ČESKÉ VYSOKÉ
UČENÍ TECHNICKÉ
V PRAZE

FAKULTA
STROJNÍ

DISERTAČNÍ PRÁCE

Diagnostika poruch neurčitých systémů pomocí markovských řetězců a EMD

DOKTORSKÝ STUDIJNÍ PROGRAM
STROJNÍ INŽENÝRSTVÍ
OBOR TECHNICKÁ KYBERNETIKA

ŠKOLITEL

Prof. Ing. Milan Hofreiter, CSc.

2017

Ing. Pavel Trnka

Prohlášení

Prohlašuji, že jsem disertační práci vypracoval samostatně a že jsem použil jen uvedených zdrojů.

V Praze dne 29.9.2017

……

Ing. Pavel Trnka

Anotace

Diagnostika poruch neurčitých systémů pomocí markovských řetězců a EMD

Diagnostika poruch má trvalý význam ve všech oblastech lidského konání. Vývoj technologií v dnešní době postupuje ke stále složitějším systémům řízení, které se vyznačují rostoucí mírou nezávislosti na lidské obsluze. Proto nabývá na významu podpora spolehlivosti a bezpečnosti provozu autonomních systémů. Diagnostika poruch založená na markovském modelu sledovaného procesu se obecně řadí mezi pravděpodobnostní modelovací techniky využívající Bayesovského pravděpodobnostního přístupu. Díky svým vlastnostem se markovský model jeví jako silný nástroj použitelný pro širokou škálu průmyslových i jiných aplikací.

Tato disertační práce předkládá několik postupů, které vedou ke zvýšení rychlosti, úspěšnosti a obecně ke zlepšení detekce a identifikace poruch diagnostickým systémem založeným na markovském modelu.

Annotation

Fault diagnostics of uncertain systems using Markov chains and EMD

Fault Diagnosis has lasting importance in all fields of human endeavor. Today's technology development is progressing towards more and more complex control systems, which are characterized by a growing degree of independence from human service. It is therefore important to promote the reliability and safety of operation of autonomous systems. Fault diagnosis based on the Markov model of the observed process is generally one of probability modeling techniques that are using Bayesian probability approach. Thanks to its features, the Markov model appears to be a powerful tool that can be used for a wide range of industrial and other applications.

This dissertation presents several procedures that increase the speed, success rate, and that generally improve the detection and identification of faults by a Markov model based diagnostic system.

Obsah

Seznam použité symboliky

Seznam symbolů

k	diskrétní čas
t	(spojitý) čas
a_k	veličina a v diskrétním čase k
a_i^j	časová posloupnost veličiny a v intervalu od diskrétního času i do diskrétního času $j \geq i$, tedy $a_i^j = a_i, a_{i+1}, \ldots, a_j$.
a^k	zkrácený zápis časové posloupnosti pro $i = 1$, tedy $a^k = a_1^k$
φ_a	obor (množina) hodnot veličiny a
$\varphi_a \times \varphi_b$	kartézský součin množin φ_a a φ_b
$\{a_1, a_2, \ldots, a_n\}$	množina tvořená výčtem prvků
$\{a : V(a)\}$	množina všech prvků a, které mají vlastnost $V(a)$
$f : \varphi_a \to \varphi_b$	zobrazení z φ_a na φ_b; funkce f definovaná na množině φ_a nabývá hodnot z množiny φ_b tak, že každému $a \in \varphi_a$ je přiřazen právě jeden prvek $b \in \varphi_b$
N	množina všech přirozených čísel; $N = \{1,2,3,\ldots\}$
R	množina všech reálných čísel
R^ρ	ρ rozměrný euklidovský prostor; $R^\rho = \underset{\rho}{\times} R$ (ρ-násobný kartézský součin)
$p(.\|.)$	podmíněná hustota pravděpodobnosti, kde tečky vyhrazují místa pro vložení konkrétních argumentů; argument zapsaný na místě první tečky představuje náhodný jev s daným rozložením pravděpodobnosti, druhý argument představuje podmínku tohoto rozložení, tedy veličiny se známou hodnotou.
$p(.,.)$	sdružená hustota pravděpodobnosti dvou náhodných veličin
$\equiv$	je definován jako
$\propto$	je proporcionální k; značí rovnost až na normalizující faktor
$\Gamma(a)$	Eulerova gama funkce; $\Gamma(a) = \int_0^\infty e^{-t} \cdot t^{a+1} \cdot dt$
$\delta(a,b)$	Kroneckerův delta operátor; $\delta(a,b) = 1$ pro $a = b$; $\delta(a,b) = 0$ pro $a \neq b$

$\mathbf{A}_{\langle a,b \rangle}$	matice $\mathbf{A}$, která má rozměry a řádků a b sloupců; v tomto zápisu je vektory možno formálně chápat jako jednorozměrné matice, tedy např. sloupcový vektor $v_{\langle a,1 \rangle}$
$\mathbf{A}^T$	transponovaná matice
$\mathbf{A}^{-1}$	inverzní matice k matici $\mathbf{A}$

Seznam zkratek

FDI	detekce a identifikace poruch (Fault Detection and Identification)
RV	regresní vektor
FT	Fourierova transformace (Fourier Transform)
DFT	diskrétní Fourierova transformace (Discrete FT)
FFT	rychlá Fourierova transformace (Fast FT)
HT	Hilbertova transformace (Hilbert Transform)
HHT	Hilbert-Huangova transformace (Hilbert-Huang Transform)
EMD	empirická modální dekompozice (Empirical Mode Decomposition)
IMF	vlastní modální funkce (Intrinsic Modal Function)

1 Úvod

Moderní trendy v průmyslu směřují ke stále větší automatizaci a autonomii technologických procesů s cílem omezit potenciálně nebezpečný lidský faktor a přesunout co největší část činností člověka z role výkonné do role rozhodovací. Určitou roli zde hrají i ekonomická hlediska. Autonomní proces vyžaduje na klíčových místech vysoce kvalifikované odborníky, těch je však potřeba relativně málo, například pro obsluhu a pro údržbu automatických řídicích systémů. Pro ostatní činnosti související s provozem takových systémů pak obslužný personál vystačí pouze se základním proškolením. Automatizovaný proces také zajišťuje rovnoměrné výkony a snadněji umožní trvalý provoz v režimu 7/24, což je důležité zejména v oblasti výroby.

Technologický proces s větší autonomií ale vyžaduje účinný nástroj, který dokáže co nejrychleji rozpoznat odchylky od běžného požadovaného chování, zjistit příčinu a zajistit včasnou nápravu. Včasné rozpoznání, určení a případné odstranění poruchy pomůže vyhnout se úplnému selhání systému, což při dnešní provázanosti systémových zařízení může znamenat předcházení nenapravitelných škod, úrazů případně ztrát na životech. Vyhnout se chybám musí nejen technologické procesy, které jsou tradičně vnímány jako nebezpečné (např. jaderné reaktory, chemická zařízení nebo letové provozy), ale zvyšují se i nároky na bezpečnost systémů používaných ve výrobních provozech nebo systémů, se kterými se setkáváme v běžném životě, např. v automobilech, vlacích, „chytrých“ domech apod.

Počátky moderní diagnostiky poruch se datují od první poloviny 70. let 20. století. Objem i intenzita výzkumu v této problematice i dnes neustále narůstá jak v teorii, tak v praxi. Se vznikem moderní teorie řízení v 90. letech a s obrovským rozmachem výpočetních technologií začíná používání technik matematického modelování, stavové analýzy a identifikace parametrů. To otevřelo cestu k predikci chování i složitých nelineárních systémů a výrazně se zvýšila efektivita diagnostických metod.

V úloze diagnostiky poruch je snahou odhalit co nejrychleji vzniklou odchylku od požadované funkce technologického procesu a zároveň zajistit co nejmenší počet chybných rozhodnutí, neboť po rozpoznání poruchy obvykle následují opatření, jejichž cílem je minimalizovat nepříznivé následky poruchy. Proto je důležité objevenou poruchu také správně lokalizovat a popsat. Monitorovací systém, který slouží k rozpoznávání poruch sledovaného zařízení, k jejich lokalizaci a ke stanovení jejich závažnosti, se obvykle označuje „systém diagnostiky poruch“ (Fault Diagnosis System). Problém diagnostiky poruch se skládá z úkolů detekce, lokalizace (izolace) a identifikace. Proto se často setkáváme také s pojmem systém FDI (Fault Detection and Isolation). *Detekce* poruchy spočívá v prostém rozhodnutí, jestli se proces nachází v poruchovém stavu. Poruchový stav přitom neznamená nutně úplný kolaps technologického procesu, ale obecně jakékoliv dostatečně významné odchýlení od požadované funkce. *Lokalizace* poruchy představuje úkol co nejpřesněji určit příčinu poruchy. Příkladem může být vadný senzor, strukturální změna konstrukce sledovaného zařízení atd. *Identifikace* poruchy má za cíl pokusit se blíže specifikovat rozsah a druh vzniklé poruchy.

Markovské řetězce jsou speciální stochastické modely, které jsou charakterizovány tzv. markovskou vlastností (nebo markovostí). Zjednodušeně řečeno markovost znamená, že pravděpodobnost aktuálního stavu procesu není určena celou jeho předešlou historií, ale pouze bezprostředně předcházejícím stavem. Řízené markovské

řetězce mají rozdělení pravděpodobnosti aktuálního stavu určené bezprostředně předcházejícím stavem a aktuální hodnotou vstupu. Hlavním kladem modelování s markovskými řetězci je možnost pracovat se silně nelineárními procesy a poměrně snadná identifikace. Nevýhodou, zejména při práci v reálném čase, je velký objem zpracovávaných dat a velké rozměry zpracovávaných matic. Mírným omezením může být také primárně diskrétní charakter markovského modelu. Diagnostika poruch s markovskými řetězci spočívá v použití modifikovaného markovského modelu, jehož regresní vektor je sestaven ze vstupních a výstupních veličin sledovaného procesu a výstupem modelu je veličina, která klasifikuje poruchový stav procesu. Takto sestavený model lze chápat jako bayesovský klasifikátor stavů technologického procesu. Markovský systém FDI může například vracet pro bezporuchový stav procesu hodnotu 0 a pro jednotlivé známé druhy poruch hodnoty 1, 2, atd.

Pro identifikaci markovského řetězce je potřeba mít dostatečně velké množství dat ze všech předpokládaných poruchových, ale i bezporuchových stavů procesu. Tato data slouží k vytvoření statistik závislostí mezi aktuální kombinací hodnot sledovaných veličin a mezi poruchovým stavem procesu. Klíčový význam pro dobré fungování FDI má vhodná volba struktury regresního vektoru, která do značné míry reprezentuje apriorní znalost o dynamických vlastnostech sledovaného technologického procesu.

Hilbert-Huangova transformace (HHT) byla představena v roce 1998 a brzy si získala značnou oblibu v nejrůznějších odvětvích (námořnictví, meteorologie, zpracování zvuku, letecký průmysl, ekonomie atd.). [48] Patří do kategorie časově-frekvenčních metod analýzy signálu. HHT spočívá v rozložení zkoumaného signálu na složky pomocí empirické modální dekompozice (Empirical Mode Decomposition – EMD) a v následné aplikaci Hilbertovy transformace pro nalezení okamžitých frekvencí a okamžitých amplitud kmitání signálu.

Hilbertova transformace umožní nalézt okamžitou frekvenci (obecně) kmitavého signálu pro každý bod časového průběhu. Aby byly výsledky HT rozumně použitelné, je třeba, aby měl signál co nejužší frekvenční spektrum a střední hodnotu blízkou nule. Úpravu signálu do vhodného tvaru zajistí EMD, která ze signálu extrahuje složky nazvané vlastní modální funkce (Intrinsic Modal Functions - IMF). Ty reprezentují mody kmitání signálu a velmi dobře splňují uvedené požadavky. Jednotlivé IMF se blíží ke vzájemné ortogonalitě. Pro potřeby FDI s markovskými řetězci je také důležité, že IMF mohou být dobrým zdrojem dat pro návrh regresního vektoru.

