušenosti získané z příkladů a zabývejme se závěry, které z nich mohou být vyvozeny.

1) Pro zajištění konsistence při použití metody nejmenších čtverců je rozhodující, jakým způsobem vstupuje šum do systému. Tento fakt ukazuje na význam a požadavky na strukturu modelu M.

2) Pokud jde o experimentální podmínky X bylo vidět, že je důležité, aby vstupní signál byl trvale budicí. To zhruba řečeno znamená, že všechny módy systému by měly být vybuzeny a dostatečně se měnit v průběhu identifikačního experimentu.

3) Jestliže experimentální podmínky zahrnují zpětnou vazbu od $y(t)$ na $u(t)$, je v některých případech nemožné identifikovat parametry systému. Na druhé straně, jestliže přidáme časově proměnný referenční signál působící na systém, parametry lze snadno nalézt.

Jistě není třeba zdůrazňovat, že předchozí tvrzení nebyla striktně dokázána. Vycházejí z jednoduchých příkladů. Nicméně, lze ukázat, že tyto závěry platí i pro mnohem obecnější podmínky.

Více o zavedených pojmech a experimentálních podmínkách lze nalézt v [24],[20],[37].

Kapitola 3

Neparametrické metody

3.1 Úvod

Identifikační metody jsou často rozdělovány na neparametrické a parametrické. Zatímco parametrické metody poskytují modely ve formě rovnic, neparametrické identifikační metody jsou charakteristické tím, že výsledné modely jsou křivky nebo funkce. Neparametrické metody jsou mnohdy označovány jako klasické metody identifikace, jako zdůraznění toho, že se obecně jedná o straší přístup než jsou metody parametrické. Neparametrické metody jsou obecně koncepčně i výpočetně jednodušší, avšak v porovnaní s parametrickými metodami nenabízejí tak široké možnosti. Proto se budeme neparametrickým metodám věnovat pouze okrajově a to jen v této kapitole. Zbylé kapitoly skript budou věnovány metodám parametrickým. Nakonec poznamenejme, že členění metod na parametrické a neparametrické je dáno spíše tradičním názvoslovím než skutečným významovým obsahem těchto pojmů. Neparametrické metody lze totiž v jistém smyslu chápat též jako parametrické.

Po tomto stručném úvodu můžeme přejít k výkladu vybraných neparametrických metod. Postupně se budeme věnovat metodám založeným na:

- frekvenční analýze,
- přechodové analýze,
- korelační analýze,
- spektrální analýze.

3.2 Frekvenční analýza

Nejdříve se zabývejme frekvenční analýzou. V tomto případě je vhodné používat modely spojité v čase a vyjít z modelu

$$Y(p) = G(p)U(p) \tag{3.2.1}$$

kde $Y(p)$ je Laplaceova transformace výstupního signálu $y(t)$,
$U(p)$ je Laplaceova transformace vstupního signálu $u(t)$,
$G(p)$ je přenos systému.

Jestliže vstupem systému je signál typu sinus

$$u(t) = a \cdot \sin(\omega t) \tag{3.2.2}$$

a systém je asymptoticky stabilní, pak výstup v ustáleném stavu bude

$$y(t) = b \cdot \sin(\omega t + \varphi) \tag{3.2.3}$$

kde

$$\begin{aligned} b &= a \mid G(i\omega) \mid \\ \varphi &= \arg[G(i\omega)] \end{aligned} \tag{3.2.4}$$

o čemž se můžeme snadno přesvědčit následujícím způsobem. Nechť systém je reprezentován váhovou funkcí $h(t)$

$$y(t) = \int_0^t h(\tau)u(t-\tau)d\tau$$

kde $h(\tau)$ je funkce jejíž Laplaceova transformace je $G(p)$. Zaveďme

$$G_t(p) = \int_0^t h(\tau)e^{-p\tau}d\tau$$

Protože $sin\omega t = \frac{1}{2i}(e^{i\omega t} - e^{-i\omega t})$, pak

$$\begin{aligned} y(t) &= \frac{a}{2i}\int_0^t h(\tau)[e^{i\omega(t-\tau)} - e^{-i\omega(t-\tau)}]d\tau \\ &= \frac{a}{2i}[e^{i\omega t}G_t(i\omega) - e^{-i\omega t}G_t(-i\omega)] \\ &= \frac{a}{2i} \mid G_t(i\omega) \mid [e^{i\omega t}e^{i\arg G_t(i\omega)} - e^{-i\omega t}e^{-i\arg G_t(i\omega)}] \\ &= a \mid G_t(i\omega) \mid sin(\omega t + \arg G_t(i\omega)) \end{aligned}$$

Když $t \to \infty$, pak $G_t(i\omega) \to G(i\omega)$.

Měřením (či odhadem na základě měření) amplitud a a b i fázové diference φ pro zvolené ω pak můžeme nalézt komplexní proměnnou $G(i\omega)$ z (3.2.4). Jestliže tento postup je opakován pro řadu frekvencí můžeme získat dosti dobrou grafickou reprezentaci $G(i\omega)$ jako funkci ω. Pro klasický návrh řídicích systémů je pak velmi vhodná Bodeho logaritmická frekvenční charakteristika nebo Nyquistova křivka.

Nastíněný postup je však dosti citlivý na šum. V této jednoduché podobě může být v praxi použit jen zřídka. Naznačme proč. Předpokládejme, že skutečný systém může být popsán rovnicí

$$Y(p) = G(p)U(p) + E(p) \tag{3.2.5}$$

s $e(t)$ představující stochastickou poruchu a $E(p)$ její Laplaceovu transformaci. Pak místo (3.2.3) dostaneme

$$y(t) = b \cdot \sin(\omega t + \varphi) + e(t) \tag{3.2.6}$$

a v důsledku přítomnosti šumu bude těžké odhadnout přesně amplitudu b a fázovou diferenci φ. Pomocí korelační techniky však předchozí postup můžeme zdokonalit. Vynásobením výstupu

$\sin(\omega t)$ a $\cos(\omega t)$ a integrací, získáme signály y_s a y_c. Aplikace této techniky na (3.2.6) vede na

$$\begin{aligned} y_s(T) &= \int_0^T y(t)\sin(\omega t)dt \\ &= \int_0^T b\sin(\omega t+\varphi)\sin(\omega t)dt + \int_0^T e(t)\sin(\omega t)dt \\ &= \frac{bT}{2}\cos(\varphi) - \frac{b}{2}\int_0^T \cos(2\omega t+\varphi)dt + \int_0^T e(t)\sin(\omega t)dt \end{aligned} \tag{3.2.7}$$

$$\begin{aligned} y_c(T) &= \int_0^T y(t)\cos(\omega t)dt \\ &= \int_0^T b\sin(\omega t+\varphi)\cos(\omega t)dt + \int_0^T e(t)\cos(\omega t)dt \\ &= \frac{bT}{2}\sin(\varphi) - \frac{b}{2}\int_0^T \sin(2\omega+\varphi)dt + \int_0^T e(t)\cos(\omega t)dt \end{aligned} \tag{3.2.8}$$

Jestliže měření neobsahuje šum ($e(t)=0$) a doba integrace T je násobek periody funkce $\sin(\cdot)$, to jest $T = k2\pi/\omega$, dostaneme

$$\begin{aligned} y_s(T) &= \frac{bT}{2}\cos(\varphi) \\ y_c(t) &= \frac{bT}{2}\sin(\varphi) \end{aligned} \tag{3.2.9}$$

Z těchto vztahů můžeme určit b a φ. Pak $\mid G(i\omega) \mid$ můžeme vypočítat podle (3.2.4). Poznamenejme, že (3.2.4) a (3.2.9) implikují

$$\begin{aligned} y_s(T) &= \frac{aT}{2} ReG(i\omega) \\ y_c(T) &= \frac{aT}{2} ImG(i\omega) \end{aligned} \tag{3.2.10}$$

což je užitečné pro popis $G(i\omega)$.

Intuitivně je zřejmé, že přístup využívající integrál lépe omezuje vliv šumu než základní metoda frekvenční analýzy. Důvod je jednoduše v tom, že využíváme informaci z delšího časového intervalu.

Pro aplikaci frekvenční analýzy tak jak je v (3.2.5)-(3.2.10), lze použít komerčně dostupná zařízení. Nevýhoda spojená s aplikací frekvenční analýzy spočívá v tom, že často vyžaduje časově dlouhé experimenty. Připomeňme, že pro každou zpracovávanou frekvenci, musíme nejprve dostat systém do „stacionární fáze“ a pak provést integraci.

3.3 Přechodová analýza

V tomto přístupu je modelem buď přechodová charakteristika nebo impulsní charakteristika. To znamená, že vstupem do reálného systému, který se nachází v ustáleném stavu, je jednotkový

skok nebo impuls a odezva na výstupu je zaznamenána. Z těchto křivek pak chceme najít model nízkého řádu (1. nebo 2.) typu (3.2.1). Takový přístup může být dosti citlivý na šum. Na druhé straně často lze jednoduše použít. Přinejmenším může dát první hrubý model, který ukazuje na zisk, dominantní časovou konstantu, tlumení a případně časové zpoždění. Použití impulsu jako vstupu je běžné v určitých aplikacích, např. když vstup je vstříknutí nějaké látky. To je typické v „tekoucích systémech". Souhrnně řečeno, přechodová analýza je vhodný prostředek pro získání hrubých modelů, ale její použitelnost je limitována. Přechodovou a frekvenční analýzu proto můžeme chápat jako základní prostředky pro návrh jednoduchých regulátorů (např. ladění PID regulátorů Ziegler-Nicholsovým pravidlem).

3.4 Korelační analýza

Třetí přístup, který budeme diskutovat, je korelační analýza. Použitá forma modelu je v tomto případě

$$y(t) = \sum_{k=0}^{\infty} h(k)u(t-k) + v(t) \tag{3.4.1}$$

kde $h(k)$ je váhová sekvence a $v(t)$ představuje poruchu. Předpokládejme, že vstup je stacionární stochastický proces, který je nezávislý na poruše. Pak pro (auto)kovarianční funkce platí následující vztah (Wiener-Hopfova rovnice)

$$r_{yu}(\tau) = \sum_{k=0}^{\infty} h(k)r_u(\tau-k) \tag{3.4.2}$$

kde $r_{yu}(\tau) = Ey(t+\tau)u(t)$ a $r_u(\tau) = Eu(t+\tau)u(t)$ jsou členy autokovarianční funkce v (3.4.2), které mohou být odhadnuty z dat

$$\hat{r}_{yu}(\tau) = \frac{1}{N - \max(\tau,0) + \min(\tau,0)} \sum_{t=1-\min(\tau,0)}^{N-\max(\tau,0)} y(t+\tau)u(t) \quad \tau = 0, \pm 1, \pm 2, ...$$

$$\hat{r}_u(\tau) = \frac{1}{N-\tau} \sum_{t=1}^{N-\tau} u(t+\tau)u(t) \quad \hat{r}_u(-\tau) = \hat{r}_u(\tau) \quad \tau = 0, 1, 2... \tag{3.4.3}$$

Pak odhad váhové funkce $\{h(k)\}$ je určen, v principu, řešením soustavy s nekonečným počtem lineárních rovnic

$$\hat{r}_{yu}(\tau) = \sum_{k=0}^{\infty} \hat{h}(k)\hat{r}_u(\tau-k) \tag{3.4.4}$$

Soustava lineárních rovnic (3.4.4) se významným způsobem zjednoduší, jestliže budeme používat jako vstup bílý šum, pro který platí $r_u(\tau) = 0$ pro $\tau \neq 0$. Tudíž z (3.4.2) je zřejmé, že

$$h(k) = r_{yu}(k)/r_u(0) \tag{3.4.5}$$

což lze snadno spočítat na základě odhadu $\hat{r}_{yu}(k)$ (3.4.3). Poznamenejme, že jsou k dispozici zařízení provádějící tyto operace automaticky.

I při použití bílého šumu jako vstupního signálu máme nekonečný počet rovnic, které umožní odhadnout váhovou sekvenci $\hat{h}(k)$. Z praktických důvodů je vhodnější použít tzv. useknutou váhovou funkci. To vede na lineární systém konečného řádu. Jestliže položíme

$$h(k) = 0 \quad \text{pro} \quad k \geq M \tag{3.4.6}$$

pak dostaneme model s konečnou impulsní odezvou (FIR, z anglického „finite impulse response“). Takové modely se často používají v oblasti zpracování signálů. Celé číslo M by mělo být vybráno velké ve srovnání s dominantní časovou konstantou systému. Pak (3.4.6) bude dobrá aproximace. Využitím (3.4.6) přejde (3.4.4) na

$$\hat{r}_{yu}(\tau) = \sum_{k=0}^{M-1} \hat{h}(k)\hat{r}_u(\tau - k) \tag{3.4.7}$$

Rozepíšeme-li tento vztah pro $\tau = 0, 1, 2, ..., M-1$, dostaneme soustavu lineárních rovnic

$$\begin{bmatrix} \hat{r}_{yu}(0) \\ \vdots \\ \hat{r}_{yu}(M-1) \end{bmatrix} = \begin{bmatrix} \hat{r}_u(0) & \dots & \hat{r}_u(M-1) \\ \hat{r}_u(1) & \dots & \dots \\ \vdots & \ddots & \dots \\ \hat{r}_u(M-1) & \dots & \hat{r}_u(0) \end{bmatrix} \begin{bmatrix} \hat{h}(0) \\ \vdots \\ \hat{h}(M-1) \end{bmatrix} \tag{3.4.8}$$

Pokud je jako vstup uvažován bílý šum, pak matice soustavy diagonální. Poznamenejme, že s korelační analýzou jsme se setkali již v kapitole druhé a v kapitole čtvrté bude vidět, jak postupovat v případě, kdy bude počet různých hodnot τ větší než je M, což odpovídá přeurčené soustavě lineárních rovnic.

3.5 Spektrální analýza

Poslední neparametrická metoda, kterou se budeme zabývat, je založená na spektrální analýze. Věnujme nejdříve pozornost vztahu mezi korelační funkcí a spektrální hustotou a výpočtu spektrální hustoty signálu po průchodu lineárním systémem.
Nechť $u(t)$ skalární stacionární stochastický proces. Předpokládejme, že jeho střední hodnota je m_u a autokovarianční funkce je

$$r_u(\tau) = E[u(t+\tau) - m_u][u(t) - m_u] \tag{3.5.1}$$

Dle definice je spektrální hustota stochastického procesu

$$\phi_u(\omega) \stackrel{\triangle}{=} \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} r_u(\tau) e^{-i\tau\omega} \tag{3.5.2}$$

Inverzní vztah k (3.5.2), který umožňuje výpočet autokovarianční funkce ze spektrální hustoty je

$$r_u(\tau) = \int_{-\pi}^{\pi} \phi_u(\omega) e^{i\tau\omega} d\omega \tag{3.5.3}$$

To můžeme snadno ověřit dosazením z (3.5.2) do (3.5.3)

$$\int_{-\pi}^{\pi} \frac{1}{2\pi} \sum_{\tau'=-\infty}^{\infty} r_u(\tau') e^{-\tau'\omega} e^{i\tau\omega} d\omega = \frac{1}{2\pi} \sum_{\tau'=-\infty}^{\infty} r_u(\tau') \int_{-\pi}^{\pi} e^{i(\tau-\tau')\omega} d\omega$$
$$= \sum_{\tau'=-\infty}^{\infty} r_u(\tau') \delta_{\tau,\tau'} = r_u(\tau)$$

kde $\delta_{\tau,\tau'} = 0$ pro $\tau \neq \tau'$ a $\delta_{\tau,\tau'} = 1$ pro $\tau = \tau'$.

Poznamenejme, že předpokládáme periodu vzorkování $T_s = 1$ a pak standardní interval pro ω je $(-\pi, \pi)$. V případě, že máme obecné T_s, pak Nyquistova frekvence je $\omega_N = \pi/T_s$ a integrační interval pro ω je $\omega \in (-\omega_N, \omega_N)$.

Nyní se zaměříme na odvození vztahu pro výpočet spektrální hustoty filtrovaného signálu. Uvažujme lineární filtraci $u(t)$ (průchod $u(t)$ lineárním systémem) podle vztahu

$$y(t) = \sum_{k=0}^{\infty} h(k) u(t-k) \tag{3.5.4}$$

kde $y(t)$ skalární signál a $\{h(k)\}$ je posloupnost definující váhovou funkci. Předpokládejme, že filtr v (3.5.4) je stabilní, to jest $\| h(k) \| \to 0$, když $k \to \infty$. V takovém případě $y(t)$ je stacionární signál. V následujícím vypočteme charakteristiky tohoto signálu, a to střední hodnotu m_y, vzájemnou kovarianční funkci $r_{yu}(\tau)$, vzájemnou spektrální hustotu $\phi_{yu}(\omega)$, kovarianční funkci $r_y(\tau)$ a spektrální hustotu $\phi_y(\omega)$. K tomu je vhodné zavést operátor přenosu

$$H(q^{-1}) = \sum_{k=0}^{\infty} h(k) q^{-k}, \tag{3.5.5}$$

kde q^{-k} je operátor zpětného k-krokového posunu. Využitím (3.5.5) lze model (3.5.4) zapsat jako

$$y(t) = H(q^{-1}) u(t) \tag{3.5.6}$$

Střední hodnotu $y(t)$ najdeme velmi lehce z (3.5.4)

$$m_y = Ey(t) = \sum_{k=0}^{\infty} h(k) Eu(t-k) = \sum_{k=0}^{\infty} h(k) m_u = H(1) m_u \tag{3.5.7}$$

přičemž $H(1)$ můžeme interpretovat jako zesílení filtru v ustáleném stavu. Před výpočtem autokovarianční funkce nejdříve definujeme

$$\tilde{y}(t) \stackrel{\triangle}{=} y(t) - m_y \quad \tilde{u}(t) \stackrel{\triangle}{=} u(t) - m_u$$

Vztah mezi $\tilde{y}(t)$ a $\tilde{u}(t)$ můžeme vypočítat takto

$$\tilde{y}(t) = \sum_{k=0}^{\infty} h(k) u(t-k) - \sum_{k=0}^{\infty} h(k) m_u = \sum_{k=0}^{\infty} h(k) \tilde{u}(t-k) = H(q^{-1}) \tilde{u}(t) \tag{3.5.8}$$

Při výpočtu kovariančních funkcí pro snazší zápis předpokládejme, že $m_u = 0$.
Nejdříve vypočteme kovarianční funkci pro $y(t)$.

$$\begin{aligned} r_y(\tau) &= Ey(t+\tau)y(t) \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h(j)Eu(t+\tau-j)u(t-k)h(k) \\ &= \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h(j)r_u(\tau-j+k)h(k) \end{aligned} \tag{3.5.9}$$

Nyní využitím definičního vztahu (3.5.2) a vztahu (3.5.9) vypočteme spektrální hustotu.

$$\begin{aligned} \phi_y(\omega) &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} r_y(\tau)e^{-i\tau\omega} \\ &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h(j)e^{-ij\omega}r_u(\tau-j+k)e^{-i(\tau-j+k)\omega}h(k)e^{ik\omega} \\ &= \frac{1}{2\pi}\sum_{j=0}^{\infty}\sum_{k=0}^{\infty} h(j)e^{-ij\omega}[\sum_{\tau'=-\infty}^{\infty} r_u(\tau')e^{-i\tau'\omega}]h(k)e^{ik\omega} \\ &= [\sum_{j=0}^{\infty} h(j)e^{-ij\omega}]\phi_u(\omega)[\sum_{k=0}^{\infty} h(k)e^{ik\omega}] \end{aligned}$$

nebo-li

$$\phi_y(\omega) = H(e^{-i\omega})\phi_u(\omega)H(e^{i\omega}) \tag{3.5.10}$$

Tento vztah popisuje, jak spektrální hustota výstupního signálu závisí na přenosové funkci $H(e^{i\omega})$ a spektrální hustotě vstupního signálu $\phi_u(\omega)$.

Zbývá vypočítat vzájemnou kovarianční funkci a vzájemnou spektrální hustotu.

$$\begin{aligned} r_{yu}(\tau) &= Ey(t+\tau)u(t) \\ &= \sum_{j=0}^{\infty} h(j)Eu(t+\tau-j)u(t) \\ &= \sum_{j=0}^{\infty} h(j)r_u(\tau-j) \end{aligned} \tag{3.5.11}$$

$$\begin{aligned} \phi_{yu}(\omega) &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} r_{yu}(\tau)e^{-i\tau\omega} \\ &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\sum_{j=0}^{\infty} h(j)e^{-ij\omega}r_u(\tau-j)e^{-i(\tau-j)\omega} \\ &= \sum_{j=0}^{\infty} h(j)e^{-ij\omega}[\frac{1}{2\pi}\sum_{\tau'=-\infty}^{\infty} r_u(\tau')e^{-i\tau'\omega}] \end{aligned}$$

nebo-li

$$\phi_{yu}(\omega) = H(e^{-i\omega})\phi_u(\omega) \tag{3.5.12}$$

kde

$$\begin{aligned}
\phi_{yu}(\omega) &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} r_{yu}(\tau)e^{-i\tau\omega} \\
\phi_u(\omega) &= \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty} r_u(\tau)e^{-i\tau\omega} \\
H(e^{-i\omega}) &= \sum_{k=0}^{\infty} h(k)e^{-ik\omega}
\end{aligned} \tag{3.5.13}$$

Přenosová funkce $H(e^{-i\omega})$ může být odhadnuta z

$$\hat{H}(e^{-i\omega}) = \hat{\phi}_{yu}(\omega)/\hat{\phi}_u(\omega) \tag{3.5.14}$$

kde odhady spektrálních hustot lze určit jako

$$\hat{\phi}_{yu}(\omega) = \frac{1}{2\pi}\sum_{\tau=-N}^{N} \hat{r}_{yu}(\tau)e^{-i\tau\omega} \tag{3.5.15}$$

$$\hat{\phi}_u(\omega) = \frac{1}{2\pi}\sum_{\tau=-N}^{N} \hat{r}_u(\tau)e^{-i\tau\omega} \tag{3.5.16}$$

Výpočet (3.5.15) by mohl být proveden následujícím způsobem. Dosazením za $\hat{r}_{yu}(\tau)$ dostaneme

$$\hat{\phi}_{yu}(\omega) = \frac{1}{2\pi N}\sum_{\tau=-N}^{N}\sum_{t=1-\min(\tau,0)}^{N-\max(\tau,0)} y(t+\tau)u(t)e^{-i\tau\omega}$$

Zavedením substituce $s = t + \tau$ dostaneme

$$\begin{aligned}
\hat{\phi}_{yu}(\omega) &= \frac{1}{2\pi N}\sum_{s=1}^{N}\sum_{t=1}^{N} y(s)u(t)e^{-is\omega}e^{it\omega} \\
&= \frac{1}{2\pi N}Y_N(\omega)U_N(-\omega)
\end{aligned} \tag{3.5.17}$$

kde

$$\begin{aligned}
Y_N(\omega) &= \sum_{s=1}^{N} y(s)e^{-is\omega} \\
U_N(\omega) &= \sum_{s=1}^{N} u(s)e^{-is\omega}
\end{aligned} \tag{3.5.18}$$

jsou diskrétní Fourierovy transformace sekvencí $\{y(t)\}$ a $\{u(t)\}$. Pro $\omega = 0, 2\pi/N, 4\pi/N, ..., \pi$ mohou být vypočteny rychlou Fourierovou transformací (FFT, z anglického „fast Fourier transform“).

Obdobně

$$\hat{\phi}_u(\omega) = \frac{1}{2\pi N} U_N(\omega) U_N(-\omega) = \frac{1}{2\pi N} \mid U_N(\omega) \mid^2 \tag{3.5.19}$$

Takový odhad spektrální hustoty se nazývá periodogram. Z (3.5.14), (3.5.17), (3.5.19) zjistíme odhad přenosové funkce

$$\hat{H}(e^{-i\omega}) = Y_N(\omega)/U_N(\omega) \tag{3.5.20}$$

Předchozí postup k získání odhadu spektrálních hustot a následně odhadu přenosové funkce vede k nepříliš přesným výsledkům. Jedním z důvodů pro toto chování je, že odhad $\hat{r}_{yu}(\tau)$ bude dost nepřesný pro velké hodnoty τ, ale jednotlivé $\hat{r}_{yu}(\tau)$ jsou váženy se stejnou vahou (v (3.5.15)). Dále v (3.5.15) je sčítáno $2N + 1$ členů. Přestože estimační chyba každého členu se blíží nule pro $N \to \infty$, není žádná jistota, že celková estimační chyba součtu také směřuje k nule. Tyto problémy mohou být překonány, jestliže v (3.5.15) členy s velkými hodnotami τ mají odlišnou váhu. Takže místo (3.5.15) je vhodnější použít následující odhad pro vzájemné spektrum

$$\hat{\phi}_{yu}(\omega) = \frac{1}{2\pi} \sum_{\tau=-N}^{N} \hat{r}_{yu}(\tau) w(\tau) e^{-i\tau\omega} \tag{3.5.21}$$

kde $w(\tau)$ je tzv. zpožďovací okénko. Obecně je okénko voleno tak, aby bylo rovno jedné pro $\tau = 0$, a klesalo s rostoucím τ, a bylo rovno nule pro „velké hodnoty“ τ. „Velké hodnoty“ znamená 5-10 procent z počtu dat N. Analogické úvahy platí pro $\hat{\phi}_u(\omega)$. V literatuře jsou uváděny nejčastěji tyto typy okének:

$$\begin{aligned} w_1(\tau) = 1 \quad &\mid \tau \mid \leq M \quad \text{pravoúhlé} \\ w_1(\tau) = 0 \quad &\mid \tau \mid > M \end{aligned} \tag{3.5.22}$$

$$\begin{aligned} w_2(\tau) = 1 - \mid \tau \mid /M \quad &\mid \tau \mid \leq M \quad \text{Barlettovo} \\ w_2(\tau) = 0 \quad &\mid \tau \mid > M \end{aligned} \tag{3.5.23}$$

$$w_3(\tau) = \frac{1}{2}\left(1 + \cos\left(\frac{\pi\tau}{M}\right)\right) \quad \mid \tau \mid \leq M \quad \text{Hammingovo}$$

$$w_3(\tau) = 0 \quad \mid \tau \mid > M \tag{3.5.24}$$

Spektrální analýza nemá žádné speciální omezení na vstupní signál kromě toho, že musí být nekorelovaný s poruchou. Proto je poměrně oblíbená pro nejrůznější aplikace, od analýzy řeči a testování mechanických vibrací až ke geofyzikálním výzkumům, nehledě na užití při analýze a syntéze řídicích systémů.

3.6 Shrnutí

Na závěr této kapitoly charakterizujme jednotlivé metody:

- Frekvenční analýza vyžaduje dosti dlouhé identifikační experimenty zvláště tehdy, jestliže chceme použít nějakou techniku na redukci šumu. Vstupní signál je typu sinus.
- Přechodová analýza je snadno aplikovatelná, ale velmi citlivá na šum, a tudíž můžeme očekávat pouze hrubý model. Vstupní signál je skok nebo impuls.
- Korelační analýza je založena na použití bílého šumu jako vstupu. Výsledným modelem je váhová funkce. Je málo citlivá na aditivní šum ovlivňující výstupní signál.
- Spektrální analýza je necitlivá na šum. Může být použita pro poměrně libovolný vstupní signál. Přenosová funkce je získána ve formě logaritmické frekvenční charakteristiky (nebo jiné ekvivalentní formy). Pro získání rozumně přesného odhadu musí být použito zpožďovací okénko.

Neparametrickými identifikačními metodami se zabývají např. práce [2], [10], [15], [20], [25] - [31], [58], [59].

Kapitola 4

Lineární regrese

4.1 Metoda nejmenších čtverců

V této kapitole se budeme věnovat lineární regresi, která je velmi často používaná a to nejen ve statistice. Struktura modelu pro lineární regresi může být vyjádřena vztahem

$$y(t) = \varphi^T(t)\Theta \tag{4.1.1}$$

kde $y(t)$ je měřitelná veličina, $\varphi(t)$ je známý vektor dimenze n a Θ je vektor neznámých parametrů dimenze n. Prvky vektoru $\varphi(t)$ jsou často nazývány regresní proměnné nebo prostě regresory. Vektor Θ nazývejme vektor parametrů. Model (4.1.1) lze přímočarým způsobem zobecnit na vícerozměrný případ. Model s více měřitelnými veličinami lze zapsat jako

$$y(t) = \Phi^T(t)\Theta \tag{4.1.2}$$

kde $y(t)$ je vektor dimenze p, tj. $dim(y(t)) = p$, $\Phi(t)$ je matice dimenze n/p a Θ je vektor dimenze n. Je velmi přirozené interpretovat t jako čas, ale pro statické modely to není nutné, jak bude v této kapitole ukázáno.

Nyní uvedeme několik příkladů lineárních regresních modelů.

Příklad 4.1.1 (Polynomiální trend)
Nechť model je polynomiální trend a má formu

$$y(t) = a_0 + a_1 t + ... + a_r t^r$$

kde r je řád polynomu a $a_0, ..., a_r$ jsou neznámé koeficienty. Výše uvedenou formu modelu (4.4.1) získáme, pokud zavedeme značení

$$\varphi(t) = [1, t, ..., t^r]^T$$
$$\Theta = [a_0, a_1, ..., a_r]^T$$

Tento model může být použit k popisu časových řad, kde řád polynomu r je volen jako celé číslo. Když $r = 0$ model popisuje pouze konstantu, pro $r = 1$ popisuje už jakoukoliv lineární funkci, pro $r = 2$ kvadratickou atd. Všimněme si tedy, že ačkoliv se jedná o lineární regresi, výsledný model může popisovat nelineární funkci. Význam slova „lineární“ tak zdůrazňuje

lineární závislost výstupní veličiny $y(t)$ na hledaných parametrech Θ.

Příklad 4.1.2 (Vážený součet exponenciál)
Při analýze přechodových procesů je vhodný model

$$y(t) = b_1 e^{-k_1 t} + ... + b_n e^{-k_n t}$$

Předpokládejme, že $k_1, ..., k_n$ jsou známé inverzní časové konstanty, ale váhy $b_1, ..., b_n$ jsou neznámé. V tomto případě můžeme položit

$$\varphi(t) = [e^{-k_1 t}, ..., e^{-k_n t}]^T$$
$$\Theta = [b_1, ..., b_n]^T$$

Příklad 4.1.3 (Váhová funkce)
Model dynamického systému založený na „useknuté“ váhové funkci. Takový model lze vyjádřit vztahem

$$y(t) = h_0 u(t) + h_1 u(t-1) + ... + h_{M-1} u(t-M+1)$$

Vstupní signál $u(t)$ je zaznamenán během experimentu a může být proto považován za známý. V tomto případě

$$\varphi(t) = [u(t), u(t-1), ..., u(t-M+1)]^T$$
$$\Theta = (h_0, ..., h_{M-1})^T$$

Abychom dostali dostatečně přesný popis dynamiky systému, vyžaduje tento model často mnoho parametrů (M může být 20-50, v oblasti aplikací zpracování signálů několik set nebo dokonce tisíc). Nicméně takový přístup je koncepčně dosti jednoduchý a patří do rámce této kapitoly.

Po uvedení několika příkladů regresních modelů se nyní budeme věnovat odhadu parametrů. Chceme nalézt odhad $\hat{\Theta}$ vektoru parametrů Θ z měření $y(1), y(2), \ldots, y(N)$ a regresorů $\varphi(1), \varphi(1), \ldots, \varphi(N)$. Za tohoto předpokladu můžeme sestavit soustavu lineárních rovnic

$$\begin{aligned} y(1) &= \varphi^T(1)\Theta \\ y(2) &= \varphi^T(2)\Theta \\ &\vdots \\ y(N) &= \varphi^T(N)\Theta \end{aligned}$$

V maticovém zápisu pak

$$Y = \Phi\Theta \tag{4.1.3}$$

kde

$$Y = \begin{bmatrix} y(1) \\ . \\ . \\ . \\ y(N) \end{bmatrix} \quad \text{vektor dimenze } N$$

$$\Phi = \begin{bmatrix} \varphi^T(1) \\ . \\ . \\ . \\ \varphi^T(N) \end{bmatrix} \quad \text{matice } N/n \tag{4.1.4}$$

Jedna možnost jak nalézt Θ z (4.1.3), je samozřejmě použít takové množství měření, aby $N = n$. Pak Φ bude čtvercová matice. Jestliže je tato matice regulární, má pak soustava lineárních rovnic (4.1.3) jednoznačné řešení. S ohledem na výskyt šumu, poruch, nedokonalého modelu je v praxi rozumné použít větší počet dat než je n. Pak je možné očekávat i zlepšený odhad. Avšak pro $N > n$ je soustava lineárních rovnic (4.1.3) přeurčená a ve většině případů nemá jednoznačné řešení. Z těchto důvodů je vhodné zavést chyby rovnice

$$\epsilon(t) = y(t) - \varphi^T(t)\Theta \tag{4.1.5}$$

a z těchto hodnot vytvořit vektor ϵ

$$\epsilon = \begin{bmatrix} \epsilon(1) \\ . \\ . \\ . \\ \epsilon(N) \end{bmatrix}$$

Ve statistické literatuře jsou chyby rovnice nazývány rezidua. Odhad Θ ve smyslu nejmenších čtverců je definován jako vektor $\hat{\Theta}$, který minimalizuje ztrátovou funkci

$$V(\Theta) = \frac{1}{2}\sum_{t=1}^{N}\epsilon^2(t) = \frac{1}{2}\epsilon^T\epsilon = \frac{1}{2} \parallel \epsilon \parallel^2 \tag{4.1.6}$$

kde $\| \, . \, \|$ označuje Euklidovu normu. Z (4.1.5) je zřejmé, že chyba rovnice $\epsilon(t)$ je lineární funkcí vektoru parametrů Θ.

Řešení tohoto optimalizačního problému ukazuje následující věta.

Věta 4.1.1 Uvažujme ztrátovou funkci $V(\Theta)$ určenou (4.1.5) a (4.1.6). Předpokládejme, že matice $\Phi^T\Phi$ je pozitivně definitní. Pak $V(\Theta)$ má jediné minimum

$$\begin{aligned}\hat{\Theta} &= (\Phi^T\Phi)^{-1}\Phi^T Y \\ &= \Phi^\dagger Y \end{aligned} \tag{4.1.7}$$

Odpovídající minimální hodnota $V(\Theta)$ je pak

$$\begin{aligned}\min_{\Theta} V(\Theta) = V(\hat{\Theta}) &= \frac{1}{2}[Y^TY - Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^TY] \\ &= \frac{1}{2}Y^T[I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T]Y\end{aligned} \tag{4.1.8}$$

Poznamenejme, že výraz

- $\Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T$ se označuje jako Moore-Penroseova pseudoinverse či zobecněná inverse obdélníkové matice Φ. Pseudoinverse přechází v inverzi, pokud Φ je čtvercová regulární matice a
- $\Phi^\perp = [I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T]$ značí ortogonální projekci s vlastností $\Phi^\perp\Phi = 0$,

kde I_N značí jednotkovou matici dimenze N.

Důkaz. Použitím (4.1.3), (4.1.5) a (4.1.6) můžeme nalézt explicitní vyjádření ztrátové funkce $V(\Theta)$. Lze vidět, že $V(\Theta)$ jako funkce Θ má kvadratický, lineární a konstantní člen. Tudíž je možné použít techniku doplnění na čtverec. Máme

$$\epsilon = Y - \Phi\Theta$$

a

$$\begin{aligned}V(\Theta) &= \frac{1}{2}[Y - \Phi\Theta]^T[Y - \Phi\Theta] \\ &= \frac{1}{2}[\Theta^T\Phi^T\Phi\Theta - \Theta^T\Phi^TY - Y^T\Phi\Theta + Y^TY]\end{aligned}$$

Pak

$$\begin{aligned}V(\Theta) &= \frac{1}{2}[\Theta - (\Phi^T\Phi)^{-1}\Phi^TY]^T\Phi^T\Phi[\Theta - (\Phi^T\Phi)^{-1}\Phi^TY] \\ &+ \frac{1}{2}[Y^TY - Y^T\Phi(\Phi^T\Phi)^{-1}\Phi^TY]\end{aligned}$$

kde druhý člen nezávisí na Θ. Protože $\Phi^T\Phi$ je podle předpokladu pozitivně definitní, bude první člen vždy větší nebo roven nule. Tudíž můžeme minimalizovat $V(\Theta)$ položením prvního členu rovno nule. Ale tím dostaneme přesně (4.1.7) a zároveň okamžitě i minimální hodnotou ztrátové funkce.

Poznámka . Matice $\Phi^T\Phi$, jak vyplývá z konstrukce, je vždy nonnegativně definitní (nebo-li pozitivně semi-definitiní). Jestliže je singulární (pozitivně semidefinitní), předchozí výpočty

neplatí. Pak můžeme vypočítat gradient ztrátové funkce. Dostaneme

$$0 = \frac{dV(\Theta)}{d\Theta} = -Y^T\Phi + \Theta^T(\Phi^T\Phi)$$

nebo-li

$$(\Phi^T\Phi)\Theta = \Phi^T Y \tag{4.1.9}$$

Když $\Phi^T\Phi$ je singulární, pak Φ nemá plnou hodnost (to jest hodnost $\Phi < n$). Jestliže však jsou experiment a struktura modelu dobře zvoleny a je dostatečný počet rovnic N, pak matice Φ by měla mít plnou hodnost.

Příklad 4.1.4. Uvažujme model (4.1.1) ve tvaru

$$y(t) = b.$$

To znamená, že chceme odhadnout konstantu z měření, které mohou být pod vlivem šumu. Dostaneme

$$\varphi(t) = 1 \qquad \Theta = b$$

$$\Phi = \begin{bmatrix} 1 \\ . \\ . \\ . \\ 1 \end{bmatrix}$$

a tudíž (4.1.7) je

$$\hat{\Theta} = \frac{1}{N}[y(1) + ... + y(N)]$$

tedy jednoduše aritmetický průměr všech měření.

Poznámka . V případě, že je k dispozici méně rovnic N než je neznámých n, tj. $N < n$, existuje nekonečně mnoho řešení soustavy (4.1.3) pro odhad parametrů Θ. Uvažujme pro jednoduchost následující nedourčenou soustavu N (deterministických) lineárních rovnic o n neznámých

$$Y = \Phi\Theta$$

kde $N < n$ a dimenze Θ je N. Jednou z možností pro nalezení odhadu Θ je pak zvolit $n - N$ hledaných parametrů na základě apriorní znalosti a zbylé dopočítat "standardní" metodou nejmenších čtverců. Druhou možností, při nedostatku apriorních znalostí, je následující odhad

$$\hat{\Theta}_{\text{minNorm}} = \Phi^T(\Phi\Phi^T)^{-1}Y \tag{4.1.10}$$

který je optimální dle následujícího kriteria

$$\hat{\Theta}_{\text{minNorm}} = \arg\min_{\Theta} \|\Theta\|^2 \tag{4.1.11}$$

Platnost tohoto tvrzení lze snadno ověřit následujícím výpočtem. Uvažujme libovolný vektor Θ splňující rovnost $Y = \Phi\Theta$, pak lze psát

$$\begin{aligned}
\|\Theta\|^2 &= \Theta^T\Theta \\
&= (\Theta - \hat{\Theta}_{\text{minNorm}} + \hat{\Theta}_{\text{minNorm}})^T(\Theta - \hat{\Theta}_{\text{minNorm}} + \hat{\Theta}_{\text{minNorm}}) \\
&= (\Theta - \hat{\Theta}_{\text{minNorm}})^T(\Theta - \hat{\Theta}_{\text{minNorm}}) + \hat{\Theta}^T_{\text{minNorm}}\hat{\Theta}_{\text{minNorm}} \\
&= \|\Theta - \hat{\Theta}_{\text{minNorm}}\|^2 + \|\hat{\Theta}_{\text{minNorm}}\|^2 \\
&\geq \|\hat{\Theta}_{\text{minNorm}}\|^2
\end{aligned} \tag{4.1.12}$$

kde byla použita rovnost

$$\begin{aligned}
(\Theta - \hat{\Theta}_{\text{minNorm}})^T\hat{\Theta}_{\text{minNorm}} &= (\Theta - \hat{\Theta}_{\text{minNorm}})^T\Phi^T(\Phi\Phi^T)^{-1}Y \\
&= (\Phi(\Theta - \hat{\Theta}_{\text{minNorm}}))^T(\Phi\Phi^T)^{-1}Y \\
&= (\Phi\Theta - \Phi\hat{\Theta}_{\text{minNorm}})^T(\Phi\Phi^T)^{-1}Y \\
&= (Y - \Phi\Phi^T(\Phi\Phi^T)^{-1}Y)^T(\Phi\Phi^T)^{-1}Y \\
&= 0
\end{aligned} \tag{4.1.13}$$

Odhad (4.1.10) lze tedy chápat jako řešení optimalizačního problému (4.1.11) s omezením $Y = \Phi\Theta$ a odhad je tedy *stranný*. Poznamenejme ještě, vzhledem k vlastnostem kvadratické funkce, která je minimalizována, nebudou typicky jednotlivé elementy odhadu $\hat{\Theta}_{\text{minNorm}}$ (4.1.10) nabývat „extrémních“ hodnot (např. první element vektoru $\hat{\Theta}_{\text{minNorm}}(1)$ blízký nule, zatímco $\hat{\Theta}_{\text{minNorm}}(2)$ v řádech tisíců), ale spíše elementy odhadu budou nabývat „podobných“ hodnot. To může být užitečné z hlediska numerického.

4.2 Analýza

Nyní se budeme věnovat statistickým vlastnostem odhadu (4.1.7). Nejprve stanovme předpoklady na data. Nechť data vyhovují následujícímu vztahu

$$y(t) = \varphi^T(t)\Theta_0 + e(t) \tag{4.2.1}$$

kde Θ_0 je vektor skutečných parametrů a $\varphi^T(t)$ je deterministický vektor regresorů. Dále předpokládejme, že $e(t)$ je stochastická absolutně náhodná proměnná s nulovou střední hodnotou a variancí λ^2, tj. bílý šum. Maticový zápis rovnice (4.2.1) je

$$Y = \Phi\Theta_0 + e \tag{4.2.2}$$

kde

$$e = \begin{bmatrix} e(1) \\ . \\ . \\ . \\ e(N) \end{bmatrix}$$

Statistické vlastnosti odhadu pak ukazuje následující věta.

Věta 4.2.1. Uvažujme odhad (4.1.7). Předpokládejme, že data splňují (4.2.1), kde $e(t)$ je bílý šum s nulovou střední hodnotou a variancí λ^2. Pak

i) $\hat{\Theta}$ je nestranný odhad Θ_0

ii) Kovarianční matice $\hat{\Theta}$ je $\lambda^2(\Phi^T\Phi)^{-1}$

iii) Nestranný odhad λ^2 je dán $s^2 = 2V(\hat{\Theta})/(N-n)$.

Důkaz. Použitím rovnic (4.1.7) a (4.2.1) dostaneme

$$\hat{\Theta} = (\Phi^T\Phi)^{-1}\Phi^T\{\Phi\Theta_0 + e\} = \Theta_0 + (\Phi^T\Phi)^{-1}\Phi^T e$$

a tudíž

$$E\hat{\Theta} = \Theta_0 + (\Phi^T\Phi)^{-1}\Phi^T E[e] = \Theta_0$$

což dokazuje bod i).

Pro důkaz bodu ii) poznamenejme, že předpoklad na bílý šum implikuje $Eee^T = \lambda^2 I$. Pak

$E(\hat{\Theta} - \Theta_0)(\hat{\Theta} - \Theta_0)^T = E[(\Phi^T\Phi)^{-1}\Phi^T e][(\Phi^T\Phi)^{-1}\Phi^T e]^T = (\Phi^T\Phi)^{-1}\Phi^T E[ee^T]\Phi(\Phi^T\Phi)^{-1} = (\Phi^T\Phi)^{-1}\Phi^T\lambda^2 I\Phi(\Phi^T\Phi)^{-1} = \lambda^2(\Phi^T\Phi)^{-1}$

což dokazuje bod ii).

Minimální hodnota $V(\hat{\Theta})$ ztrátové funkce může být podle (4.1.8) a (4.2.2) zapsána

$$\begin{aligned} V(\hat{\Theta}) &= \frac{1}{2}[\Phi\Theta_0 + e]^T[I - \Phi(\Phi^T\Phi)^{-1}\Phi^T][\Phi\Theta_0 + e] \\ &= \frac{1}{2}e^T[I - \Phi(\Phi^T\Phi)^{-1}\Phi^T]e \end{aligned}$$

Střední hodnota odhadu s^2 může pak být vypočtena takto:

$$\begin{aligned} E[s^2] &= 2E[V(\hat{\Theta})/(N-n)] = E[tr\{e^T[I - \Phi(\Phi^T\Phi)^{-1}\Phi^T]e\}/(N-n)] \\ &= E[tr\{[I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T]ee^T\}/(N-n)] \\ &= tr[I_N - \Phi(\Phi^T\Phi)^{-1}\Phi^T]\lambda^2 I_N/(N-n) \\ &= [trI_N - tr\{\Phi(\Phi^T\Phi)^{-1}\Phi^T\}]\lambda^2/(N-n) \\ &= [trI_N - tr\{(\Phi^T\Phi)^{-1}(\Phi^T\Phi)\}]\lambda^2/(N-n) \\ &= [trI_N - trI_n]\lambda^2/(N-n) = (N-n)\lambda^2/(N-n) = \lambda^2 \end{aligned}$$

Při výpočtech $I_N(I_n)$ označuje identickou matici řádu $N(n)$ a symbol tr je stopa matice. Ukázali jsme, že s^2 je nestranný odhad λ^2, což dokazuje bod iii).

Poznámka . Poznamenejme, že v důkazu je nezbytné, aby Φ byla deterministická matice. Pak bylo možné vytknout matici Φ ze střední hodnoty a dokázat zmíněné statistické vlastnosti

odhadu. Při uvažování stochastické matice Φ je odvykle nezbytné diskutovat vlastnosti odhadu za předpokladu nekonečně mnoha dat, jak bylo naznačeno v kapitole 2.5.

Ve větě 4.2.1 jsme předpokládali, že porucha $e(t)$ ve (4.2.1) je bílý šum, to jest, že $e(t)$ a $e(s)$ jsou nekorelované pro všechna $t \neq s$. Zkoumejme nyní co se stane, jestliže předpoklad není splněn. Předpokládejme, že (4.2.1) platí, ale

$$E[ee^T] = R \tag{4.2.3}$$

kde R je pozitivně definitní matice s nenulovými prvky mimo diagonálu. Z důkazu věty je vidět, že $\hat{\Theta}$ bude stále nestranný odhad Θ_0. Avšak kovarianční matice $\hat{\Theta}$ bude nyní

$$cov[\hat{\Theta}] = (\Phi^T\Phi)^{-1}\Phi^T R\Phi(\Phi^T\Phi)^{-1} \tag{4.2.4}$$

4.3 Nejlepší lineární nestranný odhad

Rozšiřme třídu identifikačních metod a uvažujme obecný lineární odhad Θ. Takový odhad lze zapsat ve tvaru

$$\hat{\Theta} = ZY, \tag{4.3.1}$$

kde Z je matice n/N závislá na matici regresorů Φ. Všimněme si, že odhad ve smyslu nejmenších čtverců (4.1.7) je speciální případ (4.3.1), který dostaneme pro

$$Z = \Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T$$

Ukážeme nyní, jak vybrat Z, aby odhad byl nejen nestranný, ale aby měl i minimální kovarianční matici. Takový odhad je nazýván jako nejlepší lineární nestranný odhad (BLUE, z anglického „best linear unbiased estimate“) nebo také Markovský odhad.

Věta 4.3.1. Uvažujme odhad (4.3.1). Předpokládejme, že data splňují (4.2.1), popř. (4.2.3). Nechť

$$Z^* = (\Phi^T R^{-1}\Phi)^{-1}\Phi^T R^{-1} \tag{4.3.2}$$

Pak odhad $\hat{\Theta}^* = Z^*Y$ je nestranný odhad Θ_0. Dále, kovarianční matice chyby odhadu je minimální ve smyslu

$$cov_{Z^*}[\hat{\Theta}^*] = (\Phi^T R^{-1}\Phi)^{-1} \leq cov_Z[\hat{\Theta}] \tag{4.3.3}$$

pro všechny nestranné lineární odhady $\hat{\Theta}$ (4.3.1). Poznamenejme, že maticová nerovnost $P_1 \leq P_2$ značí, že $P_2 - P_1$ je pozitivně semi-definitní matice.

Důkaz. Požadavek na nestranný odhad je

$$\Theta_0 = E[\hat{\Theta}] = E[Z(\Phi\Theta_0 + e)] = Z\Phi\Theta_0$$

Jelikož Θ_0 je libovolné, musí platit

$$Z\Phi = I \tag{4.3.4}$$

Poznamenejme, že výběr (4.3.2) splňuje (4.3.4). Pak kovarianční matice, v obecném případě, bude

$$cov_Z[\hat{\Theta}] = E[(ZY - \Theta_0)(ZY - \Theta_0)^T] = ZRZ^T \tag{4.3.5}$$

Jestliže vybereme $Z = Z^*$ ze (4.3.2) dostaneme

$$cov_{Z^*}[\hat{\Theta}^*] = (\Phi^T R^{-1}\Phi)^{-1}\Phi^T R^{-1} R R^{-1}\Phi(\Phi^T R^{-1}\Phi)^{-1} = (\Phi^T R^{-1}\Phi)^{-1} \tag{4.3.6}$$

z čehož vyplývá rovnost v (4.3.3). Zabývejme se nyní nerovností v (4.3.3).
Nechť $\tilde{\Theta}$ označuje chybu odhadu $\hat{\Theta} - \Theta_0$ pro obecný odhad ve smyslu (4.3.1) a $\tilde{\Theta}^*$ pro odhad $\hat{\Theta}^*$ spočtený na základě (4.3.2). Pak dostaneme

$$cov_Z[\hat{\Theta}] = E[\tilde{\Theta}\tilde{\Theta}^T] = E[\tilde{\Theta} - \tilde{\Theta}^*][\tilde{\Theta} - \tilde{\Theta}^*]^T + E[\tilde{\Theta}\tilde{\Theta}^{*T}] + E[\tilde{\Theta}^*\tilde{\Theta}^T] - E[\tilde{\Theta}^*\tilde{\Theta}^{*T}] \tag{4.3.7}$$

Avšak již víme, že

$$E[\tilde{\Theta}^*\tilde{\Theta}^{*T}] = (\Phi^T R^{-1}\Phi)^{-1}$$

Snadno dostaneme

$$\begin{aligned} E[\tilde{\Theta}\tilde{\Theta}^{*T}] &= E[Zee^T(Z^*)^T] = ZR(Z^*)^T = ZRR^{-1}\Phi(\Phi^T R^{-1}\Phi)^{-1} \\ &= Z\Phi(\Phi^T R^{-1}\Phi)^{-1} = (\Phi^T R^{-1}\Phi)^{-1} = E\tilde{\Theta}^*\tilde{\Theta}^{*T} \\ &= \left(E[\tilde{\Theta}^*\tilde{\Theta}^T]\right)^T \end{aligned}$$

Při výpočtech jsme použili (4.3.4). Nyní (4.3.7) dává

$$cov_Z[\hat{\Theta}] = E[\tilde{\Theta} - \tilde{\Theta}^*][\tilde{\Theta} - \tilde{\Theta}^*]^T + (\Phi^T R^{-1}\Phi)^{-1} \geq (\Phi^T R^{-1}\Phi)^{-1}$$

což dokazuje (4.3.3).
Poznámka . Předpokládejme $R = \lambda^2 I$. Pak dostaneme $Z^* = (\Phi^T\Phi)^{-1}\Phi^T$ Odhad parametrů pak přejde na odhad ve smyslu nejmenších čtverců (4.1.7).

Poznámka . Odhad (4.3.2) je v literatuře mnohdy označován jako odhad ve smyslu vážených nejmenších čtverců.

Poznámka . Může existovat nelineární odhad s lepší přesností? Lze ukázat, že pokud $e(t)$ je gaussovské, pak lineární odhad (4.3.1) s (4.3.2) je nejlepší mezi všemi nelineárními nestrannými odhady.

V následujícím příkladu budeme ilustrovat BLUE (4.3.1), (4.3.2).

Příklad 4.3.1. Nechť model je dán

$$y(t) = b_0 + e(t)$$

kde chyby měření jsou nezávislé, ale mají *odlišné* variance, takže

$$Ee^2(t) = \lambda_t$$

Tudíž máme

$$\Phi = \begin{bmatrix} 1 \\ . \\ . \\ . \\ 1 \end{bmatrix} \qquad R = \begin{bmatrix} \lambda_1^2 & & & & 0 \\ & \lambda_2^2 & & & \\ & & . & & \\ & & & . & \\ & & & & . \\ 0 & & & & \lambda_N^2 \end{bmatrix}$$

Pak odhad parametru b ve smyslu BLUE je

$$\hat{\Theta} = \hat{b} = \frac{1}{\sum_{j=1}^{N}(1/\lambda_j^2)} \sum_{i=1}^{N}(1/\lambda_i^2) y(i)$$

Jedná se o vážený aritmetický průměr jednotlivých pozorování. Poznamenejme, že váhy u $y(i)$

$$\frac{1}{\sum_{j=1}^{N}(1/\lambda_j^2)} 1/\lambda_i^2$$

jsou malé, jestliže měření je nepřesné (λ_i je velké) a naopak.

4.4 Výpočetní detaily

V této podkapitole se budeme zabývat numerickými aspekty výpočtu odhadu ve smyslu nejmenších čtverců (4.1.7) na základě soustavy rovnic (4.1.9). Povšimneme si následujících přístupů:

- řešení normálních rovnic,
- ortogonální triangularizace,
- rekurzivní algoritmy.

První přístup je stanovit matice $\Phi^T\Phi$ a $\Phi^T Y$ a pak řešit soustavu normálních rovnic

$$(\Phi^T\Phi)\Theta = \Phi^T Y \tag{4.4.1}$$

To je samozřejmě jednoduchý, přímočarý přístup, ale je citlivý na numerické chyby při zaokrouhlování. Pro ilustraci předpokládejme, že

$$\Phi = \begin{bmatrix} 1 & 1-\epsilon \\ 1 & 1+\epsilon \end{bmatrix} \quad Y = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \tag{4.4.2}$$

kde ϵ je malé číslo. Pak normální rovnice (4.4.1) budou

$$\begin{bmatrix} 2 & 2 \\ 2 & 2+2\epsilon^2 \end{bmatrix} \Theta = \begin{bmatrix} 6 \\ 6+2\epsilon \end{bmatrix} \tag{4.4.3}$$

a exaktní řešení je

$$\Theta = \begin{bmatrix} 3 - 1/\epsilon \\ 1/\epsilon \end{bmatrix}$$

Dále pak předpokládejme, že přesnost numerických výpočtů je η (to jest nemůžeme rozlišit mezi 1 a $1+\eta$. Nechť $\epsilon^2 < \eta,\ \epsilon > \eta$. Pak numerické chyby způsobí, že normální rovnice

$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \Theta = \begin{bmatrix} 6 \\ 6+2\epsilon \end{bmatrix}$$

nemají žádné řešení! Mohou být konstruovány i jiné příklady na ukázku, jak zaokrouhlovací chyby drasticky ovlivňují řešení. V takových případech má matice Φ sníženou hodnost, to jest sloupce jsou většinou lineárně závislé.

Druhý přístup, ortogonální triangularizace, je také známý jako QR metoda. Idea je následující. Pomocí QR metody spočítáme rozklad matice Φ na matici ortogonální Q a matici čtvercovou trojúhelníkovou R splňující následující rovnost

$$\Phi = Q^T \begin{bmatrix} R \\ 0 \end{bmatrix} \tag{4.4.4}$$

Poznamenejme, že pro ortogonální matici Q platí $Q^{-1} = Q$, $Q^T Q = I$ a tedy i $QQ^T = I$. Pak, namísto původní přeurčené soustavy lineárních rovnic

$$\Phi\Theta = Y \tag{4.4.5}$$

po vynásobení zleva ortogonální maticí Q dostaneme

$$Q\Phi\Theta = QY \tag{4.4.6}$$

To neovlivní ztrátovou funkci (4.1.6), protože

$$\| QY - Q\Phi\Theta \|^2 = \| Q(Y-\Phi\Theta) \|^2 = (Y-\Phi\Theta)^T Q^T Q(Y-\Phi\Theta) = (Y-\Phi\Theta)^T (Y-\Phi\Theta) = \| Y-\Phi\Theta \|^2$$

Členy na pravé a levé stranně rovnice (4.4.5) pak můžeme napsat ve formě

$$Q\Phi = \begin{bmatrix} R \\ 0 \end{bmatrix} \qquad QY = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \tag{4.4.7}$$

což následně vede k následujícímu tvaru ztrátové funkce

$$V(\Theta) = \| Q\Phi\Theta - QY \|^2 = \| R\Theta - z_1 \|^2 + \| z_2 \|^2 \tag{4.4.8}$$

která je minimalizována, když

$$R\Theta = z_1 \tag{4.4.9}$$

Lze vidět, minimální hodnota, kterou může kriteriální funkce (4.4.7) nabýt je

$$\min_{\Theta} V(\Theta) = \| z_2 \|^2 = z_2^T z_2 \tag{4.4.10}$$

Za předpokladu, že R je regulární (což je ekvivalentní, že Φ má plnou hodnost a také $\Phi^T\Phi$ je regulární) lze snadno řešit lineární soustavu (4.4.8) zpětným chodem substituční metody, protože R je *trojúhelníková* matice.

Metoda QR vyžaduje přibližně dvanáctkrát více operací než přímé řešení normálních rovnic. Její výhodou je, že je mnohem méně citlivá na chyby zaokrouhlení. Předpokládejme, že relativní

chyby v datech jsou velikosti ϵ a že přesnost operace je η. Pak, abychom se vyhnuli nerozumným chybám ve výsledku, měli bychom požadovat, aby $\eta < \epsilon^2$ pro normální rovnice, zatímco pro QR metodu je postačující $\eta < \epsilon$.

Třetí přístup je použít rekurzivní algoritmy. Matematický popis a analýza těchto algoritmů jsou uvedeny v kapitole 8. Idea je přepsat odhad (4.1.7) do formy

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + K(t)[y(t) - \varphi^T(t)\hat{\Theta}(t-1)] \tag{4.4.11}$$

kde $\hat{\Theta}(t)$ označuje odhad založený na t rovnicích (t-tý řádek v Φ). Výraz $y(t) - \varphi^T(t)\hat{\Theta}(t-1)$ může být chápán jako chyba predikce. Ukazuje, jak dobře může být predikováno skutečné měření $y(t)$ vektorem parametrů $\hat{\Theta}(t-1)$, který je získán z předchozích dat. Pro vektor zisku $K(t)$ je typické, že s rostoucím t klesá jako $1/t$. Rozdíl mezi jednotlivými rekurzivními algoritmy je právě v tom, jak vybírají $K(t)$.

4.5 Metoda nejmenších čtverců s lineárním omezením

Až doposud jsme sa zabývali metodou nejmenších čtverců bez ohledu na význam či interpretaci odhadovaných parametrů formujících vektor Θ. Předpokládali jsme, že parametry mohou nabýt libovolných hodnot v oboru reálných čísel. V mnoha situacích však tento předpoklad není platný a odhadované parametry mohou ležet jen v jisté podmnožině reálných čísel. Jako příklad uveďme dvě následující situace:

a) Předpokládejme určování polohy (longitudinální, laterální i výškové) objektu na základě satelitních měření systému GPS. Typicky je tato úloha řešena metodou nejmenších čtverců, kdy hledaná poloha objektu je součástí vektoru neznámých parametrů a vektor regresorů je dán známou polohou satelitů. Takovéto řešení je plně dostačující např. v situaci, kdy určujeme polohu letadla, které se v principu může pohybovat kdekoliv. Pokud bychom ale uvažovali úlohu hledání pozice vlaku, situace je už jiná; vlak nemůže být kdekoliv, ale pouze na kolejích. Z důvodu zašuměných satelitních měření lze jen stěží předpokládat, že odhadnutá pozice vlaku pomocí doposud uvažované metody nejmenších čtverců bude odpovídat umístění kolejí a velmi pravděpodobně bude odhad pouze v jejich okolí.

   Jedno z možných řešení tohoto problému je použít metodu nejmenších čtverců s rovnostním omezením. Rovnostní omezení, definující podmnožinu reálných čísel, může být v tomto případě chápáno jako nějaká křivka v horizontální rovině popsaná vztahem

$$g(\Theta) = 0$$

   dávající do vztahu longitudinální a laterální pozici kolejí, resp. vlaku.

b) Předpokládejme úlohu, kdy cílem je odhad koncentrace chemických látek v roztoku na základě měření elektrických vlastností roztoku. Koncentrace jako taková může nabývat jen kladných hodnot. Avšak z důvodu nepřesných měření může odhad koncentrace látky vyjít záporný, což nemůže odpovídat realitě.

   Možné řešení této úlohy spočívá v použití metody nejmenších čtverců s explicitním uvažováním nerovnostního omezení, které je, v tomto případě, dáno nerovností

$$\Theta \geq 0$$

Podobných situací, kdy hledáme odhad na jisté podmnožině reálných čísel, bychom v aplikacích našli bezpočet. Proto byla značná pozornost věnována odhadu při explicitním uvažování jistého omezení hledaných parametrů. V následující částech je proto formulována úloha odhadu stavu s rovnostním a nerovnostním lineárním omezením a ukázáno její řešení.

4.5.1 Rovnostní omezení

Uvažujme soustavu N lineárních rovnic o n neznámých zapsaných v již dříve uvažovaném maticovém zápisu

$$Y = \Phi\Theta + E \tag{4.5.12}$$

kde Φ je matice s plnou sloupcovou hodností. Mějme také danou množinu m lineárních omezení hledaných parametrů dané lineární maticovou rovnicí

$$D\Theta = d \tag{4.5.13}$$

kde $m < n$, D je známá matice o rozměrech m/n s plnou řádkovou hodností a d je známý vektor o m prvcích. Cílem je nalézt odhad vektoru parametrů $\hat{\Theta}$ ve smyslu nejmenších čtverců s ohledem na omezení (4.5.13), tj. hledáme následující odhad

$$\hat{\Theta} = \arg\min_{\Theta} \tfrac{1}{2}[Y - \Phi\Theta]^T[Y - \Phi\Theta] \text{ s.t. } D\Theta = d \tag{4.5.14}$$

kde zkratka s.t. značí „s ohledem na“ (z anglického „subject to“). V literatuře existuje několik postupů řešení optimalizačního problému (4.5.14), avšak zde bude pozornost věnována třem významným zástupcům, a to (i) reformulace a algebraické řešení, (ii) aproximační řešení metodou vážených nejmenších čtverců a (iii) minimalizace použitím Lagrangeových multiplikátorů.

Reformulace a algebraické řešení vztahu (4.5.14) je založena na přímočaré myšlence redukce počtu lineárních rovnic (4.5.12) pomocí rovnic daných omezení (4.5.13) a následném řešení soustavy N o $n - m$ neznámých standardní metodou nejmenších čtverců. Nejlépe lze tento přístup ilustrovat následujícím jednoduchým příkladem, kdy uvažujeme dvě rovnice o dvou neznámých $\Theta = [\Theta_1, \Theta_2]^T$

$$y(1) = [1,\ 2]\Theta + e(1) \tag{4.5.15}$$
$$y(2) = [2,\ 3]\Theta + e(2) \tag{4.5.16}$$

a ohledem na omezení

$$\Theta_1 + \Theta_2 = 3 \tag{4.5.17}$$

Pak z rovnice omezení (4.5.17) lze vyjádřit např. $\Theta_1 = 3 - \Theta_2$ a dosadit do (4.5.15), což vede na soustavu dvou rovnic o jedné neznámé přímo řešitelnou dříve představenou metodou nejmenších čtverců. Ačkoliv se jedná o přímočaré řešení, pro úlohy větších rozměrů je nevhodné.

Aproximační řešení metodou vážených nejmenších čtverců je založeno na rozšíření soustavy rovnic (4.5.12) o rovnice dané omezením (4.5.13), tj. na řešení následující rozšířené soustavy $N + m$ rovnic

$$\underbrace{\begin{bmatrix} Y \\ d \end{bmatrix}}_{\tilde{Y}} = \underbrace{\begin{bmatrix} \Phi \\ D \end{bmatrix}}_{\tilde{\Phi}} \Theta + \underbrace{\begin{bmatrix} E \\ 0 \end{bmatrix}}_{\tilde{E}} \tag{4.5.18}$$

Takto definovaná soustava rovnic přirozeně vede na metodu vážených nejmenších čtverců poskytujících odhad

$$\hat{\Theta} = (\tilde{\Phi}^T W \tilde{\Phi})^{-1} \tilde{\Phi}^T W \tilde{Y} \tag{4.5.19}$$

Váhová matice W je uživatelem volená jako diagonální matice, kde prvních N diagonálních prvků je zvoleno jako "malá" čísla a posledních m prvků jako "velká" čísla. Motivace pro tuto volbu vychází z faktu, že rovnostní omezení lze chápat jako naprosto přesná měření, jimž při řešení věříme nejvíce.
Ačkoliv se opět jedná o jednoduché řešení za pomoci již představeného standardního algoritmu, má několik nevýhod. První spočívá ve vhodné volbě váhové matice. Pokud jsou diagonální prvky zvoleny příliš odlišně, bude řešení rovnice (4.5.19) špatně podmíněné a může vést k numerickým problémům. Z podstaty věci, také nelze očekávat, že výsledný odhad bude plně vyhovovat omezení. Jak daleko bude odhad od splnění omezení závisí na konkrétní volbě váhové matice.

Minimalizace použitím Lagrangeových multiplikátorů vychází z teorie hledání optima při daném rovnostním omezení, kde je kriteriální funkce nejmenších čtverců (4.1.6) rozšířena o rovnostní omezení (4.5.13). Rozšířená kriteriální funkce pro minimalizaci je dána

$$L(\Theta, \lambda) = \tfrac{1}{2}[Y - \Phi\Theta]^T[Y - \Phi\Theta] + \lambda^T(D\Theta - d) \tag{4.5.20}$$

kde λ je m dimenzionální vektor neznámých Lagrangeových multiplikátorů. Minimum (4.5.20) s ohledem na parametry Θ a multiplikátory λ vychází z parciálních derivací kritéria (4.5.20)

$$\frac{\partial L(\Theta, \lambda)}{\partial \Theta} = -(Y - \Phi\Theta)^T\Phi + \lambda^T D \stackrel{!}{=} 0 \tag{4.5.21}$$

$$\frac{\partial L(\Theta, \lambda)}{\partial \lambda} = D\Theta - d \stackrel{!}{=} 0 \tag{4.5.22}$$

které mohou být zapsány ve formě soustavy lineárních rovnic

$$\begin{bmatrix} \Phi^T\Phi & D^T \\ D & 0 \end{bmatrix} \begin{bmatrix} \Theta \\ \lambda \end{bmatrix} = \begin{bmatrix} \Phi^T Y \\ d \end{bmatrix} \tag{4.5.23}$$

Z rovnice (4.5.21), s použitím substituce $P = (\Phi^T\Phi)^{-1}$, plyne

$$\hat{\Theta} = P(\Phi^T Y - D^T\lambda) \tag{4.5.24}$$

Dosazením (4.5.24) do (4.5.22) vede na vztah pro výpočet multiplikátorů

$$\lambda = (DPD^T)^{-1}(DP\Phi^T Y - d) \tag{4.5.25}$$

jehož dosazením do (4.5.24) získáme finální vztah pro odhad parametrů ve smyslu nejmenších čtverců při respektování omezení

$$\hat{\Theta} = \left(P - PD^T(DPD^T)^{-1}DP\right)\Phi^T Y + PD^T(DPD^T)^{-1}d \tag{4.5.26}$$

Řešení uvažovaného problému pomocí Lagrangeových multiplikátorů sice vede na komplikovanější vztah, avšak pro jeho použití není nutná specifikace jakéhokoliv parametru uživatelem a výsledný odhad zaručeně splňuje omezení. Pokud by bylo zapotřebí, lze k tomuto odhadu dopočítat i kovarianční matici chyby odhadu, jak je ukázáno např. v [62].

4.5.2 Nerovnostní omezení

V případě nerovnostního omezení opět uvažujeme soustavu lineárních rovnic (4.5.12), avšak omezení je dáno soustavou nerovnic zapsaných v maticové formě jako

$$d_L \leq D\Theta \leq d_U \tag{4.5.27}$$

kde D je známá matice o rozměrech m/n a d_L, d_U jsou známé vektory o m prvcích. Cílem je nalézt odhad vektoru parametrů $\hat{\Theta}$ ve smyslu nejmenších čtverců s ohledem na omezení (4.5.27), tj. hledáme následující odhad

$$\hat{\Theta} = \arg\min_{\Theta} \tfrac{1}{2}[Y - \Phi\Theta]^T[Y - \Phi\Theta] \text{ s.t. } d_L \leq D\Theta \leq d_U \tag{4.5.28}$$

Tato formulace přímo vede na úlohu kvadratického programování, jako podskupinu metod programování nelineárního. Bližší představení metod kvadratického programování by bylo nad rámec těchto skript i předmětu, proto se omezíme jen na konstatování, že úlohu kvadratického programování lze řešit mnoha softwarovými nástroji, mezi kterými můžeme zmínit program MATLAB® a jeho funkci `lsqlin`. Detailní informace o metodách kvadratického programování v rámci metody nejmenších čtverců lze najít např. v [63].

4.6 Shrnutí

Proveďme shrnutí této kapitoly. Definovali jsme lineární regresní modely a odvodili odhad neznámých parametrů ve smyslu nejmenších čtverců. Pak jsme se zabývali statistickými vlastnostmi. Byl využit významný předpoklad, že regresní proměnné jsou deterministické a známé funkce. Poté jsme rozšířili analýzu na obecné lineární nestranné odhady. Konkrétně jsme odvodili odhady parametrů mající nejmenší kovarianční matici a odhady s ohledem na lineární omezení.

Lineární regrese je velmi prozkoumaná oblast. Rozsáhlá a kvalitní publikace v češtině je [32] a v angličtině např. [62]. Z oblasti ekonometrické literatury či časových řad např. uveďme [10], [20], [33] - [36].

Kapitola 5

Parametrizace modelů

5.1 Klasifikace modelů

V této kapitole se budeme zabývat významem a rolí struktury modelu M. Zavedeme obecný popis lineárních modelů a výklad podpoříme řadou příkladů.

Nejprve se však budeme věnovat obecným poznámkám týkajících se klasifikace modelů. Poznamenejme, že tohoto tématu jsme se dotkli již v 1. kapitole. Klasifikace může být provedena podle různých kritérií. Můžeme rozlišovat:

- „Intuitivní“ či „mentální“ modely (např. pokud nejsou mraky na obloze, nebude pršet),
- Slovně popsané modely (např. různé pranostiky; poznamenejme, že slovo pranostika pochází z řeckého slova „prognósis“, tedy předpověď),
- Modely ve formě tabulek či grafů (např. impulsní charakteristika),
- Matematické modely (např. algebraická, diferenční, či diferenciální rovnice),
- Fyzikální modely (např. funkční model technologického procesu, který ma stejné významné charakteristiky jako zkoumaný reálný objekt).

Náplní těchto skript je tvorba *matematických* modelů. Zájem o matematické modely je ve vědních disciplínách motivován různými důvody, např. potřebou získat popis zkoumaného fyzikálního jevu nebo procesu pro predikci vývoje či získáním prostředku pro návrh regulátoru, filtru nebo detektoru poruch v systému. Matematické modely mohou být odvozeny dvěma způsoby (jak již bylo řečeno v 1. kapitole) a to:

- Matematickým modelováním, které se odvolává na základní zákony fyziky, matematiky, chemie, biologie, ekonomie atd. Často využívá základních rovnic rovnováhy jako „čistá akumulace = vstupní tok - výstupní tok“, což může být aplikováno na různé proměnné jako energie, hmota, peníze atd.
- Identifikací, která se zabývá navrhováním (dynamických) modelů z experimentálních dat mnohdy bez znalosti vnitřní principů zkoumaného systému, např. z důvodu jeho složitosti. Jako příklad může být uveden vývoj cen akcií, který je ovlivněn nejen racionálním, ale i velmi nepředvídatelným lidským chováním znemožňujícím použít matematické modelování. Proces identifikace v sobě zahrnuje získání dat, určení vhodné formy modelu,

návrh identifikačního experimentu, určení parametrů modelu a ověření tohoto modelu.

Matematické modely dynamických systémů mohou být klasifikovány opět z různých hledisek. V následující části ukážeme řadu alternativních modelů, které jsou nejčastěji používány.

- Modely s jedním vstupem a s jedním výstupem (SISO) - modely s více vstupy více výstupy (MIMO)
  SISO modely se používají k popisu procesů, kde existuje vliv jednoho vstupu (v angličtině je používán termín „single input“) na jeden výstup (v angličtině „single output“). Jestliže je použito více proměnných, hovoříme o modelech s více vstupy (v angličtině „multi input“) a více výstupy (v angličtině „multi output“) MIMO. Většina teoretických výsledků platí pro MIMO modely, ačkoliv pro ilustraci často používáme modely SISO. Poznamenejme, že modely s více vstupy a jedním výstupem (MISO) jsou ve většině případů snadno odvoditelné obdobně jako modely SISO, zatímco modely s dvěma nebo více výstupy přinášejí již strukturální problémy.

- Lineární modely - nelineární modely. Model je lineární, jestliže je např. popsán lineární diferenční rovnicí. Až na některé výjimky se budeme zabývat v 1. díle skript pouze lineárními modely, zatímco ve 2. díle i nelineárními.

- Parametrické modely - neparametrické modely. Parametrický model může být popsán množinou parametrů. Jednoduché parametrické modely jsme použili již v kapitole 2. a 4. a na tyto modely se soustředíme i v následujících kapitolách. Ve 3. kapitole byly uvedeny příklady neparametrických modelů, které jsou reprezentovány grafem.

- Časově invariantní modely - časově variantní modely. Časově invariantní modely jsou určitě nejběžnější. Pro časově variantní modely je potřeba použít speciální identifikační metody. V případech, kdy parametry modelu se s časem mění, mluvíme o sledování parametrů nebo identifikaci v reálném čase nebo též o filtraci, jak bude vidět ve 2. díle skript.

- Modely v časové oblasti - modely ve frekvenční oblasti. Typickými příklady modelů v časové oblasti jsou diferenciální a diferenční rovnice, zatímco spektrální a logaritmická frekvenční charakteristika jsou příklady modelů ve frekvenční oblasti. Převážná část skript se zabývá modely v časové oblasti.

- Modely diskrétní v čase - modely spojité v čase. Model diskrétní v čase popisuje vztah mezi vstupy a výstupy pouze v určitých časových okamžicích. Budeme předpokládat ekvidistantní dobu mezi těmito okamžiky a tuto dobu uvažujme jako časovou jednotku. Tudíž pro modely diskrétní v čase, které mají dominantní postavení v těchto skriptech, se užívá čas t=1,2,... nebo k=1,2.... Poznamenejme, že modely spojité v čase ve formě diferenciálních rovnic, mohou velmi dobře vyhovovat datům diskrétním v čase.

- Modely se soustředěnými parametry - modely s rozloženými parametry. Modely se soustředěnými parametry jsou popsány nebo založeny na konečném počtu obyčejných diferenciálních nebo diferenčních rovnic. Jestliže tento počet rovnic je nekonečný nebo model je založen na parciálních diferenciálních rovnicích, pak mluvíme o modelu s rozloženými parametry. V těchto skriptech se omezíme na modely se soustředěnými parametry.

- Deterministické modely - stochastické modely. U deterministického modelu výstup může být zhruba řečeno přesně vypočten, jakmile vstupní signál je známý. Naopak stochastický model obsahuje náhodné složky popisující poruchy, které znemožňují takový přesný výpočet. Ve skriptech se používají převážně stochastické modely.

Toto členění jistě není konečné. Mohli bychom pokračovat v klasifikaci modelů např. na fenomenologické, tj. vstupně-výstupní, modely, které jsou používány převážně v tomto dílu a stavové modely, které budou převážně využívány ve 2. dílu těchto skript.

5.2 Struktura modelu

Budeme používat následující obecnou formu struktury modelu

$$M(\Theta) : y(t) = G(q^{-1};\Theta)u(t) + H(q^{-1};\Theta)e(t) \tag{5.2.1}$$

$$Ee(t)e^T(s) = \Lambda(\Theta)\delta_{t,s}$$

V (5.2.1) $y(t)$ je ny dimenzionální výstup v čase t a $u(t)$ je nu dimenzionální vstup. Dále $e(t)$ je bílý ne dimenzionální šum, tedy posloupnost nezávislých náhodných proměnných, zde navíc se stejným rozložením a s nulovou střední hodnotou. V (5.2.1) je $G(q^{-1};\Theta)$ ny/nu dimenzionální a $H(q^{-1};\Theta)$ ny/ne dimenzionální filtr. Argument q^{-1} označuje operátor zpětného posuvu, to jest $q^{-1}u(t) = u(t-1)$ atd. a $\delta_{t,s} = 0$ pro $t \neq s$ a $\delta_{t,s} = 1$ pro $t = s$. Filtry $G(q^{-1};\Theta)$ a $H(q^{-1};\Theta)$, stejně jako kovarianční matice šumu $\Lambda(\Theta)$, jsou funkce vektoru parametrů Θ, který předpokládáme $n\Theta$ dimenzionální. Často musíme omezit Θ, aby leželo v podmnožině $R^{n\Theta}$. Tato podmnožina je dána

$$\begin{aligned} D \;=\; & \{\Theta \mid H^{-1}(q^{-1};\Theta) \;\text{ a }\; H^{-1}(q^{-1};\Theta)G(q^{-1};\Theta) \\ & \text{jsou asymptoticky stabilní,} \;\; \Lambda(\Theta) \;\text{ je pozitivně semi-definitní}\} \end{aligned} \tag{5.2.2}$$

Důvody pro zavedení těchto omezení v definici podmnožiny D budou zřejmé v následující kapitole, kde bude ukázáno, že pro $\Theta \in D$ lze najít optimální prediktor $y(t)$ za podmínky známých dat $\{y(t-1), u(t-1), y(t-2), u(t-2), \ldots, y(1), u(1)\}$. Poznamenejme, že $H(q^{-1};\Theta)$ nemusí být vždy uvažováno jako asymptoticky stabilní. Modely s nestabilním $H(q^{-1};\Theta)$ mohou být užitečné pro popis např. driftu. Stacionární poruchy lze modelovat fakticky bez omezení modelem (5.2.2), protože víme, že podle věty o spektrální faktorizaci, každý stacionární proces může být chápán jako výstup systému, který je reprezentován přenosovým operátorem $H(q^{-1};\theta)$, na jehož vstupu je bílý šum, přičemž $H(q^{-1};\Theta)$ lze vždy najít tak, aby splňoval (5.2.2).

V následující části představíme na příkladech celou řadu typických a často používaných struktur modelu, které jsou speciálními příklady obecného lineárního modelu (5.2.1), a zároveň popíšeme, jak $G(q^{-1};\Theta), H(q^{-1};\Theta)$ a $\Lambda(\Theta)$ závisí na Θ.

Příklad 5.2.1 (ARMAX model)
Nechť $y(t)$ a $u(t)$ jsou skalární signály a uvažujme strukturu modelu

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \tag{5.2.3}$$

kde

$$\begin{aligned} A(q^{-1}) &= 1 + a_1q^{-1} + ... + a_{na}q^{-na} \\ B(q^{-1}) &= b_1q^{-1} + ... + b_{nb}q^{-nb} \\ C(q^{-1}) &= 1 + c_1q^{-1} + ... + c_{nc}q^{-nc} \end{aligned} \tag{5.2.4}$$

Vektor parametrů je dán

$$\Theta = [a_1, \ldots, a_{na}, b_1, \ldots, b_{nb}, c_1, \ldots, c_{nc}]^T \tag{5.2.5}$$

Model (5.2.3) může být explicitně zapsán jako diferenční rovnice

$$y(t)+a_1y(t-1)+...+a_{na}y(t-na) = b_1u(t-1)+...+b_{nb}u(t-nb)+e(t)+c_1e(t-1)+...+c_{nc}e(t-nc) \tag{5.2.6}$$

avšak vyjádření (5.2.3) využívající polynomiální formalismus je pro teoretické použití vhodnější.
Je možné rozšířit vektor parametrů o varianci šumu

$$\lambda^2 = Ee^2(t) \tag{5.2.7}$$

pak

$$\Theta_e = [\Theta^T, \lambda^2]^T$$

je nový vektor parametrů dimenze $na + nb + nc + 1$.
Poznamenejme, že (5.2.3) může být snadno přepsáno ve smyslu relace (5.2.1), to jest

$$y(t) = \frac{B(q^{-1})}{A(q^{-1})}u(t) + \frac{C(q^{-1})}{A(q^{-1})}e(t)$$

Pak pro strukturu modelu (5.2.3) dostaneme

$$\begin{aligned} G(q^{-1};\Theta) &= \frac{B(q^{-1})}{A(q^{-1})} \\ H(q^{-1};\Theta) &= \frac{C(q^{-1})}{A(q^{-1})} \end{aligned} \tag{5.2.8}$$

$$\Lambda(\Theta) = \lambda^2$$

Množina D je v tomto případě zřejmě:

$$D = \{\Theta \mid \text{ Polynom } C(q^{-1}) \text{ má všechny nuly vně jednotkového kruhu}\}.$$

Běžnější formulace požadavků na $\Theta \in D$ je, že reciproký polynom

$$C^*(q) = q^{nc} + c_1q^{nc-1} + ... + c_{nc} = q^{nc}C(q^{-1})$$

má všechny kořeny uvnitř jednotkového kruhu.
Poznámka . q má pro nás zejména význam operátoru časového posunu, tj. $q^{-1}y(t) = y(t-1)$ a $qy(t) = y(t+1)$. Není nutné jej chápat nezbytně jako operátor z transformace.

Všimněme si několika důležitých speciálních případů (5.2.3).

- Pro $nb = nc = 0$ dostaneme autoregresní model AR (z anglického „autoregressive“). (Pak je modelována čistá časová posloupnost, časová řada, kde žádný vstupní signál není uvažován). V tomto případě dostaneme

$$\begin{aligned} A(q^{-1})y(t) &= e(t) \\ \Theta &= [a_1, \ldots, a_{na}]^T \end{aligned} \tag{5.2.9}$$

- Pro $na = nb = 0$ dostaneme klouzavý průměr MA (z anglického „moving average“). Pak máme

$$y(t) = C(q^{-1})e(t) \tag{5.2.10}$$
$$\Theta = [c_1, \ldots, c_{nc}]^T$$

- Pro $nb = 0$ dostaneme ARMA model

$$A(q^{-1})y(t) = C(q^{-1})e(t) \tag{5.2.11}$$
$$\Theta = [a_1, \ldots, a_{na}, c_1, \ldots, c_{nc}]^T$$

- Další speciální případ je, když $nc = 0$. Struktura modelu pak přejde

$$A(q^{-1})y(t) = B(q^{-1})u(t) + e(t) \tag{5.2.12}$$
$$\Theta = [a_1, \ldots, a_{na}, b_1, \ldots, b_{nb}]^T$$

Tento model je někdy nazýván ARX (X je z anglického „exogenous“ a souvisí s používáním exogenních proměnných v ekonometrii, řízený autoregresní model). Tato struktura může být v jistém smyslu chápána jako lineární regrese. Model (5.2.12) lze vskutku ekvivalentně zapsat ve formě

$$y(t) = \varphi^T(t)\Theta + e(t) \tag{5.2.13}$$

kde

$$\varphi(t) = [-y(t-1), \ldots, -y(t-na), u(t-1), \ldots, u(t-nb)]^T$$

Nicméně poznamenejme, že regresory (prvky $\varphi(t)$) zde nejsou deterministické funkce. To znamená, že analýza provedená ve 4. kapitole pro lineární regresní modely, nemůže být aplikována na případ (5.2.13). Důvodem je, že Φ není deterministická matice a nelze tedy obecně v důkazu věty 4.2.1 vytknout $\Phi^\dagger$ ze členu $E[\Phi^\dagger e]$, jak bylo ilustrováno i v podkapitole 2.5.

Příklad 5.2.2 (Obecná struktura modelu SISO)
Struktura ARMAX (5.2.3) je do určité míry obecná. Každý lineární systém konečného řádu se stacionární poruchou mající racionální spektrální hustotu může být popsán ARMAX modelem. (Samozřejmě na, nb, nc i Θ a λ nabývají určitých hodnot, které jsou různé případ od případu.) Nicméně, pro parametrizaci lineárních systémů konečného řádu existují i obecnější cesty. Zaveďme následující strukturu

$$A(q^{-1})y(t) = \frac{B(q^{-1})}{F(q^{-1})}u(t) + \frac{C(q^{-1})}{D(q^{-1})}e(t), \quad E(e^2(t)) = \lambda^2 \tag{5.2.14}$$

kde $A(q^{-1}), B(q^{-1}), C(q^{-1})$ jsou jako v (5.2.4) a

$$\begin{aligned} D(q^{-1}) &= 1 + d_1 q^{-1} + \ldots + d_{nd} q^{-nd} \\ F(q^{-1}) &= 1 + f_1 q^{-1} + \ldots + f_{nf} q^{-nf} \end{aligned} \tag{5.2.15}$$

Vektor parametrů je v tomto případě

$$\Theta = [a_1, \ldots, a_{na}, b_1, \ldots, b_{nb}, c_1, \ldots, c_{nc}, d_1, \ldots, d_{nd}, f_1, \ldots, f_{nf}]^T \tag{5.2.16}$$

Faktem je, že zřídka v praxi existuje důvod pro použití této struktury modelu v jeho obecné podobě. Při praktickém použití je jeden nebo více polynomů rovno jedné. Např. ARMAX model dostaneme pro $nd = nf = 0$, to jest $D(q^{-1}) = F(q^{-1}) = 1$. Význam formy (5.2.14) spočívá v její obecnosti, protože zahrnuje řadu důležitých forem jako speciální případy. To znamená, že můžeme popisovat a analyzovat, v případě použití (5.2.14), několik případů najednou. Při použití (5.2.14) dostaneme

$$G(q^{-1};\Theta) = \frac{B(q^{-1})}{A(q^{-1})F(q^{-1})} \tag{5.2.17}$$

$$H(q^{-1};\Theta) = \frac{C(q^{-1})}{A(q^{-1})D(q^{-1})}$$

$$\Lambda(\Theta) = \lambda^2$$

Ukažme v následujícím některé speciální případy

- Jestliže $nd = nf = 0$ dostaneme ARMAX model

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \tag{5.2.18}$$

  a jeho různé varianty, které jsou diskutovány v příkladu 5.1.1.

- Jestliže $na = nc = nd = 0$ dostaneme

$$y(t) = \frac{B(q^{-1})}{F(q^{-1})}u(t) + e(t) \tag{5.2.19}$$

  V tomto případě $H(q^{-1};\Theta) = 1$. Můžeme také říci, že v (5.2.19) jsou popsány poruchy bez dynamiky a že vstupují přímo na výstup. Strukturu (5.2.19) budeme někdy nazývat model chyby výstupu, protože (5.2.19) implikuje, že

$$e(t) = y(t) - \frac{B(q^{-1})}{F(q^{-1})}u(t) \tag{5.2.20}$$

  je chyba výstupu, to jest rozdíl mezi měřitelným výstupem $y(t)$ a „deterministickou výstupní komponentou $B(q^{-1})/F(q^{-1})u(t)$“.

- Jestliže $na = 0$ dostaneme strukturu

$$y(t) = \frac{B(q^{-1})}{F(q^{-1})}u(t) + \frac{C(q^{-1})}{D(q^{-1})}e(t) \tag{5.2.21}$$

  Speciální vlastností této struktury je, že $G(q^{-1};\Theta)$ a $H(q^{-1};\Theta)$ nemají oproti (5.2.17) žádné společné parametry. V určitých případech, to může mít pozitivní vliv na výsledky identifikace v určitých případech.

Doposud jsme se zabývali pouze lineárními modely. V mnohých případech však může vzniknout nelineární dynamika. Někdy nelinearita vznikne pouze nelineární transformací signálů obsažených v modelu. Takové případy ovšem mohou být zahrnuty do lineárního rámce pouhým předefinováním veličin. Demonstrujme tuto situaci na Hammersteinově modelu.

Příklad 5.2.3 (Hammersteinův model)
Uvažujme skalární model

$$A(q^{-1})y(t) = B_1(q^{-1})u(t) + B_2(q^{-1})u^2(t) + ... + B_m(q^{-1})u^m(t) + e(t) \tag{5.2.22}$$

Nelinearity v tomto modelu vznikají jako mocniny vstupu $u(t)$. Definováním nového umělého vstupu

$$\bar{u}(t) = \begin{bmatrix} u(t) \\ u^2(t) \\ \vdots \\ u^m(t) \end{bmatrix}$$

a položením

$$\bar{B}(q^{-1}) = [B_1(q^{-1}), B_2(q^{-2}), \ldots, B_m(q^{-m})]$$

můžeme zapsat model (5.2.22) takto:

$$A(q^{-1})y(t) = \bar{B}(q^{-1})\bar{u}(t) + e(t) \tag{5.2.23}$$

Model (5.2.23) je již standardní lineární model, jestliže chápeme $\bar{u}(t)$ jako vstupní signál. Takže použitím $\bar{u}(t)$ jako vstupu jsme získali lineární model. To znamená, že pro odhad parametrů můžeme využít stejnou techniku jako pro lineární modely. Poznamenejme, že identifikaci nelineárních modelů je věnována pozornost v kapitole 9.

Poznámka . Jak již bylo zmíněno, operátor q^{-1} značí zpětný posuv v čase, tj. $q^{-1}u(t) = u(t-1)$. Pokud však provedeme substituci $q^{-1} = e^{-i\omega\tau}$, kde i značí imaginární proměnnou, ω frekvenci signálu a τ periodu vzorkování, pak lze převést model (5.2.14) v časové oblasti do oblasti frekvenční. Model ve frekvenční oblasti je vhodný např. pro výpočet ustáleného zesílení, které je dáno modelem při uvažování $\omega = 0$, tj. kdy $q^{-1} = e^0 = 1$. Modely ve frekvenční oblasti jsou krátce diskutovány v kapitole 3.

5.3 Jednoznačnost

Až doposud jsme se zabývali popisem obecné struktury modelu (5.2.1) pro lineární modely. Ukázali jsme také několik typických příkladů. Věnujme se nyní otázce, jak určit vhodnou strukturu modelu z daných dat. V této chvíli považujme za postačující uvést, že je více faktorů, které mají vliv na výběr struktury modelu. Uveďme čtyři nejdůležitější:

- Flexibilita. Struktura modelu by měla umožnit úplný popis dynamického chování zkoumaného systému ve všech jeho pracovních módech, které mohou být očekávány při běhu systému. Důležitý je jak počet volných parametrů, tak způsob jak jsou vloženy do modelu.
- Úspornost. Struktura modelu by měla být úsporná. To znamená, že model by měl obsahovat nejmenší počet volných parametrů, které jsou zapotřebí k adekvátní reprezentaci skutečného systému.
- Algoritmická složitost. Některé identifikační metody, jako např. metoda chyby predikce (z anglického názvu „prediction error method“, PEM) viz 6. kapitola, mohou být aplikovány pro různé struktury modelů. Avšak vybraná forma struktury může značně ovlivnit množství potřebných výpočetních operací.

– Vlastnosti kriteriální funkce. Asymptotické vlastnosti odhadů získaných metodou chyby predikce významně závisí na ztrátové funkci. Výskyt lokálních minim i neexistence jediného globálního minima jsou velmi závislé na použité struktuře modelu a i kriteriální funkci.

Měli bychom si všimnout jedné otázky, která se vztahuje k úspornosti a to, za jakých podmínek může být daný systém adekvátně a jednoznačně popsán v rámci určité struktury modelu. Abychom mohli tento problém formalizovat, musíme zavést na systém generující data $u(1), y(1), u(2), y(2), ...$ nějaké předpoklady. Ovšem takové předpoklady jsou potřebné pouze pro analýzu. Aplikace identifikační techniky není striktně závislá na platnosti takových předpokladů, protože je nutno vzít v úvahu řadu dalších okolností (inženýrský pohled).
Předpokládejme tedy, že skutečný systém je lineární, diskrétní a že je rušen signálem s racionální spektrální hustotou. Takže

$$y(t) = G_s(q^{-1})u(t) + H_s(q^{-1})e_s(t) \tag{5.3.1}$$
$$Ee_s(t)e_s^T(l) = \Lambda_s\delta_{t,l}$$

kde index s označuje „skutečný“ systém a $\delta_{t,l} = 0$ pro $t \neq l$ a $\delta_{t,l} = 1$ pro $t = l$. Zaveďme nyní množinu parametrů Θ

$$D_T(S, M) = \{\Theta \mid G_s(q^{-1}) \equiv G(q^{-1}; \Theta), H_s(q^{-1}) \equiv H(q^{-1}; \Theta), \quad \Lambda_s = \Lambda(\Theta)\} \tag{5.3.2}$$

Můžeme říci, že množina $D_T(S, M)$ se skládá z takových parametrů, pro které struktura modelu umožňuje dokonalý popis skutečného systému. Mohou vzniknout tři situace:

– Množina $D_T(S, M)$ je prázdná. To znamená, že nemůže být získán přesný popis systému v M, bez ohledu na výběr parametrů. Můžeme říci, že struktura modelu byla vybrána příliš „malá“ nebo má příliš málo parametrů k adekvátnímu popisu systému. Jako příklad zde můžeme uvést situaci z 2. kapitoly, kdy generátor dat byl systém S_2 ve struktuře ARMAX s parametry $\Theta = \{a, b, c\}$, avšak model použitý metodou nejmenších čtverců byl ve struktuře ARX odpovídající struktuře S_1, tj. pouze s parametry $\Theta = \{a, b\}$.

– Množina $D_T(S, M)$ obsahuje jeden bod. Označme jej Θ^*. To je samozřejmě ideální situace. Vektor Θ^* bude skutečný vektor parametrů. Jako příklad zde můžeme uvést opět situaci z 2. kapitoly, kdy jak generátor dat, tak i model použitý metodou nejmenších čtverců, byl ve struktuře ARX popisu S_1. Tedy, jak generátor, tak i model uvažovaly stejný vektor parametrů $\Theta = \{a, b\}$.

– Množina $D_T(S, M)$ obsahuje několik bodů. Neexistuje jediný model v rámci modelů dávajících přesný popis systému. Při hledání odhadů parametrů můžeme pak očekávat numerické problémy. Taková situace je někdy označována jako přeparametrizace. Pokud bychom se opět vrátili k modelům v kapitole 2, tato situace by nastala, pokud by generátor byl dán systémem S_1 s jednoduchou ARX strukturou, kdežto model pro metodu nejmenších čtverců by byl S_2, tj. měl by složitější strukturu s více parametry.

Poznamenejme, že v reálných aplikacích bez dostatečného fyzikálního náhledu týkající se chování zkoumaného systému, lze prostřední situaci považovat za velmi nepravděpodobnou.

5.4 Identifikovatelnost

V této části kapitoly budeme zkoumat pojem identifikovatelnost. Tento pojem může být zaveden různými způsoby, ale s ohledem na naše cíle je vhodný následující způsob.

Použijeme-li v parametrickém modelu strukturu M, identifikační metodu I a experimentální podmínky X, výsledný odhad budeme nazývat $\hat{\Theta}(N; S, M, I, X)$. Zřejmě odhad nezávisí pouze na I, M, X, ale také na množství dat N a skutečném systému S. Pak říkáme, že systém S je identifikovatelný systém za podmínek M, I a X, zkráceně $SI(M, I, X)$, jestliže

$$\hat{\Theta}(N; S, M, I, X) \to D_T(S, M), \quad N \to \infty \tag{5.4.1}$$

Pro $SI(M, I, X)$ je zejména požadováno, aby množina byla $D_T(S, M)$ neprázdná. Jestliže tato množina obsahuje více než jeden bod, pak (5.4.1) musí být interpretováno jako

$$\lim_{N\to\infty} \inf_{\Theta \in D_T(S,M)} \| \hat{\Theta}(N; S, M, I, X) - \Theta \| = 0 \tag{5.4.2}$$

Můžeme dále říkat, že systém S je parametricky identifikovatelný za podmínky M, I a X, zkráceně $PI(M, I, X)$, jestliže $SI(M, I, X)$ a $D_T(S, M)$ se skládají přesně z jednoho prvku. To je ideální případ. Jestliže systém je $PI(M, I, X)$, pak odhad parametrů bude jediný pro velké N a také konsistentní, to jest konverguje ke skutečné hodnotě, jak je dáno v definici $D_T(S, M)$.

5.5 Chyba modelu

Výše představené pojmy *jednoznačnost* a *identifikovatelnost* jsou velmi důležité z pohledu teoretické analýzy identifikačních metod. Z praktického pohledu, kdy neznáme skutečnou strukturu systému (tj. generátoru dat), jsou však obtížně vyhodnotitelné a namísto těchto pojmů se můžeme setkat s pojmem *chyba modelu* (v anglicky psané literatuře označované jako „model error“ [62], [71]). Chyba modelu je definována jako rozdíl mezi výstupem reálného systému a predikcí jeho výstupu určeného na základě modelu a minulých dat[1]. Chyba modelu vzniká ze dvou důvodů [71]:

1. Model není dostatečně *flexibilní* a neumožňuje dostatečně přesný popis chování systému. Jako příklad můžeme uvést situaci, kdy systém je druhého řádu, zatímco uvažovaný model je řádu prvního. Tato chyba je označována jako *bias modelu*.

2. Model není dostatečně *jednoduchý*, tj. řád modelu je nerealisticky vysoký, a je k dispozici málo trénovacích dat. Pak uvažovaná struktura modelu má mnoho stupňů volnosti a výsledný identifikovaný model je výrazně ovlivněn případnými odhlehlými měřeními (neboli tzv, outliery), na který se model identifikací adaptuje. Čím více je model adaptován na danou trénovací množinu dat, tím větší chybu modelu lze očekávat (jak při úloze interpolace, tak i extrapolace[2]). Tato chyba, která roste s rostoucím zvoleným řádem modelu, je označována jako *variance modelu*.

[1] Z hlediska terminologie zavedené v následujících kapitolách, lze chybu modelu chápat jako chybu predikce výstupu.

[2] Pod pojmem interpolace lze rozumět vyhodnocení modelu v bodě, který spadá do rozsahu trénovacích dat, zatímco pod pojmem extrapolace lze rozumět vyhodnocení modelu v bodě, který spadá mimo rozsah trénovacích dat.

Při výběru či volbě řádu (a struktury) modelu je tak nutné vzít v potaz jak vliv biasu, tak i variance modelu, a najít takový řád modelu minimalizující celkovou chybu. Volba vhodné struktury modelu je tak *rekurzivní* proces.

5.6 Shrnutí

Tato kapitola se zabývala různými aspekty výběru struktury modelu. Nejprve jsme popsali různé způsoby klasifikace struktur modelů. Pak jsme zavedli obecnou formu struktury modelu pro lineární systémy a omezení na parametry modelu. Na příkladech byla ukázána řada dobře známých struktur, které mohou být chápány jako speciální případ obecné formy (5.2.1). Konečně byly zavedeny některé pojmy týkající se identifikovatelnosti.

Ucelený výklad parametrizace je dán např. v [18]. Parametrizací se zabývají dále např. práce [20], [38], [40], [41]. Algoritmy spektrální faktorizace poskytuje [39].

Kapitola 6

Metoda chyby predikce

V této a příští kapitole se budeme zabývat identifikačními metodami, které jsou aplikovatelné pro širší třídu modelů než tomu bylo v případě metody nejmenších čtverců. Nyní se budeme věnovat tzv. metodám chyby predikce, v 7. kapitole metodám přídavné proměnné.

Metoda chyby predikce, v anglicky psané literatuře označované jako "prediction error method, PEM", může být v jistém smyslu chápána jako rozšíření základní myšlenky metody nejmenších čtverců pro složitější modely. Ve 2. kapitole bylo ilustrováno, že metoda nejmenších čtverců poskytuje asymptoticky nestranné odhady jen pro modely ve struktuře ARX. Pro složitější modely, např. ARMAX model obsahující korelovanou poruchu, jsou odhady poskytnuté metodou nejmenších čtverců stranné. Základní myšlenka v této kapitole představené metody chyby predikce spočívá v nalezení optimálního prediktoru pro uvažovaný (libovolně složitý) model a pak v následném odhadu parametrů prediktoru vedoucí na minimální hodnotu kvadrátu chyby predikce. Tedy metodu chyby predikce musíme chápat spíše jako koncept, zahrnující mimo jiné i metodu nejmenších čtverců jako speciální případ, než jako jeden konkrétní algoritmus.

6.1 Optimální predikce

Použití modelu získaného z identifikace je různé. Záleží na cíli, příčině modelování. Společná vlastnost mnohých aplikací je použití modelu pro predikci. Poznamenejme, že použití modelu pro predikci je často neoddělitelné od funkce modelu jako základu pro syntézu řídicího či estimačního systému. V takovém případě je významné vědět v čase t, jak bude pravděpodobně vypadat výstup v čase $t+1$, abychom mohli navrhnout vhodný řídicí zásah, to jest určit vstup $u(t)$.

Intuitivně má tedy smysl určit vektor parametrů Θ uvažovaného modelu tak, aby chyba predikce

$$\epsilon(t,\Theta) = y(t) - \hat{y}(t \mid t-1;\Theta) \tag{6.1.1}$$

byla malá.

V (6.1.1) $\hat{y}(t \mid t-1;\Theta)$ označuje optimální predikci $y(t)$ ve smyslu střední kvadratické chyby z dat až do času $t-1$ (to jest $y(t-1), u(t-1), y(t-2), u(t-2), \ldots, y(1), u(1)$) založenou na modelu s vektorem parametrů Θ. Otázka zní, jak najít postup výpočtu $\hat{y}(t \mid t-1;\Theta)$. Pro ilustraci uveďme nejdříve příklad. Pak se budeme zabývat optimální predikcí pro model s obecnou lineární strukturou.

Příklad 6.1.1 (Predikce pro ARMAX model)
Uvažujme tuto strukturu modelu

$$y(t) + ay(t-1) = bu(t-1) + e(t) + ce(t-1) \tag{6.1.2}$$

kde $\{e(t)\}$ je bílý šum s nulovou střední hodnotou a známou variancí λ^2. Vektor parametrů je

$$\Theta = [a, b, c]^T$$

a je taktéž znám. Výstup v čase t pak zřejmě splňuje

$$y(t) = [-ay(t-1) + bu(t-1) + ce(t-1)] + [e(t)] \tag{6.1.3}$$

Dva členy na pravé straně (6.1.3) jsou nezávislé, jelikož $e(t)$ je bílý šum. Jestliže $y^*(t)$ je predikce $y(t)$ (založená na datech až do času $(t-1)$), dostaneme

$$E[y(t) - y^*(t)]^2 = E\left[(-ay(t-1) + bu(t-1) + ce(t-1) - y^*(t))^2\right] + \lambda^2 \geq \lambda^2 \tag{6.1.4}$$

Takže zřejmě dolní mez variance chyby predikce je λ^2. Optimální predikce $\hat{y}(t \mid t-1; \theta)$ je charakteristická tím, že variance příslušné chyby predikce je minimální. Můžeme získat tedy rovnost v (6.1.4)? Jestliže ano, pak predikce musí splňovat

$$\hat{y}(t \mid t-1; \Theta) = -ay(t-1) + bu(t-1) + ce(t-1) \tag{6.1.5}$$

Problém s predikcí ve formě (6.1.5) je samozřejmě v tom, že veličina $e(t-1)$ není měřitelná. Takže odsud nelze přímo generovat $\hat{y}(t \mid t-1; \Theta)$. Avšak lze provést rekonstrukci neměřitelné poruchy $e(t-1)$ z naměřených dat $y(t-2), u(t-2), ..., y(1), u(1)$ jak bude ukázáno dále. Užitím (6.1.2) dostaneme

$$\begin{aligned}
\hat{y}(t \mid t-1; \Theta) &= -ay(t-1) + bu(t-1) + c[y(t-1) + ay(t-2) - bu(t-2) - ce(t-2)] \\
&= (c-a)y(t-1) + acy(t-2) + bu(t-1) - bcu(t-2) \\
&- c^2[y(t-2) + ay(t-3) - bu(t-3) - ce(t-3)] \\
&= (c-a)y(t-1) + (ac - c^2)y(t-2) - ac^2y(t-3) + bu(t-1) \\
&- bcu(t-2) + bc^2u(t-3) + c^3e(t-3) \\
&= ... = \sum_{i=1}^{t-1}(c-a)(-c)^{i-1}y(t-i) - a(-c)^{t-1}y(0) \\
&+ b\sum_{i=1}^{t}(-c)^{i-1}u(t-i) - (-c)^t e(0)
\end{aligned} \tag{6.1.6}$$

Jestliže předpokládáme, že $\mid c \mid < 1$, což je splněno podle definice pro $\Theta \in D$, můžeme pro velké t zanedbat poslední člen v (6.1.6), jehož vliv s přibývajícím časem klesá. Zanedbáním posledního členu, nabývá prediktor (6.1.6) následující formy

$$\hat{y}(t \mid t-1; \Theta) = \sum_{i=1}^{t-1}(c-a)(-c)^{i-1}y(t-i) - a(-c)^{t-1}y(0) + b\sum_{i=1}^{t}(-c)^{i-1}u(t-i) \tag{6.1.7}$$

která je již realizovatelná, tj. závisí pouze na známých veličinách. Výraz (6.1.7) však není vhodný pro praktickou implementaci, protože v každém časovém okamžiku vyžaduje zpracování celé sekvence měřených dat. Pro nalezení vhodnějšího algoritmu je užitečné rozepsat součet na pravé

straně, a pak je zřejmé, že přičtením výrazu $c\hat{y}(t-1 \mid t-2;\Theta)$ na obě strany (6.1.7) se pravá strana zjednoduší a dostaneme

$$\hat{y}(t \mid t-1;\Theta) + c\hat{y}(t-1 \mid t-2;\Theta) = (c-a)y(t-1) + bu(t-1) \tag{6.1.8}$$

což umožňuje jednoduchý rekurzivní výpočet optimální predikce. Chyba predikce $\epsilon(t,\Theta)$, bude popsána podobnou rekurzí. Snadno získáme

$$\begin{aligned} \epsilon(t,\Theta) + c\epsilon(t-1,\Theta) &= y(t) + cy(t-1) - [(c-a)y(t-1) + bu(t-1)] \\ &= y(t) + ay(t-1) - bu(t-1) \end{aligned} \tag{6.1.9}$$

Rekurze (6.1.9) vyžaduje použít počáteční hodnotu $\epsilon(0,\Theta)$, která je však neznámá. Uvedený problém lze však vyřešit ve většině případů položením $\epsilon(0,\Theta) = 0$. Protože podle předpokladu $\mid c \mid < 1$, efekt této nepřesnosti bude mít s přibývajícím časem klesající význam.

Uveďme k předchozímu několik poznámek. Nejprve si povšimněme podobnosti mezi modelem (6.1.2) a rovnicí (6.1.9) pro chybu predikce $\epsilon(t,\Theta)$! Dále je vhodné ukázat, že optimální prediktor je možné odvodit mnohem snáze pomocí a v kompaktní formě použitím polynomiálního formalismu. Výstup modelu (6.1.2) může být zapsán

$$\begin{aligned} y(t) &= \frac{bq^{-1}}{1+aq^{-1}}u(t) + \frac{1+cq^{-1}}{1+aq^{-1}}e(t) \\ &= \left(\frac{bq^{-1}}{1+aq^{-1}}u(t) + \frac{(c-a)q^{-1}}{1+aq^{-1}}e(t)\right) + e(t) \\ &= \left(\frac{bq^{-1}}{1+aq^{-1}}u(t) + \frac{(c-a)q^{-1}}{1+aq^{-1}}\frac{1}{1+cq^{-1}}[(1+aq^{-1})y(t) - bq^{-1}u(t)]\right) + e(t) \\ &= \left(\frac{bq^{-1}}{(1+aq^{-1})(1+cq^{-1})}[(1+cq^{-1}) - (c-a)q^{-1}]u(t) + \frac{(c-a)q^{-1}}{1+cq^{-1}}y(t)\right) + e(t) \end{aligned}$$

a tedy dostaneme

$$\hat{y}(t \mid t-1;\Theta) = \frac{bq^{-1}}{1+cq^{-1}}u(t) + \frac{(c-a)q^{-1}}{1+cq^{-1}}y(t) \tag{6.1.10}$$

což je jen jiná forma (6.1.8). Poznamenejme, že (6.1.10) je vlastně filtr s nekonečnou impulsní odezvou a v takovém případě by měla být známa data na nekonečném intervalu, aby statistické vlastnosti chyby predikce $\epsilon(t,\Theta)$ (6.1.1) nabyly ustáleného stavu. Pak vliv počátečních hodnot $\epsilon(0,\Theta)$ v (6.1.9) zanedbatelný.
Přestože polynomiální formalismus umožňuje rychlé a elegantní odvození optimální predikce, pro praktickou implementaci prediktoru může být diferenční rovnice ve formě (6.1.8) vhodnější. Například, pro realizaci prediktoru v prostředí MATLAB® je vhodná diferenční rovnice (6.1.8), zatímco pro realizaci v grafickém prostředí MATLAB® Simulink je vhodná forma s přenosovými funkcemi (6.1.10). Samozřejmě výsledek (6.1.10) může být snadno přepsán na diferenční rovnici (6.1.8) a naopak.

Nyní uvažujme obecný lineární model

$$\begin{aligned} y(t) &= G(q^{-1};\Theta)u(t) + H(q^{-1};\Theta)e(t) \\ Ee(t)e^T(s) &= \Lambda(\Theta)\delta_{t,s}, \end{aligned} \tag{6.1.11}$$

který byl zaveden v (5.2.1). Předpokládejme, že $G(0;\Theta)=0, H(0;\Theta)=I$ a nechť $H^{-1}(q^{-1};\Theta)$ a $H^{-1}(q^{-1};\Theta)G(q^{-1};\Theta)$ jsou asymptoticky stabilní, to jest $\Theta \in D$, jak je definováno v (5.2.2). Optimální predikci je pak možno snadno nalézt s využitím polynomiálního formalismu následujícím způsobem. Model (6.1.11) lze psát

$$\begin{aligned} y(t) &= G(q^{-1};\Theta)u(t) + [H(q^{-1};\Theta) - I]e(t) + e(t) \\ &= \{G(q^{-1};\Theta)u(t) + [H(q^{-1};\Theta) - I]H^{-1}(q^{-1};\Theta)[y(t) - G(q^{-1};\Theta)u(t)]\} + e(t) \\ &= \{H^{-1}(q^{-1};\Theta)G(q^{-1};\Theta)u(t) + [I - H^{-1}(q^{-1};\Theta)]y(t)\} + e(t) \end{aligned} \tag{6.1.12}$$

Tudíž, výsledný tvar jednokrokového prediktoru je

$$\hat{y}(t \mid t-1;\Theta) = H^{-1}(q^{-1};\Theta)G(q^{-1};\Theta)u(t) + [I - H^{-1}(q^{-1};\Theta)]y(t) \tag{6.1.13}$$

$$\hat{y}(t \mid t-1;\Theta) = L_1(q^{-1};\Theta)y(t) + L_2(q^{-1};\Theta)u(t) \tag{6.1.14}$$

kde

$$L_1(q^{-1};\Theta) = [I - H^{-1}(q^{-1};\Theta)]$$
$$L_2(q^{-1};\Theta) = H^{-1}(q^{-1};\Theta)G(q^{-1};\Theta)$$

Chyba predikce pak zřejmě splňuje následující vztah

$$\epsilon(t,\Theta) = e(t) = H^{-1}(q^{-1};\Theta)[y(t) - G(q^{-1};\Theta)u(t)] \tag{6.1.15}$$

Poznamenejme, že rovnost $\epsilon(t,\Theta) = e(t)$ je nutné chápat jako asymptotickou, tj. kdy $t \to \infty$, pokud nejsou zcela shodné počáteční podmínky systému (generátoru dat) a prediktoru. Také bychom měli zmínit, že použití množiny D, která byla zavedena v (5.2.2) přesně znamená, že jsme se omezili na takové hodnoty Θ, že predikce (6.1.13) je asymptoticky stabilní. Pro konkrétní případ uvažovaný v příkladu 6.1.1 máme

$$G(q^{-1};\Theta) = \frac{bq^{-1}}{1+aq^{-1}} \quad H(q^{-1};\Theta) = \frac{1+cq^{-1}}{1+aq^{-1}}.$$

což, vzhledem k (6.1.13), vede na výraz

$$\begin{aligned} \hat{y}(t \mid t-1;\Theta) &= \frac{1+aq^{-1}}{1+cq^{-1}}\frac{bq^{-1}}{1+aq^{-1}}u(t) + \left(1 - \frac{1+aq^{-1}}{1+cq^{-1}}\right)y(t) \\ &= \frac{bq^{-1}}{1+cq^{-1}}u(t) + \frac{(c-a)q^{-1}}{1+cq^{-1}}y(t) \end{aligned}$$

plně shodný se vztahem (6.1.10).

Bylo ukázáno, jak vypočítat optimální predikci a chybu predikce pro obecný lineární model. K odvození predikce bylo využito předpokladu, že $e(t)$ v (6.1.11) je bílý šum. Predikci (6.1.13) můžeme použít i tehdy, jestliže tento předpoklad nebude splněn. Pak ovšem samozřejmě nedostaneme optimální prediktor.

6.2 Analýza metody nejmenších čtverců

Ve 4. kapitole jsme použili metodu nejmenších čtverců na lineární statické regresní modely. Odvozené výsledky závisí pouze na algebraické struktuře. V této subkapitole ukážeme aplikaci metody i na dynamické modely. Bude vidět, že statistická analýza ze 4. kapitoly automaticky neplatí pro tento typ modelů.
Pro aplikaci lineární regrese na dynamické modely uvažujme

$$A(q^{-1})y(t) = B(q^{-1})u(t) + \epsilon(t) \tag{6.2.1}$$

kde

$$A(q^{-1}) = 1 + a_1 q^{-1} + ... + a_{na} q^{-na}$$
$$B(q^{-1}) = b_1 q^{-1} + ... + b_{nb} q^{-nb}$$

V (6.2.1) člen $\epsilon(t)$ označuje chybu rovnice. Model (6.2.1) lze ekvivalentně vyjádřit jako

$$y(t) = \varphi^T(t)\Theta + \epsilon(t) \tag{6.2.2}$$

kde

$$\begin{aligned} \varphi^T(t) &= [-y(t-1), \ldots, -y(t-na), u(t-1), \ldots, u(t-nb)] \\ \Theta &= [a_1, \ldots, a_{na}, b_1, \ldots, b_{nb}]^T \end{aligned}$$

Model (6.2.2) je formálně shodný s těmi modely, kterými jsme se zabývali ve 4. kapitole. Víme proto, že optimální odhad parametrů ve smyslu minimalizace kritéria

$$V_N(\Theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t) \tag{6.2.3}$$

je

$$\hat{\Theta} = [\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)]^{-1}[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)y(t)] \tag{6.2.4}$$

Tato identifikační metoda je známa pod názvem metoda nejmenších čtverců. Někdy se používá též název "metoda chyby rovnice", protože se minimalizuje chyba rovnice.

Je zřejmé, že diskuse týkající se algoritmu výpočtu $\hat{\Theta}$ provedená ve 4. kapitole platí i zde, jelikož algebraická struktura odhadu je stejná. Pro statistické vlastnosti odhadu je však rozhodující, zda $\varphi(t)$ je předem daná, jako tomu bylo ve 4. kapitole nebo zda se jedná o realizaci stochastického procesu, tak jako v (6.2.2).
Proveďme analýzu odhadu (6.2.4) pro model (6.2.1), (6.2.2). Předpokládejme, že data splňují

$$A_0(q^{-1})y(t) = B_0(q^{-1})u(t) + v(t) \tag{6.2.5}$$

nebo-li

$$y(t) = \varphi^T(t)\Theta_0 + v(t) \tag{6.2.6}$$

kde Θ_0 je vektor skutečných parametrů a $\{v(t)\}$ je stacionární stochastický proces nezávislý na vstupním signálu.

Kvalitu odhadu $\hat{\Theta}$ jsme v kapitole 4 posuzovali prostřednictvím strannosti a konzistence odhadu. Pro analýzu kvality odhadu metodou nejmenších čtverců je tedy vhodné zavést následující chybu odhadu.

$$\begin{aligned}
\hat{\Theta} - \Theta_0 &= \left(\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)\right)^{-1}\left(\frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t) - \{\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)\}\Theta_0\right) \\
&= \left(\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)\right)^{-1}\left(\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\{\varphi^T\Theta_0 + v(t)\} - \{\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)\}\Theta_0\right) \\
&= \left(\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)\right)^{-1}\left(\frac{1}{N}\sum_{t=1}^{N}\varphi(t)v(t)\right) \qquad (6.2.7)
\end{aligned}$$

Nyní je zřejmé, že odhad $\hat{\Theta}$ je konsistentní ($\hat{\Theta} \to \Theta_0$, když $N \to \infty$), jestliže

$$E[\varphi(t)\varphi^T(t)] \quad \text{je nonsingulární} \qquad (6.2.8)$$

a

$$E[\varphi(t)v(t)] = 0 \qquad (6.2.9)$$

Podmínka (6.2.8) je splněna v mnoha případech. Nicméně jsou případy, kdy tomu tak není:

- vstup není trvale budicí alespoň dostatečného řádu,
- výstup systému neobsahuje šum ($v(t) \equiv 0$), popř. řád modelu je vybrán příliš vysoký,
- vstup $u(t)$ je generován lineární zpětnou vazbou od výstupu příliš nízkého řádu.

Na rozdíl od podmínky (6.2.8), není podmínka (6.2.9) splněna ve většině případů. Výjimkou je ta situace, kdy porucha $v(t)$ je bílý šum se střední hodnotou nula, pak je rovnost (6.2.9) jistě splněna. Avšak, jestliže $v(t)$ není bílý šum, pak bude nejspíše korelován s minulými výstupy, protože $y(t)$ závisí podle (6.2.5) na $v(s)$ pro $s \leq t$, a tedy vztah (6.2.9) neplatí. Jak bylo výše ukázáno, i když metoda nejmenších čtverců je relativně jednoduchá, poskytuje konsistentní parametrické odhady při splnění dosti tvrdých podmínek. V některých případech problémy s konsistencí odhadu mohou být tolerovány. Například, jestliže je poměr signál šum velký, bude strannost odhadu malá. Rovněž při syntéze regulátoru, který je založen na identifikovaném modelu, může být strannost přijatelná, protože dobře navržený regulátor by měl mít uzavřený systém necitlivý na parametrické změny v otevřené smyčce. Avšak v jiných situacích může být značně důležité požadovat konsistentní odhady parametrů. Proto jsou v této a následující kapitole ukázány dvě modifikace metody nejmenších čtverců, které dávají konsistentní odhad při mnohem slabších omezeních. O jaké modifikace jde:

- Minimalizace chyby predikce zvolenou (detailní) strukturu modelu *odpovídající* generátoru dat. Tato myšlenka vede na metodu chyby predikce, kterou se budeme zabývat v této kapitole.
- Úprava rovnic spojených s odhadem metodou nejmenších čtverců (zejména pak vztah (6.2.4)). Tato idea vede na metodu přídavné proměnné, která je diskutována v následující kapitole.

6.3 Popis metody chyby predikce

Princip identifikace pomocí metody chyby predikce je založen na takovém výběru vektoru parametrů Θ, aby chyby predikce $\{\epsilon(t,\Theta)\}$ byly minimální. Formalizace takové myšlenky může být následující. Předpokládejme známá nebo změřená data $\{u(1), y(1), \ldots, u(N), y(N)\}$ (tedy N vzorkovacích okamžiků). Aplikací (6.1.15) můžeme vypočítat chyby predikce $\{\epsilon(1,\Theta), \ldots, \epsilon(N,\Theta)\}$. Utvořme výběrovou kovarianční matici

$$R_N(\Theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon(t,\Theta)\epsilon^T(t,\Theta) \tag{6.3.1}$$

Jestliže systém má pouze jeden výstup ($ny = 1$), pak $\epsilon(t,\Theta)$ je skalár a stejně tak i $R_N(\Theta)$. Pak můžeme $R_N(\Theta)$ chápat nejen jako ztrátovou funkci, ale i jako vhodné kriterium k minimalizaci. Pro MIMO systém je $R_N(\Theta)$ pozitivně semi-definitní matice a jako kriterium můžeme uvažovat

$$V_N(\Theta) = h(R_N(\Theta)) \tag{6.3.2}$$

kde $h(\cdot)$ je vhodná skalární funkce splňující určité podmínky. Funkci $V_N(\Theta)$ budeme nazývat ztrátovou funkcí. Poznamenejme, že počet dat N jsme použili jako index u ztrátové funkce. Požadavkem na funkci $h(\cdot)$ je, aby byla spojitá a platilo

$$h(Q + \triangle Q) \geq h(Q) \tag{6.3.3}$$

kde Q a $\triangle Q$ jsou pozitivně semi-definitní matice.
Nyní v příkladu ukážeme některé možnosti výběru $h(Q)$.

Příklad 6.3.1 (Kriteriální funkce) Jedna možnost výběru $h(Q)$ je

$$h_1(Q) = \mathrm{tr}(SQ) \tag{6.3.4}$$

kde S je symetrická pozitivně definitní váhová matice a funkce $\mathrm{tr}(\cdot)$ značí stopu matice. Pak můžeme zapsat S jako $S = GG^T$ a dosazením do (6.3.3) dostaneme

$$h_1(Q + \triangle Q) - h_1(Q) = trS\triangle Q = trGG^T\triangle Q = trG^T\triangle QG \geq 0$$

takže podmínka (6.3.3) je vždy splněna.

Jiná možnost výběru $h(Q)$ je

$$h_2(Q) = \det(Q) \tag{6.3.5}$$

kde funkce $\det(\cdot)$ znamená determinant.

Nyní shrňme vše co bylo doposud řečeno o metodě chyby predikce. Metoda může být popsána takto:

- Výběr struktury modelu ve formě (6.1.11) a formy prediktoru (6.1.13).
- Výběr ztrátové funkce $h(\cdot)$, viz (6.3.2).
- Určení odhadu parametrů $\hat{\Theta}$, pro který ztrátová funkce $h(R_N(\Theta))$ nabývá (globální) minimum, tedy $\hat{\Theta} = \arg\min_\Theta h(R_N(\Theta))$, což znamená hodnotu, která minimalizuje $h(R_N(\Theta))$. Minimalizace je ve většině případů numerická, a tak je ztrátová funkce vyhodnocována pro iteračně získávané odhady vektoru Θ na základě příslušné sekvence chyby predikce $\{\epsilon(t,\Theta)\}_{t=1}^N$ spočtené dle (6.1.15) a aktuálního odhadu parametrů. Výběrovou kovarianční matici $R_N(\Theta)$ pak můžeme vypočítat podle (6.3.1).

Z předchozích bodů charakterizujících metodu chyby predikce je zřejmé, že se jedná o postup, který v sobě zahrnuje množství speciálních operací definovaných konkrétním výběrem struktury modelu, prediktoru, ztrátové funkce a způsobu minimalizace. Právě pro takové konkrétní vymezení byly navrženy postupy odhadu parametrů a jsou známy jako samostatné metody, což vedlo k určité roztříštěnosti pohledu na identifikaci jako takovou. Protože však všechny tyto metody jsou založeny na výše uvedeném přístupu, je možné je chápat jako jeden postup označovaný metoda (metody) chyby predikce.
Jako příklad různé volby ztrátové funkce si uveďme pro skalární případ (např. systém s jedním vstupem a jedním výstupem) nejjednodušší možnou

$$V_N(\Theta) = \frac{1}{N}\sum_{t=1}^{N}\epsilon^2(t,\Theta) \tag{6.3.6}$$

kdy všechny chyby predikce mají stejnou váhu a

$$V_N(\Theta) = \frac{1}{N}\sum_{t=1}^{N}\beta(t)\epsilon^2(t,\Theta) \tag{6.3.7}$$

kdy jsou explicitně váženy funkcí $\beta(t)$. Tato rozdílnost se pak projeví i v názvu identifikační metody, jak bude vidět později.

Za příklad volby struktury modelu nám poslouží ARX struktura, pro kterou vektor regresorů je dán

$$\varphi(t) = [-y(t-1), -y(t-2), \ldots, -y(t-na), u(t-1), \ldots, u(t-nb)]^T$$

Optimální predikce je pak

$$\hat{y}(t \mid t-1;\Theta) = (1 - A(q^{-1}))y(t) + B(q^{-1})u(t) = \varphi^T(t)\Theta$$

tedy

$$L_1(q^{-1};\Theta) = 1 - A(q^{-1}), \quad L_2(q^{-1};\Theta) = A(q^{-1})B(q^{-1})/A(q^{-1}) = B(q^{-1}) \tag{6.3.8}$$

a chyba predikce

$$\epsilon(t,\Theta) = y(t) - \hat{y}(t \mid t-1;\Theta) \tag{6.3.9}$$

Jestliže stanovíme ztrátovou funkci

$$V_N(\Theta) = \frac{1}{N}\sum_{t=1}^{N}[y(t) - \varphi^T(t)\Theta]^2 = \frac{1}{N}\sum_{t=1}^{N}\epsilon^2(t,\Theta) = R_N(\Theta) \tag{6.3.10}$$

pak optimální odhad parametrů můžeme vypočítat analyticky

$$\hat{\Theta} = \arg\min_{\Theta} V_N(\Theta) = [\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)]^{-1}\frac{1}{N}\sum_{t=1}^{N}\varphi(t)y(t) \tag{6.3.11}$$

a lze konstatovat, že metoda chyby predikce použitá na tuto strukturu modelu je známa též jako metoda nejmenších čtverců. Pokud bychom uvažovali kriterium ve formě

$$V_N(\Theta) = \frac{1}{N}\sum_{t=1}^{N}\beta(t)[y(t) - \varphi^T(t)\Theta]^2 \tag{6.3.12}$$

výsledný odhad parametrů by byl

$$\hat{\Theta} = [\frac{1}{N}\sum_{t=1}^{N}\beta(t)\varphi(t)\varphi^T(t)]^{-1}\frac{1}{N}\sum_{t=1}^{N}\beta(t)\varphi(t)y(t) \tag{6.3.13}$$

Tento postup bývá označován jako metoda vážených nejmenších čtverců.

V podkapitole 6.2 bylo ukázáno, za jakých podmínek metoda nejmenších čtverců poskytuje konsistentní odhady. V případě, že $v(t)$ v (6.2.5) není bílý šum nelze obecně zaručit konsistentní odhad. V následujícím si ukážeme dvě situace, kdy lze aplikovat metodu nejmenších čtverců i přesto, že $v(t)$ v (6.2.5) není bílý.

Příklad 6.3.1. Předpokládejme, že a_i, b_i v (6.2.5) jsou neznámé a $v(t)$ je výstup ze známého filtru, tedy mějme

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \tag{6.3.14}$$

$A(q^{-1}), B(q^{-1})$ jsou polynomy s neznámými koeficienty, zatímco $C(q^{-1})$ je známý a $\{e(t)\}$ je bílý šum. Pak přestože $v(t) = C(q^{-1})e(t)$ není bílý proces, můžeme použít metodu nejmenších čtverců na upravený model

$$A(q^{-1})y^F(t) = B(q^{-1})u^F(t) + e(t) \tag{6.3.15}$$

kde $C(q^{-1})y^F(t) = y(t)$ a $C(q^{-1})u^F(t) = u(t)$. Poznamenejme však, že toto není příliš reálná situace.

Příklad 6.3.2. Předpokládejme nyní v (6.2.5) $v(t)$ ve tvaru

$$D(q^{-1})v(t) = e(t) \tag{6.3.16}$$

kde $D(q^{-1})$ je stupně nd. Pak

$$A(q^{-1})y(t) = B(q^{-1})u(t) + \frac{1}{D(q^{-1})}e(t) \tag{6.3.17}$$

Tento vztah ale můžeme přepsat na

$$A(q^{-1})D(q^{-1})y(t) = B(q^{-1})D(q^{-1})u(t) + e(t) \tag{6.3.18}$$

Aplikací metody nejmenších čtverců na odhad parametrů polynomu $A(q^{-1})D(q^{-1})$ řádu $na+nd$ a $B(q^{-1})D(q^{-1})$ řádu $nb+nd$ dostaneme konsistentní odhady $A(q^{-1})D(q^{-1})$ a $B(q^{-1})D(q^{-1})$ a přenosová funkce z u na y

$$\frac{B(q^{-1})D(q^{-1})}{A(q^{-1})D(q^{-1})} = \frac{B(q^{-1})}{A(q^{-1})}$$

bude odhadnuta správně (za předpokladu $t \to \infty$).

Příklad 6.3.3. Volme tuto strukturu modelu

$$y(t) = G(q^{-1};\Theta)u(t) + e(t) \tag{6.3.19}$$

Pak chyba predikce vypočtená podle (6.1.15) bude

$$\epsilon(t,\theta) = y(t) - G(q^{-1};\Theta)u(t)$$

to jest rozdíl mezi měřeným výstupem $y(t)$ a „modelovaným deterministickým výstupem“ $G(q^{-1};\Theta)u(t)$. V takovém případě PEM je často nazývána *metoda výstupní chyby* (v anglicky psané literatuře označované „output error method“).

Příklad 6.3.4. V předchozím textu při popisu PEM jsme se doposud nezabývali vztahem PEM a metody maximální věrohodnosti ML (v anglicky psané literatuře označované „maximum likelihood method“).
K vyjasnění této problematiky zaveďme další předpoklad na šum v modelu (6.1.11). Předpokládejme, že má gaussovské rozložení. Odhad Θ ve smyslu maximální věrohodnosti získáme maximalizací věrohodnostní funkce, v tomto případě hustoty pravděpodobnosti pozorování podmíněné vektorem parametrů Θ. Protože existuje vzájemně jednoznačná transformace mezi $\{y(t)\}$ a $\{e(t)\}$ daná (6.1.11), můžeme rovněž dobře vyjít i z hustoty pravděpodobnosti poruchy (6.1.15). Užitím ny rozměrného gaussovského rozdělení nalezneme, že věrohodnostní funkce je dána

$$L(\Theta) = \frac{1}{(2\pi)^{Nny/2}[det\Lambda(\Theta)]^{N/2}} exp[-1/2\sum_{t=1}^{N}\epsilon^T(t,\Theta)\Lambda^{-1}(\Theta)\epsilon(t,\Theta)] \tag{6.3.20}$$

nebo po zlogaritmování obou stran přirozeným logaritmem dostaneme

$$logL(\Theta) = -\frac{1}{2}\sum_{t=1}^{N}\epsilon^T(t,\Theta)\Lambda^{-1}(\Theta)\epsilon(t,\theta) - \frac{N}{2}\log\det\Lambda(\Theta) + \text{ konstanta} \tag{6.3.21}$$

Předpokládejme, že $\epsilon(t,\Theta)$ a $\Lambda(\Theta)$ nemají společné parametry. Pro zjednodušení zápisu považujme parametry v Λ za parametry, které jsou navíc k Θ, a proto vynecháme argument Θ v $\Lambda(\Theta)$. Pro jednoduchost uvažujme pouze skalární případ ($ny = 1, \epsilon(t,\Theta)$ a Λ jsou skaláry). Odhady $\hat{\Theta}$, $\hat{\Lambda}$ získané metodou ML, maximalizují $L(\Theta,\Lambda)$. Nejdříve maximalizujme podle Λ výraz

$$\begin{aligned} logL(\Theta,\Lambda) &= -\frac{1}{2}\frac{1}{\Lambda}\sum_{t=1}^{N}\epsilon^2(t,\Theta) - \frac{N}{2}\log\Lambda + \text{ konstanta} \\ &= -\frac{N}{2}[R_N(\Theta)/\Lambda + \log\Lambda] + \text{ konstanta} \end{aligned} \tag{6.3.22}$$

Parciální derivaci podle Λ položíme rovnu nule.

$$-\frac{N}{2}[-R_N(\Theta)/\Lambda^2 + 1/\Lambda] = 0$$

Z předchozí rovnice vyplývá, že odhad Λ je

$$\hat{\Lambda} = R_N(\Theta) \tag{6.3.23}$$

Poznamenejme k tomu, že platí $\frac{\partial^2}{\partial\Lambda^2}logL < 0$, takže se jedná skutečně o maximum. Dosazením (6.3.23) do (6.3.22) dostaneme

$$logL(\Theta,\hat{\Lambda}) = -\frac{N}{2}\log(R_N(\Theta)) + \text{ jiná konstanta}$$

Odhad $\hat{\Theta}$ pak získáme maximalizací $-\log(R_N(\Theta))$ vzhledem k Θ. Bod, ve kterém $R_N(\Theta)$ nabývá minimální hodnoty, je odhad $\hat{\Theta}$ a minimální hodnota $R_N(\Theta)$ bude odhad $\hat{\Lambda}$.

Předchozí analýza demonstrovala zajímavý vztah mezi PEM a metodou ML. Jestliže předpokládáme, že poruchy v modelu jsou gaussovské, pak metoda ML přejde na metodu chyby predikce.

6.4 Analýza

V předchozí části kapitoly jsme se zabývali metodou chyby predikce. Tato metoda, či spíše přístup, i další parametrické metody identifikace, které zahrnuje, představuje zobrazení množiny dat do vektoru parametrů $\hat{\Theta}_N \in D$

$$\{y(1), u(1), ..., y(N), u(N)\} \to \hat{\Theta}_N \in D \tag{6.4.1}$$

Zkoumání vlastností tohoto zobrazení může být prováděno v principu dvěma způsoby:

1. Generovat data se známými charakteristikami. Provést zobrazení (6.4.1) s využitím konkrétní identifikační metody a ohodnotit vlastnosti odhadnutých parametrů $\hat{\Theta}_N$. Tento postup bývá nazýván simulace.
2. Předpokládat vlastnost dat a pokusit se odvodit jaké vlastnosti bude mít odhad $\hat{\Theta}_N$. Tento postup se nazývá analýza.

Připomeňme si, že oba postupy jsme prováděli na jednoduchých příkladech ve 2. kapitole.

Nyní se budeme věnovat analýze metody chyby predikce, konkrétně odhadu parametrů $\hat{\Theta}_N$, když N se blíží k nekonečnu. Protože teoretická analýza je poměrně náročná, omezíme se zde pouze na výsledky, které poskytuje. Předpokládejme

P1 Data $\{u(t), y(t)\}$ jsou stacionární procesy.

P2 Vstup je trvale budicí signál.

P3 $V_N''(\Theta)$ je nonsingulární matice přinejmenším v blízkosti bodu minimalizujícího $V_N(\Theta)$.

P4 Filtry $G(q^{-1}; \Theta)$ a $H(q^{-1}; \Theta)$ jsou hladké (diferencovatelné) funkce vektoru parametrů Θ.

P5 Množina $D_T(S, M)$ zavedená v 5. kapitole obsahuje přesně jeden bod.

Předpoklad P5 je potřebný jen někdy. Pro N blížící se nekonečnu výběrové kovariance podle ergodické teorie konvergují k odpovídající střední hodnotě. Protože funkce $h(\cdot)$ je spojitá, lze tvrdit, že

$$V_N(\Theta) = h(R_N(\Theta)) \to h(R_\infty(\Theta)) \stackrel{\triangle}{=} V_\infty(\Theta) \quad \text{pro} \quad N \to \infty \tag{6.4.2}$$

kde

$$R_\infty = E[\epsilon(t, \Theta)\epsilon^T(t, \Theta)] \tag{6.4.3}$$

Bylo dokázáno, že (6.4.2) konverguje stejnoměrně. Pak $\hat{\Theta}_N$ konverguje k bodu minimalizujícímu $V_\infty(\Theta)$. Označme tento bod Θ^*. Poznamenejme, že pro tento výsledek není třeba předpoklad P5. Jestliže totiž množina $D_T(S, M)$ je prázdná, pak získáme „nejrozumnější“ aproximaci. Vektor parametrů Θ je takový, že chyba predikce $\epsilon(t, \Theta)$ má nejmenší možnou varianci. Dále

předpokládejme, že množina $D_T(S,M)$ je neprázdná. Nechť Θ_0 je libovolný prvek $D_T(S,M)$. To znamená, že skutečný systém splňuje

$$y(t) = G(q^{-1};\Theta_0)u(t) + H(q^{-1};\Theta_0)e(t) \quad E[e(t)e^T(t)] = \Lambda(\Theta_0) \tag{6.4.4}$$

Kdyby $D_T(S,M)$ měla pouze jeden bod (P5), pak Θ_0 lze považovat za skutečný vektor parametrů. Zkoumejme nyní minima $R_\infty(\Theta)$. Z (6.1.1), (6.1.12), (6.1.15) a (6.4.4) vyplývá, že

$$\begin{aligned}\epsilon(t,\Theta) &= H^{-1}(q^{-1};\Theta)[G(q^{-1};\Theta_0)u(t) + H(q^{-1};\Theta_0)e(t) - G(q^{-1};\Theta)u(t)] \\ &= H^{-1}(q^{-1};\Theta)[G(q^{-1};\Theta_0) - G(q^{-1}\Theta)]u(t) \\ &+ H^{-1}(q^{-1};\Theta)H(q^{-1};\Theta_0)e(t)\end{aligned} \tag{6.4.5}$$

Protože $G(0,\Theta) = 0,\ H(0;\Theta) = H^{-1}(0;\Theta) = I$ pro všechny Θ dostaneme

$$\epsilon(t,\Theta) = e(t) + \text{ člen nezávislý na } e(t)$$

Tudíž

$$R_\infty(\Theta) = E[\epsilon(t,\Theta)\epsilon^T(t,\Theta)] \geq E[e(t)e^T(t)] = \Lambda(\Theta_0) \tag{6.4.6}$$

za předpokladu, že případná zpětná vazba je kauzální. Jelikož (6.4.6) dává spodní mez dosažitelnosti pro $\Theta = \Theta_0$, můžeme učinit závěr, že limitní odhad Θ^* je roven Θ_0. Dokážeme, že pro systém pracující v otevřené smyčce jsou prvky minimalizující R_∞, body Θ_0 z D_T. Pro systém se zpětnou vazbou je problém složitější.
Nechť systém pracuje v otevřené smyčce a $u(t)$ a $e(s)$ jsou nezávislé pro všechna t a s. Pak $R_\infty(\Theta) = \Lambda(\Theta_0)$ implikuje, viz (6.4.5), dva vztahy

$$H^{-1}(q^{-1};\Theta)[G(q^{-1};\Theta_0) - G(q^{-1};\Theta)]u(t) = 0$$

$$H^{-1}(q^{-1};\Theta)H(q^{-1};\Theta_0) = I$$

Z druhé relace zřejmě vyplývá, že

$$H(q^{-1};\Theta) = H(q^{-1};\Theta_0)$$

Pokud platí P2, pak lze z první identity odvodit, že

$$G(q^{-1};\Theta) = G(q^{-1};\Theta_0)$$

Tudíž máme $\Theta \in D_T(S,M)$. Předchozí výsledek pak znamená, že odhad $\hat{\Theta}_N$ ve smyslu chyby predikce je konsistentní. Poznamenejme, že při splnění P1-P4 je systém identifikovatelný a jestliže je splněn i P5, pak systém je parametricky identifikovatelný.

Po této analýze o limitě $\hat{\Theta}_N$ budeme pokračovat ve zkoumání limity rozložení. Bude ukázáno, že odhad $\hat{\Theta}_N$ má asymptoticky gaussovské rozložení. Odhad $\hat{\Theta}_N$ je bod, ve kterém ztrátová funkce $V_N(\Theta)$ nabývá svého minima. Předpokládejme, že množina $D_T(S,M)$ má přesně jeden bod, to jest existuje jediný správný vektor Θ_0 (předpoklad P5). Proveďme Taylorův rozvoj $V_N'^T(\hat{\Theta}_N)$ v bodě Θ_0 a ponechme pouze první dva členy

$$V_N'^T(\hat{\Theta}_N) \approx V_N'^T(\Theta_0) + V_N''(\Theta_0)(\hat{\Theta}_N - \Theta_0)$$

Jelikož v bodě extrému musí být první derivace nulová, tj. $V_N'^T(\hat{\Theta}_N) = 0$, získáme

$$0 \approx V_N'^T(\Theta_0) + V_\infty''(\Theta_0)(\hat{\Theta}_N - \Theta_0) \tag{6.4.7}$$

kde $V_N''(\Theta_0)$ bylo nahrazeno limitou $V_\infty''(\Theta_0)$, která platí s pravděpodobností 1.

Zde V' označuje gradient V a V'' je matice druhých derivací. Z (6.4.7) dostaneme pro velké N

$$\sqrt{N}(\hat{\Theta}_N - \Theta_0) \approx -[V_\infty''(\Theta_0)]^{-1}[\sqrt{N}V_N'^T(\Theta_0)] \tag{6.4.8}$$

Matice $V_\infty''(\Theta_0)$ je deterministická. Vektor $\sqrt{N}\ V_N'^T(\Theta_0)$ je však náhodná proměnná, která má asymptoticky gaussovské rozložení se střední hodnotou nula a kovarianční maticí P (viz znění centrální limitní věty). Pak můžeme psát

$$\sqrt{N}(\hat{\Theta}_N - \Theta_0) \xrightarrow{\text{dist}} N(0, P) \tag{6.4.9}$$

kde symbol $\xrightarrow{\text{dist}}$ znamená konvergenci v rozložení a

$$P = [V_\infty''(\Theta_0)]^{-1} P_0 [V_\infty''(\Theta_0)]^{-1} \tag{6.4.10}$$

V případě skalárního systému můžeme matici P lze vyhodnotit následujícím způsobem. Uvažujme kritérium

$$V_N(\Theta) = \frac{1}{N}\sum_{t=1}^{N} \epsilon^2(t, \Theta), \qquad V_\infty(\Theta) = E\epsilon^2(t, \Theta) \tag{6.4.11}$$

a zaveďme $n\Theta$ dimenzionální vektor $\psi(t, \Theta) = -(\frac{\partial \epsilon(t,\Theta)}{\partial \Theta})^T$. Bylo dokázáno (důkazem se nebudeme zabývat), že

$$P = \Lambda[E\psi(t, \Theta_0)\psi^T(t, \Theta_0)]^{-1} \tag{6.4.12}$$

Jako odhad P, dále značený $\hat{P}$, můžeme tedy použít

$$\hat{P} = \hat{\Lambda}[\frac{1}{N}\sum_{t=1}^{N} \psi(t, \hat{\Theta}_N)\psi^T(t, \hat{\Theta}_N)]^{-1} \tag{6.4.13}$$

To znamená, že (6.4.13) poskytuje obraz o kvalitě odhadu parametrů a lze jej vypočítat z dat. Ukažme výpočet odhadu matice P na jednodušších příkladech.

Příklad 6.4.1. (Lineární regrese)
Předpokládejme, že struktura modelu je

$$A(q^{-1})y(t) = B(q^{-1})u(t) + e(t)$$

kterou zapišme ve formě

$$y(t) = \varphi^T(t)\Theta + e(t)$$

Pak máme $\epsilon(t) = y(t) - \varphi^T(t)\Theta$. To znamená, že

$$\psi(t, \Theta) = -\left(\frac{\partial \epsilon(t)}{\partial \Theta}\right)^T = \varphi(t)$$

Takže z (6.4.12) dostaneme

$$P = \Lambda[E\varphi(t)\varphi^T(t)]^{-1} \tag{6.4.14}$$

Zajímavé je srovnat (6.4.14) s odpovídajícím výsledkem pro statický případ ze 4. kapitoly. S nyní uvažovanou notací a pro konečné množství dat máme

i) $\hat{\Theta}$ je nestranné,

ii) $\sqrt{N}(\hat{\Theta} - \Theta_0)$ má gaussovské rozložení $N(0, P)$, kde $P = \Lambda[\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)]^{-1}$.

V dynamickém případě tyto výsledky exaktně neplatí pro konečné N. Místo nich máme asymptotické výsledky:

i) $\hat{\Theta}$ je konsistentní,

ii) $\sqrt{N}(\hat{\Theta} - \Theta_0)$ je asymptoticky gaussovské rozložení $N(0, P)$, kde $P = \Lambda[E\varphi(t)\varphi^T(t)]^{-1}$.

Konkrétně v případě, že máme ARX model 1. řádu (a systém), pak

$$\varphi(t) = [-y(t-1), u(t-1)]^T$$

a P v (6.4.14) bude

$$P = \Lambda \begin{bmatrix} E[y^2(t-1)] & -E[y(t-1)u(t-1)] \\ -E[y(t-1)u(t-1)] & E[u^2(t-1)] \end{bmatrix}^{-1} \tag{6.4.15}$$

Příklad 6.4.2 (ARMA proces 1. řádu)
Mějme strukturu modelu

$$y(t) + ay(t-1) = e(t) + ce(t-1), \qquad E[e^2(t)] = \Lambda,\ \Theta = [a\ c]^T \tag{6.4.16}$$

Pak máme

$$\epsilon(t,\Theta) = \frac{1 + aq^{-1}}{1 + cq^{-1}} y(t)$$

$$\frac{\partial \epsilon(t,\Theta)}{\partial a}(t,\Theta) = \frac{q^{-1}}{1 + cq^{-1}} y(t) = q^{-1} y^F(t)$$

$$\frac{\partial \epsilon}{\partial c}(t,\Theta) = -\frac{1 + aq^{-1}}{(1 + cq^{-1})^2} q^{-1} y(t) = -\frac{q^{-1}}{1 + cq^{-1}} \epsilon(t,\Theta) = -q^{-1}\epsilon^F(t)$$

Tudíž dostaneme

$$P = \Lambda \begin{bmatrix} E[(y^F(t-1))^2] & -E[y^F(t-1)\epsilon^F(t-1)] \\ -E[y^F(t-1)\epsilon^F(t-1)] & E[(\epsilon^F(t-1))^2] \end{bmatrix}^{-1}$$

kde

$$y^F(t) = \frac{1}{1 + cq^{-1}} y(t), \qquad \epsilon^F(t) = -\frac{1}{1 + cq^{-1}} \epsilon(t)$$

a kde byla pro jednoduchost zanedbána explicitní závislost filtrovaných veličin na vektoru parametrů Θ.
Ve výpočtu matice P můžeme dále pokračovat. Pro $\Theta = \Theta_0$ dostaneme, s vyžitím (6.4.16), následující filtrované veličiny

$$y^F(t) = \frac{1}{1 + aq^{-1}} e(t), \quad \epsilon^F(t) = -\frac{1}{1 + cq^{-1}} e(t)$$

Nyní musíme vypočítat postupně $E[(y^F(t-1))^2],\ E[y^F(t-1)\epsilon^F(t-1)]$ a $E[(\epsilon^F(t-1))^2]$. Takže

$$E[(y^F(t-1))^2] = E[(-ay^F(t-2)+e(t-1))^2] = a^2E[(y^F(t-2))^2]+\Lambda$$

$$E[(y^F(t-1))^2] - a^2E[(y^F(t-1))^2] = \Lambda$$

$$E[(y^F(t-1))^2] = \frac{\Lambda}{1-a^2}$$

$$\begin{aligned} E[y^F(t-1)\epsilon^F(t-1)] &= E[-ay^F(t-2)+e(t-1)][-c\epsilon^F(t-2)+e(t-1)] \\ &= ac[Ey^F(t-2)\epsilon^F(t-2)]+\Lambda \\ E[y^F(t-1)\epsilon^F(t-1)] &= \frac{\Lambda}{1-ac} \end{aligned}$$

$$\begin{aligned} E[(\epsilon^F(t-1))^2] &= E[-c\epsilon^F(t-2)+e(t-1)]^2 = c^2E[(\epsilon^F(t-1))^2]+\Lambda \\ E[(\epsilon^F(t-1))^2] &= \frac{\Lambda}{1-c^2} \end{aligned}$$

Vypočtené výrazy dosadíme do matice P a dostaneme

$$P = \Lambda \begin{bmatrix} \Lambda/(1-a^2) & -\Lambda/(1-ac) \\ -\Lambda/(1-ac) & \Lambda/(1-c^2) \end{bmatrix}^{-1}$$

$$= \frac{1}{(c-a)^2} \begin{bmatrix} (1-a^2)(1-ac)^2 & (1-a^2)(1-ac)(1-c^2) \\ (1-a^2)(1-ac)(1-c^2) & (1-ac)^2(1-c^2) \end{bmatrix}$$

Matice P je nezávislá na varianci šumu Λ. Povšimněme si, že prvky kovarianční matice budou velmi velké, když c je blízko a. Všimněme si také, že pro $c = a$, kdy výstup skutečného systému je bílý šum, je model přeparametrizován. V takové situaci nemůžeme očekávat konvergenci odhadu $\hat{\Theta}_N$ k určitému bodu. Pro tento případ nebude mít asymptotická ztrátová funkce jediné minimum.

6.5 Výpočetní aspekty minimalizace a příklad implementace

V této části se budeme zabývat některými výpočetními aspekty minimalizace kriteriální funkce metody chyby predikce. Vyjma lineárního regresního modelu ve struktuře ARX, kde $\epsilon(t,\Theta)$ závisí *lineárně* na Θ, musí být často minimalizace $V_N(\Theta)$ provedena numerickým způsobem. Běžně používaný postup hledání extrému představuje algoritmus typu Newton-Raphson, kde iterační krok nabývá formy:

$$\hat{\Theta}^{(k+1)} = \hat{\Theta}^{(k)} - \alpha_k[V_N''(\hat{\Theta}^{(k)})]^{-1}V_N'^T(\hat{\Theta}^{(k)}) \tag{6.5.1}$$

Zde $\hat{\Theta}^{(k)}$ představuje odhad v k-té iteraci při hledání extrému. Skalární veličina α_k určuje délku kroku při hledání minima v (6.5.1), a tím ovlivňuje rychlost postupu. Základní verze algoritmu používá $\alpha_k = 1$, avšak při praktickém použití je vhodná proměnná délka kroku. Důvodem může být potřeba zajistit, aby $\Theta^{(k+1)} \in D$ pro všechna k nebo zlepšit konvergenci (6.5.1) a vybrat α_k tak, aby

$$\alpha_k = \arg\min_{\alpha} \left(V_N(\hat{\Theta}^{(k)}) - \alpha[V_N''(\hat{\Theta}^{(k)})]^{-1}V_N'^T(\hat{\Theta}^{(k)}) \right) \tag{6.5.2}$$

Definujme vektor prvních derivací vektoru chyby predikce $\epsilon(t,\Theta)$

$$\psi(t,\Theta) = -(\frac{\partial \epsilon(t,\Theta)}{\partial \Theta})^T \tag{6.5.3}$$

Pak, s ohledem na (6.4.11) snadno získáme první i druhou derivaci kriteriální funkce $V_N(\Theta)$ z (6.5.1) ve formě

$$V'_N(\Theta) = -\frac{2}{N}\sum_{t=1}^{N}\epsilon(t,\Theta)\psi^T(t,\Theta) \tag{6.5.4}$$

$$V''_N(\Theta) = \frac{2}{N}\sum_{t=1}^{N}\psi(t,\Theta)\psi^T(t,\Theta) + \frac{2}{N}\sum_{t=1}^{N}\epsilon(t,\Theta)\frac{\partial^2}{\partial\Theta^2}\epsilon(t,\Theta) \tag{6.5.5}$$

Protože pro $N \to \infty$ se $\epsilon(t,\Theta)$ blíží k bílému šumu, lze (6.5.5) aproximovat na

$$V''_N(\Theta) \sim \frac{2}{N}\sum_{t=1}^{N}\psi(t,\Theta)\psi^T(t,\Theta) \tag{6.5.6}$$

To nám umožní výrazně zjednodušit výpočet a (6.5.1) nabývá tvaru

$$\hat{\Theta}^{(k+1)} = \hat{\Theta}^{(k)} + \alpha_k[\sum_{t=1}^{N}\psi(t,\hat{\Theta}^{(k)})\psi^T(t,\hat{\Theta}^{(k)})]^{-1}[\sum_{t=1}^{N}\epsilon(t,\hat{\Theta}^{(k)})\psi^T(t,\hat{\Theta}^{(k)})] \tag{6.5.7}$$

Tento postup je nazýván Gauss-Newtonův algoritmus.

Příklad 6.5.1. (Algoritmus metody chyby predikce pro model ve struktuře ARMA)
Vraťme se nyní k formulaci problému v příkladu 6.4.2, tedy k odhadu parametrů ARMA modelu metodou chyby predikce. Jak již bylo řečeno, metoda chyby predikce by měla být spíše chápána jako myšlenkový koncept, který vede na mnoho různých identifikačních metod při uvažování konkrétní struktury modelu. V tomto příkladě bude navržena metoda chyby predikce pro ARMA model formou algoritmu, který může být použit jako základ pro implementaci. Algoritmus metody chyby predikce může být dán následujícími čtyřmi kroky.

Algoritmus metody chyby predikce pro model ve struktuře ARMA

(i) Předpokládejme sekvenci měřených dat $\{y(t)\}_{t=1}^{N}$. Z důvodu nastartování rekurzivní minimalizace (6.5.7) zvolme počáteční odhad hledaných parametrů a a c, tj., definujme vektor $\hat{\Theta}^{(1)} = [\hat{a}^{(1)}, \hat{c}^{(1)}]^T$. Nastavme $k = 1$ a definujme si práh pro ukončení rekurze δ_{thr}, popř. maximální počet iterací minimalizačního algoritmu k_{max}.

(ii) Abychom mohli provést iterační krok optimalizačního algoritmu a minimalizovat ztrátovou funkci (6.3.2) algoritmem (6.5.7) musíme spočítat vektor parciálních derivací $\psi(t,\Theta)$ (6.5.3) pro všechny časové okamžiky. Vektor je formován, jak je ukázáno v příkladu 6.4.2, filtrovanými veličinami $y^F(t)$ a $\epsilon^F(t)$, které lze, v k-té iteraci rekurze, spočítat jako

$$y^F(t) = -\hat{c}^{(k)}y^F(t-1) + y(t), \forall t \tag{6.5.8}$$

$$\epsilon^F(t) = -\hat{c}^{(k)}\epsilon^F(t-1) + \epsilon(t), \forall t \tag{6.5.9}$$

kde chyba predikce, vycházející ze vztahu (6.1. 15), je dána

$$\epsilon(t) = -\hat{c}^{(k)}\epsilon(t-1) + y(t) + \hat{a}^{(k)}y(t-1), \forall t \tag{6.5.10}$$

Počáteční podmínky rekurzivních vztahů (6.5.8)–(6.5.10) mohou být zvoleny jako nulové.

(iii) Odhad parametrů pro následující iteraci $k+1$, tj. $\hat{\Theta}^{(k+1)}$, je dán vztahem (6.5.7), kde

$$\psi(t, \hat{\Theta}^{(k)}) = [y^F(t-1), \epsilon^F(t-1)]^T \tag{6.5.11}$$

(iv) Pokud $\|\hat{\Theta}^{(k+1)} - \hat{\Theta}^{(k)}\| > \delta_{thr}$ a $k < k_{max}$, pak algoritmus pokračuje krokem (ii) s $k \leftarrow k+1$. Jinak algoritmus končí a poslední $\hat{\Theta}^{(k+1)}$ je považováno za výsledný odhad parametrů metodou chyby predikce.

Na závěr poznamenejme, že alternativa k (6.5.1) či (6.5.7) jsou rekurzivní algoritmy identifikace, kterými se budeme zabývat v 8. kapitole.

6.6 Shrnutí

Cílem této kapitoly bylo přiblížit metodu chyby predikce. K tomu bylo zapotřebí nejdříve zvládnout techniku výpočtu optimální predikce. Po analýze metody nejmenších čtverců jsme pak přistoupili k formulaci základních kroků charakterizujících PEM. Na příkladech byl dále ukázán vztah k některým identifikačním metodám. Závěr kapitoly byl věnován analýze metody a výpočetním aspektům minimalizace.

Podrobná analýza vlastností metod chyby predikce je uvedena např. v [15], [20], [42], [43] a možné optimalizační postupy v [44], [57].

Kapitola 7

Metoda přídavné proměnné

7.1 Základní verze metody přídavné proměnné

Metoda nejmenších čtverců, kterou jsme popisovali ve 2. a 4. kapitole umožňuje získat analytickým způsobem optimální odhad parametrů. Přitažlivost takového řešení spočívá na existenci jednoho extrému, globálního minima. Nicméně tato atraktivnost řešené úlohy parametrické identifikace byla podmíněna tvrdými podmínkami kladených na systém, bez jejichž splnění nelze získat konsistentní odhady. Proto byla v předcházející kapitole představena metoda chyby predikce, která poskytuje konsistentní odhad parametrů pro širokou třídu modelů. Metoda chyby predikce, podobně jako metoda nejmenších čtverců, byla založena na minimalizaci kvadrátu chyby predikce.
Oproti tomu, metoda přídavné proměnné, kterou se budeme zabývat v této kapitole, je založena na zcela jiné základní myšlence, kdy nehledáme nejlepší odhad minimalizující nějakou kriteriální (často kvadratickou) funkci, ale snažíme se najít jakýkoliv nestranný odhad parametrů. Tato metoda je svázána s výrazně měkčími předpoklady než metoda nejmenších čtverců a, oproti metodě chyby predikce, je jednorázová, a tudíž snadno aplikovatelná. Je též velmi rozpracovaná, ale v této práci se budeme věnovat pouze jejím základům, které jsou vhodné pro představení hlavní idee tohoto přístupu. Pro metodu přídavné proměnné se, v anglicky psané literatuře, vžilo označení "instrumental variable method, IVM".

Metoda přídavné proměnné je používána k odhadu dynamiky systému (přenosu od vstupu $u(t)$ na výstup $y(t)$) a neumožňuje zjistit vlastnosti šumu. Předpokládejme strukturu modelu, se kterou jsme se již vícekrát setkali

$$y(t) = \varphi^T(t)\Theta + \epsilon(t) \tag{7.1.1}$$

kde $y(t)$ je skalární výstup v čase t. Vektor regresorů $\varphi(t)$ obsahuje zpožděné výstupy a vstupy, Θ je $n\Theta$ dimenzionální vektor parametrů a konečně $\epsilon(t)$ je chyba rovnice. Z 5. kapitoly víme, že model (7.1.1) mohl vzniknout z

$$A(q^{-1})y(t) = B(q^{-1})u(t) + \epsilon(t) \tag{7.1.2}$$

kde

$$A(q^{-1}) = 1 + a_1 q^{-1} + \ldots + a_{na} q^{-na}$$

$$B(q^{-1}) = b_1 q^{-1} + \ldots + b_{nb} q^{-nb}$$

Vektor parametrů je určen

$$\Theta = [a_1, a_2, \ldots, a_{na}, b_1, \ldots, b_{nb}]^T \tag{7.1.3}$$

a vektor regresorů

$$\varphi(t) = [-y(t-1), \ldots, -y(t-na), u(t-1), \ldots, u(t-nb)]^T \tag{7.1.4}$$

Dále předpokládejme, že systém je popsán

$$y(t) = \varphi^T(t)\Theta_0 + v(t) \tag{7.1.5}$$

kde $v(t)$ je stochastická porucha.
Nejprve se zabývejme odhadem parametrů Θ ve smyslu nejmenších čtverců. Ten, jak víme z předchozích kapitol, je

$$\begin{aligned}\hat{\Theta} &= [\sum_{t=1}^{N} \varphi(t)\varphi^T(t)]^{-1}[\sum_{t=1}^{N} \varphi(t)y(t)] \\ &= [\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)]^{-1}[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)y(t)]\end{aligned} \tag{7.1.6}$$

Využitím popisu systému (7.1.5) v (7.1.6) dostaneme

$$\begin{aligned}\hat{\Theta} &= [\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)]^{-1}[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)\Theta_0] \\ &+ [\frac{1}{N}\sum_{t=1}^{N} \varphi(t)\varphi^T(t)]^{-1}[\frac{1}{N}\sum_{t=1}^{N} \varphi(t)v(t)]\end{aligned}$$

Když N se blíží nekonečnu, lze předchozí vztah zapsat jako

$$\hat{\Theta} - \Theta_0 = [E\varphi(t)\varphi^T(t)]^{-1}E\varphi(t)v(t) \tag{7.1.7}$$

To ale znamená, že odhad (7.1.6) je obecně asymptoticky stranný a nekonsistentní. Výjimku samozřejmě tvoří případ, kdy

$$E\varphi(t)v(t) = 0 \tag{7.1.8}$$

Všimněme si však, že $\varphi(t)$ závisí na výstupu, a tudíž závisí implicitně také na starých hodnotách $v(t)$. To znamená, že (7.1.8) je dosti tvrdý požadavek. Je těžké splnit (7.1.8) kromě situace, kdy $v(t)$ je bílý šum se střední hodnotou nula. V příkladech ve druhé kapitole jsme podobný případ vyšetřovali. Právě tento moment při hledání odhadu je rozhodující pro motivaci zavedení metody přídavné proměnné (v anglicky psané literatuře označované jako „instrumental variable“, IV).
Základní idea charakterizující IV odhad vektoru parametrů Θ_0 může být objasněna několika způsoby. Přímočarý a jednoduchý způsob je tento. Předpokládejme, že $\zeta(t)$ je známý $n\zeta$ dimenzionální vektor (v případě, že bychom uvažovali vícerozměrný výstup např. $\dim(y(t)) = ny$, pak $\zeta(t)$ by byla matice $n\zeta/ny$), jehož prvky jsou jistě nekorelované s náhodným procesem $v(t)$. Tuto vlastnost můžeme využít k odhadu parametrů v (7.1.1). Požadujeme, aby

$$\frac{1}{N}\sum_{t=1}^{N} \zeta(t)v(t) = \frac{1}{N}\sum_{t=1}^{N} \zeta(t)[y(t) - \varphi^T(t)\Theta] = 0 \tag{7.1.9}$$

pro $N \to \infty$. Dostali jsme soustavu lineárních rovnic pro neznámou Θ. Jestliže $n\zeta = n\Theta$ (a tedy i $n\zeta = n\varphi$), pak z (7.1.9) snadno dostaneme tzv. základní IV odhad Θ, a to

$$\hat{\Theta} = [\frac{1}{N}\sum_{t=1}^{N}\zeta(t)\varphi^T(t)]^{-1}[\frac{1}{N}\sum_{t=1}^{N}\zeta(t)y(t)] \tag{7.1.10}$$

který může být v maticovém zápisu obdobném k metodě nejmenších čtverců (4.1.7) zapsán jako

$$\hat{\Theta} = (\Psi^T\Phi)^{-1}\Psi^T Y \tag{7.1.11}$$

kde řádky matice Ψ jsou dány vektorem přídavné proměnné $\zeta(t)$ v po sobě jdoucích časových okamžicích t, obdobně jak je tomu u matice Φ a vektoru regresorů $\varphi(t)$.
Samozřejmě předpokládáme, že inverze existuje. Prvky vektoru přídavné proměnné $\zeta(t)$ jsou obyčejně nazývány instrumenty. Mohou být vybrány různě, jak ukážeme v následující podkapitole, ale vždy s podmínkou na zajištění konsistentního odhadu parametrů. Vektor přídavné proměnné (vektor instrumentů) by tedy měl splňovat

$$E[\zeta(t)\varphi^T(t)] \sim \frac{1}{N}\sum_{t=1}^{N}\zeta(t)\varphi^T(t) \tag{7.1.12}$$

je nonsingulární matice a

$$E[\zeta(t)v(t)] \sim \frac{1}{N}\sum_{t=1}^{N}\zeta(t)v(t) = 0 \tag{7.1.13}$$

Tyto vztahy znamenají, že instrumenty musí být korelované s regresory, ale nesmí být korelovány se šumem.

7.2 Výběr přídavné proměnné

Předpokládejme strukturu modelu (7.1.2), tedy

$$A(q^{-1})y(t) = B(q^{-1})u(t) + \epsilon(t) \tag{7.2.1}$$

a systém typu (7.1.5)

$$A_0(q^{-1})y(t) = B_0(q^{-1})u(t) + v(t)$$

kde koeficienty polynomů $A_0(q^{-1})$ a $B_0(q^{-1})$ tvoří vektor Θ_0. Přirozená možnost, jak generovat instrumenty je následující

$$\zeta(t) = K(q^{-1})[-x(t-1), -x(t-2), \ldots, x(t-na), u(t-1), \ldots, u(t-nb)]^T \tag{7.2.2}$$

kde $K(q^{-1})$ je lineární filtr a $x(t)$ je výstup lineárního systému na jehož vstupu je $u(t)$

$$N(q^{-1})x(t) = M(q^{-1})u(t) \tag{7.2.3}$$

kde

$$N(q^{-1}) = 1 + n_1q^{-1} + \ldots + n_{nn}q^{-nn}$$

$$M(q^{-1}) = m_0 + m_1 q^{-1} + \ldots + m_{nm} q^{-nm}$$

Je zřejmé, že $\zeta(t)$ je získáno z minulých vstupů, které jsou lineárně filtrovány, což lze vyjádřit jako

$$\zeta(t) = \zeta(t, u(t-1), u(t-2), \ldots)$$

Jestliže vstup je generován v otevřené smyčce, pak nezávisí na $v(t)$ ze systému a (7.1.13) platí. Protože vektory φ a ζ jsou generovány ze stejné vstupní sekvence, můžeme očekávat, že (7.1.12) by mělo též být splněno. Uveďme nyní nějaké speciální případy volby filtrů $N(q^{-1})$, $M(q^{-1})$ a $K(q^{-1})$.

Příklad 7.2.1 Konkrétně bychom mohli postupovat např. tímto způsobem. Nejdříve použít metodu nejmenších čtverců na odhad parametrů v $A(q^{-1})$ a $B(q^{-1})$ modelu (7.2.1) a tento odhad využít pro stanovení $N(q^{-1})$ a $M(q^{-1})$ v (7.2.3). Dále položením $K(q^{-1}) = 1$ pak dostaneme z (7.2.2) vektor instrumentů.

Příklad 7.2.2 Jiná jednoduchá možnost je následující. Volme $N(q^{-1}) = 1$ a $M(q^{-1}) = -q^{-nb}$. Pak po přeuspořádání prvků v ζ dostaneme

$$\zeta(t) = [u(t-1), u(t-2), \ldots, u(t-na-nb)]^T$$

Poznamenejme, že přeuspořádání vektoru instrumentů nemá žádný vliv na odhad. Změna instrumentů z $\zeta(t)$ na $T\zeta(t)$, kde T je nonsingulární matice, nezmění parametrický odhad (7.1.10).

Poznámka . Uvažujme systém ve formě ARMAX, tj.

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t)$$

Metoda přídavné proměnné je vhodná pro nalezení nestranného odhadu parametrů polynomů $A(q^{-1})$, $B(q^{-1})$. Parametry polynomu $C(q^{-1})$ nelze obecně najít. Avšak, v literatuře je věnováno hodně pozornosti tzv. modelu s výstupní chybou, který již byl krátce diskutován a ilustrován v kapitole 2. *Model s výstupní chybou* je ARMAX model, kde platí rovnost

$$C(q^{-1}) = A(q^{-1}) \tag{7.2.4}$$

tzn. že identifikací polynomu $A(q^{-1})$ metodou přídavné proměnné získáme automaticky i odhad polynomu $C(q^{-1})$.

7.3 Yule-Walkerovy rovnice

V této části se budeme zabývat problémem, jak použít metodu přídavné proměnné na odhad parametrů časové řady a jak souvisí tzv. Yule-Walkerovy rovnice s technikou přídavné proměnné. Uvažujme problém odhadu autoregresních parametrů skalárního ARMA procesu definovaného již v 5. kapitole jako speciální případ obecné struktury modelu. Nechť tedy

$$A(q^{-1})y(t) = C(q^{-1})e(t) \tag{7.3.1}$$

kde

$$A(q^{-1}) = 1 + a_1 q^{-1} + \ldots + a_{na} q^{-na}$$

$$C(q^{-1}) = 1 + c_1 q^{-1} + \ldots + c_{nc} q^{-nc}$$
$$E[e(t)] = 0, \qquad E[e(t)e(s)] = \lambda^2 \delta_{t,s}$$

Dále definujme

$$r_k \stackrel{\triangle}{=} E[y(t)y(t-k)] \quad k = 0, \pm 1, \pm 2, ... \tag{7.3.2}$$

Všimněme si, že pro $k > nc$ platí

$$E[C(q^{-1})e(t)y(t-k)] = 0 \tag{7.3.3}$$

což lze považovat za základní vlastnost umožňující odvození Yule-Walkerových rovnic.
Nyní, vynásobíme-li obě strany rovnice (7.3.1) veličinou $y(t-k)$ a provedeme-li operaci ustřednění, dostaneme

$$r_k + a_1 r_{k-1} + ... + a_{na} r_{k-na} = 0 \quad \text{pro} \quad k = nc+1, nc+2, ... \tag{7.3.4}$$

To je soustava lineárních rovnic pro parametry $\{a_i\}$ (koeficienty autoregresního polynomu). Rovnice (7.3.4) jsou nazývány Yule-Walkerovy rovnice. Zapišme prvních m rovnic maticovým způsobem

$$Ra = -r \tag{7.3.5}$$

kde

$$a = [a_1, a_2, ..., a_{na}]^T$$
$$r = [r_{nc+1}, ..., r_{nc+m}]^T$$

$$R = \begin{bmatrix} r_{nc} & \cdots & r_{nc+1-na} \\ \vdots & & \vdots \\ r_{nc+m-1} & \cdots & r_{nc+m-na} \end{bmatrix}$$

Počet rovnic m volíme jako celé číslo tak, aby platil vztah $m \geq na$.

Lze dokázat, že matice R má plnou hodnost. Prvky kovarianční matice R a vektoru r lze odhadnout z množiny pozorování $\{y(1), y(2), ..., y(N)\}$. Nechť $\hat{R}$ a $\hat{r}$ označují odhady R a r. Pak $\hat{a}$, odhad a, dostaneme ze vztahu

$$\hat{R}\hat{a} = -\hat{r} \tag{7.3.6}$$

a bude konsistentní. Poznamenejme, že pro

$m = na$	je $\hat{a}$ nazýván minimální Yule-Walkerův odhad.
$m > na$	je přeurčený systém rovnic a musíme hledat $\hat{a}$ ve smyslu nejmenších čtverců, minimalizující $\parallel \hat{R}\hat{a} + \hat{r} \parallel_Q^2$. V tomto případě $\hat{a}$ je nazýván přeurčený Yule-Walkerův odhad. Odhad, který minimalizuje $\parallel \hat{R}\hat{a} + \hat{r} \parallel_Q^2$ pro nějakou pozitivně definitní matici $Q \neq I$, je nazýván vážený přeurčený Yule-Walkerův odhad.

Zřejmě se jedná o jistý typ korelační techniky, se kterou jsme se setkali v této kapitole pod názvem metoda přídavné proměnné. Jestliže totiž zapíšeme $\hat{R}$ a $\hat{r}$ takto

$$\hat{R} = \frac{1}{N - n_{\max}} \sum_{t=n_{\max}}^{N} \begin{bmatrix} y(t-nc-1) \\ \vdots \\ y(t-nc-m) \end{bmatrix} [y(t-1), \ldots, y(t-na)] \tag{7.3.7}$$

$$\hat{r} = \frac{1}{N - n_{\max}} \sum_{t=n_{\max}}^{N} \begin{bmatrix} y(t-nc-1) \\ \vdots \\ y(t-nc-m) \end{bmatrix} y(t) \tag{7.3.8}$$

kde $n_{\max} = \max\{nc + m - 1, na + 1\}$, pak odhad $\hat{a}$ může být interpretován jako odhad získaný technikou přídavné proměnné pro systém

$$y(t) = -[y(t-1), \ldots, y(t-na)]a + C(q^{-1})e(t)$$

a vektor přídavné proměnné

$$\zeta(t) = [-y(t-nc-1), \ldots, -y(t-nc-m)]^T$$

Yule-Walkerovy rovnice nám tedy umožňují nalézt odhad vektoru a, aniž známe koeficienty polynomu $C(q^{-1})$. Podobně jako u metody přídavné proměnné se nedozvíme jaké jsou vlastnosti šumu. To je ale pochopitelné, protože jsme ukázali na úzký vztah mezi těmito dvěma přístupy.

Na závěr této podkapitoly ještě ukážeme, jak budou výše uvedené vztahy zjednodušeny pro autoregresní proces. V tomto případě lze rovněž snadno získat odhad λ^2. Můžeme totiž zapsat

$$\lambda^2 \stackrel{\triangle}{=} E[e^2(t)] = E[e(t)A(q^{-1})y(t)] \tag{7.3.9}$$

protože pro autoregresní proces platí

$$A(q^{-1})y(t) = e(t) \tag{7.3.10}$$

Protože $\{e(t)\}$ je bílý šum, tedy platí $E[A(q^{-1})y(t)e(t)] = E[y(t)e(t)]$, výraz ze (7.3.9) můžeme dále upravit

$$\begin{aligned} E[e(t)A(q^{-1})y(t)] &= E[e(t)y(t)] = E[A(q^{-1})y(t)y(t)] \\ &= r_0 + a_1 r_1 + \ldots + a_{na} r_{na} \\ &= \lambda^2 \end{aligned} \tag{7.3.11}$$

což lze dále psát

$$\begin{aligned} -r_0 = -E[y(t)y(t)] &= -E[(-a_1 y(t-1) - a_2 y(t-2) - \ldots + a_{na} y(t-na) + e(t))y(t)] \\ &= a_1 r_1 + a_2 r_2 + \ldots + a_{na} r_{na} - \lambda^2 \end{aligned}$$

Jednokroková korelace je pak, vzhledem ke skutečnosti $E[e(t)y(t-1)] = 0$, dána

$$\begin{aligned} -r_1 = -E[y(t)y(t-1)] &= -E[(-a_1 y(t-1) - a_2 y(t-2) - \ldots - a_{na} y(t-na) + e(t))y(t-1)] \\ &= a_1 r_0 + a_2 r_1 + \ldots + a_{na} r_{na-1} \end{aligned}$$

Dvoukroková korelace je

$$\begin{aligned} -r_2 = -E[y(t)y(t-2)] &= -E[(-a_1 y(t-1) - a_2 y(t-2) - \ldots - a_{na} y(t-na) + e(t))y(t-2)] \\ &= a_1 r_1 + a_2 r_0 + \ldots + a_{na} r_{na-2} \end{aligned}$$

a tak můžeme pokračovat až k na-krokové korelaci

$$-r_{na} = -E[y(t)y(t-na)] = -E[(-a_1y(t-1) - a_2y(t-2) - \ldots - a_{na}y(t-na) + e(t))y(t-na)]$$
$$= a_1r_{na-1} + a_2r_{na-2} + \ldots + a_{na}r_0$$

Pak konsistentní odhad parametrů $\{a_i\}$ a λ^2 AR procesu lze najít z následující soustavy rovnic, vycházející z předchozích rovnic pro korelaci výstupu, tj.

$$\begin{bmatrix} -1 & r_1 & r_2 & \cdots & r_{na} \\ 0 & r_0 & r_1 & \cdots & r_{na-1} \\ 0 & r_1 & r_0 & \cdots & r_{na-2} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & r_{na-1} & r_{na-2} & \cdots & r_0 \end{bmatrix} \begin{bmatrix} \lambda^2 \\ a_1 \\ \vdots \\ a_{na} \end{bmatrix} = \begin{bmatrix} -r_0 \\ -r_1 \\ -r_2 \\ \vdots \\ -r_{na} \end{bmatrix} \tag{7.3.12}$$

přičemž r_i jsou korelace, které mohou být snadno odhadnuty z dostupných měření $\{y(1), \ldots, y(N)\}$ dle vztahu

$$\hat{r}_i = \frac{1}{N-i} \sum_{t=i+1}^{N} y(t)y(t-i) \tag{7.3.13}$$

7.4 Modifikované verze metody přídavné proměnné

Metoda přídavné proměnné zahrnuje řadu modifikací. Základní verze, která byla uvedena v podkapitole 7.1 může být zobecněna následujícím způsobem. Předpokládejme, že $n\zeta \geq n\Theta$, $F(q^{-1})$ je nějaký asymptoticky stabilní filtr a $\|x\|_Q^2 = x^TQx$, kde Q je pozitivně definitní váhová matice. Pak

$$\hat{\Theta} = \arg\min_{\Theta} \|[\sum_{t=1}^{N} \zeta(t)F(q^{-1})\varphi^T(t)]\theta - [\sum_{t=1}^{N} \zeta(t)F(q^{-1})y(t)]\|_Q^2 \tag{7.4.1}$$

je odhad podle metody rozšířené přídavné proměnné. Jestliže $n\zeta = n\Theta$, $F(q^{-1}) = 1$ a $Q = 1$ pak odhad (7.4.1) přejde na základní verzi metody z 7.1, vztah (7.1.10).
Při podrobnějším pohledu na (7.4.1) zjistíme, že se jedná o řešení přeurčené soustavy lineárních rovnic ve smyslu vážených nejmenších čtverců. K dispozici je $n\zeta$ rovnic pro $n\Theta$ neznámých. Explicitní řešení (7.4.1) lze vyjádřit následujícím způsobem:

$$\hat{\Theta} = (R_N^TQR_N)^{-1}R_N^TQr_N \tag{7.4.2}$$

kde

$$R_N = \frac{1}{N} \sum_{t=1}^{N} \zeta(t)F(q^{-1})\varphi^T(t) \tag{7.4.3}$$

$$r_N = \frac{1}{N} \sum_{t=1}^{N} \zeta(t)F(q^{-1})y(t) \tag{7.4.4}$$

Pro praktický výpočet odhadu parametrů se však nepoužívá přímo výraz (7.4.2), ale numericky stabilnější verze.

Na závěr této podkapitoly se ještě krátce věnujme teoretické analýze metody. Za poměrně slabých předpokladů je možné dokázat, že metoda přídavné proměnné produkuje konsistentní odhad (to jest $\lim_{N\to\infty} \hat{\Theta} = \Theta_0$), jestliže

– matice R má plnou hodnost ($= n\Theta$) a

– $E\zeta(t)F(q^{-1})v(t) = 0$,

kde $R = E\zeta(t)F(q^{-1})\tilde{\varphi}^T(t) = \lim_{N\to\infty} R_N = E\zeta(t)F(q^{-1})\varphi^T(t)$. Vektor $\tilde{\varphi}(t)$ je část $\varphi(t)$ "bez šumu". Jestliže $\varphi(t)$ obsahuje zpožděné vstupy a výstupy a porucha $v(\cdot)$ je odečtena od prvků $\varphi(t)$, pak získáme $\tilde{\varphi}(t)$ "bez šumu". To můžeme dokumentovat na následujícím příkladu. Uvažujme systém

$$A(q^{-1})y(t) = B(q^{-1})u(t) + v(t)$$

pro který

$$\varphi^T(t) = [-y(t-1), ..., -y(t-na), u(t-1), ..., u(t-nb)]^T$$

pak $\tilde{\varphi}^T(t)$ je

$$\tilde{\varphi}^T(t) = [-x(t-1), ..., -x(t-na), u(t-1), ..., u(t-nb)]^T$$

kde $x(t)$ je část výstupu „bez šumu“, která je dána

$$A(q^{-1})x(t) = B(q^{-1})u(t)$$

Kovarianční matice odhadu, řekněme P_{IV}, samozřejmě závisí na výběru $\zeta(t), F(q^{-1})$ a Q. Proto můžeme též zkoumat, pro jaký výběr dostaneme minimální matici P_{IV}. K tomu nám slouží tzv. optimální metoda přídavné proměnné.

7.5 Shrnutí

Bylo ukázáno, že metoda přídavné proměnné je vhodný nástroj pro odhad parametrů dynamického systému i pro situace, kdy systém obsahuje barevný šum. Základní idea metody přídavné proměnné je blízká postupu, který vede na tzv. Yule-Walkerovy rovnice. Kromě základní metody byla v této kapitole naznačena i složitější, modifikovaná verze a na příkladech byly ukázány možnosti volby přídavné proměnné.

Pro hlubší studium metody přídavné proměnné (IVM) lze doporučit [19], [45]-[47].

Kapitola 8

Rekurzivní metody identifikace

8.1 Úvod

Jak dobře víme, identifikace se zabývá hledáním postupů k nalezení modelu systému z experimentálních dat. Doposud jsme v podstatě rozlišovali metody identifikace na neparametrické a parametrické. Nicméně soubor naměřených dat byl v obou případech jednorázově zpracován. Proto v takových případech mluvíme o jednorázové identifikaci, která je charakteristická tím, že nejdříve jsou naměřena všechna data a pak jsou najednou zpracována. Někdy je takový postup označován také jako off-line identifikace, tedy identifikace neprobíhající "průběžně v čase". Pod pojmem rekurzivní identifikace či rekurzivní identifikační metody se rozumí výpočet parametrů, který zpracovává data rekurzivně. Někdy je též používán pojem on-line. To znamená, že jestliže odhad $\hat{\Theta}(t-1)$ je založen na datech až do okamžiku $t-1$, pak odhad $\hat{\Theta}(t)$ je vypočítán na základě $\hat{\Theta}(t-1)$ a využitím nové další informace, měření v čase t. Celý 2. díl těchto skript bude založen na rekurzivních výpočtech, rekurzivním zpracování měření. Poznamenejme, že je samozřejmě rozdíl mezi rekurzivním zpracováním dat a rekurzivním zpracováním dat v reálném čase. Zatímco v prvním případě můžeme mít na mysli postup, kdy nejprve naměříme soubor dat, a pak provedeme rekurzivní výpočet parametrů (data postupně zpracujeme), tak ve druhém případě jednoznačně požadujeme v reálném čase získaná data též okamžitě v reálném čase zpracovat. Nicméně v obou případech se jedná o rekurzivní postup. Z výše uvedeného vyplývá, že identifikační metody či spíše způsoby zpracování dat můžeme též dělit na jednorázové a rekurzivní. V této kapitole se budeme věnovat právě rekurzivním postupům.

Rekurzivní identifikační metody mají následující obecné vlastnosti:

- Představují jádro (centrální část) adaptivních systémů používaných k automatickému řízení nebo pro zpracování signálů, kdy řízení a filtrace je založena na aktuálním modelu reálného systému.
- Požadavky na paměť jsou výrazně menší než v případě jednorázové identifikace, protože data jsou v případě on-line identifikace okamžitě zpracována.
- Jsou vhodné pro práci v reálném čase, např. sledování proměnných parametrů v čase je zajištěno pouhou modifikací základních algoritmů určených pro odhad konstantních parametrů.
- Jsou nedílnou součástí metod zabývajících se hledáním významných poruch či změn na sledovaném reálném systému.

Vraťme se k výše uvedeným bodům poněkud podrobněji. Většina adaptivních systémů je založena na rekurzivní identifikaci, a to buď explicitní, kdy cílem je získat model reálného systému nebo implicitní, kdy cílem je získat přímo model regulátoru nebo procesoru (systému na zpracování signálu). Pak samozřejmě v každém okamžiku musí být dostupný aktuální model. V případě, že systém je t-variantní, pak i odhadované parametry regulátoru nebo procesoru se budou v čase měnit. Informační tok, v tomto případě reálný systém (proces) $\rightarrow$ model$\rightarrow$ regulátor, se bude neustále opakovat, a proto říkáme, že regulátor se adaptuje na měnící se podmínky (charakteristiky) řízeného procesu. Výraznou vlastností rekurzivních identifikačních metod pracujících v reálném čase je, že množství paměti pro data se s časem nezvyšuje. Požadavky na paměť jsou tedy malé a konstantní. Jestliže budeme aplikovat rekurzivní algoritmy navržené pro systémy s konstantními parametry na systémy s proměnnými parametry, pak samozřejmě $\hat{\Theta}(t)$ nebude konvergovat pro t blížící se nekonečnu, tak jak jsme doposud byli zvyklí u jednorázových metod. Rovněž významnou roli hrají rekurzivní algoritmy v metodách zabývajících se sledováním systémů, které čas od času výrazně mění své charakteristiky. K zjištění takových změn, ať již se jedná o změny dynamiky systému nebo změny vlastností šumu, se používají algoritmy detekce chyb. Pro tyto účely je nutno rekurzivní identifikační metody upravit tak, aby vyhovovaly specifické úloze např. detekce okamžiku změny.

Rekurzivní identifikační metody lze odvodit z jednorázových identifikačních metod, i když je někdy nutné použít jisté aproximace. V následující části budeme též při odvození rekurzivních algoritmů vycházet z off-line přístupů.