Vlastní rozklad signálu probíhá v iteračním algoritmu prosévání (Sifting), kde se postupnými úpravami odstraňují rušivé trendy, dokud průběžné reziduum signálu nevyhovuje definici vlastní modální funkce (tj. funkce s obálkami lokálních extrémů symetrickými kolem nulové střední hodnoty, počet lokálních extrémů lišící se maximálně o 1 od počtu průchodů nulou). Algoritmus prosévání se opakuje tak dlouho, dokud je možné ze signálu získávat další IMF. Zbytkové reziduum, ze kterého již nelze žádnou další IMF získat, reprezentuje celkový trend signálu v rámci zkoumané oblasti dat.

2 Stav problematiky

V této kapitole uvedu základní informace a terminologii související s diagnostikou poruch a stručně nastíním historický vývoj, který se odehrál v této oblasti. Představím potřebný myšlenkový a matematický aparát, který mi posloužil jako východisko pro konkretizaci cílů disertační práce a jejich naplnění. V části 2.1 se věnuji diagnostice poruch z obecného hlediska. V části 2.2 se zabývám stochastickým modelováním dynamických systémů se zaměřením na třídu markovských stochastických modelů a s využitím bayesovského přístupu. V části 2.3 se zabývám vybranými metodami časově frekvenčního rozkladu signálu a jejich využitím pro diagnostiku poruch.

2.1 Diagnostika poruch

Diagnostiku poruch lze chápat jako speciální případ detekce a analýzy změn sledovaného objektu, přičemž určité způsoby jeho chování jsou nežádoucí – z hlediska funkce objektu se jedná o poruchy.

2.1.1 Základní pojmy

„Porucha“ (Fault) je nečekaná a nežádoucí změna funkce technologického procesu, která ale nemusí nutně znamenat fyzické poškození nebo dokonce havárii. V takovém případě se obvykle hovoří spíše o „selhání“ (Failure). V případě poruchy může proces za určitých podmínek dále fungovat, v jeho chování se však vyskytují neočekávané změny, které je třeba v co nejkratším čase odstranit, aby se předešlo nežádoucím následkům.

V úloze detekce poruch je snahou odhalit vzniklou chybu co nejrychleji a zároveň zajistit co nejmenší počet chybných rozhodnutí, neboť po rozpoznání chyby obvykle následují opatření, jejichž cílem je minimalizovat následky poruchy. Proto je důležité objevenou poruchu správně lokalizovat a popsat.

Monitorovací systém, který slouží k rozpoznávání poruch sledovaného procesu, k jejich lokalizaci a ke stanovení jejich závažnosti, se označuje „systém diagnostiky poruch“ (Fault Diagnosis System). [1], [7], [12]

2.1.2 Funkce systému diagnostiky poruch

Systém diagnostiky poruch obvykle plní následující základní funkce :

- **Detekce poruchy** (Fault Detection) – činí „černobílé“ rozhodnutí, jestli se ve funkci sledovaného technologického procesu vyskytla porucha, nebo jestli je všechno v pořádku.
- **Lokalizace (izolace) poruchy** (Fault Isolation) – snaží se co nejpřesněji lokalizovat, kde porucha vzniká (např. který senzor je poškozen, který ovládací prvek byl chybně nastaven apod.).
- **Identifikace poruchy** (Fault Identification) – odhaduje rozsah a druh poruchy.

Funkce detekce a lokalizace obvykle považujeme za klíčové, zatímco funkce identifikace nemusí být kritická. Diagnostika poruch se proto obvykle označuje zkratkou FDI (Fault Detection and Isolation) [15], [21], [29], [31], [32], [45].

Rozdělení dílčích úloh v současných diagnostických systémech již není tak striktní a jednotlivé funkce je nutno chápat spíše jako různé úhly pohledu na činnost systému diagnostiky poruch jako celku. Proto se v posledních letech prosazuje také termín Fault Detection and Diagnosis (FDD) [7], [17], [28] nebo prostě Fault Diagnosis [1], [16], [18], [19], [23], [24], [25], [26], [27], [30], [33].

2.1.3 Obecný průběh FDI

Průběh diagnostiky poruch (FDI) se dá obecně rozdělit do dvou základních fází, které se během činnosti sledovaného technologického procesu neustále střídají.

- **Tvorba rezidua** (Residual Generation) – V této fázi se pomocí generátoru reziduí vytváří signál (reziduum), který informuje o přítomnosti poruchy. Reziduum se tvoří na základě aktuálních vstupních a výstupních veličin technologického procesu a je obvykle nulové v bezporuchovém stavu a výrazně nenulové v případě poruchy.
- **Proces rozhodování** (Decision Making) – V této fázi se následně na základě vygenerovaných reziduí rozhoduje, zda (a případně jaká) porucha nastala.

2.1.4 Rozdělení poruch

Poruchy lze rozdělovat podle různých kritérií. Pro technologický proces (systém) popsaný analytickým modelem je možné použít rozdělení podle *způsobu ovlivnění matematického modelu*. Další možností je dělení podle *zdrojů poruchy*. (Podrobněji [7].)

Aditivní poruchy

Jsou to šumy, poruchy akčních členů, poruchy senzorů či neznámé poruchy. V rovnici systému se projevují jako aditivní členy a jsou nezávislé na velikosti měřených veličin.

V Laplaceově obrazu funkce popisující systém mohou být aditivní poruchy zapsány například takto:

$$H(s)\cdot y(s)=G(s)\cdot u(s)+H(s)\cdot d_a(s)+F(s)\cdot d_s(s)+K(s)\cdot \xi(s), \qquad (2.1.1)$$

kde $H(s)$, $G(s)$, $F(s)$ a $K(s)$ jsou matice se známými koeficienty, $y(s)$ je vektor výstupních veličin, $u(s)$ vektor vstupních veličin, $d_a(s)$ jsou poruchy akčních členů, $d_s(s)$ poruchy senzorů a $\xi(s)$ jsou neznámé poruchy a šum. Symbol s reprezentuje argument Laplaceovy transformace.

Multiplikativní poruchy

Jedná se o změny struktury rovnic popisujících systém, které se dají vyjádřit změnou parametrů. Tyto změny závisí na velikosti měřených veličin, jak je vidět v následujícím vztahu:

$$(H(s)+\Delta H(s))\cdot y(s)=(G(s)+\Delta G(s))\cdot u(s), \qquad (2.1.2)$$

kde $H(s)$ a $G(s)$ jsou matice se známými koeficienty, $y(s)$ je vektor výstupních veličin, $u(s)$ vektor vstupních veličin, $\Delta H(s)$ a $\Delta G(s)$ jsou matice s neznámými koeficienty reprezentující poruchu.

Rozdělení podle zdrojů poruch

Podle zdrojů se poruchy obvykle dělí do následujících čtyř skupin:

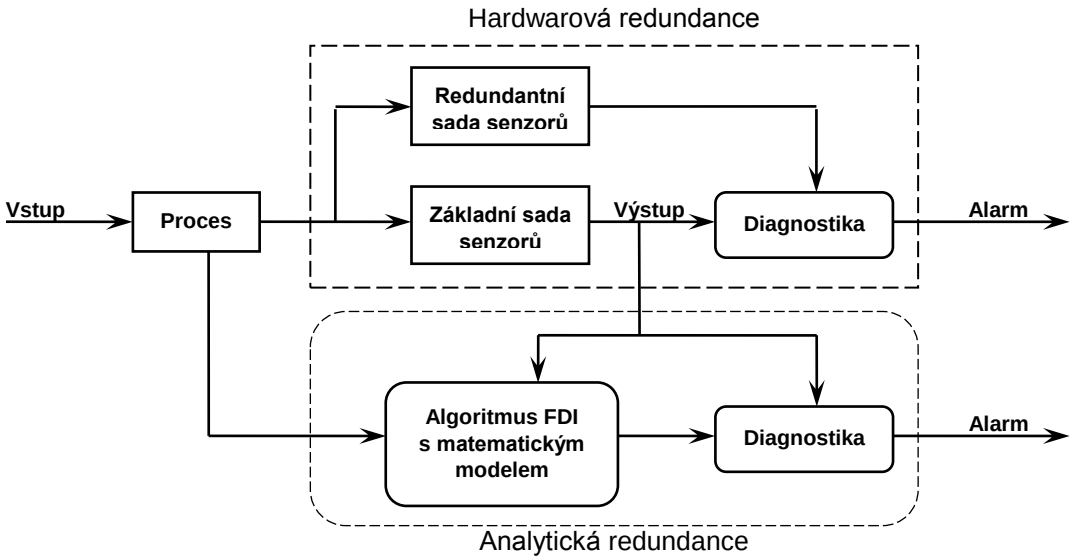
- **Změna parametrů rovnice** – jedná se o změnu některého z parametrů v rovnici popisující systém, například změna charakteristiky odporového vinutí.
- **Změna struktury rovnice** – odstranění nebo přidání členu rovnice nebo vznik nové vazby mezi rovnicemi, např. porušení těsnosti tlakového potrubí.

- **Poruchy akčních členů či senzorů**.
- **Šumy** – mohou vzniknout uvnitř systému nebo na senzorech.
- **Jiné náhodné poruchy** – jedná se o poruchy, které nelze popsat ani přesně definovat.

2.1.5 Základní přístupy k diagnostice poruch

Nejběžnějším diagnostickým postupem je sledování úrovní nebo trendů vybraných signálů a přiřazení činností ke zvoleným rozsahům. Při této metodě mohou vzniknout falešné poplachy vyvolané zašuměním, kolísáním signálu či změnou pracovního bodu. Kromě toho jedna porucha ovlivní velké množství systémových signálů, takže je obtížné ji lokalizovat.

Tyto problémy je možné výrazně omezit s pomocí testování shody v množině systémových signálů. K tomu je třeba vytvořit matematický model funkčních závislostí mezi jednotlivými signály získanými z procesu (včetně vstupů či výstupů).



obr. 2.1.1 – Srovnání hardwarové a analytické redundance. (převzato z [7])

Hardwarová (fyzická) redundance

Diagnostické metody založené na hardwarové (fyzické) nadbytečnosti (Hardware Redundancy) využívají vícenásobných senzorů, akčních členů, počítačů (a softwarů) k měření dílčích proměnných. Na základě rozhodovacího schématu použitého na redundantní systém se určuje, zda a kdy došlo k poruše a hledají se nejpravděpodobnější zdroje poruchy.

HW redundance se využívá například v dopravních prostředcích (vlaky, letadla) nebo v jaderných reaktorech. Výhodou tohoto řešení je rychlost a jednoduchost rozpoznání poruchy. Nevýhodou je vyšší cena a složitost systému, redundantní prvky také zabírají více místa.

Analytická (funkční) redundance

Princip analytické (funkční) nadbytečnosti (Analytical Redundancy) využívá redundantních funkčních vztahů mezi veličinami měřenými na sledovaném technologickém procesu (např. vstupy/výstupy, vstupy/vstupy, výstupy/výstupy,...).

Místo zdvojování každého prvku zvlášť se vzájemně porovnávají (Cross Checking) hodnoty měřené na různých senzorech v rámci celého systému.

obr. 2.1.1 ukazuje srovnání schémat hardwarové a analytické redundance.

V případě analytické redundance se nepoužívá žádný nadbytečný hardware, který by zaváděl dodatečné riziko hardwarových poruch. Analyticky redundantní systém je tedy potenciálně spolehlivější, podmínkou je však kvalitní model schopný obsáhnout podstatné děje uvnitř sledovaného technologického procesu.

V praxi se často aplikuje kombinace obou metod tak, aby se v maximální míře využily jejich výhody a minimalizovaly nevýhody. Například kriticky důležitý senzor se použije dvoj- či trojnásobný a pro podchycení komplexnějších poruch se aplikuje matematický model FDI.

2.1.6 Historie modelově orientované diagnostiky poruch

Beard-Jonesův detekční filtr

V roce 1971 přinesl Beard na Massachusetts Institute of Technology (MIT) myšlenku nahrazení hardwarové redundance analytickou redundancí. Podle ní byly vytvořeny filtry pro diagnostiku poruch generující orientovaná rezidua pro FDI. V roce 1973 přeformuloval Jones tento postup na geometrickou interpretaci. Tato myšlenková linie posléze vyústila do podoby nazvané Beard-Jonesův filtr diagnostiky poruch nebo jen Beardův filtr diagnostiky poruch. Tento postup byl později ještě upravován a vylepšován například Chungem a Speyerem v roce 1998 [8].