8.2 Rekurzivní metoda nejmenších čtverců

V této části se budeme zabývat odvozením rekurzivní metody nejmenších čtverců nebo zkráceně rekurzivních nejmenších čtverců. Vyjdeme z výsledků, které známe již ze 4. a 6. kapitoly. Nicméně pro účely odvození rekurzivního algoritmu si základní jednorázový postup připomeneme.
Předpokládejme pro skalární systém a ztrátovou funkci

$$V_t(\Theta) = \sum_{s=1}^{t} (y(s) - \varphi^T(s)\Theta)^2 \tag{8.2.1}$$

které je analogické kritériu ze 4. kapitoly. Zde však budeme používat t jako označení současného okamžiku místo N, abychom si připravili půdu pro návrh rekurzivní verze metody nejmenších čtverců vhodné pro zpracování dat i v reálném čase. Protože nemusí být vhodné vážit rozdíl mezi naměřenou veličinou $y(s)$ a predikovanou veličinou $\hat{y}(s) = \varphi^T(s)\Theta$ pro každý okamžik stejně použijeme modifikované kritérium

$$V_t(\Theta) = \sum_{s=1}^{t} (y(s) - \varphi^T(s)\Theta)^2 \alpha_s, \tag{8.2.2}$$

kde α_s je váhový koeficient větší než nula a může např. odpovídat inverzi variance šumu měření, tj. inverzi $\lambda_s^2 = E[e(s)e(s)]$, pokud je známa. Úpravou (8.2.2) dostaneme ($\varphi^T(s)\Theta = \Theta^T\varphi(s)$)

$$\begin{aligned} V_t(\Theta) &= \Theta^T[\sum_{s=1}^{t} \alpha_s \varphi(s)\varphi^T(s)]\Theta - \Theta^T[\sum_{s=1}^{t} \alpha_s \varphi(s) y(s)] \\ &- [\sum_{s=1}^{t} \alpha_s y(s)\varphi^T(s)]\Theta + \sum_{s=1}^{t} \alpha_s y^2(s) \end{aligned}$$

Minimalizace ztrátové funkce, tedy derivace $V_t(\Theta)$ podle parametrů, vede na soustavu

$$2\Theta^T[\sum_{s=1}^{t}\alpha_s\varphi(s)\varphi^T(s)] - 2[\sum_{s=1}^{t}\alpha_s\varphi(s)y(s)] = 0$$

z čehož vyplývá

$$[\sum_{s=1}^{t}\alpha_s\varphi(s)\varphi^T(s)]\hat{\Theta}(t) = [\sum_{s=1}^{t}\alpha_s\varphi(s)y(s)] \tag{8.2.3}$$

S předchozím vztahem jsme se v principu setkali již dříve. Jedná se o tzv. normální rovnice, jejichž řešením dostaneme odhad parametrů jednorázovým výpočtem.
Nyní přejděme k odvození rekurzivního algoritmu. Zaveďme

$$\bar{R}(t) \stackrel{\triangle}{=} \sum_{s=1}^{t}\alpha_s\varphi(s)\varphi^T(s) \tag{8.2.4}$$

pak můžeme (8.2.3) zapsat

$$\bar{R}(t)\hat{\Theta}(t) = \sum_{s=1}^{t}\alpha_s\varphi(s)y(s) \tag{8.2.5}$$

Roztržením $\bar{R}(t)$ na současnost a minulost dostaneme

$$\bar{R}(t) = \bar{R}(t-1) + \alpha_t\varphi(t)\varphi^T(t) \tag{8.2.6}$$

Odhad parametrů pak z (8.2.5) a (8.2.6) bude

$$\begin{aligned}
\hat{\Theta}(t) &= \bar{R}^{-1}(t)\left(\sum_{s=1}^{t-1}\alpha_s\varphi(s)y(s) + \alpha_t\varphi(t)y(t)\right) \\
&= \bar{R}^{-1}(t)(\bar{R}(t-1)\hat{\Theta}(t-1) + \alpha_t\varphi(t)y(t)) \\
&= \bar{R}^{-1}(t)\left((\bar{R}(t) - \alpha_t\varphi(t)\varphi^T(t))\hat{\Theta}(t-1) + \alpha_t\varphi(t)y(t)\right)
\end{aligned}$$

z čehož vyplývá

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + \alpha_t\bar{R}^{-1}(t)\varphi(t)[y(t) - \varphi^T(t)\hat{\Theta}(t-1)] \tag{8.2.7}$$

Z předchozího vztahu je zřejmé, že odhad parametrů v čase t, $\hat{\Theta}(t)$ je vypočítán z odhadu parametrů v čase $(t-1)$, $\hat{\Theta}(t-1)$ a vážené chyby predikce. Otázka nyní je, jak se vyvíjí v čase matice $\bar{R}(t)$. Zaveďme $R(t) \stackrel{\triangle}{=} \bar{R}(t)/t$, pak

$$\begin{aligned}
R(t) &= \frac{1}{t}[\bar{R}(t-1) + \alpha_t\varphi(t)\varphi^T(t)] \\
&= \frac{1}{t}[(t-1)R(t-1) + \alpha_t\varphi(t)\varphi^T(t)] \\
&= \frac{t-1}{t}R(t-1) + \frac{\alpha_t}{t}\varphi(t)\varphi^T(t) + \frac{1}{t}R(t-1) - \frac{1}{t}R(t-1)
\end{aligned}$$

což vede na

$$R(t) = R(t-1) + \frac{1}{t}[\alpha_t\varphi(t)\varphi^T(t) - R(t-1)] \tag{8.2.8}$$

Po dosazení za $\bar{R}(t)$ do (8.2.7) dostaneme

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + \frac{\alpha_t}{t} R^{-1}(t)\varphi(t)[y(t) - \varphi^T(t)\hat{\Theta}(t-1)] \tag{8.2.9}$$

Poslední dva vztahy (8.2.9) a (8.2.8) bychom mohli pokládat za rekurzivní algoritmus výpočtu odhadu parametrů. Avšak z těchto rovnic je rovněž zřejmé, že v (8.2.9) potřebujeme invertovat matice R, což je nepříjemná numerická operace. Vraťme se proto zpět k rovnici (8.2.7) a ukažme alternativní postup odvození vedoucí k odstranění této nepříjemnosti. Zaveďme

$$P(t) \stackrel{\triangle}{=} \bar{R}^{-1}(t) = [\sum_{s=1}^{t} \alpha_s \varphi(s)\varphi^T(s)]^{-1} \tag{8.2.10}$$

Pak

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + \alpha_t P(t)\varphi(t)[y(t) - \varphi^T(t)\hat{\Theta}(t-1)] \tag{8.2.11}$$

$$P^{-1}(t) = P^{-1}(t-1) + \alpha_t \varphi(t)\varphi^T(t) \tag{8.2.12}$$

Vztah (8.2.12) můžeme upravit pomocí tzv. maticové inverzní identity

$$[A + BCD]^{-1} = A^{-1} - A^{-1}B[C^{-1} + DA^{-1}B]^{-1}DA^{-1}$$

ze které položením za $A = P^{-1}(t-1),\ B = \varphi(t),\ C = \alpha_t,\ D = \varphi^T(t)$ dostaneme

$$\begin{aligned} P(t) &= [A + BCD]^{-1} \\ P(t) &= P(t-1) - P(t-1)\varphi(t)[\frac{1}{\alpha_t} + \varphi^T(t)P(t-1)\varphi(t)]^{-1}\varphi^T(t)P(t-1) \end{aligned} \tag{8.2.13}$$

Protože inverze se v předchozí rovnici týká skalární veličiny, jedná se zde pouze o dělení číslem. Nyní již zbývá pouze vypočítat $\alpha_t P(t)\varphi(t)$ v rovnici (8.2.11). Pro výpočet využijeme (8.2.13).

$$\begin{aligned} \alpha_t P(t)\varphi(t) &= \alpha_t P(t-1)\varphi(t) - \alpha_t P(t-1)\varphi(t)\varphi^T(t)P(t-1)\varphi(t)/[1/\alpha_t + \varphi^T(t)P(t-1)\varphi(t)] \\ &= P(t-1)\varphi(t)\{\alpha_t - \alpha_t\varphi^T(t)P(t-1)\varphi(t)/[\frac{1}{\alpha_t} + \varphi^T(t)P(t-1)\varphi(t)]\} \end{aligned}$$

Odsud již po snadné úpravě dostaneme

$$\alpha_t P(t)\varphi(t) = P(t-1)\varphi(t)/[\frac{1}{\alpha_t} + \varphi^T(t)P(t-1)\varphi(t)] \tag{8.2.14}$$

Shrňme nyní dosažené výsledky. Pro strukturu modelu

$$y(t) = \varphi^T(t)\Theta + v(t) \tag{8.2.15}$$

která zahrnuje např. systém typu ARX

$$A(q^{-1})y(t) = B(q^{-1})u(t) + e(t) \tag{8.2.16}$$

a identifikační metodu nejmenší čtverce definovanou minimalizací kritéria

$$V_t(\Theta) = \sum_{s=1}^{t} [y(s) - \varphi^T(s)\Theta]^2 \alpha_s \tag{8.2.17}$$

bude jednorázový odhad $\hat{\Theta}(t)$ definován

$$\hat{\Theta}(t) = [\sum_{s=1}^{t} \alpha_s \varphi(s)\varphi^T(s)]^{-1}[\sum_{s=1}^{t} \alpha_s \varphi(s)y(s)] \tag{8.2.18}$$

a rekurzivní odhad popsán vztahy

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + L(t)\epsilon(t) \tag{8.2.19}$$

$$\epsilon(t) = y(t) - \varphi^T(t)\hat{\Theta}(t-1) \tag{8.2.20}$$

$$L(t) = \frac{P(t-1)\varphi(t)}{\frac{1}{\alpha_t} + \varphi^T(t)P(t-1)\varphi(t)} \tag{8.2.21}$$

$$P(t) = P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{\frac{1}{\alpha_t} + \varphi^T(t)P(t-1)\varphi(t)} \tag{8.2.22}$$

Poznamenejme, že $\epsilon(t)$ často bývá označováno jako chyba predikce a vektor $L(t)$ v (8.2.19) je zisk ukazující, jak bude chyba predikce modifikovat jednotlivé prvky vektoru parametrů. Všimněme si také, že matice $P(t)$ s zvyšujícím se t monotónně klesá a pro $t \to \infty$ se bude blížit nule. Pak i zisk estimátoru $L(t)$ bude nulový a odhad parametrů $\hat{\Theta}(t)$ ustane. Tato situace může nastat i v konečném časovém intervalu, a to z důvodu omezené přesnosti výpočtu výpočetních prostředků.

Poznámka . Váhový koeficient α_t může být volen jako libovolné kladné reálné číslo. Pokud však bude zvolen jako

$$\alpha_t = 1/\lambda_t^2$$

kde λ_t^2 je variance šumu, tj. $\lambda^2 = E[e(t)^2]$, pak, po odeznění vlivu počátečních podmínek, je výsledný odhad nejlepší lineární nestranný odhad (BLUE), tak jak byl diskutován v kapitole 4.3, a matice $P(t)$ může být chápána jako kovarianční matice chyby odhadu. Poznamenejme rovněž, že rekurzivní metoda nejmenších čtverců s diskutovanou volbou váhového koeficientu je zjednodušenou verzí Kalmanova filtru, jakožto optimálního estimátoru stavu pro lineární stochastické systémy, konfigurovaného pro odhad konstantních parametrů. Tato skutečnost je detilně diskutována v návazném díle těchto skript.

Poznámka . Abychom mohli předchozí algoritmus použít, musíme stanovit počáteční podmínky. Účinek počátečních podmínek $\Theta(0)$, $P(0)$, na odhad $\hat{\Theta}(t)$ bude zřejmý z následující analýzy. Uvažujme algoritmus (8.2.19)-(8.2.22) s $\alpha_t = 1$ pro všechna t. Z rovnice (8.2.12) platí

$$P^{-1}(t) = P^{-1}(0) + \sum_{s=1}^{t} \varphi(s)\varphi^T(s)$$

Zavedením pomocné veličiny $x(t)$ ve tvaru

$$x(t) \stackrel{\triangle}{=} P^{-1}(t)\hat{\Theta}(t) \tag{8.2.23}$$

a využitím (8.2.11), (8.2.12), (8.2.20) dostaneme

$$\begin{aligned} x(t) &= P^{-1}(t)\hat{\Theta}(t-1) + \varphi(t)\epsilon(t) \\ &= [P^{-1}(t-1) + \varphi(t)\varphi^T(t)]\hat{\Theta}(t-1) + \varphi(t)[y(t) - \varphi^T(t)\hat{\Theta}(t-1)] \\ &= x(t-1) + \varphi(t)y(t) \\ &= x(0) + \sum_{s=1}^{t} \varphi(s)y(s) \end{aligned}$$

Z (8.2.23) můžeme vyjádřit $\hat{\Theta}(t)$ jako

$$\hat{\Theta}(t) = P(t)x(t)$$

a po dosazení za $P(t)$ z (8.2.12) a $x(t)$ z předposledního vztahu získáme

$$\hat{\Theta}(t) = [P^{-1}(0) + \sum_{s=1}^{t} \varphi(s)\varphi^T(s)]^{-1}[P^{-1}(0)\hat{\Theta}(0) + \sum_{s=1}^{t} \varphi(s)y(s)] \tag{8.2.24}$$

který je v úzkém vztahu s jednorázovou metodou nejmenších čtverců. Nyní je již zřejmé, že pro $P^{-1}(0)$ malé bude odhad přibližně roven jednorázovému odhadu. Odtud plyne často doporučovaný postup volit v případě, že nemáme apriorní informace o parametrech, volit

$$\hat{\theta}(0) = 0 \qquad a \qquad P(0) = \rho I$$

kde ρ je nějaké „velké“ číslo.

Na závěr této podkapitoly se krátce zmíníme o jiné možnosti jednoduché volby váhového koeficientu α_s v kritériu při požadavku exponenciální rychlosti zapomínání starých dat. Tento požadavek je motivován situací, kdy řešitel úlohy identifikace předpokládá možnost *pomalé* změny parametrů (doposud jsme vždy předpokládali konstantní parametry), a proto starým datům přiřazuje menší vypovídací schopnost o aktuální hodnotě parametrů. Pak α_s ve (8.2.12) můžeme definovat

$$\alpha_s = \tau^{t-s}$$

kde τ je tzv. faktor zapomínání v tomto případě konstantní a jeho hodnota je menší nebo rovna jedné. Často se vybírá z intervalu 0.95-0.99. Volba faktoru zapomínání může vycházet např. z apriorní znalosti o rychlosti změny (tj. o časové konstantě) hledaných parametrů. Obecně se dá říci, že čím menší je τ, tím rychlejší je zapomínání.
Tím, že takto zvolený parametr α_s explicitně závisí na posledním časovém okamžiku t, nelze jej jednoduše dosadit do rekurzivního algoritmu metody nejmenších čtverců (8.2.19)-(8.2.22). Je nutné tedy algoritmus znovu odvodit. Nicméně, odvození rekurzivní metody nejmenších čtverců s váhovým koeficientem závislým na t je do značné míry analogické k odvození algoritmu (8.2.19)-(8.2.22), a proto zde uvedeme jen finální algoritmus rekurzivní metody nejmenších čtverců s exponenciálním zapomínáním dat, který je

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + L(t)\epsilon(t) \tag{8.2.25}$$

$$\epsilon(t) = y(t) - \varphi^T(t)\hat{\Theta}(t-1)$$

$$L(t) = P(t)\varphi(t) = P(t-1)\varphi(t)/[\tau + \varphi^T(t)P(t-1)\varphi(t)]$$

$$P(t) = \{P(t-1) - P(t-1)\varphi(t)\varphi^T(t)P(t-1)/[\tau + \varphi^T(t)P(t-1)\varphi(t)]\}/\tau$$

Všimněme si, že na rozdíl od neváženého verze, může matice $P(t)$ i růst s přibývajícím počtem měření.
V současnosti jsou známy i další postupy pro odhad pomalu se měnících parametrů, např. tzv. směrové zapomínání nebo Kalmanův filtr, který bude detailně popsán ve 2. díle skript a je založen na odlišném popisu neznámých parametrů (parametry jsou uvažovány jako náhodné veličiny).

8.3 Rekurzivní metoda přídavné proměnné

V kapitole 7. jsme se zabývali především základní variantou metody přídavné proměnné. Výpočet odhadu parametrů vycházel ze vztahu

$$\hat{\Theta}(t) = [\sum_{s=1}^{t} \zeta(s)\varphi^T(s)]^{-1}[\zeta(s)y(s)] \tag{8.3.1}$$

Je zřejmé, že tento výraz je velmi strukturálně podobný s (8.2.3) pro $\alpha_s = 1$ a analogickým postupem jako u odvození rekurzivních nejmenších čtverců bychom dostali následující rekurzivní algoritmus odhadu parametrů.

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + L(t)\epsilon(t) \tag{8.3.2}$$

$$\epsilon(t) = y(t) - \varphi^T(t)\hat{\Theta}(t-1)$$

$$L(t) = P(t)\zeta(t) = P(t-1)\zeta(t)/[1 + \varphi^T(t)P(t-1)\zeta(t)]$$

$$P(t) = P(t-1) - P(t-1)\zeta(t)\varphi^T(t)P(t-1)/[1 + \varphi^T(t)P(t-1)\zeta(t)]$$

Jediný rozdíl mezi jednorázovou i rekurzivní metodou nejmenších čtverců a metodou přídavné proměnné rovněž v obou variantách je v záměně vektoru $\varphi(t)$ za vektor přídavné proměnné $\zeta(t)$. Ale $\varphi^T(t)$ zůstává stejné. Pro rekurzivní algoritmus můžeme provést např. následující volbu vektoru přídavné proměnné

$$\zeta^T(t) = [-y_m(t-1), \ldots, -y_m(t-na), u(t-1), \ldots, u(t-nb)]$$

kde $y_m(t)$ je výstup deterministického systému

$$y_m(t) + \hat{a}_1(t)y_(t-1) + \ldots + \hat{a}_{na}(t)y_m(t-na) = \hat{b}_1(t)u(t-1) + \ldots + \hat{b}_{nb}(t)u(t-nb)$$

s proměnnými parametry, které generuje rekurzivní algoritmus identifikace.

8.4 Rekurzivní metoda chyby predikce

Metodou chyby predikce jsme se již zabývali v 6. kapitole. Cílem této subkapitoly je získat rekurzivní algoritmus výpočtu odhadu metodou chyby predikce pro obecný model, s jedním vstupem a jedním výstupem. Toto omezení není samozřejmě nutné, ale pro zlepšení porozumění při odvození je vhodné. Začneme volbou ztrátové funkce. Uvažujme

$$V_t(\Theta) = \frac{1}{2}\sum_{s=1}^{t} \lambda^{t-s}\epsilon^2(s) \tag{8.4.1}$$

Pro $\lambda = 1$ se ztrátová funkce redukuje na standardní tvar se stejnou vahou pro všechna data. Takto uvažujeme exponenciální zapomínání dat. Poznamenejme, že odhad $\hat{\Theta}_t$ získaný z jednorázové identifikace minimalizující (8.4.1) nelze najít analyticky, kromě speciálních případů zkoumaných již dříve v textu. Proto musí být prováděna numerická optimalizace. To je rovněž důvod, proč nelze odvodit rekurzivní algoritmus přesně odpovídající jednorázové identifikaci. Bude potřeba provádět jisté aproximace. Předpokládejme, že $\hat{\Theta}(t-1)$ minimalizuje $V_{t-1}(\Theta)$ a

že minimum $V_t(\Theta)$ je blízko $\hat{\Theta}(t-1)$. Pak je přijatelné aproximovat $V_t(\Theta)$ Taylorovou řadou okolo bodu $\hat{\Theta}(t-1)$. Proveďme Taylorův rozvoj a ponechme pouze první tři členy.

$$\begin{aligned} V_t(\Theta) &\sim V_t(\hat{\Theta}(t-1)) + V_t'^T(\hat{\Theta}(t-1))(\Theta - \hat{\Theta}(t-1)) \\ &+ \frac{1}{2}(\Theta - \hat{\Theta}(t-1))^T V_t''(\hat{\Theta}(t-1))(\Theta - \hat{\Theta}(t-1)) \end{aligned} \tag{8.4.2}$$

Pravá strana (8.4. 2) je kvadratická funkce vzhledem k Θ. Při hledání extrému postupujme standardním způsobem, tedy první derivaci aproximace v (8.4. 2) položíme rovnu nule

$$V_t'^T(\hat{\Theta}(t-1)) + V_t''(\hat{\Theta}(t-1))[\Theta - \hat{\Theta}(t-1)] = 0 \tag{8.4.3}$$

Nový optimální odhad v čase t pak můžeme stanovit z (8.4.3)

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) - [V_t''(\hat{\Theta}(t-1))]^{-1}[V_t'(\hat{\Theta}^T(t-1))]^T \tag{8.4.4}$$

Tento optimalizační postup odpovídá tzv. Newton-Raphson algoritmu pro hledání extrému funkce. Nicméně rekurzivní vztah musíme získat i pro derivace ztrátové funkce $V_t(\Theta)$. Vyjděme z (8.4.1) a postupně derivujme $V_t(\Theta)$

$$V_t(\Theta) = \lambda V_{t-1}(\Theta) + \frac{1}{2}\epsilon^2(t)$$

$$V_t'(\Theta) = \lambda V_{t-1}'(\Theta) + \epsilon(t)\frac{\partial \epsilon(t)}{\partial \Theta} = \lambda V_{t-1}'(\Theta) + \epsilon(t)\epsilon'(t)$$

$$V_t''(\Theta) = \lambda V_{t-1}''(\Theta) + \epsilon(t)\epsilon(t)'' + \epsilon'(t)\epsilon'(t)$$

Proveďme následující aproximace

$$V_{t-1}'(\hat{\Theta}(t-1)) = 0 \tag{8.4.5}$$

$$V_{t-1}''(\hat{\Theta}(t-1)) = V_{t-1}''(\hat{\Theta}(t-2)) \tag{8.4.6}$$

$$\epsilon(t)\epsilon''(t) = 0 \tag{8.4.7}$$

Důvod pro zavedení (8.4.5) je v předpokladu, že funkce $V_{t-1}(\Theta)$ nabývá minimální hodnoty v bodě $\hat{\Theta}(t-1)$. Vztah (8.4.6) říká, že druhá derivace se skoro nemění s Θ a konečně (8.4.7) podporuje skutečnost, že pro hledané parametry (parametry systému) bude $\epsilon(t)$ bílý šum ($\hat{\Theta} \to \Theta_0 \Rightarrow \epsilon(t, \Theta_0)$ je bílý šum) a $E\epsilon(t)\epsilon''(t) = 0$. Použitím (8.4.5)-(8.4.7) na výše vypočtené V_t' a V_t'' dostaneme

$$V_t'(\hat{\Theta}(t-1)) = -\epsilon(t, \hat{\Theta}(t-1))\psi(t, \hat{\Theta}(t-1)) \tag{8.4.8}$$

$$V_t''(\hat{\Theta}(t-1)) = \lambda V_{t-1}''(\hat{\Theta}(t-2)) + \psi(t, \hat{\Theta}(t-1))\psi^T(t, \hat{\Theta}(t-1)) \tag{8.4.9}$$

kde

$$\psi(t, \Theta) = -[\frac{\partial \epsilon(t, \Theta)}{\partial \Theta}]^T \tag{8.4.10}$$

Vztah pro odhadované parametry (8.4.4) můžeme již konkretizovat dosazením z (8.4.8)-(8.4.10) a tak dostaneme

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + [V_t^{''}(\hat{\Theta}((t-1)]^{-1}\epsilon(t, \hat{\Theta}(t-1))\psi(t, \hat{\Theta}(t-1)) \tag{8.4.11}$$

Zaveďme $\bar{R}(t) \stackrel{\triangle}{=} V_t^{''}(\hat{\Theta}(t-1))$, pak (8.4.11) a (8.4.9) nabude tvar

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + \bar{R}^{-1}(t)\psi(t, \hat{\Theta}(t-1))\epsilon(t, \hat{\Theta}(t-1)) \tag{8.4.12}$$

$$\bar{R}(t) = \lambda\bar{R}(t-1) + \psi(t, \hat{\Theta}(t-1))\psi^T(t, \hat{\Theta}(t-1)) \tag{8.4.13}$$

V tuto chvíli si musíme položit otázku, jak vypočítáme $\psi(t, \hat{\Theta}(t-1))$ a $\epsilon(t, \hat{\Theta}(t-1))$. Pro výpočet těchto veličin již nutně potřebujeme znát konkrétní strukturu modelu. Pro některé struktury vyžaduje výpočet $\epsilon(t, \hat{\Theta}(t-1))$ zpracování všech dat až do času t a podobně pro $\psi(t, \hat{\Theta}(t-1))$. Abychom tyto veličiny mohli počítat pouze z posledních dat, proveďme poslední aproximaci

$$\epsilon(t) \sim \epsilon(t, \hat{\Theta}(t-1))$$

$$\psi(t) \sim -[\epsilon'(t, \hat{\Theta}(t-1))]^T$$

která umožní vyhodnocení veličin $\epsilon(t)$, $\psi(t)$ průběžně v čase. Konkrétní forma a způsob jejich implementace však záleží na aktuální struktuře modelu, jak bude vidět v příkladu 8.4.1.

Další problém, který přináší (8.4.12) a (8.4.13) je opět nekompatibilita těchto rovnic vzhledem k matici $\bar{R}(t)$, protože v (8.4.12) je její inverze. Podobný jev nastal i při odvozování metody rekurzivních nejmenších čtverců. Pro odstranění inverze proto můžeme použít stejný postup jako v podkapitole 8.2.

Zaveďme

$$P(t) = [V_t^{''}(\hat{\Theta}(t-1))]^{-1} \tag{8.4.14}$$

pak z (8.4.9) plyne

$$P^{-1}(t) = \lambda P^{-1}(t-1) + \psi(t)\psi^T(t) \tag{8.4.15}$$

Předchozí vztah můžeme použitím maticové inverzní identity zapsat (podobně jako z (8.2.12) na (8.2.13))

$$P(t) = \{P(t-1) - P(t-1)\psi(t)\psi^T(t)P(t-1)/[\lambda + \psi^T(t)P(t-1)\psi(t)]\}/\lambda \tag{8.4.16}$$

Tím jsme odstranili nutnost provádět v každém kroku inverzi matice dimenze $n\Theta/n\Theta$, protože v (8.4.16) je nutné pouze dělit číslem. Nyní již lze předvést výsledný rekurzivní algoritmus metody chyby predikce:

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + L(t)\epsilon(t) \tag{8.4.17}$$

$$L(t) = P(t)\psi(t) \tag{8.4.18}$$

$$P(t) = \{P(t-1) - P(t-1)\psi(t)\psi^T(t)P(t-1)/[\lambda + \psi^T(t)P(t-1)\psi(t)]\}/\lambda \tag{8.4.19}$$

nebo po úpravě výpočtu ziskového vektoru $L(t)$ podobně jako u nejmenších čtverců

$$L(t) = P(t-1)\psi(t)/[\lambda + \psi^T(t)P(t-1)\psi(t)] \tag{8.4.20}$$

Algoritmus (8.4.17)-(8.4.19) případně (8.4.20) je použitelný na různé struktury modelů. Výpočet $\epsilon(t)$ a $\psi(t)$ bude závislý na konkrétní struktuře. Například pro ARX model bude tento algoritmus totožný s (8.2.19)-(8.2.22) pro nejmenší čtverce. To znamená, že $\psi(t) = \varphi(t)$ a $\epsilon(t) = y(t) - \varphi^T(t)\hat{\Theta}(t-1)$. Pro ARMAX model však bude výpočet těchto veličin složitější. Věnujme se proto podrobně aplikaci odvozeného algoritmu pro ARMAX model.

Příklad 8.4.1 Aplikace rekurzivní metody chyby predikce na ARMAX model.
Předpokládejme ARMAX strukturu

$$A(q^{-1})y(t) = B(q^{-1})u(t) + C(q^{-1})e(t) \tag{8.4.21}$$

kde

$$A(q^{-1}) = 1 + a_1q^{-1} + ... + a_{na}q^{-na}$$
$$B(q^{-1}) = b_1q^{-1} + ... + b_{nb}q^{-nb}$$
$$C(q^{-1}) = 1 + c_1q^{-1} + ... + c_{nc}q^{-nc}$$

Z 6. kapitoly víme, že

$$C(q^{-1})\epsilon(t,\Theta) = A(q^{-1})y(t) - B(q^{-1})u(t) \tag{8.4.22}$$

Parciální derivace $\epsilon(t,\Theta)$ podle parametru a_i, kde $i = 1, ..., na$, splňuje

$$C(q^{-1})\frac{\partial\epsilon(t,\Theta)}{\partial a_i} = y(t-i) \tag{8.4.23}$$

nebo-li

$$\frac{\partial\epsilon(t,\Theta)}{\partial a_i} = \frac{1}{C(q^{-1})}y(t-i)$$

Obdobně parciální derivace podle parametrů b_i, kde $i = 1, ..., nb$, splňuje

$$C(q^{-1})\frac{\partial\epsilon(t,\Theta)}{\partial b_i} = -u(t-i) \tag{8.4.24}$$

a konečně analogicky pro parametry c_i, kde $i = 1, ..., nc$, lze odvodit

$$C(q^{-1})\frac{\partial\epsilon(t,\Theta)}{\partial c_i} = -\epsilon(t-i,\Theta) \tag{8.4.25}$$

Abychom mohli předchozí výsledky zapsat kompaktním způsobem, definujeme filtrované signály

$$y^F(t) \triangleq \frac{1}{C(q^{-1})}y(t) \tag{8.4.26}$$

$$u^F(t) \triangleq \frac{1}{C(q^{-1})}u(t) \tag{8.4.27}$$

$$\epsilon^F(t) \triangleq \frac{1}{C(q^{-1})}\epsilon(t) \tag{8.4.28}$$

pak derivace pro všechny parametry je dána

$$\psi(t,\Theta) = -[y^F(t-1),\ldots,y^F(t-na),-u^F(t-1),\ldots,-u^F(t-nb),-\epsilon^F(t-1),\ldots,-\epsilon^F(t-nc)]^T \tag{8.4.29}$$

Kromě případu, kdy $C(q^{-1}) = 1$ jsou (8.4.22)-(8.4.29) filtry s nekonečnou impulsní odezvou (vstupem je $u(\cdot)$ a $y(\cdot)$ a výstupem $\epsilon(\cdot)$ a $\epsilon'(\cdot)$). Také je zřejmé, že k výpočtu $\epsilon(t,\Theta)$ i $\epsilon'(t,\Theta)$ pro libovolnou hodnotou Θ je zapotřebí zpracovat všechna data až do okamžiku t. Vhodný způsob, jak aproximovat výpočet těchto hodnot je nahradit při výpočtu $\epsilon(t,\Theta)$ neznámé parametry jejich odhady v okamžiku $t-1$, tedy

$$\begin{aligned}\epsilon(t) &= y(t) + \hat{a}_1(t-1)y(t-1) + ... + \hat{a}_n(t-1)y(t-na) \\ &- \hat{b}_1(t-1)u(t-1) - ... - \hat{b}_n(t-1)u(t-nb) \\ &- \hat{c}_1(t-1)\epsilon(t-1) - ... - \hat{c}_n(t-1)\epsilon(t-nc)\end{aligned} \tag{8.4.30}$$

a podobně pro prvky $\psi(t)$ s tím, že odhad lze použít až po okamžik t

$$y^F(t) = y(t) - \hat{c}_1(t)y^F(t-1) - ... - \hat{c}_{nc}(t)y^F(t-nc) \tag{8.4.31}$$

$$u^F(t) = u(t) - \hat{c}_1(t)u^F(t-1) - ... - \hat{c}_{nc}(t)u^F(t-nc) \tag{8.4.32}$$

$$\epsilon^F(t) = \epsilon(t) - \hat{c}_1(t)\epsilon^F(t-1) - ... - \hat{c}_{nc}(t)\epsilon^F(t-nc) \tag{8.4.33}$$

Poznamenejme, že „přesný“ výpočet $\epsilon(s,\cdot)$ a $\epsilon'(s,\cdot)$ by vyžadoval filtraci s pevným $\hat{\Theta}(t-1)$ od $t=1$ do $t=s$. Takto provedeme filtraci pouze jednou s inicializací v každém časovém okamžiku s použitím aktuálního odhadu parametrů $\hat{\Theta}(t-1)$ z předchozích hodnot ϵ a ϵ'. Nyní již můžeme realizovat obecný algoritmus (8.4.17)-(8.4.20), protože veličiny, které souvisí se strukturou modelu ϵ a ψ jsou již známy.

Na závěr poznamenejme, že $\epsilon(t)$ z (8.4.7) bychom mohli počítat „lépe“ než v (8.4.30), protože pro výpočet $\epsilon(t-1)$ v (8.4.30) můžeme použít již odhad $\hat{\Theta}(t-1)$ místo $\hat{\Theta}(t-2)$ atd. To by umožnilo nalézt modifikovaný algoritmus s lepšími vlastnostmi než má výše uvedený.

V dalším příkladu ukážeme, jaký je vztah mezi značně populární metodou pseudolineární regrese a metodou chyby predikce při aplikaci na ARMAX model.

Příklad 8.4.2 Pseudolineární regrese (nebo-li rozšířená metoda nejmenších čtverců) pro ARMAX model.
Zapišme ARMAX model (8.4.21) jako lineární regresi

$$y(t) = \varphi^T(t)\Theta + e(t)$$

kde

$$\varphi(t) = [-y(t-1),\ldots,-y(t-na),u(t-1),\ldots,u(t-nb),e(t-1),\ldots,e(t-nc)]^T$$

$$\Theta = [a_1,\ldots,a_{na},b_1,\ldots,b_{nb},c_1,\ldots,c_{nc}]^T$$

Samozřejmě nelze aplikovat metodu nejmenších čtverců, protože regresory $\{e(t-1), \ldots, e(t-nc)\}$ nejsou známy. Avšak jestliže je nahradíme odhadnutou chybou predikce $\epsilon(t)$, dostaneme rekurzivní algoritmus (pro $\alpha_s = 1$ v kritériu)

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + L(t)\epsilon(t)$$
$$\epsilon(t) = y(t) - \varphi^T(t)\hat{\Theta}(t-1)$$
$$L(t) = P(t-1)\varphi(t)/[1 + \varphi^T(t)P(t-1)\varphi(t)]$$
$$P(t) = P(t-1) - P(t-1)\varphi(t)\varphi^T(t)P(t-1)/[1 + \varphi^T(t)P(t-1)\varphi(t)]$$
$$\varphi(t) = [-y(t-1)... - y(t-n)u(t-1)...u(t-n)\epsilon(t-1)...\epsilon(t-n)]^T$$

Ačkoliv tento algoritmus reprezentující metodu pseudolineární regrese je velmi podobný rekurzivní metodě chyby predikce, je zde však přece rozdíl v absenci filtru ve vektoru regresorů. Rovněž vlastnosti spojené s konvergencí a kvalitou odhadu jsou výrazně horší. Pro zajištění globální konvergence by muselo platit, že

$$Re[\frac{1}{C_0(e^{i\omega})} - \frac{1}{2}] > 0$$

pro všechny ω, což je samozřejmě splněno jen pro některé $C_0(q^{-1})$ [15].

8.5 Metoda stochastické aproximace

Pojem stochastická aproximace byl zaveden ve statistice pro sekvenční odhad parametrů. Nejdříve se pokusíme ukázat hlavní myšlenku stochastické aproximace a její vztah k rekurzivní identifikaci. Uvažujme model

$$y(t) = \varphi^T(t)\Theta + v(t) \tag{8.5.1}$$

kde $y(t)$ a $\varphi(t)$ jsou měřené veličiny, Θ je vektor parametrů, který má být určen a proměnná $v(t)$ je chyba rovnice. Je přirozené vybrat Θ tak, aby variance $v(t)$ byla minimalizována to jest řešíme optimalizační problém

$$\min_{\Theta} V(\Theta)$$

kde

$$V(\Theta) = \frac{1}{2}E[y(t) - \varphi^T(t)\Theta]^2 \tag{8.5.2}$$

Protože $V(\Theta)$ je kvadratická v Θ lze $min\ V(\Theta)$ nalézt řešením

$$[-\frac{d}{d\Theta}V(\Theta)]^T = E\varphi(t)[y(t) - \varphi^T(t)\Theta] = 0$$

Jelikož neznáme hustoty pravděpodobnosti veličin $y(t)$ ani $\varphi(t)$ nelze střední hodnotu ohodnotit. Samozřejmě můžeme střední hodnotu nahradit v tomto smyslu

$$E[f(t)] \approx (1/N)\sum_{t=1}^{N} f(t)$$

a pak vlastně přejít na metodu nejmenších čtverců.

Nyní se věnujme typické formulaci problému řešeného stochastickou aproximací. Nechť $\{e(t)\}$ je sekvence náhodných proměnných se stejným rozložením. Dále předpokládejme, že je dána funkce $Q(x, e(t))$ proměnných $e(t),\ x$ a hledáme řešení rovnice

$$E[Q(x, e(t))] = f(x) = 0,$$

kde E označuje střední hodnotu přes $e(t)$. Řešitel úlohy nezná rozložení $e(t)$. Rovněž přesný tvar funkce Q může být neznámý. Avšak realizace $Q(x, e)$ jsou pozorovány nebo mohou být nějak zjištěny pro vybraná x. Řečeno jinými slovy, uživatel vybere x a dostane realizaci $Q(x, e(t))$. Snadno se lze přesvědčit, že problém nalezení parametrů regresního modelu je speciálním případem této obecné úlohy. Stačí, pokud zavedeme

$$x = \Theta$$
$$e(t) = [y(t), \varphi^T(t)]^T$$
$$Q(x, e(t)) = \varphi(t)[y(t) - \varphi^T(t)\Theta]$$

V tomto případě $e(t)$ je měřeno (nebo známo) a $Q(x, e(t))$ je známá funkce x a $e(t)$, ale rozložení $e(t)$ je neznámé. Vraťme se však k obecné rovnici

$$E[Q(x, e(t))] = f(x) = 0$$

Otázkou zůstává jak generovat posloupnost aproximativních řešení $x(t)$ pro $t = 1, 2, \ldots$, na základě pozorovaných realizací $Q(x(t), e(t))$ tak, aby byla zajištěna konvergence k řešení. Myšlenkově jednoduchý a na provedení zdlouhavý by byl následující postup. Použít nějaké x a provést velký počet pozorování, pak odhadnout střední hodnotu jako průměr. Tímto bychom získali odhad hodnoty $f(x)$ pro jedno konkrétní x a postup opakovat pro jiné x tak dlouho, dokud nenajdeme řešení. Tato cesta je samozřejmě velmi neefektivní, protože vlastně nesystematicky procházíme možné x. Efektivnější postup navrhli Robbins a Monro ve formě následujícího rekurzivního řešení

$$\hat{x}(t) = \hat{x}(t-1) + \gamma(t)Q(\hat{x}(t-1), e(t)) \tag{8.5.3}$$

kde $\{\gamma(t)\}$ je sekvence reálných čísel s následujícími vlastnostmi

$$\gamma(t) \geq 0, \quad \sum_{t=1}^{\infty} \gamma(t) = \infty, \quad \sum_{t=1}^{\infty} \gamma^2(t) < \infty$$

Těmto podmínkám vyhovuje např. $\gamma(t) = 1/t$. Bylo dokázáno, že $\hat{x}(t)$ bude konvergovat k řešení rovnice za určitých předpokladů. Typický předpoklad je, že $\{e(t)\}$ je sekvence nezávislých náhodných vektorů, což v naší aplikaci na regresní model nemusí být splněno.

Vraťme se opět k lineární regresi (8.5.1). Algoritmus stochastické aproximace založený na (8.5.3) bude pro (8.5.1)

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + \gamma(t)\varphi(t)\left(y(t) - \varphi^T(t)\hat{\Theta}(t-1)\right) \tag{8.5.4}$$

Tento algoritmus byl velmi oblíbený v oblasti adaptivního zpracování signálů.

Vraťme se k původnímu minimalizačnímu problému (8.5.2). Obecně můžeme říci, že chceme minimalizovat

$$\min_x V(x) \quad \text{pro} \quad V(x) = E[H(x, e(t))]$$

Nechť

$$-\frac{d}{dx}V(x) = f^T(x)$$

a dále předpokládejme, že gradient

$$-\frac{\partial}{\partial x}H(x, e(t)) = Q^T(x, e(t))$$

lze vyjádřit pro vybrané x. Pak řešení minimalizačního problému je převedeno na řešení rovnice

$$0 = (-\frac{d}{dx}V(x))^T = f(x) = E[Q(x, e(t))]$$

ve které byla provedena záměna operací ustřednění a derivace. Tedy nastavení x je prováděno ve směru negativního gradientu, a proto se někdy používá pro algoritmy typu Robbins-Monro označení stochastické gradientní metody.
Stochastické gradientní metody lze chápat jako stochastickou analogii metody největšího spádu „steepest descent“ pro numerickou minimalizaci deterministické funkce. Tato metoda pracuje takto

$$x^{(t+1)} = x^{(t)} - \gamma^{(t)} \left(\frac{d}{dx}V(x)\right)^T |_{x=x^{(t)}}$$

kde $\gamma^{(t)}$ je vhodně vybrané kladné číslo a $x^{(t)}$ označuje t-tou iteraci. Je dobře známo, že tento postup není příliš účinný, když se iterace blíží minimu. Takzvané Newtonovy metody poskytují lepší výsledky. V těchto metodách je směr postupu, záporný gradient, nahrazen za

$$\left(-\frac{d^2}{dx^2}V(x)\right)^{-1} \left(\frac{d}{dx}V(x)\right)^T \tag{8.5.5}$$

Pak iterační schema bude mít tvar

$$x^{(t+1)} = x^{(t)} - \gamma(t) \left(\frac{d^2}{dx^2}V(x)\right)^{-1} \left(\frac{d}{dx}V(x)\right)^T |_{x=x^{(t)}}$$

Stochastická varianta Newtonovy metody spočívá v konstrukci aproximace $V''(x)$ na základě předchozích pozorování. Označíme-li odhad druhé derivace $\bar{V}''$ dostaneme přirozenou variantu metody stochastického gradientu

$$\hat{x}(t) = \hat{x}(t-1) + \gamma(t) \left(\bar{V}''(\hat{x}(t-1), e^t)\right)^{-1} Q(\hat{x}(t-1), e(t))$$

kde $e^t = [e(t), e(t-1), ..., e(1)]$. Tato iterace je základem stochastického Newtonova algoritmu.

Odvoďme Newtonův algoritmus pro regresní model, tedy pro (8.5.1). Nejprve zjistíme, že pro kvadratické kritérium je

$$\frac{d^2}{d\Theta^2}V(\Theta) = E\varphi(t)\varphi^T(t)$$

Matici druhých derivací lze určit jako řešení R z

$$E[\varphi(t)\varphi^T(t) - R] = 0 \tag{8.5.6}$$

Aplikací Robbins-Monro postupu na (8.5.6) dostaneme

$$R(t) = R(t-1) + \gamma(t)(\varphi(t)\varphi^T(t) - R(t-1))$$

Odhad $d^2V(\Theta)/d\Theta^2$ v čase t je tedy $R(t)$. Využitím tohoto odhadu získáme stochastický Newtonův algoritmus

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + \gamma(t)R^{-1}(t)\varphi(t)(y(t) - \varphi^T(t)\hat{\Theta}(t-1)) \tag{8.5.7}$$

který je úzce svázán s metodou rekurzivních nejmenších čtverců. Pro $\gamma((t) = 1/t$ a $\alpha_t = 1$ dostaneme rekurzivní algoritmus nejmenších čtverců z podkapitoly 8.2.

8.6 Numerické ošetření rekurzivních algoritmů

Na závěr této kapitoly si všimněme jednoho problému spojeného s realizací všech, v této kapitole představených, rekurzivních algoritmů. Problém se týká výpočtu matice $P(t)$, která je dána rozdílem dvou matic. Vlivem zaokrouhlovacích chyb a při delším chodu rekurze může dojít ke ztrátě pozitivní (semi-)definitnosti této matice, která je základem pro stanovení zisku rekurzivních algoritmů, a to ve svém důsledku obrátí směr hledání extrému, díky čemuž algoritmus diverguje. Proto byly vyvinuty algoritmy, které toto nebezpečí odstraňují. Princip algoritmů bude představen na výpočtu matice $P(t)$ (8.4.19) v rekurzivní verzi metody chyby predikce při uvažování $\lambda = 1$. Pro ostatní rekurzivní identifikační metody je postup analogický.

První z možností jsou odmocninové filtry, které se dělí na tzv. UDU^T filtry a SS^T filtry. První z nich rozkládají matici P na součin diagonální D a trojúhelníkových matic U, U^T, tedy $P = UDU^T$. Druhý způsob spočívá v rozkladu matice $P = SS^T$, kde na formu matice nejsou v principu kladeny žádné předpoklady. Druhý způsob rozkladu má obdobné numerické vlastnosti, avšak vede na jednodušší výsledný algoritmus. Proto zde bude pro ilustraci uveden bez detailního odvození.

Definujeme matici $S(t)$ takto

$$P(t) \stackrel{\triangle}{=} S(t)S^T(t) \tag{8.6.1}$$

Hledejme časový vývoj $S(t)$ místo $P(t)$. Připomeňme si základní rovnici pro výpočet $P(t)$ při $\lambda = 1$

$$P(t) = P(t-1) - P(t-1)\psi(t)\psi^T(t)P(t-1)/[1 + \psi^T(t)P(t-1)\psi(t)]$$

Výpočet $S(t)$ se pak skládá z těchto výsledných vztahů

$$f(t) = S^T(t-1)\psi(t)$$

$$\beta(t) = 1 + f^T(t)f(t)$$

$$\alpha(t) = 1/[\beta(t) + \sqrt{\beta(t)}]$$

$$K(t) = S(t-1)f(t)$$

$$S(t) = S(t-1) - \alpha(t)K(t)f^T(t)$$

Zisk $L(t)$ se vypočítá z $K(t)$

$$L(t) = K(t)/\beta(t)$$

Z důvodu ušetření numerických operací je pak vhodné výpočet parametrů provést takto

$$\hat{\Theta}(t) = \hat{\Theta}(t-1) + K(t)[\epsilon(t)/\beta(t)]$$

což ušetří $n\Theta - 1$ dělení, oproti výpočtu $L(t)$ a následnému násobení s $\epsilon(t)$.

Druhá možnost, jak se vypořádat s možnou ztrátou pozitivní semi-definitnosti matice $P(t)$, spočívá v použití tzv. Josephovy formy pro výpočet matice $P(t)$, kterou lze zapsat ve tvaru

$$P(t) = (I - L(t)\psi^T(t))P(t-1)(I - L(t)\psi^T(t))^T + L(t)L^T(t) \tag{8.6.2}$$

kde I je jednotková matice o rozměrech $n\Theta/n\Theta$ a zisk $L(t)$ je dán rovnicí (8.4.20), tedy vztahem

$$L(t) = P(t-1)\psi(t)/(1 + \psi^T(t)P(t-1)\psi(t))$$

Jednoduchým výpočtem lze ověřit, že vztahy (8.6.2) a (8.4.19) jsou identické. Roznásobením a dosazením vztahu pro výpočet zisku $L(t)$ do (8.6.2) lze psát

$$\begin{aligned}
P(t) &= P(t-1) - \underbrace{P(t-1)\psi L^T}_{\text{a}} - \underbrace{L\psi^T P(t-1)}_{\text{b}} + \underbrace{L\psi^T P(t-1)\psi L^T}_{\text{c}} + \underbrace{LL^T}_{\text{d}} \\
&= P(t-1) - \underbrace{2\frac{P(t-1)\psi\psi^T P(t-1)}{\zeta}}_{\text{a+b}} + \underbrace{\frac{P(t-1)\psi(\psi^T P(t-1)\psi+1)\psi^T P(t-1)}{\zeta^2}}_{\text{c+d}} \\
&= P(t-1) - 2\frac{P(t-1)\psi\psi^T P(t-1)}{\zeta} + \frac{P(t-1)\psi\psi^T P(t-1)}{\zeta} \\
&= P(t-1) - P(t-1)\psi(t)\psi^T(t)P(t-1)/[1 + \psi^T(t)P(t-1)\psi(t)]
\end{aligned} \tag{8.6.3}$$

kde $\psi = \psi(t)$, $L = L(t)$ a $\zeta = \zeta(t) = (1 + \psi^T(t)P(t-1)\psi(t))$.

Ze vztahů (8.6.1) a (8.6.2) je již patrné, že při výpočtu matice $P(t)$ nedochází k odečtu dvou matic a výsledná matice je tak zcela jistě pozitivně semi-definitní.

8.7 Shrnutí

Rekurzivní algoritmy identifikace jsou velmi vhodné pro nejrůznější aplikace. V této kapitole byly postupně odvozeny nejdůležitější rekurzivní postupy: rekurzivní metoda nejmenších čtverců, rekurzivní metoda přídavné proměnné, rekurzivní metoda chyby predikce, včetně aplikace na některé problémy. V závěru kapitoly byla naznačena metoda stochastické aproximace a provedeno jedno z možných numerických ošetření rekurzivních algoritmů.

Rekurzivní identifikační metody jsou velmi zajímavé nejenom z hlediska identifikace, ale i pro adaptivní systémy řízení či zpracování signálu. Uveďme např. alespoň tyto publikace [1], [4], [7], [10], [11], [18], [20], [48]-[51], [56]. Postupy pro numerické ošetření algoritmů jsou uvedeny v [10], [52], [54], [57]. Aplikací rekurzivní identifikace v úloze detekce chyb v systému se zabývá [53]. Algoritmy umožňující sledovat proměnlivé parametry reprezentuje např. [55] a jsou uvedeny rovněž ve 2. díle skript.

Kapitola 9

Identifikace nelineárních systémů

Předchozí kapitoly se věnovaly identifikaci lineárních systémů. Ovšem v mnoha situacích je chování systému nelineární a jeho popis modelem lineárním je nedostatečný, tj. lineární model vede na velkou chybu predikce výstupu nelineárního systému. Značná pozornost je proto věnována popisu (či modelování) nelineárních dynamických systémů.
Cílem této kapitoly je představit některé ze základních konceptů a idejí identifikace parametrů nelineárních vstupně-výstupních modelů.

9.1 Nelineární vstupně-výstupní model a formulace problému

Uvažujme nelineární systém, který je popsán následující rovnicí

$$\begin{aligned} y(t) = g_0\Big(&y(t-1), y(t-2), \ldots, y(t-n_{a,0}),\\ &u(t-1), \ldots, u(t-n_{b,0}),\\ &e(t-1), \ldots, e(t-n_{c,0}); \Theta_0\Big) + e(t) \end{aligned} \tag{9.1.1}$$

kde proměnné $y(t)$, $u(t)$ a $e(t)$ jsou definovány v souladu se zavedeným značením, tj. jako výstup, vstup a porucha a funkce $g_0(\cdot;\Theta_0)$ je neznámá, obecně nelineární, funkce závisející na množině neznámých parametrů Θ_0.
Protože funkce $g_0(\cdot;\Theta)$ může být velmi složitá funkce s neznámou a mnohdy stěží zjistitelnou strukturou, tak *cíl nelineární identifikace* je nalézt model

$$\begin{aligned} y(t) = g\Big(&y(t-1), y(t-2), \ldots, y(t-n_a),\\ &u(t-1), \ldots, u(t-n_b),\\ &\epsilon(t-1), \ldots, \epsilon(t-n_c); \Theta\Big) + \epsilon(t) \end{aligned} \tag{9.1.2}$$

dostatečně přesně[1] popisující (aproximující) chování systému (9.1.1), kde proměnnou $\epsilon(t)$ můžeme chápat jako chybu predikce nebo chybu modelu. Na rozdíl od lineární identifikace, kde jsme většinou předpokládali shodný popis systému (9.1.1) i modelu (9.1.2), zde tento předpoklad povětšinou neplatí a funkce $g_0(\cdot;\Theta_0)$ v (9.1.1) a $g(\cdot;\Theta)$ v (9.1.2) *nemusí být*[2] shodné. V rámci

[1] Pod pojmem „dostatečně přesný“ rozumíme odhad např. s minimální variancí chyby predikce modelu v *dané* struktuře.

[2] Pokud je opuštěn předpoklad shodné struktury popisu systému a modelu, tak pojmy jako strannost a konzistence odhadu postrádají svůj smysl.

úlohy identifikace nelineárních systémů je tedy nutné se navíc zabývat i vhodnou volbou (či definicí) nelineární funkce modelu $g(\cdot;\Theta)$. To lze provést např. na základě nějaké apriorní znalosti či zkušenosti ohledně systému a odhadu parametrů Θ z dostupných dat $\{y(t), u(t)\}$.
V následující části se zaměříme na dva základní principy identifikace nelineárních systémů, a to na popis nelineárního chování pomocí

- identifikace nelineárního modelu s lineární funkcí odhadovaných parametrů, kde budeme uvažovat následující tři přístupy:
  - množiny lineárních modelů,
  - nelineárního rozšíření třídy modelů ARMAX,
  - neuronové sítě a
- identifikace nelineárního modelu s nelineární funkcí odhadovaných parametrů.

9.2 Identifikace nelineárního modelu s lineární funkcí odhadovaných parametrů

Začněme s představením identifikačních metod, kde struktura systému (9.1.1), tj. nelineární funkce $g_0(\cdot;\Theta_0)$, je neznámá. V této části představené identifikační techniky jsou pak založeny na takové volbě struktury modelu (9.1.2), kde odhadované parametry Θ vystupují v lineární formě vzhledem k dostupným měřením. Výhoda tohoto přístupu spočívá v tom, že pro odhad parametrů nelineárního modelu můžeme využít známé techniky z identifikace lineárních systémů (např. metodu nejmenších čtverců, přídavné proměnné či chyby predikce).

9.2.1 Identifikace po částech lineárního modelu

Tento přístup je založen na jednoduché myšlence spočívající v identifikaci množiny lokálních lineárních modelů, které aproximují nelineární model v několika pracovních bodech [70]. Hlavní předností tohoto přístupu je možnost využití široké palety identifikačních a řídicích algoritmů pro lineární modely (z oblasti identifikačních algoritmů lze zmínit MNČ, IVM, PEM představené v předcházejících kapitolách). Na druhou stranu za nevýhodu lze považovat nutnost určení počtu pracovních bodů, ve kterých bude provedena identifikace, jejich specifikaci a rovněž i určení indikátorů, které umožní spolehlivé přepínaní[3] lokálních lineárních modelů použitých např. pro predikci nebo řízení výstupu systému.

9.2.2 Identifikace modelu ve struktuře NARMAX

Převážná část modelů uvažovaných v kapitolách 4 až 7 byla reprezentována lineárními modely ve struktuře ARMAX (např. model (5.2.3)), kde výstup systému $y(t)$ je lineární kombinace výstupů v předchozích časových okamžicích $\{y(t-i)\}_{i=1}^{na}$, předchozích vstupů $\{u(t-i)\}_{i=1}^{nb}$ a poruch $\{e(t-i)\}_{i=0}^{nc}$. Pro popis nelineárních systémů bylo v 80. letech minulého století navrženo rozšíření struktury ARMAX zahrnující i nelineární (typicky *polynomiální*) funkce signálů $\{y(t), u(t), e(t)\}$. Tato rozšířená třída modelů je často označována zkratkou NARMAX z anglického názvu „nonlinear autoregressive moving average model with exogenous inputs“ [70, 71].

[3] Přepínání, nebo-li volba, modelu může být založena např. na definici pracovních intervalů vstupních a výstupních signálů (např. intervaly pro měření teploty) nebo na hodnotách externího signálu, který nemusí přímo vstupovat do modelu, ale systém nějakým způsobem ovlivňuje (např. tlak, vlhkost, roční období).

Vzhledem k obecnému modelu (9.1.2), polynomiální NARMAX model má následující strukturu [70]

$$y(t) = \theta_0 + \sum_{i=1}^{n} \theta_i x_i(t) + \sum_{i=1}^{n}\sum_{j=1}^{n} \theta_{ij} x_i(t)x_j(t) + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} \theta_{ijk} x_i(t)x_j(t)x_k(t) + \ldots + \epsilon(t) \tag{9.2.3}$$

kde $n = na + nb + nc$, vektorová proměnná $x(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^T$ je definována jako

$$x_m(t) = \begin{cases} y(t-m), & 1 \le m \le na \\ u(t-(m-na)), & na+1 \le m \le na+nb \\ \epsilon(t-(m-na-nb)), & na+nb+1 \le m \le n \end{cases} \tag{9.2.4}$$

tj. $x(t)$ zahrnuje minulé vstupně-výstupní signály $\{y(t), u(t), \epsilon(t)\}$ a vektor neznámých parametrů je dán

$$\Theta = [\theta_0, \theta_1, \ldots, \theta_n, \theta_{11}, \theta_{12}, \ldots, \theta_{nn}, \theta_{111}, \ldots, \theta_{nnn}, \ldots]^T \tag{9.2.5}$$

Poznamenejme, že nejvyšší uvažovaný stupeň polynomu určuje i stupeň modelu. Jako *příklad* polynomiálního modelu můžeme uvést např. následující NARMAX model druhého stupně

$$\begin{aligned} y(t) = {} & \theta_0 + \theta_1 y(t-1) + \theta_2 u(t-1) + \theta_{11}(y(t-1))^2 \\ & + \theta_{12} y(t-1)u(t-2) + \epsilon(t) + \theta_{31}\epsilon(t-1)\epsilon(t-2) \end{aligned} \tag{9.2.6}$$

NARMAX model je tedy nelineární (konkrétně polynomiální) funkcí vstupně-výstupních proměnných, avšak je *lineární* funkcí vzhledem k parametrům ve vektoru Θ. To jest, model (9.2.6) lze zapsat v nám již známé lineární struktuře

$$y(t) = [1, y(t-1), u(t-1), (y(t-1))^2, y(t-1)u(t-2), \epsilon(t-1)\epsilon(t-2)]\Theta + \epsilon(t) \tag{9.2.7}$$

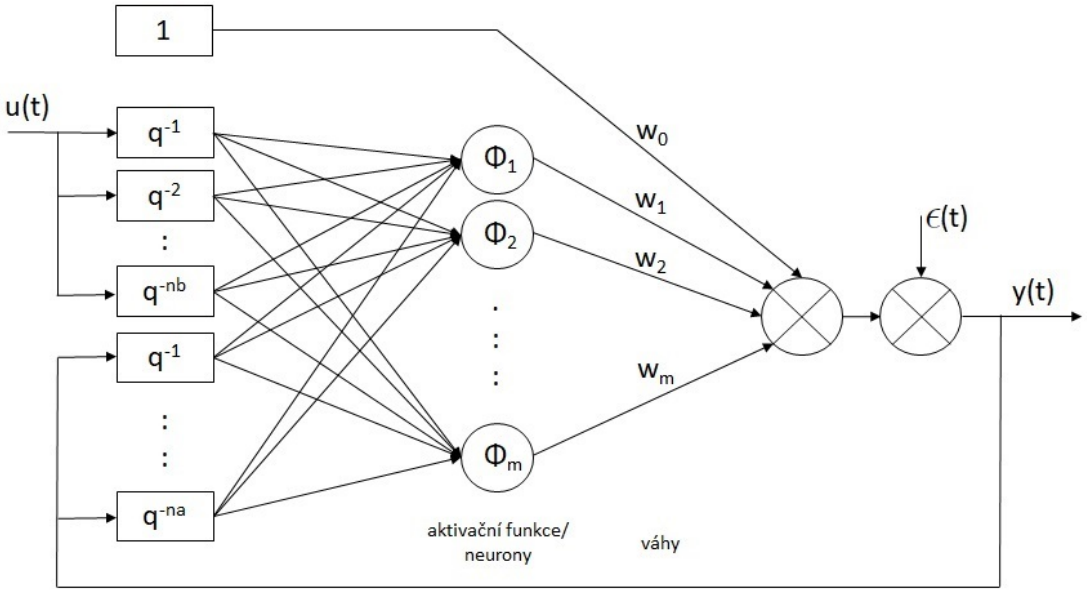
kde pro odhad vektoru parametrů $\Theta = [\theta_0, \theta_1, \theta_2, \theta_{11}, \theta_{12}, \theta_{31}]^T$ můžeme použít dříve představenou metodu rozšířených nejmenších čtverců, metodu chyby predikce nebo i metodu přídavné proměnné (pro odhad některých neznámých parametrů).

Pro úspěšnou identifikaci je klíčová rozumná volba struktury modelu (podobně jako tomu bylo u lineárních modelů, kde volba struktury byla diskutována v kapitole 5). Struktura modelu by měla být dostatečně *bohatá*, aby popsala chování systému s dostatečnou přesností. Zároveň by struktura měla být *úsporná*, aby identifikovaný model nebyl zbytečně složitý. Složitý nebo příliš jednoduchý model může vést k velkým chybám. Volba struktury vychází z apriorní znalosti o systému a je často v rukách návrháře.

Poznámka . Třída NARMAX modelů rozšiřuje nebo zobecňuje mnoho nelineárních modelů, které byly v literatuře rozvíjeny od konce 19. století. Jako příklad zde můžeme uvést Hammersteinův, Wienerův nebo Volterrův model [71]. Poznamenejme rovněž, že se statickými modely ve struktuře NARMAX jsme se již krátce setkali v úvodu kapitoly 4 a v kapitole 5.

Poznámka . Speciálním případem NARMAX modelu, je model NARX, kde porucha ovlivňující výstup systému je uvažována jako bílá (tj. NARX je nelineární obdoba modelu ARX). NARX model lze tedy formálně zapsat jako model (9.2.3), kde $n = na+nb$ a proměnná $x(t)$ je definována jako

$$x_m(t) = \begin{cases} y(t-m), & 1 \le m \le na \\ u(t-(m-na)), & na+1 \le m \le n \end{cases} \tag{9.2.8}$$



Obrázek 9.1: Ilustrace struktury neuronové sítě modelující systém v úloze identifikace.

Pro odhad parametrů NARX modelu je vhodná metoda nejmenších čtverců (v rekurzivní i jednorázové podobě).

9.2.3 Identifikace modelu ve formě neuronových sítí

Identifikační metody založené na neuronových sítích jsou intenzivně rozvíjeny a aplikovány od 90. let minulého století v mnoha oblastech zahrnujících např. zpracování signálů, řízení, rozpoznávání řeči nebo obrázků. Neuronové sítě tak hrají důležitou roli v systémech využívajících umělé inteligence.
Tím, že neuronové sítě byly rozvíjeny v mnoha různorodých oblastech, je používaná terminologie odlišná od terminologie používané v oblasti identifikace systémů. Proto terminologie použitá pro představení neuronových sítí bude odlišná od terminologie v předchozích kapitolách těchto skript. Na závěr však používané termíny v neuronových sítích a identifikaci sjednotíme.

Neuronové sítě, v anglicky psané literatuře označované pojmem „neural networks“, jsou používány k aproximaci nelineárních funkcí, popř. popisu chování nelineárního systému, na základě měřených dat. Za základní strukturu neuronové sítě je považována *síť s jednou skrytou vrstvou*. V anglicky psané literatuře se lze setkat s pojmy „single-hidden-layer network“ nebo „single-layer network“ (SLN) [70]. Tato základní struktura je založena na následujícím matematickém modelu

$$y(t) = w_0 + \sum_{i=1}^{m} w_i \phi_i(x(t)) + \epsilon(t) \tag{9.2.9}$$

kde proměnná $x(t)$ obsahující minulé vstupy a výstupy je definována vztahem (9.2.8), $\phi_i(\cdot)$ je tzv. aktivační funkce (v anglicky psané literatuře označované jako „activation function“), $\{w_i\}_{i=0}^{m}$ je množina vah a m je počet aktivačních funkcí tvořící jednotlivé neurony[4]. Pro lepší představu je neuronová síť (9.2.9) znázorněna na obrázku 9.1.

[4]Matematicky formulované neurony tvořené nelineární funkcí byly ve 50. letech minulého století použity k popisu chování vazeb mezi biologickými neurony a vstupními a výstupními signály [71].

Aktivační funkce jsou voleny uživatelem. V literatuře bylo navrženo několik vhodných aktivačních funkcí [70, 71]. Mezi nejpoužívanější můžeme zařadit následující

- sigmoida popsaná

$$\phi_i(x) = \frac{1}{1 + e^{-p_i^T x}} \tag{9.2.10}$$

  kde $p_i = [p_{i,1}, p_{i,2}, \ldots, p_{i,n}]$ je vektor parametrů aktivační funkce shodné dimenze jakou má vektor x, tj. $\dim(p_i) = dim(x) = n$, a každý element vektoru p_i je kladný, tj. $p_i > 0, \forall i$,

- hyperbolickou tangenciálu popsanou

$$\phi_i(x) = \tanh(p_i^T x) = \frac{e^{p_i^T x} - e^{-p_i^T x}}{e^{p_i^T x} + e^{-p_i^T x}} \tag{9.2.11}$$

  kde pro parametry platí $p_{i,j} > 0, \forall j$,

- gaussovskou funkci popsanou

$$\phi_i(x) = e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} \tag{9.2.12}$$

  kde $\Sigma_i \in \mathbb{R}^{n \times n}$ je diagonální čtvercová matice s vektorem $[\sigma_{i,1}^2, \sigma_{i,2}^2, \ldots, \sigma_{i,n}^2]^T$ na diagonále a $\mu_i \in \mathbb{R}^n$. Elementy matice Σ_i a vektoru μ_i tvoří vektor parametrů p_i.

Aby bylo možné pro odhad neznámých parametrů modelu opět využít lineárních estimačních algoritmů, musí uživatel při modelování systému neuronovými sítěmi zvolit (nebo-li definovat) aktivační funkci a m vektorů parametrů p_i pro $i = 1, 2, \ldots, m$. Každá aktivační funkce v (9.2.9) je definována jiným vektorem parametrů, tj. $\phi_i(x(t)) = \phi_i(x(t); p_i)$. Počet a typ neuronů a příslušné vektory parametrů by měly být voleny tak, aby výsledný model byl schopný dostatečně přesně popsat chování systému (9.1.1) a obsáhnout rozsah vstupních a výstupních veličin. Jedná se tak o velmi důležitou volbu návrháře. Ilustrace volby parametrů neuronové sítě je dána v následující části.

Cílem identifikace modelu ve formě neuronových sítí je následně nalézt odhad vah neuronové sítě $\{w_i\}_{i=0}^m$ na základě vstupních a výstupních dat a definovaných aktivačních funkcí. Tento proces se označuje jako *trénování neuronové sítě* nebo *učení* (v anglicky psané literatuře „network training“ nebo „learning“).
Pokud se podíváme na strukturu modelu (9.2.9), je patrné, že se při daných aktivačních funkcích jedná o lineární strukturu, kterou můžeme opět zapsat ve formě lineárního regresního modelu (4.1.1), tj.

$$y(t) = \underbrace{[1, \phi_1(x(t)), \ldots, \phi_m(x(t))]}_{\varphi^T(t)} \Theta + \epsilon(t) \tag{9.2.13}$$

kde $\Theta = [w_0, w_1, \ldots, w_m]$. Pro trénování neuronové sítě, tj. pro odhad vah sumarizovaných ve vektoru Θ, tak můžeme použít nám známou metodu nejmenších čtverců.

Poznámka . Terminologie používaná v neuronových sítích se odlišuje od té, která je používána v oblasti identifikace systémů. Ale po představení neuronových sítí můžeme najít terminologické ekvivalenty, tj. váhy jsou neznámá (hledané) parametry modelu, aktivační funkce jsou bázové funkce a trénování sítě je vlastně proces odhadu parametrů. Poznamenejme, že neuronová síť

s gaussovskou aktivační funkcí je v anglicky psané literatuře označována jako „radial basis function (RBF) network“.

Poznámka . Jak již bylo zmíněno, klíčem k úspěchu je vhodná volba aktivačních funkcí a parametrů neuronové sítě. Pokud nemáme žádnou apriorní informaci, která by umožnila rozumnou volbu parametrů, můžeme tyto parametry odhadovat společně s váhami. V tomto případě se už ale jedná o nelineární optimalizační úlohu a musíme použít nelineární estimátor. Pro sítě, kde se odhadují jak váhy, tak i parametry platí, že je nutné použít nelineární trénovací algoritmy [70, 71]. Ty jsou často založeny na nelineárních metodách pro odhad stavu a jsou diskutovány a ilustrovány v následujícím díle skript věnovaném odhadu stavu.

Poznámka . Neuronová síť může mít více vrstev (úrovní) neuronů. Takováto síť se označuje jako *neuronová síť s více vrstvami* (v anglicky psané literatuře označovaná pojmem multi-layer neural network) [70, 71]. Množství vrstev je obecně svázáno s nelinearitou systému.

Poznámka . Všimněme si, že NARX model je vlastně vážený součet polynomiálních funkcí minulých vstupů a výstupů systému, kdežto neuronová síť je vážený součet aktivačních funkcí minulých vstupů a výstupů. Aktivační funkce mohou být jak polynomiální, tak i exponenciální, nebo jejich kombinace.

Poznámka . V literatuře se lze setkat i s pojmem „wavelet network“. Zde se jedná v podstatě o neuronovou síť se speciální volbou aktivačních funkcí [70].

9.2.4 Ilustrace nelineárních identifikačních metod

Závěrem sekce věnované identifikaci nelineárních systémů se budeme věnovat ilustraci použití představených identifikačních metod. Uvažujme následující *nelineární* systém generující data

$$S : y(t+1) = 0.5 + 0.1y(t) + u(t) - 0.2y(t)u(t) + 0.8\sin(y(t)) + e(t), t = 0, 1, \ldots, \tau \quad (9.2.14)$$

kde vstupní signál $u(t)$ je bílý gaussovský proces s nulovou střední hodnotou a jednotkovou variancí, tj.

$$u(t) \sim \mathcal{N}\{u(t); 0, 1\}, \forall t \quad (9.2.15)$$

a porucha $e(t) \sim \mathcal{N}\{u(t); 0, \sigma^2\}, \forall t$, je taktéž bílý proces s $\sigma^2 = 0.01$.

Cílem je na základě měřených vstupně-výstupních dat $\{y(t), u(t)\}$ najít *model* (tj. strukturu i parametry) systému. V tomto příkladu jsou uvažovány tři modely:

i) ARX model prvního řádu (5.2.12) s $na = 1$ a $nb = 1$ zapsaný v lineární regresní struktuře

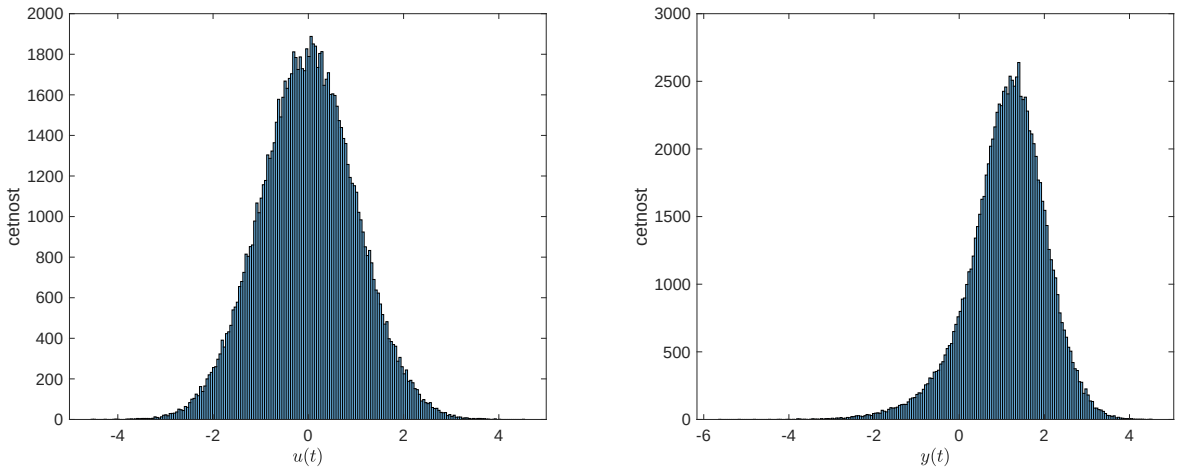
$$M_{\text{ARX}} : y(t+1) = [1, y(t), u(t)]\Theta_{\text{ARX}} + \epsilon(t) \quad (9.2.16)$$

kde hledaný vektor parametrů je $\Theta_{\text{ARX}} = [\theta_0, \theta_1, \theta_2]^T$ a $\epsilon(t)$ je chyba modelu (či predikce), která je identifikační metodou minimalizována.

ii) NARX model prvního řádu a třetího stupně (9.2.3) s (9.2.8), $na = 1$ a $nb = 1$ zapsaný v lineární regresní struktuře

$$M_{\text{NARX}} : y(t+1) = [1, y(t), u(t), (y(t))^2, (u(t))^2, y(t)u(t), (y(t))^3, (u(t))^3, \ldots, (y(t))^2u(t), y(t)(u(t))^2]\Theta_{\text{NARX}} + \epsilon(t) \quad (9.2.17)$$

kde hledaný vektor parametrů je $\Theta_{\text{NARX}} = [\theta_0, \ldots, \theta_9]^T$.



Obrázek 9.2: Histogram vstupního a výstupního signálu přes všechny časové okamžiky pro systém S (9.2.14).

iii) Model ve formě neuronové sítě (9.2.9) zapsaný v lineární regresní struktuře

$$M_{\mathrm{NN}} : y(t+1) = [1, \phi_1(y(t), u(t)), \phi_2(y(t), u(t)), \ldots, \phi_m(y(t), u(t))]\Theta_{\mathrm{NN}} + \epsilon(t) \quad (9.2.18)$$

kde $m = 221$, hledaný vektor parametrů je $\Theta_{\mathrm{NN}} = [\theta_0, \ldots, \theta_{221}]^T$ a aktivační funkce je zvolena jako Gaussova funkce (9.2.12)

$$\phi_i(y(t), u(t)) = \left(e^{-\frac{1}{2}(y(t)-\mu_{i,1})^2/\sigma_{i,1}^2}\right)\left(e^{-\frac{1}{2}(u(t)-\mu_{i,2})^2/\sigma_{i,2}^2}\right) \quad (9.2.19)$$

tj. s uživatelem volenými parametry $p_i = \{\mu_{i,1}, \sigma_{i,1}^2, \mu_{i,2}, \sigma_{i,2}^2\}$, kde parametry $\mu_{i,1}, \sigma_{i,1}^2$ jsou vytaženy k výstupnímu signálu $y(t)$ a parametry $\mu_{i,2}, \sigma_{i,2}^2$ jsou vztaženy k vstupnímu signálu $u(t)$.

Uvažujme, že máme k dispozici $\tau = 10^5$ měřených dat. Pro ilustraci jsou ukázány histogramy vstupního a výstupního signálu přes všechny časové okamžiky na obrázku 9.2. Odhadu parametrů ARX a NARX modelu již nestojí nic v cestě, máme definovanou strukturu modelů (9.2.16), (9.2.17) a máme k dispozici měřená data. Snadno tedy určíme soustavu τ lineárních rovnic a odhadneme vektor hledaných parametrů Θ_{ARX} a Θ_{NARX} metodou nejmenších čtverců. Avšak pro odhad parametrů Θ_{NN} modelu ve formě neuronových sítí ještě musíme specifikovat množiny parametrů $p_i, \forall i$, pro všechny aktivační funkce. K tomu nám dopomůžou histogramy vstupního a výstupního signálu. Z nich lze vyčíst, že vstupní signál je *zhruba* v rozmezí $u(t) \in (-3, 3)$ a výstupní signál v rozmezí $y(t) \in (-4, 4)$. Proto zvolme následující hodnoty parametrů aktivačních funkcí, které pokrývají definiční obor obou signálů (při použití notace programového prostředí MATLAB®):

i) $\mu_{i,1} \in \{-4 : 0.5 : 4\}$, tj. 17 hodnot pro pokrytí rozsahu signálu $y(t)$,

ii) $\mu_{i,2} \in \{-3 : 0.5 : 3\}$, tj. 13 hodnot pro pokrytí rozsahu signálu $u(t)$,

iii) $\sigma_{i,1} = 0.5$,

iv) $\sigma_{i,2} = 0.5$,

	ARX	NARX	NN
MSE	0.116	0.0145	0.0138

Tabulka 9.1: Střední kvadratická chyba modelů v nelineární identifikaci pro systém (9.2.14).

což vede na celkem $m = 17 \times 13 = 221$ unikátních vektorů parametrů p_i, a tím i na m unikátních aktivačních funkcí $\phi_i(y(t), u(t))$ definovaných v (9.2.18). Tímto máme rovněž plně specifikovanou strukturu modelu s neuronovými sítěmi a metodou nejmenších čtverců již lze snadno odhadnout vektor parametrů Θ_{NN}. Všimněme si, že variance gaussovské funkce σ_i je zvolena jako krok mřížky pro umístění středních hodnot μ_i. Důvodem této volby je to, že gaussovská funkce nabývá významných hodnot na intervalu $(\mu_i - \sigma_i, \mu_i + \sigma_i)$, a tím je tedy i zajištěno pokrytí definičního oboru signálů gaussovskou funkcí mezi definovanými body mřížky.

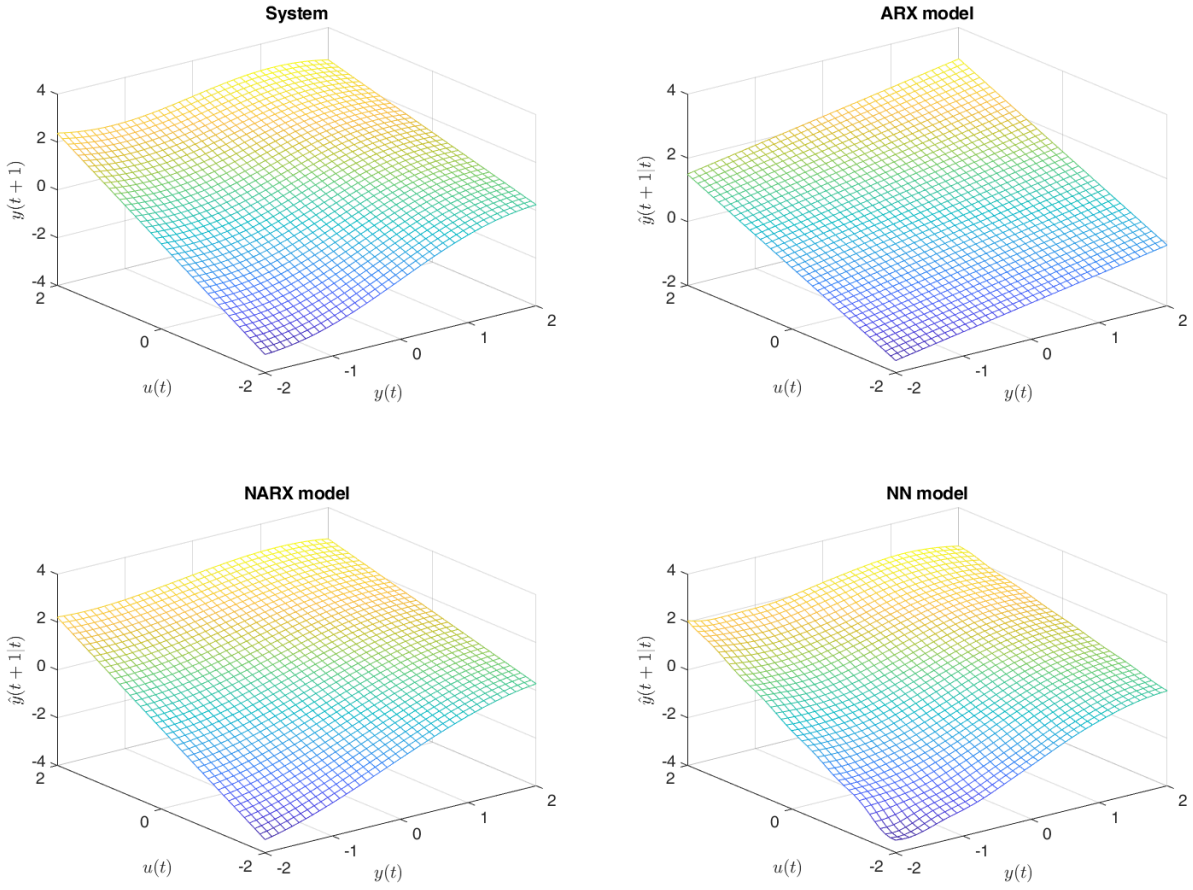
Pro porovnání a ohodnocení kvality identifikovaných modelů M_{ARX}, M_{NARX} a M_{NN} můžeme (pro jednu naměřenou trajektorii) použít střední kvadratickou chybu predikce výstupu definovanou jako

$$MSE = \frac{1}{\tau}\sum_{i=1}^{\tau}(y(t) - \hat{y}(t|t-1;\hat{\Theta}))^2 = \frac{1}{\tau}\sum_{i=1}^{\tau}(\tilde{y}(t|t-1;\hat{\Theta}))^2 \tag{9.2.20}$$

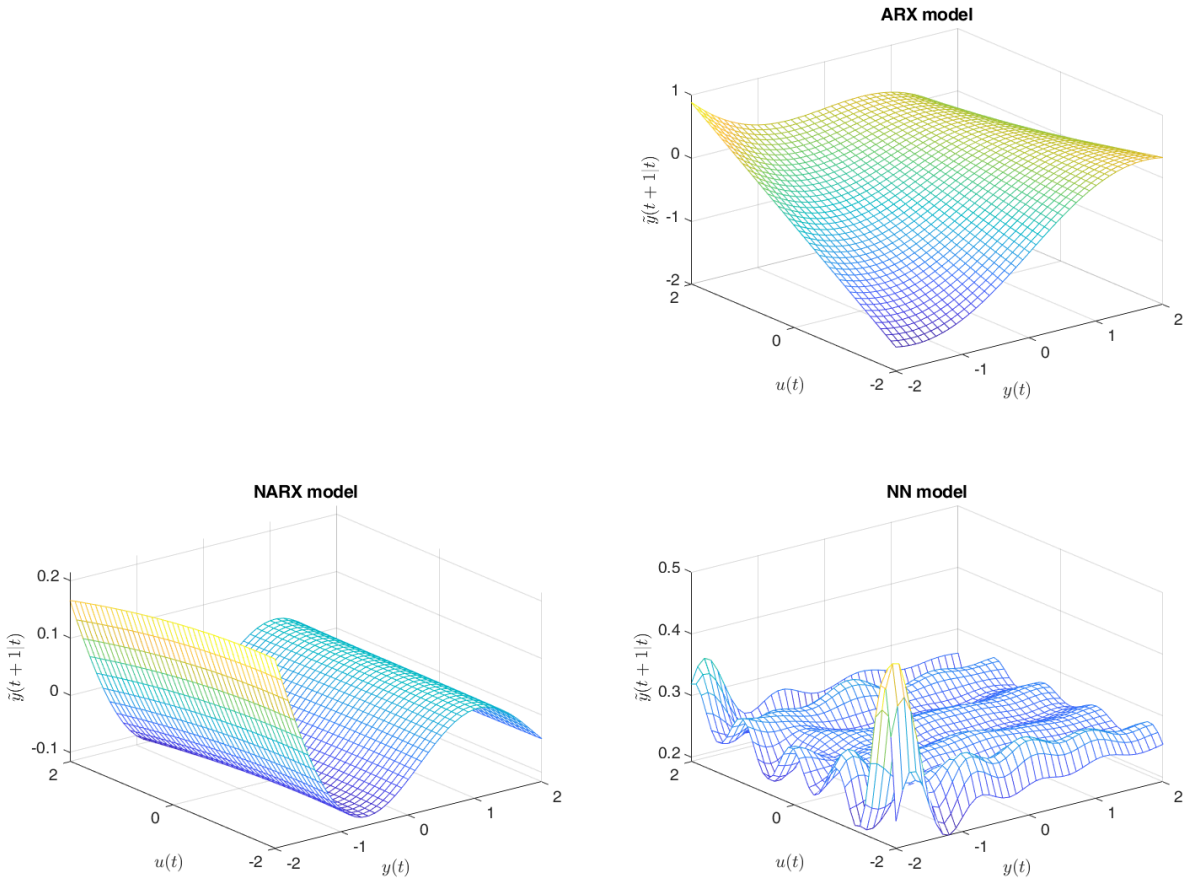
kde $y(t)$ je skutečný výstup systému S (9.2.14), $\hat{y}(t|t-1;\hat{\Theta})$ je jednokroková predikce výstupu systému spočtená na základě uvažovaných modelů (např. model M_{ARX} (9.2.16)), kde namísto parametrů Θ_{ARX} byly použity identifikované parametry metodou nejmenších čtverců $\hat{\Theta}_{\mathrm{ARX}}$, a $\tilde{y}(t|t-1;\hat{\Theta}) = y(t) - \hat{y}(t|t-1;\hat{\Theta})$ je chyba predikce. Výsledky jsou shrnuty v tabulce 9.1. Z nich je patrné, že nejméně kvalitní odhad poskytuje ARX model. Tato skutečnost je pochopitelná, protože ARX model je model lineární a snažíme se popsat chování nelineárního systému. Výrazně lépe popisují systém nelineární modely NARX a NN. Zde je střední kvadratická chyba odhadu blízká varianci chyby měření $\sigma^2 = 0.01$, tj. oba modely poskytují téměř optimální predikci. Model NN je v tomto ohledu o něco kvalitnější, avšak obsahuje daleko větší počet neznámých parametrů a také vyžaduje větší interakci s návrhářem kvůli specifikaci parametrů aktivačních funkcí.

Pro úplnost ještě vykreslíme funkční hodnotu výstupu systému $y(t+1)$ v závislosti na hodnotách $y(t) \in (-2, 2)$ a $u(t) \in (-2, 2)$ *bez* uvažování šumu $e(t)$. To samé vykreslíme i pro modely ARX, NARX a NN, kde vektory parametrů jsou odhadnuty z dat. Ideálně bychom chtěli, aby modely co nejpřesněji popisovaly systém. Průběhy jsou vykresleny na obrázku 9.3. Chyba modelu oproti systému je pak vykreslena na obrázku 9.4.

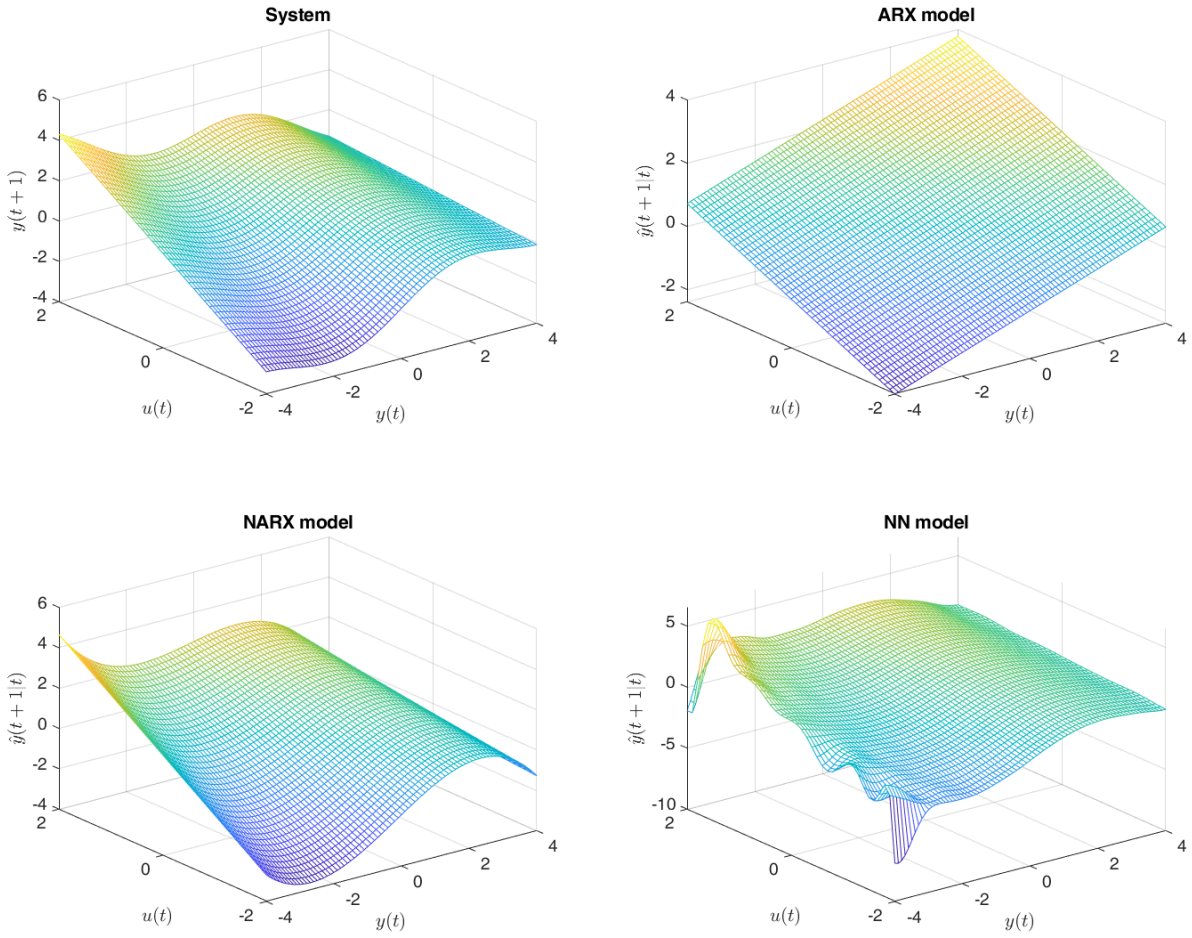
Z obrázků lze vyčíst, že modely NARX i NN poskytují kvalitní aproximaci systému, avšak chyba každého modelu má jiný charakter. Kvalitní aproximace je však v oblasti, kde byl dostatečný počet trénovacích dat (měření). Pro zvětšující se obor hodnot pro $y(t)$ a $u(t)$ se kvalita aproximace zhoršuje (zejména na okrajích oboru hodnot) v důsledku menšího počtu trénovacích dat (z histogramů 9.2 lze vidět, že na okrajích uvažovaných oborů hodnot je malé množství dostupných dat). Zhoršující se kvalita aproximace pro větší obor hodnot $y(t) \in (-4, 4)$ a $u(t) \in (-4, 4)$ je ilustrována na obrázku 9.5. Všimněme si tedy, že identifikované modely mají *lokální* platnost, která vychází z použitých experimentálních (trénovacích) dat. Pro vstupní signál v jiném rozsahu bychom dostali jiné modely.



Obrázek 9.3: Výstup systému $y(t+1)$ a modelů $\hat{y}(t+1|t)$ v závislosti na $y(t)$ a $u(t)$ pro (9.2.14).



Obrázek 9.4: Chyba výstupu modelů $\tilde{y}(t+1|t)$ v závislosti na $y(t)$ a $u(t)$ pro (9.2.14).



Obrázek 9.5: Výstup systému $y(t+1)$ a modelů $\hat{y}(t+1|t)$ v závislosti na $y(t)$ a $u(t)$ pro (9.2.14) (větší obor hodnot).

9.3 Identifikace nelineárního modelu se známou nelineární funkcí odhadovaných parametrů

Doposud jsme v této kapitole uvažovali situaci, kdy struktura ani parametry systému (9.1.1) nejsou známy. Pak představený přístup k identifikaci byl založen na vhodné volbě struktury modelu (např. NARMAX (9.2.3) či neuronové sítě (9.2.9)), která je typicky nelineární vzhledem k měřeným veličinám, ale lineární vzhledem k parametrům. V důsledku tak pro odhad parametrů nelineárních modelů využíváme známé techniky z oblasti identifikace lineárních systémů.
Může však nastat situace, kdy struktura systému, tj. funkce $g_0(\cdot;\Theta_0)$, je známá a nelineární vzhledem k hledaným parametrům Θ_0. Jedním z příkladů takového systému, je navigační systém pro určení polohy objektu na základě satelitních měření (tzv. pseudo-vzdáleností mezi objektem a daným satelitem) [62], [63], kde struktura plyne z fyzikální podstaty funkce navigačního systému. Zjednodušený model měřené pseudo-vzdálenosti mezi objektem a satelitem můžeme napsat ve formě

$$y(t) = \sqrt{(p_x(t)-\theta_1)^2 + (p_y(t)-\theta_2)^2 + (p_z(t)-\theta_3)^2} + \epsilon(t) \tag{9.3.21}$$

kde $p(t) = [p_x(t), p_y(t), p_z(t)]^T$ je *známá* poloha satelitu[5] a $\Theta = [\theta_1, \theta_2, \theta_3]^3$ je *hledaná* poloha objektu v kartézských souřadnicích, která je pro jednoduchost uvažována konstantní. Model je tedy nelineární k hledaným parametrů a můžeme ho pro jednoduchost formálně zapsat jako

$$y(t) = g(p(t);\Theta) + \epsilon(t) \tag{9.3.22}$$

Zdůrazněme, že model (9.3.22) nelze zapsat ve formě lineárního regresního modelu (4.1.1) jako tomu bylo v případě NARMAX modelu či neuronových sítí.

Možné řešení identifikace nelineárně svázaných parametrů Θ modelu (9.3.22) spočívá v linearizaci nelineární funkce $g(p(t);\Theta)$ v okolí jednoho či více uživatelem zvolených linearizačních bodů Θ_{LIN} např. pomocí Taylorova rozvoje prvního řádu [62]. Využití Taylorova rozvoje nelineární funkce $g(p(t);\Theta)$ umožní zapsat model (9.3.22) ve formě

$$y(t) \approx \Theta_{\text{LIN}} + G(p(t);\Theta_{\text{LIN}})\,(\Theta - \Theta_{\text{LIN}}) + \epsilon(t) \tag{9.3.23}$$

kde $G(\Theta_{\text{LIN}}) = G(p(t);\Theta_{\text{LIN}}) = \frac{dg(p(t);\Theta)}{d\Theta}|_{\Theta=\Theta_{\text{LIN}}}$ je známý Jacobián (tj. matice prvních derivací) nelineární funkce $g(p;\Theta)$ vyhodnocený v linearizačním bodě Θ_{LIN}. Protože linerizační bod je známý lze přepsat (9.3.23) do formy

$$(y(t) - \Theta_{\text{LIN}}) + G(\Theta_{\text{LIN}})\Theta_{\text{LIN}} \approx G(\Theta_{\text{LIN}})\Theta + \epsilon(t) \tag{9.3.24}$$

která je již lineární vzhledem k hledaným parametrům Θ a pro odhad parametrů můžeme použít identifikační techniky z oblasti lineárních systémů, jakou je např. metoda nejmenších čtverců.

Poznámka . Taylorův rozvoj je jen jednou z mnoha technik linearizace nelineární funkce. Mezi jinými můžeme např. zmínit Stirlingovu interpolaci nebo statistickou linearizaci. Poznamenejme, že různým technikám linearizace je věnována obsáhlá diskuze v druhém díle skript.