Stochastické systémy FDI

Souběžně s vývojem Beard-Jonesova filtru se od počátku 70. let rozvíjely také statistické postupy. Mehra a Peschon představili v roce 1971 obecný postup FDI využívající přírůstků (nebo reziduí) generovaných Kálmánovým filtrem. Poruchy jsou rozpoznávány statistickým testováním na „bělost", střední hodnotu a kovarianci reziduí.

Willsky a Jones poté vyvinuli strategii FDI využívající zobecněnou míru pravděpodobnosti (GLR – Generalized Likelihood Ratio) k testování na reziduích generovaných Kálmánovým filtrem pro rozpoznávání poruch. V dobře známém sborníku [40] prezentoval Willsky (1976) klíčové koncepty analytické redundance v modelově orientovaných FDI se zvláštním zaměřením na stochastické systémy a detekci skoků. Basseville pokračoval v této myšlenkové linii a zaměřil se na problémy detekce, odhadu a diagnostiky změn dynamických vlastností signálů a systémů se zaměřením na statistické metody detekce, aby poskytl obecný rámec pro detekci změn signálů a systémů.

Vývoj statistického odhadu později shrnuli ve vynikající práci Tzafestas a Watanabe v roce 1990 [28]. Byly zde diskutovány také další metody odhadu jako třeba znalostní techniky. Nejvýznamnějším přínosem uvedené práce bylo velmi kvalitní zmapování známých stochastických technik. Základní metody statistického odhadování je možno nalézt v [14].

Metoda vícemodelového adaptivního filtru zavádí vícenásobné testování hypotéz na rezidua generovaná skupinou Kálmánových filtrů.

Od roku 1996 je kladen stále větší důraz na statistické metody založené na použití analýzy hlavních prvků (viz např. Zhang, Martin a Morris [13]).

FDI odhad s pozorovatelem

Clark s kolektivem aplikovali jako první Luenbergerovy pozorovatele v oblasti detekce poruch a následně vyvinuli různé postupy pro lokalizaci poruch senzorů.

Pozici metod využívajících pozorovatele upevnila Frankova práce 1987 [33]. V tomto sborníku je uvedeno mnoho různých postupů využívajících lineární i nelineární pozorovatele a také několik příkladů použití.

Metoda paritních vztahů pro FDI

Metodu paritních vztahů pro generování reziduí (paritní vektor), založenou na kontrole konzistence mezi vstupními a výstupními daty systému přes časové okno, původně navrhl Mironovski (1979), přestože tehdy použil jinou terminologii. Jeho práce bohužel nebyla široce publikována a tak se jí nedostalo zasloužené pozornosti. Postup později nezávisle na Mironovskim formulovali Chow a Willsky v roce 1984 [35].

Následně použili postup v mnoha rozličných variantách i další autoři. Například Gertler přenesl metodu návrhu paritních vztahů do Z-oblasti. Chen a Zhang vyvinuli v roce 1990 metodu detekce a identifikace chyb stochastických systémů založenou na myšlence přímého rozvoje paritních vektorů používané v hardwarové redundanci.

Pozdější výzkumy založené na metodě paritních vztahů popisuje například Gertler (1998) [7].

Metoda odhadu parametrů pro FDI

Jednou z důležitých metod diagnostiky poruch je použití odhadu parametrů, které je založeno přímo na technikách identifikace systémů. Tento postup poprvé nastínili v roce 1979 Bakiotis, Raymond a Rault [39] a v roce 1982 Geiger.

Isermann a kol. následovali tento směr výzkumu od počátku 80. let. V roce 1984 Isermann ukázal, že porucha systému může být rozpoznána za použití odhadu neměřitelných parametrů procesu a/nebo jeho stavových proměnných [36]. Ve své práci definoval zobecněnou strukturu diagnostiky poruch založené na modelu procesu a na neměřitelných veličinách.

Na tuto strukturu se později ve svých pracích odkazovalo mnoho autorů, například Frank (1990) [26]. V roce 1987 zveřejnil Isermann svoje zkušenosti s použitím odhadu parametrů pro účely diagnostiky poruch procesů [34]. V roce 1990 zveřejnili Isermann a Freyermuth studii o on-line systémech diagnostiky poruch využívajících kombinace odhadu parametrů a heuristické analýzy procesů [27]. V roce následovala 1991 další práce, která přinesla přehled metod odhadu parametrů vycházející z mnoha reálných i laboratorních aplikací zaměřený na praktické použití.

Další zajímavé výsledky ukázali Isermann a Ballé ve své práci z roku 1997 [11].

Dvoustavová modelově orientovaná struktura FDI

Chow a Willsky definovali v letech 1980 až 1984 modelově orientovanou diagnostiku poruch jako proces pouze se dvěma stavy:

1) tvorba reziduí,
2) rozhodovací proces (včetně ohodnocení reziduí).

Tento dvoustavový proces se dodnes uznává jako standardní procedura pro modelově orientované rozpoznávání poruch [35].

Problém robustnosti v diagnostice poruch

V roce 1981 upozornil Leininger na dopad špatného modelu na kvalitu procesu rozpoznávání chyb [38]. První pokusy o zlepšení robustnosti diagnostických metod s pozorovatelem provedli Frank a Keller (1981) [37].

Robustní FDI s pozorovatelem neznámých vstupů

Watanabe a Himmelblau představili v roce 1982 metodu detekce robustních senzorů využívající pozorovatele neznámých vstupů (UIO – Unknown Input Observer), která řešila problém robustnosti FDI. Robustními diagnostikami poruch založenými na UIO se na Duisburské universitě v Německu široce zabýval Frankův tým, který také na toto téma bohatě publikoval (viz např. [26]). Metody FDI využívající UIO podrobně zpracovali také Chen a Patton 1999 [1].

Chen a Zhang navrhli schéma robustní lokalizace poruch akčních členů a demonstrovali jeho použití v chemickém provozu [22] (1991). Ge a Fang vyvinuli metodu robustních komponent, která využívala tzv. metodu robustního pozorování a byla principiálně podobná metodám UIO (1989) [29].

Významným příspěvkem je také metoda lokalizace poruch akčních členů, kterou navrhl v roce 1987 Viswanadham se svými spolupracovníky, problém robustnosti v něm však bohužel nebyl uvažován [32].

Robustní FDI s pomocí přiřazení vlastní struktury

Metodu využívající přiřazování vlastní struktury navrhli Patton, Willcox a Winter. Skupina profesora Pattona se jí v širokém rozsahu zabývala a přinesla mnoho výsledků ([12], [1]).

Optimální paritní vztahy pro robustní FDI

V roce 1986 vyvinuli Lou, Willsky a Verghese strategii návrhu „optimálně robustních paritních vztahů“ k rozpoznávání poruch v systémech reprezentovaných více modely.

Vedeni stejnou myšlenkou řešili Wünnenberg a Frank návrh optimálních paritních vztahů adaptací modifikovaného kriteria, což je poměr vlivu neurčitosti modelu a vlivu poruchy [26].

Gertler s kolegy navrhli na podobném principu postup návrhu robustních paritních vztahů využívající koncept „ortogonálních paritních vztahů“ [19], [21].

Návrh ve frekvenční oblasti pro modelově orientované FDI

Viswanadham, Tailor a Luce představili v roce 1987 novou metodu tvorby reziduí založenou na faktorizaci přechodové matice systému. Tento postup později rozvinuli Ding a Frank (1990) a poté přepracovali Kinnaert a Peng (1995). V této podobě je dnes postup známý jako Metoda tvorby reziduí ve frekvenční oblasti [15].

Viswanadham a Minto se věnovali problému robustnosti této metody a v roce 1988 navrhli řešení, jak zvýšit robustnost tvorby reziduí ve frekvenční oblasti za použití H_∞ optimalizačních technik. Ding a Frank přispěli k tomuto tématu sérií článků v letech 1991 až 1994 (viz např. [16], [20], [26]). K řešení problému robustnosti FDI ve frekvenční oblasti byla použita také technika robustního návrhu zvaná „μ syntéza“.

I v posledních letech je o metody návrhu FDI ve frekvenční oblasti stále velký zájem [1].

Robustní ohodnocení reziduí a adaptivní práh

V případech, kdy není možné zvýšit robustnost reziduí vůči neurčitosti systému, je možné docílit robustního rozpoznávání poruch nasazením robustního rozhodování s adaptivními prahy. Koncept prahového řadiče (Treshold Selector) určeného ke generování adaptivních prahů zavedli v roce 1988 Emami-Naeini, Akhter a Rock. V roce 1989 představil Clark funkční metodu pro generování adaptivních prahů.

Metodami stanovení adaptivních prahů se od poloviny 90. let intenzivně zabývali Ding a Frank [16], [6].

Myšlenku dosažení robustnosti ve fázi procesu rozhodování nadále rozvinul Frank a kol. zavedením fuzzy logiky při ohodnocování reziduí a v procesu rozhodování [9]. V roce 1998 ukázali, že optimalizací funkce ohodnocení reziduí v návrhu pozorovatelů je možno získat robustní diagnostiku poruch [6].

Modelování neurčitosti pro robustní FDI

Aby bylo možné vyřešit problém robustnosti FDI, je nutné vytvořit matematickou reprezentaci popisující neurčitost modelování. Patton a Chen navrhli několik postupů, jak reprezentovat neurčitost modelování z různých zdrojů jako aditivní poruchy s předpokládanou distribuční maticí. Robustnosti FDI je tak dosaženo postupy „rozdělení" poruch. Toto je dodnes jeden z nejvýznamnějších přínosů v robustní diagnostice poruch. Nejrobustnější metody tvorby reziduí jsou založeny na předpokladu, že matice rozdělení poruch jsou známé, nicméně tento předpoklad není splněn pro většinu reálných systémů.

Pattonovy a Chenovy články otevřely cestu praktickému využití robustních technik rozpoznávání poruch (viz např. [12]).

Rychlý rozvoj modelově orientovaných FDI

Největšího rozmachu dosáhly modelově orientované FDI v období mezi koncem 80. a začátkem 90. let minulého století. Tehdy byly formulovány základní definice a byla navržena vlastní struktura modelově orientovaného systému FDI.

Roku 1988 popsal Gertler základní myšlenky a klíčové formulace [31]. Některá témata, která řešil ve své práci, jako jsou například podmínky lokalizovatelnosti, citlivost či robustnost, neztratila dodnes svůj význam.

Sborník autorů Pattona, Franka a Clarka vydaný v roce 1989 zmapoval hlavní modelově orientované metody diagnostiky poruch 80. let včetně velkého množství příkladů použití [30]. Odborníci se na tuto knihu stále ještě často odkazují.

Povzbuzen úspěchem sborníku zveřejnil Frank v roce 1990 pojednání, ve kterém nastínil principy a nejdůležitější postupy modelově orientované tvorby reziduí s využitím identifikace parametrů a odhadu stavů [26]. Zvláštní důraz kladl na nejnovější pokusy se zvyšováním robustnosti vůči neurčitosti modelu a také se zamýšlel nad možností společného použití modelově orientovaných a znalostních technik v oblasti FDI.

Aby mohly různé metody modelově orientované FDI tvořit jednotný systém, bylo nutné popsat jejich vzájemné vztahy a souvislosti.

Jako první nastínili spojitost mezi postupy s paritními vztahy a postupy pracujícími s pozorovatelem roku 1991 Patton a Chen [24]. Shodnost přístupů založených

na pozorovateli a přístupů využívajících paritních vztahů potvrdili formálně Gertler a DiPierro v roce 1997 [10].

V úvodní přednášce na konferenci IFAC Symposium: SAFEPROCESS'91 v Basileji v roce 1991 představili Patton a Chen společný formát paritního prostoru (Parity Space Format), který v sobě sdružoval přístupy založené na pozorovateli s přístupy využívajícími paritních vztahů [24]. Modelově orientovanou diagnózu poruch přeformulovali na tvorbu a analýzu reziduálních signálů v paritním prostoru. V přednášce Patton a Chen poskytli obecný rámec generátorů reziduí a zavedli některé důležité definice a předvedli na dvou názorných příkladech robustní metody diagnostiky poruch. O rok později navázali prací, ve které se zabývali metodami syntézy pro generátory reziduí se zvláštním zaměřením na letectví.

Na konferenci SAFEPROCESS'91 prezentoval Frank svůj názor na stupňující se robustnost metod FDI založených na pozorovatelích s rozpojováním poruch, na optimálních paritních vztazích, na pozorovatelích H∞ a na použití adaptivního prahu [20].