Poznámka . Při identifikaci linearizovaných modelů je nutné vzít v potaz vliv linearizační chyby. Někdy linearizační chyba může být zanedbatelná (např. vzhledem k chybě měření), jindy může

[5] Index t nemusí v případě satelitní navigace nutně znamenat časový index měření, ale může být chápán i jako index satelitu.

být natolik významná, že může výrazným způsobem ovlivnit kvalitu výsledných odhadů.

Poznámka . Jinou možností je ignorovat znalost o struktuře systému, tj. o funkci $g(\cdot;\Theta)$ a použít nějaký obecný nelineární regresní model, jakým je např. neuronová síť. V tomto případě však můžeme očekávat model s větší chybou predikce výstupu.

9.4 Shrnutí a zhodnocení výsledků

V této kapitole jsme se seznámili s modelováním a identifikací parametrů nelineárního systému. Pozornost byla upřena zejména na popis systému modelem ve struktuře NARMAX, modelem ve formě neuronových sítí a modelem s nelineární funkcí hledaných parametrů. Identifikace nelineárních systémů je velmi aktuální a aktivně rozvíjené téma s aplikacemi napříč téměř všemi obory lidského snažení. Pro další studium identifikace nelineárních systémů lze doporučit knihy [15], [62], [70]-[73], kde kromě detailního rozboru identifikačních metod lze najít i další typy umožňující vhodnou volbu struktury modelu. Poznamenejme také, že metody pro identifikaci nelineárních systémů jsou implementovány v mnoha programových balíčcích, které jsou volně dostupné.

Kapitola 10

Identifikace parametrů lineárních stavových modelů

V předcházejících kapitolách jsme se soustředili zejména na identifikaci vstupně-výstupních modelů. Moderní algoritmy pro automatické řízení, detekci poruch a predikci výstupu systému jsou však mnohdy založeny na popisu systému stavovým modelem. Stavový model lze chápat jako alternativu k modelu vstupně-výstupnímu, která umožní popsat systém a jeho vnitřní strukturu detailněji. Stavový model, či jeho struktura, tak často vychází z matematicko-fyzikálního modelování uvažovaného systému. Jako příklad zde můžeme uvést kinematický model pohybu tělesa [63].
Cílem této kapitoly je představit základní koncepty a ideje metod pro identifikaci parametrů stavových modelů.

10.1 Stavový model a formulace problému

Časově invariatní stavový model je dán následujícími rovnicemi

$$x_{k+1} = Fx_k + Gu_k + w_k \tag{10.1.1}$$
$$z_k = Hx_k + Ju_k + v_k \tag{10.1.2}$$

kde

- $x_k \in \mathbb{R}^{n_x}$ je neznámý stav systému dimenze n_x,
- $u_k \in \mathbb{R}^{n_u}$ je známý vstup dimenze n_u,
- $z_k \in \mathbb{R}^{n_z}$ je dostupné měření dimenze n_z,
- $w_k \in \mathbb{R}^{n_x}$ je neznámý stavový šum dimenze n_x,
- $v_k \in \mathbb{R}^{n_z}$ je neznámý šum v rovnici měření dimenze n_z,
- $F \in \mathbb{R}^{n_x \times n_x}$, $G \in \mathbb{R}^{n_x \times n_u}$, $H \in \mathbb{R}^{n_z \times n_x}$ a $J \in \mathbb{R}^{n_z \times n_u}$ jsou matice modelu (jmenovitě matice dynamiky, vstupní matice, výstupní matice a matice přímého působení vstupu na výstup),
- $k = 0, 1, 2, \ldots, \tau$ značí časový okamžik.

Rovnice (10.1.1) popisuje (modeluje) vývoj neznámého stavu v čase a nazývá se *stavová rovnice*. Rovnice (10.1.2) popisuje vztah mezi neznámým stavem a dostupným měřením a nazývá se *rovnice měření*.

Obecným *cílem* úlohy odhadu parametrů stavového modelu je nalézt nejen odhad matic F, G, H a J, ale i dimenzi stavového vektoru n_x statistickým zpracováním dostupných vstupně-výstupních dat $u^\tau = [u_0, u_1, \ldots, u_\tau]$ a $z^\tau = [z_0, z_1, \ldots, z_\tau]$.

Poznámka . Detailní informace a příklady stavového modelu lze nalézt ve druhém díle skript, který je věnován estimačním a identifikačním metodám založených na stavovém modelu. V druhém díle skript jsou tak rovněž uvedeny i postupy k převodu stavového modelu na vstupně-výstupní a naopak. Zatímco převod stavového modelu na vstupně-výstupní je jednoznačný, přechod ze vstupně-výstupního modelu vede na nekonečně mnoho modelů stavových. Proto, parametry stavového modelu nemusí být vždy fyzikálně interpretovatelné.

10.2 Přímá identifikace: Metoda podprostorů

Rozvoj identifikační metody podprostorů, v anglicky psané literatuře označované jako „subspace identification“, začal v devadesátých letech minulého století. Tyto metody umožňují odhad parametrů stavového modelu na základě výpočtu řádkových či sloupcových podprostorů vhodně strukturovaných matic obsahujících dostupná vstupně-výstupní data [61], [65].
V literatuře lze najít několik přístupů k návrhu této identifikační metody, které se v podstatě liší volbou hledaných podprostorů [66]. V této kapitole se budeme věnovat metodě v literatuře označované zkratkou MOESP (pocházející z anglického označení metody „multivariable output-error state space“) [61]. Metodu představíme ve třech krocích, nejprve pro odhad parametrů deterministického autonomního stavového modelu, pak metodu rozšíříme o možnost uvažování vstupního signálu a nakonec představíme identifikační metodu při explicitním uvažování poruch (tj. při uvažování stochastického stavového modelu).

10.2.1 Autonomní deterministický model: Ilustrace základního konceptu a idejí

Pro ilustraci základního konceptu identifikační metody podprostorů uvažujme nejprve deterministický autonomní model druhého řádu popsaný modelem

$$x_{k+1} = Fx_k \tag{10.2.3}$$
$$z_k = Hx_k \tag{10.2.4}$$

kde $n_x = 2$, $n_z = 1$. Předpokládejme dále, že systém je pozorovatelný, tj. (rozšířená) matice pozorovatelnosti

$$O_s = \begin{bmatrix} H \\ HF \\ HF^2 \\ \vdots \\ HF^{(s-1)} \end{bmatrix} \tag{10.2.5}$$

má plnou hodnost, tj. $rank(O_s) = n_x$, a $s \geq n_x$. Značení hodnosti matice vychází z pojmu „rank“ používaného v anglicky psané literatuře.

Identifikační metody podprostorů jsou založeny na výpočtu podprostorů (řádkového či sloupcového) matic v tzv. Hankelově struktuře[1] konstruovaných z dostupných dat. Uvažujme $\tau = 4$ dostupných měření $\{z_k\}_{k=0}^{4}$ a zvolme parametr $s = 3$. Pak Hankelova matice výstupu může být definována jako

$$Z = \begin{bmatrix} z_0, z_1, z_2 \\ z_1, z_2, z_3 \\ z_2, z_3, z_4 \end{bmatrix} \tag{10.2.6}$$

Hankelovu matici lze, vzhledem k popisu systému (10.2.3), (10.2.4), zapsat pomocí matice pozorovatelnosti (10.2.5) následujícím způsobem

$$Z = O_s X \tag{10.2.7}$$

kde rozšířená matice stavu X je dána

$$\begin{aligned} X &= [x_0, x_1, x_2] \\ &= [x_0, Fx_0, F^2 x_0] \end{aligned} \tag{10.2.8}$$

Klíčová myšlenka umožňující návrh identifikační metody je založena na skutečnosti, že Hankelova matice výstupu Z má hodnost shodnou s dimenzí stavu, tj.

$$rank(Z) = n_x \tag{10.2.9}$$

Důkaz vztahu (10.2.9) vychází ze Sylvestrova kriteria a je dán níže. Pak lze, bez dopadu na hodnost, vybrat z matic Z a X pouze n_x sloupců. Vzhledem k uvažovanému příkladu můžeme tedy vybrat např. první dva sloupce a psát (při použití notace programového prostředí MATLAB®)

$$Z(:, 1:2) = O_s X(:, 1:2) \tag{10.2.10}$$

$$\begin{bmatrix} z_0, z_1 \\ z_1, z_2 \\ z_2, z_3 \end{bmatrix} = \begin{bmatrix} O_s(1,1), O_s(1,2) \\ O_s(2,1), O_s(2,2) \\ O_s(3,1), O_s(3,2) \end{bmatrix} \begin{bmatrix} X(1,1), X(1,2) \\ X(2,1), X(2,2) \end{bmatrix} \tag{10.2.11}$$

kde značení $O_s(i,j)$ znamená prvek matice O_s v i-té řádce a j-tém sloupci a $O_s(:,j)$ znamená j-tý sloupec matice O_s, tj. $O_s(:,j) = [O_s(1,j), O_s(2,j), O_s(3,j)]^T$. Tedy, *čtvercovou* matici $X(1:2, 1:2)$ lze zapsat jako $X = [x_0, x_1] = [X(:,1), X(:,2)]$. Roznásobením pravé strany (10.2.11) získáme následující důležitý vztah

$$Z(:, 1:2) = [O_s(:,1)X(1,1) + O_s(:,2)X(2,1), O_s(:,1)X(1,2) + O_s(:,2)X(2,2)] \tag{10.2.12}$$

který říká, že sloupce matice výstupů Z jsou lineární kombinací sloupců matice pozorovatelnosti O_s. To znamená, že matice O_s a $Z(:, 1:2)$ mají shodný sloupcový podprostor, tj. matice jsou shodné až na transformační matici danou X. Sloupcový podprostor je v anglicky psané literatuře označován pojmem „range“ a tedy lze psát

$$range(Z(:, 1:2)) = range(O_s) \tag{10.2.13}$$

Všimněme si rovněž, že matice X je determinována počáteční podmínkou systému x_0 a má plnou hodnost n_x.

Vztahy (10.2.10) a (10.2.13) jsou *zásadními* vztahy, ze kterých vychází celý koncept identifikační metody podprostorů. Věnujme se jim proto podrobněji. Matice výstupu Z na levé straně

[1] Pro prvky Hankelovy matice Z platí $Z(i,j) = Z(i+k, j-k)$ pro $i \leq j$ a $k = 0, 1, \ldots, j-i$.

(10.2.10) je *známá*, protože je tvořena dostupnými měřeními. Naproti tomu matice pozorovatelnosti O_s a matice stavu X tvořící pravou stranu rovnice jsou *neznámé*, jelikož závisí na hledaných maticích popisu systému F, H a neznámém počátečním stavu.
Avšak ze vztahu (10.2.13) víme, že matice O_s a $Z(:, 1:2)$ mají stejný sloupcový podprostor a ten jsme schopni snadno určit singulárním rozkladem známé matice Z. Singulární rozklad v redukované formě, v anglicky psané literatuře označovaný pojmem „singular value decomposition" (SVD), matice $Z(:, 1:2)$ vede na

$$Z(:, 1:2) = USV^T \tag{10.2.14}$$

kde $U \in \mathbb{R}^{s \times n_x} = \mathbb{R}^{3\times 2}$ je ortonormální matice vlastních (bázových) vektorů reprezentující sloupcový podprostor matice, $V \in \mathbb{R}^{n_x \times s} = \mathbb{R}^{2\times 3}$ je ortonormální matice vlastních (bázových) vektorů reprezentující řádkový podprostor matice a $S \in \mathbb{R}^{n_x \times n_x} = \mathbb{R}^{2\times 2}$ je diagonální matice singulárních čísel. Dle (10.2.13) mají matice $Z(:, 1:2)$ a O_s stejný sloupcový podprostor a tedy lze, vzhledem k (10.2.10), psát

$$U = O_s T \tag{10.2.15}$$

kde T je *neznámá* regulární transformační matice. Vztah (10.2.15) pro matici U lze dále upravit jako

$$U = \begin{bmatrix} H \\ HF \\ HF^2 \end{bmatrix} T = \begin{bmatrix} HT \\ HFT \\ HF^2T \end{bmatrix} = \begin{bmatrix} HT \\ HTT^{-1}FT \\ HTT^{-1}FTT^{-1}FT \end{bmatrix} = \begin{bmatrix} H_T \\ H_T F_T \\ H_T F_T^2 \end{bmatrix} \tag{10.2.16}$$

kde matice

$$F_T = T^{-1}FT \tag{10.2.17}$$
$$H_T = HT \tag{10.2.18}$$

jsou transformované matice popisu systému (10.2.3), (10.2.4).

Ortogonální matice U, která je *známá*, je tak funkcí transformovaných matic popisu systému F_T a H_T. Transformované matice lze proto určit, s přihlédnutím k (10.2.16), následujícím způsobem:

(1) *transformovaná matice měření* H_T je, vzhledem k dimenzi měření $n_z = 1$, přímo dána první řádkou matice U, tj.

$$H_T = U(1, :) \tag{10.2.19}$$

(2) *transformovaná matice dynamiky* F_T je, vzhledem ke splněné podmínce $s > n_x$ a rovnici (10.2.16), dána řešením následující rovnice

$$\begin{bmatrix} H_T \\ H_T F_T \end{bmatrix} F_T = \begin{bmatrix} H_T F_T \\ H_T F_T^2 \end{bmatrix} \tag{10.2.20}$$
$$U(1:2, :) F_T = U(2:3, :) \tag{10.2.21}$$

Tedy, matici dynamiky spočteme v našem případě jako

$$F_T = (U(1:2, :))^{-1} U(2:3, :) \tag{10.2.22}$$

Zaměřme se na vztahy (10.2.7), (10.2.8), (10.2.14) a (10.2.15). Vztahy říkají, že

- *známá* matice výstupu Z je funkcí neznámé matice pozorovatelnosti O_s a neznámé matice stavu X,
- ortogonální matice U je funkcí známé matice výstupu stavu Z, a tudíž je *známá*,
- ortogonální matice U je funkcí neznámé matice pozorovatelnosti O_s a neznámé transformační matice T.

V důsledku můžeme říci, že transformační matice T je funkcí matice stavu X (tedy i neznámé počáteční podmínky systému x_0) a je neznámá. *Nelze* proto získat původní stavovou reprezentaci (10.2.3), (10.2.4) danou maticemi F a H. Avšak, obě reprezentace, tj. F, H a F_T, H_T, vedou na stejný vstupně-výstupní model a obě matice dynamiky mají tedy shodná vlastní čísla.

Definice 9.1.1 Sylvesterovo pravidlo (zjednodušené) [61]: Uvažujme dvě matice $A \in \mathbb{R}^{m\times n}$ a $B \in \mathbb{R}^{n\times p}$, kde $n \leq p$ a $n \leq m$, mající plnou hodnost n. Pak Sylvesterovo pravidlo říká, že

$$rank(AB) = n \tag{10.2.23}$$

Definice 9.1.2 Základní podprostory matice [61]: Uvažujme matici $A \in \mathbb{R}^{m\times n}$, která reprezentuje lineární transformaci z n-dimenzionálního prostoru do prostoru m-dimenzionálního. Pak existují čtyři základní podprostory této matice, a to

- sloupcový podprostor matice definovaný

$$range(A) = \{y \in \mathbb{R}^m | y = Ax, \forall x \in \mathbb{R}^n\} \tag{10.2.24}$$

- řádkový podprostor matice definovaný

$$range(A^T) = \{x \in \mathbb{R}^n | x = A^T y, \forall y \in \mathbb{R}^m\} \tag{10.2.25}$$

- jádro matice, v anglicky psané literatuře označované jako „kernel, null space“, definované

$$ker(A) = \{x \in \mathbb{R}^n | Ax = 0, \forall x \in \mathbb{R}^n\} \tag{10.2.26}$$

- levé jádro matice, v anglicky psané literatuře označované jako „cokernel, left null space“, definované

$$ker(A^T) = \{y \in \mathbb{R}^m | A^T y = 0, \forall x \in \mathbb{R}^n\} \tag{10.2.27}$$

Definice 9.1.3 Singulární rozklad [61, 67]: Každou matici $A \in \mathbb{R}^{m\times n}$ s hodnotí $rank(A) = r$ lze následujícím způsobem dekomponovat

$$A = USV^T \tag{10.2.28}$$

kde matice $U \in \mathbb{R}^{m\times m}$ a $V \in \mathbb{R}^{n\times n}$ jsou ortonormální matice, tj. $UU^T = I$ a $VV^T = I$, a matice $S \in \mathbb{R}^{m\times n}$ je diagonální matice s nezápornými reálnými čísly na diagonále, pro které platí

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_{min(m,n)} = 0 \tag{10.2.29}$$

Čísla $\{\sigma_i\}_{i=1}^r$ označujeme jako singulární čísla, sloupce matice U jako levé singulární vektory a sloupce matice V (tj. řádky matice V^T) jako pravé singulární vektory. Sloupce matice U tak můžou být chápány jako vlastní vektory matice AA^T a sloupce matice V jako vlastní vektory matice A^TA.

Pokud pro hodnost matice A označenou proměnou r platí $r < m$ a $r < n$, pak singulární dekompozice vede na speciální formu

$$A = \begin{bmatrix} U_r & U_{(m-r)} \end{bmatrix} \begin{bmatrix} S_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{(n-r)}^T \end{bmatrix} = U_r S_r V_r^T \tag{10.2.30}$$

kde $U_r \in \mathbb{R}^{m\times r}$, $U_{(m-r)} \in \mathbb{R}^{m\times(m-r)}$, $S_r \in \mathbb{R}^{r\times r}$, $V_r \in \mathbb{R}^{n\times r}$ a $V_{(n-r)} \in \mathbb{R}^{n\times(n-r)}$. Trojice matic U_r, S_r, V_r je označována jako singulární rozklad matice v *redukované* formě.

Poznámka . Singulární dekompozice, či rozklad, je numericky dobře podmíněná faktorizace, která je dostupná v mnoha statistických programech, jakým je např. MATLAB®. Dekompozice je obvykle dostupná jak v plné tak i redukované formě, viz funkce `svd`.

10.2.2 Autonomní deterministický model: Obecná metoda MOESP

Uvažujme obecný autonomní stavový model (10.2.3), (10.2.4) bez omezení na dimenzi stavu a měření. Definujme rovněž dva parametry s a N, které definují rozměr Hankelovy matice výstupu

$$Z_{0,s,N} = \begin{bmatrix} z_0 & z_1 & \dots & z_{N-1} \\ z_1 & z_2 & \dots & z_N \\ \vdots & \vdots & \ddots & \vdots \\ z_{s-1} & z_s & \dots & z_{N+s-2} \end{bmatrix} \tag{10.2.31}$$

a které splňují následující nerovnosti

$$s > n_x \tag{10.2.32}$$

$$N \geq s \tag{10.2.33}$$

Pak lze, podobně jako tomu bylo v případě (10.2.7), psát

$$Z_{0,s,N} = O_s X_{0,N} \tag{10.2.34}$$

kde

$$X_{0,N} = [x_0, Fx_0, F^2x_0, \dots, F^{N-1}x_0] \tag{10.2.35}$$

Na základě vztahu (10.2.34) lze za určitých mírných podmínek ukázat, že matice $Z_{0,s,N}$ a $O_s X_{0,N}$ mají stejné sloupcové podprostory, tedy

$$range(Z_{0,s,N}) = range(O_s X_{0,N}) \tag{10.2.36}$$

To znamená, že sloupcový podprostor Hankelovy matice je shodný se sloupcovým podprostorem matice pozorovatelnosti O_s a tedy lze psát

$$U_{n_x} = O_s T = \begin{bmatrix} H_T \\ H_T F_T \\ \vdots \\ H_T F_T^{s-1} \end{bmatrix} \tag{10.2.37}$$

kde $T \in \mathbb{R}^{n_x\times n_x}$ je neznámá transformační matice obecně neznámé dimenze n_x, $F_T = T^{-1}FT$, $H_T = HT$ a $U_{n_x} \in \mathbb{R}^{sn_z\times n_x}$ je matice levých singulárních vektorů spočtená následující redukovanou singulární dekompozicí

$$Z_{0,s,N} = U_{n_x} S_{n_x} V_{n_x}^T \tag{10.2.38}$$

Majíce spočtenou matici levých singulárních vektorů U_{n_x}, transformované matice systému F_T, H_T mohou být dle (10.2.37) určeny následovně:

(1) matice měření $H_T = U_{n_x}(1 : n_z, :)$,

(2) matice dynamiky $F_T = (U_{n_x}(1 : (s-1)n_z, :))^{\dagger} U_{n_x}(n_z + 1 : sn_z, :)$.

10.2.3 Deterministický model

Přejděme nyní k deterministickému stavovému modelu s uvažováním vstupu ve formě

$$x_{k+1} = Fx_k + Gu_k \tag{10.2.39}$$

$$z_k = Hx_k + Ju_k \tag{10.2.40}$$

Podobně jako v případě autonomního modelu, definujme nyní rovnici odpovídající Hankelově matici výstupu (10.2.31) pro model (10.2.39), (10.2.40). Rovnice pro matici výstupu nabývá následujícího tvaru

$$Z_{0,s,N} = O_s X_{0,N} + C_s U_{0,s,N} \tag{10.2.41}$$

kde matice O_s, $X_{0,N}$ jsou definovány shodně s předchozí částí, matice $U_{0,s,N}$ je Hankelova matice vstupu definovaná shodně s $Z_{0,s,N}$ a matice C_s je definována jako

$$C_s = \begin{bmatrix} J & 0 & 0 & \cdots & 0 \\ HG & J & 0 & \cdots & 0 \\ HFG & HG & J & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ HF^{s-2}G & HF^{s-3}G & \dots & HG & J \end{bmatrix} \tag{10.2.42}$$

Aby bylo možné použít podobný přístup pro výpočet transformovaných matic systému jako v případě autonomního modelu, je vhodné upravit rovnici (10.2.41) do formy (10.2.34), tj. „odstranit“ součet matic na pravé straně. To lze provést například s využitím projekční matice[2] $U^{\perp}_{0,s,N}$ definované jako

$$U^{\perp}_{0,s,N} = I_N - U^T_{0,s,N} \left(U_{0,s,N} U^T_{0,s,N}\right)^{-1} U_{0,s,N} \tag{10.2.43}$$

která má následující důležitou vlastnost

$$U_{0,s,N} U^{\perp}_{0,s,N} = 0 \tag{10.2.44}$$

Vynásobením (10.2.41) projekční maticí $U^{\perp}_{0,s,N}$ získáme, díky rovnosti (10.2.44), vztah

$$Z_{0,s,N} U^{\perp}_{0,s,N} = O_s X_{0,N} U^{\perp}_{0,s,N} \tag{10.2.45}$$

který je formálně shodný se vztahem (10.2.34) odvozeným pro autonomní model. Všimněme si, že vynásobením (10.2.41) projekční maticí $U^{\perp}_{0,s,N}$ jsme odstranili vliv vstupu na výstup.
Za předpokladu vhodného[3] vstupního signálu lze ukázat [61], že platí

$$rank(Z_{0,s,N} U^{\perp}_{0,s,N}) = rank(O_s) = n_x \tag{10.2.46}$$

a tedy, že sloupcové podprostory uvažovaných matic jsou shodné, tj.

$$range(Z_{0,s,N} U^{\perp}_{0,s,N}) = range(O_s) \tag{10.2.47}$$

Výpočet *transformovaných* matic modelu F_T, H_T je pak následující:

[2] Nutnou podmínkou pro výpočet inverze součinu matic $U_{0,s,N} U^T_{0,s,N}$ je podmínka (10.2.33). Aby však výsledná projekční matice nebyla nulová je nutné, aby matice $U_{0,s,N}$ nebyla čtvercová, tj. pro existenci nenulové projekční matice je nutná podmínka $N > s$.

[3] Pojem „vhodný signál“ bude diskutován později.

(1) Matice F_T, H_T vychází opět z redukované singulární dekompozice

$$Z_{0,s,N}U_{0,s,N}^{\perp} = U_{n_x}S_{n_x}V_{n_x}^T \tag{10.2.48}$$

a vypočteme je tedy shodně s předchozí částí z matice U_{n_x}.

(2) Při známých (tj. již odhadnutých) maticích F_T, H_T, výpočet matic G_T, J_T a vektoru[4] $x_{T,0}$ vychází ze soustavy *lineárních* rovnic

$$z_k = H_T F_T^k x_{T,0} + \left(\sum_{t=0}^{k-1} u_t^T \otimes H_T F_T^{k-t-1}\right)(G_T)_s + \left(u_k^T \otimes I_{n_z}\right)(J_T)_s \tag{10.2.49}$$

pro $k = 0, 1, 2, \ldots, \tau$, které přímo vychází z popisu modelu (10.2.39), (10.2.40). V rovnici (10.2.49) je použita notace $(A)_s$ označující vektor, který je tvořen sloupci matice A naskládanými pod sebe, a symbol $\otimes$ znamenající Kroneckerův součin[5] [69].

Poznámka . V odvození metody podprostorů pro model (10.2.39), (10.2.40) jsme předpokládali vhodný, tj. dostatečně bohatý vstupní signál u_k. Podobně jako tomu bylo u identifikace vstupně-výstupních modelů, i u identifikace stavových modelů je nutné, aby vstupní signál byl trvale budící daného řádu. Dostatečně bohatý vstupní signál zaručí, že existuje inverze matice $U_{0,s,N}U_{0,s,N}^T$, která je nutná pro výpočet projekční matice $U_{0,s,N}^{\perp}$ (10.2.43).
Detailnější diskuzi ohledně vstupního signálu lze najít v [61], kde je odvozena i verze identifikační metody podprostorů vhodná při uvažování vstupního signálu ve formě jednotkového impulsu.

Poznámka . Podobně jako tomu bylo u návrhu metody přídavné proměnné diskutované v kapitole 7, odvození metody podprostorů nevychází primárně z minimalizace kriteriální funkce. Avšak, jak bylo ukázáno v [61], získaný odhad matic systému lze chápat jako odhad minimalizující následující *kvadratickou* kriteriální funkci

$$\min_{C_s} \|Z_{0,s,N} - C_s U_{0,s,N}\| \tag{10.2.50}$$

Všimněme si, že matice C_s je funkcí všech matic systému F, G, H a J.

Poznámka . Singulární dekompozice použitá pro výpočet vlastních vektorů je spolehlivá metoda, která však pro velké Hankelovy matice může být velmi výpočetně náročná. Jako alternativu lze použít výpočetně přijatelnější RQ dekompozici [61].

10.2.4 Stochastický model

Doposud jsme při odvození metody podprostorů předpokládali deterministický model, tj. model, který nebyl ovlivněn šumem. V této sekci opět rozšíříme třídu uvažovaných modelů. Budeme uvažovat model s chybou výstupu, se kterým jsme se již setkali v kapitolách 2 a 7.
Uvažujme tedy stavový stochastický model

$$x_{k+1} = Fx_k + Gu_k \tag{10.2.51}$$
$$z_k = Hx_k + Ju_k + v_k \tag{10.2.52}$$

[4] Jako vedlejší produkt výpočtu matic je vypočten i neznámý počáteční stav transformovaného systému $x_{T,0} = T^{-1}x_0$.

[5] S Kroneckerovou algebrou se detailněji seznámíme v druhém díle skript.

pro který rovnice odpovídající Hankelově matici výstupu je ve tvaru

$$Z_{0,s,N} = O_s X_{0,N} + C_s U_{0,s,N} + V_{0,s,N} \tag{10.2.53}$$

kde nová matice $V_{0,s,N}$ je Hankelova matice šumu měření v_k definovaná analogicky k $Z_{0,s,N}$. Šum v rovnici měření je v tomto případě předpokládán jako absolutně náhodný signál nezávislý na vstupu u_k (tj. jako bílý šum).
V případě stochastického modelu s chybou výstupu (10.2.51), (10.2.52) lze použít analogický postup jako pro deterministický model (10.2.39), (10.2.40) představený v předchozí kapitole. Dle důkazu v [61] platí, že odhady matic F_T, G_T, H_T a J_T konvergují ke skutečným hodnotám s $N \to \infty$, tj. metoda podprostorů poskytuje v tomto případě konzistentní odhady.

Identifikační metoda podprostorů je značně teoreticky rozpracovaná. Metoda byla navržena i pro modely s explicitním uvažováním šumu v rovnici stavu nebo pro situace, kdy šumy nejsou bílé, ale jsou v čase či navzájem korelované. Vlastnosti šumů lze v některým případech taktéž identifikovat. Další informace k identifikační metodě podprostorů lze nalézt např. v [61], [65], [66].

10.2.5 Ilustrace metody podprostorů

Ilustrujme použití metody podprostorů pro identifikaci deterministického stavového modelu (10.2.39), (10.2.40) s maticemi

$$F = \begin{bmatrix} 0.8 & 0.5 \\ -0.7 & 0.1 \end{bmatrix},\ G = \begin{bmatrix} 1 \\ 1 \end{bmatrix},\ H = \begin{bmatrix} 0.5 & 1 \end{bmatrix},\ J = 0 \tag{10.2.54}$$

kde dimenze stavu a měření jsou $n_x = 2, n_z = 1, k = 0, 1, \ldots, \tau, \tau = 9$, počáteční stav $x_0 = [0, 0]^T$ a vstupní signál je generován jako gaussovský pseudo-náhodný proces s nulovou střední hodnotou a variancí 1, tj. $E[u_k] = 0$ a $var[u_k] = 1, \forall k$.
Použití metody podprostorů je svázáno s definicí dvou parametrů s a N, které ve svém důsledku definují dimenzi Hankelovy matice vstupu $U_{0,s,N}$ a výstupu $Z_{0,s,N}$. V tomto příkladě zvolme $s = 3$ a $N = 6$.
Identifikace parametrů stavového modelu je dána následujícími kroky:

(1) Z dostupných vstupních a výstupních dat z_k a u_k vytvořme Hankelovy matice $U_{0,s,N}$ a výstupu $Z_{0,s,N}$ ve struktuře (10.2.31).

(2) Dle (10.2.43) vypočtěme projekční matici vstupu $U^{\perp}_{0,s,N}$.

(3) Vypočtěme matice U, S a V singulární dekompozicí součinu matic $Z_{0,s,N}U^{\perp}_{0,s,N}$, která má stejný sloupcový prostor jako rozšířená matice pozorovatelnosti O_s (viz (10.2.45) a následná diskuze).

(4) Stanovme dimenzi identifikovaného modelu. Zde můžeme využít buď případnou apriorní znalost o systému nebo můžeme analyzovat singulární čísla uspořádané na diagonále matice S. Za vhodný řád systému/modelu můžeme považovat počet významných (tj. nenulových) singulárních čísel. V našem příkladě, pro uvažovanou realizaci, nabývá matice S následujících hodnot

$$S = \begin{bmatrix} 2.68 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.45 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{10.2.55}$$

a tedy řád modelu zvolíme $n_x = 2$.

(5) Určeme redukovanou formu U_r matice levých singulárních vektorů U dle (10.2.30), tj.

$$U_r = \begin{bmatrix} 0.94 & -0.14 \\ 0.31 & 0.70 \\ -0.13 & 0.69 \end{bmatrix} \tag{10.2.56}$$

Matice U_r má n_x sloupců a dle (10.2.16) má stejný sloupcový podprostor jako matice pozorovatelnosti O_s. Tudíž, pro $s = 3, n_z = 1$, platí

$$U_r = \begin{bmatrix} H_T \\ H_T F_T \\ H_T F_T^2 \end{bmatrix} \tag{10.2.57}$$

kde $H_T = HT$, $F_T = T^{-1}FT$ a T je neznámá transformační matice.

(6) Ze vztahu (10.2.57) plyne, že transformovaná matice měření H_T je dána první řádkou matice U_r, tj.

$$H_T = U_r(1,:) = [0.94, -0.14] \tag{10.2.58}$$

a transformovanou matice dynamiky F_T lze spočítat vzhledem ke struktuře (10.2.57) a vztahu (10.2.22) jako

$$F_T = (U_r(1:2,:))^{-1} U_r(2:3,:) = \begin{bmatrix} 0.28 & 0.84 \\ -0.30 & 0.62 \end{bmatrix} \tag{10.2.59}$$

(7) Poslední transformovanou matici G_T lze dopočítat z (10.2.49) vedoucí, pro první čtyři časové okamžiky, na následující soustavu rovnic (Kroneckerův součin přechází v standardní součin pro skalární vstup $u(t)$)

$$\begin{bmatrix} z_0 \\ z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} H_T & 0_{1\times 2} \\ H_T F_T & u_0 H_T \\ H_T F_T^2 & u_0 H_T F_T + u_1 H_T \\ H_T F_T^2 & u_0 H_T F_T^2 + u_1 H_T F_T + u_2 H_T \end{bmatrix} \begin{bmatrix} x_{T,0} \\ G_T \end{bmatrix} \tag{10.2.60}$$

Výsledný odhad matice G_T je

$$G_T = \begin{bmatrix} 1.50 \\ -0.59 \end{bmatrix} \tag{10.2.61}$$

Všimněme si, že vedlejší efekt odhadu G_T je i odhad transformovaného počátečního stavu $x_{T,0} = T^{-1}x_0$.

Poznamenejme také, že můžeme samozřejmě sestavit tolik lineárních rovnic (10.2.49), kolik máme měření, tj. τ. V našem příkladě však uvažujeme deterministický systém a pro odhad dvou neznámých veličin, jmenovitě $x_{T,k}(0), G_T$, které obsahují celkem čtyři prvky, jsou postačující právě čtyři rovnice.

Skutečné matice F, G, H (10.2.54) a identifikované transformované matice F_T (10.2.59), G_T (10.2.61), H_T (10.2.58) nejsou číselně stejné. Jsou však svázány přes transformační matici T, kterou lze snadno vypočítat ze znalosti původních a transformovaných matic a s přihlédnutím ke vztahu (10.2.37) jako

$$T = O_s^{\dagger} U_{n_x} = \begin{bmatrix} 0.04 & -1.59 \\ 0.92 & 0.65 \end{bmatrix} \tag{10.2.62}$$

Oba stavové modely samozřejmě také vedou na stejný vstupně-výstupní model v ARX struktuře

$$z_k - 0.9 z_{k-1} + 0.43 z_{k-2} = 1.5 u_{k-1} - 1.3 u_{k-2} \tag{10.2.63}$$

a obě matice dynamiky F a F_T mají shodná vlastní čísla.

10.3 Nepřímá identifikace

Nepřímá identifikace spočívá v identifikaci vstupně-výstupního modelu (např. ve struktuře ARX, ARMAX) a jeho následném převodu na model stavový. Převod je detailně diskutován ve druhém díle skript.

10.4 Metody odhadu stavu v úloze identifikace

Další možnost odhadu parametrů stavového modelu spočívá ve využití technik z oblasti odhadu stavu (např. využití Kalmanova filtru). Tyto techniky jsou vhodné pro identifikaci parametrů široké množiny modelů, ať již vstupně-výstupních či stavových, lineárních či nelineárních, popř. časově variantních či invariantních. Detailní popis metod odhadu stavu spolu s diskuzí o jejich využití v identifikaci lze nalézt v druhém díle těchto skript.

10.5 Shrnutí a zhodnocení výsledků

V této kapitole byla upřena pozornost na přestavení stavových modelů a identifikaci jejich parametrů zejména metodou podprostorů. Kapitola tak může být brána jako pozvolný přechod k druhému dílu skript, které se věnují algoritmům využívající stavové modely.
Metoda podprostorů je relativně nová identifikační metoda, která byla v posledních letech značně rozpracována jak v teoretické, tak i algoritmické rovině. V literatuře lze proto najít nepřeberné množství verzí metody podprostorů lišících se zejména výpočtem ortogonálních matic. Velký zájem o tyto metody byl rovněž motivací pro vznik mnoha programových nástrojů, kde různé verze metody podprostorů byly implementovány. Tyto nástroje jsou volně dostupné pro populární (statistické) programové prostředí, jakým je např. produkt MATLAB® nebo jazyk Python. Jako vhodnou literaturu pro další studium identifikace stavových modelů lze doporučit např. knihy [61], [65].

Kapitola 11

Závěr

Obsah tohoto dílu skript můžeme chápat jako úvod do problematiky identifikace systémů. Po úvodních motivačních příkladech byly krátce představeny neparametrické metody. Hlavní pozornost byla věnována parametrické jednorázové identifikaci lineárních stochastických systémů. Odvozené algoritmy jednorázové identifikace pak byly převedeny do více atraktivního rekurzivního rámce.

Pro studium i aplikace identifikace je velmi důležité mít nejen teoretické zázemí, ale i kvalitní programové vybavení umožňující zpracování dat, použití identifikačních metod, grafické znázornění průběhů sledovaných veličin a další služby pokud možno v interaktivním režimu. Za takový prostředek můžeme dnes považovat produkt MATLAB® se speciálními programovými balíky na identifikaci. Samostatné zkušenosti s realizací identifikačních algoritmů získané např. při laboratorních cvičeních výrazně pomáhají pochopit vlastnosti metod a vedou k získání potřebné jistoty na poli identifikace.

Pro hlubší teoretické poznání identifikačních přístupů a jednotlivých metod je nutné věnovat mnohem větší pozornost než v těchto skriptech teoretické analýze, umožňující najít podmínky konsistence odhadů, podmínky identifikovatelnosti při použití zpětné vazby, řád modelu atd. Zdárná aplikace identifikačních přístupů vyžaduje testování správnosti předpokladů spojených s určitým postupem. Sem patří test linearity, časové invariance, existence zpětné vazby atd.

Stále větší význam má identifikace systémů v oblasti zpracování signálů, časových řad, detekce chyb, adaptivní predikci a adaptivního řízení. Jádrem každého adaptivního systému ať už z oblasti zpracování signálů nebo automatického řízení je právě identifikační algoritmus reprezentující nutnost kontinuálního poznávání.

Literatura

[1] Strejc, V.: Teorie automatického řízení II (přednášky). Skriptum ČVUT Praha, 1988, 205s.

[2] Šutek, L.-Varga, M.: Experimentální metódy identifikácie. Veda, Bratislava, 1981, 197s.

[3] John, J.-Horáček, P.: Identifikace a modelování. Skriptum ČVUT Praha, 1982, 119s.

[4] Eck, V.: Identifikace a modelování. Skriptum ČVUT Praha, 1989, 215s.

[5] Hudzovič, P.: Identifikácia a modelovanie. Skriptum SVŠT Bratislava, 1981s.

[6] Murgaš, J.-Hejda, I.: Adaptivne riadenie technologickych procesov. Skriptum STU Bratislava, 1993, 176s.

[7] Šimandl, M.: Adaptivní systémy. Skriptum ZČU Plzeň, 1993, 133s.

[8] Beneš, J.-Žampa, P.: Stochastické systémy a jejich řízení, Skriptum ČVUT Praha, 1976, 201s.

[9] Strejc, V.: Stavová teorie lineárního diskrétního řízení. Academia Praha, 1978, 374s.

[10] Havlena, V.-Štecha, J.: Moderní teorie řízení, Skriptum ČVUT Praha, 1994, 289s.

[11] Peterka, V.: Číslicové řízení procesů s náhodnými poruchami a neurčitými charakteristikami. Doktorská práce. ÚTIA Praha, 1975.

[12] Hušek, R.: Základy ekonometrie. Skriptum VŠE Praha, 1986, 219s.

[13] Aström, K.J.-Eykhoff, P.: System identification - a survey. Automatica, Vol. 7, 31–38.

[14] Eykhoff, P.: System identification: Parameter and State Estimation. Wiley, London, 1974.

[15] Ljung, L.: System Identification: Theory for the User, 2nd Edition. Prentice Hall, New Jersey, 1999.

[16] Norton, J.P.: An Introduction to Identification. Academic Press, New York, 1986.

[17] Goodwin, G.C.-Payne, R.L.: Dynamic System Identification: Experiment Design and Data Analysis. Academic Press, New York, 1977.

[18] Ljung, L.-Södeström, T.: Theory and Practice of Recursive Identification. MIT Press, Cambridge, 1983.

[19] Söderström, T.-Stoica, P.: Instrumental variable methods for System Identification. (Lecture Notes in Control and Information Sciences, 57). Springer Verlag, Berlin, 1983.

[20] Söderström, T.-Stoica, P.: System Identification, Prentice Hall, New York, 1989.

[21] Isermann, R.: Identification Dynamischer Systeme. Springer Verlag, Berlin, 1987.

[22] Wellstead, P.E.: Introduction to Physical System, System Modelling . Academic Press, London, 1979.

[23] Marcus-Roberts, H.-Thompson, M. (eds.): Life Science Models. Springer Verlag, New York, 1976.

[24] Ljung, L.: On the consistency of prediction error identification methods. In: R.K. Mehra, D.G.Laimotis (eds.), System Identification - Advances and Case Studies. Academic Press, New York, 1976.

[25] Glover, K.: Identification: Frequency-domain methods. In: M.Singh (ed.) Systems and Control Encyklopedia, Pergamon, Oxford, 1987.

[26] Rake, H.: Identification: transient-and frequency-responce methods. In: M.Singh (ed.), Systems and Control Encyclopedia, Pergamon, Oxford, 1987.

[27] Jenkins, G.M.-Watts, D.G.: Spectral Analysis and Its Applications. Holden-Day, San Francisco, 1969.

[28] Priestley, M.B.: Spectral Analysis and Time Series. Academic Press, London, 1982.

[29] Wellstead, P.E.: Non-parametric methods of System Identification, Automatica, Vol. 17, 55–69.

[30] Bergland, G.D.: A guided tour of the fast Fourier transform, IEEE Spectrum, Vol. 6, 41–52.

[31] Bendat, J.S.-Piersol, A.G.: Engineering Applications of Correlation and Spectral Analysis. Wiley-Interscience, New York, 1980.

[32] Rao, R.C.: Lineární metody statistická indukce a jejich aplikace. Academia Praha, 1978, 666s.

[33] Brockwell, P.J., Davis R.A.: Time Series: Theory and Methods. Springer Verlag, 1991, 577s.

[34] Maddala, G.S.: Econometrics. Mc GRAW-HILL, London, 1988, 515s.

[35] Meloun, M.-Militky, J.: Statistická zpracování dat. Edice Plus, Praha, 1994.

[36] Cipra, T.: Analýza časových řad v ekonomii. SNTL/ALFA, Praha, 1986.

[37] Goodwin, G.C.-Sin, K.S.: Adaptive Filtering, Prediction and Control, Prentice Hall, Englewood Cliffs, 1984.

[38] Guidorzi, R.: Invariants and canonical forms for systems structural and parametric identification. Automatica, Vol. 17, 117–133.

[39] Kučera, V.: Discrete Linear Control. Wiley, Chichester 1979.

[40] Gevers, M.-Wetz, V.: Parametrization issues in system identification. Preprints of IFAC 10th World Congress, Münich, 1987.

[41] Nguyen, V.-Wood, E.: Review and unification of linear identifiability concepts. SIAM Review, Vol. 24, 34–51.

[42] Ljung, L.: Convergence analysis of parametric identifications methods. IEEE Transactions on Automatic Control, Vol. AC-23, 770–783.

[43] Caines, P.: Prediction error identification methods for stationary stochastic processes, IEEE Transactions on Automatic Control, Vol. AC-21, 1976, 500–505.

[44] Dennis, J.-Schnabel, R.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice Hall, Englewood Cliffs, 1983.

[45] Wong, K.-Polak, E.: Identification of linear discrete time systems using the instrumental variable approach. IEEE Transactions on Automatic Control, Vol. AC-12, 1967, 707–718.

[46] Young, P.: Some observations on instrumental variable methods of time series analysis. International Journal of Control, Vol. 23, 1976, 593–612.

[47] Stoica, P.-Söderström, T.: Optimal instrumental variable estimation and approximate implementation. IEEE Transactions on Automatic Control, Vol. AC-28, 1983, 757–772.

[48] Saridis, G.: Comparison of six on-line identification algorithms, Automatica, Vol. 10, 1974, 69–79.

[49] Aström, K.-Wittenmark, B.: Adaptive Control. Addison-Wesley, 1988

[50] Widrow, B.-Stearns, S.: Adaptive Signal Processing, Prentice Hall, Englewood Cliffs, 1985.

[51] Ljung, L.: Analysis of a general recursive prediction error identification algorithm. Automatica, Vol. 17, 1981, 89–100.

[52] Bierman, G.: Factorization Methods for Discrete Sequential Estimation. Academic Press, New York, 1977.

[53] Basseville, M.-Benveniste, A. (eds.): Detection of Abrupt Changes in Signals and Dynamical Systems. Springer Verlag, Berlin, 1986.

[54] Peterka, V.: A square root filter for real time multivariable regression. Kybernetika, Vol. 11, 1975, 53–67.

[55] Kulhavý, R.: Restricted Exponencial Forgetting in Real-time Identification. Automatica 23, 1987, 589–600.

[56] Kárný, M. a kol.: Design of Linear Quadratic Adaptive Control: Theory and Algorithms for Practice - Suplement to Kybernetika, Vol. 21, No 3,4,5,6, 1985.

[57] Ralston, A.: Základy numerické matematiky. Academia, Praha 1973, 635s.

[58] Štecha, J.: Teorie automatického řízení I. Skriptum ČVUT Praha, 1990.

[59] Spíral, L.-Rada, V.-Žampa, P.:Experimentální vyšetření a vyhodnocení regulačních obvodů. Skriptum VŠSE Plzeň, 1968.

[60] Simon, D.: Optimal State Estimation: Kalman, H-infinity, and Nonlinear Approaches, CRC Press, 2012.

[61] Verhaegen, M.-Verdult, V.: Filtering and System Identification, Cambridge University Press, 2007.

[62] Gibbs, B.: Advanced Kalman Filtering, Least-Squares and Modelling, Wiley, 2011.

[63] Crassidis, J. L.-Junkins, J. L.: Optimal Estimation of Dynamic Systems, CRC Press, 2012.

[64] Björck, Å: Numerical Methods for Least-Squares Problems, SIAM, 1996.

[65] van Overshee, P. and de Moor, B.: Subspace Identification for Linear Systems: Theory-Implementation-Applications, Kluwer Academic Publishers, 1996.

[66] Qin, S. J.: An overview of subspace identification. Computers and Chemical Engineering, Vol. 30, 2006, 1502–1513.

[67] Strang, G.: Introduction to Linear Algebra, 4th Edition, Wellesley–Cambridge Press, 2009.

[68] Shynk, J. J.: Probability, Random Variables, and Random Processes: Theory and Signal Processing Applications, Wiley, 2013.

[69] Brewer, J. W.: Kronecker products and matrix calculus in system theory. IEEE Transactions on Circuits and Systems, Vol. 25, No. 9, 1978, pp. 772—781.

[70] Billings, S. A.: Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temoporal Domains, Wiley, 2013.

[71] Isermann, R.-Münchhof, M.: Identification of Dynamic Systems: An Introduction with Applications, Springer, 2011.

[72] Haykin, S.: Neural Networks: A Comprehensive Foundation, Prentice Hall, 1998.

[73] Haykin, S.: Kalman Filtering and Neural Networks, Wiley-Interscience, 2001.

DÍL DRUHÝ
FILTRACE

Obsah

Kapitola 1

Úvod

V tomto dílu se v převážné míře budeme zabývat úlohou estimace stochastických procesů, která je alternativně nazývána úloha filtrace. Estimaci parametrů nebo též parametrickou identifikaci bychom mnohdy mohli chápat jako speciální případ úlohy filtrace.

Vývoj teorie filtrace můžeme rozdělit do tří vývojových etap. První etapa je založena na Wiener-Kolmogorovově teorii a je reprezentována především Wienerovým filtrem odvozeným v práci [1] pro stacionární stochastické procesy. Hlavními prostředky k syntéze filtru byly spektrální faktorizace, Fourierova transformace a integrální rovnice. Uvedené matematické prostředky komplikovaly jeho širší pochopení a aplikaci Wienerovy filtrace. Základem pro druhou etapu, která se zabývá i nestacionárními stochastickými procesy a užívá stavovou teorii je Kalmanova teorie filtrace [2]. Třetí etapa se zabývá teorií a aproximačním řešením problému nelineární filtrace [3], [4], [5], jejíž další vývoj podnítily významnou měrou sympozia o nelineární filtraci a jejich aplikacích [6], [7].

Rostoucí zájem o nelineární filtraci však neznamená, že první dvě etapy jsou uzavřeny. Např. technika syntézy Wienerova filtru je nyní založena na polynomiálním přístupu [8], [9] a jeho použití je velmi populární pro zpracování a estimaci signálů. Rovněž Kalmanův filtr, obecněji, kalmanovský přístup má široké uplatnění a je hluboce rozpracováván [9]-[12], [84]. Tento přístup proniká i do oblastí zdánlivě mimo oblast automatického řízení či teorie systémů jako je např. ekonometrie a finanční management, kde pro modelování časových řad se používá stavová teorie a k estimaci následně Kalmanův filtr [13]. Linearita a gaussovost, což jsou předpoklady pro exaktní použití tohoto přístupu, jsou však velmi omezující pro další širší uplatnění. Proto bude v této práci věnována značná pozornost situacím, které překračují rámec linearity a gaussovosti. Estimace stavu nelineárních a negaussovských systémů vede na problém nelineární estimace. Přístupy, které se tímto problémem zabývají můžeme rozdělit na lokální a globální [14], pokud sledujeme hledisko platnosti získaných výsledků ve stavovém prostoru nebo na analytické a numerické, pokud sledujeme způsob řešení bayesovských rekurzivních vztahů [15].

V této práci se soustředíme jak na analytický, tak i numerický přístup a Kalmanův filtr nám mnohdy poslouží často jako základní stavební kámen při návrhu estimačních algoritmů pro nelineární a negaussovské systémy. Uvedený postup zároveň vytvoří příznivé podmínky i pro řešení úloh identifikace, detekce chyby, odhadů skokově se měnících parametrů, modelování hrubých chyb měření, syntézy robustních filtrů a dalších.

Cílem tohoto dílu skript je vytvořit ucelený pohled na modelování a estimaci stavu (diskrétních) stochastických systémů v rámci analytického přístupu k syntéze estimačních algoritmů. Při

zpracování tématu použijeme zpravidla induktivní přístup. Základem pro syntézu všech estimačních algoritmů bude bayesovský přístup. Rovněž snaha o plynulý přirozený přechod od lineární filtrace k nelineární filtraci bude evidentní.

Tento díl je členěn do devíti kapitol a jejich obsah je následující. Ve druhé kapitole je formulována úloha estimace stavu pro obecný diskrétní stochastický systém. Třetí kapitola se zabývá odvozením Kalmanova filtru z bayesovských vztahů, protože použití standardních postupů využívajících podmínky ortogonality nebo využití známých vztahů o náhodných veličinách nejsou z pohledu záměrů této práce konstruktivní. Syntéza estimačních algoritmů pro složitější problémy bude využívat řadu vztahů a obratů právě z tohoto odvození. Čtvrtá kapitola bude věnována klasickým i moderním lokálním nelineárním filtrům, a to postupně rozšířenému Kalmanovu filtru, filtru druhého řádu, diferenčnímu filtru, unscentovanému filtru i iteračnímu filtru. V páté kapitole odvodíme jednak globální filtr s vícenásobnou linearizací, jednak filtr pro negaussovský systém. Šestá kapitola je zaměřena na modelování specifických jevů takových, jako jsou skokové změny stavu či parametrů, modelování hrubých chyb měření a následně pak se zabývá odhadem stavu systémů obsahujících tyto speciální jevy. Sedmá kapitola je věnována numerickému řešení bayesovských vztahů a je odvozena metoda bodových mas. V osmé kapitole jsou ukázány další možnosti použití nelineární filtrace jako např. při rozhodování. Taktéž je ukázána spojitost mezi oblastí identifikace a filtrace. Konečně v deváté kapitole je provedeno stručné zhodnocení.

Kapitola 2

Problém modelování a estimace

Cílem modelování systémů v technických i netechnických oblastech a z pohledu této práce především dynamických systémů je postavit model systému. Tímto úkolem se zabývá jednak matematické modelování, které využívá známých fyzikálních, ekonomických a dalších zákonů a jednak identifikace systémů založená na zpracování experimentálních dat. Mnohdy nejúčelnější bývá kombinace těchto přístupů.
Modely jsou základním opěrným bodem v úlohách rozhodování, estimace nebo-li odhadu a řízení. Úspěšnost a složitost řešení těchto úloh je závislá na vhodně postaveném modelu. Modely dynamických systémů mohou mít různý charakter podle způsobu pohledu a popisu. Můžeme například používat modely verbální (učitel autoškoly popisuje slovně chování auta), modely ve formě grafů a tabulek nebo modely založené na matematických rovnicích. Právě posledně jmenované, konkrétněji diferenční stochastické rovnice, budeme nejčastěji používat v této práci, i když někdy budou motivovány verbálním popisem sledovaných jevů.
S ohledem na zaměření tohoto dílu se nyní stručně budeme zabývat vývojem estimační teorie včetně modelů, které jsou součástí formulace estimačních úloh.

2.1 Základní úvahy

Estimací parametrů modelu systému využitím pozorovaných dat se zabývali již babylónští astronomové 300 let před naším letopočtem. Rovněž později byly astronomické studie stimulem pro rozvoj této disciplíny. Připomeňme alespoň jména Cotes, Euler, Bernoulli. Důležitý výsledek pro estimační teorii a pro tuto práci byl dán Thomasem Bayesem v roce 1761, který formuloval Bayesovo pravidlo. Základy nejznámější a nejpoužívanější metody odhadu metody nejmenších čtverců položil v roce 1806 Legendre a 1809 Gauss. V roce 1835 Cauchy zkoumal problém řešení soustavy lineárních rovnic a ukázal, že čtvercovou nesymetrickou matici lze zapsat jako součin dvou trojúhelníkových matic. Za významný příspěvek do teorie odhadu lze jistě pokládat práce Fishera, zvláště pak metodu maximální věrohodnosti z roku 1911.
Výrazný zlom v náhledu na estimační problém chápaný jako problém estimace parametrů nastává ve 40tých letech s příchodem Wiener Kolmogorovovy teorie a následné aplikaci v komunikační a vojenské technice. Wiener [1] chápe pozorovaná data jako součet signálu a šumu a předpokládá, že signál i šum jsou stacionární náhodné procesy. Estimátor-filtr snažící se oddělit signál od šumu z měřených dat a generující odhadovaný signál je navrhován pomocí Fourierovy transformace a spektrální faktorizace. Vznik stavové teorie systémů v rámci teorie automatického řízení, která nastupuje po klasické teorii řízení založené na vstupně-výstupních charakteristikách se odrazil i v estimační teorii. V roce 1960 publikoval Kalman článek o

rekurzivní filtraci [2] využívající stavový popis systému poskytující optimální odhad stavu pro t-variantní lineární gaussovské systémy. Můžeme říci, že s méně náročným matematickým aparátem řeší složitější problém než Wiener. Historický vývoj estimace i s bibliografickými údaji je kvalitně popsán v [16].

Po tomto stručném velmi obecném představení vývoje teorie estimace se nyní budeme věnovat úloze estimace a modelování konkrétněji. Uvažujme jako přirozený základ pro diskusi o odhadu parametrů a odhadu stavu následující rovnici

$$z_k = y_k + v_k \quad k = 0, 1, ... \tag{2.1.1}$$

kde z_k je měření na systému dostupné v čase t_k
y_k je výstup systému v čase t_k
v_k reprezentuje šum měření neboli poruchu v čase t_k

Problém je najít odhad y_k z měření z_k pro všechna k. Je zřejmé, že rozumný odhad by mohl být

$$\hat{y}_k \stackrel{\triangle}{=} z_k \tag{2.1.2}$$

Chyba odhadu by v tomto případě byla

$$\tilde{y}_k \stackrel{\triangle}{=} y_k - \hat{y}_k = z_k - v_k - z_k = -v_k \tag{2.1.3}$$

Cíl estimační teorie je navrhnout takový postup při stanovení odhadu, který bude redukovat chybu odhadu na menší, než kterou dává primitivní estimátor (2.1.2). Chyba $\tilde{y}_k$ nemůže být určena explicitně, protože ani y_k ani v_k nejsou známy. Proto je třeba znát více o signálu y_k.
Předpokládejme nejdříve, že $y_k = \theta$ tj. výstup je konstantní signál. Pak měření vyhovuje rovnici

$$z_k = \theta + v_k \quad k = 0, 1, 2, ..., N \tag{2.1.4}$$

Odhad θ můžeme provést následujícím způsobem. Sečteme všechna měření a součet vydělíme jejich počtem, tedy

$$(\sum_{k=0}^{N} z_k)/(N+1) = \theta + (\sum_{k=0}^{N} v_k)/(N+1)$$

Rozumný odhad θ by pak mohl být

$$\hat{\theta} \stackrel{\triangle}{=} (\sum_{k=0}^{N} z_k)/(N+1) \tag{2.1.5}$$

a chyba odhadu

$$\tilde{\theta} = \theta - \hat{\theta} = -(\sum_{k=0}^{N} v_k)/(N+1)$$

Jestliže však v_k je také konstantní pro všechna k tj. $v_k = v$, pak k žádnému zlepšení odhadu vzhledem k odhadu definovanému v (2.1.2) nedojde, přestože jsme věděli více o výstupu y_k a

použili statistický přístup. Ukazuje se, že je důležité pro zlepšení odhadu, aby signál a šum měli rozdílné charakteristiky.
Z předchozího je zřejmé, že velikost chyby odhadu může být redukována průměrováním měření pouze za předpokladu, že šum bude mít např. tu vlastnost, že bude měnit znaménko. Kdyby šum měl průměrnou hodnotu nula a počet měření by byl dostatečně velký, pak odhad bude blízko průměrné hodnotě měření. Jinými slovy, tento postup při syntéze estimátoru transformujícího měření na odhadované parametry může významně zmenšit chybu odhadu, redukovat vliv šumu na odhad, i když šum bude zatěžovat měření velmi rozdílnými hodnotami. Ukazuje se, že problém modelování v souvislosti s úlohou odhadu je vhodné chápat strukturálně, tedy z pohledu vztahů mezi veličinami a zároveň je nutné věnovat pozornost i bližší specifikaci poruch, např. využitím teorie pravděpodobnosti.

2.2 Strukturální modelování

Rozšiřme úvahy z předcházející části, kdy výstup systému byla konstanta a rovnice (2.1.1) obsahovala pouze skalární veličiny, na situaci popsanou rovnicí

$$y_k = h_k(\Theta), \quad k = 0, 1, 2, \ldots \tag{2.2.1}$$

kde $h_k(\cdot)$ je známá vektorová funkce a y_k, z_k a v_k jsou vektory. Často se rovněž předpokládá speciální případ ve tvaru

$$y_k = H_k\Theta, \quad k = 0, 1, 2, \ldots \tag{2.2.2}$$

kde H_k je známá matice příslušných dimenzí. Rovnice (2.2.1) definuje pak nelineární model pro problém estimace parametrů Θ a (2.2.2) lineární problém.

Mnohdy však není možné uvažovat neznámé jako konstantní veličiny, protože se jedná o proměnné v čase, a tudíž výstupní signál by měl být reprezentován spíše takto

$$y_k = h_k(x_k) \tag{2.2.3}$$

kde $h_k(\cdot)$ je opět známá funkce a vývoj vektorové proměnné x_k je popsán diferenční rovnicí

$$x_{k+1} = f_k(x_k) + w_k \tag{2.2.4}$$

kde $f_k(\cdot)$ je známá vektorová funkce
w_k je neznámý vektor reprezentující šum v čase t_k.

Proměnná x_k je označována jako stav a (2.2.3), (2.2.4) definují model pro problém nelineární estimace stavu. Poznamenejme však, že takto vymezený problém nelineární estimace stavu není zcela vyčerpávajícím způsobem charakterizován a k této problematice se ještě vrátíme po přesnějším vymezení stavu x_k a šumů $\{w_k\}, \{v_k\}$. Šum w_k ve stavové rovnici reprezentuje neznámé okolnosti, neurčitosti ovlivňující dynamický vývoj stavu a z pohledu teorie řízení ho můžeme chápat jako neznámý, neměřitelný vstupní signál působící na systém definovaný rovnicemi (2.2.3), (2.2.4).

Problém lineární estimace stavu pak může vzniknout jako speciální případ předchozí situace a je založen na modelu

$$x_{k+1} = F_k x_k + w_k \tag{2.2.5}$$

$$y_k = H_k x_k \tag{2.2.6}$$

kde F_k a H_k jsou známé matice příslušných dimenzí.

Je zřejmé, že problém odhadu stavu a odhadu parametrů spolu úzce souvisí. Problém estimace stavu zavádí přídavné struktury pro vývoj stavu v čase. Povšimněme si, že pro

$$x_{k+1} = x_k = \Theta \tag{2.2.7}$$

přechází (2.2.4) na (2.2.1). Dále, jestliže

$$x_{k+1} = f_k(x_k) \tag{2.2.8}$$

tedy stavový šum je nulový pro všechna k, pak problém estimace stavu může být rovněž chápán jako problém estimace parametrů. Vysvětlení je jednoduché. Protože x_k může být formálně vyjádřeno jako funkce počátečního stavu

$$x_k = \phi_{k,0}(x_0) \tag{2.2.9}$$

pak parametr lze chápat jako počáteční stav zavedený do (2.2.1).
Na závěr této sekce poznamenejme, že šum v_k v rovnici měření (2.1.1) i stavový šum w_k ve (2.2.4) působí aditivně. Zobecnění je jistě možné, např. vývoj stavu by byl popsán vztahem

$$x_{k+1} = f_k(x_k, w_k)$$

kde $f_k(\cdot, \cdot)$ je známá funkce. Toto zvýšení obecnosti však nese i značné zvýšení složitosti řešení estimační úlohy. Obdobně rovnici měření (2.1.1) lze zobecnit na tvar

$$z_k = h_k(x_k, v_k)$$

kde $h_k(\cdot, \cdot)$ je známá funkce. S dalším alternativním vyjádřením, které se využívá jako základní struktura v rámci identifikace systémů a zpracování signálů se setkáme ve 3. kapitole.

Dosud jsme se věnovali modelování struktury systémů, jakožto významného aspektu modelování a konstituování úlohy estimace. Poznamenejme, že metoda nejmenších čtverců využívá právě strukturální vlastnosti, což způsobuje jednoduchost při aplikaci, ale zároveň vznikají problémy s kvalitativním ohodnocením přesnosti odhadu, jelikož poruchy nejsou specifikovány. V následující sekci se budeme věnovat specifikaci vlastností poruch, tedy šumu měření a stavového šumu.

2.3 Pravděpodobnostní modelování

K popisu neurčitosti se tradičně používá počtu pravděpodobnosti. Alternativní možností je množinový přístup [17], [18], případně jiné techniky. V této práci budeme využívat výhradně pravděpodobnostní přístup.

Z předchozí sekce o strukturálním modelování vyplývá, že bude existovat značná závislost mezi jednotlivými vzorky signálu. Na druhé straně by bylo žádoucí, aby šum, který nyní budeme

považovat za stochastický proces, nevykazoval žádné možnosti predikce. Nejlépe, aby hustota pravděpodobnosti náhodného procesu $v^N = (v_0, v_1, ..., v_N)$, šumu měření, splňovala vlastnost

$$p(v^N) = p(v_0, v_1, ..., v_N) = p(v_0).p(v_1)...p(v_N) \tag{2.3.1}$$

Takovýto stochastický proces bývá označován jako bílý šum, neboli absolutně nezávislý proces a tedy též nepredikovatelný proces. Z hlediska diskuse v sekci 2.1 týkající se rozdílných vlastností signálu a šumu pak "ideální"situace pro úlohu estimace nastane v případě, kdy uvažujeme konstantní parametry a bílý šum, jelikož rozdíl mezi signálem a šumem je největší. Hlavní výsledky estimační teorie jsou založeny právě na takových předpokladech.

Co se týče popisu stavového šumu, zde je situace jednoduchá, a to z toho důvodu, že pokud chápeme stav systému v tradičním smyslu, pak, zhruba řečeno, stav x_k musí obsahovat veškerou informaci o minulosti do času t_k, která je potřebná pro určení jeho budoucího vývoje, a tudíž stavový šum w_k nesmí vykazovat žádnou závislost do minulosti. Tudíž požadujeme, aby opět hustota pravděpodobnosti náhodného procesu w_k splňovala vztah

$$p(w^k) = p(w_0, w_1, ..., w_k) = p(w_0).p(w_1)...p(w_k) \tag{2.3.2}$$

neboli, aby se jednalo o bílý šum.
Výše uvedená vlastnost stavu pak dále implikuje i nutnost vzájemné nezávislosti stavového šumu, šumu měření a počátečního stavu x_0, který se v této pravděpodobnostní interpretaci stává náhodnou veličinou.
Poznamenejme, že v tomto případě x_k pro $k = 0, 1, 2, ...$ reprezentuje markovský proces. Pokud by tyto předpoklady nebyly splněny je vhodnější považovat za stav dvojici (x_k, z_k), kde x_k je neznámá složka stavu a z_k je známá složka stavu. K tomuto problému se ještě podrobně vrátíme ve třetí kapitole. Alternativním pojetím pojmu stav se zabývá moderní teorie systémů, která je prezentována v [19].
Z předchozí části implicitně vyplývá, že teoreticky požadujeme znalost hustot pravděpodobnosti stavového a výstupního šumu. Z praktického hlediska je však znalost těchto hustot problematická. Nicméně na tuto skutečnost můžeme nahlížet následujícím způsobem: 1. Předpokládat, že $p(v_k)$ je gaussovské rozložení s jistou střední hodnotou a kovariancí s odůvodněním, že gaussovské rozložení má největší entropii (míra nepořádku je největší). 2. Neuvažovat žádné rozložení, ale předpokládat pouze znalost prvních dvou momentů. Odhad těchto momentů šumu měření lze provést zavedením známého, pouze pro tento odhad, signálu y_k.

Problém estimace parametrů.

Mějme vektor měření z_k popsaný rovnicí

$$z_k = h_k(\Theta) + v_k \quad k = 0, 1, 2, ..., N, \tag{2.3.3}$$

kde $E[v_k] = 0, E[v_k v_l^T] = R_k \delta_{kj}, \quad \delta_{kj} = 1, k = j; \quad \delta_{kj} = 0, k \neq j$
$h_k(\cdot)$ je známá vektorová funkce.

Cílem je určit odhad neznámých konstantních parametrů Θ.
Při řešení této úlohy se někdy předpokládá, na rozdíl od 1. dílu, že Θ je náhodný vektor se střední hodnotou $\hat{\Theta}'_0$ a kovariancí P'_0, kde Θ a v_k jsou nezávislé. K diskusi takového postupu se ještě vrátíme.

Nyní formulujeme problém estimace časově proměnných parametrů nazývaných stavové proměnné.

Problém estimace stavu

Mějme vektor měření z_k popsaný

$$z_k = h_k(x_k) + v_k \quad k = 0, 1, 2, ..., N \tag{2.3.4}$$

kde $E[v_k] = 0,\ E[v_k v_j^T] = R_k \delta_{kj}$
$h_k(\cdot)$ je známá vektorová funkce

a vektor stavu se vyvíjí podle rovnice

$$x_{k+1} = f_k(x_k) + w_k \quad k = 0, 1, 2, ..., N \tag{2.3.5}$$

kde $E[w_k] = 0, E[w_k w_j^T] = Q_k \delta_{kj}$
$f_k(\cdot)$ známá vektorová funkce

Cílem je odhadnout stav x_k.

Poznamenejme, že pokud stavový šum bude nulový, pak problém estimace stavu můžeme převést na problém estimace parametrů, jak již bylo naznačeno dříve.

V následující části se budeme věnovat bayesovskému přístupu k řešení estimačního problému.

2.4 Bayesovský přístup

Předchozí formulace úlohy estimace je základem pro různé přístupy k řešení problému estimace. V této práci budeme preferovat výhradně bayesovský přístup. K přiblížení tohoto přístupu použijeme konfrontaci s klasickým postupem k problému odhadu.

Uvažujme náhodný vektor $z = [z_1, z_2, ..., z_N]$ s hustotou pravděpodobnosti $p(z; \Theta)$, kde $\Theta = [\Theta_1, \Theta_2, ..., \Theta_l]^T$ je vektor parametrů. Při klasickém přístupu k problému odhadu Θ považujeme parametry za neznámé konstanty a k závěrům o Θ použijeme pouze z a tvar rozdělení z. Při bayesovském přístupu k závěrům o parametru Θ, tentokrát chápaného jako náhodná proměnná použijeme kromě toho ještě apriorní informaci o parametru Θ, kterou máme k dispozici nezávisle na realizaci z. Apriorní informace se vyjadřuje předpokladem, že Θ je náhodný vektor s jistým rozložením. Tato informace může mít objektivní i subjektivní charakter. K objasnění objektivní i subjektivní apriorní informace použijeme dva hypotetické příklady.

Příklad 2.4.1 Vyšetřením vestibulárního ústrojí se má rozhodnout, zda pacient trpí poruchou tohoto ústrojí. Z předchozích výzkumů vyplývá, že touto chorobou trpí 10% populace. To je možné pokládat za objektivní apriorní informaci.

Příklad 2.4.2 Biolog má odhadnout jistou konstantu Θ. Má určitou představu o možných hodnotách Θ, ale dává jim různou váhu. Považuje je za náhodné veličiny s určitou pravděpodobností. Avšak jiný biolog pro stejnou úlohu přiřadí podle své zkušenosti těmto hodnotám jiné

pravděpodobnosti. Jedná se tedy o subjektivní apriorní informaci.

Použitím klasického a bayesovského přístupu můžeme dostat, jak lze tušit z předchozího, velmi různé výsledky odhadu. Dále lze ukázat, že i když jsou oba přístupy odlišné v chápání neznámých parametrů, protože v klasickém postupu chápeme neznámý parametr jako konstantu, zatímco bayesovský přístup pracuje s náhodnými veličinami, můžeme tyto přístupy sblížit tím, že apriorní rozdělení parametru bude rovnoměrné na nekonečném intervalu nebo gaussovské s kovariancí jdoucí do nekonečna, tedy rozložení nepreferující žádné hodnoty parametrů. K tomuto problému se ještě vrátíme při diskusi odhadů ve smyslu maximální věrohodnosti a maximální aposteriorní pravděpodobnosti v následující sekci. Co se týče významu subjektivní a objektivní apriorní informace u bayesovského přístupu na aposteriorní odhad poznamenejme, že při opakovaném použití tohoto přístupu vliv této informace klesá.

Z předchozí diskuse vyplývá, že bayesovský přístup je vhodný nástroj k řešení estimačních úloh, protože umožňuje pracovat s apriorní informací a zároveň zahrnuje v případě potřeby i klasický postup.
Nyní můžeme přistoupit k formulaci obecného řešení problému estimace parametrů.
Nechť $z^k = [z_0, z_1, ..., z_k]$ obsahuje naměřené hodnoty v čase $t_0, t_1, ...t_k$. Předpokládejme, že Θ a v_k jsou spojité náhodné veličiny se známými hustotami pravděpodobnosti. Pak aposteriorní hustota pravděpodobnosti Θ pro dané z^k je podle Bayesova pravidla

$$p(\Theta \mid z^k) = \frac{p(z^k \mid \Theta).p(\Theta)}{p(z^k)} \tag{2.4.1}$$

kde $p(z^k) = \int p(z^k \mid \Theta).p(\Theta)d\Theta$.

Za předpokladu, že Θ a $v^k = (v_0, v_1, ..., v_k)$ jsou nezávislé, $p(z^k \mid \Theta)$ je známa z $p(v^k)$, protože vycházíme z platnosti modelu měření

$$z_k = h_k(\Theta) + v_k \quad k = 0, 1, 2, ... \tag{2.4.2}$$

a tedy

$$p(z^k \mid \Theta) = p_v(z^k - h(\Theta)) \tag{2.4.3}$$

kde $p_v()$ označuje hustotu pravděpodobnosti šumu měření. Později tento postup budeme často používat i v jiných situacích, ale hustoty již nebudeme značit speciálním symbolem.
Poznamenejme, že hustota $p(z^k)$ nezávisí na Θ a slouží jako normalizační konstanta. Nalezení $p(\Theta \mid z^k)$ umožní zjistit efekt měření na zvýšení informace o Θ a poskytne úplný pravděpodobnostní popis parametru Θ.
Po formulaci přístupu k řešení parametrického odhadu přejděme k řešení problému estimace stavu. Pro odhad stavu se parametry, nyní chápané jako stavové proměnné, mění v čase. Při použití bayesovských vztahů je proto potřebné zavést čas, ve kterém bude probíhat odhad v závislosti na měření. Místo pojmů odhad, estimace se v tomto případě často používá pojem filtrace, a to v širším slova smyslu, jakožto úlohy určení podmíněné hustoty pravděpodobnosti $p(x_k \mid z^l)$ nebo odhadu x_k na základě pozorování z^l a v užším smyslu, právě když $k = l$. Pokud $l > k$ jedná se o úlohu vyhlazování a pro $k > l$ je zaveden pojem predikce. V této práci se budeme věnovat převážně úlohám filtrace a predikce. Protože se stav v čase mění, je přirozené vyžadovat jeho odhad také v každém časovém okamžiku, což nás vede k formulaci rekurzivní filtrace následujícím způsobem:
Mějme odhad x_{k-1} založený na měření z_{k-1}, řekněme $\hat{x}_{k-1}$. Určeme odhad $\hat{x}_k$ z $\hat{x}_{k-1}$ a z^k. Tento

problém je obecně řešen užitím Bayesova pravidla. Získáme vztahy, které jsou podobné (2.4.1), ale jsou doplněny konvoluční rovnicí, která popisuje účinek změny stavových proměnných v čase. Abychom dostali rekurzivní řešení, je vhodné vyžadovat, aby $\{v_k\}$ byl bílý šum na rozdíl od (2.4.1), kde tato vlastnost vyžadována nebyla. Dále předpokládejme, že bílé šumy $\{v_k\}, \{w_k\}$ jsou nezávislé a rovněž jsou nezávislé na x_0. Pak aposteriorní hustota pravděpodobnosti nebo alternativně filtrační hustota pravděpodobnosti může být určena rekurzivně

$$p(x_k \mid z^k) = \frac{p(z_k \mid x_k).p(x_k \mid z^{k-1})}{p(z_k \mid z^{k-1})} \tag{2.4.4}$$

kde $p(z_k \mid z^{k-1}) = \int p(z_k \mid x_k)p(x_k \mid z^{k-1})dx_k$ je normalizační konstanta.

Hustota $p(x_k \mid z^k)$ je ve stejné formě jako v (2.4.1), kde $p(\Theta)$ je nahrazeno $p(x_k \mid z^{k-1})$. Hustota $p(z_k \mid x_k)$ je určena hustotou $p(v_k)$. Rovnice (2.3.5) nám poslouží k určení $p(x_k \mid z^{k-1})$, které je dáno

$$p(x_k \mid z^{k-1}) = \int p(x_k, x_{k-1} \mid z^{k-1})dx_{k-1} = \int p(x_k \mid x_{k-1}).p(x_{k-1} \mid z^{k-1})dx_{k-1} \tag{2.4.5}$$

přičemž $p(x_k \mid x_{k-1})$ je určeno $p(w_{k-1})$. Rovnice (2.4.5) vyžaduje konvoluci $p(x_k \mid x_{k-1})$ a $p(x_{k-1} \mid z^{k-1})$. Rekurze (2.4.4) a (2.4.5) může být odstartována aplikací Bayesova pravidla a apriorního popisu x_0, tedy

$$p(x_0 \mid z^0) = \frac{p(z_0 \mid x_0).p(x_0)}{p(z^0)} \tag{2.4.6}$$

Můžeme konstatovat, že t-variantní stav vyžaduje přídavnou strukturu s porovnáním s problémem odhadu parametrů. Zároveň je zřejmé, že estimace stavu daná rovnicemi (2.4.4)-(2.4.6) v sobě zahrnuje úlohu estimace parametrů jako speciální případ. Na závěr této kapitoly provedeme shrnutí formulace a řešení problému estimace stavu.

Obecné řešení problému estimace stavu: shrnutí
Nechť vektor stavu se vyvíjí podle následujícího vztahu

$$x_{k+1} = f_k(x_k) + w_k \quad k = 0, 1, 2, ... \tag{2.4.7}$$

kde x_k je nx dimenzionální vektor stavu systému v čase t_k,
w_k je nx dimenzionální stavový šum působící na systém v čase t, kde $t_k \leq t < t_{k+1}$,
$f_k(\cdot)$ je známá vektorová funkce příslušné dimenze.
Náhodný proces $\{w_k\}$ je bílý šum se známou hustotou pravděpodobnosti $p(w_k)$ a rovněž hustota pravděpodobnosti počátečního stavu $p(x_0)$ je známa.
Stav systému je sledován pomocí měřených hodnot z_k, které jsou ve známém vztahu k x_k, ale jsou kontaminovány šumem měření

$$z_k = h_k(x_k) + v_k \quad k = 0, 1, 2, ... \tag{2.4.8}$$

kde z_k je nz dimenzionální vektor známých měřených dat v čase t_k a v_k je nz dimenzionální vektor šumu měření ovlivňující data v čase t_k. Náhodný proces $\{v_k\}$ je bílý šum se známou hustotou pravděpodobnosti $p(v_k)$. Procesy $\{w_k\}, \{v_k\}$ a náhodná veličina x_0 jsou navzájem nezávislé.

Cílem je určení podmíněné hustoty pravděpodobnosti $p(x_k \mid z^l)$.

Protože budeme často pracovat s úlohou filtrace tedy pro $l = k$ a jednokrokové predikce $l = k - 1$, pak rekurzivní vztahy budou mít tvar:

filtrační hustota

$$p(x_k \mid z^k) = \frac{p(x_k \mid z^{k-1}).p(z_k \mid x_k)}{p(z_k \mid z^{k-1})} \tag{2.4.9}$$

prediktivní hustota

$$p(x_k \mid z^{k-1}) = \int_{-\infty}^{\infty} p(x_{k-1} \mid z^{k-1}) p(x_k \mid x_{k-1}) dx_{k-1} \tag{2.4.10}$$

kde

$$p(z_k \mid z^{k-1}) = \int_{-\infty}^{\infty} p(x_k \mid z^{k-1}) p(z_k \mid x_k) dx_k \tag{2.4.11}$$

$$p(x_0 \mid z^{-1}) \stackrel{\triangle}{=} p(x_0)$$

Vztahy (2.4.9), (2.4.10) jsou známé bayesovské rekurzivní vztahy.
Z teoretického hlediska jsou předchozí vztahy úplné řešení problému estimace stavu, protože poskytují úplný pravděpodobnostní popis náhodných stavových veličin ve formě hustot pravděpodobnosti. Avšak v mnoha případech bychom potřebovali spíše konkrétní číselně vyjádřené odhady a výsledek ve formě funkce, hustoty pravděpodobnosti, je nevhodný. Rovněž řešení funkcionálních bayesovských vztahů (2.4.9)-(2.4.10) je obecně analyticky neschůdné a právě situace, kdy analytická řešitelnost je zachována, budou obsahem následujících kapitol.

Poslední sekce této kapitoly bude věnována bodovým odhadům, tedy odhadům, kdy výsledkem estimace není funkce, ale vektor čísel.

2.5 Bodové odhady

Předpokládejme, že máme z^N měření proměnné z_k a hledáme bodový odhad spojité neznámé veličiny Θ. Jak již bylo řečeno v předchozích sekcích mohou být použity dva odlišné přístupy: klasický a bayesovský. Klasický přístup bude uvažovat Θ jako neznámou deterministickou konstantu, zatímco bayesovský přístup chápe Θ jako stochastickou proměnnou. Rozdíl mezi těmito přístupy je zřejmý, jestliže napíšeme hustotu pravděpodobnosti pro z_k

$$p(z_k; \Theta) \; a \; p(z_k \mid \Theta)$$

Tyto dvě funkce jsou ve skutečnosti nerozlišitelné, ačkoliv jejich interpretace je zcela odlišná. První funkce je parametrická hustota pro konstantní parametr Θ, zatímco druhá je podmíněná hustota pravděpodobnosti. V další části budeme používat zápis druhý i pro parametrickou hustotu.

V současné teorii bodových odhadů je navrženo velké množství kriterií, podle kterých je možné stanovit optimální odhad. Všimněme si pouze některých často používaných optimálních bodových odhadů, a to ve smyslu

- maximální věrohodnosti, $\hat{\Theta}^{ML}$

$$\hat{\Theta}^{ML} = arg\ max_{\Theta}\ \ p(z^N \mid \Theta) \tag{2.5.1}$$

- maximální aposteriorní pravděpodobnosti, $\hat{\Theta}^{MAP}$

$$\hat{\Theta}^{MAP} = arg\ max_{\Theta}\ \ p(\Theta \mid z^N) \tag{2.5.2}$$

- podmíněné střední hodnoty, $\hat{\Theta}^{E}$

$$\hat{\Theta}^{E} = \int_{-\infty}^{\infty} \Theta p(\Theta \mid z^k) d\Theta \tag{2.5.3}$$

- medián, $\hat{\Theta}^{ME}$

$$\int_{-\infty}^{\hat{\Theta}^{ME}} p(\Theta \mid z^k) d\Theta = \int_{\hat{\Theta}^{ME}}^{\infty} \ p(\Theta \mid z^k) d\Theta \tag{2.5.4}$$

Výběr optimálního estimátoru

$$\hat{\Theta}^{*} = t(z^N) \tag{2.5.5}$$

je tedy vlastně dán funkcí viz (2.5.1)-(2.5.4).
V úvodu sekce (2.4) jsme částečně zkoumali vztah mezi odhadem podle maximální věrohodnosti (klasický přístup k odhadu) a maximální aposteriorní pravděpodobnosti (bayesovský přístup). Jejich úzký vztah můžeme nyní snadno prokázat z Bayesova pravidla

$$p(A, B) = p(A \mid B).p(B) \tag{2.5.6}$$

Pro tento případ plyne

$$p(\Theta \mid z^N) = \frac{p(\Theta, z^N)}{p(z^N)} = \frac{p(\Theta)}{p(z^N)} p(z^N \mid \Theta) \tag{2.5.7}$$

Protože $p(z^N)$ může být chápáno jako konstanta při odhadu Θ, pak $p(\Theta \mid z^N)$ a $p(z^N \mid \Theta)$ se odlišují v tomto vztahu pouze apriorním rozdělením $p(\Theta)$. Jestliže apriorní rozdělení je neinformativní vzhledem k Θ to jest $p(\Theta) = C$ v případě, že Θ je diskrétní v úrovni - nabývá konečného počtu hodnot nebo má normální rozložení s kovariancí blížící se nekonečnu, pak

$$\hat{\Theta}^{MAP} = \hat{\Theta}^{ML}$$

a $\hat{\Theta}^{ML}$ je pouze „speciální“ případ odhadu MAP. Nicméně na rozdílnou interpretaci těchto přístupů nesmíme zapomenout.

Při generování bodového optimálního odhadu stavu bychom mohli (2.5.2)-(2.5.4) zapsat analogickým způsobem. Kritérium maximální věrohodnosti se v tomto případě nedá uplatnit, jelikož je v rozporu s předpokladem na charakter odhadované veličiny, stavu, který je chápán jako stochastický proces. Optimální bodový odhad stavu je pak ve smyslu

- a maximální aposteriorní pravděpodobnosti $\hat{x}_k^{MAP}$

$$\hat{x}_k^{MAP} = arg\ max_{x_k}\ p(x_k \mid z^k) \tag{2.5.8}$$

- podmíněné střední hodnoty $\hat{x}_k^E$

$$\hat{x}_k^E = \int_{-\infty}^{\infty} x_k p(x_k \mid z^k) dx_k$$

- medián $\hat{x}_k^{ME}$

$$\int_{-\infty}^{\hat{x}^{ME}} p(x_k \mid z^k) dx_k = \int_{\hat{x}_k^{ME}}^{\infty} p(x_k \mid z^k) dx_k$$

Tyto bodové odhady bychom mohli používat po příslušné úpravě i pro diskrétní náhodné veličiny. Pouze podmíněná střední hodnota použít nelze, protože odhad by nemusel patřit do množiny přípustných hodnot odhadované diskrétní (v úrovni) náhodné veličiny.

Bodové odhady pro gaussovskou hustotu pravděpodobnosti budou shodné, avšak pro negaussovské hustoty se mohou diametrálně lišit. Tento fakt si potvrdíme v následujícím příkladu.

Příklad 2.5.1 Vypočtěme bodové odhady $\hat{x}_k^{MAP}, \hat{x}_k^E, \hat{x}_k^{ME}$, jestliže filtrační hustota pravděpodobnosti $p(x_k \mid z^k)$ je dána tímto předpisem

$$\begin{array}{lll} p(x_k \mid z^k) & = 0,5 - \varepsilon & x_k \in < 0,1 > \\ & = 0,25 - \varepsilon & x_k \in < 1,3 > \\ & = 1 & x_k \in < 6, 6 + 3\varepsilon > \ pro\ \varepsilon \to 0 \end{array}$$

- Je zřejmé, že $\hat{x}_k^{MAP} \in < 6, 6 + 3\varepsilon >$
- Výpočet $\hat{x}^E$ bude složitější. Vyjdeme z definice, tudíž $\hat{x}_k^E = \int_0^{6+3\varepsilon} x_k p(x_k \mid z^k) dx_k = \int_0^1 (0,5 - \varepsilon) x_k dx_k + \int_1^3 (0,25 - \varepsilon) x_k dx_k + \int_6^{6+3\varepsilon} x_k dx_k = 1,25$
- Zbývá vypočítat medián. Musí platit, že $\int_{-\infty}^{\hat{x}_k^{ME}} p(x_k \mid z^k) dx_k = \int_{\hat{x}_k^{ME}}^{\infty} p(x_k \mid z^k) dx_k$ Odsud je zřejmé, že pro uvažovanou $p(x_k \mid z^k)$ dostaneme optimální bodový odhad ve smyslu medián v bodě $\hat{x}_k^{ME} = 1$.

Předchozí příklad ukazuje, že bodové odhady přinášejí výrazně jednodušší výsledek estimační úlohy v porovnání s hustotou pravděpodobnosti, ale v případě negaussovské hustoty pravděpodobnosti výrazně rozdílné výsledky a samozřejmě nepředstavují úplný pravděpodobnostní popis. Z tohoto důvodu a z faktu, že při znalosti hustoty pravděpodobnosti můžeme určit jakýkoliv bodový odhad, budeme v následujících kapitolách preferovat v převážné míře výpočet filtrační a prediktivní hustoty pravděpodobnosti.

Kapitola 3

Kalmanův filtr

Nový pohled na teorii filtrace reprezentované Wienerovou teorií [1] přinesl v roce 1960 Kalman [2]. Provedl rozbor metod řešení Wienerova problému a formuloval limitující faktory, které vážně ovlivňují praktickou použitelnost. Z těchto faktorů připomeňme alespoň omezení na stacionární procesy a značnou matematickou náročnost při odvození způsobující neprůhlednost z inženýrského pohledu. Kalmanův přístup k problému lineární filtrace tyto nevýhody eliminuje. K odvození filtru využívá ortogonální projekci a pracuje v časové oblasti se stavovými veličinami na rozdíl od Wienerovy teorie založené na spektrální faktorizaci a frekvenčním popisu.

Kalmanův filtr je chápán jako lineární rekurzivní algoritmus generující nestranný odhad ve smyslu podmínění střední hodnoty (minimální variance) neznámého stavu dynamického systému ze zašuměných dat získávaných v diskrétních časových okamžicích [2], [4], [9]-[13], [84]. Jeho uplatnění najdeme především v řídicích, komunikačních či monitorovacích systémech [5], [11], [20]. O rostoucí popularitě svědčí i pomalý průnik do ekonomických či ekologických disciplín [21], [22].

Cílem této kapitoly je především odvodit Kalmanův filtr využitím bayesovského přístupu. Vztahy získané během odvození a pohled do vnitřního mechanismu bayesovského přístupu budou nezbytné v dalších kapitolách práce. Součástí této kapitoly však bude i řada dalších úloh souvisejících s analýzou či využitím kalmanovské filtrace.

3.1 Lineární gaussovský systém

V této sekci se budeme zabývat lineárním systémem, speciálním případem (2.4.7), (2.4.8), který je definován následujícími vztahy

$$x_{k+1} = F_k x_k + w_k \qquad k = 0, 1, 2, ... \tag{3.1.1}$$

$$z_k = H_k x_k + v_k \qquad k = 0, 1, 2, ... \tag{3.1.2}$$

kde F_k je nx/nx dimenzionální matice
H_k je nz/nx dimenzionální matice

Rovnice (3.1.1) představuje stavovou rovnici popisující vývoj stavu x_k a (3.1.2) je rovnice měření definující vztah mezi výstupem systému $y_k = H_k x_k$, šumem v_k a měřením z_k. Za neměřitelný vstup systému budeme považovat šum w_k. Předpokládejme, že w_k a v_k pro $k = 0, 1, 2, ...$ jsou

procesy s těmito vlastnostmi

1. $\{v_k\}$ a $\{w_k\}$ jsou bílé šumy

2. $\{v_k\}$ a $\{w_k\}$ jsou gaussovské s nulovou střední hodnotou

3. $\{v_k\}$ a $\{w_k\}$ jsou nezávislé

kde symbol $\{v_k\}$ představuje náhodný proces v_k pro $k = 0, 1, 2....$

To znamená, že z první vlastnosti vyplývá kromě jiného

$$E[w_k w_l^T] = E[w_k]E[w_l^T]$$
$$E[v_k v_l^T] = E[v_k]E[v_l^T]$$

Z druhé vlastnosti dostaneme

$$E[w_k w_l^T] = 0 \qquad k \neq l$$
$$E[v_k v_l^T] = 0 \qquad k \neq l$$

a z druhé a třetí plyne

$$E[w_k v_l^T] = 0 \quad \forall k, l$$

Budeme dále předpokládat, že

$$E[v_k v_l^T] = R_k \delta_{k,l}$$

$$E[w_k w_l^T] = Q_k \delta_{k,l}$$

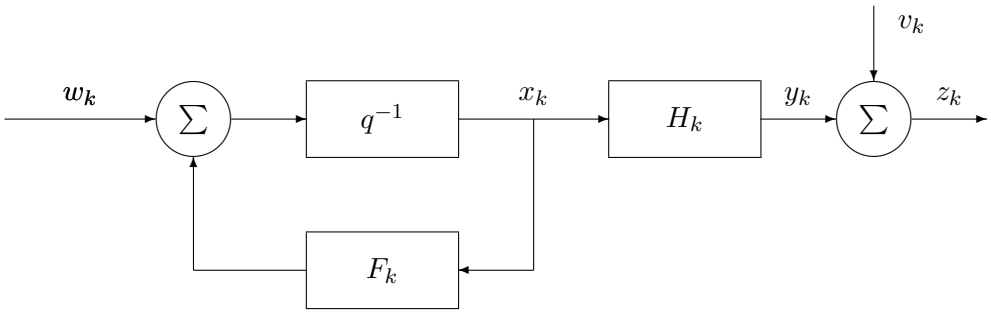
kde R_k a Q_k jsou symetrické nonnegativně definitní matice pro všechna k a počáteční stav x_0 je gaussovská náhodná veličina nezávislá na $\{v_k\}$ a $\{w_k\}$ se známou střední hodnotou $\bar{x}_0$ a známou kovariancí Σ_0

$$E[x_0] = \bar{x}_0 \qquad E[(x_0 - \bar{x}_0)(x_0 - \bar{x}_0)^T] = \Sigma_0$$

Signálový model uvažovaného systému je na obr. 3.1.1, kde q^{-1} je operátor zpětného posuvu definovaný $q^{-1}x(t) = x(t-1)$.

Připomeňme, že problém filtrace spočívá v generování odhadu stavu x_k na základě pozorování $z_0, z_1, ..., z_k$. Ještě před tím, než se tímto problémem budeme zabývat, si však všimneme samotného náhodného procesu $\{x_k\}$ a ukážeme, že se jedná o gaussmarkovský proces.
Je zřejmé, že x_k je náhodná veličina a z (3.1.1) plyne, že

$$x_k = \phi_{k,0} x_0 + \sum_{l=0}^{k-1} \phi_{k,l+1} w_l \tag{3.1.3}$$



Obrázek 3.1.1: Signálový model lineárního systému

kde

$$\phi_{k,l} = F_{k-1}F_{k-2}...F_l \qquad k > l \tag{3.1.4}$$

$$\phi_{k,k} = I \tag{3.1.5}$$

To znamená, že x_k je lineární kombinace náhodných vektorů $x_0, w_0, w_1, ..., w_{k-1}$ a jelikož mají normální rozdělení, pak i náhodná veličina x_k je popsána gaussovským rozdělením. Protože předchozí postup platí pro každé k, pak je zřejmé, že $\{x_k\}$ je gaussovský proces.
Nyní ukážeme, že $\{x_k\}$ je též markovský proces.
Pro markovský proces platí, že pro $k_1 < k_2 < ... < k_m < k$

$$p(x_k \mid x_{km}, ..., x_{k2}, x_{k1}) = p(x_k \mid x_{km})$$

Z (3.1.1) vyplývá, že

$$x_k = \phi_{k,km}x_{km} + \sum_{l=km}^{k-1} \phi_{k,l+1}w_l \tag{3.1.6}$$

Zřejmě w_l nejsou závislé na $x_{k1}, x_{k2}, ..., x_{km}$, a tudíž znalost $x_{k1}, x_{k2}, ..., x_{km}$ nepřináší novou informaci o w_l. To znamená, že $\{x_k\}$ je markovský proces. Tato vlastnost je důsledkem kauzality systému (3.1.1.) a skutečnosti, že $\{w_k\}$ je bílý šum.

Dále si všimneme náhodného procesu $\{z_k\}$. Zřejmě se jedná o gaussovský proces. Ale o markovský proces nejde, jelikož výstup systému y_k je v obecném případě korelovaný proces s hloubkou závislosti větší než jedna.

Jelikož $\{x_k\}$ i $\{z_k\}$ jsou gaussovské procesy, pro jejich popis vystačíme s prvními dvěma momenty. V následující části se budeme zabývat jejich výpočtem.
Z (3.1.3) je zřejmé, že

$$E[x_k] = \phi_{k,0}\bar{x}_0 \tag{3.1.7}$$

a z (3.1.2) vyplývá

$$E[z_k] = H_k E[x_k] \tag{3.1.8}$$

Před výpočtem kovarianční funkce stavu si nejdříve označme

$$\bar{x}_k = E[x_k] \tag{3.1.9}$$

$$\Sigma_{k,l} = E\{[x_k - \bar{x}_k][x_l - \bar{x}_l]^T\} \quad \text{pro } k > l \tag{3.1.10}$$

Výpočet je jednoduchý. Z (3.1.3) a (3.1.7) dostaneme

$\Sigma_{k,l} = E\{[\phi_{k,0}(x_0 - \bar{x}_0) + \sum_{i=0}^{k-1} \phi_{k,i+1} w_i][\phi_{l,0}(x_0 - \bar{x}_0) + \sum_{j=0}^{l-1} \phi_{l,j+1} w_j]^T\}$

Protože $x_0 - \bar{x}_0, w_0, ..., w_{k-1}$ jsou nezávislé, můžeme předchozí vztah zjednodušit na

$$\begin{aligned} \Sigma_{k,l} &= \phi_{k,0} E\{[x_0 - \bar{x}_0][x_0 - \bar{x}_0]^T\}\phi_{l,0}^T + \sum_{i=0}^{l-1} \phi_{k,i+1} Q_i \phi_{l,i+1}^T \\ &= \phi_{k,l}\{\phi_{l,0}\Sigma_0\phi_{l,0}^T + \sum_{i=0}^{l-1} \phi_{l,i+1} Q_i \phi_{l,i+1}^T\} \end{aligned} \tag{3.1.11}$$

a pro případ $k = l$

$$\Sigma_{k,k} = \phi_{k,0}\Sigma_0\phi_{k,0}^T + \sum_{i=0}^{k-1} \phi_{k,i+1} Q_i \phi_{k,i+1}^T \tag{3.1.12}$$

Z (3.1.1) však též vyplývá

$$\Sigma_{k+1,k+1} = F_k \Sigma_{k,k} F_k^T + Q_k \tag{3.1.13}$$

což umožňuje rekurzivní výpočet $\Sigma_{k,k}$. Využitím $\Sigma_{k,k}$ a vztahu (3.1.11) pak můžeme vyjádřit matice $\Sigma_{k,l}$ pro všechna k, l a $\Sigma_{0,0} = \Sigma_0$

$$\Sigma_{k,l} = \phi_{k,l}.\Sigma_{l,l} \quad k \geq l \tag{3.1.14}$$

a protože $\Sigma_{k,l} = \Sigma_{l,k}^T$, pak

$$\Sigma_{k,l} = \Sigma_{k,k}\phi_{l,k}^T \quad k \leq l \tag{3.1.15}$$

Rovnice (3.1.13)-(3.1.15) umožňují určit kovarianci stavu.

Přejděme k výpočtu kovarianční funkce měření. Označme $E[z_k] = \bar{z}_k$. Protože

$$z_k = H_k x_k + v_k$$

pak

$$\begin{aligned} E\{[z_k - \bar{z}_k][z_l - \bar{z}_l]^T\} &= E\{H_k[x_k - \bar{x}_k][x_l - \bar{x}_l]^T H^T\} + E\{H_k[x_k - \bar{x}_k]v_l^T\} \\ &+ E\{v_k[x_l - \bar{x}_l]^T H_k^T\} + E[v_k v_l^T] \end{aligned} \tag{3.1.16}$$

Protože podle předpokladů $\{v_k\}$ je nezávislé na x_0 a $\{w_k\}$, pak je také nezávislé na $\{x_k - \bar{x}_k\}$, a tudíž druhý a třetí člen na pravé straně předchozí rovnice jsou nulové. Kovariance po úpravě

bude určena následujícím vztahem

$$cov(z_k, z_l) = H_k \phi_{k,l} \Sigma_{l,l} H_l + R_k \delta_{k,l} \quad k \geq l \tag{3.1.17}$$

Tím jsme dokončili popis náhodných procesů $\{x_k\}$ a $\{z_k\}$. Bylo vidět, jak předpoklady na strukturu systému dané rovnicemi (3.1.1), (3.1.2), na počáteční stav a náhodné procesy $\{v_k\}, \{w_k\}$ ovlivňují výpočet charakteristik náhodných procesů $\{x_k\}, \{z_k\}$ a jak se při výpočtu projevují.
Poznamenejme, že někdy je možné se setkat s požadavkem, aby $E[w_k v_l^T] = S_k \delta_{k,l}$. Tento fakt ovlivní $cov(z_k, z_l)$ a tedy (3.1.16), (3.1.17) a rovněž způsobí, že $p(x_{k+1} \mid x_k, z_k) \neq p(x_{k+1} \mid x_k)$, jak plyne ze vztahů (3.1.1), (3.1.2), jelikož $E[w_k v_k^T] = S_k$. To je důvod, proč bychom v tomto případě měli pod pojmem stav chápat dvojici $[x_k, z_k]$, přičemž x_k chápat jako neznámou složku stavu a z_k jako známou složku stavu. Nicméně předpoklad, že $S_k = 0$ může být bez ztráty obecnosti zaveden, jelikož můžeme provést následující transformaci v případě, že $S_k \neq 0$.
Nechť

$$x_{k+1} = F_k x_k + w_k \tag{3.1.18}$$

$$z_k = H_k x_k + v_k \tag{3.1.19}$$

a všechny specifikace veličin v těchto vztazích jsou shodné s předchozím, pouze

$$cov(w_k, v_k^T) = S_k \tag{3.1.20}$$

Jestliže R_k^{-1} existuje, můžeme (3.1.18) upravit na

$$\begin{aligned} x_{k+1} &= F_k x_k + w_k - S_k R_k^{-1}[z_k - z_k] = F_k x_k + w_k - S_k R_k^{-1}[H_k x_k + v_k - z_k] \\ &= (F_k - S_k R_k^{-1} H_k) x_k + S_k R_k^{-1} z_k + w_k - S_k R_k^{-1} v_k \end{aligned}$$

Označme

$$\bar{F}_k = F_k - S_k R_k^{-1} H_k$$
$$\bar{w}_k = w_k - S_k R_k^{-1} v_k$$

pak

$$x_{k+1} = \bar{F}_k x_k + S_k R_k^{-1} z_k + \bar{w}_k \tag{3.1.21}$$

Nyní vypočtěme $cov(\bar{w}_k, \bar{w}_k), cov(\bar{w}_k, v_k^T)$ a $cov(v_k, \bar{w}_k^T)$

$cov(\bar{w}_k, \bar{w}_k) = E\{[w_k - S_k R_k^{-1} v_k][(w_k - S_k R_k^{-1} v_k)^T] = Q_k - S_k R_k^{-1} S_k^T - S_k R_k^{-1} S_k^T + S_k R_k^{-1} R_k R_k^{-1} S_k^T = Q_k - S_k R_k^{-1} S_k^T$

$cov(\bar{w}_k, v_k) = E\{[w_k - S_k R_k^{-1} v_k] v_k^T\} = S_k - S_k R_k^{-1} R_k = 0$

$cov(v_k, \bar{w}_k^T) = E[v_k (w_k - S_k R_k^{-1} v_k)^T] = S_k^T - R_k R_k^{-1} S_k^T = 0$

To znamená, že "nový"stavový šum $\bar{w}_k$ není závislý na v_k a systém (3.1.21), (3.1.19) je obsahově shodný s (3.1.1) a (3.1.2). Odlišuje se pouze ve členu $S_k R_k^{-1} z_k$ v(3.1.21), ale to je známá veličina

v čase k.

V této sekci jsme ukázali, že vlastnosti náhodných procesů $\{x_k\}$ a $\{z_k\}$ jsme schopni popsat pomocí střední hodnoty a kovariančních matice pro všechny časové okamžiky. V následující kapitole se budeme zabývat návrhem filtru poskytující odhad stavu na základě dostupných měření a modelu systému.

3.2 Bayesovský přístup k syntéze filtru

V této sekci se soustředíme na syntézu filtru generující odhad stavu lineárního gaussovského systému (3.1.1), (3.1.2) v bayesovském rámci. Představíme dvě techniky návrhu filtru. Zatímco první technika důsledně vychází z bayesovských vztahů (2.4.9), (2.4.10), druhá staví na základních větách pravděpodobnosti a statistiky. Každý z přístupů nabídne jiný pohled na podstatu Kalmanova filtru a umožní, v následujících kapitolách, odlišné přístupy k rozšíření aplikovatelnosti myšlenky Kalmanova filtru pro nelineární systémy.

3.2.1 Přímý přístup k syntéze filtru

V této sekci se soustředíme na syntézu filtru důsledně vycházející z bayesovských vztahů (2.4.9), (2.4.10) a generující odhad stavu lineárního gaussovského systému ze sekce 3.1.
Nejdříve budeme počítat filtrační hustotu pravděpodobnosti. K tomu potřebujeme znát podle (2.4.9) podmíněnou prediktivní hustotu pravděpodobnosti. Začneme pro $k = 0$. Pak prediktivní hustota $p(x_0 \mid z^{-1})$ je dána známým apriorním rozložením $p(x_0) \stackrel{\triangle}{=} p(x_0 \mid z^{-1})$, tj.

$$p(x_0 \mid z^{-1}) = p(x_0) = N(x_0 : \hat{x}'_0, P'_0) \tag{3.2.1}$$

kde $\hat{x}'_0 = E[x_0 \mid z^{-1}]$ je apriorní střední hodnota a $P'_0 = cov[x_0 \mid z^{-1}]$ je apriorní kovarianční matice chyby odhadu stavu. Značení $N(x_0 : \hat{x}'_0, P'_0)$ pak označuje, že náhodná veličina x_0 má normální rozložení se střední hodnotou $\hat{x}'_0$ a kovariancí P'_0.

Hustotu $p(z_k \mid x_k)$ snadno získáme z (3.1.2) a základních vět teorie pravděpodobnosti ve tvaru

$$p(z_0 \mid x_0) = N(z_0 : H_0 x_0, R_0) \tag{3.2.2}$$

Bayesův vztah obsahuje ještě tzv. normalizační konstantu, hustotu pravděpodobnosti $p(z_0 \mid z^{-1})$. Získáme ji opět snadno z (3.2.1) a aplikací základů teorie pravděpodobnosti

$$p(z_0 \mid z^{-1}) = N(z_0 : H_0 \hat{x}'_0, H_0 P'_0 H_0^T + R_0) \tag{3.2.3}$$

protože

$$\begin{aligned} E[z_0 \mid z^{-1}] = E[H_0 x_0 + v_0 \mid z^{-1}] &= H_0 \hat{x}'_0 \\ E[(z_0 - \hat{z}_0)(z_0 - \hat{z}_0)^T \mid z^{-1}] &= E[(H_0 x_0 + v_0 - H_0 \hat{x}'_0)(H_0 x_0 + v_0 - H_0 \hat{x}'_0)^T \mid z^{-1}] \\ &= H_0 P'_0 H_0^T + R_0. \end{aligned}$$

Nyní můžeme dosadit do (2.4.9)

$$
\begin{aligned}
p(x_0 \mid z^0) &= \frac{p(x_0 \mid z^{-1})p(z_0 \mid x_0)}{p(z_0 \mid z^{-1})} = \frac{N(z_0 : H_0x_0, R_0)N(x_0 : \hat{x}'_0, P'_0)}{N(z_0 : H_0\hat{x}'_0, H_0P'_0H_0^T + R_0)} \\
&= \frac{1}{(2\pi)^{nz/2}(detR_0)^{1/2}} e^{-\frac{1}{2}[z_0 - H_0x_0]^T R_0^{-1}[z_0 - H_0x_0]} \\
&\times \frac{1}{(2\pi)^{nx/2}(detP'_0)^{1/2}} e^{-\frac{1}{2}[x_0 - \hat{x}'_0]^T P_0'^{-1}[x_0 - \hat{x}'_0]} \\
&\times \frac{(2\pi)^{nz/2}[det(H_0P'_0H_0^T + R_0)]^{1/2}}{e^{-\frac{1}{2}[z_0 - H\hat{x}'_0]^T[H_0P'_0H_0^T + R_0]^{-1}[z_0 - H\hat{x}'_0]}}
\end{aligned} \tag{3.2.4}
$$

Nejdříve si všimneme konstanty před exponenciální částí, řekněme *konst*, která se rovná

$$
konst = \frac{1}{(2\pi)^{nx/2}} . \frac{[det(H_0P'_0H_0^T + R_0)]^{1/2}}{(detR_0)^{1/2} . (detP'_0)^{1/2}} \tag{3.2.5}
$$

Nyní přejděme k exponenciální části (3.2.4). Pro jednodušší zápis nebudeme používat časové indexy. Exponent pak je

$$
\begin{aligned}
- \quad & \frac{1}{2}[z - Hx]^T R^{-1}[z - Hx] - \frac{1}{2}[x - \hat{x}']^T P'^{-1}[x - \hat{x}'] \\
+ \quad & \frac{1}{2}[z - H\hat{x}']^T[HP'H^T + R]^{-1}[z - H\hat{x}'] = -\frac{1}{2}x^T[H^TR^{-1}H + P'^{-1}]x \\
+ \quad & \frac{1}{2}x^T[H^TR^{-1}z + P'^{-1}\hat{x}'] + \frac{1}{2}[z^TR^{-1}H + \hat{x}'^TP'^{-1}]x - \frac{1}{2}z^TR^{-1}z \\
+ \quad & \frac{1}{2}[z - H\hat{x}']^T[HP'H^T + R]^{-1}[z - H\hat{x}'] - \frac{1}{2}\hat{x}'^TP'^{-1}\hat{x}'
\end{aligned} \tag{3.2.6}
$$

Ukážeme, že pravou stranu (3.2.6) lze upravit na čtverec

$$
\begin{aligned}
& & - \quad & \frac{1}{2}(x - \hat{x})^T P^{-1}(x - \hat{x}) = -\frac{1}{2}x^TP^{-1}x \\
+\frac{1}{2}\hat{x}^TP^{-1}x & & + \quad & \frac{1}{2}x^TP^{-1}\hat{x} - \frac{1}{2}\hat{x}^TP^{-1}\hat{x}
\end{aligned} \tag{3.2.7}
$$

Protože úprava musí být platná pro všechna x, pak

$$
P^{-1} = H^TR^{-1}H + P'^{-1} \tag{3.2.8}
$$

a zároveň

$$
P^{-1}\hat{x} = H^TR^{-1}z + P'^{-1}\hat{x}' \tag{3.2.9}
$$

Úpravou (3.2.9) dostaneme

$$
\hat{x} = PH^TR^{-1}z + PP'^{-1}\hat{x}' \tag{3.2.10}
$$

Dosazením z (3.2.8) do (3. 2.10) získáme

$$
\hat{x} = PH^TR^{-1}z + P(P^{-1} - H^TR^{-1}H)\hat{x}' = \hat{x}' + PH^TR^{-1}(z - H\hat{x}') \tag{3.2.11}
$$

Aplikací maticového inverzního lemma

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(DA^{-1}B + C^{-1})^{-1}DA^{-1} \tag{3.2.12}$$

na (3.2.8) dostaneme

$$P = P' - P'H^T[HP'H^T + R]^{-1}HP' \tag{3.2.13}$$

To znamená, že PH^TR^{-1} z (3.2.11) se rovná

$$\begin{aligned} PH^TR^{-1} &= P'H^TR^{-1} - P'H^T[HP'H^T + R]^{-1}HP'H^TR^{-1} \\ &= P'H^T[I - (HP'H^T + R)^{-1}HP'H^T]R^{-1} \\ &= P'H^T[(HP'H^T + R)^{-1}(HP'H^T + R) \\ &- (HP'H^T + R)^{-1}HP'H^T]R^{-1} \\ &= P'H^T[(HP'H^T + R)^{-1}(HP'H^T + R - HP'H^T)]R^{-1} \\ &= P'H^T[HP'H^T + R]^{-1} \end{aligned} \tag{3.2.14}$$

Nyní je zřejmé, že (3.2.11) můžeme zapsat využitím (3.2.14) takto

$$\hat{x} = \hat{x}' + P'H^T[HP'H^T + R]^{-1}(z - H\hat{x}') \tag{3.2.15}$$

Zbývá prověřit, že skutečně poslední tři členy na pravé straně (3.2.6) se rovnají poslednímu členu na pravé straně (3.2.7). Tedy

$$\begin{aligned} -\frac{1}{2}z^TR^{-1}z &+ \frac{1}{2}[z - H\hat{x}']^T[HP'H^T + R]^{-1}[z - H\hat{x}'] \\ &- \frac{1}{2}\hat{x}'^TP'^{-1}\hat{x}' = -\frac{1}{2}\hat{x}^TP^{-1}\hat{x} \end{aligned} \tag{3.2.16}$$

Dosaďme do pravé strany (3.2.16) z (3.2.15) a (3.2.8). Pak

$$\begin{aligned} \hat{x}^TP^{-1}\hat{x} &= [\hat{x}' + P'H^T(HP'H^T + R)^{-1}(z - H\hat{x}')]^T[H^TR^{-1}H + P'^{-1}] \\ &. \quad [\hat{x}' + P'H^T(HP'H^T + R)^{-1}(z - H\hat{x}')] \end{aligned} \tag{3.2.17}$$

Označme

$$\begin{aligned} K &= P'H^T(HP'H^T + R)^{-1} \\ J &= H^TR^{-1}H + P'^{-1} \end{aligned} \tag{3.2.18}$$

a dosaďme do (3.2.17)

$$-\frac{1}{2}\hat{x}^TP^{-1}\hat{x} = -\frac{1}{2}[Kz + (I - KH)\hat{x}']^TJ[Kz + (I - KH)\hat{x}'] \tag{3.2.19}$$

Upravme levou stranu (3.2.16)

$$\begin{aligned} &- z^TR^{-1}z + [z - H\hat{x}']^T[HP'H^T + R]^{-1}[z - H\hat{x}'] - \hat{x}'P'^{-1}\hat{x}' = \\ &- z^T[R^{-1} - (HP'H^T + R)^{-1}]z - \hat{x}'H^T(HP'H^T + R)^{-1}z \\ &- z^T(HP'H^T + R)^{-1}H\hat{x} + \hat{x}'[H^T(HP'H^T + R)^{-1}H \\ &- P'^{-1}]\hat{x}' \end{aligned} \tag{3.2.20}$$

Nyní odečteme kvadratické členy u z na pravých stranách rovnic (3.2.19) a (3.2.20)

$$\begin{aligned}
K^T J K &- [R^{-1} - (HP'H^T + R)^{-1}] \\
&= (HP'H^T + R)^{-1} HP'(H^T R^{-1} H + P'^{-1}) P'H^T (HPH^T + R^{-1}) \\
&- R^{-1} + (HP'H^T + R)^{-1} \\
&= (HP'H^T + R)^{-1}[HP'(H^T R^{-1} H + P'^{-1}) P'H^T (HP'H^T + R)^{-1} - (HP'H^T + R)^{-1} + I] \\
&= (HP'H^T + R)^{-1}[HP'H^T R^{-1} HP'H^T + HP'P'^{-1} P'H^T \\
&- (HP'H^T + R) R^{-1} (HP'H^T + R) + (HP'H^T + R)](HP'H^T + R)^{-1} \\
&= (HP'H^T + R)^{-1}[HP'H^T R^{-1} HP'H^T + HP'H^T \\
&+ HP'H^T + R - HP'H^T R^{-1} HP'H^T \\
&- HP'H^T - HP'H^T - R][HP'H^T + R]^{-1} = 0
\end{aligned}$$

Smíšené členy se musí také rovnat

$$\begin{aligned}
K^T J(I - KH) &- (HP'H^T + R)^{-1} H = (HP'H^T + R)^{-1} HP'(H^T R^{-1} H \\
&+ P'^{-1})[I - P'H^T (HP'H^T + R)^{-1} H] - (HP'H^T + R)^{-1} H \\
&= (HP'H^T + R)^{-1}(HP'H^T R^{-1} H + H)[I - P'H^T (HP'H^T + R)^{-1} H] \\
&- (HP'H^T + R)^{-1} H = (HP'H^T + R)^{-1}[HP'H^T R^{-1} H + H \\
&- HP'H^T R^{-1} HP'H^T (HP'H^T + R)^{-1} H \\
&- HP'H^T (HP'H^T + R)^{-1} H] - (HP'H^T + R)^{-1} H \\
&= (HP'H^T + R)^{-1}\{[HP'H^T R^{-1} H + H \\
&- HP'H^T R^{-1} HP'H^T (HP'H^T + R)^{-1} H - HP'H^T (HP'H^T \\
&+ R)^{-1} H] - H\} = (HP'H^T + R)^{-1} HP'H^T [R^{-1} \\
&- R^{-1} HP'H^T (HP'H^T + R)^{-1} - (HP'H^T + R)^{-1}] H \\
&= (HP'H^T + R)^{-1}[R^{-1} - R^{-1}(HP'H^T + R - R)(HP'H^T + R)^{-1} \\
&- (HP'H^T + R)^{-1}] H = (HP'H^T + R)^{-1} HP'H^T [R^{-1} \\
&- R^{-1} + (HP'H^T + R)^{-1} - (HP'H^T + R)^{-1}] H = 0
\end{aligned}$$

Konečně, kvadratické členy u $\hat{x}'$ po odečtení musí být též rovny nule.

$$
\begin{aligned}
(I &- H^T K^T)J(I-KH) + H^T[HP'H^T+R]^{-1}H - P'^{-1} = J - H^T K^T J \\
&- JKH + H^T K^T KH + H^T[HP'H^T+R]^{-1}H - P'^{-1} = H^T R^{-1} H \\
&+ P'^{-1} - H^T(HP'H^T+R)^{-1}HP'(H^T R^{-1}H + P'^{-1}) \\
&- (H^T R^{-1}H + P'^{-1})P'H^T(HP'H^T+R)^{-1}H \\
&+ H^T(HP'H^T+R)^{-1}HP'(H^T R^{-1}H + P'^{-1})P'H^T(HP'H^T \\
&+ R)^{-1}H + H^T[HP'H^T+R]^{-1}H - P'^{-1} = H^T R^{-1} H \\
&- H^T(HP'H^T+R)^{-1}HP'H^T R^{-1}H - H^T(HP'H^T+R)^{-1}H \\
&- H^T R^{-1}HP'H^T(HP'H^T+R)^{-1}H - H^T(HP'H^T+R)^{-1}H \\
&+ H^T(HP'H^T+R)^{-1}HP'H^T R^{-1}HP'H^T(HP'H^T+R)^{-1}H \\
&+ H^T(HP'H^T+R)^{-1}HP'H^T(HP'H^T+R)^{-1}H \\
&- H^T[HP'H^T+R]^{-1}H = H^T[HP'H^T+R]^{-1}[(HP'H^T+R)R^{-1}(HP'H^T+R) \\
&- HP'H^T R^{-1}(HP'H^T+R) - (HP'H^T+R) - (HP'H^T \\
&+ R)R^{-1}HP'H^T - (HP'H^T+R) + HP'H^T R^{-1}HP'H^T + HP'H^T \\
&+ (HP'H^T+R)][HP'H^T+R]^{-1}H = H^T[HP'H^T \\
&+ R]^{-1}[(HP'H^T+R)R^{-1}(HP'H^T+R) \\
&- HP'H^T R^{-1}(HP'H^T+R) - (HP'H^T+R) \\
&- (HP'H^T+R)R^{-1}HP'H^T + HP'H^T R^{-1}HP'H^T \\
&+ HP'H^T][HP'H^T+R]^{-1}H = H^T[HP'H^T+R]^{-1}[(HP'H^T \\
&+ R)R^{-1}HP'H^T + HP'H^T + R - HP'H^T R^{-1}HP'H^T - HP'H^T - (HP'H^T+R) \\
&- (HP'H^T+R)R^{-1}HP'H^T + HP'H^T R^{-1}HP'H^T + HP'H^T][HP'H^T+R]^{-1}H^T = 0
\end{aligned}
$$

Výraz (3.2.16) je splněn.

Nyní se musíme vrátit k (3.2.1) a prověřit, že platí

$$\frac{1}{(2\pi)^{nx/2}(detP)^{1/2}} = \frac{1}{(2\pi)^{nx/2}} \frac{[det(HP'H^T+R)]^{1/2}}{(detR)^{1/2}(detP')^{1/2}}$$

nebo-li

$$(detP)^{1/2} = \frac{(detR)^{1/2}(detP')^{1/2}}{[det(HP'H^T+R)]^{1/2}} \tag{3.2.21}$$

Z(3.2.13) víme, že

$$
\begin{aligned}
(detP) &= \{det[P' - P'H^T(R+HP'H^T)^{-1}HP']\} \\
&= (detP')\{det[I - H^T(R+HP'H^T)^{-1}HP']\}
\end{aligned} \tag{3.2.22}
$$

Pro další úpravu použijeme vztah

$$det(I_n + AB) = det(I_p + BA) \tag{3.2.23}$$

kde A je matice dimenze n/p a B matice dimenze p/n.
Položíme-li $HP' = B$ $(dim\ nz/nx)$ a $A = -H^T(R+HP'H^T)^{-1}$ $(dim\ nx/nz)$ můžeme upravit

(3.2.22) na

$$\begin{aligned} detP &= (detP')\{det[I_{nz} - HP'H^T(R + HP'H^T)^{-1}]\} \\ &= (detP')\{det[(R + HP'H^T)(R + HP'H^T)^{-1} \\ &- HP'H^T(R + HP'H^T)^{-1}]\} \\ &= (detP')\{det[R(R + HP'H^T)^{-1}]\} \\ &= \frac{detP'detR}{det(R + HP'H^T)} \end{aligned} \tag{3.2.24}$$

Dosazením (3.2.24) do (3.2.21) je zřejmé, že rovnost platí.

Po těchto výpočtech je jasné, že filtrační hustota pravděpodobnosti je normální rozložení se střední hodnotou danou rovnicí (3.2.15), tedy

$$\hat{x}_0 = \hat{x}'_0 + P'_0 H_0^T [H_0 P'_0 H_0^T + R_0]^{-1} [z_0 - H_0 \hat{x}'_0] \tag{3.2.25}$$

a kovariancí danou rovnicí (3.2.13), tedy

$$P_0 = P'_0 - P'_0 H_0^T [H_0 P'_0 H_0^T + R_0]^{-1} H_0 P'_0 \tag{3.2.26}$$

Filtrační hustota pravděpodobnosti je pak

$$p(x_0 \mid z_0) = N(x_0 : \hat{x}_0, P_0) \tag{3.2.27}$$

V následující části této sekce se budeme věnovat odvození prediktivní hustoty pravděpodobnosti $p(x_1 \mid z^0)$ tedy pro $k = 1$.Vyjdeme ze vztahu (2.4.10). K výpočtu potřebujeme znát filtrační hustotu pravděpodobnosti $p(x_0 \mid z^0)$ a přechodovou hustotu pravděpodobnosti $p(x_1 \mid x_0)$. Filtrační hustotu známe z předchozího odvození, vztah (3.2.27) a přechodovou hustotu snadno vypočteme ze znalosti (3.1.1) a základních vět teorie pravděpodobnosti. Takže přechodová hustota je stanovena následujícím vztahem

$$p(x_1 \mid x_0) = N(x_1 : F_0 x_0, Q_0) \tag{3.2.28}$$

kde

$$\begin{aligned} E[x_1 \mid x_0] &= E[F_0 x_0 + w_0 \mid x_0] = F_0 x_0 \\ E[(x_1 - E[x_1 \mid x_0])(x_1 - E[x_1 \mid x_0])^T \mid x_0] &= E[(F_0 x_0 - F_0 x_0 + w_0)(F_0 x_0 - F_0 x_0 + w_0)^T \mid x_0] \\ &= E[w_0 w_0^T \mid x_0] = Q_0 \end{aligned}$$

Nyní dosadíme do (2.4.10) z (3.2.27) a (3.2.28).

$$
\begin{aligned}
p(x_1 \mid z^0) &= \int N(x_0 : \hat{x}_0, P_0) N(x_1 : F_0 x_0, Q_0) dx_0 \\
&= \int \frac{1}{(2\pi)^{nx/2}.(detP_0)^{1/2}} e^{-\frac{1}{2}[x_0-\hat{x}_0]^T P_0^{-1}[x_0-\hat{x}_0]} \\
&\times \quad \frac{1}{(2\pi)^{nx/2}(detQ_0)^{1/2}} e^{-\frac{1}{2}[x_1-F_0x_0]^T Q_0^{-1}[x_1-F_0x_0]} dx_0 \\
&= \quad \frac{1}{(2\pi)^n x (detP_0)^{1/2}(detQ_0)^{1/2}} \\
&\times \quad \int e^{-\frac{1}{2}\{[x_0-\hat{x}_0]^T P_0^{-1}[x_0-\hat{x}_0]+[x_1-F_0x_0]^T Q_0^{-1}[x_1-F_0x_0]\}} dx_0 \qquad (3.2.29)
\end{aligned}
$$

Exponent v (3.2.29) můžeme upravit

$$
\begin{aligned}
- \quad & \frac{1}{2}\{[x_0-\hat{x}_0]^T P_0^{-1}[x_0-\hat{x}_0] + [x_1-F_0x_0]^T Q_0^{-1}[x_1-F_0x_0]\} = \\
- \quad & \frac{1}{2} x_0^T [P_0^{-1} + F_0^T Q_0^{-1} F_0] x_0 - \frac{1}{2} x_0^T [-P_0^{-1}\hat{x}_0 - F_0^T Q_0^{-1} x_1] \\
- \quad & \frac{1}{2}[-\hat{x}_0^T P_0^{-1} - x_1^T Q_0^{-1} F_0] x_0 - \frac{1}{2}[\hat{x}_0^T P_0^{-1}\hat{x}_0 + x_1^T Q_0^{-1} x_1] = \\
- \quad & \frac{1}{2}\{x_0^T A^{-1} x_0 - x_0^T b - b^T x_0 + c\} \qquad (3.2.30)
\end{aligned}
$$

kde

$$A^{-1} = P_0^{-1} + F_0^T Q_0^{-1} F_0$$

$$b = P_0^{-1}\hat{x}_0 + F_0^T Q_0^{-1} x_1$$

$$c = \hat{x}_0^T P_0^{-1}\hat{x}_0 + x_1^T Q_0^{-1} x_1$$

Využitím (3.2.30) můžeme integrál v (3.2.29) zapsat a upravit následujícím způsobem

$$
\begin{aligned}
&\int e^{-\frac{1}{2}(x_0^T A^{-1} x_0 - x_0^T b - b^T x_0 + c)} dx_0 = \\
&= \int \frac{(2\pi)^{nx/2}(detA)^{1/2}}{(2\pi)^{nx/2}(detA)^{1/2}} e^{-\frac{1}{2}(x_0-Ab)^T A^{-1}(x_0-Ab)} . e^{\frac{1}{2}(b^T Ab - c)} dx_0 \\
&= (2\pi)^{nx/2}(detA)^{1/2} e^{\frac{1}{2}(b^T Ab-c)} \int \frac{1}{(2\pi)^{nx/2}(detA)^{1/2}} e^{-\frac{1}{2}(x_0-Ab)^T A^{-1}(x_0-Ab)} dx_0 \\
&= (2\pi)^{nx/2}(detA)^{1/2} e^{\frac{1}{2}(b^T Ab-c)} \qquad (3.2.31)
\end{aligned}
$$

Využili jsme skutečnosti, že integrand je gaussovské rozložení a integrál z hustoty je roven jedné. Vztah (3.2.31) můžeme dosadit do (3.2.29) a dostaneme

$$p(x_1 \mid z^0) = \frac{(2\pi)^{nx/2}(detA)^{1/2}}{(2\pi)^{nx}(detP_0)^{1/2}(detQ_0)^{1/2}} e^{\frac{1}{2}(b^T Ab-c)} \qquad (3.2.32)$$

Nyní ukážeme, že exponent v (3.2.32) lze upravit na čtverec vzhledem k x_1. Dosazením z (3.2.31) dostaneme

$$
\begin{aligned}
b^T Ab - c &= (P_0^{-1}\hat{x}_0 + F_0^T Q_0^{-1} x_1)^T (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} (P_0^{-1}\hat{x}_0 \\
&+ F_0^T Q_0^{-1} x_1) - \hat{x}_0^T P_0^{-1}\hat{x}_0 - x_1^T Q_0^{-1} x_1 \\
&= x_1^T [Q_0^{-1} F_0 (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} F_0^T Q_0^{-1} \\
&- Q_0^{-1}] x_1 + x_1^T [Q_0^{-1} F_0 (P_0^{-1} + F_O^T Q_0^{-1} F_0)^{-1} P_0^{-1} \hat{x}_0] \\
&+ [\hat{x}_0^T P_0^{-1} (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} F_0^T Q_0^{-1}] x_1 \\
&+ \hat{x}_0^T [P_0^{-1} (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} P_0^{-1} - P_0^{-1}] \hat{x}_0
\end{aligned} \tag{3.2.33}
$$

Pro další úpravy použijeme maticové inverzní lemma a dokážeme, že platí

$$
b^T Ab - c = -(x_1 - \hat{x}_1')^T P_1'^{-1} (x_1 - \hat{x}_1') \tag{3.2.34}
$$

Srovnáním (3.2.34) a (3.2.33) musí platit

$$
-P_1'^{-1} = Q_0^{-1} F_0 (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} F_0^T Q_0^{-1} - Q_0^{-1} = -(Q_0 + F_0 P_0 F_0^T)^{-1} \tag{3.2.35}
$$

$$
\begin{aligned}
-P_1'^{-1}\hat{x}_1' &= -Q_0^{-1} F_0 (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} P_0^{-1} \hat{x}_0 \\
&= -Q_0^{-1} F_0 [P_0 - P_0 F_0^T (F_0 P_0 F_0^T + Q_0)^{-1} F_0 P_0] P_0^{-1} \hat{x}_0 \\
&= -Q_0^{-1} [F_0 - F_0 P_0 F_0^T (F_0 P_0 F_0^T + Q_0)^{-1} F_0] \hat{x}_0 \\
&= -Q_0^{-1} [I - F_0 P_0 F_0^T (F_0 P_0 F_0^T + Q_0)^{-1}] F_0 \hat{x}_0 \\
&= -Q_0^{-1} [(F_0 P_0 F_0^T + Q_0)(F_0 P_0 F_0^T + Q_0)^{-1} \\
&- F_0 P_0 F_0^T (F_0 P_0 F_0^T + Q_0)^{-1}] F_0 \hat{x}_0 \\
&= -Q_0^{-1} [Q_0 (F_0 P_0 F_0^T + Q_0)^{-1}] F_0 \hat{x}_0 \\
&= -(F_0 P_0 F_0^T + Q_0)^{-1} F_0 \hat{x}_0 \\
&= -P_1'^{-1} F_0 \hat{x}_0
\end{aligned} \tag{3.2.36}
$$

Tedy

$$
\hat{x}_1' = F_0 \hat{x}_0 \tag{3.2.37}
$$

Porovnáním dalšího smíšeného členu z (3.2.34) a (3.2.33)

$$
\begin{aligned}
+\hat{x}_1'^T P_1'^{-1} &= \hat{x}_0^T P_0^{-1} (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} \\
+\hat{x}_1' P_1'^{-1} &= \hat{x}_0^T P_0^{-1} [P_0 - P_0 F_0^T (Q_0 + F_0 P_0 F^T)^{-1} F_0 P_0] F_0^T Q_0^{-1} \\
+\hat{x}_1'^T P_1'^{-1} &= \hat{x}_0^T F_0^T [I - (Q_0 + F_0 P_0 F^T)^{-1} F_0 P_0 F_0^T] Q_0^{-1} \\
+\hat{x}_1'^T P_1'^{-1} &= \hat{x}_0^T F_0^T [(Q_0 + F_0 P_0 F^T)(Q_0 + F_0 P_0 F^T)^{-1} \\
&- (Q_0 + F_0 P_0 F^T)^{-1} F_0 P_0 F_0^T] Q_0^{-1} \\
+\hat{x}_1'^T P_1'^{-1} &= \hat{x}_0^T F_0^T (Q_0 + F_0 P_0 F^T)^{-1} \\
F_0 \hat{x}_0 &= \hat{x}_1'
\end{aligned} \tag{3.2.38}
$$

Zbývá prověřit poslední člen z (3.2.35) a (3.2.36). Musí platit

$$
\hat{x}_0^T [P_0^{-1} (P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1} - P_0^{-1}] \hat{x}_0 = \hat{x}_1'^T P_1'^{-1} \hat{x}_1'
$$

Po úpravě vnitřní závorky dostaneme

$$\begin{aligned}
\hat{x}_1'^T P_1'^{-1} \hat{x}_1' &= \hat{x}_0^T \{P_0^{-1}[P_0 - P_0 F^T (Q_0 + F_0 P_0 F_0^T)^{-1} F P_0] P_0^{-1} - P_0^{-1}\} \hat{x}_0 \\
\hat{x}_1'^T P_1'^{-1} \hat{x}_1' &= \hat{x}_0^T [P_0^{-1} - F^T (Q_0 + F_0 P_0 F_0^T)^{-1} F - P_0^{-1}] \hat{x}_0 \\
\hat{x}_1'^T P_1'^{-1} \hat{x}_1' &= \hat{x}_0^T F^T P_1'^{-1} F \hat{x}_0 \\
\hat{x}_1'^T P_1'^{-1} \hat{x}_1' &= \hat{x}_1'^T P_1'^{-1} \hat{x}_1' \qquad (3.2.39)
\end{aligned}$$

Exponent (3.2.32) je kvadratická forma

$$\frac{1}{2}(b^T A b - c) = -\frac{1}{2}(x_1 - \hat{x}_1') P_1'^{-1} (x_1 - \hat{x}_1') \qquad (3.2.40)$$

kde $\hat{x}_1' = F_0 \hat{x}_0$, $P_1' = F_0 P_0 F_0^T + Q_0$.

Nyní se vrátíme k (3.2.32), k úpravě konstanty před exponenciálou. Konstantu lze upravit následujícím způsobem

$$\begin{aligned}
\frac{(detA)^{\frac{1}{2}}}{(2\pi)^{\frac{nx}{2}} (detP_0)^{\frac{1}{2}} (detQ_0)^{\frac{1}{2}}} &= \frac{[det(P_0^{-1} + F_0^T Q_0^{-1} F_0)^{-1}]^{\frac{1}{2}}}{(2\pi)^{\frac{nx}{2}} (detP_0)^{\frac{1}{2}} (detQ_0)^{\frac{1}{2}}} \\
&= \frac{1}{(2\pi)^{\frac{nx}{2}} (detQ_0)^{\frac{1}{2}} [det(I + P_0 F_0^T Q_0^{-1} F_0)]^{\frac{1}{2}}} \\
&= \frac{1}{(2\pi)^{\frac{nx}{2}} (detQ_O)^{\frac{1}{2}} [det(I + F_0 P_0 F_0^T Q_0^{-1})]^{\frac{1}{2}}} \\
&= \frac{1}{(2\pi)^{\frac{nx}{2}} [det(Q_0 + F_0 P_0 F_0^T)]^{\frac{1}{2}}} \\
&= \frac{1}{(2\pi)^{\frac{nx}{2}} (detP_1')^{\frac{1}{2}}} \qquad (3.2.41)
\end{aligned}$$

Konečně je odvození uzavřeno. Výraz (3.2.29) s ohledem na (3.2.13), (3.2.31), (3.2.34), (3.2.40), (3.2.41) můžeme zapsat takto

$$p(x_1 \mid z^0) = N(x_1 : \hat{x}_1', P_1') \qquad (3.2.42)$$

kde $\hat{x}_1' = F_0 \hat{x}_0$
$P_1' = F_0 P_0 F_0^T + Q_0$

Prediktivní hustota pravděpodobnosti obdobně jako filtrační hustota pravděpodobnosti je gaussovská. Je vidět, že prediktivní hustota v $k = 1$ je opět gaussovská jako v $k = 0$. To znamená, že výpočet je platný formálně pro jakékoliv k. Shrňme proto dosažený výsledek.
Výpočet filtrační hustoty pravděpodobnosti:

$$p(x_k \mid z^k) = N(x_k : \hat{x}_k, P_k) \qquad k = 0, 1, 2, ... \qquad (3.2.43)$$

$$\hat{x}_k = \hat{x}_k' + P_k' H_k^T [R_k + H_k P_k' H_k^T]^{-1} [z_k - H_k \hat{x}_k'] \qquad (3.2.44)$$

$$P_k = P_k' - P_k' H_k^T [R_k + H_k P_k' H_k^T]^{-1} H_k P_k' \qquad (3.2.45)$$

Výpočet prediktivní hustoty pravděpodobnosti:

$$p(x_{k+1} \mid z^k) = N(x_{k+1} : \hat{x}'_{k+1}, P'_{k+1}) \quad k = 0, 1, 2, .. \tag{3.2.46}$$

$$\hat{x}'_{k+1} = F_k \hat{x}_k \tag{3.2.47}$$

$$P'_{k+1} = F_k P_k F_k^T + Q_k \tag{3.2.48}$$

Poznamenejme, že předpokládáme

$$p(x_0 \mid z^{-1}) = p(x_0) = N(x_0 : \hat{x}'_0, P'_0) \tag{3.2.49}$$

Rovnice (3.2.43)-(3.2.49) definují Kalmanův filtr.

Na závěr poznamenejme, že kdybychom uvažovali v rovnici (3.1.1) i vstupní signál u_k, tedy

$$x_{k+1} = F_k x_k + u_k + w_k$$

známý v kroku k, celé odvození bude analogické a jediná změna, která nastane ve výsledných vztazích (3.2.43)-(3.2.49) by nastala v (3.2.47), která by nabyla tvar

$$\hat{x}'_{k+1} = F_k \hat{x}_k + u_k$$

což je přirozené s ohledem na charakter signálu. Analogicky bychom postupovali při změně předpokladu o nulové střední hodnotě stavového šumu a šumu měření. Pak bychom dostali místo (3.2.47)

$$\hat{x}'_{k+1} = F_k \hat{x}_k + \hat{w}_k$$

kde $E[w_k] = \hat{w}_k$ a místo (3.2.45)

$$\hat{x}_k = \hat{x}'_k + P'_k H_k^T [R_k + H_k P'_k H_k^T]^{-1} [z_k - H_k \hat{x}'_k - \hat{v}_k]$$

kde $E[v_k] = \hat{v}_k$.

3.2.2 Nepřímý přístup k syntéze filtru

Přímý přístup představený v předchozí sekci je založen na úpravě bayesovských vztahů pomocí základních matematických operací. Relativní složitost odvození je pak důsledek zvoleného přístupu. Proto lze v literatuře najít i alternativní, ve svém důsledku výrazně snazší, přístup k odvození Kalmanova filtru, který vychází ze základních vět pravděpodobnosti a statistiky. Tedy, než přejdeme k vlastnímu odvození, připomeňme si základní vlastnosti veličin popsaných gaussovskou sdruženou hustotou pravděpodobnosti.

Věta 3.2.1 Uvažujme *nezávislé* náhodné veličiny x a v popsané gaussovskými hustotami pravděpodobnosti $p(x) = N(x : \hat{x}', P'_x)$ a $p(v) = N(v : 0, R)$. Dále uvažujme náhodnou veličinu y danou následující lineární transformací

$$y = Gx + v \tag{3.2.50}$$

kde G je známá matice vhodné dimenze. Pak sdružená hustota pravděpodobnosti veličin x a y je dána

$$\begin{aligned} p\left(x,y\right) &= p(y|x)p(x) \\ &= N\left(\begin{bmatrix}x\\y\end{bmatrix} : \begin{bmatrix}\hat{x}'\\\hat{y}'\end{bmatrix}, \begin{bmatrix}P'_x & P'_{xy}\\P'_{yx} & P'_y\end{bmatrix}\right) \\ &= N\left(\begin{bmatrix}x\\y\end{bmatrix} : \begin{bmatrix}\hat{x}'\\G\hat{x}'\end{bmatrix}, \begin{bmatrix}P'_x & P'_xG^T\\GP'_x & GP'_xG^T+R\end{bmatrix}\right) \end{aligned} \tag{3.2.51}$$

kde pravděpodobnost $p(y|x)$, jak plyne z (3.2.50), je

$$p(y|x) = p_v(y - Gx) = N(y : Gx, R) \tag{3.2.52}$$

Marginální hustota y je

$$\begin{aligned} p(y) &= \int p(x,y)dx = N(y : \hat{y}', P'_y) \\ &= N(y : G\hat{x}', GP'_xG^T + R) \end{aligned} \tag{3.2.53}$$

Věta 3.2.2 Uvažujme sdruženě gaussovské náhodné veličiny x a y

$$p\left(x,y\right) = N\left(\begin{bmatrix}x\\y\end{bmatrix} : \begin{bmatrix}\hat{x}'\\\hat{y}'\end{bmatrix}, \begin{bmatrix}P'_x & P'_{xy}\\P'_{yx} & P'_y\end{bmatrix}\right) \tag{3.2.54}$$

kde $P'_{xy} = (P'_{yx})^T$. Pak podmíněná hustota pravděpodobnosti veličiny x při známé hodnotě náhodné veličiny y je

$$p(x|y) = N(x : \hat{x}' + P'_{xy}(P'_y)^{-1}(y - \hat{y}'), P'_x - P'_{xy}(P'_y)^{-1}P'_{yx}) \tag{3.2.55}$$

Tedy, předpokládejme danou počáteční prediktivní hustotu pravděpodobnosti

$$p(x_0|z^{-1}) = N(x_0 : \hat{x}'_0, P'_0) \tag{3.2.56}$$

a rovnici měření (3.1.2) vedoucí na gaussovskou hustotu pravděpodobnosti měření

$$p(z_0|x_0) = N(z_0 : Hx_0, R_0) \tag{3.2.57}$$

a začněme odvozením *filtračního kroku* z Bayesova vztahu (2.4.9) v čase $k = 0$

$$\begin{aligned} p(x_0|z^0) &= \frac{p(x_0, z_0|z^{-1})}{p(z_0|z^{-1})} \\ &= \frac{p(z_0|x_0)p(x_0|z^{-1})}{p(z_0|z^{-1})} \\ &= \frac{p(z_0|x_0)p(x_0|z^{-1})}{\int p(z_0|x_0)p(x_0|z^{-1})dx_0} \end{aligned} \tag{3.2.58}$$

Dle (3.2.56), (3.2.57) a věty 3.2.1, sdružená prediktivní hustota pravděpodobnosti stavu x_0 a měření z_0 je dána

$$\begin{aligned} p(x_0, z_0|z^{-1}) &= p(z_0|x_0)p(x_0|z^{-1}) \\ &= N\left(\begin{bmatrix}x_0\\z_0\end{bmatrix} : \begin{bmatrix}\hat{x}'_0\\\hat{z}'_0\end{bmatrix}, \begin{bmatrix}P'_0 & P'_{xz,0}\\P'_{zx,0} & P'_{z,0}\end{bmatrix}\right) \end{aligned} \tag{3.2.59}$$

kde

$$\hat{z}'_0 = E[z_0|z^{-1}] \tag{3.2.60}$$
$$= E[H_0x_0 + v_0|z^{-1}] = E[H_0x_0|z^{-1}] + E[v_0]$$
$$= H_0\hat{x}'_0 \tag{3.2.61}$$
$$P'_{z,0} = cov[z_0|z^{-1}] = E[(z_0 - \hat{z}'_0)(z_0 - \hat{z}'_0)^T|z^{-1}] \tag{3.2.62}$$
$$= E[(H_0x_0 + v_0 - H_0\hat{x}'_0)(H_0x_0 + v_0 - H_0\hat{x}'_0)^T|z^{-1}]$$
$$= H_0E[(x_0 - \hat{x}'_0)(x_0 - \hat{x}'_0)^T|z^{-1}]H_0^T + E[v_0v_0^T]$$
$$= H_0P'_0H_0^T + R_0 \tag{3.2.63}$$
$$P'_{xz,0} = cov[x_0, z_0|z^{-1}] = E[(x_0 - \hat{x}'_0)(z_0 - \hat{z}'_0)^T|z^{-1}] \tag{3.2.64}$$
$$= E[(x_0 - \hat{x}'_0)(x_0 - \hat{x}'_0)^T|z^{-1}]H_0^T$$
$$= P'_0H_0^T \tag{3.2.65}$$

jsou prediktivní momenty odhadu měření a stavu. Ve filtračním kroku estimátoru předpokládáme dostupné měření z_0 a zajímá nás, jak znalost měření ovlivní prediktivní odhad stavu $p(x_0|z^{-1})$. K tomu můžeme využít větu 3.2.2, která vede na finální tvar filtrační hustoty pravděpodobnosti odhadu stavu $p(x_0|z_0, z^{-1}) = p(x_0|z^0) = N(x_0 : \hat{x}_0, P_0)$ s momenty

$$\hat{x}_0 = \hat{x}'_0 + P'_{xz,0}(P'_{z,0})^{-1}(z_0 - \hat{z}'_0) \tag{3.2.66}$$
$$P_0 = P'_0 - P'_{xz,0}(P'_{z,0})^{-1}(P'_{xz,0})^T \tag{3.2.67}$$

Pokračujme odvozením *prediktivního kroku* z Chapman-Kolmogorovovy rovnice (2.4.10) v čase $k = 1$

$$p(x_1|z^0) = \int p(x_1, x_0|z^0)dx_0$$
$$= \int p(x_1|x_0)p(x_0|z^0)dx_0 \tag{3.2.68}$$

Za předpokladu gaussovské filtrační hustoty počáteční hustotu pravděpodobnosti

$$p(x_0|z^0) = N(x : \hat{x}_0, P_0) \tag{3.2.69}$$

a gaussovské hustoty pravděpodobnosti stavu

$$p(x_1|x_0) = N(x_1 : Fx_0, Q_0) \tag{3.2.70}$$

danou rovnicí dynamiky (3.1.1), prediktivní hustota stavu je, dle výpočtu marginální hustoty pravděpodobnosti ve větě 3.2.1, opět gaussovská $p(x_1|z^0) = N(x_1 : \hat{x}'_1, P'_1)$ s momenty

$$\hat{x}'_1 = E[x_1|z^0] \tag{3.2.71}$$
$$= E[F_0x_0 + w_0|z^0]$$
$$= F_0\hat{x}_0 \tag{3.2.72}$$
$$P'_1 = cov[x_1|z^0] \tag{3.2.73}$$
$$= E[(F_0x_0 + w_0 - F_0\hat{x}_0)(F_0x_0 + w_0 - F_0\hat{x}_0)^T|z^0]$$
$$= F_0P_0F_0^T + Q_0 \tag{3.2.74}$$

Prediktivní hustota je opět gaussovská a tedy výpočet filtračních a prediktivních hustot pravděpodobnosti probíhá analogicky pro jakýkoliv časový okamžik k.

Porovnáním vztahů (3.2.44), (3.2.45) a (3.2.66), (3.2.67) a (3.2.47), (3.2.48) a (3.2.72), (3.2.74) vidíme, že oba přístupy k návrhu Kalmanova filtru vedou k stejnému algoritmu.

3.3 Vlastnosti filtru

V předchozích dvou sekcích této kapitoly byl definován lineární gaussovský systém a navržen rekurzivní algoritmus generující odhad neznámého stavu tohoto systému. Za nejvýznamnější lze považovat skutečnost, že hustoty pravděpodobnosti (3.2.43), (3.2.46) jsou *reprodukovatelné*, tj. při uvažování Gaussovské počáteční podmínky (3.2.1) a lineárního Gaussovského systému (3.1.1), (3.1.2) jsou všechny filtrační a prediktivní podmíněné hustoty odhadu stavu (3.2.58), (3.2.68) taktéž gaussovské. Jelikož se jedná o gaussovské hustoty pravděpodobnosti, první a druhý moment postačuje k úplnému popisu stavu (3.2.44), (3.2.45), (3.2.47), (3.2.48).

3.3.1 Explicitní jednokrokový prediktor

Bayesovský přístup nás přirozeně přivedl k rovnicím reprezentujícím filtraci (3.2.44), (3.2.45) a predikci (3.2.47), (3.2.48). Jestliže (3.2.44) dosadíme do (3.2.47) a (3.2.45) do (3.2.48), dostaneme

$$\hat{x}'_{k+1} = F_k[\hat{x}'_k + P'_k H_k^T (R_k + H_k P'_k H_k^T)^{-1}(z_k - H_k \hat{x}'_k)] \tag{3.3.1}$$

$$P'_{k+1} = F_k[P'_k - P'_k H_k^T (R_k + H_k P'_k H_k^T)^{-1} H_k P'_k] F_k^T + Q_k \tag{3.3.2}$$

Tyto vztahy jsou často označovány jako rovnice Kalmanova filtru a rovnice (3.3.2) je nazývána diferenční Riccatiho rovnicí. V tomto případě se jedná o filtraci v širším slova smyslu, neboť striktně vzato se, dle klasifikace v sekci 2.4, jedná o jednokrokový prediktor. Výrazy

$$K_k = P'_k H_k^T (R_k + H_k P'_k H_k^T)^{-1} \tag{3.3.3}$$

nebo

$$K'_k = F_k P'_k H_k^T (R_k + H_k P'_k H_k^T)^{-1} \tag{3.3.4}$$

jsou označovány jako Kalmanův zisk. Z porovnání vztahů (3.2.44), (3.2.45) a (3.2.66), (3.2.67) plyne, že zisk estimátoru K_k (3.3.3) může být zapsán jako následující součin kovariančních matic predikce měření a stavu

$$K_k = P'_{xz,k}(P'_{z,k})^{-1}$$

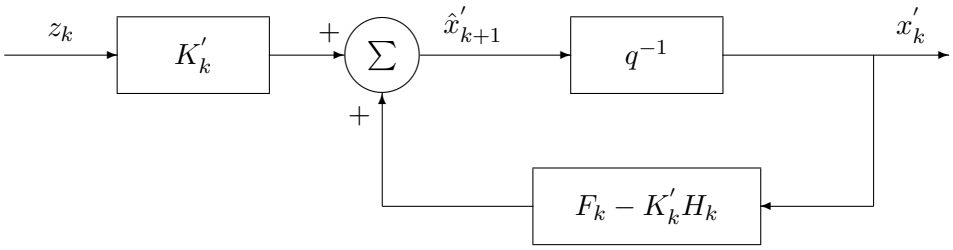
Tento vztah bude užitečný při návrhu některých estimátorů pro nelineární systémy, které jsou diskutovány v následujících kapitolách.

Povšimněme si, že vztah (3.3.1) můžeme alternativně vyjádřit následujícím způsobem

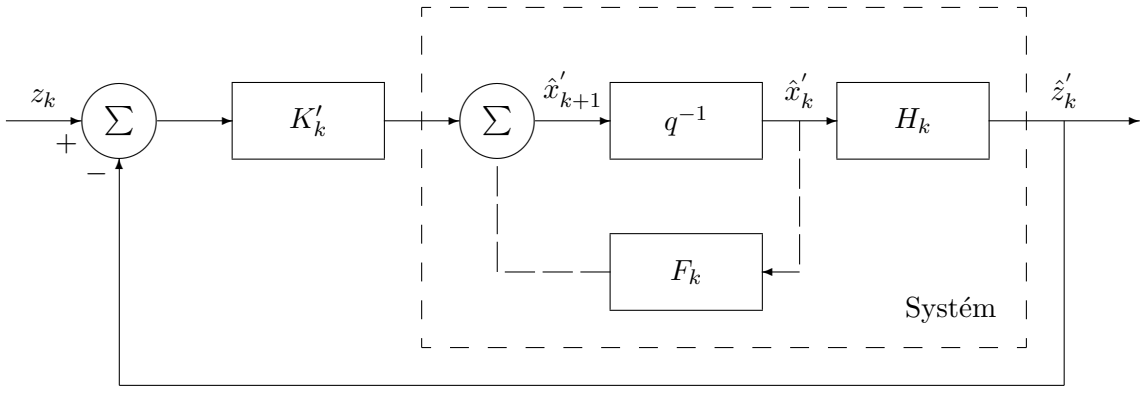
$$\hat{x}'_{k+1} = (F_k - K'_k H_k)\hat{x}'_k + K'_k z_k \tag{3.3.5}$$

Z tohoto vztahu je více zřejmé, že Kalmanův filtr je lineární, t-variantní systém, na jehož vstupu je proces $\{z_k\}$ reprezentující měření a výstup je optimální odhad stavu $\hat{x}'_k = E[x_k \mid z^{k-1}]$. Zajímavé je, že Kalmanův filtr bude t-variantní systém i v případě, že procesy $\{w_k\}$, $\{v_k\}$ budou stacionární a systém, jehož stav je odhadován bude t-invariantní. Tedy $F_k = F, H_k = H, Q_k = Q, R_k = R$ pro všechna k. Rozdíl mezi (3.3.5) a (3.3.1) je více zřejmý z grafického vyjádření struktury prediktorů na obr. 3.3.1 a 3.3.2. Povšimněme si též, že strukturu ilustrovanou na obr. 3.3.2 můžeme interpretovat jako zpětnovazební systém, jehož cílem je regulace chyby predikce měření, tj. inovace $z_k - \hat{z}'_k$, do nuly.

Vraťme se k rovnicím Kalmanova filtru (3.3.1), (3.3.2). Toto vyjádření je jistě výhodné v případě, že filtr je součástí řídicího systému. Výpočet predikované hodnoty stavu probíhá



Obrázek 3.3.1: Signálový model prediktoru podle (3.3.5).



Obrázek 3.3.2: Signálový model prediktoru podle (3.3.1) obsahující „kopii“ struktury systému, jehož stav je odhadován.

mezi okamžiky vzorkování a tudíž v mnoha úlohách neovlivní dobu výpočtu řídicí veličiny. V této práci však budeme preferovat popis Kalmanova filtru ve formě (3.2.43)-(3.2.49), protože lépe vyjadřuje vnitřní práci filtru svým členěním na filtrační a prediktivní část.

3.3.2 Kalmanův filtr jako lineární estimátor s minimální variancí

K odvození filtru jsme vyšli z bayesovských rekurzivních vztahů (2.4.9) a (2.4.10). Dokázali jsme, že pro filtrační hustotu platí

$$N(x_k : \hat{x}_k, P_k) = \frac{N(x_k : \hat{x}'_k, P'_k)N(z_k : H_k x_k, R_k)}{N(z_k : H_k \hat{x}'_k, H_k P'_k H_k^T + R_k)} \tag{3.3.6}$$

a pro prediktivní hustotu platí

$$N(x_{k+1} : \hat{x}'_{k+1}, P'_{k+1}) = \int N(x_k : \hat{x}_k, P_k)N(x_{k+1} : F_k x_k, Q_k)dx_k \tag{3.3.7}$$

Právě tyto vztahy spolu s přirozeným členěním estimačních algoritmů na filtrační a prediktivní část budeme s výhodou často používat v dalších kapitolách.

Sledujme nyní rovnice Kalmanova filtru z pohledu závislosti podmíněné střední hodnoty a kovariance stavu na měření. Snadno nahlédneme, že odvozený filtr patří do třídy lineárních filtrů. Totiž z (3.2.44), (3.3.1) je zřejmé, že odhady $\hat{x}_k, \hat{x}'_{k+1}$ jsou lineární funkcí měření a z (3.2.43), (3.2.48) vyplývá, že vývoj P_k, P'_{k+1} není závislý na měření. Díky nezávislosti na měření můžeme v případě potřeby vypočítat P_k, P'_{k+1}, K_k, K'_k již v $k = 0$ pro všechna k. Tyto vlastnosti jsou velmi vyjímečné a jsou umožněny předpoklady na systém (sekce 3.1), tj. předpokladem lineárního gaussovského systému.

Poznamenejme také, že algoritmus Kalmanova filtru může být odvozen alternativním způsobem, a to minimalizací následující kriteriální funkce

$$V(\hat{x}_k) = E[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T]$$

Tento vztah ukazuje, že je možné chápat Kalmanův filtr jako lineární estimátor poskytující odhady s minimální variancí. Je vhodné podotknout, že odvození filtru skrze minimalizaci kriteria nevyžaduje předpoklad gaussovských náhodných veličin, tak jak je tomu u bayesovského přístupu. Odvození, které může být nalezeno v [9], [10], [84], v tomto případě spočívá v nalezení rovnic pro rekurzivní výpočet prvních dvou momentů odhadu stavu bez ohledu na hustotu pravděpodobnosti.

Poznámka . Dle zvoleného kriteria se, v anglicky psané literatuře, pro výsledný filtr vžil název „linear minimum mean square error estimator“ (LMMSE).

3.3.3 Konvergence

Pro praktické využití Kalmanova filtru je vhodné znát limitní vlastnosti odhadu a podmínky konvergence filtru. V této části proto uvedeme několik základních vlastností odhadu Kalmanova filtru, který je odvozen pro t-invariantní lineární gaussovský systém. Detailnější diskuze spolu s důkazy pak může být nalezena např. v [9].

Uvažujeme-li t-invariantní systém, tj. $F_k = F, H_k = H, Q_k = Q$ a $R_k = R, \forall k$, pak *diferenční* Riccatiho rovnice (3.3.2), která charakterizuje chybu prediktivního odhadu $\hat{x}'_{k+1}$, nabývá tvaru

$$P'_{k+1} = F[P'_k - P'_k H^T (R + H P'_k H^T)^{-1} H P'_k] F^T + Q$$

Definujme rozklad S_Q kovarianční matice Q tak, že $Q = S_Q S_Q^T$. Lze dokázat, že pokud dvojice (H, F) je pozorovatelná a dvojice (F, S_Q) je dosažitelná, pak existuje jediné *pozitivně definitní* řešení *algebraické* Riccatiho rovnice

$$P' = F[P' - P' H^T (R + H P' H^T)^{-1} H P'] F^T + Q$$

kde $\lim_{k\to\infty} P'_k = P$. Dále, pokud počáteční podmínka P'_0 je pozitivně definitní, pak řešení diferenční Riccatiho rovnice konverguje k řešení algebraické Riccatiho rovnice.

Při splnění výše uvedených podmínek, tj. pozorovatelnost dvojice (H, F) a dosažitelnost dvojice (F, S_Q), kovarianční matice chyby prediktivního odhadu konverguje k ustálené hodnotě. První z podmínek je vcelku přirozená. Abychom mohli odhadnout stav, je nutné, aby systém, tj. stav systému v každém okamžiku, byl pozorovatelný. Pokud některá složka stavu nebude pozorovatelná, nelze ani očekávat, že kovarianční matice P'_k se ustálí. Spíše, odhadovaná variance, příslušná nepozorovatelné složce stavu, poroste s přibývajícím časem. Druhá podmínka je možná více překvapující. Ta říká, že požadujeme, aby stavový šum ovlivnil všechny složky stavu. Pokusíme se ilustrovat význam druhé podmínky na příkladu. Předpokládejme, že existuje nějaká složka stavu neovlivněná stavovým šumem. Tuto složku stavu pak můžeme interpretovat jako neznámou, avšak deterministicky popsanou, proměnnou (nebo-li, využijeme-li názvosloví z metody nejmenších čtverců, parametr). Pak lze neznámý parametr odhadnout, pro $k \to \infty$, "naprosto" přesně. Variance příslušná této složce tedy bude konvergovat do nuly, a tak výsledná matice P' nebude pozitivně definitní, ale semi-definitní.

3.3.4 Konzistence odhadu

Kalmanův filtr poskytuje odhady stavu ve formě podmíněných momentů, a to filtrační střední hodnoty $\hat{x}_k$ (3.2.44) a kovarianční matice P_k (3.2.45) a prediktivní střední hodnoty $\hat{x}'_k$ (3.2.47) a kovarianční matice P'_k (3.2.48). Pokud jsou splněny podmínky kladené na model diskutované v předchozí části a model použitý pro návrh Kalmanova filtru odpovídá realitě, tj. generátoru dat, pak kovarianční matice přesně charakterizuje chybu odhadu střední hodnoty. Tato vlastnost se obecně nazývá konzistencí odhadu. Avšak z důvodu výskytu nemodelovaných (náhlých) chyb v měření, např. z důvodu poruchy senzoru, může dojít ke ztrátě konzistence odhadu, což, v mnoha aplikacích, může vést k fatálním důsledkům. V této části si představíme základní techniku pro monitorování konzistence odhadu formou jednoduchého příkladu.

Motivace . Dle mezinárodního standardu o navigačních systémech v letectví [86] může být navigační systém založen na Kalmanově filtru. Jako příklad takového systému uveďme inerciální navigační systém Honeywell Laseref VI. Takový navigační systém poskytuje nejen odhad polohy, rychlosti a natočení letadla (tzv. navigační informaci), ale i příslušné kovarianční matice charakterizující chybu navigační informace. Navigační informace spolu s kovariančními maticemi je pak využita pro plánování následné trajektorie letu. Nedetekovaný nekonzistentní odhad pak může vést nejen k neefektivnímu plánování letu, ale i k havárii letadla.

Příklad 3.3.1 Předpokládejme skalární lineární gaussovský t-invariantní systém popsaný modelem (3.1.1) a (3.1.2) generující skuteční stav x_k a předpokládejme Kalmanův filtr (3.2.43)–(3.2.49), navržený pro tento model, který poskytuje prediktivní odhad ve formě střední hodnoty

$\hat{x}'_k$ a variance P'_k pro $k = 0, 1, \ldots, T$. Předpokládejme dále, že vliv počátečních podmínek filtru je zanedbatelný a existuje ustálené řešení Riccatiho rovnice (3.3.2) pro prediktivní varianci P'. Říkáme, že prediktivní odhad je *konzistentní* pokud

$$P' = M'$$

kde M' je odhad variance chyby odhadu

$$M' = \frac{1}{T}\sum_{k=1}^{T}(x_k - \hat{x}'_k)^2$$

pro $T \to \infty$. Pokud platí $P' > M'$, říkáme, že odhad je *pesimistický*, tzn. variance poskytovaná filtrem je větší než aktuální chyba odhadu střední hodnoty, a pokud $P' < M'$, tak odhad je *optimistický*, tzn. filtr poskytuje varianci, která je menší než aktuální chyba odhadu stavu. Zejména optimistický odhad tak může být chápán jako hlavní nebezpečí při využití odhadu v aplikacích se zvýšenými požadavky na bezpečnost.
Za předpokladu nominálních podmínek, tj. kdy model použitý pro návrh filtru odpovídá generátoru dat, bude odhad poskytovaný Kalmanovo filtrem konzistentní. Vinou nemodelovaných či neočekávaných chyb ovlivňujících systém se však může odhad stát nekonzistentním. Jako jednu z příčin ztráty konzistence odhadu můžeme uvést nedetekovanou poruchu senzoru či náhlou změnu pracovním podmínek (např. okolní teplotu). Pak skutečná rovnice měření může být zapsána

$$z_k = Hx_k + b_k + v_k \tag{3.3.8}$$

kde b_k představuje poruchu o neznámých vlastnostech. Avšak, Kalmanův filtr je navržen za předpokladu platnosti rovnice (3.1.2), tj. bez vědomí přítomnosti poruchy b_k.
Základní technika pro detekci nemodelovaných chyb v měření, a tím i nekonzistentního odhadu, je založena na statistickém testu hypotéz. Definujme nulovou hypotézu

- H_0 : prediktivní odhad stavu je konzistentní a rovnice měření z_k (3.1.2) je platná, a tedy jednokroková predikce měření je, dle (3.2.3), dána hustotou $p(z_k|z^{k-1}) = N(z_k : \hat{z}'_k, P'_{z,k})$ se střední hodnotou $\hat{z}'_k = E[z_k|z^{k-1}] = H\hat{x}'_k$ a variancí $P'_{z,k} = var[z_k|z^{k-1}] = H^2P' + R$

kterou budeme testovat oproti alternativní hypotéze

- H_1 : prediktivní odhad stavu není konzistentní a rovnice měření (3.1.2) není platná,

při dané hladině významnosti α. Hladina významnosti, či pravděpodobnost chyby I. druhu, reprezentuje pravděpodobnost, kdy nulová hypotéza je neoprávněně zamítnuta. Vlastní test platnosti hypotézy H_0 je pak dán následujícím algoritmem.

Algoritmus pro detekci nemodelovaných chyb v měření pomocí Kalmanova filtru

---

(i) Specifikujme hladinu významnosti α. Typicky je volena jako malé číslo, např. $\alpha \in \langle 0,0001;\ 0,01\rangle$ v závislosti na aplikaci.

(ii) V časovém okamžiku k, spočtěme kvantily $q_{k,\frac{\alpha}{2}}$ a $q_{k,1-\frac{\alpha}{2}}$ normálního rozdělení $p(z_k|z^{k-1})$ pro pravděpodobnost $\frac{\alpha}{2}$ a $(1 - \frac{\alpha}{2})$. K výpočtu může být použita např. funkce `norminv` z prostředí MATLAB®.

(iii) Porovnejme aktuální měření s kvantily. Pokud

$$z_k \in \langle q_{k,\frac{\alpha}{2}}, q_{k,1-\frac{\alpha}{2}} \rangle \tag{3.3.9}$$

pak hypotéza H_0 je považována za platnou a aktuální měření z_k tedy není ovlivněno nemodelovanou chybou. Pokud však

$$z_k \notin \langle q_{k,\frac{\alpha}{2}}, q_{k,1-\frac{\alpha}{2}} \rangle \tag{3.3.10}$$

pak hypotéza H_0 je zamítnuta. V této situaci je vystaveno varování, že měření z_k je ovlivněno nemodelovanou chybou a nemělo by být použito ve filtru.

Uvažujeme-li T dostupných měření a platnost hypotézy H_0, pak bychom měli pozorovat následující (očekávaný) počet varování

$$T_{alert} = \alpha \times T$$

která lze chápat jako planá varování (v anglicky psané literatuře označovaná jako "false alert"). Pokud platí hypotéza H_1, tj. je zde nemodelovaná chyba b_k, budeme pozorovat větší počet varování než T_{alert}.
Představená metoda pro detekci chyb v rovnici měření je z podstaty vhodnější pro detekci tzv. hrubých chyb v měření. Pro detekci tzv. pomalu rostoucích chyb při daném omezení na detekční čas byly navrženy jiné techniky. Širší diskuze o detekci nekonzistentního chování filtru může být nalezena např. v [87], [89].

Poznámka . Všimněme si, že v oblasti odhadu stavu se používá jiná definice konzistence odhadu, než tomu je u identifikačních metod.

3.3.5 Numerická stabilita a výpočetní nároky

Vyjma konzistence a konvergence odhadu je, při implementaci filtru, nutné brát v potaz také případné numerické chyby a jejich dopad kvalitu odhadu. Zejména musí být zajištěno, aby případné zaokrouhlovací chyby nevedly ke ztrátě symetričnosti a pozitivní semi-definitnosti kovariančních matic chyby odhadu stavu. Z hlediska citlivosti k numerickým chybám při výpočtu, nejproblematičtějším je výpočet filtrační kovarianční matice P_k (3.2.45), a to z důvodu odčítání dvou maticových členů. Proto byla v literatuře věnována velká pozornost odvození numericky stabilních verzí Kalmanova filtru.
Podobně jak tomu bylo u rekurzivních identifikačních technik, i zde lze rozlišit dva hlavní přístupy. Jeden spočívá v přímém rekurzivním výpočtu odmocnin kovariančních matic, tj. namísto filtračních a prediktivních matic P_k a P'_k jsou rekurzivně počítány matice S_k a S'_k, pro než platí

$$P_k = S_k S_k^T$$
$$P'_k = S'_k (S'_k)^T$$

Výsledné kovarianční matice jsou pak zaručeně symetrické a pozitivně semi-definitní. Druhý způsob zvyšující numerickou stabilitu Kalmanova filtru je výpočet filtrační kovarianční matice v tzv. Josephově formě

$$P_k = (I - K_k H_k) P'_k (I - K_k H_k)^T + K_k R_k K_k^T$$

kde K_k je Kalmanův zisk daný (3.3.3).
Poznamenejme, že i v případě použití Josephovy formy je nutné počítat inverzi matice $H_k P'_k H_k^T + R_k$, která je nutná pro výpočet Kalmanova zisku. Inverze matice je obecně náročná a na numerické chyby citlivá operace. Proto byly navrženy verze Kalmanova filtru se sekvenčním zpracováním dílčích měření ve vektoru z_k, kde inverze matice je nahrazena sekvencí podílů. Detailní algoritmy mohou být nalezeny např. v [87], [96].

3.3.6 Inovační forma

Na závěr této sekce se vrátíme ke vztahům (3.3.1), (3.3.4). Dosazením (3.3.4) do (3.3.1) dostaneme

$$\hat{x}'_{k+1} = F_k \hat{x}'_k + K'_k (z_k - H_k \hat{x}'_k) \tag{3.3.11}$$

Definujeme $\tilde{z}_k$ takto

$$\tilde{z}_k \stackrel{\triangle}{=} z_k - H_k \hat{x}'_k \tag{3.3.12}$$

Proces $\{\tilde{z}_k\}$, označovaný jako inovační posloupnost, umožní vyjádřit alternativní popis Kalmanova filtru, který bude užitečný při převodu mezi stavovým a vstupně-výstupní modelem. Tedy

$$\hat{x}'_{k+1} = F_k \hat{x}'_k + K'_k \tilde{z}_k \tag{3.3.13}$$

$$z_k = H_k \hat{x}'_k + \tilde{z}_k \tag{3.3.14}$$

Tento alternativní model budeme označovat pojmem inovační model. Inovační model tedy generuje z_k pomocí inovace $\tilde{z}_k$. Můžeme však vyjádřit i obrácenou závislost. Generovat $\tilde{z}_k$ pomocí z_k, pak

$$\hat{x}'_{k+1} = (F_k - K'_k H_k) \hat{x}'_k + K'_k z_k \tag{3.3.15}$$

$$\tilde{z}_k = z_k - H_k \hat{x}'_k \tag{3.3.16}$$

Rovnice (3.3.15), (3.3.16) definují systém, jehož vstup je z_k a výstup $\tilde{z}_k$. Vztahy (3.3.13), (3.3.14) a (3.3.5), (3.3.16) tak lze chápat jako alternativní způsoby zápisu optimálního lineárního filtru.

V této sekci jsme si všimli některých důležitých vlastností Kalmanova filtru a zdůraznili jeho vyjímečné vlastnosti, které jsou spojeny s předpoklady na systém týkající se linearity a gaussovosti. V následující sekci ukážeme vztah mezi stochastickými fenomenologickými modely a stavovými modely, který odvodíme využitím inovačního modelu neboli jedné z alternativních reprezentací Kalmanova filtru.

3.4 Převod stavového modelu na fenomenologický a naopak

V teorii automatického řízení se často kromě stavových modelů používají i modely fenomenologické, nebo-li vstupně-výstupní, [19], [24], [25]. Rovněž práce z oblastí zpracování signálu [26], [27], identifikace systémů [28] a časových řad [21], [26], [29] využívají fenomenologických modelů. Proto je užitečné ukázat, jak lze přejít od stochastických stavových modelů k fenomenologickým a naopak.

Uvažujme následující fenomenologický ARMA model

$$z_k = \frac{C(q^{-1})}{D(q^{-1})} e_k \tag{3.4.1}$$

kde q^{-1} je operátor zpětného posunu ($q^{-1} z_k = z_{k-1}$)
$C(q^{-1}), D(q^{-1})$ jsou polynomy
$C(q^{-1}) = 1 + c_1 q^{-1} + c_2 q^{-2} + ... + c_{nc} q^{-nc}$
$D(q^{-1}) = 1 + d_1 q^{-1} + d_2 q^{-2} + ... + c_{nd} q^{-nd}$
e_k je bílý šum s nulovou střední hodnotou a kovariancí R_e.

Vztah (3.4.1) můžeme zapsat nyní též takto

$$z_k = G(q^{-1}; \Theta) e_k \tag{3.4.2}$$

kde Θ je vektor parametrů obsahující koeficienty polynomů $C(q^{-1}), D(q^{-1})$ a $G(q^{-1}; \Theta)$ je racionální lomená funkce $C(q^{-1})/D(q^{-1})$.
V případě, že měření z_k a šum e_k jsou nz dimenzionální vektory, pak $G(\cdot;\cdot)$ je matice nz/nz, jejímiž prvky jsou racionální funkce. V teorii časových řad je proces $\{z_k\}$ nazýván ARMA proces, jehož speciálními případy je autoregresní proces AR ($C(q^{-1}) = 1$) a klouzavý průměr MA ($D(q^{-1}) = 1$).

V další části se budeme zabývat problémem jak lze přejít od stavového modelu typu (3.1.1), (3.1.2) k fenomenologickému modelu typu (3.4.2).

Uvažujme lineární gaussovský model (3.1.1), (3.1.2), který je t-invariantní, tj.

$$x_{k+1} = F x_k + w_k \tag{3.4.3}$$

$$z_k = H x_k + v_k \tag{3.4.4}$$

kde $\{w_k\}, \{v_k\}$ jsou bílé, vzájemně nezávislé šumy se střední hodnotou nula a kovariancemi

$$\begin{aligned} cov[w_k] &= Q \\ cov[v_k] &= R \\ cov[w_k, v_k] &= 0 \end{aligned}$$

a s počátečním stavem x_0 nezávislým na $\{w_k\}, \{v_k\}$ se střední hodnotou a kovariancí

$$\begin{aligned} E[x_0 \mid z^{-1}] &= \hat{x}'_0 \\ cov[x_0 \mid z^{-1}] &= P'_0 \end{aligned}$$

Připomeňme, že $\{w_k\}$ je nx dimenzionální proces a $\{v_k\}$ nz dimenzionální proces. Z pohledu na (3.4.2) je zřejmé, že zde vystupuje pouze jeden náhodný proces, který je nz dimenzionální. Je tedy převod (3.4.3), (3.4.4) na (3.4.2), tj. přepočet matic stavového modelu F, H, Q, a R na vektor parametrů vstupně-výstupního modelu Θ, možný? Nyní ukážeme, že převod je snadno proveditelný využitím inovačního modelu.

Inovační model pro (3.4.3), (3.4.4) je podle sekce 3.3

$$\hat{x}'_{k+1} = F \hat{x}'_k + K' \tilde{z}_k \tag{3.4.5}$$

$$z_k = H\hat{x}'_k + \tilde{z}_k \tag{3.4.6}$$

kde

$$\tilde{z}_k = z_k - H\hat{x}'_k = H(x_k - \hat{x}'_k) + v_k \tag{3.4.7}$$

$$K' = FPH^T[R + HPH^T]^{-1} \tag{3.4.8}$$

$$P = FPF^T + Q - FPH^T[R + HPH^T]^{-1}HPF^T \tag{3.4.9}$$

Z (3.4.5) a (3.4.6) je zřejmé, že v obou rovnicích je jediný šum, inovace $\tilde{z}_k$ dimenze nz, a tudíž hlavní problém je odstraněn. K (3.4.9) pouze poznamenejme, že se jedná o ustálené řešení Riccatiho rovnice (3.3.2).
Nyní je již snadné zjistit $G(\cdot;\cdot)$ v (3.4.2). Na $\tilde{z}_k$ se lze dívat jako na vstupní známý signál, a tudíž se vlastně jedná o standardní postup transformace deterministických systémů ze stavové reprezentace na vstupně-výstupní (fenomenologickou). Po krátkých úpravách (3.4.5) a (3.4.6) dostaneme

$$\begin{aligned}
(qI - F)\hat{x}'_k &= K'\tilde{z}_k \\
\hat{x}'_k &= (qI - F)^{-1}K'\tilde{z}_k \\
z_k &= H(qI - F)^{-1}K'\tilde{z}_k + \tilde{z}_k \\
&= (I + H(qI - F)^{-1}K')\tilde{z}_k
\end{aligned} \tag{3.4.10}$$

Vyjádření pro $G(q^{-1};\Theta)$, které je porovnáním (3.4.2) a (3.4.10), je pak

$$G(q^{-1};\Theta) = I + H(qI - F)^{-1}K' \tag{3.4.11}$$

Z (3.4.7) je zřejmé, že $R_e = HP'H^T + R$. Převod je tím ukončen. Výstupy stavového modelu (3.4.3), (3.4.4) a fenomenologického modelu (3.4.2) s (3.4.11) mají tedy shodné statistické vlastnosti.
Ukázali jsme použití Kalmanova filtru na úlohu, která zdánlivě nemá nic společného s estimací stavu. Na závěr poznamenejme, že převod tímto způsobem lze uskutečnit, jestliže (3.4.9) má řešení. S podmínkami řešitelnosti Riccatiho rovnice jsme se krátce seznámili v kapitole 3.3.3.

Další část této podkapitoly věnujeme opačnému problému. Budeme předpokládat model typu (3.4.2) a chceme přejít ke stavové reprezentaci. Připomeňme, že stavových reprezentací může být k jedné reprezentaci fenomenologické nekonečně mnoho.
V následujícím příkladu ukážeme převod ARMA modelu 1. řádu na stavový model.

Příklad 3.4.1 Uvažujme ARMA proces $\{z_k\}$

$$z_k + dz_{k-1} = e_k + ce_{k-1} \tag{3.4.12}$$

nebo vyjádřeno ve formě (3.4.2)

$$z_k = \frac{1 + cq^{-1}}{1 + dq^{-1}}e_k \tag{3.4.13}$$

Výraz (3.4.13) můžeme upravit

$$\begin{aligned}
z_k &= \frac{1 + cq^{-1}}{1 + dq^{-1}}e_k \\
&= e_k + \frac{(c - d)q^{-1}}{1 + dq^{-1}}e_k
\end{aligned} \tag{3.4.14}$$

Položme

$$x_k \triangleq \frac{(c-d)q^{-1}}{1+dq^{-1}} e_k \tag{3.4.15}$$

pak dosazením (3.4.15) do (3.4.14) dostaneme

$$z_k = x_k + e_k \tag{3.4.16}$$

a z (3.4.15) dále plyne

$$x_{k+1} = -dx_k + (c-d)e_k \tag{3.4.17}$$

Rovnice (3.4.17) a (3.4.16) představují stavový model.
Tento stavový model obsahuje závislost stavového šumu $(c-d)e_k$ a šumu měření e_k. Pokud bychom chtěli tuto závislost odstranit, můžeme použít postup z podkapitoly 3.1 a dostaneme

$$x_{k+1} = -cx_k + (c-d)z_k \tag{3.4.18}$$

$$z_k = x_k + e_k \tag{3.4.19}$$

Postup uvedený v příkladu 3.4.1 můžeme zobecnit pro ARMA model libovolného řádu. Pro jednoduchost uvažujme pouze jednodimenzionální z_k a e_k. Nechť

$$D(q^{-1})z_k = C(q^{-1})e_k \tag{3.4.20}$$

kde $nd = nc$ Poznamenejme, že předpoklad stejných řádů není omezující a byl zaveden jen pro zjednodušení zápisu. Zavedením substituce

$$x_k = \frac{C(q^{-1}) - D(q^{-1})}{D(q^{-1})} e_k$$

a úpravou (3.4.20) dostaneme

$$\begin{aligned} z_k &= \frac{C(q^{-1})}{D(q^{-1})})e_k \\ &= (1 + \frac{C(q^{-1}) - D(q^{-1})}{D(q^{-1})})e_k \\ &= (1 + \frac{(c_1-d_1)q^{-1} + \ldots + (c_{nc}-d_{nd})q^{-nc}}{1 + d_1q^{-1} + \ldots + d_{nd}q^{-nd}})e_k \\ &= e_k + Hx_k \end{aligned} \tag{3.4.21}$$

kde $H = [1, 0, \ldots, 0]$ je matice rozměru $1 \times nc$. Pak

$$x_{k+1} = Fx_k + Le_k \tag{3.4.22}$$

kde

$$F = \begin{bmatrix} -d_1 & 1 & \ldots & \ldots & 0 \\ -d_2 & 0 & 1 & \ldots & 0 \\ \vdots & & & & \vdots \\ \ldots & \ldots & \ldots & \ldots & 1 \\ -d_{nd} & \ldots & \ldots & \ldots & 0 \end{bmatrix} \quad L = \begin{bmatrix} c_1 & - & d_1 \\ c_2 & - & d_2 \\ \vdots & & \vdots \\ c_{nc} & - & d_{nd} \end{bmatrix}$$

Tedy, vztahy (3.4.22) a (3.4.21) definují stav a stavový model. Tento model lze upravit opět tak, aby stavový šum a šum měření byly nezávislé. Po úpravě dostaneme

$$x_{k+1} = \bar{F}x_k + Lz_k \tag{3.4.23}$$

$$z_k = Hx_k + e_k \tag{3.4.24}$$

kde

$$\bar{F} = \begin{bmatrix} -c_1 & 1 & \dots & \dots & 0 \\ -c_2 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots \\ \dots & \dots & \dots & \dots & 1 \\ -c_{nc} & \dots & \dots & \dots & 0 \end{bmatrix}$$

Převody z fenomenologického popisu na stavový lze nalézt také v [23], [24].

V této sekci jsme ukázali, že při přechodu z fenomenologické reprezentace na stavovou je spíše problém formální související s volbou báze stavového prostoru. Opačný postup je složitější, protože musíme vypočítat nejdříve inovační reprezentaci, tedy vlastně aplikovat Kalmanův filtr.

3.5 Úloha predikce a vyhlazování

Doposud jsme se zabývali estimačním problémem zaměřeným na úlohu filtrace a jednokrokové predikce. V této sekci se zaměříme na problém vyhlazování a obecný problém predikce pro signály definované v sekci 3.1 rovnicemi (3.1.1), (3.1.2). Nejdříve se budeme zajímat o vyhlazování stavu predikci stavu, pak o vyhlazování stavu.

Začněme tedy predikcí stavu. Nalezení odhadu x_{k+l} využitím z^k a (3.1.1), (3.1.2), kde $l \geq 1$ je problém predikce. Předpokládejme, že systém je t-invariantní a šumy stacionární. Tedy $F_k = F, H_k = H, Q_k = Q, R_k = R$ pro všechna k. Pak ze (3.2.48) vyplývá podmíněná střední hodnota jako optimální bodový odhad

$$E[x_{k+l} \mid z^k] = FE[x_{k+l-1} \mid z^k] \tag{3.5.1}$$

Opakovaným využitím (3.5.1) dostaneme

$$E[x_{k+l} \mid z^k] = F^l E[x_k \mid z^k] \tag{3.5.2}$$

Podmíněnou kovarianci získáme opět lehce využitím (3.2.49).

$$\begin{aligned} E[(x_{k+l} &- E[x_{k+l} \mid z^k])(x_{k+l} - E[x_{k+l} \mid z^k])^T \mid z^k] \\ &= E[(x_{k+l} - \hat{x}'_{k+l|k})(x_{k+l} - \hat{x}'_{k+l|k})^T \mid z^k] \\ &= FE[(x_{k+l-1} - E[x_{k+l-1} \mid z^k])(x_{k+l-1} - E[x_{k+l-1} \mid z^k])^T \mid z^k]F^T + Q \end{aligned} \tag{3.5.3}$$

Pro snazší zápis definujeme pro $l \geq 1$

$$E[x_{k+l} \mid z^k] \stackrel{\triangle}{=} \hat{x}'_{k+l} \tag{3.5.4}$$

$$E[(x_{k+l} - E[x_{k+l} \mid z^k])(x_{k+1} - E[x_{k+l} \mid z^k])^T \mid z^k] \stackrel{\triangle}{=} P'_{k+l} \tag{3.5.5}$$

Opakovaným použitím (3.5.3) a využitím značení (3.5.4),(3.5.5) dostaneme

$$P'_{k+l} = F^l P_k (F^T)^l + \sum_{i=0}^{l-1} F^i Q (F^T)^i \tag{3.5.6}$$

Můžeme konstatovat, že Kalmanův filtr a vztahy (3.5.2), (3.5.6) řeší uvažovaný problém predikce.

Nyní se budeme věnovat úloze vyhlazování. Za stejných předpokladů jako v předchozí úloze predikce chceme nalézt odhad x_0 na základě pozorování z^k, pro $k > 0$ a znalosti (3.1.1), (3.1.2). Okamžik, pro který požadujeme odhad, je fixován, ale počet pozorování k se mění. Pro řešení tohoto problému je zde výhodné rozšířit stavový vektor o počáteční stav a následně zavést rozšířený stavový model

$$\bar{x}_k \stackrel{\triangle}{=} [x_k^T, x_0^T]^T \quad k \geq 0 \tag{3.5.7}$$

$$\bar{x}_{k+1} = \begin{bmatrix} F & 0 \\ 0 & I \end{bmatrix} \bar{x}_k + \begin{bmatrix} w_k \\ 0 \end{bmatrix} \tag{3.5.8}$$

$$z_k = [H \quad 0]\bar{x}_k + v_k \tag{3.5.9}$$

Je zřejmé, že na odhad stavu $\bar{x}_k$ můžeme použít Kalmanův filtr, a tudíž získáme i odhad x_0, který je podle (3.5.7) částí $\bar{x}_k$.
Využijeme vztahu (3.3.2) pro prediktivní kovarianci a s ohledem na (3.5.7)-(3.5.9) dostaneme

$$\begin{bmatrix} P_{k+1}^{11'} & P_{k+1}^{12'} \\ P_{k+1}^{12^{T'}} & P_{k+1}^{22'} \end{bmatrix} = \begin{bmatrix} F & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} P_k^{11'} & P_k^{12'} \\ P_k^{12^{T'}} & P_k^{22'} \end{bmatrix} \begin{bmatrix} F^T & 0 \\ 0 & I \end{bmatrix}$$
$$- \begin{bmatrix} K_k^{1'} \\ K_k^{2'} \end{bmatrix} [H \quad 0] \begin{bmatrix} P_k^{11'} & P_k^{12'} \\ P_k^{12'^T} & P_k^{22'} \end{bmatrix} \begin{bmatrix} F^T & 0 \\ 0 & I \end{bmatrix} + \begin{bmatrix} Q & 0 \\ 0 & 0 \end{bmatrix} \tag{3.5.10}$$

kde

$$\begin{bmatrix} K_k^{1'} \\ K_k^{2'} \end{bmatrix} = \begin{bmatrix} F & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} P_k^{11'} & P_k^{12'} \\ P_k^{12'^T} & P_k^{22'} \end{bmatrix} \begin{bmatrix} H^T \\ 0 \end{bmatrix} [H P_k^{11'} H^T + R]^{-1} \tag{3.5.11}$$

Z (3.5.10), (3.5.11) zjistíme, že $P_{k+1}^{11'}$ je P'_{k+1} ze (3.3.2). Další blok kovarianční matice pak je

$$P_{k+1}^{22'} = P_k^{22'} - K_k^{2'} H P_k^{12'}$$

Dosazením z (3.5.11) za $K_k^{2'}$ dostaneme

$$\begin{aligned} P_{k+1}^{22'} &= P_k^{22'} - P_k^{12'^T} H^T (H P_k^{11'} H^T + R)^{-1} H P_k^{12'} \\ K_k^{2'} &= P_k^{12'^T} H^T (H P_k^{11'} H^T + R)^{-1} \end{aligned} \tag{3.5.12}$$

Zbývá vyjádřit $P_{k+1}^{12'}$

$$P_{k+1}^{12'} = FP_k^{12'} - K_k^{1'} HP_k^{12'}$$

Dosazením z (3.5.11) získáme

$$P_{k+1}^{12'} = F\left(I - P_k^{11'} H^T (HP_k^{11'} H^T + R)^{-1} H\right) P_k^{12'} \tag{3.5.13}$$

Nyní použijeme vztah (3.3.1) s ohledem na (3.5.7)-(3.5.9) a dostaneme

$$\hat{\bar{x}}'_{k+1} = \begin{bmatrix} F & 0 \\ 0 & I \end{bmatrix} \hat{\bar{x}}'_k + \begin{bmatrix} K_k^{1'} \\ K_k^{2'} \end{bmatrix} (z_k - H\hat{x}'_k) \tag{3.5.14}$$

Víme, že

$$\hat{\bar{x}}'_{k+1} = E[\bar{x}_{k+1} \mid z^k]$$

kde

$$\bar{x}_{k+1} = \begin{bmatrix} x_{k+1} \\ x_0 \end{bmatrix}$$

takže jsme vypočetli i $E(x_0 \mid z^k)$. Z (3.5.14) tedy plyne

$$E[x_0 \mid z^k] = E[x_0 \mid z^{k-1}] + K_k^{2'}(z_k - H\hat{x}'_k) \tag{3.5.15}$$

nebo po úpravě

$$E[x_0 \mid z^k] = E[x_0 \mid z^{-1}] + \sum_{i=0}^{k} K_i^{2'}(z_i - H\hat{x}'_i) \tag{3.5.16}$$

Při řešení tohoto problému jsme opět použili Kalmanův filtr. Zajímavé je, že výsledný odhad podle (3.5.16) je vlastně součet apriorního odhadu a váženého součtu inovací.