Gertler zde prezentoval návod k metodám syntézy generátorů reziduí [21]. Přehledně a systematicky předvedl nejznámější metody generování reziduí jako např. paritní rovnice, diagnostické pozorovatele nebo Kálmánův filtr. V pojednání nechyběly ani příklady pro porovnání jednotlivých metod.

Isermann představil návod k metodám odhadu parametrů zaměřený na praktickou realizaci [23].

V roce 1993 proběhla ve francouzském Labarrère další významná konference věnovaná diagnostice poruch – TOOLDIAG'93. Frank zde rozebíral možnosti zlepšení robustnosti rozhodování s použitím fuzzy logiky. Gertler a Kunwert představili návrh generátoru rozpojených reziduí, a to úplný i s aproximovanými poruchami, se zaměřením na metody návrhu paritních vztahů v Z-doméně. Patton se zabýval robustností řídicích systémů odolných vůči chybám včetně problematiky diagnostiky a rekonfigurace. Zdůrazňoval, že propojením diagnostiky poruch a regulátoru v analýze i návrhu se podaří nejlépe optimalizovat stabilitu i výkon systému řízení odolného k chybám.

Vzhledem k velkému množství existujících metod FDI může být obtížné zvolit v dané situaci tu nejvhodnější. Patton, Chen a Nielsen vydali v roce 1994 směrnice pro inženýry zabývající se tímto problémem [18]. Isermann řešil použitelnost různých metod FDI podle jejich požadavků a výsledků při simulacích [17].

2.2 Markovské řetězce

Vlastnosti většiny reálných systémů nejsou stálé, ale v průběhu času se mírně mění či kolísají. Příčiny změn vlastností reálného systému mohou být buď systematické (např. postupné opotřebení strojních součástí, degradace vlastností senzorů apod.), nebo náhodné (působení poruch, šumů, kolísání parametrů okolního prostředí atd.). Tato práce se zaměřuje zejména na procesy, které vykazují druhý případ proměnlivosti svých vlastností. Díky stochastickému charakteru takového procesu je obtížné vytvořit pro něj dostatečně kvalitní deterministický model založený na matematicko-fyzikální analýze. V takovém případě je výhodnější použít pravděpodobnostní model.

Stochastické (pravděpodobnostní) modely jsou založeny na dlouhodobém sběru známých vstupů a výstupů technologického procesu a jejich pravděpodobnostní vyhodnocování (viz např.[3], [4], [5] a [1]).

Pro účely diagnostiky poruch se na přítomnost chyby usuzuje na základě srovnání modelu a reálného systému.

Důležité vztahy bayesovské statistiky

V dalším textu budou často používány dva následující vztahy z teorie pravděpodobnosti. Mějme tři náhodné veličiny a, b, c a symbol $p(\cdot|\cdot)$ reprezentující podmíněnou pravděpodobnost. Pak platí

pravidlo násobení

$$p(a,b|c) = p(a|b,c)\cdot p(b|c) \tag{2.2.1}$$

a **pravidlo marginalizace**

$$p(b|c) = \int p(a,b|c)da\,. \tag{2.2.2}$$

S pomocí uvedených vztahů je možné odvodit **Bayesův vzorec**

$$p(a|b,c) = \frac{p(b|a,c)\cdot p(a|c)}{p(b|c)} = \frac{p(b|a,c)\cdot p(a|c)}{\int p(b|a,c)\cdot p(a|c)da} \propto p(b|a,c)\cdot p(a|c)\,, \tag{2.2.3}$$

který je základem Bayesovské statistiky. Symbol proporcionality (úměrnosti) $\propto$ znamená rovnost až na konstantu úměrnosti.

Podle Bayesovského přístupu se vychází z předpokládaného modelu stochastického procesu, který je průběžně korigován porovnáním s reálnými daty. Tento způsob zajišťuje pružné přizpůsobování modelu pozorovanému chování reálného procesu a případně také jeho změnám, zároveň však umožňuje využít zkušenosti tvůrce modelu, který může dát modelu předpokládanou strukturu, nastavit výchozí hodnoty parametrů a zohlednit další apriorní znalosti o modelovaném procesu [2], [41], [42].

2.2.1 Pravděpodobnostní model

Pravděpodobnostní přístup, na rozdíl od deterministického, se nesnaží popsat reálný systém pouze pomocí přesně definovaných jednoznačných pravidel (například vzorce matematicko-fyzikální analýzy), ale přidává k popisu určitého jevu ještě další informaci vyjadřující jeho míru očekávání. To nám umožní kvalitativně vyhodnocovat důvěryhodnost dosažených výsledků.

Pravděpodobnostní (stochastický) popis procesu

Předpokládejme stochastický proces diskrétní v čase i hodnotách, na který působí časová posloupnost vstupů $\{v_k,\ \ k=1,2,\ldots\}$ a jehož odezvou je posloupnost výstupů $\{y_k,\ \ k=1,2,\ldots\}$ pozorovaných v diskrétním čase $k=1,2,\ldots$. Dále předpokládejme, že jedinou dostupnou informaci o aktuálním stavu procesu je možné získat jeho vnějším pozorováním. Zavedeme označení pro vstupy a výstupy procesu v diskrétním čase k

$$D_k = \{y_k, v_k\}\,. \tag{2.2.4}$$

Pro získání úplného popisu procesu v diskrétním čase pro časové období od počátečního diskrétního okamžiku $k = k_0 + 1$ do okamžiku $k = k_K$ bychom museli znát všechny hodnoty $D_{k_0+1}^{k_K} = \left\{ D_{k_0+1}, D_{k_0+2}, \ldots, D_{k_K-1}, D_{k_K} \right\}$ pro jakoukoliv možnou realizovatelnou strategii řízení. Dosáhnout takového požadavku by bylo velmi obtížné. Je však možno navrhnout model procesu, který přiřadí libovolné přípustné strategii hustotu pravděpodobnosti

$$p\left(D_{k_0+1}^{k_K} \middle| D^{k_0} \right), \tag{2.2.5}$$

tedy podmíněnou pravděpodobnost, že proběhne posloupnost $D_{k_0+1}^{k_K}$, jestliže známe předchozí průběh chování procesu D^{k_0} až do okamžiku $k = k_0$ včetně. Tato podmínka je snadno dosažitelná například zavedením předpokladu $k_0 = 0$, takže D^{k_0} je prázdná množina.
Výraz (2.2.5) převedeme opakovaným použitím pravidla násobení (pravidlo řetězení, [1], [61]) a dosazením z (2.2.4) na tvar

$$p\left(D_{k_0+1}^{k_K} \middle| D^{k_0} \right) = \prod_{k=k_0+1}^{k_K} p\left(y_k \middle| v_k, D^{k-1} \right) \cdot p\left(v_k \middle| D^{k-1} \right), \tag{2.2.6}$$

kde hustota pravděpodobnosti $p\left(v_k \middle| D^{k-1} \right)$ vyjadřuje obecně stochastický popis strategie řízení, kterou je vstup v_k určen na základě známé předchozího historie procesu. [1], [61]

Hustota pravděpodobnosti

$$p\left(y_k \middle| v_k, D^{k-1} \right);\ k = k_0 + 1, k_0 + 2, \ldots, k_K \tag{2.2.7}$$

udává závislost výstupu y_k na známém předešlém průběhu chování procesu a na aktuálním vstupu v_k.

Množina pravděpodobností (2.2.7) tvoří **úplný vnější pravděpodobnostní popis** řízeného procesu a představuje tak jeho obecný **stochastický model**. [61]

Identifikace stochastické soustavy

Úplný vnější popis procesu (2.2.7) získáme s využitím pravděpodobnostních modelů. Přitom lze dokázat (viz [61]), že není nutné zvolit jediný nejlepší model. Stačí stanovit množinu hypotéz o struktuře a parametrech stochastického modelu a nalézt rozdělení pravděpodobnosti mezi přípustnými hypotézami.

Matematický popis soustavy nazývaný stochastický model soustavy [44], [61] je charakterizován dvojicí

$$M \equiv \left\{ {}_i H, {}_i K \right\}, \tag{2.2.8}$$

kde ${}_i H$ je *i*-tá hypotéza ze souboru všech možných hypotéz o struktuře modelu a ${}_i K$ představuje hodnotu parametrů modelu s touto předpokládanou strukturou.

Označíme-li množinu všech předpokládaných možných modelů $\varphi_M, M \in \varphi_M$, pak podmíněnou pravděpodobnost výstupů nezávislou na konkrétním modelu (2.2.7) je možné za pomoci pravidel marginalizace a násobení a za předpokladu **přirozených podmínek řízení**[1] [41], [61] vyjádřit vztahem

$$p\left(y_k \middle| v_k, D^{k-1}\right) = \int_{\varphi_M} p\left(y_k \middle| v_k, D^{k-1}, M\right) \cdot p\left(M \middle| D^{k-1}\right) \cdot dM, \tag{2.2.9}$$

kde rozdělení pravděpodobnosti $p\left(y_k \middle| v_k, D^{k-1}, M\right)$ představuje (známý) odhad chování řízeného systému při použití určitého modelu M a výraz $p\left(M \middle| D^{k-1}\right)$ představuje prozatím neznámou podmíněnou pravděpodobnost, že tento model nejlépe vystihuje chování řízeného procesu, jestliže byla pozorována historie procesu D^{k-1} [61], [44].

Hustotu pravděpodobnosti $p\left(M \middle| D^{k-1}\right)$ je možné vyjádřit za pomoci pravidla násobení vztahem

$$p\left(M \middle| D^{k-1}\right) = \frac{p\left(D_{k_0+1}^{k-1} \middle| D^{k_0}, M\right) \cdot p\left(M \middle| D^{k_0}\right)}{p\left(D_{k_0+1}^{k-1} \middle| D^{k_0}\right)} = \frac{p\left(M \middle| D^{k_0}\right) \cdot \prod_{\kappa=k_0+1}^{k-1} p\left(D_\kappa \middle| D^{\kappa-1}, M\right)}{p\left(D_{k_0+1}^{k-1} \middle| D^{k_0}\right)}, \tag{2.2.10}$$

který dále upravíme dosazením z (2.2.4) a aplikací přirozených podmínek řízení na

$$p\left(M \middle| D^{k-1}\right) = p\left(M \middle| D^{k_0}\right) \cdot \prod_{\kappa=k_0+1}^{k-1} p\left(y_\kappa \middle| v_\kappa, D^{\kappa-1}, M\right) \cdot \frac{\prod_{\kappa=k_0+1}^{k-1} p\left(v_\kappa \middle| D^{\kappa-1}\right)}{p\left(D_{k_0+1}^{k-1} \middle| D^{k_0}\right)}. \tag{2.2.11}$$

Zlomek na pravé straně obsahuje pouze výrazy konstantní vzhledem k M a můžeme psát

$$p\left(M \middle| D^{k-1}\right) \propto p\left(M \middle| D^{k_0}\right) \cdot \prod_{\kappa=k_0+1}^{k-1} p\left(y_\kappa \middle| v_\kappa, D^{\kappa-1}, M\right), \tag{2.2.12}$$

kde symbol $\propto$ představuje proporcionalitu, tedy rovnost až na normalizující konstantu. Výrazy $p\left(y_k \middle| v_k, D^{k-1}, M\right)$ jsou známé hustoty pravděpodobnosti z (2.2.9) v diskrétních časech $\kappa = k_0+1, k_0+2, \ldots, k-1$ a $p\left(M \middle| D^{k_0}\right)$ je apriorně zvolené počáteční rozdělení pravděpodobnosti na množině předpokládaných možných modelů φ_M. Takto apriorně zvolená hustota pravděpodobnosti odráží *subjektivní* míru naší důvěry v jednotlivé modely a průběžný výpočet hustoty pravděpodobnosti (2.2.12) je možné interpretovat

[1] Přirozené podmínky řízení lze slovně vyjádřit tak, že řídicí systém nemá jinou informaci o řízeném procesu, než může získat jeho vnějším pozorováním a strategie řízení tedy nezávisí na zvoleném modelu, ale pouze na pozorované historii procesu, tedy $p\left(v_k \middle| D^{k-1}, M\right) = p\left(M \middle| D^{k-1}\right)$. Důsledkem je také skutečnost, že nelze získat ani informace o struktuře a parametrech modelu pouze pozorováním vstupů procesu a tedy $p\left(M \middle| v_k, D^{k-1}\right) = p\left(M \middle| D^{k-1}\right)$. Podrobněji viz [41], [61].

jako adaptivní přizpůsobování modelu projevům reálné soustavy na základě průběžného sledování *objektivních* dat z jejího provozu.