Poznámka . V podkapitole 3.3.3 byly diskutovány podmínky, kdy řešení Riccatiho rovnice (3.5.10) konverguje k ustálené pozitivně *definitní* matice. Jedna z podmínek říkala, že je nutné, aby stavový šum ovlivňoval všechny složky stavu. Tato podmínka není zcela zřejmě splněna pro rozšířený model (3.5.8), (3.5.9). Jako důsledek nelze očekávat, že výsledná prediktivní kovarianční matice P'_{k+1} (3.5.10), pro $k \to \infty$, bude pozitivně definitní. Místo toho, matice bude P'_{k+1} konvergovat k pozitivně *semi-definitní* matici, kde $P_{k+1}^{12'} = 0$ a $P_{k+1}^{22'} = 0$ pro $k \to \infty$. Ač se takové chování filtru může zdát překvapivé, je zcela správné. Počáteční stav x_0, který je odhadován, je v rozšířeném modelu (3.5.8) modelován jako neznámá konstanta (není zde žádná komponenta stavového šumu). Konstantu je filtr schopen naprosto přesně odhadnout, tj. odhadnout s nulovou kovarianční maticí chyby odhadu $P_{k+1}^{22'} = 0$, za podmínky nekonečně mnoha měření. Poznamenejme, že podobnou „schopnost“ mají i identifikační techniky, jako například metoda nejmenších čtverců, které byly představeny v prvním díle skript. Důvod pro, v limitním případě, nulovou křížovou kovarianční matici $P_{k+1}^{12'}$ lze najít v tom, že pro rostoucí k, klesá množství informace o počátečním stavu, které je obsaženo v měření. Pro $k \to \infty$ je aktuální stav x_k naprosto nezávislý na počátečním stavu x_0, proto jejich vzájemná „korelace“,

vyjádřená $P_k^{12'}$, je nulová. Povšimněme si také, že pro $P_k^{12'} = 0$ je zisk $K_k^{2'}$ umožňující filtrační odhad počátečního stavu x_0 nulový, a dojde tedy k zastavení časového vývoje odhadu počáteční podmínky.

Výše představený přístup k odhadu „minulého“ stavu je vhodný pokud nás zajímá vyhlazený stav v jednom konkrétním okamžiku. Pokud by nás zajímal vyhlazený stav ve vícero časových okamžicích, je uvedený způsob odhadu nevhodný. Proto byla navržena celá řada alternativních přístupů k vyhlazování, kde jako jeden z nejpoužívanějších můžeme uvést tzv. Rauch-Tung-Striebelův vyhlazovač. Ten je vhodný zejména pro ty úlohy, kdy chceme vyhladit stav ve všech časových okamžicích. Je založen na přímém a zpětném chodu, přičemž přímý chod zajišťují rovnice Kalmanova filtru pro filtraci (3.2.44), (3.2.45) a predikci (3.2.47), (3.2.48). Zpětný chod je pak určen následujícími vztahy (pro t-invariantní systém).

Rauch-Tung-Striebelův vyhlazovač

$$P_k^v = P_k - P_k F^T (P'_{k+1})^{-1}(P'_{k+1} - P^v_{k+1})(P'_{k+1})^{-1} F P_k \tag{3.5.17}$$

$$K_k^v = P_k F^T (P'_{k+1})^{-1} \tag{3.5.18}$$

$$\hat{x}_k^v = \hat{x}_k + K_k^v(\hat{x}^v_{k+1} - \hat{x}'_{k+1}) \tag{3.5.19}$$

$$P_k^v = P_k - K_k^v(P'_{k+1} - P^v_{k+1})K_k^{vT} \tag{3.5.20}$$

kde $k = N-1, N-2, ...$ Pro měření $z_0, z_1, ..., z_N$ bude počáteční podmínka $P_N^v = P_N$ a $\hat{x}_N^v = \hat{x}_N$, tj. je dána posledním dostupným filtračním odhadem. Symboly s indexem v jsou vyhlazené veličiny. Tento algoritmus, daný vztahy (3.5.17)–(3.5.20), je výhodnější než řada dalších verzí vyhlazovačů především z algoritmického hlediska. Lze ukázat, že pro lineární gaussovský systém je i vyhlazený odhad gaussovský.

Poznamenejme, že odvození Rauch-Tung-Striebelova vyhlazovače vychází z následujících vztahů pro rekurzivní výpočet vyhlazené hustoty pravděpodobnosti $p(x_k|z^l)$, kde $l > k$. Tedy vyhlazenou hustotu lze zapsat

$$\begin{aligned} p(x_k|z^l) &= \int p(x_k, x_{k+1}|z_0^l)dx_{k+1} \\ &= \int p(x_k, x_{k+1}|z_0^k, z_{k+1}^l)dx_{k+1} \\ &= \int p(x_k|x_{k+1}, z_0^k, z_{k+1}^l)p(x_{k+1}|z_0^l)dx_{k+1} \end{aligned}$$

kde značení $z_k^l = [z_k^T, \ldots, z_l^T]$. Protože stav x_k nezávisí na budoucím měření, můžeme psát

$$\begin{aligned} p(x_k|z^l) &= \int p(x_k|x_{k+1}, z_0^k)p(x_{k+1}|z_0^l)dx_{k+1} \\ &= \int \frac{p(x_k|z_0^k)p(x_{k+1}|x_k, z_0^k)}{p(x_{k+1}|z_0^k)} p(x_{k+1}|z_0^l)dx_{k+1} \\ &= p(x_k|z_0^k) \int \frac{p(x_{k+1}|x_k)}{p(x_{k+1}|z_0^k)} p(x_{k+1}|z_0^l)dx_{k+1} \end{aligned}$$

kde $p(x_k|z_0^k)$ je filtrační a $p(x_{k+1}|z_0^k)$ je prediktivní hustota z Kalmanova filtru, $p(x_{k+1}|x_k)$ je přechodová hustota z rovnice (3.1.1) a $p(x_{k+1}|z_0^l)$ je vyhlazená hustota z předchozího časového okamžiku. Počáteční podmínka této rekurze tak je filtrační hustota $p(x_l|z_0^l)$. Další podrobnosti je možné najít v [84], [88].

Poznámka . Lze očekávat, a vztah pro výpočet prediktivní kovarianční matice (3.5.6) to potvrzuje, že v predikci dojde k růstu kovarianční matice s rostoucím horizontem predikce, tj. variance dvou krokové predikce P'_{k+2} bude větší než jednokroková variance predikce P'_{k+1} a ta bude větší než variance filtračního odhadu P_k. Naopak, vyhlazováním stavu bude docházet k poklesu kovarianční matice, jak je i vidět ze vztahu (3.5.20).

3.6 Kalmanův filtr v úloze odhadu nelineárního a negaussovského systému

Doposud jsme se věnovali odhadu stavu za předpokladu lineárního a Gaussovského systému, pro který jsme našli exaktní řešení úlohy filtrace, predikce i vyhlazování. Pokud však opustíme některý z předpokladů, exaktní řešení úlohy odhadu stavu již není možné. Pro lepší představu o významu předpokladů nám poslouží následující dva příklady. V prvním opustíme předpoklad gaussovosti a ve druhém linearity.

Příklad 3.6.1 Uvažujme skalární rovnici měření v nultém kroku a vynechme časový index

$$z = x + v \tag{3.6.21}$$

kde x a v jsou nezávislé náhodné veličiny s rovnoměrným rozdělením na intervalu (-1,1), tedy

$$p(x) = \frac{1}{2} \qquad x \in (-1,1) \tag{3.6.22}$$

$$= 0 \qquad \text{jinak} \tag{3.6.23}$$

$$p(v) = \frac{1}{2} \qquad v \in (-1,1) \tag{3.6.24}$$

$$= 0 \qquad \text{jinak} \tag{3.6.25}$$

Odsud platí, že

$$E[x] = E[v] = 0 \tag{3.6.26}$$

$$var[x] = var[v] = 1/3 \tag{3.6.27}$$

Cílem je spočítat $p(x \mid z)$ a $E[x \mid z], var[x \mid z]$. Vyjdeme z bayesovských vztahů. Tedy

$$p(x \mid z) = \frac{p(z \mid x)p(x)}{p(z)} = \frac{p(z-x)p(x)}{\int p(z-x)p(x)dx}$$

kde

$$p(z-x) = \frac{1}{2} \qquad z \in (-1+x, 1+x) \tag{3.6.28}$$

$$= 0 \qquad \text{jinak} \tag{3.6.29}$$

Jelikož rozložení součtu dvou náhodných proměnných s rovnoměrným rozložením je tzv. trojúhelníkové rozložení, je $p(z)$ dáno vztahem

$$\begin{aligned} p(z) &= \frac{z+2}{4} \qquad z \in (-2,0\rangle \\ &= \frac{2-z}{4} \qquad z \in (0,2) \\ &= 0 \qquad \text{jinak} \end{aligned}$$

Aby zápis byl kompaktní, využijeme funkce sign.

$$\begin{aligned} p(z \mid x) &= [sign(1+z-x) - sign(-1+z-x)]/4 \\ p(x) &= [sign(x+1) - sign(x-1)]/4 \\ p(z) &= \frac{1}{8}[2 - zsign(z)][sign(z+2) - sign(z-2)] \end{aligned}$$

Nyní již můžeme vyjádřit $p(x \mid z)$, tedy

$$p(x \mid z) = \frac{[sign(1+z-x) - sign(-1+z-x)][sign(x+1) - sign(x-1)]}{2[2 - zsign(z)][sign(z+2) - sign(z-2)]} \tag{3.6.30}$$

Při analýze čitatele (3.6.30) zjistíme, že $p(x \mid z)$ je nenulová nad oblastí ABCD v obr. 3.6.1.

Dále zjistíme, že $p(x \mid z)$ je rovnoměrné rozložení (čitatel konstantní, mění se pouze jmenovatel). Z toho vyplývá, že

$$E[x \mid z] = z/2 \tag{3.6.31}$$

a podmíněná kovariance pro $z \in (0,2)$

$$E[(x - E[x \mid z])^2 \mid z] = E[x^2 \mid z] - z^2/4$$

Protože víme, že hustota je rovnoměrná, snadno spočteme $E[x^2 \mid z]$

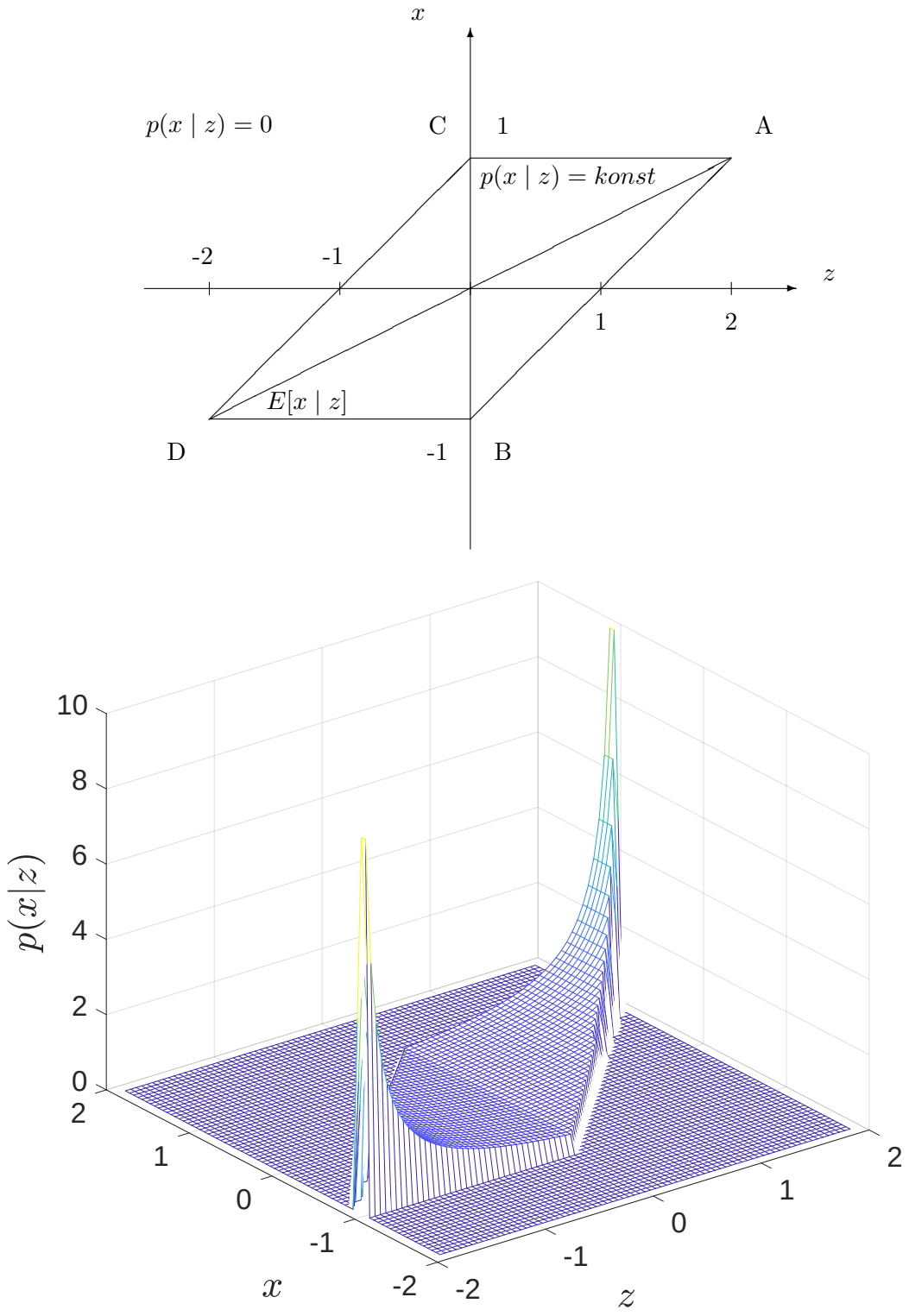
$$E[x^2 \mid z] = \int_{z-1}^{1} x^2 \frac{1}{1-(z-1)} dx = \frac{1}{2-z}[\frac{1}{3} - \frac{(z-1)^3}{3}]$$

Tudíž $cov(x \mid z)$ je určena

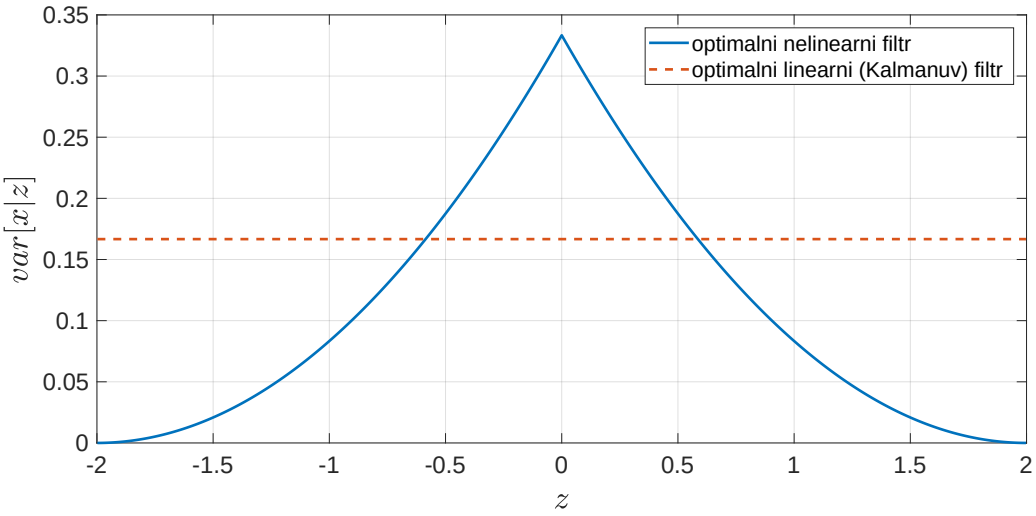
$$var[x \mid z] = \frac{1}{3(2-z)}[1 - (z-1)^3] - \frac{z^2}{4} \tag{3.6.32}$$

Analogicky můžeme vypočítat podmíněnou kovarianci pro $z \in (-2,0)$. Dostaneme

$$\begin{aligned} var[x \mid z] &= E[x^2 \mid z] - z^2/4 \\ &= \int_{-1}^{1+z} x^2 \frac{1}{1+(z+1)} dx - z^2/4 \\ &= \frac{1}{3(z+2)}[1 + (z+1)^3] - \frac{z^2}{4} \end{aligned} \tag{3.6.33}$$



Obrázek 3.6.1: Podmíněná hustota pravděpodobnosti s ilustrací střední hodnoty (v horním obrázku)



Obrázek 3.6.2: Závislost podmíněné variance odhadu na hodnotě měření

Z (3.6.32) a (3.6.33) vyplývá, že podmíněná kovariance je v tomto případě *nelineární* funkcí měření. Dále je vhodné si všimnout, že na rozdíl od lineárního gaussovského systému, nedochází k reprodukovatelnosti hustot, tj. filtrační hustota $p(x|z)$ již není rovnoměrná. Také, pro úplný popis filtrační hustoty již nepostačují první dva momenty.

Pro úplnost se ještě vraťme ke KF a jeho interpretaci jako estimátoru s minimální variancí (viz sekce 3.3.2). Při této interpretaci lze navrhnout (*lineární*) KF pro uvažovaný problém nezávisle na distribuci apriorní informace stavu $p(x)$ a šumu v rovnici měření $p(v)$. Pak lze, vzhledem k lineární rovnici měření (3.6.21), filtrační krok KF (3.2.66), (3.2.67) zapsat

$$E_{\mathrm{KF}}[x|z] = \hat{x} = \hat{x}' + K(z - \hat{x}') \tag{3.6.34}$$

$$var_{\mathrm{KF}}[x|z] = P = P' - \frac{(P')^2}{P'+R} \tag{3.6.35}$$

kde $K = \frac{P'}{P'+R}$, $\hat{x}' = E[x] = 0$, $P' = var[x] = \frac{1}{3}$ a $R = var[v] = \frac{1}{3}$. Dosazením a úpravou vztahu pro výpočet střední hodnoty, lze (3.6.34) přepsat do formy

$$\hat{x} = 0.5z \tag{3.6.36}$$

která je shodná s optimálním řešením dle Bayesových vztahů (3.6.31). Výpočet variance KF však vede na jinou hodnotu, a to na

$$P = \frac{1}{3} - \frac{1}{9}\left(\frac{2}{3}\right) = \frac{1}{6} \tag{3.6.37}$$

která je, na rozdíl od optimální variance (3.6.32), nezávislá na aktuální hodnotě měření.
Pro lepší představu o (ne)závislosti podmíněné variance optimálního *nelineárního* filtru (3.6.32) a nejlepšího *lineárního* filtru (3.6.37) na měření se podívejme na obr. 3.6.2. Na obrázku jsou vykresleny podmíněné variance v závislosti na měření. Lze vidět výraznou závislost variance ma měření v případě optimálního filtru (3.6.32). Tu lze intuitivně vysvětlit pomocí následujících limitních situací:

- Situace 1: Předpokládejme měření $z \to 2$, pak, vzhledem k definicím rovnoměrně distribuovaných hustot (3.6.22)–(3.6.25), stav i šum měření musel nabýt hodnoty asymptoticky

blízké 1. Podmíněný odhad stavu pak musí nabývat hodnoty $E[x|z] = z/2 \Longrightarrow 1$ s „velkou“ jistotou, tj. s variancí $var[x|z] \Longrightarrow 0$.

- Situace 2: Předpokládejme měření $z = 0$, pak stav mohl nabýt libovolné hodnoty z intervalu $x \in (0, 1)$ a šum opačné hodnoty z intervalu $v \in (-1, 0)$, popř. obráceně. V tomto případě tak měření nepřináší žádnou novou informaci a podmíněný odhad $E[x|z] = 0$ má varianci chyby odhadu $var[x|z] = 1/3$, což je variance P' apriorní hustoty pravděpodobnosti $p(x)$.

Tedy, čím vyšší hodnota měření z je, tím vyšší informaci přináší. Tato skutečnost, však, není vzata v potaz lineárním, tedy Kalmanovo, filtrem.
Za zmínku rovněž stojí fakt, že střední kvadratická chyba odhadu a příslušná (střední, nebo-li průměrná) variance přes *všechna* měření je pro nelineární filtr

$$MSE = E[(x - E[x|z])^2] = 1/6, \tag{3.6.38}$$
$$AVAR = E[var[x|z]] = 1/6 \tag{3.6.39}$$

shodná a je rovněž shodná i se statistikami lineárního KF

$$MSE_{\mathrm{KF}} = E[(x - \hat{x})^2] = 1/6, \tag{3.6.40}$$
$$AVAR_{\mathrm{KF}} = E[P] = P = 1/6. \tag{3.6.41}$$

To značí že oba filtry poskytují konzistentní odhad (viz sekce 3.3.4) s minimální *nepodmíněnou* variancí AVAR. Právě minimalizace nepodmíněné variance AVAR je základem odvození KF optimalizačním přístupem, tj. minimalizací kriteriální funkce $V(\hat{x}_k)$ definované v sekci 3.3.2.

Příklad 3.6.2 Uvažujme skalární nelineární stavovou rovnici a rovnici měření

$$\begin{aligned} x_{k+1} &= f_k x_k + g_k x_k^2 + w_k \\ z_k &= h_k(x_k) + v_k \end{aligned} \tag{3.6.42}$$

kde $\{w_k\}$ je bílý gaussovský proces s nulovou střední hodnotou a variancí Q_k. Předpokládejme, že

$$\begin{aligned} E[x_k \mid z^k] &= \hat{x}_k \\ E[(x_k - \hat{x}_k)^2 \mid z^k] &= P_k \end{aligned}$$

Cílem je vypočítat

$$E[x_{k+1} \mid z^k] = \hat{x}'_{k+1} \quad a \quad var[x_{k+1} \mid z^k] = P'_{k+1}$$

Začneme střední hodnotou

$$\begin{aligned} \hat{x}'_{k+1} &= f_k E[x_k \mid z^k] + g_k E[x_k^2 \mid z^k] \\ &= f_k \hat{x}_k + g_k (P_k + \hat{x}_k^2) \end{aligned} \tag{3.6.43}$$

Můžeme konstatovat, že pro výpočet predikované střední hodnoty potřebujeme filtrační střední hodnotu a varianci, tedy první dva momenty.

Nyní vypočítáme $var[x_{k+1} \mid z^k]$. Nechť

$$\begin{aligned} \tilde{x}'_{k+1} &\stackrel{\triangle}{=} x_{k+1} - \hat{x}'_{k+1} \\ \tilde{x}_k &\stackrel{\triangle}{=} x_k - \hat{x}_k \end{aligned}$$

Pak, využitím vztahu $x_k^2 - \hat{x}_k^2 = (x_k - \hat{x}_k)^2 + 2x_k\hat{x}_k$ získáme

$$\begin{aligned} \tilde{x}'_{k+1} &= f_k\tilde{x}_k + w_k - g_kP_k + g_k(x_k^2 - \hat{x}_k^2) \\ &= (f_k + 2g_k\hat{x}_k)\tilde{x}_k + g_k\tilde{x}_k^2 - g_kP_k + w_k \end{aligned}$$

$$\begin{aligned} E[\tilde{x}_{k+1}^{'2} \mid z^k] &= (f_k + 2g_k\hat{x}_k)^2 E[\tilde{x}_k^2 \mid z^k] + g_k^2 E[\tilde{x}_k^4 \mid z^k] \\ &+ g_k^2P_k^2 + E[w_k^2 \mid z^k] + 2g_k(f_k + 2g_k\hat{x}_k)E[\tilde{x}_k^3 \mid z^k] \\ &- 2g_kP_k(f_k + 2g_k\hat{x}_k)E[\tilde{x}_k \mid z^k] \\ &- 2g_k^2P_kE[\tilde{x}_k^2 \mid z^k] = (f_k + 2g_k\hat{x}_k)^2P_k + g_k^2\gamma_k \\ &- g_k^2P_k + Q_k + 2g_k(f_k + 2g_k\hat{x}_k)\delta_k \end{aligned} \tag{3.6.44}$$

kde $\gamma_k \stackrel{\triangle}{=} E[\tilde{x}_k^4 \mid z^k]$ a $\delta_k \stackrel{\triangle}{=} E[\tilde{x}_k^3 \mid z^k]$.

Podmíněná variance $var[x_{k+1} \mid z^k]$ je podle (3.6.44) funkcí prvních čtyř „filtračních“ momentů. Dochází tedy k jevu, který bývá označován jako zacyklenost momentů. Povšimněme si také, že i při uvažování gaussovské filtrační hustoty, prediktivní hustota již není gaussovská, tj. ani v tomto případě nedochází k reprodukovatelnosti hustot.

Předchozí dva příklady podtrhly jedinečnost exaktního řešení úlohy odhadu stavu lineárního Gaussovského systému a naznačily, že pro nelineární či negaussovský systém bude řešení bayesovských vztahů složitější a většinou se neobejde bez aproximací. V následujících kapitolách se zaměříme na úlohu odhadu stavu nelineárních a negaussovských systémů. Nejprve se budeme věnovat koncepčně jednodušším technikám, které staví a rozšiřují aplikovatelnost Kalmanova filtru, avšak poskytují omezenou kvalitu odhadu. Tyto techniky budeme souhrnně nazývat lokálními filtry. Pak upřeme pozornost na globální filtry, které poskytují kvalitativně lepší odhady, však za cenu výrazně vyšší teoretické i výpočetní složitosti.

Kapitola 4

Lokální filtry

V předchozí kapitole věnované kalmanovské filtraci, nebo-li odhadu stavu lineárního gaussovského systému, jsme se setkali s lineární filtrací. Z rovnic Kalmanova filtru snadno nahlédneme, že odhad stavu je zde získán lineární transformací měření, aniž bychom prováděli jakoukoliv modifikaci základní úlohy, nebo aproximaci řešení. Tato elegantně formulovaná a elegantně řešená úloha však z hlediska širokého uplatnění naráží na bariéru svých předpokladů, a to linearity a gaussovosti. V této kapitole se budeme věnovat situaci, kdy předpoklad linearity je opuštěn, ale základní myšlenkový postup pro syntézu filtru bude shodný s předchozí kapitolou, tedy použijeme opět bayesovský přístup. Problematikou lokálních filtrů se zabývají např. [4], [5], [14], [35], [36], [90]-[95]. Souvislost rozšířeného Kalmanova filtru s parametrickými metodami identifikace, zejména s metodou chyby predikce je ukázána v [28].

4.1 Rozšířený Kalmanův filtr

Jak již bylo řečeno, v této kapitole opustíme rámec linearity. Uvažujme tedy systém, který můžeme chápat jako zobecnění (3.1.1), (3.1.2), definovaný následujícími vztahy

$$x_{k+1} = f_k(x_k) + w_k \tag{4.1.1}$$

$$z_k = h_k(x_k) + v_k \tag{4.1.2}$$

kde $f_k(\cdot), h_k(\cdot)$ jsou známé nelineární diferencovatelné vektorové funkce příslušných dimenzí, bílé šumy $\{w_k\}, \{v_k\}$ a počáteční stav mají stejné vlastnosti jako v předchozí kapitole a jsou popsány

$$p(x_0) = N(x_0 : \hat{x}'_0, P_0) \tag{4.1.3}$$

$$p(w_k) = N(w_k : 0, Q_k) \tag{4.1.4}$$

$$p(v_k) = N(v_k : 0, R_k) \tag{4.1.5}$$

pro všechna k. Filtr navrhneme opět dvěma přístupy, přímým a nepřímým, které byly představeny při návrhu Kalmanova filtru.

4.1.1 Přímý přístup

Z rovnice (4.1.1) můžeme snadno zjistit, že přechodová hustota pravděpodobnosti má tvar

$$p(x_{k+1} \mid x_k) = N(x_{k+1} : f_k(x_k), Q_k) \tag{4.1.6}$$

a z rovnice (4.1.2) určíme hustotu pravděpodobnosti měření

$$p(z_k \mid x_k) = N(z_k : h_k(x_k), R_k) \tag{4.1.7}$$

Nyní máme vše připraveno pro aplikaci bayesovského přístupu k výpočtu podmíněných filtračních a prediktivních hustot. Při odvození začneme opět od počátečního časového okamžiku jako v předchozí kapitole a budeme opět dosazovat do bayesovských rekurzivních vztahů. Otázka zní, jak se změní $p(x_0)$ na základě informace z měření z_0. Tedy

$$p(x_0 \mid z^0) = \frac{N(x_0 : \hat{x}'_0, P'_0)N(z_0 : h_0(x_0), R_0)}{\int N(x_0 : \hat{x}'_0, P'_0)N(z_0 : h_0(x_0), R_0)dx_0} \tag{4.1.8}$$

Výrazný rozdíl mezi (4.1.8) a (3.2.4) z pohledu explicitního nalezení filtrační hustoty je přítomnost nelinearity $h_0(x_0)$ v (4.1.8), která znemožňuje použít analogický postup jako v podkapitole 3.2 vedoucí na exaktní řešení. Odsud je zřejmé, že pro explicitní výpočet nebo jinými slovy pro zachování analytického řešení (4.1.8) musíme provést jistý zásah do formulace úlohy. Nabízí se jednoduché řešení, linearizovat funkci $h_0(\cdot)$ a přejít formálně na shodnou úlohu jako v podkapitole 3.2. Otázkou zůstává, v jakém bodě linearizaci provést. Protože však předpokládáme gaussovské rozložení počátečního stavu, pak za nejlepší odhad počátečního stavu můžeme jistě uvažovat hodnotu $\hat{x}'_0$.
Uvažujme tedy Taylorův rozvoj $h_0(x_0)$ v okolí bodu $\hat{x}'_0 = E[x_0 \mid z^{-1}]$ a ponechme pouze první dva členy, jelikož chceme lineární aproximaci. Dostaneme

$$h_0(x_0) = [h_{1,0}(x_0), h_{2,0}(x_0), \ldots, h_{nz,0}(x_0)]^T \simeq h_0(\hat{x}'_0) + H_0(\hat{x}'_0)(x_0 - \hat{x}'_0) \tag{4.1.9}$$

kde

$$H_0(\hat{x}'_0) = \frac{\partial h_0(x_0)}{\partial x_0}|_{x_0=\hat{x}'_0} = \begin{bmatrix} \frac{\partial h_{1,k}(x_0)}{\partial x_1} & \frac{\partial h_{1,k}(x_0)}{\partial x_2} & \cdots & \frac{\partial h_{1,k}(x_0)}{\partial x_n} \\ \vdots & \vdots & \vdots & \\ \frac{\partial h_{nz,k}(x_0)}{\partial x_1} & \cdots & \cdots & \frac{\partial h_{nz,k}(x_0)}{\partial x_n} \end{bmatrix} |_{x_0=\hat{x}'_0}$$

je matice nz/nx prvních derivací, tzv. Jacobián, vyhodnocený v bodě aktuálně nejlepšího dostupného odhadu stavu $\hat{x}'_0$. Je zřejmé, že pravou stranu (4.1.9) můžeme zapsat i takto

$$H_0(\hat{x}'_0)x_0 + h_0(\hat{x}'_0) - H_0(\hat{x}'_0)\hat{x}'_0$$

kde první člen je lineární funkce x_0 a další dva členy jsou známé veličiny v $k = -1$, a tudíž nemohou přinášet žádné komplikace. Dosaďme proto (4.1.9) do (4.1.8), pak dostaneme

$$p_A(x_0 \mid z^0) = \frac{N(x_0 : \hat{x}'_0, P'_0)N(z_0 : h_0(\hat{x}'_0) + H_0(\hat{x}'_0)(x_0 - \hat{x}'_0), R_0)}{\int N(x_0 : \hat{x}'_0, P'_0)N(z_0 : h_0(\hat{x}'_0) + H_0(\hat{x}'_0)(x_0 - \hat{x}'_0), R_0)dx_0} \tag{4.1.10}$$

Označení $p_A(\cdot)$ v (4.1.10) označuje, že na rozdíl od (4.1.8) tato hustota již bude pouze aproximovat skutečnou hustotu $p(x_0 \mid z^0)$, kterou však explicitně neznáme.
Nyní můžeme $p_A(x_0 \mid z^0)$ vypočítat explicitně, analogicky jako v podkapitole 3.2. Odsud dostaneme aproximační filtrační hustotu pravděpodobnosti mající gaussovské rozložení tvaru

$$p_A(x_0 \mid z^0) = N(x_0 : \hat{x}_0, P_0) \tag{4.1.11}$$
$$\hat{x}_0 = \hat{x}'_0 + P'_0 H_0^T(\hat{x}'_0)[H_0(\hat{x}'_0)P'_0 H_0^T(\hat{x}'_0) + R_0]^{-1}[z_0 - h_0(\hat{x}'_0)]$$
$$P_0 = P'_0 - P'_0 H_0^T(\hat{x}'_0)[H_0(\hat{x}'_0)P'_0 H_0^T(\hat{x}'_0) + R_0]^{-1}H_0(\hat{x}'_0)P'_0$$

Dalším krokem syntézy filtru je nalezení prediktivní hustoty pravděpodobnosti $p(x_1 \mid z^0)$. Víme, že

$$p(x_1 \mid z^0) = \int p(x_0 \mid z^0) p(x_1 \mid x_0) dx_0 \tag{4.1.12}$$

Přechodovou hustotu známe a aproximaci filtrační hustoty též, tudíž dosaďme do (4.1.12) z (4.1.11) a (4.1.6). Dostaneme

$$p_A(x_1 \mid z^0) = \int N(x_0 : \hat{x}_0, P_0) N(x_1 : f_0(x_0), Q_0) dx_0 \tag{4.1.13}$$

Opět jsme použili označení $p_A(x_1 \mid z^0)$, protože se jedná o aproximaci neznámé $p(x_1 \mid z^0)$ (z důvodu použití aproximativní filtrační hustoty $p_A(x_1 \mid z^0)$). Výpočet integrálu na pravé straně analogickým způsobem jako v podkapitole 3.2 komplikuje nelineární funkce $f_0(x_0)$. Linearizací této funkce, obdobně jako v kroku filtrace, pak můžeme získat stejným postupem jako v podkapitole 3.2 explicitní analytické řešení. Proveďme Taylorův rozvoj funkce $f_0(x_0)$ na okolí $\hat{x}_0$ a ponechme opět jako v kroku filtrace první dva členy. Dostaneme

$$f_0(x_0) \simeq f_0(\hat{x}_0) + F_0(\hat{x}_0)(x_0 - \hat{x}_0) \tag{4.1.14}$$

Poznamenejme, že nyní jsme zvolili jako linearizační bod filtrační odhad $\hat{x}_0$, protože obsahuje již více informací o x_0 než $\hat{x}'_0$ (viz 4.1.11). Pro úplnost dodejme, že

$$F_0(\hat{x}_0) = \frac{\partial f(x_0)}{\partial x_0}|_{x_0 = \hat{x}_0}$$

je matice nx/nx prvních derivací. Dosazením (4.1.14) do (4.1.13) dostaneme

$$p_A(x_1 \mid z^0) = \int N(x_0 : \hat{x}_0, P_0) N(x_1 : f_0(\hat{x}_0) + F_0(\hat{x}_0)(x_0 - \hat{x}_0), Q_0) dx_0 \tag{4.1.15}$$

Výraz $f_0(\hat{x}_0) - F_0(\hat{x}_0)\hat{x}_0$ je v okamžiku $k = 0$ známý, tudíž z hlediska výpočtu integrálu jsme opět přešli na případ řešený v podkapitole 3.2 začínající vztahem (3.2.29). Analogickým výpočtem jako v této sekci dostaneme aproximační prediktivní hustotu pravděpodobnosti ve tvaru

$$\begin{aligned} p_A(x_1 \mid z^0) &= N(x_1 : \hat{x}'_1, P'_1) \\ \hat{x}'_1 &= f_0(\hat{x}_0) \\ P'_1 &= F_0(\hat{x}_0) P_0 F_0^T(\hat{x}_0) + Q_0 \end{aligned} \tag{4.1.16}$$

Všimněme si, že prediktivní krok, tj. výpočet rovnice (4.1.12) pomocí (4.1.15), je založen na dvou po sobě jdoucích aproximacích. První je dána použitím aproximované hustoty filtrační hustoty $p_A(x_1 \mid z^0)$ vedoucí na (4.1.13). Druhá aproximace plyne z použití linearizované funkce (4.1.14), a tím i aproximativní přechodové hustoty vedoucí na finální vztah (4.1.15).

Doposud jsme provedli výpočet filtrační hustoty pro $k = 0$ a prediktivní hustoty pro $k = 1$. Stejně můžeme pokračovat pro libovolné k, to jest můžeme zapsat obecné vztahy pro $p_A(x_k \mid z^k)$ a $p_A(x_{k+1} \mid z^k)$. Shrňme tedy výsledky:

Aproximační filtrační hustota pravděpodobnosti:

$$p_A(x_k \mid z^k) = N(x_k : \hat{x}_k, P_k) \tag{4.1.17}$$

$$\hat{x}_k = \hat{x}'_k + P'_k H_k^T(\hat{x}'_k)[H_k(\hat{x}'_k)P'_k H_k^T(\hat{x}'_k) + R_k]^{-1}[z_k - h_k(\hat{x}'_k)]$$
$$P_k = P'_k - P'_k H_k^T(\hat{x}'_k)[H_k(\hat{x}'_k)P'_k H_k^T(\hat{x}'_k) + R_k]^{-1} H_k(\hat{x}'_k)P'_k$$

Aproximační prediktivní hustota pravděpodobnosti:

$$\begin{aligned} p_A(x_{k+1} \mid z^k) &= N(x_{k+1} : \hat{x}'_{k+1}, P'_{k+1}) \\ \hat{x}'_{k+1} &= f_k(\hat{x}_k) \\ P'_{k+1} &= F_k(\hat{x}_k)P_k F_k^T(\hat{x}_k) + Q_k \end{aligned} \tag{4.1.18}$$

4.1.2 Nepřímý přístup

Nepřímý přístup vychází z algoritmu Kalmanova filtru (3.2.66), (3.2.67), (3.2.71) a (3.2.73), kde prediktivní momenty měření a stavu jsou vypočteny při uvažovaní linearizovaných funkcí v rovnici měření, tj. (4.1.9), a dynamiky, tj. (4.1.14).
Tedy, aproximativní filtrační hustota stavu v časovém okamžiku k je dána $p_A(x_k|z^k) = N(x_k : \hat{x}_k, P_k)$ s momenty

$$\hat{x}_k = \hat{x}'_k + P'_{xz,k}(P'_{z,k})^{-1}(z_k - \hat{z}'_k) \tag{4.1.19}$$
$$P_k = P'_k - P'_{xz,k}(P'_{z,k})^{-1}(P'_{xz,k})^T \tag{4.1.20}$$

kde momenty predikce měření, s uvažováním linearizace nelineární funkce v rovnici měření (4.1.9), jsou

$$\begin{aligned} \hat{z}'_k &= E[z_k|z^{k-1}] \\ &= E[h_k(x_k)|z^{k-1}] + E[v_k|z^{k-1}] \\ &= E[h_k(x_k)|z^{k-1}] + E[v_k] \\ &\approx E[h_k(\hat{x}'_k) + H_k(\hat{x}'_k)(x_k - \hat{x}'_k)|z^{k-1}] \\ &= h_k(\hat{x}'_k) \end{aligned} \tag{4.1.21}$$

$$\begin{aligned} P'_{z,k} &= cov[z_k|z^{k-1}] = E[(z_k - E[z_k|z^{k-1}])(z_k - E[z_k|z^{k-1}])^T|z^{k-1}] \\ &= E[(h_k(x_k) + v_k - E[z_k|z^{k-1}])(h_k(x_k) + v_k - E[z_k|z^{k-1}])^T|z^{k-1}] \\ &\approx E[(h_k(\hat{x}'_k) + H_k(\hat{x}'_k)(x_k - \hat{x}'_k) + v_k - h_k(\hat{x}'_k))(h_k(\hat{x}'_k) + H_k(\hat{x}'_k)(x_k - \hat{x}'_k) + v_k - h_k(\hat{x}'_k))^T|z^{k-1}] \\ &= E[(H_k(\hat{x}'_k)(x_k - \hat{x}'_k))(H_k(\hat{x}'_k)(x_k - \hat{x}'_k))^T|z^{k-1}] + E[v_k v_k^T] \\ &= H_k(\hat{x}'_k)P'_k(H_k(\hat{x}'_k))^T + R_k \end{aligned} \tag{4.1.22}$$

$$\begin{aligned} P'_{xz,k} &= cov[x_k, z_k|z^{k-1}] = E[(x_k - \hat{x}'_k)(z_k - E[z_k|z^{-1}])^T|z^{k-1}] \\ &\approx E[(x_k - \hat{x}'_k)(H_k(\hat{x}'_k)(x_k - \hat{x}'_k))^T|z^{k-1}] \\ &= P'_k(H_k(\hat{x}'_k))^T \end{aligned} \tag{4.1.23}$$

Poznamenejme, že díky použitým předpokladům a aproximacím, konkrétně

- předpokladem sdruženě gaussovské prediktivní hustoty $p(x_k, z_k|z^{k-1})$, umožňující odvození vztahů pro lineární filtr (4.1.19), (4.1.20), který není obecně platný při uvažování nelineárního popisu systému, a
- použitím linearizace Taylorovo rozvojem prvního řádu při výpočtu prediktivních momentů měření (4.1.21)–(4.1.23)

je výsledná filtrační hustota pouze gaussovskou aproximací skutečné, obecně negaussovské, podmíněné hustoty.
Prediktivní hustota v čase $k+1$ je aproximována gaussovskou hustotou $p_A(x_{k+1}|z^k) = N(x_{k+1} : \hat{x}'_{k+1}, P'_{k+1})$ s aproximativními prediktivními momenty stavu

$$\begin{aligned}
\hat{x}'_{k+1} &= E[x_{k+1}|z^k] \\
&= E[f_k(x_k) + w_k|z^k] \\
&= E[f_k(x_k)|z^k] \\
&\approx E[f_k(\hat{x}_k) + F_k(\hat{x}_k)(x_k - \hat{x}_k)|z^k] \\
&= f_k(\hat{x}_k) \qquad (4.1.24) \\
P'_{k+1} &= cov[x_{k+1}|z^k] \\
&\approx E[(F_k(\hat{x}_k)(x_k - \hat{x}_k))(F_k(\hat{x}_k)(x_k - \hat{x}_k))^T|z^k] + E[w_k w_k^T] \\
&= F_k(\hat{x}_k)P_k(F_k(\hat{x}_k))^T + Q_k \qquad (4.1.25)
\end{aligned}$$

4.1.3 Vlastnosti filtru a odhadu

Filtr, který byl právě odvozen, definovaný rovnicemi (4.1.17) a (4.1.18) je nazýván rozšířený Kalmanův filtr (v anglicky psané literatuře se setkáme s pojmem „extended Kalman filter“ (EKF)). Jaké jsou hlavní odlišnosti tohoto filtru od Kalmanova filtru.

1. Kovarianční matice P_k, P'_{k+1} není možné počítat off-line, protože jsou funkcemi měření na rozdíl od kovariančních matic generovaných Kalmanovým filtrem, které nejsou na měření závislé. Předpočítávání tedy není možné.

2. Filtrační a prediktivní hustoty jsou pouze aproximace skutečných hustot daných exaktním řešením problému na rozdíl od hustot generovaných Kalmanovým filtrem, které představují exaktní řešení problému. Jelikož zvolená aproximace je prováděna v jednom bodě stavového prostoru, mají výsledky tohoto filtru pouze lokální platnost. Proto tento filtr můžeme chápat jako lokální. Zároveň poznamenejme, že v tomto případě není, z důvodu aproximací, zaručena konvergence odhadu stavu, tak jak byla diskutována pro Kalmanův filtr.

Rozšířený Kalmanův filtr se často využívá v technických aplikacích, např. v úloze navigace a sledování [86], [87], ale též spolu s bayesovským přístupem v ekonometrii, např. [37], [38]. Poznamenejme, že v případě výskytu neznámých parametrů v (4.1.1), (4.1.2), můžeme rozšířit stavový vektor o tyto parametry a pak odhadovat rozšířený stav pomocí rozšířeného Kalmanova filtru.

Poznámka . Jak jsme v této kapitole viděli, nepřímý přístup k návrhu rozšířeného Kalmanova filtru vede opět ke stejným vztahům jako přímý přístup, avšak umožňuje snazší odvození. Proto v následujících kapitolách bude uvažován pouze nepřímý způsob odvození ostatních lokálních filtrů.

4.2 Filtr druhého řádu

Linearizace nelineárních funkcí Taylorovo rozvojem prvního řádu může být v mnoha situacích nedostatečně přesná a ve svém důsledku může vést k divergenci odhadu filtru. Proto byl navržen

filtr druhého řádu, v anglicky psané literatuře označovaný „second-order (extended) Kalman filter“, kde nelineární funkce jsou aproximovány Taylorovou řadou ovšem s ponecháním i členů druhého řádu.
Uvažujme tedy následující aproximaci nelineární funkce $h_k(x_k)$ Taylorovo rozvojem druhého řádu

$$h_k(x_k) \simeq h_k(\hat{x}'_k) + H_k(\hat{x}'_k)(x_k - \hat{x}'_k) + \frac{1}{2}\bar{h}_k \tag{4.2.1}$$

kde $\hat{x}'_k, h_k(\cdot), H_k(\cdot)$ jsou shodné s podkapitolou 4.1 a $\bar{h}_k$ je vektor kvadratických členů závisejících na Hessiánu, tj. na druhých derivacích nelineární funkce v rovnici měření $h_k(\cdot)$. Vektor kvadratických členů je definován pro

$$h_k(\cdot) \stackrel{\triangle}{=} \begin{bmatrix} h_{1,k}(\cdot) \\ h_{2,k}(\cdot) \\ \vdots \\ h_{nz,k}(\cdot) \end{bmatrix}$$

následujícím způsobem

$$M_{i,k} \stackrel{\triangle}{=} \frac{\partial^2 h_{i,k}}{\partial x_k^2} \mid \hat{x}'_k \; = \begin{bmatrix} \frac{\partial^2 h_{i,k}}{\partial x_1 \partial x_1} & \frac{\partial^2 h_{i,k}}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 h_{i,k}}{\partial x_1 \partial x_{nx}} \\ \frac{\partial^2 h_{i,k}}{\partial x_2 \partial x_1} & \frac{\partial^2 h_{i,k}}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 h_{i,k}}{\partial x_2 \partial x_{nx}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 h_{i,k}}{\partial x_{nx} \partial x_1} & \cdots & \cdots & \frac{\partial^2 h_{i,k}}{\partial x_{nx} \partial x_{nx}} \end{bmatrix}, \quad i = 1, 2, ..., nz \tag{4.2.2}$$

$$\bar{h}_{i,k} \stackrel{\triangle}{=} (x_k - \hat{x}'_k)^T M_{i,k} (x_k - \hat{x}'_k) \tag{4.2.3}$$

$$\bar{h}_k \stackrel{\triangle}{=} \begin{bmatrix} \bar{h}_{1,k} \\ \bar{h}_{2,k} \\ \vdots \\ \bar{h}_{nz,k} \end{bmatrix}$$

Pak můžeme aproximaci $h_k(x_k)$ zapsat následujícím způsobem

$$h_k(x_k) \simeq h_k(\hat{x}'_k) + H_k(\hat{x}'_k)(x_k - \hat{x}'_k) + \frac{1}{2}\bar{h}_k \tag{4.2.4}$$

Protože v (4.2.3) se vyskytují členy druhého řádu, analytický výpočet filtrační hustoty by nebyl možný. Když však nahradíme kvadratickou formu v (4.2.3) její očekávanou hodnotou, dostaneme

$$\bar{h}_{i,k} = (x_k - \hat{x}'_k)^T M_{i,k} (x_k - \hat{x}'_k) \simeq tr(P'_k M_{i,k}) \stackrel{\triangle}{=} \bar{h}_{a,i,k} \tag{4.2.5}$$

Označme nyní

$$\bar{h}_{a,k} = [\bar{h}_{a,1,k}, \bar{h}_{a,2,k}, ..., \bar{h}_{a,nz,k}]^T \tag{4.2.6}$$

a (4.2.4) lze zapsat

$$h_k(x_k) \simeq h_k(\hat{x}'_k) + H_k(\hat{x}'_k)(x_k - \hat{x}'_k) + \frac{1}{2}\bar{h}_{a,k} \tag{4.2.7}$$

Z (4.2.7) je vidět, že se jedná o lineární funkci x_k, protože všechny další členy jsou v kroku $k-1$ známy. Můžeme tedy použít opět analogicky jako v kapitole třetí stejný způsob odvození

filtrační hustoty pravděpodobnosti a pro časový okamžik k dostaneme

$$p_A(x_k \mid z^k) = N(x_k : \hat{x}_k, P_k) \tag{4.2.8}$$

$$\hat{x}_k = \hat{x}'_k + P'_k H_k^T(\hat{x}'_k)[H_k(\hat{x}'_k)P'_k H_k^T(\hat{x}'_k) + R_k]^{-1}[z_k - h_k(\hat{x}'_k) - \frac{1}{2}\bar{h}_{ak}] \tag{4.2.9}$$

$$P_k = P'_k - P'_k H_k^T(\hat{x}'_k)[H_k(\hat{x}'_k)P'_k H_k^T(\hat{x}'_k) + R_k]^{-1} H(\hat{x}'_k)P'_k \tag{4.2.10}$$

Vidíme, že oproti rozšířenému Kalmanovu filtru dochází ke změně v inovační posloupnosti v rovnici (4.2.9).

Pro výpočet predikce pokračujme v duchu předchozí části této podkapitoly. Vyjádřeme Taylorův rozvoj funkce $f_k(\cdot)$

$$f_k(x_k) \simeq f_k(\hat{x}_k) + F_k(\hat{x}_k)(x_k - \hat{x}_k) + \frac{1}{2}\bar{f}_k. \tag{4.2.11}$$

kde

$$\bar{f}_k \triangleq \begin{bmatrix} \bar{f}_{1,k} \\ \bar{f}_{2,k} \\ \vdots \\ \bar{f}_{nx,k} \end{bmatrix}$$

$$\bar{f}_{i,k} \triangleq (x_k - \hat{x}_k)^T N_{i,k}(x_k - \hat{x}_k) \tag{4.2.12}$$

$$N_{i,k} \triangleq \frac{\partial^2 f_{i,k}}{\partial x_k^2} \mid \hat{x}_k \; = \begin{bmatrix} \frac{\partial^2 f_{i,k}}{\partial x_1 \partial x_1} & \frac{\partial^2 f_{i,k}}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f_{i,k}}{\partial x_1 \partial x_{nx}} \\ \frac{\partial^2 f_{i,k}}{\partial x_2 \partial x_1} & \frac{\partial^2 f_{i,k}}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f_{i,k}}{\partial x_2 \partial x_{nx}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f_{i,k}}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 f_{i,k}}{\partial x_{nx} \partial x_{nx}} \end{bmatrix}, \quad i = 1, 2, ..., nx \tag{4.2.13}$$

Odstraněním nelinearit ustředněním (4.2.12) dostaneme

$$\bar{f}_{a,i,k} = tr(P_k N_{i,k}) \tag{4.2.14}$$

$$\bar{f}_{a,k} = [\bar{f}_{a,1,k}, \bar{f}_{a,2,k}, ..., \bar{f}_{a,nx,k}]^T \tag{4.2.15}$$

Nyní můžeme přepsat (4.2.11) na

$$f_k(x_k) \simeq f_k(\hat{x}_k) + F_k(\hat{x}_k)(x_k - \hat{x}_k) + \frac{1}{2}\bar{f}_{a,k} \tag{4.2.16}$$

To je opět lineární funkce x_k, protože všechny ostatní veličiny jsou v kroku k známé. Tudíž můžeme dosadit do bayesovských vztahů a analogicky jako ve třetí kapitole odvodit prediktivní hustotu pravděpodobnosti.

$$p_A(x_{k+1} \mid z^k) = N(x_{k+1} : \hat{x}'_{k+1}, P'_{k+1}) \tag{4.2.17}$$

$$\hat{x}'_{k+1} = f_k(\hat{x}_k) + \frac{1}{2}\bar{f}_{a,k} \tag{4.2.18}$$

$$P'_{k+1} = F_k(\hat{x}_k)P_kF_k^T(\hat{x}_k) + Q_k \tag{4.2.19}$$

Rovnice pro kovarianční matici je formálně shodná s rozšířeným Kalmanovo filtrem a střední hodnota je modifikována o člen závisející na druhých derivacích $\bar{f}_{ak}$. Připomeňme, že se jedná opět o lokální filtr a získané hustoty pravděpodobnosti ve formě gaussovských rozložení jsou pouze aproximace skutečných explicitně neznámých hustot pravděpodobnosti. Poznamenejme, že v literatuře lze najít i jiné způsoby odvození vedoucí na algoritmus, kde i kovarianční matice závisí na Hessiánu nelineárních funkcí, např. [88].

4.3 Diferenční filtr prvního řádu

Doposud představené lokální filtry byly navrženy v 60. či 70. letech 20. století. Návrh filtrů byl založen na Taylorově rozvoji nelineárních funkcí v popisu systému, tedy na nutnosti výpočtu prvních, popř. druhých derivací nelineárních funkcí. Pro některé funkce však derivace nemusí být definována či její výpočet může být zdlouhavý (např. pro mnohorozměrné modely). V tom lze spatřit hlavní motivaci návrhu, tzv. bezderivačních lokálních filtrů, které jsou od začátku 21. století v literatuře intenzivně rozvíjeny. Základní idee bezderivačních filtrů budou představeny v této a v následujících kapitolách.

Začněme diferenčním filtrem prvního řádu [90], [93]. Tento filtr, v anglicky psané literatuře označovaný jako „divided diference filter“, je alternativou k rozšířenému Kalmanovu filtru, kdy namísto Taylorova rozvoje je použita tzv. Stirlingova interpolace k linearizaci nelineárních funkcí. Stirlingovu interpolaci si můžeme představit jako Taylorův rozvoj, kde derivace jsou nahrazeny diferencemi.
Pro odvození filtračního kroku předpokládejme dostupný prediktivní odhad stavu

$$p(x_k|z^{k-1}) = N(x_k : \hat{x}'_k, P'_k) \tag{4.3.1}$$

a zaveďme následující lineární transformaci stavu

$$\chi_k = (S'_k)^{-1}x_k \tag{4.3.2}$$

kde S'_k je rozklad (nebo-li odmocnina) prediktivní kovarianční matice P'_k takový, že

$$P'_k = S'_k(S'_k)^T \tag{4.3.3}$$

Odmocnina může být spočtena např. Choleského dekompozicí nebo singulárním rozkladem (v prostředí MATLAB® lze využít funkce `chol` nebo `svd`). S ohledem na (4.3.1) je prediktivní hustota pravděpodobnosti χ_k (4.3.2) dána normální hustotou pravděpodobnosti

$$p(\chi_k|z^{k-1}) = N(\chi_k : \hat{\chi}'_k, I) \tag{4.3.4}$$

kde $\hat{\chi}'_k = (S'_k)^{-1}\hat{x}'_k$. Význam této transformace spočívá v tom, že prvky vektoru χ_k jsou nezávislé a mají jednotkovou varianci. V konečném důsledku nám tato transformace umožní odlišný a v literatuře preferovaný způsob odvození, než tomu bylo u rozšířeného Kalmanova filtru.
S přihlédnutím k transformaci (4.3.2) může být rovnice měření (4.1.2) zapsána ve formě

$$\begin{aligned} z_k &= h_k(x_k) + v_k \\ &= h_k(S'_k\chi_k) + v_k \\ &= \tilde{h}_k(\chi_k) + v_k \end{aligned} \tag{4.3.5}$$

Stirlingova interpolace prvního řádu nelineární funkce v rovnici měření (4.3.5) v okolí linearizačního bodu $\hat{\chi}'_k$ je dána

$$\tilde{h}_k(\chi_k) \approx \tilde{h}_k(\hat{\chi}'_k) + \tilde{H}_k(\hat{\chi}'_k, \kappa)(\chi_k - \hat{\chi}'_k) \tag{4.3.6}$$

kde matice nz/nx prvních diferencí je

$$\tilde{H}_k(\hat{\chi}'_k, \kappa) = \begin{bmatrix} \frac{\tilde{h}_{1,k}(\hat{\chi}'_k+\kappa e_1)-\tilde{h}_{1,k}(\hat{\chi}'_k-\kappa e_1)}{2\kappa} & \cdots & \frac{\tilde{h}_{1,k}(\hat{\chi}'_k+\kappa e_{nx})-\tilde{h}_{1,k}(\hat{\chi}'_k-\kappa e_{nx})}{2\kappa} \\ \vdots & \ddots & \vdots \\ \frac{\tilde{h}_{nz,k}(\hat{\chi}'_k+\kappa e_1)-\tilde{h}_{nz,k}(\hat{\chi}'_k-\kappa e_1)}{2\kappa} & \cdots & \frac{\tilde{h}_{nz,k}(\hat{\chi}'_k+\kappa e_{nx})-\tilde{h}_{nz,k}(\hat{\chi}'_k-\kappa e_{nx})}{2\kappa} \end{bmatrix} \tag{4.3.7}$$

a κ je tzv. škálovací parametr, jehož volba je diskutována později.
Pak filtrační hustota stavu v časovém okamžiku k je dána $p_A(x_k|z^k) = N(x_k : \hat{x}_k, P_k)$ s momenty

$$\hat{x}_k = \hat{x}'_k + P'_{xz,k}(P'_{z,k})^{-1}(z_k - \hat{z}'_k) \tag{4.3.8}$$

$$P_k = P'_k - P'_{xz,k}(P'_{z,k})^{-1}(P'_{xz,k})^T \tag{4.3.9}$$

kde prediktivní střední hodnota měření, s uvažováním Stirlingovy interpolace (4.3.6), je

$$\begin{aligned} \hat{z}'_k &= E[z_k|z^{k-1}] \\ &= E[h_k(x_k)|z^{k-1}] = E[\tilde{h}_k(\chi_k)|z^{k-1}] \\ &\approx E[\tilde{h}_k(\hat{\chi}'_k) + \tilde{H}_k(\hat{\chi}'_k, \kappa)(\chi_k - \hat{\chi}'_k)|z^{k-1}] \\ &= \tilde{h}_k(\hat{\chi}'_k) \end{aligned} \tag{4.3.10}$$

Tu lze, využitím (4.3.2), (4.3.5), zapsat jako funkci prediktivních momentů netransformovaného stavu

$$\begin{aligned} \hat{z}'_k &= \tilde{h}_k((S'_k)^{-1}\hat{x}'_k) \\ &= h_k(S'_k(S'_k)^{-1}\hat{x}'_k) \\ &= h_k(\hat{x}'_k) \end{aligned} \tag{4.3.11}$$

Podobně vypočteme prediktivní kovarianční matici

$$\begin{aligned} P'_{z,k} &= cov[z_k|z^{k-1}] = E[(z_k - E[z_k|z^{k-1}])(z_k - E[z_k|z^{k-1}])^T|z^{k-1}] \\ &\approx E[(\tilde{H}_k(\hat{\chi}'_k, \kappa)(\chi_k - \hat{\chi}'_k))(\tilde{H}_k(\hat{\chi}'_k, \kappa)(\chi_k - \hat{\chi}'_k)|z^{k-1}] + E[v_k v_k^T] \\ &= \tilde{H}_k(\hat{\chi}'_k, \kappa)I(\tilde{H}_k(\hat{\chi}'_k, \kappa))^T + R_k \\ &= H_k(\hat{x}'_k, \kappa)(H_k(\hat{x}'_k, \kappa))^T + R_k \end{aligned} \tag{4.3.12}$$

kde, dle (4.3.2), (4.3.5),

$$H_k(\hat{x}'_k, \kappa) = \begin{bmatrix} \frac{h_{1,k}(\hat{x}'_k+\kappa s'_{k,1})-h_{1,k}(\hat{x}'_k-\kappa s'_{k,1})}{2\kappa} & \cdots & \frac{h_{1,k}(\hat{x}'_k+\kappa s'_{k,nx})-h_{1,k}(\hat{x}'_k-\kappa s'_{k,nx})}{2\kappa} \\ \vdots & \ddots & \vdots \\ \frac{h_{nz,k}(\hat{x}'_k+\kappa s'_{k,1})-h_{nz,k}(\hat{x}'_k-\kappa s'_{k,1})}{2\kappa} & \cdots & \frac{h_{nz,k}(\hat{x}'_k+\kappa s'_{k,nx})-h_{nz,k}(\hat{x}'_k-\kappa s'_{k,nx})}{2\kappa} \end{bmatrix} \tag{4.3.13}$$

a $s'_{k,i}$ je i-tý sloupec matice S'_k. Výpočet vzájemné kovarianční matice je pak

$$\begin{aligned} P'_{xz,k} &= cov[x_k, z_k|z^{-1}] = E[(x_k - \hat{x}'_k)(z_k - E[z_k|z^{-1}])^T|z^{k-1}] \\ &\approx E[(S'_k(\chi_k - \hat{\chi}'_k))(\tilde{H}_k(\hat{\chi}'_k, \kappa)(\chi_k - \hat{\chi}'_k))^T|z^{k-1}] \\ &= S'_k(\tilde{H}_k(\hat{\chi}'_k, \kappa))^T = S'_k(H_k(\hat{x}'_k, \kappa))^T \end{aligned} \tag{4.3.14}$$

Porovnejme vztahy pro výpočet prediktivních momentů rozšířeného Kalmanova filtru (4.1.21)–(4.1.22) a diferenčního filtru (4.3.10)–(4.3.14). Vidíme, že vztahy pro výpočet střední hodnoty jsou totožné, zatímco vztahy pro výpočet kovariančních matic se významně liší. Kovarianční matice predikce měření pro rozšířený Kalmanův filtr jsou lineární funkcí prediktivní matice P'_k, zatímco pro diferenční lokální filtr jsou nelineární.
Využitím analogického přístupu k odvození prediktivní střední hodnoty měření (4.3.11) a kovarianční matice měření (4.3.12), jsou momenty prediktivní hustoty stavu pro časový okamžik $k+1$, tj., $p_A(x_{k+1}|z^k) = N(x_{k+1} : \hat{x}'_{k+1}, P'_{k+1})$, založeny na aproximaci nelineární funkce v rovnici dynamiky (4.3.1) pomocí Stirlingovy interpolace prvního řádu v okolí bodu $\hat{\chi}_k = (S_k)^{-1}\hat{x}_k$

$$\tilde{f}_k(\chi_k) \approx \tilde{f}_k(\hat{\chi}_k) + \tilde{F}_k(\hat{\chi}_k, \kappa)(\chi_k - \hat{\chi}_k) \tag{4.3.15}$$

kde $\chi_k = (S_k)^{-1}x_k$, S_k je odmocnina matice P_k taková, že $P_k = S_k(S_k)^T$, $p(\chi_k|z^{k-1}) = N(\chi_k : \hat{\chi}_k, I)$ s $\hat{\chi}_k = (S_k)^{-1}\hat{x}_k$, $\tilde{f}_k(\chi_k) = f_k(S_k\chi_k)$ a $\tilde{F}_k(\hat{\chi}'_k, \kappa)$ je matice nx/nx prvních diferencí definovaná

$$\tilde{F}_k(\hat{\chi}_k, \kappa) = \begin{bmatrix} \frac{\tilde{f}_{1,k}(\hat{\chi}_k+\kappa e_1)-\tilde{f}_{1,k}(\hat{\chi}_k-\kappa e_1)}{2\kappa} & \cdots & \frac{\tilde{f}_{1,k}(\hat{\chi}_k+\kappa e_{nx})-\tilde{f}_{1,k}(\hat{\chi}_k-\kappa e_{nx})}{2\kappa} \\ \vdots & \ddots & \vdots \\ \frac{\tilde{f}_{nx,k}(\hat{\chi}_k+\kappa e_1)-\tilde{f}_{nx,k}(\hat{\chi}_k-\kappa e_1)}{2\kappa} & \cdots & \frac{\tilde{f}_{nx,k}(\hat{\chi}_k+\kappa e_{nx})-\tilde{f}_{nx,k}(\hat{\chi}_k-\kappa e_{nx})}{2\kappa} \end{bmatrix} \tag{4.3.16}$$

Prediktivní střední hodnotu a kovarianční matici lze spočítat následujícím způsobem

$$\begin{aligned} \hat{x}'_{k+1} &= E[x_{k+1}|z^k] \\ &= E[f_k(x_k)|z^k] = E[\tilde{f}_k(\chi_k)|z^k] \\ &\approx E[\tilde{f}_k(\hat{\chi}_k) + \tilde{F}_k(\hat{\chi}_k, \kappa)(\chi_k - \hat{\chi}_k)|z^k] \\ &= \tilde{f}_k(\hat{\chi}_k) \\ &= f_k(\hat{x}_k) \end{aligned} \tag{4.3.17}$$

$$\begin{aligned} P'_{k+1} &= cov[x_{k+1}|z^{k-1}] = E[(x_{k+1} - E[x_{k+1}|z^k])(x_{k+1} - E[x_{k+1}|z^k])^T|z^k] \\ &\approx E[(\tilde{F}_k(\hat{\chi}_k, \kappa)(\chi_k - \hat{\chi}_k))(\tilde{F}_k(\hat{\chi}_k, \kappa)(\chi_k - \hat{\chi}_k))^T|z^k] + E[w_k w_k^T] \\ &= \tilde{F}_k(\hat{\chi}_k, \kappa)I(\tilde{F}_k(\hat{\chi}_k, \kappa))^T + Q_k \\ &= F_k(\hat{x}_k, \kappa)(F_k(\hat{x}_k, \kappa))^T + Q_k \end{aligned} \tag{4.3.18}$$

kde

$$F_k(\hat{x}_k, \kappa) = \begin{bmatrix} \frac{f_{1,k}(\hat{x}_k+\kappa s_{k,1})-f_{1,k}(\hat{x}_k-\kappa s_{k,1})}{2\kappa} & \cdots & \frac{f_{1,k}(\hat{x}_k+\kappa s_{k,nx})-f_{1,k}(\hat{x}_k-\kappa s_{k,nx})}{2\kappa} \\ \vdots & \ddots & \vdots \\ \frac{f_{nx,k}(\hat{x}_k+\kappa s_{k,1})-f_{nx,k}(\hat{x}_k-\kappa s_{k,1})}{2\kappa} & \cdots & \frac{f_{nx,k}(\hat{x}_k+\kappa s_{k,nx})-f_{nx,k}(\hat{x}_k-\kappa s_{k,nx})}{2\kappa} \end{bmatrix} \tag{4.3.19}$$

a $s_{k,i}$ je i-tý sloupec matice S_k.
Vztahy (4.3.8)–(4.3.14), (4.3.17) a (4.3.18) tak představují algoritmus diferenčního filtru. Pro jeho implementaci je nutné specifikovat škálovací parametr κ. Lze ukázat, že vhodnou volbou parametru je $\kappa = \sqrt{3}$ [90]. Detailní diskuze k volbě parametru spolu s doporučeními volby může být nalezena např. v [94].
Závěrem poznamenejme, že v publikaci [90] byl odvozen i diferenční filtr druhého řádu, který je založen na Stirlingově interpolaci druhého řádu a poskytuje v mnoha případech lepší kvalitu odhadu.

Poznámka . Odvození diferenčních filtrů je založeno na lineární transformaci stavu (4.3.2). Transformace může být chápána jako stochastické rozvazbení (z anglického výrazu „stochastic decoupling“) elementů náhodné veličiny x_k vedoucí na náhodnou proměnnou χ_k. Ta má pak jednotkovou kovarianční matici, což umožní snadnou a zejména konstantní volbu škálovacího parametru κ pro všechny časové okamžiky. Samozřejmě, odvození diferenčních filtrů by bylo možné i bez této transformace, avšak v tomto případě by bylo nutné volit proměnný škálovací parametr, a to právě v závislosti na aktuální kovarianční matici P'_k veličiny x_k (čím větší je P'_k, tím větší je stavový prostor, ve kterém se skutečný stav x_k může nacházet a tím je tedy nutné zvětšit oblast pomocí parametru κ, kde nelineární funkce je aproximována Stirlingovo interpolací).

Příklad 4.3.1 Stirlingova interpolace může být chápána jak Taylorův rozvoj, kde derivace jsou nahrazeny diferencemi. Pro lepší představu a ilustraci, uvažujme následující filtrační odhad

$$p(x_k|z^k) = N(x_k : \hat{x}_k, P_k) = N(x_k : 1, 1) \tag{4.3.20}$$

a následující nelineární funkci ve stavové rovnici

$$f_k(x_k) = \sin(x_k^2) \tag{4.3.21}$$

Taylorův rozvoj funkce $f_k(\mathbf{x}_k)$ v okolí nejlepšího odhadu stavu $\hat{x}_k$ vede na lineární vztah

$$\begin{aligned} f_k(x_k) &\approx f_k(\hat{x}_k) + \tfrac{df_k(x_k)}{dx_k}|_{x_k=\hat{x}_k}(x_k - \hat{x}_k) \\ &= \sin(\hat{x}_k^2) + 2\hat{x}_k \cos(\hat{x}_k^2)(x_k - \hat{x}_k) \\ &= 0.8415 + 1.0806(x_k - 1) \end{aligned} \tag{4.3.22}$$

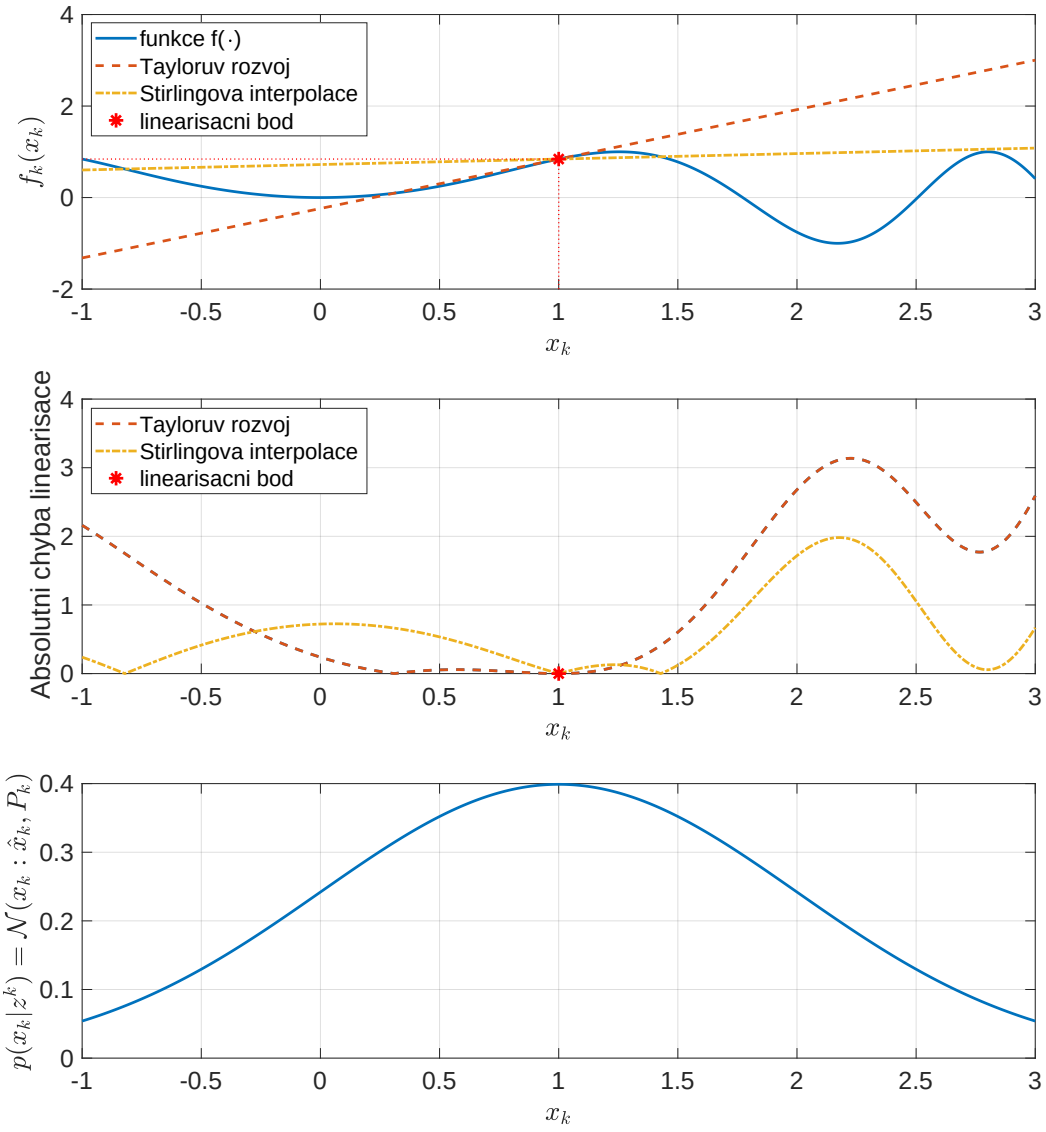
Naproti tomu Stirlingova interpolace se škálovacím parametrem $\kappa = \sqrt{3}$ vede na následující lineární funkci

$$\begin{aligned} f_k(x_k) &\approx f_k(\hat{x}_k) + F_k(\hat{x}_k, \kappa)(x_k - \hat{x}_k) \\ &= f_k(\hat{x}_k) + \frac{f_k(\hat{x}_k + \kappa S_k) - f_k(\hat{x}_k - \kappa S_k)}{2\kappa}(x_k - \hat{x}_k) \\ &= \sin(\hat{x}_k^2) + \frac{\sin\left((\hat{x}_k + \kappa S_k)^2\right) - \sin\left((\hat{x}_k - \kappa S_k)^2\right)}{2\kappa}(x_k - \hat{x}_k) \\ &= 0.8415 + 0.1196(x_k - 1) \end{aligned} \tag{4.3.23}$$

Obě linearizace tedy vedou k poměrně odlišným lineárním funkcím. Nelineární funkce, její linearizace, chyby linearizací a filtrační hustota pravděpodobnosti jsou znázorněny na obr. 4.1. Lze vidět, že Taylorův rozvoj má menší chybu v blízkém okolí linearizačního bodu, zatímco Stirlingova interpolace je „přesnější“ v širším okolí.

4.4 Unscentovaný Kalmanův filtr

Unscentovaný Kalmanův filtr, pro který se v anglicky psané literatuře vžil pojmem „unscented Kalman filter“, je dalším zástupcem bezderivačních filtrů. Unscentovaný filtr je však založen na odlišné myšlence od rozšířeného Kalmanova filtru a diferenčního filtru; při návrhu filtru není aproximována nelineární funkce, ale je aproximován popis odhadu stavu množinou deterministicky volených bodů [91], [93], [94], [88], [96], [97].



Obrázek 4.1: Ilustrace linearizace nelineární funkce Taylorovo rozvojem a Stirlingovo interpolací

Filtrační krok unscentovaného filtru spočívá opět ve výpočtu momentů aproximativní filtrační hustoty pravděpodobnosti $p_A(x_k|z^k) = N(x_k : \hat{x}_k, P_k)$, kde

$$\hat{x}_k = \hat{x}'_k + P'_{xz,k}(P'_{z,k})^{-1}(z_k - \hat{z}'_k) \tag{4.4.1}$$

$$P_k = P'_k - P'_{xz,k}(P'_{z,k})^{-1}(P'_{xz,k})^T \tag{4.4.2}$$

Výpočet prediktivních momentů odhadu měření vychází z aproximace prediktivní hustoty pravděpodobnosti $p_A(x_k|z^{k-1}) = N(x_k : \hat{x}'_k, P'_k)$ množinou deterministicky volených vážených bodů, tzv. sigma-bodů, a jejich transformace přes nezměněnou nelineární funkci v rovnici měření. Tedy, mějme prediktivní hustotu pravděpodobnosti stavu a definujme množinu $(2nx+1)$ sigma-

bodů dle následujících vztahů

$$\mathcal{X}'_{k,0} = \hat{x}'_k \tag{4.4.3}$$
$$\mathcal{X}'_{k,i} = \hat{x}'_k + \sqrt{nx+\kappa}s'_{k,i}, i = 1, \ldots, nx \tag{4.4.4}$$
$$\mathcal{X}'_{k,j} = \hat{x}'_k - \sqrt{nx+\kappa}s'_{k,j-nx}, j = nx+1, \ldots, 2nx \tag{4.4.5}$$

kde $s'_{k,i}$ je i-tý sloupec matice S'_k splňující rovnost $S'_k(S'_k)^T = P'_k$ a κ je opět škálovací parametr, jehož volba je diskutována později. Každý bod je asociován s váhou

$$\mathcal{W}_0 = \frac{\kappa}{nx+\kappa} \tag{4.4.6}$$
$$\mathcal{W}_i = \frac{1}{2(nx+\kappa)}, i = 1, \ldots, 2nx \tag{4.4.7}$$

Všimněme si klíčové vlastnosti množiny vážených sigma-bodů $\{\mathcal{X}'_{k,i}, \mathcal{W}_i\}_{i=0}^{2nx}$, kterou je zachování prvních dvou momentů prediktivní hustoty pravděpodobnosti $p_A(x_k|z^{k-1}) = N(x_k : \hat{x}'_k, P'_k)$, tj. střední hodnota a kovarianční matice množiny bodů je

$$\sum_{i=0}^{2nx} \mathcal{W}_i \mathcal{X}'_{k,i} = \hat{x}'_k$$
$$\sum_{i=0}^{2nx} \mathcal{W}_i (\mathcal{X}'_{k,i} - \hat{x}'_k)(\mathcal{X}'_{k,i} - \hat{x}'_k)^T = P'_k$$

Pak, jsou sigma-body transformovány přes nezměněnou nelineární funkci v rovnici měření (4.1.2)

$$\mathcal{Z}'_{k,i} = h_k(\mathcal{X}'_{k,i}), i = 0, 1, \ldots, 2nx \tag{4.4.8}$$

a na jejich základě jsou vypočteny požadované prediktivní charakteristiky měření dle

$$\hat{z}'_k = \sum_{i=0}^{2nx} \mathcal{W}_i \mathcal{Z}'_{k,i} \tag{4.4.9}$$
$$P'_{z,k} = \sum_{i=0}^{2nx} \mathcal{W}_i (\mathcal{Z}'_{k,i} - \hat{z}'_k)(\mathcal{Z}'_{k,i} - \hat{z}'_k)^T + R_k \tag{4.4.10}$$
$$P'_{xz,k} = \sum_{i=0}^{2nx} \mathcal{W}_i (\mathcal{X}'_{k,i} - \hat{x}'_k)(\mathcal{Z}'_{k,i} - \hat{z}'_k)^T \tag{4.4.11}$$

Prediktivní charakteristiky měření (4.4.9)–(4.4.11) jsou použity ve vztazích (4.4.1), (4.4.2) a tím končí filtrační krok unscentovaného filtru.
Výpočet momentů prediktivní hustoty pravděpodobnosti $p_A(x_{k+1}|z^k) = N(x_k : \hat{x}'_{k+1}, P'_{k+1})$ startuje výpočtem filtrační množiny sigma-bodů na základě filtračních momentů dle

$$\mathcal{X}_{k,0} = \hat{x}_k \tag{4.4.12}$$
$$\mathcal{X}_{k,i} = \hat{x}_k + \sqrt{nx+\kappa}s_{k,i}, i = 1, \ldots, nx \tag{4.4.13}$$
$$\mathcal{X}_{k,j} = \hat{x}_k - \sqrt{nx+\kappa}s_{k,j-nx}, j = nx+1, \ldots, 2nx \tag{4.4.14}$$

kde $s_{k,i}$ je i-tý sloupec matice S_k splňující rovnost $S_k S_k^T = P_k$. Poznamenejme, že váhy jsou funkcí dimenze stavu a konstantního škálovacího parametru, proto je stačí vypočítat jen jednou.

Filtrační sigma-body jsou transformovány přes nezměněnou nelineární funkci v rovnici dynamiky (4.1.1)

$$\mathcal{X}'_{k+1,i} = f_k(\mathcal{X}_{k,i}), \forall i \tag{4.4.15}$$

a na jejich základě jsou vypočteny požadované prediktivní charakteristiky stavu

$$\hat{x}'_{k+1} = \sum_{i=0}^{2nx} \mathcal{W}_i \mathcal{X}'_{k+1,i} \tag{4.4.16}$$

$$P'_{k+1} = \sum_{i=0}^{2nx} \mathcal{W}_i (\mathcal{X}'_{k+1,i} - \hat{x}'_{k+1})(\mathcal{X}'_{k+1,i} - \hat{x}'_{k+1})^T + Q_k \tag{4.4.17}$$

Vztahy (4.4.1)–(4.4.11) a (4.4.12)-(4.4.17) reprezentují algoritmus unscentovaného Kalmanova filtru. Škálovací parametr nutný pro výpočet sigma-bodů a vah lze, dle analýzy v [91], [94], zvolit $\kappa = 3 - nx$, pokud $nx \leq 3$, nebo $\kappa = 0$, pokud $nx > 3$. Závěrem poznamenejme, že kvalitou odhadu je unscentovaný filtr srovnatelný s filtrem druhého řádu.

Poznámka . Unscentovaný filtr lze považovat za zástupce širší třídy lokálních filtrů označovaných jako sigma-bodové filtry. Ty jsou založeny na aproximaci integrálu nelineární funkce s gaussovskou váhou, které se vyskytují při návrhu lokálních filtrů. Např. výpočet prediktivní střední hodnoty stavu lze zapsat

$$\hat{x}'_{k+1} = E[x_{k+1}|z^k] = E[f_k(x_k)|z^k] \tag{4.4.18}$$

$$\approx \int f_k(x_k) p_A(x_k|z^k) dx_k \tag{4.4.19}$$

$$= \int f_k(x_k) N(x_k : \hat{x}_k, P_k) dx_k \tag{4.4.20}$$

Řešení tohoto typu integrálu bylo v literatuře věnováno mnoho úsilí již od 19. století a bylo navrženo velké množství tzv. kvadraturních či kubaturních integračních pravidel. Všechna integrační pravidla lze zapsat ve tvaru

$$\int f_k(x_k) N(x_k : \hat{x}_k, P_k) dx_k \approx \sum_{i=1}^{m} \omega_i f_k(\chi_{k,i}) \tag{4.4.21}$$

kde $\{\chi_{k,i}, \omega_i\}_{i=1}^m$ je množina kvadraturních či kubaturních bodů s příslušnými váhami. Právě výpočet bodů a vah je to, co odlišuje jednotlivá integrační pravidla. Poznamenejme také, že pro určitý typ nelineárních funkcí $f_k(\cdot)$, jakými jsou například polynomiální funkce, poskytují kvadraturní či kubaturní pravidla přesné výsledky. Více o použití integračních pravidel v návrhu lokálních filtrů a jejich přesnosti lze najít např. v [92], [95].

Poznámka . Ačkoliv základní idee unscentovaného a diferenčního filtru (druhého řádu) jsou odlišné a stejně tak i způsob jejich odvození je odlišný, je pozoruhodné, že výsledné algoritmy jsou, za určitých podmínek, totožné [93].

4.5 Iterační filtr

Používaný bayesovský přístup umožňuje sledovat odděleně problém filtrace a jednokrokové predikce. Dále umožňuje sledovat, jak nová informace ve formě měření ovlivňuje filtrační

hustotu. Rovněž je zřejmé, jakým způsobem se projeví linearizace v bodě nejlepšího odhadu ve smyslu střední hodnoty při syntéze filtru. Podívejme se blíže na vztahy definující filtrační gaussovskou hustotu pro střední hodnotu a kovarianci rozšířeného Kalmanova filtru

$$\hat{x}_k = \hat{x}'_k + P'_k H_k^T(\hat{x}'_k)[H_k(\hat{x}'_k)P'_k H_k^T(\hat{x}'_k) + R_k]^{-1}[z_k - h_k(\hat{x}'_k)] \tag{4.5.1}$$

$$P_k = P'_k - P'_k H_k^T(\hat{x}'_k)[H_k(\hat{x}'_k)P'_k H_k^T(\hat{x}'_k) + R_k]^{-1} H_k(\hat{x}'_k)P'_k \tag{4.5.2}$$

Je zřejmé, že $\hat{x}_k$ obsahuje více informace o x_k než $\hat{x}'_k$. Ale linearizace byla provedena v $\hat{x}'_k$. Toho můžeme využít a, obrazně řečeno, provést linearizaci v kroku k znovu, ale v bodě $\hat{x}_k$. Tím bychom dostali novou hodnotu odhadů a tento postup bychom mohli opakovat tak dlouho, dokud by rozdíl následujících dvou odhadů nebyl menší než předem dané ϵ.
V literatuře lze nalézt mnoho přístupů realizujících zmíněný koncept. Zde uvedeme pouze výsledný algoritmus přístupu, který hledá optimální odhad stavu ve smyslu maximální věrohodnosti pomocí Newton-Raphsonova iteračního optimalizačního algoritmu. Odvození, které je mimo rozsah těchto skript, a další možné přístupy lze nalézt např. v [4], [12], [96], [97].
Rekurzivní maximalizace věrohodnostní funkce ve výsledku spočívá v iterování následujících rovnic

$$\hat{x}_k^{i+1} = \hat{x}'_k + P'_k H_k^T(\hat{x}_k^i)[H_k(\hat{x}_k^i)P'_k H_k^T(\hat{x}_k^i) + R_k]^{-1}[z_k - h_k(\hat{x}'_k) - H_k(\hat{x}_k^i)(\hat{x}'_k - \hat{x}_k^i)] \tag{4.5.3}$$

$$P_k^{i+1} = P'_k - P'_k H_k^T(\hat{x}_k^i)[H_k(\hat{x}_k^i)P'_k H_k^T(\hat{x}_k^i) + R_k]^{-1} H_k(\hat{x}_k^i)P'_k \tag{4.5.4}$$

pro $i = 1, 2, 3, \ldots, imax$ s počáteční podmínkou

$$\hat{x}_k^1 = \hat{x}'_k \tag{4.5.5}$$

Iterace bude ukončena, když

$$\|\hat{x}_k^{i+1} - \hat{x}_k^i\|_2 < \epsilon, \quad \epsilon > 0 \tag{4.5.6}$$

a hodnota $i + 1$, kdy došlo k ukončení iterace, bude značena $imax$. Tyto rovnice tak definují filtrační krok iteračního filtru. Je velmi zajímavé, že i když se jedná o odlišný přístup k návrhu filtru, filtrační krok iteračního filtru v sobě zahrnuje filtrační krok rozšířeného Kalmanova filtru jako speciální případ, když $i = 1$.
Podobnou úvahu nelze aplikovat pro predikci, protože predikce neoperuje s žádnou informací z reality. To znamená, že vztahy pro predikci budou shodné se vztahy u rozšířeného Kalmanova filtru. Tedy

$$\hat{x}'_{k+1} = f_k(\hat{x}_k) \tag{4.5.7}$$

$$P'_{k+1} = F_k(\hat{x}_k)P_k F_k^T(\hat{x}_k) + Q_k \tag{4.5.8}$$

kde

$$\hat{x}_k = \hat{x}_k^{imax} \tag{4.5.9}$$

$$P_k = P_k^{imax} \tag{4.5.10}$$

Vztahy (4.5.1)–(4.5.10) definují iterační filtr, v anglicky psané literatuře označovaného jako „iterated extended Kalman filter“. Tento filtr je jistým vylepšením rozšířeného Kalmanova filtru, jelikož zpřesňuje lokální aproximaci při výpočtu filtračního odhadu. Nicméně jedná se opět o lokální aproximaci a rovněž konvergence odhadu opět není zaručena. Poznamenejme, že v literatuře můžeme nalézt i iterační filtr využívající jiné aproximace než je Taylorův rozvoj prvního řádu, jako například unscentovanou transformaci.

Kapitola 5

Nelineární filtrace s danou strukturou hustot pravděpodobnosti

V této kapitole bude zachován základní rámec úlohy estimace využívající bayesovský přístup a zdůrazňující analytické řešení problému. Zároveň však na rozdíl od předchozí kapitoly budeme klást větší důraz na kvalitu odhadu a proto přejdeme od lokálních filtrů k filtrům globálním [6], [7], [14], [36], [42], [43], [44], [45], [46], [47]. Linearizace v jednom bodě stavového prostoru vedoucí na lokální filtry bude nahrazena vícenásobnou linearizací pokrývající větší část stavového prostoru. Odhad stavu, ať už ve smyslu získání hustoty pravděpodobnosti či bodového odhadu, bude prováděn jak pro nelineární systémy, tak pro lineární negaussovské systémy, jelikož i v těchto případech nelze vystačit se standardní lineární filtrací. Základním stavebním kamenem navrhovaných filtrů bude Kalmanův filtr a rozšířený Kalmanův filtr. Základním předpokladem pro popis náhodných veličin bude možnost zavedení hustoty ve tvaru součtu normálních rozdělení [39], [40], [41], [42].

5.1 Základní formulace úlohy

Vraťme se k systému popsanému vztahy (4.1.1), (4.1.2) (nelineární dynamika, nelineární měření) se stavovým šumem (4.1.4) a šumem měření (4.1.5), tedy k úloze, která je formulovaná v úvodu podkapitoly 4.1. Na rozdíl od úlohy v 4.1 předpokládejme, že počáteční stav je popsán hustotou pravděpodobnosti ve tvaru součtu normálních rozdělení. Takový typ rozdělení začal používat na poli estimace [41] a později [48], [39].

Uvažujme, známou hustotu pravděpodobnosti počátečního stavu ve formě směsi normálních rozdělení

$$p(x_0 \mid z^{-1}) = \sum_{i=1}^{\xi_0'} \alpha_{0i}' N(x_0 : \hat{x}_{0i}^{'}, P_{0i}') \tag{5.1.1}$$

kde α_{0i}' jsou váhové koeficienty, pro které platí

$$\sum_{i=1}^{\xi_0'} \alpha_{0i}' = 1 \quad \alpha_{0i}' > 0$$

Hustotu pravděpodobnosti (5.1.1) lze chápat buď jako

- exaktně danou hustotu (např. pro modelování odlehlých chyb měření, jak je diskutováno v podkapitole 6.5) nebo
- aproximaci skutečné negaussovské hustoty charakterizující počáteční stav systému (jak je diskutováno v podkapitole 6.4 a pracích [40], [35], [49], [115]).

Předpoklad počáteční podmínky ve formě směsi normálních rozložení nám umožní plynulý přechod od lokálních filtrů, především reprezentovaných rozšířeným Kalmanovým filtrem k filtrům kvalitativně vyššího typu.

Pro odvození filtrační a prediktivní hustoty pravděpodobnosti užijeme opět bayesovské vztahy a výsledky třetí a čtvrté kapitoly. Začneme filtrační hustotou pro $k = 0$.

Hustotu pravděpodobnosti měření můžeme snadno určit

$$p(z_0 \mid x_0) = N(z_0 : h_0(x_0), R_0) \tag{5.1.2}$$

Zkušenosti získané ve 4. kapitole nám však říkají, že nelineární funkce $h_0(\cdot)$ znemožňuje analytické řešení bayesovských vztahů. Abychom zajistili řešitelnost těchto vztahů, je třeba tuto funkci vhodně aproximovat [85]. Klíčovou otázkou aproximace je, jak vybrat linearizační bod za předpokladu počáteční podmínky ve formě směsi normálních hustot. Odpověď na tuto otázku nalezneme v následující části.