Obecně neznáme strukturu ani parametry modelu soustavy, můžeme pouze vytvořit konečnou množinu hypotéz o jeho struktuře $_iH \in {}_i\varphi,\ i=1,2,\ldots,r$ a k nim pak najít hodnoty parametrů $_iK$. Zde množiny ${}_i\varphi \subseteq \varphi_M$, $i=1,2,\ldots,r$ jsou podmnožinami množiny všech možných modelů φ_M a každá ${}_i\varphi$ definuje třídu všech modelů z φ_M, které vedou pro stejnou pozorovanou historii procesu $D_{k_0+1}^{k-1}$ na stejnou množinu hustot pravděpodobnosti $p\left(y_k \middle| v_k, D^{k-1}, {}_iK, {}_iH\right)$. Množiny ${}_i\varphi$ jsou navzájem disjunktní a budeme předpokládat

$$\varphi_M = \bigcup_{i=1}^{r} {}_i\varphi;\ 1 \le r < \infty. \tag{2.2.13}$$

Vhledem k (2.2.8) a za pomoci bayesovského přístupu můžeme využít vztah (2.2.12) k určení podmíněného rozdělení pravděpodobnosti hypotéz o struktuře modelu

$$p\left({}_iH \middle| D^{k-1}\right) \propto p\left({}_iH \middle| D^{k_0}\right) \cdot \int_{\varphi_{iK}} p\left({}_iK \middle| D^{k_0}, {}_iH\right) \cdot \prod_{\kappa=k_0+1}^{k-1} p\left(y_\kappa \middle| v_\kappa, D^{\kappa-1}, {}_iK, {}_iH\right) \cdot d\,{}_iK, \tag{2.2.14}$$

Označíme-li

$${}_iI\left(k-1, k_0\right) = \int_{\varphi_{iK}} p\left({}_iK \middle| D^{k_0}, {}_iH\right) \cdot \prod_{\kappa=k_0+1}^{k-1} p\left(y_\kappa \middle| v_\kappa, D^{\kappa-1}, {}_iK, {}_iH\right) \cdot d\,{}_iK, \tag{2.2.15}$$

a provedeme-li normalizaci výrazu (2.2.14), dostaneme podmíněnou hustotu pravděpodobnosti hypotézy o struktuře modelu

$$p\left({}_iH \middle| D^{k-1}\right) = \frac{p\left({}_iH \middle| D^{k_0}\right) \cdot {}_iI\left(k-1, k_0\right)}{\sum_{j=1}^{r} p\left({}_jH \middle| D^{k_0}\right) \cdot {}_jI\left(k-1, k_0\right)}. \tag{2.2.16}$$

Obdobně můžeme odvodit podmíněné rozdělení pravděpodobnosti pro hodnoty parametrů modelu podle i-té hypotézy

$$p\left({}_iK \middle| D^{k-1}, {}_iH\right) \propto p\left({}_iK \middle| D^{k_0}, {}_iH\right) \cdot \prod_{\kappa=k_0+1}^{k-1} p\left(y_\kappa \middle| v_\kappa, D^{\kappa-1}, {}_iK, {}_iH\right), \tag{2.2.17}$$

jehož normující členem je přímo (2.2.15) a tedy

$$p\left({}_iK \middle| D^{k-1}, {}_iH\right) = \frac{p\left({}_iK \middle| D^{k_0}, {}_iH\right) \cdot \prod_{\kappa=k_0+1}^{k-1} p\left(y_\kappa \middle| v_\kappa, D^{\kappa-1}, {}_iK, {}_iH\right)}{{}_iI\left(k-1, k_0\right)}. \tag{2.2.18}$$

Podrobnější odvození uvedených vztahů viz [61].

2.2.2 Markovský model soustavy

Protože je obecný stochastický model (2.2.9) vystavěn na bayesovském přístupu, vyžaduje ještě vložení určitých apriorních informací o struktuře a parametrech zkoumaného systému. Přitom je z hlediska praktické realizovatelnosti vhodné hledat takové řešení, aby jeho výpočetní mohutnost nerostla neomezeně s časem. Hledáme tedy takové modely (viz [61]), pro které platí

$$p\left(y_k \middle| v_k, x_{k-1}, D^{k-1}, M\right) = p\left(y_k \middle| v_k, x_{k-1}, M\right),\ k = k_0+1, k_0+2, \ldots, k_K, \tag{2.2.19}$$

tedy že aktuální výstup je podmíněně nezávislý na celé minulé historii procesu za předpokladu, že známe kromě vstupu a modelu ještě vhodnou konečně rozměrnou datovou statistiku

$$X_{k-1} : \varphi_{D^{k-1}} \to R^{\rho_x}, \tag{2.2.20}$$

kde $\varphi_{D^{k-1}}$ je množina všech možných posloupností D^{k-1}, R^{ρ_x} je ρ_x rozměrný Euklidovský prostor a x_{k-1} je ρ_x rozměrný vektor

$$x_{k-1} = X_{k-1}\left(D^{k-1}\right). \tag{2.2.21}$$

Obdobně pro odhad vhodného modelu potřebujeme ρ_s rozměrnou statistiku

$$s_{k-1} = S_{k-1}\left(D^{k-1}\right), \tag{2.2.22}$$

která představuje zobrazení

$$S_{k-1} : \varphi_{D^{k-1}} \to R^{\rho_s}, \tag{2.2.23}$$

kde R^{ρ_s} je ρ_s rozměrný Euklidovský prostor. Pokud platí

$$p\left(M \middle| D^{k-1}\right) = p\left(M \middle| s_{k-1}\right),\ k = k_0+1, k_0+2, \ldots, k_K, \tag{2.2.24}$$

nazveme statistiku s_{k-1} suficientní statistikou náhodné proměnné M.

Ze vztahu (2.2.24) vyplývá, že suficientní statistiku můžeme použít k odhadu struktury i parametrů modelu. Pro statistiky splňující předpoklad (2.2.20) také platí, že jsou spočítatelné rekurzivně, tedy že hodnotu $s_k = S_k\left(D^k\right)$ je možné určit pouze se znalostí $s_{k-1} = S_{k-1}\left(D^{k-1}\right)$ pro libovolný diskrétní čas $k = k_0+1, k_0+2, \ldots, k_K$. [61]

Uvedené vlastnosti dobře splňuje třída stochastických modelů nazvaných Markovské řetězce.

Markovost (markovská vlastnost)

Markovské řetězce jsou stochastické modely charakterizované tzv. markovskou vlastností (nebo markovostí). Markovost říká, že rozdělení pravděpodobnosti přechodu procesu do následujících stavů závisí pouze na aktuálním stavu procesu a nezávisí na stavech předchozích.

Pravděpodobnost, že proces přejde v diskrétním čase k do stavu x_k tedy není určena celou předešlou historií procesu, ale pouze bezprostředně předcházejícím stavem x_{k-1} (viz [43]). Pro rozdělení pravděpodobnosti pak platí

$$p_k\left(x_k \middle| x_{k-1}, x_{k-2}, \ldots, x_0\right) = p_k\left(x_k \middle| x_{k-1}\right). \tag{2.2.25}$$

Podmíněná hustota pravděpodobnosti (2.2.25) se nazývá pravděpodobnost přechodu (přechodová pravděpodobnost).

Pokud je rozdělení pravděpodobností (2.2.25) časově invariantní, jedná se o stacionární nebo též homogenní markovský řetězec.

Předpokládejme homogenní markovský proces, kde stav x_k může nabývat jen konečného počtu hodnot. Tyto hodnoty označme čísly z oboru přirozených čísel $\left\{i : i = 1,2,\ldots,n; i \in N\right\}$. Potom můžeme vytvořit matici přechodových pravděpodobností, ve které každý řádek odpovídá jednomu možnému bezprostředně předcházejícímu stavu x_{k-1} a každý sloupec odpovídá jednomu možnému nastávajícímu stavu x_k. Jednotlivé prvky matice udávají pravděpodobnost přechodu $p\left(j \middle| i\right)$ do stavu $x_k = j$ ze stavu $x_{k-1} = i$ v diskrétním čase k. Zde i, j představují i-tý a j-tý stav z množiny všech možných stavů procesu.

Matice má obecný tvar popsaný vztahem (2.2.26) a nazývá se **matice přechodu** (Transient Matrix) [43], [1], [61]. Platí, že všechny prvky matice jsou nezáporné a součet prvků na každém řádku je vždy roven jedné.

$$\mathbf{K}_k = \begin{bmatrix} p\left(x_k = 1 \middle| x_{k-1} = 1\right) & \cdots & p\left(x_k = N_x \middle| x_{k-1} = 1\right) \\ p\left(x_k = 1 \middle| x_{k-1} = 2\right) & \cdots & p\left(x_k = N_x \middle| x_{k-1} = 2\right) \\ \cdots & p\left(x_k = j \middle| x_{k-1} = i\right) & \cdots \\ p\left(x_k = 1 \middle| x_{k-1} = N_x\right) & \cdots & p\left(x_k = N_x \middle| x_{k-1} = N_x\right) \end{bmatrix} \tag{2.2.26}$$

Pokud vztah (2.2.25) upravíme na

$$p_k\left(x_k \middle| x_{k-1}, x_{k-2}, \ldots, x_0\right) = p_k\left(x_k \middle| x_{k-1}, x_{k-2}, \ldots, x_{k-m}\right), \tag{2.2.27}$$

vznikne markovský řetězec m-tého řádu. Pokud podmíněné pravděpodobnost. (2.2.27) závisí také na nějaké řídicí veličině (například na vstupech sledovaného procesu), jedná se o **řízený markovský řetězec m-tého řádu**

$$p_k\left(x_k \middle| r_k, x_{k-1}, x_{k-2}, \ldots, x_0\right) = p_k\left(x_k \middle| r_k, x_{k-1}, x_{k-2}, \ldots, x_{k-m}\right). \tag{2.2.28}$$

Realizace markovského modelu

Předpokládejme proces popsatelný stacionárním stochastickým modelem diskrétním v čase i v parametrech, s konstantní periodou vzorkování, s výstupem $\mathbf{y}_k$ a

vstupem $\mathbf{v}_k$. Dále předpokládejme, že vstupní i výstupní veličiny mohou nabývat pouze konečného počtu hodnot, které označíme pomocí přirozených čísel.

Vstup $\mathbf{v}_k$ je μ rozměrný vektor vstupních veličin

$$\mathbf{v}_k = \left[v_k[1], v_k[2], \ldots, v_k[\mu] \right]^T \in \varphi_v = \varphi_{v[1]} \times \varphi_{v[2]} \times \ldots \times \varphi_{v[\mu]}, \tag{2.2.29}$$

kde množina φ_v všech možných hodnot vstupního vektoru $\mathbf{v}_k$ je tvořena kartézským součinem μ množin $\varphi_{v[j]}$ možných hodnot jednotlivých vstupních veličin

$$v_k[j] \in \varphi_{v[j]} = \left\{ 1,2,\ldots,N_{v[j]} \right\},\ N_{v[j]} < \infty,\ j = 1,2,\ldots,\mu,\ k = k_0+1, k_0+2, \ldots, k_K \tag{2.2.30}$$

Obdobně výstup $\mathbf{y}_k$ je η rozměrný vektor výstupních veličin

$$\mathbf{y}_k = \left[y_k[1], y_k[2], \ldots, y_k[\eta] \right]^T \in \varphi_y = \varphi_{y[1]} \times \varphi_{y[2]} \times \ldots \times \varphi_{y[\eta]}, \tag{2.2.31}$$

kde množina φ_y všech možných hodnot výstupního vektoru $\mathbf{y}_k$ je tvořena kartézským součinem η množin $\varphi_{y[j]}$ možných hodnot jednotlivých výstupních veličin

$$y_k[j] \in \varphi_{y[j]} = \left\{ 1,2,\ldots,N_{y[j]} \right\},\ N_{y[j]} < \infty,\ j = 1,2,\ldots,\eta,\ k = k_0+1, k_0+2, \ldots, k_K \tag{2.2.32}$$

Mohutnost konečné množiny φ_y (celkový počet všech možných hodnot výstupního vektoru $\mathbf{y}_k$) označíme ρ_{φ_y}.