5.2 Filtr s vícenásobnou linearizací

Vyjděme z formulace úlohy provedené v předchozí podkapitole 5.1. Podle bayesovského vztahu a pro danou formulaci úlohy dostaneme filtrační hustotu

$$p(x_0 \mid z^0) = \frac{\overbrace{\sum_{i=1}^{\xi'_0} \alpha'_{0i} N(x_0 : \hat{x}'_{0i}, P'_{0i})}^{p(x_0|z^{-1})} \overbrace{N(z_0 : h_0(x_0), R_0)}^{p(z_0|x_0)}}{\underbrace{\int \sum_{i=1}^{\xi'_0} \alpha'_{0i} N(x_0 : \hat{x}'_{0i}, P'_{0i}) N(z_0 : h_0(x_0), R_0) dx_0}_{p(z_0|z^{-1})}} \tag{5.2.1}$$

kterou můžeme přepsat do formy

$$p(x_0 \mid z^0) = \frac{\sum_{i=1}^{\xi'_0} \alpha'_{0i} \Big(N(x_0 : \hat{x}'_{0i}, P'_{0i}) N(z_0 : h_0(x_0), R_0) \Big)}{p(z_0|z^{-1})}$$

ze které lze snáze vidět, že bayesovský vztah (5.2.1) je vážený součet ξ'_0 bayesovských vztahů použitých při odvození rozšířeného Kalmanova filtru, tj. vztahu (4.1.8). Každý dílčí vztah je však svázán s jinou apriorní gaussovskou hustotou pravděpodobnosti $N(x_0 : \hat{x}'_{0i}, P'_{0i})$.
Abychom zajistili analytickou řešitelnost tohoto vztahu, lze ξ'_0-krát použít myšlenku rozšířeného Kalmanova filtru. To znamená, že provedeme linearizaci funkce $h_0(x_0)$ pro každý gaussovský člen prediktivní hustoty pravděpodobnosti (5.1.1) *nezávisle* na ostatních a tedy Funkci $h_0(x_0)$ linearizujeme ve všech bodech $\hat{x}'_{0i}$, pro $i = 1, 2, ..., \xi'_0$. Tento postup je označován jako

vícenásobná linearizace [51]. Výraz (5.2.1) se pak změní na

$$p_A(x_0 \mid z^0) = \frac{\sum_{i=1}^{\xi_0'} \alpha_{0i}' N(x_0 : \hat{x}_{0i}', P_{0i}') N(z_0 : h_0(\hat{x}_{0i}') + H(\hat{x}_{0i}')(x_0 - \hat{x}_{0i}'), R_0)}{\int \sum_{i=1}^{\xi_0'} \alpha_{0i}' N(x_0 : \hat{x}_{0i}', P_{0i}') N(z_0 : h_0(\hat{x}_{0i}') + H(\hat{x}_{0i}')(x_0 - \hat{x}_{0i}'), R_0) dx_0} \tag{5.2.2}$$

Čitatel zlomku na pravé straně (5.2.2) obsahuje členy

$$\alpha_{0i}' N(x_0 : \hat{x}_{0i}', P_{0i}') N(z_0 : h_0(\hat{x}_{0i}') + H_0(\hat{x}_{0i}')(x_0 - \hat{x}_{0i}'), R_0)$$

které jsou shodné až na váhový koeficient α_{0i}' se členem použitým při odvození rozšířeného Kalmanova filtru ve vztahu (4.1.10). To znamená, že (5.2.2) můžeme řešit ξ_0'-krát aplikací odvození rozšířeného Kalmanova filtru. Pro i-tý rozšířený Kalmanův filtr bude, s ohledem na (4.1.8), platit $p_i(x_0 \mid z^0) p_i(z_0 \mid z^{-1}) = p_i(x_0 \mid z^{-1}) p_i(z_0 \mid x_0)$, tedy

$$\begin{aligned} & \alpha_{0i}' N(x_0 : \hat{x}_{0i}, P_{0i}) N(z_0 : h_0(\hat{x}_{0i}'), H(\hat{x}_{0i}') P_{0i}' H^T(\hat{x}_{0i}') + R_0) \\ = \; & \alpha_{0i}' N(x_0 : \hat{x}_{0i}', P_{0i}') N(z_0 : h_0(\hat{x}_{0i}') + H(\hat{x}_{0i}')(x_0 - \hat{x}_{0i}'), R_0) \end{aligned} \tag{5.2.3}$$

Pro zjednodušení zápisu definujme

$$\zeta_{0i} \stackrel{\triangle}{=} N(z_0 : h_0(\hat{x}_{0i}'), H(\hat{x}_{0i}') P_{0i}' H^T(\hat{x}_{0i}') + R_0) \tag{5.2.4}$$

a

$$\bar{\alpha}_{0i} = \alpha_{0i}' \zeta_{0i} \tag{5.2.5}$$

Pak filtrační hustotu můžeme zapsat tímto způsobem

$$p_A(x_0 \mid z^0) = \frac{\sum_{i=1}^{\xi_0} \bar{\alpha}_{0i} N(x_0 : \hat{x}_{0i}, P_{0i})}{\sum_{i=1}^{\xi_0} \bar{\alpha}_{0i}} = \sum_{i=1}^{\xi_0} \alpha_{0i} N(x_0 : \hat{x}_{0i}, P_{0i}) \tag{5.2.6}$$

kde $\hat{x}_{0i}, P_{0i}$ lze vypočítat shodně jako u rozšířeného Kalmanova filtru v sekci 4.1 a

$$\alpha_{0i} = \frac{\bar{\alpha}_{0i}}{\sum_{i=1}^{\xi_0} \bar{\alpha}_{0i}} \tag{5.2.7}$$

$$\xi_0 = \xi_0' \tag{5.2.8}$$

Tím je ukončen výpočet aproximační filtrační hustoty pravděpodobnosti. Přejděme nyní k prediktivní hustotě pravděpodobnosti pro $k = 1$.

Vyjděme z bayesovských rekurzivních vztahů. Pro prediktivní hustotu platí

$$p(x_1 \mid z^0) = \int p(x_0 \mid z^0) p(x_1 \mid x_0) dx_0 \tag{5.2.9}$$

Dosazením z (5.2.6) a (4.1.6) dostaneme

$$p_A(x_1 \mid z^0) = \int \sum_{i=1}^{\xi_1} \alpha_{0i} N(x_0 : \hat{x}_{0i}, P_{0i}) N(x_1 : f_0(x_0), Q_0) dx_0 \tag{5.2.10}$$

kde $\xi_1 = \xi_0$. Analogickou úvahou jako při stanovení filtrační hustoty (5.2.6), to jest realizací vícenásobné linearizace $f_0(x_0)$ dostaneme

$$\begin{aligned} p_A(x_1 \mid z^0) &= \int \sum_{i=1}^{\xi_1} \alpha_{0i} N(x_0 : \hat{x}_{0i}, P_{0i}) N(x_1 : f_0(\hat{x}_{0i}) \\ &+ F(\hat{x}'_{0i})(x_0 - \hat{x}'_{0i}), Q_0) dx_0 \end{aligned} \tag{5.2.11}$$

Funkce $f_0(x_0)$ byla ξ'_1-krát linearizována v bodech $\hat{x}_{0i}$, tedy

$$f_0(x_0) \simeq f_0(\hat{x}_{0i}) + F(\hat{x}'_{0i})(x_0 - \hat{x}'_{0i}), \quad i = 1, 2, ..., \xi_1$$

Ve vztahu (5.2.11) můžeme zaměnit pořadí integrace a sumace a za integrálem dostaneme pro i stejný výraz jako při odvození prediktivní hustoty v podkapitole 4.1. Výraz (5.2.11) můžeme rovnou upravit na

$$\begin{aligned} p_A(x_1 \mid z^0) &= \sum_{i=1}^{\xi_1} \alpha'_{0i} \int N(x_0 : \hat{x}_{0i}, P'_{0i}) N(x_1 : f_0(\hat{x}_{0i}) + F(\hat{x}'_{0i})(x_0 - \hat{x}_{0i}), Q_0) dx_0 \\ &= \sum_{i=1}^{\xi'_1} \alpha_{0i} N(x_1 : \hat{x}'_{1i}, P'_{1i}) \end{aligned} \tag{5.2.12}$$

kde $\alpha_{0i} = \alpha'_{0i}$, $\xi'_1 = \xi_1$ a $\hat{x}'_{1i}, P'_{1i}$ se získají z formálně shodných vztahů jako v podkapitole 4.1 u rozšířeného Kalmanova filtru.

Můžeme uzavřít, že jak (5.2.6), tak (5.2.12) jsou hustoty pravděpodobnosti stejné struktury jako hustota pravděpodobnosti počátečního stavu, a to vážený součet normálních rozložení. Předchozí postup tedy opět můžeme zobecnit pro jakýkoliv časový okamžik k. Shrňme dosažené výsledky a použijme již časový index k.

Aproximační filtrační hustota pravděpodobnosti:

$$p_k(x_k \mid z^k) = \sum_{i=1}^{\xi_k} \alpha_{ki} N(x_k : \hat{x}_{0i}, P_{ki}) \tag{5.2.13}$$

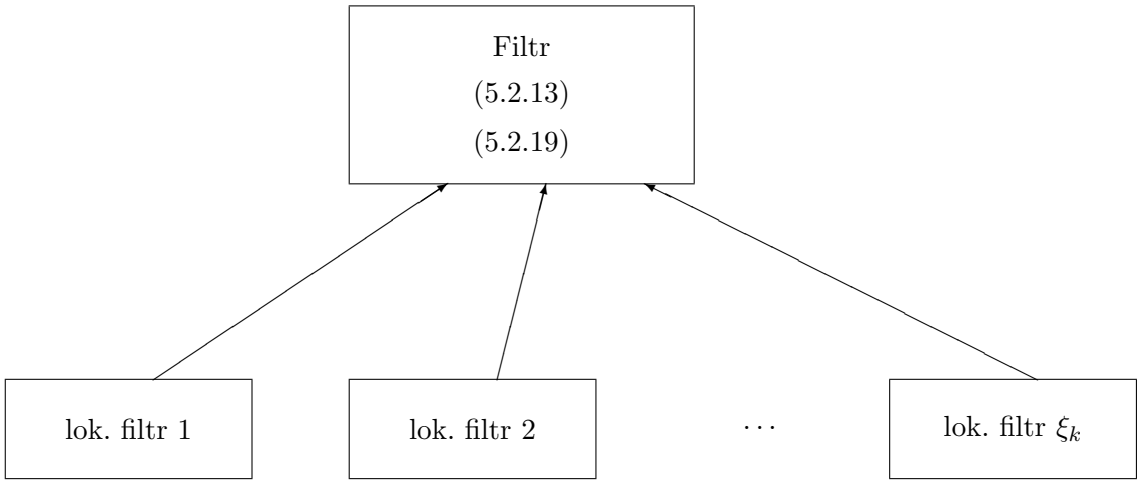
$$\hat{x}_{ki} = \hat{x}'_{ki} + P'_{ki} H^T(\hat{x}'_{ki}) [H(\hat{x}'_{ki}) P'_{ki} H^T(\hat{x}'_{ki}) + R_k]^{-1} [z_k - h_k(\hat{x}'_{ki})] \tag{5.2.14}$$

$$P_{ki} = P'_{ki} - P'_{ki} H^T(\hat{x}'_{ki}) [H(\hat{x}'_{ki}) P'_{ki} H^T(\hat{x}'_{ki}) + R_k]^{-1} H(\hat{x}'_{ki}) P'_{ki} \tag{5.2.15}$$

$$\zeta_{ki} = N(z_k : h_k(\hat{x}'_{ki}), H(\hat{x}'_{ki}) P'_{ki} H^T(\hat{x}'_{ki}) + R_k) \tag{5.2.16}$$

$$\alpha_{ki} = \frac{\alpha'_{ki} \zeta_{ki}}{\sum_{i=1}^{\xi'_k} \alpha'_{ki} \zeta_{ki}} \tag{5.2.17}$$

$$\xi_k = \xi'_k \tag{5.2.18}$$



Obrázek 5.2.1: Filtr s vícenásobnou linearizací

Aproximační prediktivní hustota pravděpodobnosti:

$$p_A(x_{k+1} \mid z^k) = \sum_{i=1}^{\xi'_{k+1}} \alpha'_{k+1,i} N(x_{k+1} : \hat{x}'_{k+1,i}, P'_{k+1,i}) \tag{5.2.19}$$

$$\hat{x}'_{k+1,i} = f_k(\hat{x}_{ki}) \tag{5.2.20}$$

$$P'_{k+1,i} = F(\hat{x}_{ki}) P_{ki} F^T(\hat{x}_{ki}) + Q_k \tag{5.2.21}$$

$$\alpha'_{k+1,i} = \alpha_{ki} \tag{5.2.22}$$

$$\xi'_{k+1} = \xi_k \tag{5.2.23}$$

Vztahy (5.2.13)-(5.2.23) definují filtr, jehož blokové schema je na obr. 5.2.1, kde je zřejmé, že informační tok je jednosměrný, a to od lokálních filtrů reprezentovaných rozšířenými Kalmanovými filtry směrem ke globálnímu filtru generujícího podmíněné hustoty pravděpodobnosti (5.2.13) a (5.2.19). Poznamenejme, že každý Kalmanův filtr vznikl linearizací nelinearit okolo jiného bodu. To je motivace pro název filtr s vícenásobnou linearizací. Poznamenejme, že tento filtr byl publikován bez odvození v [39] a v anglicky psané literatuře se můžeme setkat s názvy „Gaussian sum filter“ nebo „Gaussian mixture filter“. V literatuře se lze také setkat s přístupy k návrhu filtru s vícenásobnou linearizací využívající jiné aproximační techniky používané v lokálních filtrech, jako např. Stirlingovu interpolaci [93].

Představený filtr s vícenásobnou linearizací implicitně poskytuje odhady ve formě podmíněných hustot pravděpodobnosti. Pro následné využití odhadů je však často nutné vypočítat bodové odhady, např. střední hodnotu a příslušnou kovarianční matici chyby odhadu. Výpočet $\hat{x}_k, \hat{x}'_{k+1}, P_k, P'_{k+1}$, tedy globální střední hodnoty a kovariance, lze provést následujícím způsobem.

$$\begin{aligned}\hat{x}_k &= E[x_k \mid z^k] = \int x_k p_A(x_k \mid z^k) dx_k \\ &= \int x_k \sum_{i=1}^{\xi_k} \alpha_{ki} N(x_k : \hat{x}_{ki}, P_{ki}) dx_k \\ &= \sum_{i=1}^{\xi_k} \alpha_{ki} \hat{x}_{ki} \end{aligned} \tag{5.2.24}$$

$$\begin{aligned}\hat{x}_{k+1} &= E[x_{k+1} \mid z^k] = \int x_{k+1} p_A(x_{k+1} \mid z^k) dx_{k+1} \\ &= \int x_{k+1} \sum_{i=1}^{\xi'_{k+1}} \alpha'_{k+1} N(x_{k+1} : \hat{x}'_{k+1,i}, P'_{k+1,i}) dx_{k+1} \\ &= \sum_{i=1}^{\xi'_{k+1}} \alpha'_{k+1,i} \hat{x}'_{k+1,i} \end{aligned} \tag{5.2.25}$$

$$\begin{aligned}P_k &= E[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T \mid z^k] = \int (x_k - \hat{x}_k)(x_k - \hat{x}_k)^T p_A(x_k \mid z^k) dx_k \\ &= \int (x_k - \sum_{i=1}^{\xi_k} \alpha_{ki} \hat{x}_{ki})(x_k - \sum_{i=1}^{\xi_k} \alpha_{ki} \hat{x}_{ki})^T \sum_{i=1}^{\xi_k} \alpha_{ki} N(x_k : \hat{x}_{ki}, P_{ki}) dx_k \\ &= \sum_{i=1}^{\xi_k} \alpha_{ki} [\int x_k x_k^T N(x_k : \hat{x}_{ki}, P_{ki}) dx_k \\ &- \sum_{i=1}^{\xi_k} \alpha_{ki} \hat{x}_{ki} \int x_k^T N(x_k : \hat{x}_{ki}, P_{ki}) dx_k \\ &- \int x_k N(x_k : \hat{x}_{ki}, P_{ki}) dx_k (\sum_{i=1}^{\xi_k} \alpha_{ki} \hat{x}_{ki})^T + \hat{x}_k \hat{x}_k^T] \\ &= \sum_{i=1}^{\xi_k} \alpha_{ki} (P_{ki} + \hat{x}_{ki} \hat{x}_{ki}^T - \hat{x}_k \hat{x}_{ki}^T - \hat{x}_{ki} x_k^T + \hat{x}_k \hat{x}_k^T) \\ &= \sum_{i=1}^{\xi_k} \alpha_{ki} [P_{ki} + (\hat{x}_k - \hat{x}_{ki})(\hat{x}_k - \hat{x}_{ki})^T] \end{aligned} \tag{5.2.26}$$

$$\begin{aligned}P'_{k+1} &= E[(x_{k+1} - \hat{x}'_{k+1,i})(x_{k+1} - \hat{x}'_{k+1,i})^T \mid z^k] = ... \\ &= \sum_{i=1}^{\xi'_{k+1}} \alpha'_{k+1,i} [P'_{k+1,i} + (\hat{x}'_{k+1} - \hat{x}'_{k+1,i})(\hat{x}'_{k+1} - \hat{x}'_{k+1,i})^T] \end{aligned} \tag{5.2.27}$$

Na závěr této sekce proveďme zhodnocení odvozeného filtru:

1. Zvolené aproximace nelineárních funkcí $h_k(\cdot), f_k(\cdot)$ umožňují explicitní řešitelnost úlohy.

2. Lokální filtry jsou navzájem nezávislé. Avšak poznamenejme, že v literatuře lze najít verze filtru s vícenásobnou linearizací, kde globální odhad je propagován zpět k lokálním filtrům a tedy, lokální filtry již nejsou navzájem nezávislé [87]. V tomto případě se můžeme, v anglicky psané literatuře, setkat s pojmem „interacting multiple models“.

3. Jestliže $h_k(\cdot), f_k(\cdot)$ jsou lineární funkce, pak filtrační i prediktivní hustoty jsou vypočteny exaktně (bez jakýchkoliv aproximací), i když apriorní hustota pravděpodobnosti je negaussovská (součet normálních rozložení).

4. Aproximační filtrační a prediktivní hustoty pravděpodobnosti mají vlastnost reprodukovatelnosti, to jest v každém časovém okamžiku $k = 0, 1, 2, \ldots$ jsou dány váženým součtem normálních rozložení.

5. Jestliže hustota pravděpodobnosti počátečního stavu je gaussovská, pak filtr zdegeneruje na rozšířený Kalmanův filtr.

6. Výpočty $N(x_k : \hat{x}_{ki}, P_{ki})$ a $N(x_{k+1} : \hat{x}'_{k+1,i}, P'_{k+1,i})$ pro $i = 1, 2, ..., \xi_k$ mohou být realizovány sériově, ale i paralelně.

7. Kvůli linearizaci nelineárních funkcí jsou kovariance P_{ki} a $P'_{k+1,i}$ funkcemi měření.

8. Váhový koeficient α_{ki} je nelineární funkcí měření, neboť ζ_{ki} je nelineární funkcí měření. Jako důsledek, pak, na rozdíl od lokálních filtrů, je odhad stavu poskytovaný filtrem s vícenásobnou linearizací nelineární vzhledem k aktuálnímu měření.

5.3 Syntéza filtru pro lineární negaussovský systém

Vyjděme z předpokladu, že základní prostředek pro popis apriorních charakteristik náhodných proměnných je hustota pravděpodobnosti daná váhovým součtem normálních rozložení. Na rozdíl od předchozích dvou sekcí, kde jsme vážený součet normálních rozložení použili pro popis počátečního stavu, nyní ho použijeme i pro popis stavového šumu a šumu měření, tedy všech náhodných veličin vystupujících v modelu systému. Vážený součet normálních rozložení budeme opět chápat buď jako exaktně danou nebo jako pevnou strukturu s volnými parametry (váha, střední hodnota, kovariance) umožňující aproximaci libovolné hustoty pravděpodobnosti. Co se týče funkcí $f_k(\cdot), h_k(\cdot)$, budeme v této sekci budeme předpokládat, že jsou lineární. Na druhé straně však v porovnání s podkapitolou 5.2 zde uvažovaný problém je obecnější z pohledu předpokladů na šumy.

Definujme lineární negaussovský systém následujícími vztahy a vlastnostmi:

$$x_{k+1} = F_k x_k + w_k \tag{5.3.1}$$

$$z_k = H_k x_k + v_k \tag{5.3.2}$$

$$p(x_0 \mid z^{-1}) = \sum_{i=1}^{\xi'_0} \alpha'_{0i} N(x_0 : \hat{x}_{0i}, P'_{0i}) \tag{5.3.3}$$

$$p(w_k) = \sum_{n=1}^{q_k} \beta_{kn} N(w_k : \hat{w}_{kn}, Q_{kn}) \tag{5.3.4}$$

$$p(v_k) = \sum_{m=1}^{r_k} \gamma_{km} N(v_k : \hat{v}_{km}, R_{km}) \tag{5.3.5}$$

Opět předpokládáme, že $\{w_k\}$ a $\{v_k\}$ jsou bílé šumy, vzájemně nezávislé a nezávislé rovněž na x_0.

Pro estimaci stavu použijeme opět bayesovský přístup. Cílem je určit podmíněné hustoty pravděpodobnosti stavu. Začneme opět, jako ve všech předchozích odvozeních, v okamžiku $k = 0$ výpočtem filtrační hustoty, která vznikne z apriorní $p(x_0 \mid z^{-1})$ a zpracováním měření z_0. Pro výpočet filtrační hustoty potřebujeme nejdříve vyjádřit hustotu pravděpodobnosti měření. Tu můžeme v tomto případě snadno vyjádřit

$$p(z_0 \mid x_0) = \sum_{m=1}^{r_k} \gamma_{km} N(z_0 : H_0 x_0 + \hat{v}_{0m}, R_{0m}) \tag{5.3.6}$$

Dosazením (5.3.3) a (5.3.6) do vztahu pro filtrační hustotu a roznásobením dostaneme

$$\begin{aligned}
p(x_0 \mid z^0) &= \frac{[\sum_{i=1}^{\xi_0'} \alpha_{0i}' N(x_0 : \hat{x}_{0i}', P_{0i}')][\sum_{m=1}^{r_k} \gamma_{0m} N(z_0 : H_0 x_0 + \hat{v}_{0m}, R_{0m}]}{p(z_0 \mid z^{-1})} \\
&= \frac{\sum_{i=1}^{\xi_0'} \alpha_{0i}' \gamma_{01} N(x_0 : \hat{x}_{0i}' P_{0i}') N(z_0 : H_0 x_0 + \hat{v}_{01}, R_{01})}{p(z_0 \mid z^{-1})} \\
&+ \frac{\sum_{i=1}^{\xi_0'} \alpha_{0i}' \gamma_{02} N(x_0 : \hat{x}_{0i}', P_{0i}') N((z_0 : H_0 x_0 + \hat{v}_{02}, R_{02})}{p(z_0 \mid z^{-1})} \\
&\vdots \\
&+ \frac{\sum_{i=1}^{\xi_0'} \alpha_{0i}' \gamma_{0r_0} N(x_0 : \hat{x}_{0i}', P_{0i}') N(z_0 : H_0 x_0 + \hat{v}_{0r_0}, R_{0r_0})}{p(z_0 \mid z^{-1})}
\end{aligned} \tag{5.3.7}$$

Každý z těchto r_0 zlomků v (5.3.7) je formálně a obsahově shodný s výrazy v (3.2.4) uvažovaných při syntéze Kalmanova filtru. Proto můžeme převzít výsledky z podkapitoly 3.2 pro vyjádření zlomků v předchozím vztahu, čímž získáme

$$\begin{aligned}
p(x_0 \mid z^0) &= \frac{\sum_{i=1}^{\xi_0'} \alpha_{0i}' \gamma_{01} N(z_0 : H_0 \hat{x}_{0i}' + \hat{v}_{01}, H_0 P_{0i}' H_0^T + R_{01}) N(x_0 : \hat{x}_{0i}, P_{0i})}{p(z_0 \mid z^{-1})} \\
&+ \frac{\sum_{i=1}^{\xi_0'} \alpha_{0i}' \gamma_{02} N(z_0 : H_0 \hat{x}_{0i}' + \hat{v}_{02}, H_0 P_{0i}' H_0^T + R_{02}) N(x_0 : \hat{x}_{0(i+\xi_0')}, P_{0(i+\xi_0')})}{p(z_0 \mid z^{-1})} \\
&\vdots \\
&+ \frac{\sum_{i=1}^{\xi_0'} \alpha_{0i}' \gamma_{0r_0} N(z_0 : H_0 \hat{x}_{0i}' + \hat{v}_{0r_0}, H_0 P_{0i}' H_0 + R_{0r_0}) N(x_0 : \hat{x}_{0[i+\xi_0'(r_0-1)]}, P_{0[i+\xi_0'(r_0-1)]})}{p(z_0 \mid z^{-1})}
\end{aligned} \tag{5.3.8}$$

kde $p(z_0 \mid z^{-1})$ představuje normalizační konstantu.

Ze vztahu (5.3.8) vyplývá, že $p(x_0 \mid z^0)$ je opět vážený součet normálních rozdělení a počet členů v tomto součtu je dán součinem počtu členů v apriorní hustotě pravděpodobnosti a počtu

členů figurujících v hustotě pravděpodobnosti měření. Tedy

$$\xi_0 \stackrel{\triangle}{=} \xi_0' r_0 \tag{5.3.9}$$

Kvůli ujednodušení zápisu (5.3.8) definujme

$$\zeta_{0j} \stackrel{\triangle}{=} N(z_0 : H_0 \hat{x}_{0i}' + \hat{v}_{0m}', H_0 P_{0i}' H_0^T + R_{0m}) \tag{5.3.10}$$

$$\alpha_{0j} \stackrel{\triangle}{=} \frac{\alpha_{0i}' \gamma_{0m} \zeta_{0j}}{\sum_{j=1}^{\xi_0} \alpha_{0i}' \gamma_{0m} \zeta_{0j}} \tag{5.3.11}$$

kde indexy i a m jsou v následujícím vztahu k j

$$i = j - IFIX(\frac{j-1}{\xi_0'})\xi_0', \quad i = 1, 2, ..., \xi_0' \tag{5.3.12}$$

$$m = 1 + IFIX(\frac{j-1}{\xi_0'}), \quad m = 1, 2, ..., r_0 \tag{5.3.13}$$

kde funkce $IFIX$ ponechává celou část čísla v závorce. Platnost (5.3.12) a (5.3.13) je dostatečně zřejmá ze vztahů (5.3.7) a (5.3.8).

Nyní již můžeme (5.3.8) zapsat přehledným způsobem, a to

$$p(x_0 \mid z^0) = \sum_{j=1}^{\xi_0} \alpha_{0j} N(x_0 : \hat{x}_{0j}, P_{0j}) \tag{5.3.14}$$

kde ξ_0 je dáno (5.3.9), α_{0j} (5.3.11) a střední hodnoty $\hat{x}_{0j}$ a kovariance P_{0j} jsou analogie vztahů v podkapitole 3.2, to jest

$$\hat{x}_{0j} = \hat{x}_{0i}' + P_{0i}' H_0^T [H_0 P_{0i}' H_0^T + R_{0m}]^{-1} [z_0 - H_0 \hat{x}_{0i} - \hat{v}_{0m}] \tag{5.3.15}$$

$$P_{0j} = P_{0i}' - P_{0i}' H_0^T [H_0 P_{0i}' H_0^T + R_{0m}]^{-1} H_0 P_{0i} \tag{5.3.16}$$

a indexy se mění podle (5.3.12), (5.3.13).

Tím je výpočet filtrační hustoty pravděpodobnosti ukončen. Můžeme konstatovat, že filtrační hustota má stejnou strukturu jako apriorní hustota pravděpodobnosti.

Dále pokračujme v syntéze estimátoru výpočtem prediktivní hustoty pravděpodobnosti $p(x_1 \mid z^0)$. Potřebujeme znát přechodovou hustotu pravděpodobnosti. Z (5.3.1) a (5.3.4) je však zřejmé, že

$$p(x_1 \mid x_0) = \sum_{n=1}^{q_0} \beta_{0n} N(x_1 : F_0 x_0 + \hat{v}_{0n}, Q_{0n}) \tag{5.3.17}$$

Dosazením (5.3.17) a (5.3.14) do (3.2.29) dostaneme

$$p(x_1 \mid z^0) = \int [\sum_{j=1}^{\xi_0} \alpha_{0j} N(x_0 : \hat{x}_{0j}, P_{0j})][\sum_{n=1}^{q_0} \beta_{0n} N(x_1 : F_0 x_0 + \hat{w}_{0n}, Q_{0n}]dx_0 \tag{5.3.18}$$

Tento výraz může být rozepsán

$$\begin{aligned} p(x_1 \mid z^0) = & \sum_{j=1}^{\xi_0} \alpha_{0j}\beta_{01} \int N(x_0 : \hat{x}_{0j}, P_{0j}) N(x_1 : F_0 x_0 + \hat{w}_{01}, Q_{01}) dx_0 \\ & + \sum_{j=1}^{\xi_0} \alpha_{0j}\beta_{02} \int N(x_0 : \hat{x}_{0j}, P_{0j}) N(x_1 : F_0 x_0 + \hat{w}_{02}, Q_{02}) dx_0 \\ & \vdots \\ & + \sum_{j=1}^{\xi_0} \alpha_{0j}\beta_{0q_0} \int N(x_0 : \hat{x}_{0j}, P_{0j}) N(x_1 : F_0 x_0 + \hat{w}_{0q_0}, Q_{0q_0}) dx_0 \end{aligned} \tag{5.3.19}$$

Každý z těchto q_0 součtů je obsahově shodný s (3.2.29) vedoucí k (3.2.43), a proto (5.3.19) můžeme přímo upravit na tvar

$$\begin{aligned} p(x_1 \mid z^0) &= \sum_{j=1}^{\xi_0} \alpha_{0j}\beta_{01} N(x_1 : \hat{x}'_{1j}, P'_{1j}) \\ &+ \sum_{j=1}^{\xi_0} \alpha_{0j}\beta_{02} N(x_1 : \hat{x}'_{1,j+\xi_0} + \xi_0, P'_{1,j+\xi_0}) \\ &\vdots \\ &+ \sum_{j=1}^{\xi_0} \alpha_{0j}\beta_{0q_0} N(x_1 : \hat{x}'_{1,j+[\xi_0(q_0-1)]}, P'_{1,j+[\xi_0(q_0-1)]}) \end{aligned} \tag{5.3.20}$$

To znamená, že prediktivní hustota pravděpodobnosti má stejnou strukturu jako filtrační hustota, protože se jedná opět o vážený součet normálních rozložení. Výraz (5.3.20) můžeme zřejmě přepsat na tvar

$$p(x_1 \mid z^0) = \sum_{i=1}^{\xi'_1} \alpha'_{1i} N(x_1 : \hat{x}'_{1i}, P'_{1i}) \tag{5.3.21}$$

kde

$$\xi'_1 = \xi_0 q_0 \tag{5.3.22}$$

$$\alpha'_{1i} = \alpha_{0j}\beta_{0n} \tag{5.3.23}$$

$$\hat{x}'_{1i} = F_0 \hat{x}_{0j} + \hat{w}_{0n} \tag{5.3.24}$$

$$P'_{1i} = F_0 P_{0j} F_0^T + Q_{0n} \tag{5.3.25}$$

Indexy j, n jsou v následujícím vztahu k indexu i

$$j = i - IFIX(\frac{i-1}{\xi_0})\xi_0, \quad j = 1, 2, ..., \xi_0 \tag{5.3.26}$$

$$n = 1 + IFIX(\frac{i-1}{\xi_0}), \quad n = 1, 2, ..., q_0 \tag{5.3.27}$$

Výpočet prediktivní hustoty pravděpodobnosti je tímto ukončen. Filtrační a prediktivní hustoty byly vypočteny bez jakýchkoliv aproximací a patří do stejné třídy jako apriorní hustota. Z toho vyplývá, že dosažené výsledky můžeme formálně zobecnit pro libovolné k. Pro přehlednost opět shrňme výsledný algoritmus estimace a zapišme ho již pro časový okamžik k.

Filtrační hustota pravděpodobnosti $p(x_k \mid z^k)$ pro $k = 0, 1, ...$

$$p(x_k \mid z^k) = \sum_{j=1}^{\xi_k} \alpha_{kj} N(x_k : \hat{x}_{kj}, P_{kj}) \tag{5.3.28}$$

$$\hat{x}_{kj} = \hat{x}'_{ki} + P'_{ki} H_k^T [H_k P'_{ki} H_k^T + R_{km}]^{-1} [z_k - H_k \hat{x}_{ki} - \hat{v}_{km}] \tag{5.3.29}$$

$$P_{kj} = P'_{ki} - P'_{ki} H_k^T [H_k P'_{ki} H_k^T + R_{km}]^{-1} H_k P'_{ki} \tag{5.3.30}$$

$$i = j - IFIX(\frac{j-1}{\xi'_k})\xi'_k, \quad i = 1, 2, ..., \xi'_k \tag{5.3.31}$$

$$m = 1 + IFIX(\frac{j-1}{\xi'_k}), \quad m = 1, 1, ..., r_k \tag{5.3.32}$$

$$\zeta_{kj} = N(z_k : H_k \hat{x}'_{ki} + \hat{v}_{km}, H_k P'_{ki} H_k^T + R_{km}) \tag{5.3.33}$$

$$\alpha_{kj} = \frac{\alpha'_{ki} \gamma_{km} \zeta_{kj}}{\sum_{j=1}^{\xi_k} \alpha'_{ki} \gamma_{km} \zeta_{kj}} \tag{5.3.34}$$

$$\xi_k = \xi'_k r_k \tag{5.3.35}$$

přičemž

$$p(x_k \mid z^{-1}) = \sum_{i=1}^{\xi'_k} \alpha'_{ki} N(x_k : \hat{x}'_{ki}, P'_{ki})$$

je pro $k = 0$ známá apriorní hustota $p(x_0 \mid z^{-1})$ a pro $k > 0$ bude vypočtena, a tudíž rovněž známá.

Prediktivní hustota pravděpodobnosti $p(x_{k+1} \mid z^k)$ pro $k = 0, 1, 2, ...$

$$p(x_{k+1} \mid z^k) = \sum_{i=1}^{\xi'_{k+1}} \alpha'_{k+1,i} N(x_{k+1} : \hat{x}'_{k+1,i}, P'_{k+1,i}) \tag{5.3.36}$$

$$\hat{x}'_{k+1,i} = F_k \hat{x}_{kj} + \hat{w}_{kn} \tag{5.3.37}$$

$$P'_{k+1,i} = F_k P_{kj} F_k^T + Q_{kn} \tag{5.3.38}$$

$$j = i - IFIX(\frac{i-1}{\xi_k})\xi_k, \quad j = 1, 2, ..., \xi_k \tag{5.3.39}$$

$$n = 1 + IFIX(\frac{i-1}{\xi_k}), \quad n = 1, 2, ..., q_k \tag{5.3.40}$$

$$\alpha'_{k+1,i} = \alpha_{kj}\beta_{kn} \tag{5.3.41}$$

$$\xi'_{k+1} = \xi_k q_k \tag{5.3.42}$$

přičemž $p(x_k \mid z^k) = \sum_{j=1}^{\xi_k} \alpha_{kj} N(x_k : \hat{x}_{kj}, P_{kj})$ je vypočtená hustota. V okamžiku výpočtu prediktivní hustoty je známá.

Jaké jsou základní vlastnosti tohoto nelineárního estimátoru (5.3.28)-(5.3.42).

1. Výpočet hustot pravděpodobnosti je exaktní, neprovádí se žádná aproximace.
2. Pro $q_k = r_k = \xi'_0 = 1$ pro všechna k estimátor přejde na Kalmanův filtr.
3. Estimátor je teoretickým a praktickým mostem mezi lineární a nelineární filtrací. Umožňuje vnitřní detailní pohled na vztah mezi formulací úlohy a jejím řešením.
4. Vytváří teoretický základ pro řešení nejenom lineárního negaussovského problému, ale i pro další úlohy, které budou součástí šesté a sedmé kapitoly.
5. V obecném případě, tj. pro $r > 1$ a $q > 1$, počet členů ve váženém součtu normálních rozložení definující hustoty pravděpodobnosti roste. Redukcí počtu členů se zabývá např. [49], [52].

5.4 Syntéza filtru pro nelineární negaussovský systém

V této sekci se zaměříme na syntézu filtru pro systém, který na rozdíl od podkapitoly 5.3, obsahuje nelinearity ve stavové rovnici případně v rovnici měření. Předchozí podkapitoly pak mohou být chápány jako speciální případy této úlohy.
Pro syntézu filtru použijeme postup obdobný jako při návrhu nelineárního filtru z (5.2) a (5.3). Uvažujme tedy následující nelineární negaussovský systém.

$$x_{k+1} = f_k(x_k) + w_k \tag{5.4.1}$$

$$z_k = h_k(x_k) + v_k \tag{5.4.2}$$

$$p(x_0 \mid z^{-1}) = \sum_{i=1}^{\xi'_0} \alpha'_{0i} N(x_0 : \hat{x}_{0i}, P'_{0i}) \tag{5.4.3}$$

$$p(w_k) = \sum_{n=1}^{q_k} \beta_{kn} N(w_k : \hat{w}_{kn}, Q_{kn}) \tag{5.4.4}$$

$$p(v_k) = \sum_{m=1}^{r_k} \gamma_{km} N(v_k : \hat{v}_{km}, R_{km}) \tag{5.4.5}$$

Náhodné procesy $\{w_k\}$ a $\{v_k\}$ jsou bílé, vzájemně nezávislé a nezávislé na počátečním stavu x_0. Protože jsme provedli v sekci 5.2 odvození filtru pro nelineární gaussovský systém a v 5.3 pro lineární negaussovský systém, nebudeme zde již odvozovat filtr pro systém (5.4.1)-(5.4.5) a pouze provedeme kombinaci výsledků z 5.2 a 5.3. Rovnou pak dostaneme filtrační aproximační hustotu pro krok k.
Filtrační aproximační hustota pravděpodobnosti $p_A(x_k \mid z^k)$ pro $k = 0, 1, 2, ...$

$$p_A(x_k \mid z^k) = \sum_{j=1}^{\xi_k} \alpha_{kj} N(x_k : \hat{x}_{kj}, P_{kj}) \tag{5.4.6}$$

$$\hat{x}_{kj} = \hat{x}'_{ki} + P'_{ki} H_k^T(\hat{x}'_{ki}) [H_k(\hat{x}'_{ki}) P'_{ki} H_k^T(\hat{x}_{ki}) + R_{km}]^{-1} [z_k - h_k(\hat{x}'_{ki}) - \hat{v}_{km}] \tag{5.4.7}$$

$$P_{kj} = P'_{ki} - P'_{ki} H_k^T(\hat{x}'_{ki}) [H_k(\hat{x}'_{ki}) P'_{ki} H_k^T(\hat{x}_{ki}) + R_{km}]^{-1} H_k(\hat{x}'_{ki}) P'_{ki} \tag{5.4.8}$$

kde kombinace indexů je stejná jako v předchozí sekci. Rovněž $H_k(\cdot)$ je definováno stejně jako v podkapitole 4.1.

Prediktivní hustota pravděpodobnosti $p_A(x_{k+1} \mid z^k)$ pro $k = 0, 1, 2, ...$

$$p_A(x_{k+1} \mid z^k) = \sum_{i=1}^{\xi'_{k+1}} \alpha'_{k+1,i} N(x_{k+1} : \hat{x}'_{k+1,i}, P'_{k+1,i}) \tag{5.4.9}$$

$$\hat{x}'_{k+1,i} = f_k(\hat{x}_{kj}) + \hat{w}'_{kn} \tag{5.4.10}$$

$$P'_{k+1,i} = F_k(\hat{x}_{kj}) P_{kj} F_k^T(\hat{x}_{kj}) + Q_{kn} \tag{5.4.11}$$

kde kombinace indexů je stejná jako v předchozí sekci. Rovněž $F_k(\cdot)$ je definováno stejně jako v (4.1.14).

Podkapitola 5.4 uzavírá výklad a analytické odvození nelineárních filtrů jak pro lineární negaussovské systémy, tak pro nelineární negaussovské systémy, kde hustota pravděpodobnosti počáteční podmínky a poruch je ve speciální struktuře směsi normálních rozdělení. Tvorbě modelu směsových hustot pravděpodobnosti a příkladům použití je věnována následující kapitola. Uvažovaný popis umožňuje navrhnout estimační algoritmy jejichž činnost je průhledná a dobře interpretovatelná. Významnou předností je možnost plynulého přechodu od jednotlivých úloh estimace spojených s odhadem stavu lineárních gaussovských systémů až k úlohám estimace stavu nelineárních negaussovských systémů, které jsou součástí problému nelineární filtrace.
Technika a základní myšlenky návrhu nelineárních filtrů mohou být použity i při syntéze duálních regulátorů [53], optimálního řízení [54], [55] a predikce [57]. Alternativní obecné

postupy k aproximaci Bayesova pravidla a k rekurzivní nelineární estimaci jsou prezentovány např. [57], [58], [59], [60].

Kapitola 6

Modelování specifických jevů a estimace stavu

Předchozí kapitoly obsahují množství algoritmů estimace pro situace, kdy je potřebné překročit při modelování systémů či procesů rámec linearity a gaussovosti. V této kapitole se zaměříme na několik možných aplikací negaussovských modelů pro formulaci a řešení úloh významných z praktického hlediska.

Především se budeme věnovat problémům spojených s odhadem skokově (hrubě) se měnících parametrů či stavu [52], [61], [63] a syntézou robustních estimátorů s ohledem na hrubé, avšak řídce se vyskytující chyby měření [62], [63], [64], [65].
Bude ukázáno, že předcházející kapitoly obsahují nejenom vhodný aparát pro popis takových jevů, ale zároveň předkládají i analytické řešení estimačních úloh, které je i z inženýrského hlediska pro tyto i další situace přijatelné.

6.1 Modelování skokových změn stavu

Předpokládejme, že hustota pravděpodobnosti stavového šumu je dána součtem gaussovských rozložení

$$p(w_k) = \sum_{n=1}^{2} \beta_{kn} N(w_k : \hat{w}_{kn}, Q_{kn}) \tag{6.1.1}$$

Na rozdíl od [39] předpokládejme, že hustota pravděpodobnosti (6.1.1) není aproximace nějaké jiné, známé hustoty (např. rovnoměrného rozložení), ale že tato funkce je vytvořena takovým způsobem, že umožní popsat hrubé změny stavu. Konkretizujme tuto myšlenku.
Předpokládejme, že skaláry β_{k1}, β_{k2} vyjadřují pravděpodobnost, že nastane určitý jev b_{k1}, resp. b_{k2}. Náhodná veličina w_k je tímto jevem podmíněna a je popsána podmíněnou hustotou pravděpodobnosti. Tedy

$$p(w_k) = \sum_{n=1}^{2} p(w_k \mid b_{kn}) Pr(b_{kn}) \tag{6.1.2}$$

kde $Pr(b_{kn})$ znamená pravděpodobnost jevů b_{kn} a $p(w_k \mid b_{kn})$ je příslušná podmíněná hustota pravděpodobnosti.

Vztah (6.1.1) můžeme tedy chápat jako marginalizaci hustoty pravděpodobnosti pro spojitou náhodnou veličinu w_k a diskrétní náhodnou veličinu b_k. Alternativně vyjádřeno, z (6.1.2) a (6.1.1) vyplývá, že w_k je popsána normálním rozložením $N(w_k : \hat{w}_{k1}, Q_{k1})$ s pravděpodobností $Pr(b_1) = \beta_{k1}$ a rozložením $N(w_k : \hat{w}_{k2}, Q_{k2})$ s pravděpodobností $Pr(b_2) = \beta_{k2}$. Tedy w_k je náhodná veličina, které přísluší jisté rozložení s jistou pravděpodobností:

$w_k \to N(w_k : \hat{w}_{k1}, Q_{k1})$ s pravděpodobností β_{k1}
$w_k \to N(w_k : \hat{w}_{k2}, Q_{k2})$ s pravděpodobností β_{k2}.
$\beta_{k1} + \beta_{k2} = 1$

Tato interpretace (6.1.1) nám nyní snadno umožní postavit model šumu reprezentující skokové změny stavu.
Definujme β_{k1} jako pravděpodobnost standardní „malé“ nejistoty ovlivňující vývoj stavu a β_{k2} jako pravděpodobnost výskytu „velkých“ poruch působících následně výrazné změny stavu. To znamená, že trojice $\beta_{k1}, \hat{w}_{k1}$ a Q_1 by měla reprezentovat mírné poruchy ($\beta_{k1} \to 1$, Q_{k1} je „malá kovariance“), zatímco trojice $\beta_{k2}, \hat{w}_{k1}, Q_{k2}$ odpovídá potenciálně výrazným změnám ($\beta_{k2} \to 0, Q_{k2}$ je „velká“ kovariance) při $\hat{w}_{k1} = \hat{w}_{k2}$.
Samozřejmě, počet členů v sumě (6.1.1) může být zvýšen na $3, 4, \ldots$ podle situace a interpretace se příslušně změní.
Zavedení diskrétní náhodné veličiny nám umožnilo konstruovat model poruchy s přirozenou interpretací a využít ho pro modelování skokových změn stavu.

6.2 Modelování hrubých chyb měření

V této sekci se budeme zabývat modelováním hrubých chyb měření. Postup bude formálně analogický jako v podkapitole 6.1.
Předpokládejme, že hustota pravděpodobnosti měření může být popsána

$$p(v_k) = \sum_{m=1}^{2} \gamma_{km} N(v_k : \hat{v}_{km}, R_{km}) \tag{6.2.1}$$

Návrhář tohoto modelu může definovat γ_{k1} jako pravděpodobnost běžné poruchy měření s malou neurčitostí vyjádřenou kovariancí R_{k1} a γ_{k2} jako pravděpodobnost zřídka se vyskytující „velké“ poruchy (v anglicky psané literatuře označované jako „outliers“) vyjádřenou kovariancí R_{k2}. Neboli obdobně jako v (6.1)

$$p(v_k) = \sum_{m=1}^{2} p(v_k \mid c_m) Pr(c_m) \tag{6.2.2}$$

kde $Pr(c_m) = \gamma_{km}$ je pravděpodobnost vzniku jevu c_m a $p(v_k \mid c_m) = N(v_k : \hat{v}_{km}, R_{km})$ je příslušná podmíněná hustota pravděpodobnosti.

Trojice $\gamma_{k1}, \hat{v}_{k1}, R_{k1}$ a $\gamma_{k2}, \hat{v}_{k2}, R_{k2}$ pak modelují běžné poruchy měření a zřídka se vyskytující hrubé chyby. Samozřejmě počet členů v sumě se může zvětšit na $3, 4, \ldots$ V takovém případě by se změnila i odpovídajícím způsobem interpretace.
Poznamenejme, že tento způsob modelování poruchy měření zahrnuje i další možnosti kromě modelování hrubých chyb. Všimněme si např. možnosti zavést střední hodnoty $\hat{v}_{km} = \hat{v}_m$,

kde $\hat{v}_m$ může nabývat konečného počtu hodnot a $R_{km} = R_m = R = 0$ pro všechna k. Tímto způsobem bychom vlastně modelovali přítomnost neznámého konstantního signálu (o konečném počtu úrovní) v měření. Podobně i v podkapitole 6.1 lze využít v (6.1.1) střední hodnoty stavového šumu pro modelování různě velikých skoků.

6.3 Popis počátečního stavu

Pokračujme ve stejném duchu jako v podkapitolách 6.1 a 6.2. Uvažujme popis počátečního stavu ve formě

$$p(x_0) = \sum_{i=1}^{2} \alpha'_{0i} N(x_0 : \hat{x}'_{0i}, P'_{0i}), \tag{6.3.1}$$

kde opět $p(x_0)$ není aproximace nějaké známé hustoty počátečního stavu, ale α'_{01} je pravděpodobnost, že náhodná veličina x_0 odpovídá rozložení $N(x_0, \hat{x}'_{01}, P'_{01})$ a α'_{02} je pravděpodobnost, že x_0 odpovídá rozložení $N(x_0 : \hat{x}'_{02}, P'_{02})$.

Tuto skutečnost bychom přesněji mohli zapsat takto

$$p(x_0) = p(x_0 \mid z^{-1}) = \sum_{i=1}^{2} p(x_0 \mid a_i, z^{-1}) Pr(a_i \mid z^{-1}) \tag{6.3.2}$$

kde $Pr(a_i \mid z^{-1})$ je podmíněná pravděpodobnost výskytu jevu a_i,
$p(x_0 \mid a_i, z^{-1})$ je podmíněná hustota pravděpodobnosti x_0.

Poznamenejme, že počet členů v sumě z (6.3.1) může být zvětšen na 3,4,... a interpretován podle konkrétní volby trojice $(\alpha'_{0i}, \hat{x}'_{0i}, P'_{0i})$.
Nyní je zřejmé, že hustoty pravděpodobnosti v (6.1.1), (6.2.1) i v (6.3.1) mají stejnou formální strukturu a umožňují velmi flexibilní interpretaci pro různé typy úloh. Např. jevy $a_i,\ i = 1, 2, ..., \xi'_0$ mohou být spojeny s předpoklady o systému reprezentovaném modely M_i jak bude ukázáno v sedmé kapitole.

6.4 Aproximace hustoty pravděpodobnosti směsí normálních rozdělení

V předchozích sekcích byla pozornost upřena na situace, kdy lze poruchu ve stavu či měření nebo počáteční podmínku přesně modelovat směsí normálních rozložení. Mnohdy však hustota pravděpodobnosti poruch či počáteční podmínky je popsána jinou specifickou hustotou pravděpodobnosti. Jak příklad můžeme uvést Studentovo rozdělení či šikmé Studentovo rozdělení vyskytující se úlohách sledování [114]. V těchto situacích však můžeme aproximovat obecnou hustotu pravděpodobnosti směsí normálních rozložení, tj. např. při uvažování obecné hustoty stavového šumu $p(w_k)$ hledáme následující aproximaci

$$p(w_k) \approx \sum_{n=1}^{N_w} \beta_{kn} N(w_k : \hat{w}_{kn}, Q_{kn}) \tag{6.4.1}$$

kde hledané parametry jsou váhy β_{kn}, dílčí střední hodnoty $\hat{w}_{kn}$ a dílčí kovarianční matice Q_{kn}. V literatuře lze najít spoustu přístupů a metod k výpočtu hledaných parametrů směsi normálních rozdělení. Mezi všemi lze zmínit přístup založený na tzv. EM algoritmu (z anglického „expectation-maximization“) [115], který hledá optimální odhad, ve smyslu maximální věrohodnosti, parametrů směsové hustoty na základě náhodně generovaných vzorků z původní aproximované hustoty. Pro tento přístup je dostupný i toolbox pro prostředí MATLAB®. Jako alternativu můžeme uvést funkci „fit“ přímo z programového prostředí MATLAB®. Poznamenejme závěrem, že odhad parametrů směsové hustoty (6.4.1) se neobejde bez specifikace počtu členů N_w směsové hustoty a mnohdy i počátečních odhadů dílčích středních hodnot a kovariančních matic. Za zmínku taktéž stojí fakt, že obecnou hustotu pravděpodobnosti je možné aproximovat směsí normálních rozdělení s *libovolnou* přesností.

6.5 Modelování dalších specifických jevů

Doposud jsme se v této kapitole zabývali alternativními možnostmi interpretace formálního aparátu sloužícího jak pro popis stochastických systémů, tak pro analytický návrh estimačních algoritmů, který byl detailně formulován a používán v předchozích kapitolách.
Jak vyplývá ze podkapitol 6.1–6.4 týkajících se modelování skokových změn stavu, hrubých chyb měření a dalších úloh, je možné využít takové postupy, které jsou speciálním případem úlohy modelování a syntézy filtru pro lineární negaussovský systém (5.3.1)–(5.3.5) formulované a řešené v 5.3. Ukažme si, jak by mohli být v rámci podkapitoly 5.3 tyto speciální případy formulovány.

Odhad (sledování) skokově se měnícího stavu.
Uvažujme (5.3.1)–(5.3.5) např. s těmito specifikacemi pro $k = 0, 1, 2, \ldots$

$$F_k = 1, \qquad H_k = 1 \tag{6.5.1}$$

$$r_k = 1 \tag{6.5.2}$$

$$q_k = 2 \tag{6.5.3}$$

$$\beta_{k1} = 0.98, \qquad \beta_{k2} = 0.02 \tag{6.5.4}$$

$$Q_{k1} = 10^{-4}, \qquad Q_{k2} = 0.5 \tag{6.5.5}$$

$$\hat{w}_{k1} = 0, \qquad \hat{w}_{k2} = 0 \tag{6.5.6}$$

$$\hat{v}_k = 0, \qquad R_k = 10^{-4} \tag{6.5.7}$$

$$\hat{x}'_0 = 0, \qquad P'_0 = 100 \tag{6.5.8}$$

Pak estimační algoritmus (5.3.28)–(5.3.42) umožňuje získat odhad stavu. Jestliže budeme redukovat v každém estimačním kroku počet členů v (5.3.28) na $\xi_k = 1$, např. podle maximální hodnoty α_{kj}, bude algoritmus velmi jednoduchý, ale přesto rychle reagující na prudké změny stavu. Připomeňme, že α_{kj} zde reprezentuje pravděpodobnost jevu a_j a přísluší výskytu

náhodné veličiny x_k s rozložením $N(x_k : \hat{x}_{kj}, P_{kj})$.

Hrubé chyby měření. Uvažujme (5.3.1)–(5.3.5) např. s tím, že pro $k = 0, 1, 2, \ldots$ je dáno

$$F_k = 1, \qquad H_k = 1 \tag{6.5.9}$$

$$r_k = 2 \tag{6.5.10}$$

$$q_k = 1 \tag{6.5.11}$$

$$\gamma_{k1} = 0.98, \qquad \gamma_{k2} = 0.02 \tag{6.5.12}$$

$$R_{k1} = 10^{-4}, \qquad R_{k2} = 1 \tag{6.5.13}$$

$$\hat{v}_{k1} = 0, \qquad \hat{v}_{k2} = 0 \tag{6.5.14}$$

$$\hat{w}_k = 0, \qquad Q_k = 10^{-4} \tag{6.5.15}$$

$$\hat{x}'_0 = 0, \qquad P'_0 = 100 \tag{6.5.16}$$

Pak estimační algoritmus (5.3.28)–(5.3.42) umožňuje získat odhad stavu. Je opět výhodné provádět redukci počtu členů v (5.3.28) na $\xi_k = 1$ podle maximální α_{kj} se stejnou interpretací jako v předchozí úloze. Poznamenejme, že v tomto případě estimátor je robustní, necitlivý na hrubé chyby měření, jelikož v případě, že detekuje hrubou poruchu, zisk estimátoru se zmenší, a tudíž měřená veličina obsahující mylnou informaci není brána algoritmem "v úvahu".

Různá četnost a charakter skokových (hrubých) změn stavu.
Uvažujme (5.3.1)–(5.3.5) s tím, že chceme specifikovat parametry tohoto modelu tak, aby bylo možné vyjádřit hrubé změny pro každou stavovou proměnnou individuálně, a to jak s ohledem na četnost takových změn, tak na velikost těchto změn. Jak postavit v tomto případě model? Řešení je jednoduché. Předpokládejme, že

$$x_k \stackrel{\triangle}{=} [x_{k1}, x_{k2}, ..., x_{ks}]^T$$

tedy, že x_k je stav dimenze s. Pak hustota pravděpodobnosti stavového šumu

$$p(w_k) = \sum_{n=1}^{q_k} \beta_{kn} N(w_k : \hat{w}_{kn}, Q_{kn})$$

musí obsahovat takový počet členů v součtu, aby platilo

$$q_k \geq s + 1$$

přičemž rovnost by nastala v situaci, kdy jeden člen odpovídá standardním poruchám a další pak přísluší jednotlivým stavovým proměnným. Ukažme si tuto situaci na příkladu.

Nechť $s = 2$. Na stav působí stavový šum s těmito vlastnostmi $(\forall k)$

1. Stav x_k je ovlivněn většinou "malým"šumem např.

$$\beta_{k1} = 0.95, \quad N(w_k : \hat{w}_{k1}, Q_{k1}) \quad \begin{array}{l} Q_{k1} = 10^{-4} I \\ \hat{w}_{k1} = [0,0]^T \end{array}$$

2. Stav x_{k1} je občas pod vlivem silnější poruchy např.

$$\begin{array}{ll} \beta_{k2} = 0.04, & N(w_k : \hat{w}_{k2}, Q_{k2}) \\ & Q_{k2} = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.0001 \end{bmatrix} \\ & \hat{w}_{k2} = [0,0]^T \end{array}$$

3. Stav x_{k2} je zcela vyjímečně pod vlivem hrubé poruchy např.

$$\begin{array}{ll} \beta_{k3} = 0.01, & N(w_k : \hat{w}_{k3}, Q_{k3}) \\ & Q_{k3} = \begin{bmatrix} 0.0001 & 0 \\ 0 & 1 \end{bmatrix} \\ & \hat{w}_{k3} = [0 \;\; 0]^T \end{array}$$

Hustota pravděpodobnosti šumu w_k pro $q_k = 3$ a $s = 2$ pak je dána

$$\begin{aligned} p(w_k) &= 0.95 N(w_k : \hat{w}_{k1}, Q_{k1}) + 0.04 N(w_k : \hat{w}_{k2}, Q_{k2}) \\ &+ 0.01 N(w_k : \hat{w}_{k3}, Q_{k3}) \end{aligned} \tag{6.5.17}$$

Povšimněme si, že v tomto případě modelujeme skokové změny, které mají značný rozptyl kolem nulové hodnoty. Kdybychom chtěli vyjádřit situaci, kdy porucha nabývá nějaké konkrétní nenulové hodnoty, pak můžeme využít střední hodnoty a kovarianci volit velmi malou. Např.

$$\begin{aligned} \hat{w}_{k1} &= [0,0]^T, \qquad Q_{k1} = 10^{-4} I \\ \hat{w}_{k2} &= [1,0]^T, \qquad Q_{k2} = 10^{-4} I \\ \hat{w}_{k3} &= [0,-4]^T, \qquad Q_{k3} = 10^{-4} I \end{aligned}$$

by reprezentovalo naši apriorní představu o tom, že stav x_{k1} se občas skokově mění vlivem poruchy +1 a stav x_{k2} se vyjímečně mění vlivem poruchy -4. Nicméně standardně jsou obě složky stavu pod vlivem slabé „oboustranné“ poruchy.

Různá četnost a charakter hrubých chyb měření. Analogicky s předchozím výkladem můžeme specifikovat stochastický model pro různou četnost a charakter hrubých chyb měření.
Uvažujme opět (5.3.1)–(5.3.5) a předpokládejme, že vektor měření je dán

$$z_k = [z_{k1}, z_{k2}, ..., z_{kd}]^T$$

Pak hustota pravděpodobnosti šumu měření

$$p(v_k) = \sum_{m=1}^{r_k} \gamma_{km} N(v_k : \hat{v}_{km}, R_{km})$$

musí obsahovat takový počet členů r_k, aby platilo

$$r_k \geq d + 1$$

přičemž rovnost by nastala v situaci, kdy jeden člen odpovídá standardním poruchám a další pak přísluší jednotlivým měřením z vektoru měření. Demonstrujme tuto situaci na příkladu. Nechť $d = 2$. Měření je pro všechny časové okamžiky kontaminováno šumem následujících vlastností

1. Měření (celý vektor) je nejčastěji pod vlivem "malého"šumu např.

$$\gamma_{k1} = 0.95 \quad N(v_k : \hat{v}_{k1}, R_{k1}) \quad \begin{array}{l} R_1 = 10^{-4} I \\ \hat{v}_{k1} = [0,0]^T \end{array}$$

2. První složka vektoru měření je zcela zřídka pod vlivem hrubé poruchy, např.

$$\gamma_{k2} = 0.005 \quad \begin{array}{c} N(v_k : \hat{v}_{k2}, R_{k2}) \\ R_{k2} \left[\begin{array}{cc} 1 & 0 \\ 0 & 10^{-4} \end{array} \right] \\ \hat{w}_{k2} = [0 \;\; 0]^T \end{array}$$

3. Druhá složka vektoru měření je občas pod vlivem těžké poruchy, např.

$$\gamma_{k3} = 0.045 \quad \begin{array}{c} N(v_k : \hat{v}_{k3}, R_{k3}) \\ R_{k3} = \left[\begin{array}{cc} 10^{-4} & 0 \\ 0 & 5 \end{array} \right] \\ \hat{v}_{k3} = [0,0] \end{array}$$

Hustota pravděpodobnosti v_k pak bude mít tvar

$$\begin{aligned} p(v_k) &= 0.95 N(v_k : \hat{v}_{k1}, R_{k1}) + 0.005 N(v_k : \hat{v}_{k2}, R_{k2}) \\ &+ 0.045 N(v_k : \hat{v}_{k3}, R_{k3}) \end{aligned} \tag{6.5.18}$$

Poznamenejme, že můžeme lehce vytvořit i poruchy, které mají stranný charakter. Využitím střední hodnoty a kovarianční matice obdobně jako u skokových změn stavu. Hustota (6.5.18) by pak mohla mít např. tyto parametry

$$R_{k1} = R_{k2} = R_{k3} = 10^{-4} I$$

$$\hat{v}_{k1} = [0,0]^T$$

$$\hat{v}_{k2} =]1,0]^T$$

$$\hat{v}_{k3} = [0,5]^T$$

s touto interpretací. Vektor měření je nejčastěji pod mírnou poruchou, velmi zřídka je první složka vektoru měření ovlivněna silnou poruchou o hodnotě „zhruba“ 1 a občas je druhá složka vektoru měření kontaminována hrubou poruchou v hodnotě „zhruba“ 5.

6.6 Realizovatelné estimační algoritmy pro speciální negaussovské systémy

Jak již bylo řečeno v předchozí části této kapitoly, uvedené modely můžeme chápat jako speciální případy modelů reprezentovaných lineárním negaussovským systémem v (5.3.1)-(5.3.5). Protože v 5.3 byl pro lineární negaussovský systém uveden i estimační algoritmus, bude tento algoritmus představovat obecný algoritmus estimace i pro modely uvedené v této kapitole. Hlavním problémem obecného algoritmu estimace z 5.3 je růst počtu členů v šumu definující podmíněné hustoty pravděpodobnosti v případě, že r_k nebo q_k je větší než jedna. Nicméně tento obecný algoritmus můžeme upravit pro různé speciální případy a postupy umožňující algoritmickou únosnost a kvalitní výsledky odhadu.
Pro představu si podrobně popíšeme dva speciální algoritmy pro případ, kdy $F_k = F$ a $H_k = H$ pro všechna k.

Algoritmus A1 (lineární negaussovský systém, hrubé chyby měření - outliers)

Apriorní hustota pravděpodobnosti a hustota pravděpodobnosti šumů

$$p(x_0) = N(x_0 : \hat{x}'_0, P'_0) \tag{6.6.1}$$

$$p(w_k) = N(w_k : 0, Q) \tag{6.6.2}$$

$$p(v_k) = \gamma_1 N(v_k : \hat{v}_1, R_1) + \gamma_2 N(v_k : \hat{v}_2, R_2) \tag{6.6.3}$$

Pak filtrační hustota pravděpodobnosti pro $k = 0$ je

$$p(x_0 \mid z^0) = \alpha_{01} N(x_0 : \hat{x}_{01}, P_{01}) + \alpha_{02} N(x_0 : \hat{x}_{02}, P_{02}) \tag{6.6.4}$$

kde

$$\bar{\alpha}_{01} = \gamma_1 \zeta_{01}, \qquad \bar{\alpha}_{02} = \gamma_2 \zeta_{02}$$

$$\alpha_{01} = \frac{\bar{\alpha}_{01}}{\bar{\alpha}_{01} + \bar{\alpha}_{02}}, \quad \alpha_{02} = 1 - \alpha_{01} \tag{6.6.5}$$

$$\hat{x}_{0j} = \hat{x}_0 + K_{0j}(z_0 - H\hat{x}'_0 - \hat{v}_j) \tag{6.6.6}$$

$$K_{0j} = P'_0 H^T (H P'_0 H^T + R_j)^{-1} \tag{6.6.7}$$

$$P_{0j} = P'_0 - K_{0j} H P'_0 \tag{6.6.8}$$

$$\zeta_{0j} = N(z_0 : H\hat{x}'_0 + \hat{v}_j, H P'_0 H^T + R_j) \tag{6.6.9}$$

Nejjednodušší algoritmus redukce počtu členů filtrační hustoty pravděpodobnosti (6.6.4) je vybrat ve všech časových okamžicích k pouze člen s nejvyšší pravděpodobností nebo dva nejvíce pravděpodobné členy atd. Pro stručnost ve vyjadřování můžeme pak mluvit o redukci na jeden

člen, dva členy atd. V tomto případě proveďme redukci na jeden člen. Nejdříve vypočteme α_{01} a α_{02} a pak provedeme aproximaci exaktně vypočtené hustoty $p(x_0 \mid z^0)$ v (6.6.4) tak, že člen s menší pravděpodobností vypustíme. Tedy např. nechť $\alpha_{02} > \alpha_{01}$ pak

$$p_A(x_0 \mid z^0) \stackrel{\triangle}{=} N(x_0 : \hat{x}_{02}, P_{02}) \sim p(x_0 \mid z^0) \tag{6.6.10}$$

$$p_A(x_0 \mid z^0) \stackrel{\triangle}{=} N(x_0 : \hat{x}_0, P_0) \tag{6.6.11}$$

Prediktivní hustotu pravděpodobnosti pro $k = 1$ můžeme nyní lehce získat

$$p_A(x_1 \mid z^0) = N(x_1 : \hat{x}'_1, P'_1) \tag{6.6.12}$$

$$\hat{x}'_1 = F\hat{x}_0, \qquad P'_1 = FP_0F^T + Q \tag{6.6.13}$$

Pro kroky $k = 1, 2, \ldots$ bychom postupovali analogicky. Poznamenejme, že tento algoritmus může být použit pro eliminaci hrubých chyb měření. Detekuje tyto poruchy a přepíná na vhodný Kalmanův filtr. Algoritmus je z numerického hlediska nepatrně náročnější než Kalmanův filtr, ale oproti němu je robustní vůči hrubým chybám v měření.

Algoritmus A2 (lineární negaussovský systém, skokové změny stavu)

Apriorní hustoty pravděpodobnosti a hustoty pravděpodobnosti šumů

$$p(x_0) = N(x_0 : \hat{x}'_0, P'_0)$$

$$p(w_k) = \beta_1 N(w_k : \hat{w}_1, Q_1) + \beta_2 N(w_k : \hat{w}_2, Q_2)$$

$$p(v_k) = N(v_k : 0, R)$$

Pak filtrační hustota pro $k = 0$ je pak dána

$$p(x_0 \mid z^0) = N(x_0 : \hat{x}_0, P_0)$$

kde

$$\hat{x}_0 = \hat{x}'_0 + K_0(z_0 - H\hat{x}'_0)$$
$$K_0 = P'_0H^T(HP'_0H^T + R)^{-1}$$
$$P_0 = P'_0 - K_0HP'_0$$

Prediktivní hustota pro $k = 1$ je pak

$$p(x_1 \mid z^0) = \alpha'_{11}N(x_1 : \hat{x}'_{11}, P'_{11}) + \alpha'_{12}N(x_1 : \hat{x}'_{12}, P'_{12})$$

kde

$$\alpha'_{11} = \beta_1, \qquad \alpha'_{12} = \beta_2$$
$$\hat{x}'_{11} = F\hat{x}_0 + \hat{w}'_1$$

$$\hat{x}'_{12} = F\hat{x}_0 + \hat{w}'_2$$
$$P'_{11} = FP_0F^T + Q_1$$
$$P'_{12} = FP_0F^T + Q_2$$

Nyní můžeme pokračovat jako v algoritmu A1. Nevypočítáme $p(x_1 \mid z^1)$, ale pouze α_{11} a α_{12} a pak člen s menší pravděpodobností dále nebudeme uvažovat. Například pro $\alpha_{11} > \alpha_{12}$ pak

$$\begin{aligned} p_A(x_1 \mid z^1) &\stackrel{\triangle}{=} N(x_1 : \hat{x}_{11}, P_{11}) \sim p(x_1 \mid z^1) \\ p_A(x_1 \mid z^1) &= N(x_1 : \hat{x}_1, P_1) \end{aligned}$$

a další pokračování pro $k = 2, 3, ...$ je zřejmé.

Poznamenejme, že algoritmus může být použit pro hrubé nebo skokové změny stavu. Opět detekuje změnu a přepne na vhodný Kalmanův filtr. Vlastně dojde k zvětšení kovarianční matice a následkem toho i Kalmanova zisku. Počet numerických operací je zhruba stejný jako při použití Kalmanova filtru, ale rychlost sledování stavu se zvětší.

Na závěr této podkapitoly poznamenejme, že při redukci počtu členů ne na jeden člen jako v předchozích algoritmech, ale např. na tři členy můžeme vypočítat střední hodnotu stavu i kovarianci stavu, jelikož máme k dispozici hustoty pravděpodobnosti, a graficky vyjádřit průběh odhadu. S vyšším počtem členů lze očekávat lepší kvalitu odhadu.

6.7 Zhodnocení uvedených algoritmů estimace

Přístup uvedený v této kapitole vyniká především tím, že umožňuje jednotné a současné modelování výše uvedených jevů, které mohou být chápány jako speciální případy obecného problému modelování a estimace řešeného v kapitolách předchozích. Poznamenejme rovněž, že výsledný estimační algoritmus obsahuje jako vedlejší produkt rozhodovací mechanismy, detekující výskyt sledovaných specifických jevů.

Důležitou vlastností uvedených algoritmů je jejich přirozená fyzikální interpretace. Nabízí se srovnání s parametrickými metodami identifikace z 1. dílu skript, které při použití na odhad měnících se parametrů využívají různé typy zapomínání dat [64], [68], [69], [70], [74]. Parametrické metody identifikace z 1. dílu nepotřebují nutně tolik apriorní informace k definování úlohy, na druhé straně však nepokrývají tolik „situací“ jako postupy uvedené zde a neumožňují přirozenou cestou vložit apriorní informaci do formulace úlohy.

Kapitola 7

Numerické řešení bayesovských vztahů

V předchozích kapitolách jsme se věnovali návrhu estimátorů pro nelineární popř. negaussovský systém, který byl založen na použití aproximací umožňujících analytické řešení, v principu integrálních, bayesovských rekurzivních vztahů (2.4.9), (2.4.10). V této kapitole se soustředíme na zcela jiný přístup k návrhu estimátoru, a to na přístup založený na numerickém řešení těchto vztahů.
Základní idea numerického řešení bayesovských vztahů spočívá v aproximaci spojitého stavového prostoru konečnou (diskrétní) množinou bodů, ve kterých je podmíněná hustota pravděpodobnosti stavu vyčíslena. Tomuto přístupu byla věnována pozornost již od 70. let minulého století [98]. Významný rozvoj však nastal v 90. letech s rozvojem výpočetní techniky, kdy se numerické řešení bayesovských vztahů uplatnilo v návrhu navigačních systémů určujících horizontální polohu dopravního prostředku na základě měření nadmořské výšky, tj. vertikální pozice, a terénní mapy [99], [100], [103], [117]. Poznamenejme ještě, že přístup vedoucí na návrh estimátoru založeného na numerickém řešení bayesovských vztahů je často označován metoda bodových mas [100]. V anglicky psané literatuře se můžeme setkat s pojmy „point-mass method“, popř. „point-mass filter“.

7.1 Popis systému a formulace problému

V této kapitole budeme uvažovat systém popsaný obecně nelineárním negaussovským stavovým modelem

$$x_{k+1} = f_k(x_k) + w_k \tag{7.1.1}$$
$$z_k = h_k(x_k) + v_k \tag{7.1.2}$$

kde proměnné jsou definovány v souladu s předchozími zvyklostmi, tj. k značí časový okamžik, x_k je hledaný stav dimenze nx, z_k je dostupné měření dimenze nz, $f_k(\cdot)$, $h_k(\cdot)$ jsou známé nelineární funkce a w_k a v_k jsou neznámé poruchy popsané známými hustotami $p(w_k)$ a $p(v_k)$. Známá je i hustota pravděpodobnosti počátečního stavu $p(x_0)$. Uvažované hustoty pravděpodobnosti mohou být libovolné, avšak pro jednoduchost, jsou v této kapitole uvažovány jako gaussovské, tj.,

$$p(w_k) = N(w_k : 0, Q_k) \tag{7.1.3}$$
$$p(v_k) = N(v_k : 0, R_k) \tag{7.1.4}$$
$$p(x_0) = N(x_0 : \hat{x}'_0, P'_0) \tag{7.1.5}$$

Cílem je nalézt odhad stavu ve formě podmíněné hustoty stavu $p(x_k|z^l)$, kde $l = k$ pro filtrační a $l = k-1$ pro prediktivní hustotu, na základě numerického řešení bayesovských vztahů, tj. navrhnout metodu bodových mas.

7.2 Aproximace spojité hustoty po částech konstantní hustotou

Klíčovým rozhodnutím pro návrh metody bodových mas je způsob aproximace spojitého prostoru stavu sítí zvolených bodů a interpretací výsledné aproximativní hustoty pravděpodobnosti. V této kapitole uvažujeme aproximaci podmíněné hustoty odhadu stavu směsí rovnoměrných rozdělení definovaných v ekvidistantní[1] mřížce, která pokrývá významnou část stavového prostoru.

Předpokládejme tedy podmíněnou hustotu pravděpodobnosti $p(x_k|z^l)$, kde prostor stavu je $\mathbb{R}^{nx}$, a množinu N ekvidistantně rozložených bodů $\{\xi_k^{(i)}\}_{i=1}^N$, tzv. sítí bodů. Pak, po částech konstantní aproximace podmíněné hustoty může být zapsána jako

$$\hat{p}(x_k|z^l) = \sum_{i=1}^{N} \mathcal{P}_{k|l}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} \tag{7.2.1}$$

kde, pro $nx = 2$, platí

- $\mathcal{P}_{k|l}(\xi_k^{(i)}) = c_k^{-1}\tilde{\mathcal{P}}_{k|l}(\xi_k^{(i)})$; $\tilde{\mathcal{P}}_{k|l}(\xi_k^{(i)}) = p(\xi_k^{(i)}|z^l)$ je hodnota aproximované pravděpodobnosti v bodě $\xi_k^{(i)}$ a c_k je normalizační konstanta definovaná vzápětí,
- $\Delta_k = [\Delta_{k,1}, \Delta_{k,2}]^{\mathrm{T}}$ definuje obdélníkové[2] okolí bodu $\xi_k^{(i)}$, ve kterém je hodnota pravděpodobnosti $\mathcal{P}_{k|l}(\xi_k^{(i)})$ považována za konstantní,
- $S\{x_k; \xi_k^{(i)}, \Delta_k\}$ je výběrová funkce, která nabývá jednotkových hodnot jen a pouze v okolí Δ_k bodu $\xi_k^{(i)}$, jinak je nulová, a je definována jako

$$S\{x_k : \xi_k^{(i)}, \Delta_k\} = \begin{cases} 1, & \text{pokud } x_{k,1} \in [\xi_{k,1}^{(i)} - \frac{\Delta_{k,1}}{2}, \xi_{k,1}^{(i)} + \frac{\Delta_{k,1}}{2}] \wedge \\ & \qquad x_{k,2} \in [\xi_{k,2}^{(i)} - \frac{\Delta_{k,2}}{2}, \xi_{k,2}^{(i)} + \frac{\Delta_{k,2}}{2}] \\ 0, & \text{jinak} \end{cases} \tag{7.2.2}$$

  Výběrovou funkci můžeme tedy chápat jako nenormalizované rovnoměrné rozložení v obdélníkovém okolí uvažovaného bodu, jejíž integrál je

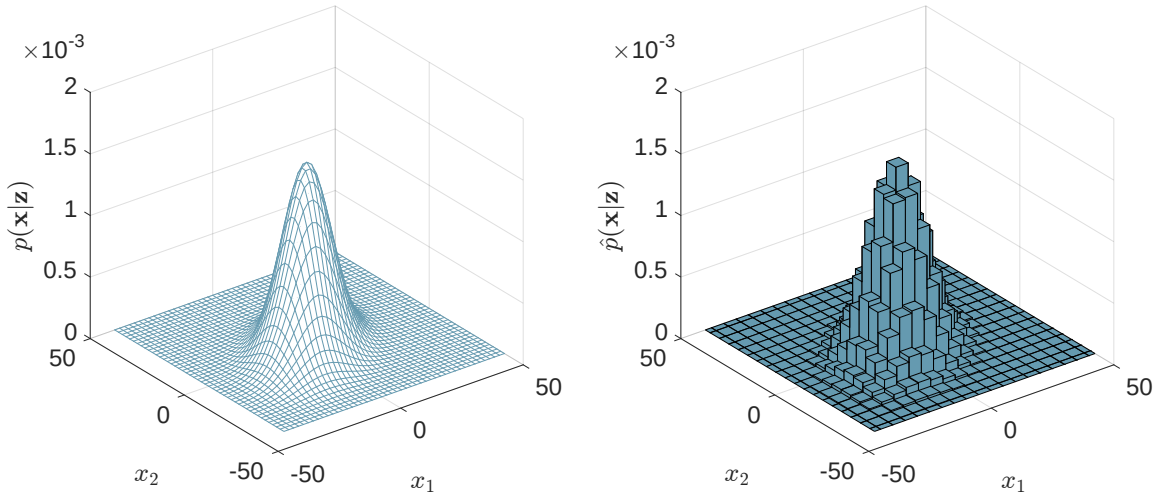
$$\int S\{x_k; \xi_k^{(i)}, \Delta_k\} dx_k = \Delta_{k,1}\Delta_{k,2} = \delta_k \tag{7.2.3}$$

Normalizační konstanta zajišťující, že integrál z aproximované hustoty $\hat{p}(x_k|z^l)$ (7.2.1) je roven jedné, je

$$\begin{aligned} c_k &= \int \sum_{i=1}^{N} \tilde{\mathcal{P}}_{k|l}(\xi_k^{(i)}) S\{x_k; \xi_k^{(i)}, \Delta_k\} dx_k \\ &= \delta_k \sum_{i=1}^{N} \tilde{\mathcal{P}}_{k|l}(\xi_k^{(i)}) \end{aligned} \tag{7.2.4}$$

[1] V ekvidistantní mřížce jsou vztálenosti mezi dvěma sousedními body konstatní.

[2] Okolí bodu si lze, pro $nx = 2$, představit jako obdélník centrovaný v daném bodě. Okolí dvou sousedních bodů se nepřekrývají.



Obrázek 7.1: Ilustrace aproximace spojité hustoty hustotou po částech konstantní.

Poznamenejme, že okolí Δ_k jednotlivých bodů jsou navržena tak, aby se nepřekrývala a těsně na sebe navazovala. Také je vhodné zmínit, že pro vhodně navrženou mřížku bude normalizační konstanta blízká jedničce.
Aproximace hustoty pravděpodobnosti $p(x_k|z^l)$ hustotou po částech konstantní $\hat{p}(x_k|z^l)$ je ilustrována na obrázku 7.1. Poznamenejme, že aproximace hustoty byla definována a ilustrována pro dimenzi stavu $nx = 2$ zejména z důvodu názornosti. Obecně lze aproximovat hustotu pravděpodobnosti stavu libovolné dimenze.

Poznámka . Volba množiny bodů aproximující stavový prostor je klíčová úloha při návrhu metody bodových mas; čím uvažujeme více bodů, hustší síť a větší část stavového prostoru pokrytou body, tím můžeme očekávat přesnější výsledky ovšem za cenu nezanedbatelně rostoucích výpočetních nároků. Obecně lze doporučit, že množina bodů by měla pokrývat takovou část stavového prostoru, kde leží významná část aproximované hustoty pravděpodobnosti. Návrh množiny je tak ovlivněn nejen vlastnostmi šumů a dynamikou systému ale i tím, zda předpokládáme, že aproximovaná hustota pravděpodobnosti je či není multimodální.

Poznámka . Poloha a tvar prediktivní i filtrační hustoty pravděpodobnosti ve stavovém prostoru se v čase vyvíjí. Proto se musí měnit i poloha množiny bodů, tj. poloha mřížky bodů, ve stavovém prostoru. Dynamika mřížky je detailněji diskutována při návrhu prediktivního kroku metody bodových mas.

7.3 Metoda bodových mas

Algoritmus metody bodových mas je dán, podobně jako ostatní dříve představené estimační algoritmy, třemi principiálními kroky, a to inicializace, filtrace a predikce, které budou postupně představeny a odvozeny.

7.3.1 Inicializace

Inicializace algoritmu začíná definicí prediktivní hustoty pravděpodobnosti v čase $k = 0$, která je rovna počáteční podmínce systému, tj.

$$p(x_0|z^{-1}) = p(x_0) \tag{7.3.1}$$

a definicí vhodné množiny bodů $\{\xi_0^{(i)}\}_{i=1}^N$. Pak počáteční spojitá prediktivní hustota je aproximována následující po částech konstantní hustotou určenou body mřížky

$$\hat{p}(x_0|z^{-1}) = \sum_{i=1}^{N} \mathcal{P}_{0|-1}(\xi_0^{(i)}) S\{x_0 : \xi_0^{(i)}, \Delta_0\} \tag{7.3.2}$$

7.3.2 Filtrace

Výpočet filtrační hustoty v časovém okamžiku k vychází z Bayesova vztahu (2.4.9) a dostupného měření z_k. Tedy, Bayesův vztah lze psát

$$\begin{aligned} p(x_k|z^k) &= \frac{p(z_k|x_k)p(x_k|z^{k-1})}{p(z_k|z^{k-1})} \\ &= \tilde{c}_k^{-1} p(z_k|x_k)p(x_k|z^{k-1}), \end{aligned} \tag{7.3.3}$$

kde

- $\tilde{c}_k = \int p(z_k|x_k)p(x_k|z^{k-1})dx_k$ je normalizační konstanta,
- $p(z_k|x_k)$ je hustota pravděpodobnosti měření, která dle (7.1.2), (7.1.4) a diskuze v kapitole 2.4, je dána

$$p(z_k|x_k) = p_{v_k}(z_k - h_k(x_k)) = N(z_k : h_k(x_k), R_k) \tag{7.3.4}$$

- $p(x_k|z^{k-1})$ je prediktivní hustota pravděpodobnosti.

Skutečná prediktivní hustota však není známa. K dipozici máme jen její po částech konstantní aproximaci tj.

$$\hat{p}(x_k|z^{k-1}) = \sum_{i=1}^{N} \mathcal{P}_{k|k-1}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} \tag{7.3.5}$$

která je známá buď z předchozího kroku algoritmu nebo je dána počáteční podmínkou (7.3.2). Dosazením této aproximativní hustoty do Bayesova vztahu (7.3.3) lze odvodit vztah pro výpočet diskrétní aproximace filtrační hustoty $p(x_k|z^k)$ ve tvaru

$$\begin{aligned} \hat{p}(x_k|z^k) &= \hat{\tilde{c}}_k^{-1} p(z_k|x_k)\hat{p}(x_k|z^{k-1}) \\ &= \hat{\tilde{c}}_k^{-1} p(z_k|x_k) \sum_{i=1}^{N} \mathcal{P}_{k|k-1}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} \end{aligned} \tag{7.3.6}$$

Vzhledem k po částech konstantní aproximaci prediktivní hustoty pravděpodobnosti, není nutné hustotu pravděpodobnosti měření vyhodnocovat v každém bodě x_k stavového prostoru. Stačí ji

vyhodnotit jen v bodech mřížky. Pak po částech konstantní aproximace filtrační hustoty nabývá tvaru

$$\hat{p}(x_k|z^k) \approx \hat{\tilde{c}}_k^{-1} \sum_{i=1}^{N} \hat{p}(z_k|x_k = \xi_k^{(i)}) \mathcal{P}_{k|k-1}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} \tag{7.3.7}$$

kde $\hat{\tilde{c}}_k$ je aproximace normalizační konstanty $\tilde{c}_k$ a aproximativní (po částech konstantní) pravděpodobnost měření $\hat{p}(z_k|x_k = \xi_k^{(i)})$ je, dle (7.3.4),

$$\begin{aligned} \hat{p}(z_k|x_k = \xi_k^{(i)}) &= N(z_k : h_k(\xi_k^{(i)}), R_k) \\ &= \tfrac{1}{(2\pi)^{nz/2}(\det R_k)^{1/2}} e^{-\frac{1}{2}\left(z_k - h_k(\xi_k^{(i)})\right)^T R_k^{-1}\left(z_k - h_k(\xi_k^{(i)})\right)} \end{aligned} \tag{7.3.8}$$

Pravděpodobnost měření (7.3.8) je možné chápat jako věrohodnost aktuálního měření z_k vzhledem k i-tému bodu mřížky $\xi_k^{(i)}$. Jinými slovy (7.3.8) říká, jaká je pravděpodobnost, že skutečný stav x_k, na jehož základě bylo realizováno dostupné aktuální měření z_k, je v okolí i-tého bodu $\xi_k^{(i)}$. Pak, filtrační hustotu můžeme zapsat ve finálním tvaru

$$\hat{p}(x_k|z^k) = \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} \tag{7.3.9}$$

kde filtrační pravděpodobnost i-tého bodu je dána součinem apriorní pravděpodobnosti a věrohodnosti měření, tj.

$$\mathcal{P}_{k|k}(\xi_k^{(i)}) = \hat{\tilde{c}}_k^{-1} p(z_k|x_k = \xi_k^{(i)}) \mathcal{P}_{k|k-1}(\xi_k^{(i)}) \tag{7.3.10}$$

Výpočtem normalizační konstanty, zajišťující jednotkový integrál hustoty,

$$\begin{aligned} \hat{\tilde{c}}_k &= \int \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k \\ &= \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) \int S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k \\ &= \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) \delta_k \end{aligned} \tag{7.3.11}$$

filtrační krok metody bodových mas končí.

Všimněme si, že ve filtračním kroku metody bodových mas nedochází ke změně polohy bodů mřížky. Mění se pouze pravděpodobnost asociovaná s každým bodem. Poloha filtrační hustoty ve stavovém prostoru se však může lišit od polohy prediktivní hustoty a tento fakt musíme vzít v potaz už při návrhu mřížky pro prediktivní hustotu (7.3.5). Poznamenejme také, že metoda bodových mas poskytuje odhad ve formě po částech konstantní hustoty pravděpodobnosti. Na základě této hustoty, můžeme vypočítat libovolný filtrační moment, např. střední hodnotu či kovarianční matici. Způsob výpočtu momentů z aproximativních po částech konstantních hustot je ilustrován později. Všimněme si, že znalost momentů není však pro běh filtru nutná.

7.3.3 Predikce

Výpočet prediktivní hustoty pro čas $k+1$ vychází z Chapman-Kolmogorovovy rovnice (2.4.10) mající tvar

$$p(x_{k+1}|z^k) = \int p(x_{k+1}|x_k)p(x_k|z^k)dx_k \tag{7.3.12}$$

kde

- $p(x_{k+1}|x_k)$ je přechodová hustota pravděpodobnosti stavu, která je, dle (7.1.1), (7.1.3) a diskuze v kapitole 2.4, dána

$$p(x_{k+1}|x_k) = p_{w_k}(x_{k+1} - f_k(x_k)) = N(x_{k+1} : f_k(x_k), Q_k) \tag{7.3.13}$$

- $p(x_k|z^k)$ je filtrační hustota pravděpodobnosti.

Skutečná filtrační hustota není známá, k dipozici je však její aproximace pomocí bodových mas daná vztahem (7.3.9). Jejím dosazením do (7.3.12) získáme vztah

$$\hat{p}(x_{k+1}|z^k) = \int p(x_{k+1}|x_k) \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k \tag{7.3.14}$$

který nelze jednoduchým způsobem dále upravit díky závislosti stavu x_{k+1} na x_k, přes který integrujeme. Abychom tedy mohli vyřešit výše uvedený integrální vztah v duchu bodových mas (tj. bez nutnosti integrace), musíme navrhnout novou mřížku bodů $\{\xi_{k+1}^{(i)}\}_{i=1}^N$ vhodně aproximující tu část stavového prostoru, kde očekáváme počítanou prediktivní hustotu.
Definujme si tedy novou mřížku bodů $\{\xi_{k+1}^{(j)}\}_{j=1}^N$. Pak můžeme aproximovat hustotu $p(x_{k+1}|x_k)$ následující po částech konstantní hustotou

$$p(x_{k+1}|x_k) \approx \hat{p}(x_{k+1}|x_k) = \sum_{j=1}^{N} p(\xi_{k+1}^{(j)}|x_k) S\{x_{k+1} : \xi_{k+1}^{(j)}, \Delta_{k+1}\} \tag{7.3.15}$$

kde $p(\xi_{k+1}^{(j)}|x_k) = c_k^{-1}\tilde{p}(\xi_{k+1}^{(j)}|x_k)$ a $c_k = \delta_{k+1}\sum_{j=1}^{N} p(\xi_{k+1}^{(j)}|x_k)$.

Dosazením (7.3.15) do (7.3.14) lze odvodit finální aproximativní vztah pro výpočet Chapman-Kolmogorovovy rovnice ve tvaru

$$\begin{aligned}
\hat{p}(x_{k+1}|z^k) &\approx \int \hat{p}(x_{k+1}|x_k) \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k \\
&= \int \sum_{j=1}^{N} p(\xi_{k+1}^{(j)}|x_k) S\{x_{k+1} : \xi_{k+1}^{(j)}, \Delta_{k+1}\} \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k \\
&\approx \int \sum_{j=1}^{N} \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) p(\xi_{k+1}^{(j)}|x_k = \xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} S\{x_{k+1} : \xi_{k+1}^{(j)}, \Delta_{k+1}\} dx_k \\
&= \sum_{i=1}^{N} \sum_{j=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) p(\xi_{k+1}^{(j)}|x_k = \xi_k^{(i)}) S\{x_{k+1} : \xi_{k+1}^{(j)}, \Delta_{k+1}\} \underbrace{\int S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k}_{\delta_k} \\
&= \sum_{j=1}^{N} \underbrace{\sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) p(\xi_{k+1}^{(j)}|x_k = \xi_k^{(i)}) \delta_k}_{\mathcal{P}_{k+1|k}(\xi_{k+1}^{(j)})} S\{x_{k+1} : \xi_{k+1}^{(j)}, \Delta_{k+1}\} \\
&= \sum_{j=1}^{N} \mathcal{P}_{k+1|k}(\xi_{k+1}^{(j)}) S\{x_{k+1} : \xi_{k+1}^{(j)}, \Delta_{k+1}\}
\end{aligned} \tag{7.3.16}$$

Výpočet prediktivní pravděpodobnosti v bodě nové mřížky, tj.

$$\mathcal{P}_{k+1|k}(\xi_{k+1}^{(j)}) = \sum_{i=1}^{N} p(\xi_{k+1}^{(j)}|x_k = \xi_k^{(i)}) \mathcal{P}_{k|k}(\xi_k^{(i)}) \delta_k, \forall j \tag{7.3.17}$$

lze interpretovat jako pravděpodobnost, že skutečný stav v čase $k+1$, tj. x_{k+1}, leží v okolí bodu j-tého budu nové mřížky $\xi_{k+1}^{(j)}$, za předpokladu, že skutečný stav x_k ležel v okolí prvního bodu původní mřížky $\xi_k^{(1)}$ s pravděpodobností $\mathcal{P}_{k|k}(\xi_k^{(1)})$, druhého bodu $\xi_k^{(2)}$ s pravděpodobností $\mathcal{P}_{k|k}(\xi_k^{(2)})$ až N-tého bodu $\xi_k^{(N)}$ s pravděpodobností $\mathcal{P}_{k|k}(\xi_k^{(N)})$. Prediktivní pravděpodobnost (7.3.17) je tedy nutno vypočítat pro každý bod nové mřížky.
Vztah (7.3.16), který ve své podstatě realizuje konvoluci, je výpočetně nejnáročnější operací filtru. Výpočetní náročnost konvoluce roste kvadraticky s počtem bodů mřížky. Proto, lze v literatuře najít aproximativní přístupy k výpočetně úspornému řešení konvoluce [99]. Alternativou je použití prostředků pro paralelní výpočet konvoluce, který může být efektivně realizován např. pomocí moderních grafických karet (výpočet $\mathcal{P}_{k+1|k}(\xi_{k+1}^{(j)})$ (7.3.17) je nezávislý pro jednotlivé body $\xi_{k+1}^{(j)}$). Poznamenejme, že nejnovější verze programu MATLAB® umožňují uživatelsky přívětivou kompilaci programu pro paralelní zpracování dat na grafických kartách.

Poznámka . Definice nové sítě bodů je důležitou operací prediktivního kroku filtru. Nová mřížka typicky vychází z původní mřížky, která je propagována do následujícího časového okamžiku za pomoci rovnice dynamiky (7.1.1), kde musíme vzít v potaz i vliv stavového šumu, zejména jeho variance. Všimněme si, že nová mřížka může mít odlišný počet bodů od původní mřížky. Avšak z hlediska konstantní výpočetní náročnosti algoritmu filtru, je vhodné uvažovat počet bodů konstantní. Návrh mřížky je detailně diskutován např. v [99], [102].

7.3.4 Výpočet momentů

Metoda bodových mas poskytuje odhad stavu ve formě hustoty pravděpodobnosti vypočtené v diskrétní mřížce bodů. Pro interpretaci a následné využití odhadu je však vhodné míti k dispozici bodový odhad stavu popř. příslušnou kovarianční matici.
Uvažujme filtrační hustotu pravděpodobnosti $\hat{p}(x_k|z^k)$ (7.3.9), pak odhad ve smyslu maximální aposteriorní pravděpodobnosti (2.5.8) je

$$\hat{x}_k^{MAP} = \xi_k^{(i^{MAP})} \tag{7.3.18}$$

kde $i^{MAP} = \arg\max_i \mathcal{P}_{k|k}^{(i)}$ je index bodu mřížky, kde hodnota filtrační hustoty nabývá maxima. Odhad ve smyslu podmíněné střední hodnoty je dán vztahem

$$\begin{aligned}
\hat{x}_k &= E[x_k|z^k] \\
&= \int x_k p(x_k|z^k) dx_k \\
&\approx \int x_k \hat{p}(x_k|z^k) dx_k \\
&= \int x_k \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k \\
&= \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) \int x_k S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k
\end{aligned} \tag{7.3.19}$$

který lze, s ohledem na definici výběrové funkce (7.2.2), zapsat jako

$$\hat{x}_k = \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) \xi_k^{(i)} \delta_k \tag{7.3.20}$$

Tedy, střední hodnota je dána váženým součtem bodů mřížky. Kovarianční matice chyby filtračního odhadu (7.3.20) je

$$\begin{aligned}
P_k &= cov[x_k|z^k] \\
&= \int (x_k - \hat{x}_k)(x_k - \hat{x}_k)^T p(x_k|z^k) dx_k \\
&\approx \int (x_k - \hat{x}_k)(x_k - \hat{x}_k)^T \hat{p}(x_k|z^k) dx_k \\
&= \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) \int (x_k - \hat{x}_k)(x_k - \hat{x}_k)^T S\{x_k : \xi_k^{(i)}, \Delta_k\} dx_k \\
&= \sum_{i=1}^{N} \mathcal{P}_{k|k}(\xi_k^{(i)}) (\xi_k^{(i)} - \hat{x}_k)(\xi_k^{(i)} - \hat{x}_k)^T \delta_k + P_k^{unif}
\end{aligned} \tag{7.3.21}$$

kde matice $P_k^{unif} = \frac{\delta_k^2}{12} I[\Delta_{k,1}, \Delta_{k,2}, \ldots, \Delta_{k,nx}]$ plyne z interpretace aproximativní hustoty jako součtu rovnoměrných rozdělení. Odvození matice pro více-dimenzionální systém může být nalezeno např. v [101]. Poznamenejme, že pro dostatečně hustou síť bodů, kde okolí bodu je dostatečně malé, lze matici P_k^{unif} zanedbat.