Takto popsaný diskrétní model je možné aplikovat i na systém se spojitými rozsahy hodnot vstupních a výstupních veličin použitím jejich vhodné kvantizace (diskretizace v hodnotách), typicky rozdělením spojitého rozsahu na konečný počet intervalů a jejich očíslováním.

Nechť uvedený technologický proces je popsatelný homogenním řízeným markovským řetězcem konečného řádu jako závislost diskrétní výstupní veličiny $\mathbf{y}_k$ na diskrétním regresním vektoru $\mathbf{z}_k$ podle vztahu

$$p\left(\mathbf{y}_k \middle| \mathbf{v}_k, D^{k-1} \right) = p\left(\mathbf{y}_k \middle| \mathbf{z}_k \right) \text{ pro } k = k_0+1, k_0+2, \ldots, k_K. \tag{2.2.33}$$

Konečně rozměrný diskrétní regresní vektor $\mathbf{z}_k$ obsahuje informaci o konečné vstupně výstupní historii technologického procesu včetně informace o jeho aktuálním vstupu

$$\mathbf{z}_k = \left\{ \mathbf{v}_k, D_{k-m}^{k-1} \right\};\ m \geq 1, \tag{2.2.34}$$

kde $m \geq 1$ udává maximální hloubku ukládaných dat a tím v souladu s (2.2.27) také řád markovského řetězce.

Protože vlastnosti modelu (2.2.33) neznáme, musíme odhadnout jeho strukturu a příslušné parametry. Je třeba stanovit obecně r hypotéz ${}_iH \quad (i = 1, 2, \ldots, r)$ o struktuře regresního vektoru ${}_i\mathbf{z}_k$ a příslušné parametry markovského řetězce, neboli určit matice pravděpodobností přechodu ${}_i\mathbf{K}$ k daným hypotézám [44], [61]. Vzhledem ke vztahům

(2.2.8), (2.2.33) je možné takto neúplně definovaný markovský model stochastického procesu vyjádřit ve tvaru

$$p\left(\mathbf{y}_k \middle|_i\mathbf{z}_k, {}_i\mathbf{K}, {}_iH\right) \text{ pro } k = k_0+1, k_0+2, \ldots, k_K, \tag{2.2.35}$$

Regresní vektor (2.2.34) má obecnou základní strukturu

$$\mathbf{z}_k = \left[v_k[1]\ldots v_k[\mu], y_{k-1}[1]\ldots y_{k-1}[\eta], v_{k-1}[1]\ldots v_{k-1}[\mu], \ldots\right]^T \tag{2.2.36}$$

a počet jeho prvků se rovná

$$\rho_z = \mu\cdot(m+1)+\eta\cdot m \tag{2.2.37}$$

Hypotézu ${}_iH \ \ (i=1,2,\ldots,r)$ o struktuře regresního vektoru ${}_i\mathbf{z}_k$ je možné chápat jako volbu vybrané kombinace několika prvků základního regresního vektoru

$${}_i\mathbf{z}_{k\langle\rho_{iz},1\rangle} = {}_i\mathbf{J}_{\langle\rho_{iz},\rho_z\rangle}\cdot\mathbf{z}_{k\langle\rho_z,1\rangle}, \tag{2.2.38}$$

kde výběrová matice ${}_i\mathbf{J}$ obsahuje na každém řádku právě jeden a v každém sloupci maximálně jeden prvek s hodnotou 1. Ostatní prvky matice ${}_i\mathbf{J}$ jsou nulové. Umístění jednotkových prvků určuje, které položky ze základního regresního vektoru $\mathbf{z}_k$ budou zahrnuty do regresního vektoru ${}_i\mathbf{z}_k$ podle hypotézy ${}_iH$. Údaje ve špičatých závorkách udávají dimenze vektorů ${}_i\mathbf{z}_k$, $\mathbf{z}_k$ a matice ${}_i\mathbf{J}$. Délka ρ_{iz} regresního vektoru ${}_i\mathbf{z}_k$ se rovná celkovému počtu prvků vybraných ze základního regresního vektoru. Pořadí prvků v regresním vektoru nemá vliv na výsledný model, stejně jako nemá význam provádět vícenásobný výběr stejného prvku.

Složení výběrového regresního vektoru ${}_i\mathbf{z}_k$ určuje strukturu markovského řetězce (= stochastického modelu) a odráží naši představu (případně znalost) o dynamických vlastnostech modelovaného procesu.

Podle (2.2.26) a (2.2.35) je možné prvky ${}_iK_{i\zeta,\upsilon}$ matice přechodu ${}_i\mathbf{K}$ příslušející hypotéze ${}_iH$ vyjádřit jako přechodové aposteriorní pravděpodobnosti toho, že vektor výstupních veličin $\mathbf{y}_k$ nabyde konkrétní hodnoty υ za podmínky, že regresní vektor ${}_i\mathbf{z}_k$ má aktuální hodnotu ${}_i\zeta$

$${}_i\mathbf{K} = \left[{}_iK_{i\zeta,\upsilon} = p\left(\mathbf{y}_k = \upsilon \middle|_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right)\right] \tag{2.2.39}$$

$$\text{pro } {}_i\zeta \in \varphi_{iz},\ \upsilon \in \varphi_y,\ k = k_0+1, k_0+2, \ldots, k_K$$

kde:

${}_i\zeta$ je index jednoznačně přiřazený konkrétní realizaci regresního vektoru ${}_i\mathbf{z}_k$ a určuje, na jakém řádku matice ${}_i\mathbf{K}$ se prvek ${}_iK_{i\zeta,\upsilon}$ nachází;

υ je index jednoznačně přiřazený konkrétní realizaci výstupního vektoru $\mathbf{y}_k$ a určuje, v jakém sloupci matice ${}_i\mathbf{K}$ se prvek ${}_iK_{i\zeta,\upsilon}$ nachází;

$\varphi_{i\mathbf{z}} = \varphi_{i\mathbf{z}[1]} \times \varphi_{i\mathbf{z}[2]} \times \ldots \times \varphi_{i\mathbf{z}[\rho_{i\mathbf{z}}]}$ je konečná množina všech možných realizací (hodnot) regresního vektoru ${}_i\mathbf{z}_k$ o délce $\rho_{i\mathbf{z}}$ sestaveného podle hypotézy ${}_iH$;

$\rho_{\varphi_{i\mathbf{z}}}$ nazveme mohutnost množiny $\varphi_{i\mathbf{z}}$, tedy celkový počet všech možných hodnot regresního vektoru ${}_i\mathbf{z}_k$;

φ_y je konečná množina všech možných realizací (hodnot) vektoru výstupu $\mathbf{y}_k$ viz (2.2.31).

Pro prvky tvořící celý řádek matice ${}_i\mathbf{K}$ použijeme souhrnné označení

$$ {}_i\mathbf{K}_{i\zeta} = \left[{}_iK_{i\zeta,1}, {}_iK_{i\zeta,2}, \ldots, {}_iK_{i\zeta,\rho_{\varphi_y}} \right] \in \varphi_{{}_iK_{i\zeta}}, \tag{2.2.40} $$

kde ρ_{φ_y} je počet prvků konečné množiny φ_y. Vzhledem k obecným vlastnostem rozdělení pravděpodobnosti a k definici (2.2.39) jsou množiny $\varphi_{{}_iK_{i\zeta}}$ možných hodnot řádků ${}_i\mathbf{K}_{i\zeta}$ definovány jako

$$ \varphi_{{}_iK_{i\zeta}} \equiv \left\{ {}_i\mathbf{K}_{i\zeta} : {}_iK_{i\zeta,\upsilon} \geq 0 \text{ pro } \upsilon \in \varphi_y, \sum_{\upsilon \in \varphi_y} {}_iK_{i\zeta,\upsilon} = 1 \right\} \text{ pro } {}_i\zeta \in \varphi_{i\mathbf{z}} \tag{2.2.41} $$

a množinu možných hodnot matice přechodu ${}_i\mathbf{K}$ tvoří jejich kartézský součin

$$ \varphi_{{}_iK} = \underset{{}_i\zeta \in \varphi_{i\mathbf{z}}}{\times} \varphi_{{}_iK_{i\zeta}} \text{ pro } {}_i\zeta \in \varphi_{i\mathbf{z}} \tag{2.2.42} $$

Za uvedených předpokladů můžeme podle [61] postupně odvodit následující vztahy (2.2.43) až (2.2.64)

Odhad pravděpodobností v matici přechodu ${}_i\mathbf{K}$ pro markovský model se strukturou podle hypotézy ${}_iH$ lze podle bayesovského přístupu určit na základě pozorovaných dat D^k jako hustotu pravděpodobnosti podle vztahu (2.2.18) modifikovaného pro markovský model, viz (2.2.33), (2.2.35)

$$ p\left({}_i\mathbf{K} \middle| D^k, {}_iH \right) \propto p\left({}_i\mathbf{K} \middle| D^{k_0}, {}_iH \right) \cdot \prod_{\kappa=k_0+1}^{k} p\left(\mathbf{y}_\kappa \middle| {}_i\mathbf{z}_\kappa, {}_i\mathbf{K}, {}_iH \right). \tag{2.2.43} $$

Vzhledem k (2.2.39) můžeme chápat průběh výpočtu součinů $\prod_{\kappa=k_0+1}^{k} p\left(\mathbf{y}_\kappa \middle| {}_i\mathbf{z}_\kappa, {}_i\mathbf{K}, {}_iH \right)$ v diskrétních časech $\kappa = k_0+1, k_0+2, \ldots, k$ jako postupné procházení přes prvky matice přechodu ${}_i\mathbf{K}$, kdy určitý prvek ${}_iK_{i\zeta,\upsilon}$ bude zahrnut v součinu tolikrát, kolikrát vyplynula kombinace hodnot ${}_i\mathbf{z}_\kappa = {}_i\zeta$, $\mathbf{y}_\kappa = \upsilon$ z naměřené datové posloupnosti $D^k_{k_0+1}$, takže

$$\prod_{\kappa=k_0+1}^{k} p\left(\mathbf{y}_\kappa \middle|_i\mathbf{z}_\kappa, {}_iK, {}_iH\right) = \prod_{{}_i\zeta\in\varphi_{iz}} \prod_{\upsilon\in\varphi_y} \prod_{\kappa=k_0+1}^{k} {}_iK_{i\zeta,\upsilon} = \prod_{{}_i\zeta\in\varphi_{iz}} \prod_{\upsilon\in\varphi_y} {}_iK_{i\zeta,\upsilon}^{{}_in^1_{i\zeta,\upsilon}(k)} \quad , \tag{2.2.44}$$

kde ${}_in^1_{i\zeta,\upsilon}(k)$ je počet událostí, kdy ${}_i\mathbf{z}_\kappa = {}_i\zeta$ a zároveň $\mathbf{y}_\kappa = \upsilon$ v průběhu časového intervalu $k_0+1<\kappa\le k$.