7.4 Shrnutí algoritmu a závěrečné poznámky

Odvození metody bodových mas ilustruje obecný přístup k numerickému řešení integrálních funkcionálních vztahů, v tomto případě bayesovských rekurzivních vztahů. I přes složitější odvození, výsledný algoritmus metody bodových mas je relativně jednoduchý a je shrnut v následujících krocích.

Algoritmus metody bodových mas

(i) Definujme počáteční (prediktivní) hustotu pravděpodobnosti $p(x_0|z^{-1})$ a mřížku bodů $\{\xi_0^{(i)}\}_{i=1}^N$. Spočtěme aproximativní prediktivní hustotu pravděpodobnosti $\hat{p}(x_0|z^{-1})$ (7.3.2). Stanovme počáteční časový okamžik $k = 0$.

(ii) Po příchodu měření z_k, spočtěme filtrační hustotu pravděpodobnosti $\hat{p}(x_k|z^k)$ dle (7.3.9) ve všech bodech mřížky s přihlédnutím ke vztahům (7.3.8), (7.3.10) a (7.3.11).

(iii) Nadefinujme novou mřížku pro časový okamžik $k+1$, tj. $\{\xi_{k+1}^{(i)}\}_{i=1}^N$ a spočtěme prediktivní hustotu pravděpodobnosti $\hat{p}(x_{k+1}|z^k)$ dle konvoluce (7.3.16) a (7.3.17).

Pro $k = k + 1$, algoritmus pokračuje krokem (ii).

Poznamenejme, že při běhu metody bodových mas je nutné kontinuálně kontrolovat, zda odhadované hustoty pravděpodobnosti jsou v oblasti stavového prostoru, který je aproximován mřížkou, popř. vlastnosti mřížky upravit [104].
I přes výrazný pokrok ve výkonu osobních i průmyslových počítačů je metoda bodových mas realizovatelná do dimenze stavu $nx = 3$. V případě některých více-dimenzionálních systémů, kdy část stavových veličin se vyvíjí dle nelineárního modelu a část dle lineárního, se nabízí možnost použití tzv. marginalizace či Rao-Blackwellizace, při návrhu metody bodových mas. Tento postup ve svém důsledku vede k estimačnímu algoritmu, kde „nelineárně modelovaná“ část stavu je odhadována výpočetně náročnou metodou bodových mas a zbylá „lineárně modelovaná“ část výpočetně úsporným Kalmanovo filtrem. Více o tomto konceptu lze nalézt např. v [104]–[107].

Kapitola 8

Využití lineární i nelineární filtrace v úlohách identifikace a rozhodování

Cílem této kapitoly je naznačit, že diskutovaný přístup k modelování a estimaci lze použít s úspěchem na řadu dalších úloh. Ukážeme si také vztah mezi identifikací a estimací.

8.1 Využití Kalmanova filtru při identifikaci systémů

V této sekci se budeme zabývat využitím Kalmanova filtru při identifikaci [31], [10], [28], [32], [33], [34]. Identifikace systémů vedle matematického modelování slouží k postavení matematického modelu reprezentujícího zkoumaný systém. Uvažujme pro jednoduchost jednodimenzionální ARX model s proměnnými koeficienty

$$A_k(q^{-1})z_k = B_k(q^{-1})u_k + v_k \tag{8.1.1}$$

kde

$$A_k(q^{-1}) = 1 + a_{k1}q^{-1} + a_{k2}q^{-2} + ... + a_{kna}q^{-na}$$
$$B_k(q^{-1}) = b_{k1}q^{-1} + b_{k2}q^{-2} + ... + b_{knb}q^{-nb}$$

a z_k je měřený výstup, u_k vstup systému a $\{v_k\}$ bílý gaussovský šum s nulovou střední hodnotou a kovariancí R_k, tedy

$$p(v_k) = N(v_k : 0, R_k) \tag{8.1.2}$$

Parametry systému neboli koeficienty polynomů $A_k(q^{-1}), B_k(q^{-1})$ se vyvíjí v čase a jsou neznámé. Definujme vektor Θ_k složený z těchto parametrů

$$\Theta_k \stackrel{\triangle}{=} [a_{k1}, a_{k2}, ..., a_{kna}, b_{k1}, b_{k2}, ..., b_{knb}]^T \tag{8.1.3}$$

Předpokládejme, že vektor parametrů Θ_k je pod vlivem gaussovského bílého šumu $\{w_k\}$ a vyvíjí se podle následujícího vztahu

$$\Theta_{k+1} = \Theta_k + w_k \tag{8.1.4}$$

kde hustota pravděpodobnosti stavového šumu je dána

$$p(w_k) = N(w_k : 0, Q_k) \tag{8.1.5}$$

Jestliže shromáždíme měřené veličiny do vektoru regresorů φ_k

$$\varphi_k \stackrel{\triangle}{=} [z_{k-1}, z_{k-2}, ..., z_{k-na}, u_{k-1}, u_{k-2}, ..., u_{k-nb}]^T \tag{8.1.6}$$

pak (8.1.1) můžeme zapsat takto

$$z_k = \varphi_k^T \Theta_k + v_k \tag{8.1.7}$$

Položíme-li dále

$$\begin{aligned} x_k &\stackrel{\triangle}{=} \Theta_k \\ H_k &\stackrel{\triangle}{=} \varphi_k^T \end{aligned} \tag{8.1.8}$$

a budeme-li předpokládat, že náhodná veličina x_{kmin} (počáteční stav) je nezávislá na $\{w_k\}$ a $\{v_k\}$, které jsou vzájemně nezávislé, a má gaussovské rozložení

$$p(x_{kmin}) = N(x_{kmin} : \hat{x}'_{kmin}, P'_{kmin}) \tag{8.1.9}$$

kde $kmin = \max\{na, nb\}$ je první časový okamžik umožňující vytvoření vektoru regresorů (8.1.6), dostaneme se formálně ke shodné formulaci úlohy kalmanovské filtrace. Tudíž pro odhad parametrů ARX modelu (8.1.1) můžeme použít Kalmanův filtr či prediktor. Povšimněme si, že v tomto případě nelze předpočítávat kovarianční matici nebo Kalmanův zisk, protože vektor regresorů není znám do budoucnosti. Na rozdíl od 1. dílu, kdy parametry byly konstanty, zde jsou chápány jako náhodné veličiny, a proto můžeme nyní úlohu odhadu parametrů pojmout obecněji, a to ve smyslu sledování měnících se parametrů (stavu).

Poznámka . Vhodným výběrem hustoty pravděpodobnosti $p(v_k)$ můžeme opět postavit model pro hrubé chyby měření (rovnice) a stanovením hustoty $p(w_k)$ tvarovat skokové změny parametrů. Pak lze použít pro odhad parametrů filtr s vícenásobnou linearizací. Simulační ověření bylo provedeno v [52], [66], [67].

Poznámka . Volba modelu dynamiky vývoje odhadovaných parametrů (8.1.4) je klíčová pro správnou funkcionalitu estimátoru. V některých situacích může být výhodnější uvažovat gauss-markovský model, který vede na omezenou varianci odhadu, na místo modelu náhodné procházky, který principiálně umožňuje nekonečně velkou varianci odhadu. Všimněme si také, že případná znalost o rychlosti a povaze změn elementů vektoru Θ_k může být reflektována ve volbě matice Q_k či v dynamice vývoje odhadovaných parametrů, jejíž volba je plně v rukách návrháře.

8.2 Využití nelineární filtrace při identifikaci systémů

V předchozí kapitole byla pozornost věnována identifikaci lineárních modelů ve struktuře ARX. Pokud je identifikovaný model popsán následujícím modelem, který je nelineární vzhledem k neznámým parametrům,

$$z_k = h_k(\Theta_k) + v_k \tag{8.2.1}$$

kde $h_k(\cdot)$ je známá nelineární funkce (implicitně závislá na známých regresorech φ_k^T) nelze již Kalmanův filtr pro odhad parametrů Θ_k použít a je nutné využít dříve představené metody nelineární filtrace. Ty zahrnují např. rozšířený Kalmanův filtr představený v kapitole

4.1, unscentovaný Kalmanův filtr představený v kapitole 4.4 nebo metodu bodových mas představenou v kapitole 7.

Poznámka . Analogický postup lze použít i při identifikaci modelu ve formě neuronových sítí, která byla představena v prvním díle skript, kdy odhadujeme nejen váhy aktivačních funkcí, ale i parametry aktivačních funkcí. Pak na identifikovaný model můžeme pohlížet jako na model s nelineární funkcí parametrů, které odhadujeme nelineárním filtrem. Poznamenejme, že aktivační funkce používané v neuronových sítích jsou poměrně snadno derivovatelné a tedy pro odhad parametrů je často používán rozšířený Kalmanův filtr. Poznamenejme rovněž, že při současném odhadu vah a parametrů aktivačních funkcí obvykle stačí uvažovat menší počet aktivačních funkcí, než je tomu v případě, kdy odhadujeme pouze váhy a parametry volíme.

8.3 Odhad vlastností poruch pomocí lineárního prediktoru

Metody odhadu stavu lze nalézt i v jiných, zdánlivě překvapivých, úlohách identifikace parametrů. Jednou z těchto úloh je i odhad vlastností poruch stochastických dynamických systémů.

Jak jsme si mohli v předchozích kapitolách všimnout, všechny doposud zmíněné metody odhadu stavu jsou založeny na známém stavovém modelu uvažovaného systému. Podobně, známý stavový model je klíčovým předpokladem i pro mnohé techniky pro návrh regulátoru či detekce poruch. Konstrukce stavového modelu stochastického dynamického systému, či některých jeho částí, však může být v mnoha případech složitá.
Zatímco model deterministické části stavového modelu, tj. funkce $f_k(\cdot)$, $h_k(\cdot)$ v rovnici dynamiky a měření (7.1.1), (7.1.2), typicky vychází z různých fyzikálních, chemických, biologických či matematických zákonů, model stochastické části, tj. popis stavového šumu w_k a šumu v rovnici měření v_k, musí být mnohdy nalezen využitím naměřených dat.
Jako příklad zde můžeme opět zmínit navigační systém letadla. Ten je založen na stavovém modelu. Jeho deterministická část popisuje teoretický vztah mezi pozicí, rychlostí a orientací letadla v závislosti na zrychlení a úhlové rychlosti letadla. Je tak, při uvažovaných podmínkách, dokonale známá. Stochastická část modelu pak bere v potaz jednak chyby ovlivňující dostupné měření, tak i nemodelované síly působící na letadlo, jejichž model by byl příliš složitý. Popis poruch nelze typicky najít analytickým způsobem, a tedy musíme jej identifikovat.
Proto, od 70. let minulého století, byla věnována značná pozornost metodám odhadující popis poruch ve stavovém modelu na základě známé deterministické části modelu a množiny naměřených dat. V literatuře lze najít široké spektrum identifikačních metod pro lineární i nelineární, časově variantní i invariantní systémy [108], [109], mnohdy založené na využití metod odhadu stavu.
V této kapitole představíme moderní metodu odhadu kovariančních matic poruch stavového modelu, která je založena na specifickém využití lineárního, avšak neoptimálního, estimátoru stavu [110]. Metodu, v anglicky psané literatuře označovanou „autocovariance least-squares method“, budeme dále nazývat jako autokovarianční metodu.

8.3.1 Popis systému a formulace problému

Uvažujme systém popsaný lineárním časově invariantním stavovým modelem

$$x_{k+1} = Fx_k + w_k \tag{8.3.1}$$

$$z_k = Hx_k + v_k \tag{8.3.2}$$

kde proměnné jsou definovány v souladu s definicí modelu pro Kalmanův filtr, tj. k značí časový okamžik, x_k je neznámý stav dimenze nx, z_k je dostupné měření dimenze nz, F, H jsou známé

matice. Předpokládáme, podobně jako u Kalmanova filtru, že dvojice F a H je pozorovatelná. Vektory w_k a v_k reprezentují stavový šum a šum v rovnici měření. Oproti Kalmanově filtru však popis poruch w_k a v_k není znám. Pouze předpokládáme, že se jedná o stacionární procesy s nulovou střední hodnotou, které jsou nezávislé na počátečního stavu x_0. Není dále předpokládaná znalost hodnoty ani momentů počátečního stavu.
Cílem je nalézt odhad kovariančních matic poruch, tj. $Q = cov[w_k], R = cov[v_k], \forall k$, za předpokladu známých matic F a H a dostupné sekvence měření $z^N = [z_0, \ldots, z_N]$.

8.3.2 Predikce stavu a měření

Uvažujme lineární systém (8.3.1), (8.3.2). Při neznalosti kovariančních matic poruch Q a R nelze navrhnout optimální prediktor (3.3.11) poskytující odhad s minimální variancí chyby odhadu. Avšak lze navrhnout *neoptimální* prediktor stavu ve formě

$$\begin{aligned} \hat{x}'_{k+1} &= F\hat{x}'_k + FK(z_k - H\hat{x}'_k) \\ &= F\hat{x}'_k + FK\tilde{z}_k \end{aligned} \tag{8.3.3}$$

kde

$$\tilde{z}_k = z_k - H\hat{x}'_k \tag{8.3.4}$$

je chyba predikce měření a K je zisk prediktoru. Zisk K a počáteční podmínka prediktoru $\hat{x}'_0$ jsou v tomto případě chápány jako uživatelem volené parametry. Počáteční podmínka může být zvolena jako libovolný reálný vektor, avšak za vhodnou volbu lze považovat $\hat{x}'_0 = 0$. Zisk může být zvolen jako reálná matice splňující podmínku, že níže definovaná matice

$$\bar{F} = F - FKH \tag{8.3.5}$$

je stabilní, tj. všechna vlastní čísla matice $\bar{F}$ jsou uvnitř jednotkového kruhu. Volba zisku je detailněji diskutována později. Pak chyba predikce stavu, definována vztahem

$$\begin{aligned} \varepsilon_{k+1} &= x_{k+1} - \hat{x}'_{k+1} \\ &= Fx_k + w_k - F\hat{x}'_k - FK(z_k - H\hat{x}'_k) \\ &= Fx_k + w_k - F\hat{x}'_k - FK(Hx_k + v_k - H\hat{x}'_k) \\ &= F(x_k - \hat{x}'_k) - FK(H(x_k - \hat{x}_k) + v_k) - w_k \\ &= (F - FKH)\varepsilon_k - FKv_k + w_k \end{aligned} \tag{8.3.6}$$

je stabilní proces, což znamená, že kovarianční variance chyby odhadu stavu konverguje ke konečné hodnotě a tedy zůstává omezená (i když výsledná kovarianční matice chyby odhadu není minimální, tak jak je tomu pro optimální Kalmanův prediktor).
S přihlédnutím k (8.3.4), (8.3.6) lze definovat chybový model predikce stavu

$$\begin{aligned} \varepsilon_{k+1} &= \overbrace{(F - FKH)}^{\bar{F}}\varepsilon_k + \overbrace{[I,\ -FK]}^{G}\overbrace{\begin{bmatrix} w_k \\ v_k \end{bmatrix}}^{\zeta_k} \\ &= \bar{F}\varepsilon_k + G\zeta_k \end{aligned} \tag{8.3.7}$$

$$\begin{aligned} \tilde{z}_k &= Hx_k + v_k - H\hat{x}_k \\ &= H\varepsilon_k + v_k \end{aligned} \tag{8.3.8}$$

který bude základem odvození autokovarianční metody. Všimněme si, že matice $\bar{F}$ a G v chybovém modelu jsou známé a ζ_k je náhodná veličina s nulovou střední hodnotou a kovarianční maticí

$$\Sigma = cov[\zeta_k] = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \tag{8.3.9}$$

8.3.3 Autokovarianční metoda pro odhad vlastností poruch

Autokovarianční metoda je založena na statistické analýze chybového modelu (8.3.7), (8.3.8) lineárního prediktoru (8.3.3). Vlastnosti chyby predikce stavu a měření v ustáleném stavu, tj. pro $k \to \infty$, kdy vliv počáteční podmínky $\hat{x}'_0$ odezněl, jsou sumarizovány v následujících odstavcích.

Chyba predikce stavu ε_k (8.3.7) je, v ustáleném stavu, stochastický proces s nulovou střední hodnotou, tj.

$$\begin{aligned} E[\varepsilon] &= E[\varepsilon_{k+1}] = E[\varepsilon_k] \\ &= \bar{F}E[\varepsilon] + GE[\zeta_k] \\ &= \tfrac{G}{I-F}E[\zeta_k] \\ &= 0 \end{aligned} \tag{8.3.10}$$

a kovarianční maticí

$$\begin{aligned} P' &= cov[\varepsilon] = cov[\varepsilon_{k+1}] = cov[\varepsilon_k] \\ &= \bar{F}cov[\varepsilon]\bar{F}^T + Gcov[\zeta_k]G^T \\ &= \bar{F}P'\bar{F}^T + G\Sigma G^T \end{aligned} \tag{8.3.11}$$

Elegantní řešení výše uvedené maticové (Lyapunovovy) lineární rovnice (8.3.11) je založeno na Kroneckerově algebře [112], zejména na vztazích

$$(ABC)_s = (C^T \otimes A)B_s \tag{8.3.12}$$

$$(A + B)_s = A_s + B_s \tag{8.3.13}$$

kde A, B, C jsou matice vhodných dimenzí, $\otimes$ značí Kroneckerův součin a $A_s = (A)_s$ značí vektor, který je dán sloupci matice A naskládaných pod sebe. Využitím vztahů (8.3.12) a (8.3.13) v (8.3.11) můžeme získat finální vztah pro ustálenou kovarianční matici chyby predikce stavu

$$P'_s = (I - \bar{F} \otimes \bar{F})^{-1}(G \otimes G)\Sigma_s \tag{8.3.14}$$

která je lineární funkcí hledaných kovariančních matic poruch Q a R tvořících vektor Σ_s.

Chyba predikce měření $\tilde{z}_k$ (8.3.8) je, v ustáleném stavu, také stochastický proces s nulovou střední hodnotou, tj.

$$\begin{aligned} E[\tilde{z}_k] &= E[H\varepsilon_k + v_k] \\ &= HE[\varepsilon_k] + E[v_k] \\ &= 0 \end{aligned} \tag{8.3.15}$$

a kovarianční maticí

$$\begin{aligned} C_0 &= E[(H\varepsilon_k + v_k)(H\varepsilon_k + v_k)^T] \\ &= HP'H^T + R \end{aligned} \tag{8.3.16}$$

kterou lze s ohledem na (8.3.12)–(8.3.14) upravit do tvaru

$$\begin{aligned} (C_0)_s &= (H \otimes H)P'_s + R_s \\ &= \underbrace{(H \otimes H)(I - \bar{F} \otimes \bar{F})^{-1}(G \otimes G)}_{\text{známá matice}} \underbrace{\Sigma_s}_{\text{hledaný vektor}} + \underbrace{R_s}_{\text{hledaný vektor}} \end{aligned} \tag{8.3.17}$$

Tím, že prediktor (8.3.3) není optimální, inovační posloupnost $\tilde{z}_k$ (8.3.8) není bílá, jak tomu bylo u Kalmanova filtru, a tedy, má nenulové následující kovarianční matice

$$\begin{aligned} C_p &= E[(H\varepsilon_k + v_k)(H\varepsilon_{k+p} + v_{k+p})^T] \\ &= E[(H\varepsilon_k + v_k)\left(H(\bar{F}^p\varepsilon_k + \bar{F}^{p-1}G\zeta_k + \ldots + G\zeta_{k+p-1}) + v_{k+p}\right)^T] \\ &= E[H\varepsilon_k(H\bar{F}^p\varepsilon_k)^T] + E[v_k(H\bar{F}^{p-1}G\zeta_k)^T] \\ &= HP'(\bar{F}^p)^T H^T - RK^T F^T(\bar{F}^{p-1})^T H^T \end{aligned} \tag{8.3.18}$$

kde $p = 1, 2, \ldots, P$, charakterizující závislost inovace mezi dvěma časovými okamžiky. Využitím (8.3.12)–(8.3.14) lze dále psát

$$\begin{aligned} (C_p)_s &= \left((H\bar{F}^p) \otimes H\right) P'_s - \left((H\bar{F}^{p-1}FK) \otimes I\right) R_s \\ &= \underbrace{\left((H\bar{F}^p) \otimes H\right)(I - \bar{F} \otimes \bar{F})^{-1}(G \otimes G)}_{\text{známá matice}} \underbrace{\Sigma_s}_{\text{hledaný vek.}} - \underbrace{\left((H\bar{F}^{p-1}FK) \otimes I\right)}_{\text{známá matice}} \underbrace{R_s}_{\text{hledaný vek.}} \end{aligned} \tag{8.3.19}$$

Vztahy (8.3.17), (8.3.19), definující autokovarianční funkci inovační posloupnosti, jsou lineární funkcí hledaných prvků kovariančních matic poruch Q a R, které formují vektory Σ_s a R_s. Na jejich základě můžeme tedy definovat soustavu lineárních rovnic ve formě

$$(C_0)_s = \underbrace{[(H \otimes H)(I - \bar{F} \otimes \bar{F})^{-1}(G \otimes G) + M]\mathcal{M}}_{\varphi^T(0)}\Theta \tag{8.3.20}$$

$$(C_1)_s = \underbrace{[((H\bar{F}) \otimes H)(I - \bar{F} \otimes \bar{F})^{-1}(G \otimes G) + ((HFK) \otimes I)M]\mathcal{M}}_{\varphi^T(1)}\Theta \tag{8.3.21}$$

$$\vdots \tag{8.3.22}$$

$$(C_P)_s = \underbrace{[((H\bar{F}^P) \otimes H)(I - \bar{F} \otimes \bar{F})^{-1}(G \otimes G) + ((H\bar{F}^{P-1}FK) \otimes I)M]\mathcal{M}}_{\varphi^T(P)}\Theta \tag{8.3.23}$$

kde Θ je vektor obsahující unikátní prvky odhadovaných matic Q a R a M a $\mathcal{M}$ jsou známé výběrové matice, tvořené jedničkami a nulami, pro které platí $R_s = M\Sigma_s$ a $\Sigma_s = \mathcal{M}\Theta$. Poznamenejme, že tvorba výběrových matic je ilustrována dále.
Soustavu rovnic (8.3.20)–(8.3.23) můžeme zapsat v kompaktní formě používané v metodě nejmenších čtverců

$$Y = \Phi\Theta \tag{8.3.24}$$

kde $Y = [((C_0)_s)^T, ((C_1)_s)^T, \ldots, ((C_P)_s)^T]^T$ a $\Phi = \begin{bmatrix} \varphi^T(0) \\ \varphi^T(1) \\ \vdots \\ \varphi^T(P) \end{bmatrix}$ je známá matice.

Pokud bychom znali autokovarianční funkci $C_p, \forall p$, definovanou (8.3.16), (8.3.18), můžeme snadno zkonstruovat vektor Y v (8.3.24) a odhadnout vektor neznámých parametrů Θ. Nicméně, autokovarianční funkce známa není. Můžeme ji však *odhadnout* na základě měřených dat.
Pokud máme sekvenci měření z^k, můžeme vypočítat predikci stavu $\hat{x}'_k, \forall k$, pomocí (8.3.3) a tím i inovační posloupnost $\tilde{z}_k, \forall k$, pomocí (8.3.4). Pak asymptoticky nestranný odhad autokovarianční funkce (8.3.16), (8.3.18) je dán

$$\hat{C}_0 = \tfrac{1}{\tau+1} \sum_{k=0}^{\tau} \tilde{z}_k \tilde{z}_k^T, \tag{8.3.25}$$

$$\hat{C}_p = \tfrac{1}{\tau-p+1} \sum_{k=0}^{\tau-p} \tilde{z}_k \tilde{z}_{k+p}^T, \forall p \tag{8.3.26}$$

Odhad autokovarianční funkce použijeme pro specifikaci vektoru

$$\hat{Y} = [((\hat{C}_0)_s)^T, ((\hat{C}_1)_s)^T, \ldots, ((\hat{C}_P)_s)^T]^T \tag{8.3.27}$$

který reprezentuje odhad vektoru Y v (8.3.24), a následně pro odhad parametrů, tj. prvků kovariančních matic poruch dle

$$\begin{aligned} \hat{\Theta} = \hat{\Theta}(Q,R) &= (\Phi^T\Phi)^{-1}\Phi^T\hat{Y} \\ &= \Phi^\dagger \hat{Y} \end{aligned} \tag{8.3.28}$$

Z prvků vektoru $\hat{\Theta}$ pak následně sestavíme odhady kovariančních matic $\hat{Q}$ a $\hat{R}$.

Příklad 8.2.1. Ilustrujme význam a tvorbu výběrových matic M a $\mathcal{M}$. Předpokládejme skalární systém, tj. $nx = nz = 1$. Pak kovarianční matice poruch Q a R jsou (skalární) variance a tedy $Q = Q_s$ a $R = R_s$. Hledaný vektor unikátních prvků variancí poruch je definován $\Theta = [Q, R]^T$. Kovarianční matice rozšířeného vektoru poruch ζ_k (8.3.9) je matice $2/2$ s variancemi na diagonále $\Sigma = \left[\begin{smallmatrix} Q & 0 \\ 0 & R \end{smallmatrix}\right]$ a vektorová forma matice je $\Sigma_s = \left[\begin{smallmatrix} Q \\ 0 \\ 0 \\ R \end{smallmatrix}\right]$. Výběrová matice M v (8.3.20)–(8.3.23) je pak

$$M = [0, 0, 0, 1] \tag{8.3.29}$$

což splňuje rovnici $R = M\Sigma_s = [0,0,0,1] \left[\begin{smallmatrix} Q \\ 0 \\ 0 \\ R \end{smallmatrix}\right]$. Výběrová matice $\mathcal{M}$ je definována jako

$$\mathcal{M} = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] \tag{8.3.30}$$

což splňuje požadovanou rovnost $\Sigma_s = \mathcal{M}\Theta = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{smallmatrix}\right] \left[\begin{smallmatrix} Q \\ R \end{smallmatrix}\right]$.

Poznámka . Zisk prediktoru K v rovnici (8.3.3) musí být volen tak, aby matice $\bar{F}$ (8.3.5) byla stabilní. Jednou z možností je definovat stabilní matici $\bar{F}$ a pak dopočítat prvky zisku K. Druhou možností, která vždy povede na stabilní matici $\bar{F}$, je definovat dvě libovolné pozitivně definitní matice Q_A a R_A příslušných dimenzí, vypočítat řešení následující algebraické Riccatiho rovnice

$$P'_A = F[P'_A - P'_A H^T (HP'H^T + R_A)^{-1} HP'_A]F^T + Q_A \tag{8.3.31}$$

a stanovit zisk prediktoru dle

$$K = P'_A H^T (H P'_A H^T + R_A)^{-1} \tag{8.3.32}$$

Poznámka . Prediktor (8.3.3) není optimální z hlediska střední kvadratické chyby, tj. lze ukázat že kovarianční matice chyby odhadu stavu (8.3.11) bude větší než ustálená kovarianční matice prediktivní chyby odhadu stavu Kalmanova filtru, resp. prediktoru (3.3.11). Avšak, jak ukazuje (8.3.10), prediktor (8.3.3) poskytuje nestranný odhad stavu i pro neoptimální zisk K.

8.3.4 Shrnutí algoritmu a závěrečné poznámky

Autokovarianční metoda pro odhad vlastností poruch může být definována následujícím algoritmem.

Algoritmus autokovarianční metody pro odhad kovariančních matic poruch

---

(i) Definujme zisk prediktoru K v (8.3.3) tak, aby matice $\bar{F} = F - FKH$ byla stabilní. Definujme počet rovnic P autokovarianční funkce (8.3.20)–(8.3.23), tak aby počet rovnic byl větší než počet odhadovaných parametrů vektoru Θ.

(ii) Na základě sekvence měření z^k, vypočtěme predikci stavu $\hat{x}'_k, \forall k$ (8.3.3) a tím i inovační posloupnost $\tilde{z}_k, \forall k$ (8.3.4).

(iii) Odhadněme autokovarianční funkci inovace dle (8.3.25) a (8.3.26) a stanovme matici Φ v (8.3.24) dle (8.3.20)–(8.3.23).

(iv) Odhadněme vektor parametrů Θ, obsahující prvky hledaných matic Q a R, pomocí metody nejmenších čtverců (8.3.28).

Představená autokovarianční metoda poskytuje asymptoticky nestranný odhad kovariančních matic poruch Q a R [110]. Metoda umožňuje odhad všech prvků matice R, avšak ne více než $nx \cdot nz$ prvků matice Q.
Návrh metody je podmíněn specifikací dvou uživatelských parametrů, a to ziskem prediktoru K a počtem rovnic P. Obě volby mají podstatný vliv na kvalitu odhadu a doporučení k jejich volbě je diskutováno např. v [113]. Poznamenejme, že vyšší počet rovnic P nevede automaticky k lepším odhadům, protože nepoužíváme váženou metodu nejmenších čtverců, tj. nemáme nejlepší nestranný lineární estimátor (tzv. BLUE).
Autokovarianční metoda patří do široké třídy tzv. korelačních metod, které jsou založeny na analýze vlastností chyby predikce měření. Korelační metody byly navrženy i pro lineární časově variantní nebo dokonce i nelineární modely [109], [111].

8.4 Vícemodelový přístup v úloze rozhodování

V této sekci budeme uvažovat situaci, kdy máme rozhodnout, jaký vektor parametrů z dané množiny možných přísluší zkoumanému systému. Jedná se tedy o rozhodnutí, který z konečné množiny modelů je správný. V takovém případě bychom mohli postupovat, při využití technik odhadu stavu, následujícím způsobem.

Předpokládejme, že vektor parametrů Θ je prvek množiny Θ_D

$$\Theta \in \Theta_D = \{\Theta_1, \Theta_2, ..., \Theta_N\}. \tag{8.4.1}$$

Dále předpokládejme, že stavový model s vektorem parametrů Θ je definován

$$x_{k+1} = F_k(\Theta)x_k + w_k \tag{8.4.2}$$

$$z_k = H_k(\Theta)x_k + v_k \tag{8.4.3}$$

Model (8.4.2), (8.4.3) je speciální případ (5.3.1), (5.3.2), jelikož předpokládáme, že $\{w_k\}$, $\{v_k\}$ jsou bílé, vzájemně nezávislé procesy s gaussovským rozložením

$$p(w_k) = N(w_k : 0, Q_k(\Theta)) \tag{8.4.4}$$

$$p(v_k) = N(v_k : 0 : R_k(\Theta)) \tag{8.4.5}$$

a

$$\begin{aligned} p(x_0) \stackrel{\triangle}{=} p(x_0 \mid z^{-1}) &= \sum_{i=1}^{N} \alpha'_{0i} N(x_0 : \hat{x}_{0i}, P_{0i}) \\ &= \sum_{i=1}^{N} P_r(\Theta_i \mid z^{-1}) p(x_0 \mid \Theta_i, z^{-1}) \end{aligned} \tag{8.4.6}$$

kde $P_r(\Theta_i \mid z^{-1})$ je podmíněná pravděpodobnost, že $\Theta = \Theta_i$ za podmínky z^{-1}.

Filtr pro uvažovaný systém lze snadno najít, protože se jedná o speciální případ úlohy řešené v podkapitole 5.3. Z (5.3.28)-(5.3.42) pak dostaneme

$$\begin{aligned} p(x_k \mid z^k) &= \sum_{j=1}^{N} P_r(\Theta_j \mid z^k) p(x_k \mid \Theta_j, z^k) \\ &= \sum_{j=1}^{N} P_r(\Theta_j \mid z^k) N(x_k : \hat{x}_{kj}, P_{kj}) \\ p(x_k \mid \Theta_j, z^{k-1}) &= N(x_k : \hat{x}'_{kj}, P'_{kj}) \end{aligned}$$

$$\hat{x}_{kj} = \hat{x}'_{kj} + K_{kj}[z_k - H_k(\Theta_j)\hat{x}'_{kj}] \tag{8.4.7}$$

$$K_{kj} = P'_{kj} H_k^T(\Theta_j)[H_k(\Theta_j) P'_{kj} H_k^T(\Theta_j) + R_k(\Theta_j)]^{-1} \tag{8.4.8}$$

$$\hat{x}'_{kj} = F_{k-1}(\Theta_j)\hat{x}_{k-1,j} \tag{8.4.9}$$

$$P_{kj} = P'_{kj} - K_{kj} H_k(\Theta_j) P'_{kj} \tag{8.4.10}$$

$$P'_{kj} = F_{k-1}(\Theta_j) P_{k-1,j} F^T_{k-1}(\Theta_j) + Q_{k-1}(\Theta_j) \tag{8.4.11}$$

Zbývá určit aposteriorní pravděpodobnost parametrů Θ_j. Analogicky jako v 5.3 můžeme získat $P_r(\Theta_j \mid z^k)$. Tedy

$$P_r(\Theta_j \mid z^k) = \frac{P_r(\Theta_j \mid z^{k-1}) \zeta_{kj}}{\sum_{j=1}^{N} P_r(\Theta_j \mid z^{k-1}) \zeta_{kj}} \tag{8.4.12}$$

kde

$$\zeta_{kj} = N(z_k : H_k(\Theta_j) \hat{x}'_{kj}, H^T_k(\Theta_j) P'_{kj} H(\Theta_j) + R_k(\Theta_j)) \tag{8.4.13}$$

$j = 1, 2, ...N$

Nyní již je zřejmé, že rozhodnutí, který parametr z dané množiny je správný získáme jako vedlejší produkt úlohy nelineární filtrace. Rovněž je zřejmé, že v této úloze není třeba žádná aproximace, jelikož $q_k = 1$ a $r_k = 1$, a tudíž počet členů ve filtrační hustotě stavu je konstantní. Je roven právě N nebo-li počtu modelů odpovídajících parametrům Θ_j, kde $j = 1, 2, ..., N$. Rozsáhlé testování tohoto přístupu je provedeno v [75]. Podobná technika je používána i v [76].
S přibývajícím počtem měření se bude jedna z aposteriorních pravděpodobností $P_r(\Theta_j \mid z^k)$ blížit jedné. Úlohu rovněž můžeme interpretovat jako rozhodnutí, který z N modelů, z nichž každý představuje systém s jiným vektorem parametrů z předkládané množiny je adekvátní realitě. Algoritmus rozhodování je pak vlastně založen na možné paralelní činnosti N Kalmanových filtrů, jejichž důvěryhodnost je měřena pravděpodobností z (8.4.12). Poznamenejme, že apriorní pravděpodobnosti $P_r(\Theta_j \mid z^{-1})$ pro $j = 1, 2, ..., N$ mohou být shodné, ale mohou rovněž reprezentovat různou důvěru v jednotlivé parametry a v přeneseném smyslu v jednotlivé modely. Speciální případ pak nastane, když máme rozhodnout pouze mezi dvěma modely nebo rozhodnout zda daný model ještě platí.
Povšimněme si, že hlavním výsledkem algoritmu z této sekce je pravděpodobnost příslušející k danému parametru (modelu) a vedlejším produktem je odhad stavu. Naopak algoritmy v kapitole páté produkují jako hlavní výsledek odhad stavu a vedlejším produktem je věrohodnost jednotlivých "modelů".
Poznamenejme, že metoda představená v této kapitole může být chápána jako alternativa k autokovarianční metodě pro odhad kovariančních matic poruch systému. Vícemodelový přístup pro odhad matic Q a R byl navržen a diskutován např. v [116].

8.5 Testování hypotéz

V této podkapitole naznačíme využití vícemodelového přístupu k testování hypotéz. Již z předchozí podkapitoly je zřejmé, že vícemodelový přístup vedoucí na problém nelineární filtrace se speciální strukturou filtru umožňuje snadné řešení úloh i z oblasti zpracování a detekce signálů. Nebudeme se tomuto problému věnovat detailně, ale uveďme alespoň základní ideu pro testování hypotéz. Uvažujme dvě hypotézy na signál

H_0: signál se rovná nule
H_1: signál se rovná pěti.

Měření z_k obsahuje signál s aditivním korelovaným šumem a aditivním bílým šumem. Nechť korelovaný šum řekněme $\{x_k\}$ je reprezentován gauss-markovským procesem se známou korelační funkcí. Předpokládejme, že je vyjádřen pomocí skalární diferenční rovnice

$$x_{k+1} = F_k x_k + w_k \tag{8.5.1}$$

kde F_k je známo, $\{w_k\}$ je bílý gaussovský šum se střední hodnotou nula a známou variancí. Pak hypotézy bychom mohli spíše popsat takto:

$$\begin{aligned} H_0 :\ & x_{k+1} = F_k x_k + w_k \\ & z_k = 0 + x_k + v_k \end{aligned}$$

$$\begin{aligned} H_1 :\ & x_{k+1} = F_k x_k + w_k \\ & z_k = 5 + x_k + v_k \end{aligned}$$

kde $\{v_k\}$ je bílý gaussovský šum $N(v_k : 0, \hat{v}_k)$ nezávislý na $\{w_k\}$.
Jestliže budeme chápat H_1 tak, že proces $\{v_k\}$ má střední hodnotu pět místo nula, pak můžeme zcela využít metodologii z předchozí podkapitoly, protože stačí použít dva Kalmanovy filtry a věrohodnost hypotézy bude dána vztahem (8.4.12) [75]. Povšimněme si podobnosti tohoto přístupu s monitorováním konzistentního odhadu Kalmanova filtru diskutovaného v kapitole 3.3.4.

Poznamenejme opět jako v předchozí podkapitole, že hlavním výsledkem je potvrzení nebo vyvrácení hypotézy a vedlejším produktem je odhad stavu.

8.6 Adaptivní systémy

Adaptivní systémy jsou chápány jako systémy, které jsou schopny samostatného ladění (adaptace) a jejich cílem je buď řízení reálných procesů nebo zpracování signálů [30], [76], [78], [79], [27]. Jádrem každého adaptivního systému je estimační algoritmus zajišťující poznávání řízeného procesu nebo systému generujícího zpracovávaný signál. Přístup k modelování a estimaci popisovaný v tomto díle umožňuje, zvládnout i takové situace, které přesahují rámec linearity a gaussovosti. Důsledkem toho pak adaptivní řízení či adaptivní zpracování signálů vytváří předpoklady pro syntézu adaptivních systémů s lepšími vyhlídkami na kvalitu řízení či zpracování signálu [53], [56], [80], [81]. Podobně můžeme argumentovat i v souvislosti se stochastickým optimálním řízením [52], [82].

Kapitola 9

Metody odhadu stavu v navigačních systémech

Navigační systém je systém určující polohu, rychlost a orientaci objektu (např. auta, letadla, či lodě) na základě měření ze senzorů, které jsou pevně spjaty s objektem. Poloha, rychlost a natočení objektu v prostoru, v anglicky psané literatuře označovány pojmy „position, velocity, attitude“, tvoří hledanou navigační informaci.
Vznik a rozvoj metod odhadu stavu je úzce spjat s rozvojem systémů pro navigaci a sledování (v anglické literatuře označované pojmy „navigation“ a „tracking“). Od šedesátých let minulého století až do dnešních dnů je tak většina navigačních a sledovacích systémů založena na metodách odhadu stavu, které umožní optimální, či téměř optimální, odhad navigační informace na základě dostupných měření a popisu dynamiky objektu.
Cílem této kapitoly je krátce představit dva navigační systémy, které pro výpočet navigační informace využívají metody odhadu stavu. Jmenovitě bude v následujících částech představen

- hybridní navigační systém zpracovávající inerciální měření a měření z globálního navigačního satelitního systému,
- terénní navigační systém kombinující inerciální měření a terénní mapu.

9.1 Integrovaný inerciální a satelitní navigace

Pojmem integrovaná nebo hybridní navigace označujeme navigační systém, který je tvořen kombinací dvou (nebo více) navigačních systémů a senzorů. Typicky jsou integrovány následující [103], [118]

- inerciální navigační systém (INS), který poskytuje odhad navigační informace na základě měření inerciálních senzorů[1],
- přijímač satelitního navigačního systému[2], který poskytuje odhad polohy navigovaného objektu. Přijímač budeme dále označovat jako přijímač GNSS z anglického výrazu „global navigation satelite system“.

[1] Za inerciální senzory považujeme akcelerometr, měřící zrychlení objektu, a gyroskop, měřící úhlovou rychlost objektu.

[2] Mezi globální satelitní navigační systémy patří americký „Global Positioning System (GPS)“, evropské Galileo, ruský GLONASS nebo čínský systém Beidou. Za regionální satelitní navigační systémy můžeme považovat indický NAVIC a japonský QZSS.

Inerciální senzory poskytují měření na vysoké frekvenci (v rozmezí 100 a 300 Hz). Se stejnou frekvencí je pak schopen INS poskytovat odhad navigační informace, nebo-li navigační řešení. V závislosti na přesnosti inicializace INS a kvalitě inerciálních senzorů dokáže INS poskytovat „přesné[3]“ navigační řešení po dobu několika sekund až minut. Na druhou stranu, GNSS přijímač poskytuje navigační řešení na nižší frekvenci s „významnou“, ale v čase konstantní, chybou. Integrovaný navigační systém kombinuje v jistém smyslu duální vlastnosti INS a GNSS přijímače a poskytuje přesné navigační řešení s omezenou chybou, která má v čase relativně konstantní vlastnosti.

9.1.1 Souřadné systémy

Navigační systémy pracují s veličinami, které jsou vyjádřeny nebo vztaženy k různým souřadným systémům (v anglicky psané literatuře označované pojmem „frame“). Mezi základní kartézské souřadné systémy lze zařadit následující:

- ECI (z anglického „Earth-centered inertial“) systém se středem v těžišti Země a s osou z směřující ve směru zemské osy, který nerotuje a ani se neposouvá vůči okolním hvězdám. Tento systém budeme v následujících výpočtech značit písmenem I, popř. pojmem „I-frame“.

- ECEF (z anglického „Earth-centered Earth fixed“) systém se středem v těžišti Země, je definován podobně jako systém ECI s tím rozdílem, že osy ECEF systému jsou fixovány se Zemí, tj. souřadný systém ECEF rotuje vůči systému ECI. Tento systém je přirozený pro popis pozice objektu a v následujících výpočtech jej budeme značit písmenem E, popř. pojmem „E-frame“.

- Lokální navigační (nebo-li tangenciální) souřadný systém (v anglicky psané literatuře označován pojmem „local navigation frame“) je systém se středem v těžišti navigovaného objektu a s osou z směřující ve směru normálového vektoru k elipsoidu modelujícího Zemi. Rovinu $x - y$ tohoto systému si tak lze přestavit jako tangenciální rovinu k elipsoidu, vůči které přirozeně vyjadřujeme orientaci objektu v prostoru. Lokální navigační souřadný systém je tedy nezávislý na orientaci objektu (jeho osy *nejsou* svázány s objektem), souřadný systém pouze sdílí těžiště s navigovaným objektem. Tento systém budeme v následujících výpočtech značit písmenem N, popř. pojmem „N-frame“.

- Souřadný systém objektu (v anglicky psané literatuře označován pojmem „body frame“) je systém se středem v těžišti navigovaného objektu, jehož osy jsou pevně však spjaty s objektem (na rozdíl od navigačního souřadného systému). Typicky, osa x souhlasí s podélnou osou navigovaného objektu. Tento systém budeme v následujících výpočtech značit písmenem B, popř. pojmem „B-frame“.

Poznamenejme, že uvedené definice souřadných systémů jsou zkratkovité a jsou určeny jen pro hrubou představu umožňující porozumět principu navigačních systémů. Přesné definice, další detaily i jiné používané souřadné systémy lze najít např. v [103].

9.1.2 Veličiny, transformace a notace

Některé veličiny používané v návrhu navigačních systémů jsou vztaženy pouze k jednomu souřadnému systému. Jako příklad zde můžeme uvést pozici navigovaného objektu.

[3] S přibývajícím časem roste chyba navigačního řešení poskytovaného INS. Pro některé složky navigační informace může chyba růst až do nekonečna.

- Pozici objektu, která je součástí navigační informace, můžeme v kartézském ECEF souřadném systému popsat vektorem o třech složkách

$$r_{\mathrm{K}}^{\mathrm{E}} = \begin{bmatrix} r_x^{\mathrm{E}} \\ r_y^{\mathrm{E}} \\ r_z^{\mathrm{E}} \end{bmatrix} \tag{9.1.1}$$

  a značí pozici navigovaného objektu vzhledem k počátku souřadného systému v těžišti Země. Složky vektoru $r_{\mathrm{K}}^{\mathrm{E}}$ (9.1.1) mohou být udávány v metrech. Pozici objektu lze však pro člověka přirozeněji popsat využitím sférických souřadnic, kdy pozice objektu je dána následujícím vektorem

$$r^{\mathrm{E}} = \begin{bmatrix} \lambda \\ \varphi \\ h \end{bmatrix} \tag{9.1.2}$$

  kde λ značí zeměpisnou *délku* (v angličtině označovanou pojmem „longitude“), φ značí zeměpisnou *šířku* (v angličtině označovanou pojmem „latitude“) a h je nadmořská *výška* (v angličtině označovaná pojmem „altitude“). První dvě veličiny jsou tak udávány ve stupních a poslední veličina v metrech. Přepočet mezi sférickými a kartézskými souřadnicemi je poměrně jednoduchý, avšak konkrétní vztahy závisí na zvoleném referenčním elipsoidu, který definuje výšku moře. Jako příklad můžeme uvést model elipsoidu WGS84 (z anglického „World Geodetic System 1984“), který využívá i satelitní systém GPS. Konkrétní vztahy pro přepočet souřadnic lze najít např. v [103].

Na druhou stranu, jiné veličiny používané v návrhu navigačního systému jsou vztaženy k více souřadným soustavám. Jako příklad veličiny vztažené ke třem souřadným soustavám zde můžeme uvést další součást navigační informace, kterou je rychlost.

- Rychlost objektu je charakterizována následujícím vektorem[4]

$$v_{\mathrm{EB}}^{\mathrm{N}} = \begin{bmatrix} v_{\mathrm{EB},x}^{\mathrm{N}} \\ v_{\mathrm{EB},y}^{\mathrm{N}} \\ v_{\mathrm{EB},z}^{\mathrm{N}} \end{bmatrix} \tag{9.1.3}$$

  kde použité značení popisuje rychlost objektu (resp. B-frame) vůči povrchu země (resp. E-frame) tak, jak se jeví pozorovateli v lokálním navigačním souřadném systému (tj. v N-frame). Navigační souřadný systém je pro vyjádření rychlosti objektu pro člověka přirozený. Jednotlivé složky vektoru rychlosti $v_{\mathrm{EB}}^{\mathrm{N}}$ (9.1.3) jsou typicky udávány v metrech za sekundu.

  Podotkněme, že horní index označuje souřadný systém, ve kterém je veličina vyjádřena. Tento systém se označuje jako referenční systém dané veličiny.

Poslední složka navigační informace je orientace objektu, která je vztažena ke dvěma souřadným soustavám udává orientaci souřadného systému objektu (tj. B-frame) vůči lokálnímu navigačnímu systému (tj. N-frame).

- Orientace objektu může být charakterizována rotační maticí $C_{\mathrm{B}}^{\mathrm{N}}$. Rotační matice je *ortonormální* matice, pro kterou platí

$$C_{\mathrm{B}}^{\mathrm{N}} = (C_{\mathrm{N}}^{\mathrm{B}})^T \tag{9.1.4}$$

$$(C_{\mathrm{B}}^{\mathrm{N}})^T = (C_{\mathrm{N}}^{\mathrm{B}})^{-1} \tag{9.1.5}$$

$$C_{\mathrm{B}}^{\mathrm{N}} C_{\mathrm{N}}^{\mathrm{B}} = C_{\mathrm{N}}^{\mathrm{N}} = I \tag{9.1.6}$$

[4]Vektor rychlosti má tři složky často označované písmeny x značící severní směr „north“, y značící východní směr „east“ a z značící směr dolů „down“. Občas se proto můžeme v literatuře setkat i s označením os N, E, D.

Rotační matice umožní změnit referenční soustavu veličiny, tj. vyjádřit daný vektor v jiné soustavě. Pokud bychom například chtěli vyjádřit vektor rychlosti $v^{\mathrm{N}}_{\mathrm{EB}}$ (9.1.3) v souřadné soustavě objektu, tj. chtěli bychom znát rychlost tak, jak by byla vnímána člověkem na palubě objektu, pak lze použít následující transformaci (přesněji rotaci)

$$v^{\mathrm{B}}_{\mathrm{EB}} = C^{\mathrm{B}}_{\mathrm{N}} v^{\mathrm{N}}_{\mathrm{EB}} \tag{9.1.7}$$

Poznamenejme, že vyjádření orientace objektu rotační maticí je vhodné z teoretického hlediska pro svoji přehlednost a možnost použití běžných maticových operací. Z pohledu aplikačního a implementačního však rotační matice nepřestavuje vhodnou volbu pro reprezentaci orientace a to ze dvou hlavních důvodů:

- Matice obsahuje 9 unikátních prvků, ačkoliv pro popis orientace objektu ve tří dimenzionálním prostoru postačují tři (Eulerovy) úhly.
- Vlivem numerických chyb dochází ke ztrátě fundamentálních vlastností rotační matice (9.1.4)–(9.1.6).

Místo rotační matice lze, při implementaci, použít pro vyjádření rotace alternativní reprezentace jakými např. jsou Eulerovy úhly, kvaterniony nebo Rodriguesovy parametry. Přepočet mezi jednotlivými reprezentacemi orientace objektu lze najít např. v [103], [118].

Poznámka . V anglicky psané literatuře se norma vektoru rychlosti označuje pojmem „speed“.

9.1.3 Stavový model

Stavový model používaný v integrovaných navigačních systémech je založen na zákonech dynamiky[5] popisující vývoj v čase hledané polohy, rychlosti a orientace navigovaného objektu[6]. Existuje mnoho modelů dynamiky, v těchto skriptech si však krátce představíme diferenciální rovnice popisující vývoj navigační informace v čase[7], které jsou vyjádřené v navigačním souřadném systému [103]

$$\begin{bmatrix} \dot{\lambda} \\ \dot{\varphi} \\ \dot{h} \\ \dot{v}^{\mathrm{N}}_{\mathrm{EB}} \\ \dot{C}^{\mathrm{N}}_{\mathrm{B}} \end{bmatrix} = \begin{bmatrix} \frac{v^{\mathrm{N}}_{\mathrm{EB},N}}{R_N(\varphi)+h} \\ \frac{v^{\mathrm{N}}_{\mathrm{EB},E}}{(R_E(\varphi)+h)\cos(\varphi)} \\ -v^{\mathrm{N}}_{\mathrm{EB},D} \\ f^{\mathrm{N}}_{\mathrm{IB}} + g^{\mathrm{N}}_{\mathrm{B}}(\varphi,h) - (\Omega^{\mathrm{N}}_{\mathrm{EN}} + 2\Omega^{\mathrm{N}}_{\mathrm{IE}}) v^{\mathrm{N}}_{\mathrm{EB}} \\ C^{\mathrm{N}}_{\mathrm{B}} \Omega^{\mathrm{B}}_{\mathrm{NB}} \end{bmatrix} \tag{9.1.8}$$

kde $R_N(\varphi)$ a $R_E(\varphi)$ reprezentuje poloměry křivosti elipsoidu aproximujícího tvar Země, $g^{\mathrm{N}}_{\mathrm{B}}(\varphi,h)$ značí vektor gravitačního pole, matice $\Omega^{\mathrm{N}}_{\mathrm{IE}}$ je antisymetrická matice vyjadřující rotaci Země (tj. systému ECEF vůči ECI) vyjádřeného v navigačním souřadném systému, tj. je funkcí zeměpisné šířky, a $\Omega^{\mathrm{N}}_{\mathrm{EN}}$ je antisymetrická matice vyjadřující rotaci navigačního souřadného systému vůči ECEF způsobenou pohybem objektu. Matice úhlových rychlostí $\Omega^{\mathrm{N}}_{\mathrm{IE}}$ je tak v anglicky psané

[5] Dynamické zákony popisují dynamiku objektu v závislosti na jeho příčinách, např. při daných působících silách na objekt, je určen jeho pohyb. Na druhou stranu, kinematické modely popisují vztah mezi pozicí, rychlostí a případně i akcelerací bez zkoumání příčin pohybu.

[6] Trojice veličin pozice, rychlost a orientace je často označována jako navigační informace.

[7] Explicitní závislost na čase, tj. např. $\lambda = \lambda(t)$, je v následující rovnicích vynechána z důvodu kompaktnosti zápisu.

literatuře označována pojmem „Earth rate“ a rotace $\Omega^{\mathrm{N}}_{\mathrm{EN}}$ jako „transport rate“. Přesné vztahy pro výpočet těchto veličin lze najít např. v [103].
Zbývající veličiny reprezentují specifickou sílu a úhlovou rychlost

$$f^{\mathrm{N}}_{\mathrm{IB}} = C^{\mathrm{N}}_{\mathrm{B}} f^{\mathrm{B}}_{\mathrm{IB}} \tag{9.1.9}$$

$$\Omega^{\mathrm{B}}_{\mathrm{NB}} = \Omega^{\mathrm{B}}_{\mathrm{IB}} - (\Omega^{\mathrm{B}}_{\mathrm{IE}} - \Omega^{\mathrm{B}}_{\mathrm{EN}}) \tag{9.1.10}$$

kde $\Omega^{\mathrm{B}}_{\mathrm{IE}} = C^{\mathrm{B}}_{\mathrm{N}}\Omega^{\mathrm{N}}_{\mathrm{IE}}C^{\mathrm{N}}_{\mathrm{B}}$, $\Omega^{\mathrm{B}}_{\mathrm{EN}} = C^{\mathrm{B}}_{\mathrm{N}}\Omega^{\mathrm{N}}_{\mathrm{EN}}C^{\mathrm{N}}_{\mathrm{B}}$, $f^{\mathrm{B}}_{\mathrm{IB}}$ je *specifická síla* v souřadném systému objektu přímo měřená akcelerometrem (přesněji ortogonálně umístěnou trojicí akcelerometrů měřící zrychlení v každé ze tří os kartézského systému) a $\Omega^{\mathrm{B}}_{\mathrm{IB}}$ je *úhlová rychlost* vyjádřená ve formě antisymetrické matice založené na měření z gyroskopu (přesněji ze soustavy tří gyroskopů). Antisymetrickou matici úhlové rychlosti $\Omega^{\mathrm{B}}_{\mathrm{IB}}$ lze z vektoru přímo měřené úhlové rychlosti $\omega^{\mathrm{B}}_{\mathrm{IB}} = [\omega^{\mathrm{B}}_{\mathrm{IB},x}, \omega^{\mathrm{B}}_{\mathrm{IB},y}, \omega^{\mathrm{B}}_{\mathrm{IB},z}]^T$ získat ze vztahu

$$\Omega^{\mathrm{B}}_{\mathrm{IB}} = \begin{bmatrix} 0 & -\omega^{\mathrm{B}}_{\mathrm{IB},z} & \omega^{\mathrm{B}}_{\mathrm{IB},y} \\ \omega^{\mathrm{B}}_{\mathrm{IB},z} & 0 & -\omega^{\mathrm{B}}_{\mathrm{IB},x} \\ -\omega^{\mathrm{B}}_{\mathrm{IB},y} & \omega^{\mathrm{B}}_{\mathrm{IB},x} & 0 \end{bmatrix} \tag{9.1.11}$$

Model (9.1.8)–(9.1.11) bude naprosto přesně popisovat dynamiku navigovaného objektu za předpokladu, že všechny veličiny jakými jsou např. poloměry křivosti, vektor gravitačního pole, rychlost rotace Země, budou přesně známy a inerciální měření, tj. $f^{\mathrm{B}}_{\mathrm{IB}}$ a $\omega^{\mathrm{B}}_{\mathrm{IB}}$, budou přesná, tj. neovlivněná šumem. V reálném aplikacích však tento předpoklad neplatí a namísto skutečných veličin musíme pracovat s jejich odhady a namísto přesných měření s reálnými měřeními, které jsou ovlivněny šumem. Proto, je dynamický model (9.1.8) nutné psát ve formě [103], [118]

$$\dot{x}(t) = f(x(t), u(t)) + w(t) \tag{9.1.12}$$

V rovnici (9.1.12) $x(t) \in \mathbb{R}^{n_x}$ značí (hledaný) stavový vektor obsahující navigační informaci, tj.,

$$x(t) = \begin{bmatrix} \lambda \\ \varphi \\ h \\ v^{\mathrm{N}}_{\mathrm{EB}} \\ q^{\mathrm{N}}_{\mathrm{B}} \end{bmatrix} \tag{9.1.13}$$

kde $q^{\mathrm{N}}_{\mathrm{B}}$ je vhodné vektorové vyjádření rotační matice $C^{\mathrm{N}}_{\mathrm{B}}$, např. výše zmíněný kvaternion, $u(t) \in \mathbb{R}^{n_u}$ značí (dostupný) vektor inerciálních měření, tj.,

$$u(t) = \begin{bmatrix} f^{\mathrm{B}}_{\mathrm{IB}} \\ \omega^{\mathrm{B}}_{\mathrm{IB}} \end{bmatrix} \tag{9.1.14}$$

a $w(t) \in \mathbb{R}^{n_x}$ je vektor šumu, jehož vlastnosti odpovídají použitým aproximacím a vlastnostem senzorů.

Přestavený model dynamiky stavu (9.1.12) je ve formě nelineární diferenční rovnice. Tu lze převést do nám již známé diskrétní formy (4.1.1) např. pomocí Eulerovy nebo Rungeovy-Kuttovy metody s ohledem na periodu vzorkování inerciálních senzorů.

Poznámka . Inerciální, a i jiné v navigaci používané, senzory jsou ovlivněny nejen bílým šumem reprezentovaným vektorem $w(t)$, ale i šumem barevným, tj. šumem v čase korelovaným. Pro tento typ šumu se vžil pojmem „bias“. Bias senzorů se typicky odhaduje spolu

s navigační informací. Je tedy nutné znát relativně přesný dynamický model biasu, který bývá ve formě Gaussovského-Markovského procesu. V případě přesnějších inerciálních senzorů je vhodné modelovat i chybu modelu gravitačního pole jako barevnou a též ji odhadovat [119, 118].

Poznámka . Připomeňme, že některé estimační techniky, např. rozšířený Kalmanův filtr, jsou založeny na linearizaci nelineární funkce ve stavovém modelu. Nabízí se tedy dvě možnosti, jak přejít od nelineární diferenciální rovnice typu (9.1.12) k rovnici lineární diferenční používané např. Kalmanovým filtrem, a to [120]

- nejprve model (9.1.12) diskretizovat a pak použít vhodnou linearizaci (např. Taylorův rozvoj nebo Stirlingovu interpolaci) nebo
- nejprve model linearizovat a pak jej exaktně zdiskretizovat.

Každý z přístupů na své výhody a nevýhody a je vhodný pro rozdílné aplikace.

Inerciální měření v navigačních systémech je tedy uvažováno jako vstup $u(t)$ stavové rovnice (9.1.12). Integrovaný navigační systém má však k dispozici ještě (přinejmenším) jeden senzor, kterým je přijímač satelitního navigačního systému. Ten může poskytovat buď přímo pozici a rychlost navigovaného objektu nebo tzv. pseudo-vzdálenosti[8] mezi navigovaným objektem (s neznámou pozicí) a satelity (se známými pozicemi). V závislosti na zvoleném výstupu rozlišujeme dva *základní* typy integrovaného navigačního systému, a to [103, 118]:

- integrovaný navigační systém s *volnou* vazbou (v anglicky psané literatuře označované jako „loosely coupled integrated navigation system“) a
- integrovaný navigační systém s *pevnou* vazbou (v anglicky psané literatuře označované jako „tightly coupled integrated navigation system“).

I přesto, že druhý jmenovaný typ navigačního systému může poskytovat kvalitnější odhady navigační informace, zaměříme se zde, z důvodu přehlednosti, na představení prvně jmenovaného typu integrace. Ta je založena na využití odhadu polohy (a případně rychlosti) poskytované GNSS přijímačem. Odhad polohy navigovaného objektu, který je poskytován GNSS přijímačem, můžeme chápat jako dostupné měření z_k. Za předpokladu, že GNSS přijímač poskytuje odhad polohy ve formě zeměpisné délky λ, zeměpisné šířky φ a nadmořské výšky h, lze rovnici měření v časovém okamžiku $k(= t_k)$ zapsat ve formě

$$\begin{aligned} z_k &= r^{\mathrm{E}}(t_k) + v_k = r_k^{\mathrm{E}} + v_k \\ &= H_k x_k + v_k \end{aligned} \tag{9.1.15}$$

kde matice měření je

$$H_k = [I_3, 0_{3,n_x-3}] \tag{9.1.16}$$

a v_k je šum vyjadřující nepřesnost v odhadu polohy GNSS přijímače.

Rovnice (9.1.12), (9.1.15) formují stavový model, který může být v principu použit pro návrh integrovaného navigačního systému. Všimněme si, že model je nelineární a tedy pro odhad stavu (tj. navigační informace) musíme použít nelineární estimační techniku.

[8] Krátkou diskuzi k pseudo-vzdálenostem v GNSS navigaci lze najít v prvním díle skript v kapitole věnované identifikaci nelineárních systémů.

Poznámka . V anglicky psané literatuře z oblasti návrhu navigačních systémů se měření z_k často nazývá pojmem „aiding source“.

Poznámka . Integrovaný navigační systém může využívat i jiné senzory jakými např. jsou barometrický nebo radarový výškoměr, magnetometr, GNSS přijímač s více anténami nebo Pitotova trubice.

9.1.4 Estimační algoritmy a aplikace

V předchozích kapitolách skript bylo představeno velké množství nelineárních estimačních algoritmů, které mohou být (a typicky jsou) použity pro odhad navigační informace o objektu [103], [118], [119], [124]. Od 70. let minulého století, se však v oblasti letectví převážně využívá *rozšířený Kalmanův filtr*, který byl představen v kapitole 4.1. Použití rozšířeného Kalmanova filtru v oblasti návrhu integrovaných navigačních systémů pro civilní letectví se řídí standardem DO-229D spravovaným institucí „Radio Technical Commission for Aeronautics (RTCA)“ [86], kde lze najít detaily k implementaci a především k validaci leteckých navigačních systémů.

V literatuře lze najít nepřeberné množství článků a patentů ilustrujících návrh a kvalitu odhadu integrovaných navigačních systémů v závislosti na dostupných senzorech, použitých estimačních algoritmech a operačních podmínkách [103], [118], [119], [121]-[124]. Například v článcích [121], [126] lze najít analýzu kvality odhadu inerciálního navigačního systému založeného na rozšířeném Kalmanově filtru na základě reálných dat pro trajektorie odpovídající standardům pro testování leteckých navigačních systémů. Výsledky ukazují, že, v závislosti na použitých senzorech, lze, pro duální GNSS přijímač, očekávat chybu odhadu pozice v řádu jednotek metrů a chybu odhadu orientace kolem půl stupně.

9.2 Terénní navigace

V předchozí části představený integrovaný navigační systém je v současné době pravděpodobně nejrozšířenější navigační systém. Za předpokladu dostupnosti signálů ze satelitní navigace poskytuje přesný a konzistentní odhad navigační informace. Bohužel, signál satelitní navigace je velmi náchylný k rušení, ať už

- úmyslného, kdy signál je záměrně rušen jiným zařízením (zde se můžeme setkat s pojmy „jamming, spoofing“) nebo
- neúmyslného, kde signál je zcela blokován nebo částečně ovlivněn např. stromy, vysokými budovami nebo tunely (zde se můžeme setkat s pojem „multipath“) nebo ovlivněn přírodními úkazy jako např. solární bouří.

Pokud je satelitní signál rušen, pak integrovaný navigační systém *nemůže* poskytovat dostatečně přesný nebo konzistentní odhad navigační informace. V tomto případě je tak nutné použít alternativní navigační systém, jakým je terénní navigace.
Terénní navigační systém, v anglicky psané literatuře označovaný pojmem „terrain-aided navigation“, je založen jen na senzorech pevně spojených s navigovaným objektem a nepřijímající žádný externí signál a terénní mapě. Často používané senzory v terénní navigaci tak jsou výškoměry, odometry nebo inerciální senzory [99], [103], [104], [117].

9.2.1 Stavový model

V literatuře lze najít poměrně velké množství přístupů k návrhu stavového modelu pro terénní navigační systém. Zde si představíme poměrně jednoduchý přístup vhodný pro automobily založený na kinematických zákonitostech a předpokladu lineárního pohybu [99], [117]. Budeme předpokládat, že máme k dispozici dva senzory, kterými jsou

- barometrický *výškoměr* poskytující měření nadmořské výšky navigovaného automobilu,
- *odometr* poskytující informaci o relativním pohyb automobilu mezi dvěma měřeními výškoměru (mezi dvěma po sobě jdoucími časovými okamžiky).

Předpokládejme, že cílem je nalézt pouze odhad *horizontální* pozice r_k^{N} a rychlosti v_k^{N} navigovaného objektu vyjádřených v kartézském lokálním souřadném systému, tj. definujme stav $x_k \in \mathbb{R}^{n_x}$, kde $n_x = 4$, v časovém okamžiku k jako

$$\begin{aligned} x_k &= [r_{k,N}^{\mathrm{N}}, r_{k,E}^{\mathrm{N}}, v_{k,N}^{\mathrm{N}}, v_{k,E}^{\mathrm{N}}]^T \\ &= [x_{1,k}, x_{2,k}, x_{3,k}, x_{4,k}]^T \end{aligned} \tag{9.2.17}$$

Pak, na základě kinematických zákonitostí můžeme definovat následující stavový model modelující (popisující) dynamiku vozidla

$$x_{k+1} = F_k x_k + \begin{bmatrix} u_k \\ 0_{2\times 1} \end{bmatrix} + w_k \tag{9.2.18}$$

kde matice dynamiky je

$$F_k = \begin{bmatrix} 1 & 0 & \Delta t_k & 0 \\ 0 & 1 & 0 & \Delta t_k \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{9.2.19}$$

a $\Delta t_k = t_{k+1} - t_k$ značí periodu vzorkování. Perioda vzorkování se může v čase měnit. Vstupní signál u_k je relativní změna pozice měřená odometrem a nejistota tohoto měření je modelována stavovým šumem w_k. Poznamenejme, že v literatuře je tento model označován jako model pohybu s téměř konstantní rychlostí (v anglicky psané literatuře jako „nearly constant velocity motion model“) [87].
Rovnici měření, která dává do vztahu měřenou veličinu z_k a neznámý stav x_k, lze zapsat jako

$$z_k = h_k(x_{1,k}, x_{2,k}) + v_k \tag{9.2.20}$$

kde z_k nadmořská výška vozidla měřená výškoměrem, $h_k(\cdot)$ je známá výšková mapa terénu a v_k je šum měření zahrnující jak samotnou chybu měření výškoměru, tak i možnou chybu terénní mapy. Všimněme si, že rovnice měření reprezentuje mapu, která nám určuje vztah mezi měřenou nadmořskou výškou a hledanou horizontální pozicí vozidla. Poznamenejme, že mapu terénu lze vnímat jako nelineární funkci horizontální pozice, která je však ve formě tabulky a ne analytické funkce, jak jsme tomu byli doposud zvyklí v předcházejících kapitolách.
Rovnice (9.2.18), (9.2.20) formují nelineární stavový model, který může být v principu použit pro návrh terénního navigačního systému.

Poznámka . Kromě použití výškoměru a terénní mapy, lze v literatuře najít i terénní navigaci založenou na porovnání měřeného magnetického, resp. gravitačního vektoru s odpovídající

mapou magnetického, resp. gravitačního pole [125].

Poznámka . Hlavní výhodou terénního navigačního systému je jeho nezávislost na externě vysílaných signálech. Avšak, terénní navigační systém bude dobře fungovat v oblastech, kde je „bohatý“ terén. V místech, kde je terén téměř nebo zcela rovný, nelze tento navigační systém použít.

9.2.2 Estimační algoritmy a aplikace

Představený model (9.2.18), (9.2.20) má lineární dynamiku, avšak „silně“ nelineární funkci v rovnici měření. Proto pro odhad stavu x_k je vhodné použít globální filtr, jakým je například metoda bodových mas představená v kapitole 7, popř. filtr s vícenásobnou linearizací diskutovaný v kapitole 5.

Příklady návrhu terénního navigačního systému založeného na metodě bodových mas spolu s vyhodnocením přesnosti a konzistence odhadu navigační informace na základě reálných i simulovaných dat lze najít např. v [117], [104]. Výsledky ukazují, že, v závislosti na přesnosti mapy, použitých senzorech a variabilitě krajiny, lze očekávat chybu odhadu horizontální pozice okolo 20 metrů.

Kapitola 10

Závěr

Úloha odhadu stavu na základě dostupných měření a dalších informací obsažených v modelu systému je významná nejen pro zjišťování neznámých veličin, rozhodování, monitorování, ale rovněž tvoří často jádro syntézy řídicích systémů. Úspěšnost, ale i složitost řešení estimační úlohy jsou výrazně určeny matematickým modelem systému, a proto modelování a estimace jsou tak těsně spojeny.

Charakteristickým rysem tohoto dílu skript je snaha o ucelený výklad problému modelování a odhadu stavu stochastických systémů. Základním omezením úlohy estimace bylo úzké propojení stavby modelu a způsobu řešení s cílem zajistit nejen analytické, ale i numerické řešení.

Opěrným bodem pro syntézu analytických estimačních algoritmů byla kalmanovská filtrace. Její detailní popis pro lineární gaussovský případ zahrnující odvození a propojení z bayesovským přístupem byl s výhodou využit i pro složitější úlohy nelineární filtrace překračující původní rámec linearity a gaussovosti.

Druhým opěrným bodem byla Magilova myšlenka [41] o konečném počtu možných hodnot parametrů použitá v adaptivním řízení a idea aproximovat hustoty pravděpodobnosti náhodných veličin součtem normálních rozložení [48], [39]. Na tomto základě byla prováděna postupně syntéza algoritmů nelineární estimace pro řadu speciálních případů až po nelineární negaussovské případy. Velmi zajímavé uplatnění estimačních algoritmů bylo uvedeno pro lineární negaussovské systémy, které umožňují přirozeně modelovat prakticky významné jevy jako hrubé chyby měření nebo skokové změny parametrů či stavu a vytvářejí přirozenou základnu pro řešení celé škály dalších speciálních úloh. Nejmarkantněji se možnosti tohoto typu modelování a estimace projevuje v oblasti rozhodování, detekce chyb, testování hypotéz a dalších oblastech např. při adaptivním zpracování signálů a adaptivním řízení.

Jako alternativa k analytickému návrh bayesovského filtru, bylo představeno numerické řešení Bayesových rekurzivních vztahů. To vedlo na metodu bodových mas, jejíž základní myšlenka spočívá v aproximaci hustot pravděpodobnosti po částech konstantními hustotami.

Pozoruhodnou skutečností předkládaných algoritmů nelineární filtrace je jejich přirozená možnost paralelní implementace, což je, v současné době, velmi rozvíjená oblast jak po stránce softwarové, tak i hardwarové.

Literatura

[1] Wiener, N.: The Extrapolation, Interpolation and Smoothing of Stationary Time Series. New York, John Wiley, 1949.

[2] Kalman, R.E.: New Results in Linear Filtering and Prediction Theory. Trans. ASME, J. Basic Eng. 83D, 1960 s. 35-45.

[3] Kushner, H.J.: Approximation to optimal non-linear filters. IEEE Trans. on Automatic Control, 12, 1967, s. 546-556.

[4] Jazwinski, A.H.: Stochastic processes and filtering theory, New York, Academic Press 1970, 376 s.

[5] Sage, A.P.-Melsa, J.L.: Estimation Theory with Applications to Communications and Control. New York, Mc Graw - Hill 1971, 529 s.

[6] Proceedings of First Symposium on Nonlinear Estimation Theory and Its Applications. San Diego, 1970.

[7] Proceedings of Second Symposium on Nonlinear Estimation Theory and Its Applications. San Diego, 1971.

[8] Ahlén, A.-Sternad, M.: Wiener Filter Design Using Polynomial Equation. Uppsala University, UPTEC 90057R, 1990.

[9] Havlena,V.-Štecha, J.:Moderní teorie řízení. Skriptum ČVUT Praha, 1994, 289s.

[10] Anderson, B.-Moore J.: Optimal Filtering. Prentice Hall, 1979, 354s.

[11] Chui, C.K.-Chen, G.: Kalman Filtering with Real-Time Applications. Second Edition, Springer Verlag, 1990, 192s.

[12] Söderström, T.: Discrete-time stochastic systems: Estimation & Control. Prentice Hall, London, 1994, 333s.

[13] Aoki, M.: State Space Modeling of Time Series. Springer Verlag, Berlin, Heidelberg, 1987, 311s.

[14] Sorenson, H.W.: On the development of practical nonlinear filters. In: D.G. Lainiotis (Ed.), Estimation Theory. American Elsevier, New York, 1974.

[15] Šimandl, M.-Mošna, J.: Analytický přístup k řešení bayesovských vztahů. In: Sborník semináře kateder kybernetiky a automatizace ČSSR, ČSVTS Plzeň, 1988.

[16] Sorenson, H.W.: Parameter Estimation. Marcel Dekker, New York, 1980, 382s.

[17] Schweppe, F.C.: Uncertain Dynamic Systems. Prentice Hall, Engelwood Cliffs, NJ, 1968.

[18] Milanese, M.-Vicino, A.: Optimal Estimation Theory for Dynamic Systems with Set Members Ship Uncertainty. Automatica, Vol. 27, 1991, 997-1009.

[19] Žampa, P.: Teorie kauzálních systémů, GR 140, KKY, ZČU v Plzni, 1992.

[20] Průcha, J.-Šimandl, M.-Škarda, Z.: The Kalman Filter Approach to Improve of Surface Temperature Diagnostic for Plasma Sprayed Particles. In: Acta Technica ČSAV, Academia, 1991.

[21] Brockwell, P.J.-Davis, R.A.: Time Series: Theory and Methods. Springer Verlag, 1991, 577s.

[22] Ng, C.N.-Young, P.C.: Recursive Estimation and Forecasting of Non-stationary Time Series. Journal of Forecasting, Vol. 9, 1990, 173-204.

[23] Strejc, V.: Stavová teorie lineárního diskrétního řízení, Academia, Praha 1978, 374s.

[24] Aström, K.J.: Introduction to Stochastic Control Theory. Academic Press 1970, 299s.

[25] Kubík, S.-Kotek, Z.-Strejc, V.-Štecha, J.: Teorie automatického řízení I. SNTL, Praha, 1982, 522s.

[26] Box, G.-Jenkins, G.: The Series Analysis Forecasting and Control. Holden-Day, San Francisco, 533s.

[27] Widrow, B.-Stearns, S.D.: Adaptive Signal Processing. Prentice Hall, London, 474s.

[28] Ljung, L.-Söderström, T.: Theory and Practice of Recursive Identification. MIT Press, Cambridge, 1983.

[29] Spall, J.C.: Bayesian Analysis of Time Series and Dynamic Models. Marcel Dekker, New York, 1988.

[30] Šimandl, M.: Adaptivní systémy. ZČU v Plzni. Skripta, 1993, 138s.

[31] Šimandl, M.: Využití některých identifikačních metod v praxi. In: Technická kybernetika a biokybernetika. ČSVTS-FEL-ČVUT, Temešvár, 1991, 18-28.

[32] Beneš, J.-Žampa, P.: Stochatické systémy a jejich řízení. ČVUT Praha, Skripta, 1976, 201s.

[33] Hrušák, J.-Mošna, J.-Janeček, E.-Šimandl, M.: A New Adaptive Controller for Cold Rolling Mills. In: Proc. of IFAC/IFIP Symposium on Real Time Digital Control Application, Guadalajara, Mexico, 1983, Oxford, England: Pergamon, 1984, 69-74.

[34] Hrušák, J.-Mošna, J.-Janeček, E.-Šimandl, M.: Strip-Thickness Adaptive Control for Cold-Rolling Mills. Problems of Control and Information Theory, 11, No 6, 1982, 455-464.

[35] Šimandl, M.: Teorie a aproximační algoritmy bayesovské filtrace. KKY, VŠSE v Plzni, 1982, 71s.

[36] Krebs, V.: Nichtlienare Filtering. Oldenbourg, München, 1980.

[37] Štecha, J.: Parameter Estimation in Nonlinear Economic Models. In: Modern Control Theory. International Summer School '92. UTIA Prague and Czech Technical University, Prague 1992.

[38] Zellner, A.: Bayesian econometrics. Econometrica. Vol. 53, No. 2, March 1985, 253-269.

[39] Alspach, D.L.-Sorenson, H.W.: Approximation of Density Function by a Sum of Gaussians for Nonlinear Bayesian Estimation. In: Proceedings of the 1st Symposium on Nonlinear Estimation Theory, San Diego, 1970.

[40] Sorenson, H.W.-Alspach, D.L.: Recursive Bayesian Using Gaussian Sums. Automatica, Vol. 7, 1971, 465-479.

[41] Magil, D.T.: Optimal adaptive estimation of sampled stochastic processes. IEEE Trans. Automatic Control, Vol. AC-10, Oct. 1965, 434-439.

[42] Lainiotis, D.G.: Partioning - A Unifying Framework for Adaptive Systems, In: Estimation. Proc. of the IEEE, 64, No. 8, 1976, 1126-1143.

[43] Sorenson, H.W.: An Overview of Filtering and Stochastic Control in Dynamic Systems. Advances in Theory and Applications. C.T. Leondes, New York, Academic Press, 1976, 1-61.

[44] Šimandl, M.: Nonlinear Filtering: Set Approach. Uppsala University, Uppsala, UPTEC 8458R, 1984.

[45] Hrušák, J.-Mošna, J.- Šimandl M.: Nonlinear Filtering: System Set Reprezentation. In: Preprints of 10th World Congress IFAC, München, 1987.

[46] Mošna, J.-Šimandl, M.: Číslicové monitorovací systémy a bayesovská filtrace. In: Sborník semináře kateder kybernetiky a automatizace ČSSR, ČSVTS Plzeň, 1988.

[47] Mošna, J.-Šimandl, M.: Bayesovská filtrace a množinová reprezentace systému. Automatizace, SNTL Praha, 37, č. 7, 1989, 182–184.

[48] Lo, J.T.: Finite - dimensional sensor orbits and optimal nonlinear filtering. IEEE Trans. Inform. Theory, IT18, 1972.

[49] Bultas, M.: Aproximační metody nelineární filtrace. Diplomová práce. KKY VŠSE v Plzni 1983.

[50] Šimandl, M.-Prantner, V.: Zobecněný rozšířený Kalmanův filtr. Rukopis VZ, KKY, VŠSE v Plzni. 1988.

[51] Alspach, D.L.-Sorenson, H.W.: Nonlinear Bayesian Estimation Using Gaussian Sum Approximations. Trans. on Automatic Control, Vol. AC-17, No. 4, August 1972.

[52] Šimandl, M.: Syntéza algoritmů nelineární estimace pro aplikace v reálném čase. Kand. dis. práce, KKY, VŠSE v Plzni, 1986.

[53] Alspach, D.L.: Dual Control on Approximate Aposteriori Density Functions. IEEE Trans. on Automatic Control, October 1972.

[54] Dajani, M.Z.-Campion, G.: Closed loop control design for nonlinear nonquadratic systems. In: Proc. of the IEEE Conference on Decision and Control. San Diego, 1973, 82–87.

[55] Namera, T.-Stubberund, A.: Gaussian Sum Approximation for Nonlinear Fixed Point Prediction. Int. J. Control, Vol. 38, No. 5, 1983, 1047-1053.

[56] Lainiotis, D.G.: Partitioning: A Unifying Framework for Adaptive Systems II. Control. Proceedings of the IEEE 64, 1976, 1182–1198.

[57] Kárný, M.-Hangos, K.H.: Approximation of the Bayes rule. In: Preprints of the 7th IFAC/IFORS Symposium on Identification and System Parameter Estimation, York, England, Vol. 1, 1985, 785–990.

[58] Kulhavý, R.: A Bayes-closed approximation of recursive nonlinear estimation. Int. J. Adaptive Control and Signal Processing, 4, 1990, 271–285.

[59] Kulhavý, R.: Differential geometry of recursive nonlinear estimation. In: Preprints of the 11th IFAC World Congress, Tallin, Estonia, Vol. 3, 1990, 113-118.

[60] Kulhavý, R.: Recursive nonlinear estimation: geometry of a space of posterior densities. Automatica, 28, 1992, 313-323.

[61] Šimandl, M.-Mošna, J.: Nelineární estimátor pro lineární negaussovský systém. In: Automatická regulace a logické řízení 87. ČSVTS Škoda Plzeň, Žinkovy. 1987.

[62] Šimandl, M.-Mošna, J.: Detekce chyby a nelineární filtrace. In: Automatická regulace a logické řízení 89. ČSVTS Škoda Plzeň, Žinkovy, 1989.

[63] Šimandl, M.: State Estimation for Nongaussian models. VZ, KKY, ZČU v Plzni, 1992.

[64] Kárný, M.: Lokální filtr necitlivý k hrubým chybám měření. In: ASž TP '90 Automatizácia a systémy riadenia technologických procesov. ČSVTS, Ústav racionalizace průmyslu, Žilina, 1990, 141-144.

[65] Tanaka, M.-Kalayana, T.: Identification and Smoothing for Linear System with Outliers and Missing Data. In: Preprints of World Congress IFAC, Vol. 3, Tallin, 1990.

[66] Šrubař, P.: Negaussovské modelování a odhad stavu. Dipl. práce, KKY, ZČU v Plzni, 1992.

[67] Šonka, M.: Zapomínání při odhadu časově proměnných parametrů. Dipl. práce. KKY ZČU v Plzni. 1992.

[68] Kulhavý, R.-Kárný, M.: Tracking of Slowly varying Parameters by Directional Forgetting. In: Preprints of 9th World Congress IFAC, Budapest, 1984, 687–692.

[69] Parkum, J.E.-Poulsen, N.K.-Holst, J.: Selective Forgetting in Adaptive Procedures. In: Preprints of 11th World Congress IFAC, Tallin, 1991, 180–186.

[70] Löhnberg, P.-Stienstta, A.: Time Varying Parameter Estimation Combining Directional and Uniform Forgetting. In: Preprints of 11th World Congress IFAC, Tallin, 1991, 232-237.

[71] Benveniste, A.-Basseville, M.-Moustakides, G.: Modelling and Monitoring of Changes in Dynamical Systems. In: Proceedings of 25th Conference on Decision and Control. Athens, Grece, December 1986, 776–782.

[72] Campo, L.-Bar-Shalom, Y.: A New Controller for Discrete-Time Stochastic Systems with Markovian Jump Parameters. In: Preprints of World Congress IFAC, Vol. 3, Tallin, 1990.

[73] Ljung, L.: Optimal and ad hoc Adaption Mechanismus. In: Proceedings ECC 91, European Control Conference, Grenoble, France, July 1-5, 1991, 2013–2020.

[74] Hägglund, T.: New Estimation Techniques for Adaptive Control. Dissertation, Lund University, 1983.

[75] Chrástecký: Detekce poruch v dynamických systémech. Dipl. práce, KKY, ZČU v Plzni, 1992.

[76] Zhang, X.J.: Auxiliary Signal Design in Fault Detection and Diagnosis. Springer Verlag, Berlin, 1989.

[77] Aström, K.J.-Wittenmark, B.: Adaptive Control. Addison-Wesley. New York, 1989.

[78] Aström, K.J.: Intelligent Control. In: Proceedings of ECC 91, European Control Conference, Grenoble, France, July 2-5, 1991, 2328-2339.

[79] Kárný, M.-Halousková, A.-Böhm, J.-Kulhavý, R.-Nedoma, P.: Design of linear quadratic adaptive control: theory and algorithms for practice. Kybernetika, Příloha k číslům 3,4,5,6, Vol. 21, 1985.

[80] Wenk, C.J.-Bar-Shalom, Y.: A Multiple Model Adaptive Dual Control Algorithm for Stochastic Systems with Unknown Parameters. IEEE Transactions on Automatic Control, Vol. AC-25, No. 4, Aug. 1980.

[81] Qi, Xiao-Jiang: A multi-model adaptive predictor for stochastic processes with Markov switching parameters. Int. J. Control, Vol. 43, No. 5, 1986, 1453–1463.

[82] Chizech, H.J.-Wilski, A.S.-Castanon, D.: Discrete time markovian-jump linear quadratic optimal control, Int. J. Control, Vol. 43, No. 1, 1986, 213-231.

[83] Kučera, V.: Algebraická teorie diskrétního řízení. Academia, Praha, 1978.

[84] Lewis, F.L.: Optimal Estimation. John Wiley, New York, 1986.

[85] Šimandl, M.: Kauzální stochastické systémy a odhad stavu. Výzkumná zpráva, ZČU KKY, 1993.

[86] RTCA: DO-229D Minimum operational performance standard for global positioning system / wide area augmentation system airborne equipment. Radio Technical Commission for Aeronautics, Standard, Dec. 2016.

[87] Bar-Shalom, Y.-Li, X.R.-Kirubarajan, T.: Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software. John Wiley, 2001.

[88] Särkkä, S.: Bayesian Filtering and Smoothing. Cambridge University Press, 2013.

[89] Punčochář, I.: Základy identifikace systémů a detekce chyb, Učební text k předmětu Základy identifikace systémů a detekce chyb, ZČU v Plzni, 2016.

[90] Nørgaard, M.-Poulsen, N.K.-Ravn, O.: New developments in state estimation for nonlinear systems, Automatica, vol. 36, no. 11, 2000, 1627–1638.

[91] Julier, S.J.-Uhlmann, J.K.: Unscented filtering and nonlinear estimation, IEEE Proceedings, vol. 92, no. 3, 2004, 401–421.

[92] Arasaratnam, I.-Haykin, S.: Cubature Kalman filters, IEEE Transactions on Automatic Control, vol. 54, no. 6, 2009, 1254–1269.

[93] Šimandl, M.-Duník, J.: Derivative-free estimation methods: New results and performance analysis, Automatica, vol. 45, no. 7, 2009, 1749–1757.

[94] Duník, J.-Šimandl, M.-Straka, O.: Unscented Kalman filter: Aspects and adaptive setting of scaling parameter, IEEE Transactions on Automatic Control, vol. 57, no. 9, 2012, 2411–2416.

[95] Duník, J.-Straka, O.-Šimandl, M.-Blasch, E.: Random-point-based filters: Analysis and comparison in target tracking, IEEE Transactions on Aerospace and Electronic Systems, vol. 51, no. 2, 2015, 303–308.

[96] Simon, D.: Optimal State Estimation: Kalman, H-infinity, and Nonlinear Approaches, CRC Press, 2012.

[97] Gibbs, B.: Advanced Kalman Filtering, Least-Squares and Modelling, Wiley, 2011.

[98] Bucy, R.S.-Senne, K.D.: Digital synthesis of nonlinear filters. Automatica, vol. 7, 1971, 287–298.

[99] Bergman, N.: Recursive Bayesian estimation: Navigation and tracking applications, Disertační práce, Linköping University, Sweden, 1999.

[100] Královec, J.: Metoda bodových mas v úloze nelineární filtrace, Disertační práce, Západočeská univerzita v Plzni, 2002.

[101] Sirola, N.: Nonlinear filtering with piecewise probability densities, Disertační práce, Tampere University of Technology, Finland, 2007.

[102] Šimandl, M.-Královec, J.-Söderström, T.: Anticipative grid design in point-mass approach to nonlinear state estimation, IEEE Transactions on Automatic Control, vol. 47, no. 4, 2002, 699–702.

[103] Groves, P.D.: Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems. Artech House, 2008.

[104] Duník, J.-Soták, M.-Veselý, M.-Hawkinson, W. J.: Apparatus and method for data-based referenced navigation, US Patent (US10444269 B2), 2019.

[105] Schön, T.-Gustafsson, F.-Nordlund, P.: Marginalized particle filters for mixed linear/nonlinear state-space models, IEEE Transactions on Signal Processing, vol. 53, no. 7, 2005, 2279—2289.

[106] Šmídl, V.-Gašperin, M.: Rao-Blackwellized point mass filter for reliable state estimation, in Proceedings of the 16th International Conference on Information Fusion, Istanbul, Turkey, 2013.

[107] Lindfors, M.-Hendeby, G.-Gustafsson, F.-Karlsson, R.: Vehicle speed tracking using chassis vibrations, in Proceedings of the 2016 IEEE Intelligent Vehicles Symposium, Gothenburg, Sweden, Jun. 2016.

[108] Mehra, R.K.: On the identification of variances and adaptive filtering, IEEE Transactions on Automatic Control, vol. 15, no. 2, 1970, 175—184.

[109] Duník, J.-Straka, O.-Kost, O.-Havlík, J.: Noise covariance matrices in state-space models: A survey and comparison - part I, International Journal of Adaptive Control and Signal Processing, vol. 31, no. 11, 2017, 1505–1543.

[110] Odelson, B.J.-Rajamani, M.R.-Rawlings, J.B.: A new autocovariance least-squares method for estimating noise covariances, Automatica, vol. 42, no. 2, 2006, 303—308.

[111] Duník, J.-Kost, O.-Straka, O.: Design of measurement difference autocovariance method for estimation of process and measurement noise covariances, Automatica, vol. 90, 2018, 16–24.

[112] Brewer, J.W.: Kronecker products and matrix calculus in system theory, IEEE Transactions on Circuits and Systems, vol. 25, no. 9, 1978, 772—781.

[113] Duník, J.-Straka, O.-Šimandl, M.: On autocovariance least-squares method for noise covariance matrices estimation, IEEE Transactions on Automatic Control, vol. 62, no. 2, 2017, 967—972.

[114] Nurminen, H.-Ardeshiri, T.-Piché, R.-Gustafsson, F.: Robust inference for state-space models with skewed measurement noise, IEEE Signal Processing Letters, vol. 22, no. 11, 2015, 1898–1902.

[115] Couvreur, C.: The EM algorithm: A guided tour, in Proceedings of the Second IEEE European Workshop on Computer-Intensive Methods in Control and Signal Processing (CMP'96), 1996, pp. 115-–120.

[116] Lainiotis, D.G.: Optimal adaptive estimation: Structure and parameters adaptation, IEEE Transactions on Automatic Control, vol. 16, no. 2, 1971, 160–170.

[117] Duník, J.-Soták, M.-Veselý, M.-Straka, O.-Hawkinson, W. J.: Design of Rao-Blackwellised point-mass filter with application in terrain-aided navigation, IEEE Transactions on Aerospace and Electronic Systems, vol. 55, no. 1, 2019, 251—272.

[118] Rogers, R.M.: Applied Mathematics in Inegrated Navigation Systems (2nd Edition). AIAA, 2003.

[119] Jekeli, C.: Inertial Navigation Systems with Geodetic Applications. Walter de Gruyter, Berlin, 2001.

[120] Straka, O.-Duník, J.-Šimandl, M.: Design of discrete second order filters for continuous-discrete models, in Proceedings of the 18th International Conference on Information Fusion, Washington, DC, USA, 2015.

[121] Maliňák, P.-Soták, M.-Kaňa, Z.-Baránek, R.-Duník, J.: Pure-inertial AHRS with adaptive elimination of non-gravitational vehicle acceleration. in Proceedings of the IEEE/ION Position Location and Navigation Symposium, Monterey, CA, USA, 2018.

[122] Duník, J.-Orejas, M.-Kaňa, Z.: Selected aspects of advanced receiver autonomous integrity monitoring application to Kalman filter based navigation filter. US Patent (US9547086 B2), 2017.

[123] Kaňa, Z.-Duník, J.-Soták, M.: Heading for a hybrid navigation solution based on magnetically calibrated measurements. European, US, Russian Patent (EP3043148 B1, US9939532 B2, RU2673504 C2), 2017.

[124] Crassidis, J. L.: Sigma-point Kalman filtering for integrated GPS and inertial navigation. IEEE Transactions on Aerospace and Electronic Systems, vol. 42, no. 2, 2006, 750–756.

[125] Musso, C.-Sacleux, B.-Bresson, A.-Allard, J.M.-Dahia, K.-Bidel, Y.-Zahzam, N.-Palmier, C.: Terrain-aided navigation with an atomic gravimeter. in Proceedings of the 22nd International Conference on Information Fusion, Ottawa, Canada, 2019.

[126] Orejas, M.-Duník, J.: Hybrid DFMC GNSS/INS to support approach iperations. in Proceedings of the 2014 International Technical Meeting of The Institute of Navigation, San Diego, California, USA, 2014.