Pokud ještě podle [61] zvolíme apriorní rozdělení pravděpodobnosti $p\left({}_i\mathbf{K}\middle|D^{k_0}, {}_iH\right)$ tak, aby platilo

$$p\left({}_i\mathbf{K}\middle|D^{k_0}, {}_iH\right) = \prod_{{}_i\zeta\in\varphi_{iz}} p\left({}_i\mathbf{K}_{i\zeta}\middle|D^{k_0}, {}_iH\right), \tag{2.2.45}$$

potom můžeme jednotlivé řádky ${}_i\mathbf{K}_{i\zeta}$ přechodové matice ${}_i\mathbf{K}$ identifikovat nezávisle, takže platí

$$p\left({}_i\mathbf{K}\middle|D^{k}, {}_iH\right) = \prod_{{}_i\zeta\in\varphi_{iz}} p\left({}_i\mathbf{K}_{i\zeta}\middle|D^{k}, {}_iH\right) = \prod_{{}_i\zeta\in\varphi_{iz}} \frac{p\left({}_i\mathbf{K}_{i\zeta}\middle|D^{k_0}, {}_iH\right)\cdot\prod_{\upsilon\in\varphi_y} {}_iK_{i\zeta,\upsilon}^{{}_in^1_{i\zeta,\upsilon}(k)}}{{}_iI_{i\zeta}\left(k,k_0\right)}, \tag{2.2.46}$$

kde

$${}_iI_{i\zeta}\left(k,k_0\right) = \int_{\varphi_{i\kappa_{i\zeta}}} p\left({}_i\mathbf{K}_{i\zeta}\middle|D^{k_0}, {}_iH\right)\cdot\prod_{\upsilon\in\varphi_y} {}_iK_{i\zeta,\upsilon}^{{}_in^1_{i\zeta,\upsilon}(k)}\cdot d\,{}_i\mathbf{K}_{i\zeta}\,. \tag{2.2.47}$$

Když zvolíme apriorní rozdělení $p\left({}_i\mathbf{K}_{i\zeta}\middle|D^{k_0}, {}_iH\right)$ z (2.2.46) ve tvaru

$$p\left({}_i\mathbf{K}_{i\zeta}\middle|D^{k_0}, {}_iH\right) = \frac{\chi\left({}_i\mathbf{K}_{i\zeta}, \varphi_{{}_iK_{i\zeta}}\right)\cdot\prod_{\upsilon\in\varphi_y} {}_iK_{i\zeta,\upsilon}^{{}_in_{i\zeta,\upsilon}(k_0)-1}}{{}_iI_{i\zeta}\left(k_0,k_0\right)}, \tag{2.2.48}$$

kde ${}_in_{i\zeta,\upsilon}\left(k_0\right)$ představuje apriorně zvolenou hodnotu, která by se dala interpretovat jako počet událostí, kdy ${}_i\mathbf{z}_\kappa = {}_i\zeta$ a zároveň $\mathbf{y}_\kappa = \upsilon$ před zahájením identifikace.

Symbolické označení $\chi\left({}_i\mathbf{K}_{i\zeta}, \varphi_{{}_iK_{i\zeta}}\right)$ odráží podmínku normování z (2.2.41)

$$\chi\left({}_i\mathbf{K}_{i\zeta}, \varphi_{{}_iK_{i\zeta}}\right) \equiv \begin{cases} 1 \text{ pro } {}_i\mathbf{K}_{i\zeta} \in \varphi_{{}_iK_{i\zeta}} \\ 0 \text{ pro } {}_i\mathbf{K}_{i\zeta} \notin \varphi_{{}_iK_{i\zeta}} \end{cases} \tag{2.2.49}$$

Z (2.2.46), (2.2.48) dostaneme aposteriorní rozdělení pravděpodobnosti

$$p\left({}_i\mathbf{K}_{i\zeta}\middle|D^{k}, {}_iH\right) = \frac{\chi\left({}_i\mathbf{K}_{i\zeta}, \varphi_{{}_iK_{i\zeta}}\right)\cdot\prod_{\upsilon=1}^{\rho_{\varphi_y}} {}_iK_{i\zeta,\upsilon}^{{}_in_{i\zeta,\upsilon}(k_0)+{}_in^1_{i\zeta,\upsilon}(k)-1}}{{}_iI_{i\zeta}\left(k,k_0\right)}, \tag{2.2.50}$$

Odkud zavedeme souhrnné označení

$$ {}_i n_{i\zeta,\upsilon}(k) = {}_i n_{i\zeta,\upsilon}(k_0) + {}_i n^1_{i\zeta,\upsilon}(k). \tag{2.2.51} $$

S využitím formální podobnosti s vícerozměrnou Eulerovou beta funkcí [45], [61], viz obecný vztah pro reálné proměnné $P_i, i = 1,2,\ldots,m$

$$ \int_{\varphi_P} \left(\prod_{i=1}^{m-1} P_i^{n_i} \right) \cdot \left(1 - \sum_{j=1}^{m-1} P_j^{n_j} \right)^{n_m} \cdot dP_1 \cdot dP_2 \cdot \ldots \cdot dP_m = \frac{\prod_{i=1}^{m} \Gamma(n_i + 1)}{\Gamma\left(\sum_{j=1}^{m} (n_i + 1) \right)}, \tag{2.2.52} $$

kde

$$ \varphi_P = \left\{ P_1, P_2, \ldots, P_m : P_i \geq 0 \text{ pro } i = 1,2,\ldots,m; \sum_{i=1}^{m} P_i = 1 \right\} \tag{2.2.53} $$

a kde $\Gamma(.)$ je Eulerova gama funkce

$$ \Gamma(x) = \int_0^{\infty} e^{-t} \cdot t^{x-1} \cdot dt, \tag{2.2.54} $$

můžeme odvodit normalizační konstantu hustoty pravděpodobnosti (2.2.50)

$$ {}_i I_{i\zeta}(k, k_0) = \int_{\varphi_{i\mathbf{K}_{i\zeta}}} \left(\prod_{\upsilon=1}^{\rho_{\varphi_y}-1} {}_i K_{i\zeta,\upsilon}^{{}_i n_{i\zeta,\upsilon}(k)-1} \right) \cdot \left(1 - \sum_{\upsilon=1}^{\rho_{\varphi_y}-1} {}_i K_{i\zeta,\upsilon}^{{}_i n_{i\zeta,\upsilon}(k)-1} \right)^{{}_i n_{i\zeta,\rho_{\varphi_y}}(k)-1} \cdot d\,{}_i\mathbf{K} = $$

$$ = \frac{\prod_{\upsilon=1}^{\rho_{\varphi_y}} \Gamma\left({}_i n_{i\zeta,\upsilon}(k) \right)}{\Gamma\left(\sum_{\upsilon=1}^{\rho_{\varphi_y}} {}_i n_{i\zeta,\upsilon}(k) \right)} . \tag{2.2.55} $$

Výsledné aposteriorní rozdělení pravděpodobnosti pro odhad matice přechodu ${}_i\mathbf{K}$ přiřazené hypotéze ${}_i H$ pak bude

$$ p\left({}_i\mathbf{K} \middle| D^k, {}_i H \right) = \left(\prod_{{}_i\zeta \in \varphi_{iz}} \frac{\Gamma\left(\sum_{\upsilon \in \varphi_y} {}_i n_{i\zeta,\upsilon}(k) \right)}{\prod_{\upsilon \in \varphi_y} \Gamma\left({}_i n_{i\zeta,\upsilon}(k) \right)} \cdot \left(\prod_{\upsilon \in \varphi_y} {}_i K_{i\zeta,\upsilon}^{{}_i n_{i\zeta,\upsilon}(k)-1} \right) \right) \cdot \chi\left({}_i\mathbf{K}, \varphi_{{}_i K} \right), \tag{2.2.56} $$

kde podobně jako v (2.2.49) je použito označení

$$ \chi\left({}_i\mathbf{K}, \varphi_{{}_i\mathbf{K}} \right) = \prod_{{}_i\zeta \in \varphi_{iz}} \chi\left({}_i\mathbf{K}_{i\zeta}, \varphi_{{}_i K_{i\zeta}} \right) \equiv \begin{cases} 1 \text{ pro } {}_i\mathbf{K} \in \varphi_{{}_i\mathbf{K}} \\ 0 \text{ pro } {}_i\mathbf{K} \notin \varphi_{{}_i\mathbf{K}} \end{cases} \tag{2.2.57} $$

Zbývá odhadnout rozdělení $p\left(\mathbf{y}_{k+1}\middle|{}_i\mathbf{z}_{k+1},{}_iH\right)$ nezávislé na konkrétních parametrech, které potřebujeme pro predikci budoucího výstupu $\mathbf{y}_{k+1}$ za podmínky, že známe aktuální vstup $\mathbf{v}_{k+1}$ a minulou historii D^k sledovaného procesu a z nich také na základě (2.2.34), (2.2.38) aktuální hodnotu regresního vektoru ${}_i\mathbf{z}_{k+1}={}_i\zeta$, jehož struktura byla zvolena podle hypotézy ${}_iH$

$$\begin{aligned} &p\left(\mathbf{y}_{k+1}=\upsilon\middle|{}_i\mathbf{z}_{k+1}={}_i\zeta,D^k,{}_iH\right)= \\ &=\int_{\varphi_{i\mathbf{K}}} p\left(\mathbf{y}_{k+1}=\upsilon\middle|{}_i\mathbf{z}_{k+1}={}_i\zeta,{}_i\mathbf{K},{}_iH\right)\cdot p\left({}_i\mathbf{K}\middle|D^k,{}_iH\right)\cdot d\,{}_i\mathbf{K} \end{aligned} \tag{2.2.58}$$

Využijeme vztahů (2.2.39), (2.2.52) a (2.2.56) a vlastnosti Eulerovy gama funkce

$$\Gamma\left(x+1\right)=x\cdot\Gamma\left(x\right). \tag{2.2.59}$$

Potom můžeme podle [61] snadno odvodit z (2.2.58) výsledné aposteriorní rozdělení

$$p\left(\mathbf{y}_{k+1}=\upsilon\middle|{}_i\mathbf{z}_{k+1}={}_i\zeta,D^k,{}_iH\right)=\frac{{}_in_{i\zeta,\upsilon}\left(k\right)}{\sum_{\upsilon_p\in\varphi_y}{}_in_{i\zeta,\upsilon_p}\left(k\right)}\,, \tag{2.2.60}$$

kde ${}_in_{i\zeta,\upsilon}\left(k\right)$ je podle (2.2.51) počet událostí, kdy vektor výstupních veličin $\mathbf{y}_\kappa=\upsilon$ a zároveň regresní vektor ${}_i\mathbf{z}_\kappa={}_i\zeta$ pro diskrétní časy $\kappa\le k$, tedy v průběhu celé minulé historie až do současného okamžiku. Výraz (2.2.60) tedy vyjadřuje relativní četnost událostí, kdy po vstupu ${}_i\zeta$ následoval výstup υ.

Obdobně by bylo možné (viz [61]) odvodit na základě známé historie procesu D^k odhad struktury regresního vektoru markovského modelu, který vede na aposteriorní rozdělení pravděpodobnosti

$$p\left({}_iH\middle|D^k\right)\propto p\left({}_iH\middle|D^{k_0}\right)\cdot\prod_{\kappa=k_0+1}^{k}\frac{{}_in_{i\mathbf{z}_\kappa,y_\kappa}\left(\kappa-1\right)}{\sum_{\upsilon_p\in\varphi_y}{}_in_{i\mathbf{z}_\kappa,\upsilon_p}\left(\kappa-1\right)}\,, \tag{2.2.61}$$

resp. na výhodnější rekurzivní tvar

$$p\left({}_iH\middle|D^k\right)\propto p\left({}_iH\middle|D^{k-1}\right)\cdot\frac{{}_in_{i\mathbf{z}_k,y_k}\left(k-1\right)}{\sum_{\upsilon_p\in\varphi_y}{}_in_{i\mathbf{z}_k,\upsilon_p}\left(k-1\right)} \tag{2.2.62}$$

pro $i=1,2,\ldots,r$, $k=k_0+1,k_0+2,\ldots,k_K$,

kde ${}_in_{i\mathbf{z}_\kappa,y_\kappa}\left(\kappa-1\right)$ představují podobně jako v (2.2.51) apriorně zvolenými hodnotami korigované počty událostí v časovém intervalu $k_0<\kappa\le k-1$, kdy regresní a výstupní vektor nabyly konkrétních hodnot ${}_i\mathbf{z}_\kappa$, $\mathbf{y}_\kappa$ získaných z naměřených dat D^κ podle hypotézy ${}_iH$.

Dalším odvozením podle [61] určíme prediktivní model procesu (2.2.9) nezávislý na hypotéze o struktuře a na parametrech, tedy aposteriorní hustotu pravděpodobnosti

$$p\left(\mathbf{y}_{k+1}\middle|{}_i\mathbf{z}_{k+1},D^k\right)=p\left(\mathbf{y}_{k+1}\middle|\mathbf{v}_{k+1},D^k\right)=\sum_{i=1}^{r}p\left(\mathbf{y}_{k+1}\middle|\mathbf{v}_{k+1},D^k,{}_iH\right)\cdot p\left({}_iH\middle|D^k\right)\propto$$

$$\propto\sum_{i=1}^{r}p\left({}_iH\middle|D^{k_0}\right)\cdot\prod_{\kappa=k_0+1}^{k+1}\frac{{}_in_{{}_iz_\kappa,y_\kappa}\left(\kappa-1\right)}{\sum_{\upsilon_p\in\varphi_y}{}_in_{{}_iz_\kappa,\upsilon_p}\left(\kappa-1\right)}\quad.\qquad(2.2.63)$$

pro $k=k_0+1,k_0+2,\ldots,k_K$,

kde $p\left({}_iH\middle|D^{k_0}\right)$ je apriorně zvolené výchozí rozdělení, pomocí kterého přiřazujeme subjektivní míru důvěry jednotlivým hypotézám o struktuře regresního vektoru. Pokud přidělíme všem hypotézám stejnou apriorní pravděpodobnost $p\left({}_iH\middle|D^{k_0}\right)=1/r$, můžeme vztah (2.2.63) zjednodušit na

$$p\left(\mathbf{y}_{k+1}\middle|\mathbf{v}_{k+1},D^k\right)\propto\sum_{i=1}^{r}\prod_{\kappa=k_0+1}^{k+1}\frac{{}_in_{{}_iz_\kappa,y_\kappa}\left(\kappa-1\right)}{\sum_{\upsilon_p\in\varphi_y}{}_in_{{}_iz_\kappa,\upsilon_p}\left(\kappa-1\right)}\ \text{pro}\ k=k_0+1,k_0+2,\ldots,k_K.\qquad(2.2.64)$$

Aposteriorní rozdělení pravděpodobnosti (2.2.63), resp. (2.2.64) ukazuje, že není nezbytně nutné rozhodnout ani o konkrétní struktuře regresního vektoru. Stačí, když sestavíme o této struktuře dostatečně obsáhlý soubor hypotéz.

Výpočet vztahů (2.2.56), (2.2.60), (2.2.62) a (2.2.63) je možné opět chápat jako průběžnou korekci naší apriorní představy o vlastnostech markovského modelu objektivními daty získanými ze sledovaného procesu.

Matice četností

Ze vztahů pro výpočet aposteriorních hustot pravděpodobnosti $p\left({}_i\mathbf{K}\middle|D^k,{}_iH\right)$, $p\left(\mathbf{y}_{k+1}\middle|{}_i\mathbf{z}_{k+1},D^k,{}_iH\right)$, $p\left({}_iH\middle|D^k\right)$ a $p\left(\mathbf{y}_{k+1}\middle|{}_i\mathbf{z}_{k+1},D^k\right)$ vyplývá (viz [61], [44], [45]), že pro odhad chování markovského řetězce stačí znát matice absolutních četností (dále jen matice četností) ${}_i\mathbf{n}\left(k\right)$, $i=1,2,\ldots,r$ určené součty

$${}_i\mathbf{n}\left(k\right)={}_i\mathbf{n}\left(k_0\right)+{}_i\mathbf{n}^1\left(k\right)\ ,\ i=1,2,\ldots,r\ ,\qquad(2.2.65)$$

kde ${}_i\mathbf{n}\left(k_0\right)$, $i=1,2,\ldots,r$ jsou matice, jejichž prvky ${}_in_{{}_i\zeta,\upsilon}\left(k_0\right)$ nabývají apriorně zvolených nezáporných hodnot a vyjadřují naši subjektivní míru důvěry, že z naměřených dat $D_{k_0+1}^k$ získáme pomocí hypotézy ${}_iH$ sdruženou dvojici $\left\{{}_i\mathbf{z}_\kappa={}_i\zeta,\mathbf{y}_\kappa=\upsilon\right\}$ a můžeme je interpretovat jako počet takových událostí ještě před zahájením identifikace, tedy v diskrétních časech $\kappa\le k_0$.

${}_i\mathbf{n}^1\left(k\right)$, $i=1,2,\ldots,r$ jsou matice, jejichž prvky ${}_in^1_{{}_i\zeta,\upsilon}\left(k\right)$ představují objektivně zjištěný počet výskytů sdružených dvojic $\left\{{}_i\mathbf{z}_\kappa={}_i\zeta,\mathbf{y}_\kappa=\upsilon\right\}$ získaných z naměřených dat

$D_{k_0+1}^{k}$ pomocí hypotézy ${}_iH$ pro diskrétní časy $k_0 < \kappa \le k$ a vzhledem k (2.2.56) je spočítáme pomocí vztahu

$${}_i n^1_{{}_i\zeta,\upsilon}(k) = \sum_{\kappa=k_0+1}^{k} \delta({}_i\zeta, {}_i\mathbf{z}_\kappa) \cdot \delta(\upsilon, \mathbf{y}_\kappa) \text{ pro } \upsilon \in \varphi_y \text{ a } {}_i\zeta \in \varphi_{{}_i z} \tag{2.2.66}$$

Symbol $\delta(\alpha,\beta)$ se nazývá Kroneckerův delta operátor a je definován předpisem

$$\delta(\alpha,\beta) \equiv \begin{cases} 1 & pro \quad \alpha = \beta \\ 0 & pro \quad \alpha \neq \beta \end{cases} \tag{2.2.67}$$

Statistika (2.2.66) je počítána jednorázově, po skončení měření. V praxi je však často výhodnější počítat ji průběžně – během sledování činnosti soustavy. Proto je výhodné převézt uvedený vztah také v rekurzivní formě

$${}_i n^1_{{}_i\zeta,\upsilon}(k) = {}_i n^1_{{}_i\zeta,\upsilon}(k-1) + \delta({}_i\zeta, {}_i\mathbf{z}_k) \cdot \delta(\upsilon, \mathbf{y}_k) \tag{2.2.68}$$

pro $i = 1, 2, \ldots, r$, $k = k_0 + 1, k_0 + 2, \ldots, k_K$, $\upsilon \in \varphi_y$ a ${}_i\zeta \in \varphi_{{}_i z}$.

Pokud se parametry stochastické soustavy průběžně mění, je možno do výpočtu statistiky (2.2.68) resp. (2.2.66) zahrnout ještě zapomínání (např. metodou exponenciálního zapomínání) pro zmenšení vlivu starších dat.

Hlavním kladem modelování s markovskými řetězci je možnost pracovat i se silně nelineárními systémy a poměrně snadná identifikace.

Nevýhodou, zejména při práci v reálném čase, je velký objem zpracovávaných dat vzhledem k velkým rozměrům zpracovávaných matic.

2.2.3 Diagnostika poruch s markovskými řetězci

Princip diagnostiky poruch s markovskými řetězci

Základní myšlenkou je vytvoření zvláštního markovského modelu (2.2.33), jehož regresní vektor je sestaven ze vstupních i výstupních veličin sledované soustavy [1]. Výstupem modelu však není odhad výstupu soustavy, nýbrž veličina, jež klasifikuje její pracovní režim, tedy poruchový stav soustavy. Markovský systém detekce a identifikace poruch přiřazuje jednotlivým druhům chybových režimů hodnoty z oboru přirozených čísel a je přirozené bezporuchovému stavu přiřadit hodnotu 0. Takto popsaný stochastický model založený na markovském řetězci označíme bayesovský klasifikátor.

Klíčový význam pro dobré fungování FDI má opět vhodná volba struktury regresního vektoru (viz [44]). Postupy pro formulace a vyhodnocování hypotéz o struktuře regresního vektoru byly popsány v předchozí kapitole.

Pro nalezení parametrů modelu a vytvoření suficientních statistik pro zvolené hypotézy použijeme opět postupy z předchozí kapitoly. K identifikaci stochastického modelu potřebujeme dostatečně velké množství dat ze všech předpokládaných poruchových, ale i bezporuchových stavů soustavy.

Pasivní diagnostika

Podle načasování učení existují obecně dva přístupy, jak provádět diagnostiku poruch [45]. Prvním z nich je postup, kdy se diagnostika provádí s použitím předem naučeného modelu.

Proces probíhá ve dvou oddělených fázích.

Fáze učení

V první fázi se musí markovský model se zvoleným regresním vektorem naučit jednorázově na základě známých naměřených dat ze soustavy. Tato data musí být vhodně zvolena, musí postihovat všechny možné poruchové i bezporuchové stavy soustavy, ke kterým by mohlo dojít. Ke každé takto získané konkrétní hodnotě regresního vektoru se přiřadí hodnota funkce stavu (poruchy) soustavy, ve kterém byla tato data zaznamenána. Tímto způsobem vznikne statistika popisující danou soustavu z hlediska diagnostiky poruch.

Fáze diagnostiky

Ve druhé fázi se již markovský model nemění. Veličiny soustavy jsou průběžně sledovány a je z nich tvořen regresní vektor. Ten se potom vyhledává v naučených datech a výstupem je pravděpodobnostní rozdělení možného výskytu pro všechny naučené poruchové a bezporuchové stavy. Z tohoto rozdělení pravděpodobnosti se vybere stav, který nejlépe odpovídá zvolenému kritériu. Nejběžnější je zvolit stav s největší pravděpodobností. Informace o stavu je následně předána k dalšímu využití.

Diagnostika a učení v reálném čase

Druhý přístup nevyžaduje předem kompletně naučený model. Data jsou sbírána průběžně v reálném čase, jsou z nich generovány regresní vektory a probíhá diagnostika stejně jako v předchozím případě. Zároveň s diagnostikou však neustále probíhá také učení podle pokynů operátora. Obvykle jsou tedy nejprve sebrána data o bezporuchovém stavu. Když nastane porucha, musí operátor ručně oznámit systému

FDI, že nastal nový poruchový stav. Systém se tak průběžně doučuje nové poruchové stavy a neustále se zlepšuje v samostatném rozpoznávání typu poruchy.

Výhodou metody je, že statistika modelu neobsahuje zbytečně stavy, které se v soustavě nevyskytují a naopak v případě, že bychom nějakou poruchu opomenuli, je možné ji dodatečně doučit. Nevýhodou je naopak dlouhé počáteční období, kdy model ještě neumí stavy rozpoznávat.

Kombinovaná diagnostika

Každý z obou uvedených přístupů má určité slabiny, proto se obvykle využívá kombinace obou. To znamená, že nejprve se model naučí na zkušebních datech. (Pokud je to možné, nasimulují se například na soustavě i nejběžnější poruchové stavy.)

Poté během provozu je možno model rovnou využívat pro FDI, zároveň však nejsou kladeny tak velké požadavky na kvalitu předběžného učení, protože se počítá s následným doučováním i později.

Redukce rozměrnosti matice přechodu

Velkým problémem metod založených na markovských řetězcích je i dnes velký rozměr statistik, resp. matice přechodu. Proto je třeba vhodným způsobem velikost statistiky omezit.

Existuje celá řada metod, jednou z nich je například *metoda aproximace predikce založené na markovských řetězcích* (AMCP) [45]. Aproximace se provádí v okolí aktuálního nově vytvořeného regresního vektoru. Nejprve proběhne *lokalizace okolí*, kdy algoritmus vyhledá všechny regresní vektory, které se nacházejí v pevně stanoveném okolí aktuálního vektoru. Následně proběhne *výpočet vah*, kdy se stanovuje význam jednotlivých nalezených vektorů podle relativní vzdálenosti. Nakonec se na základě rozdělení pravděpodobnosti nalezených okolních vektorů a jejich vah vypočte výsledné rozdělení pravděpodobnosti pro daný regresní vektor. Díky tomuto postupu stačí ukládat mnohem menší množství regresních vektorů při zachování dobré kvality FDI.

Další metodou redukce rozměrnosti markovských modelů je *metoda odhadu ideálního predikátoru* [1]. Původní rozměrný regresní vektor se rozdělí na několik subvektorů a za pomoci metody přibližného skládání stochastických modelů se následně určí věrohodný odhad ideálního prediktoru.

U spojitých soustav je možné použít metodu aproximativní identifikace [1], ve které se provede relativně hrubá diskretizace a rozdělení pravděpodobnosti se potom vyhlazuje spojitou nebo křivkovou aproximací.

Regresní vektor

Jak bylo popsáno v kapitole 2.2.2, struktura regresního vektoru obecně vyjádřeného vztahem (2.2.36) představuje podstatnou část apriorní složky markovského modelu (2.2.35) technologického procesu. Jako jednotlivé prvky regresního vektoru jsou použity vybrané signály naměřené na technologickém procesu – údaje senzorů, informace z ovládacích panelů o uživatelských zásazích apod. Nabízí se ale také možnost zpracovat tyto „surové“ signály a rozložit je vhodným způsobem na dílčí signály – složky. Může být prospěšné takovou komponentu využít při sestavení regresního vektoru a dosáhnout tak lepšího vyjádření dynamiky dějů odehrávajících se v procesu, například ke zvýraznění rozdílů mezi dvěma poruchami s podobnou dynamikou. Rozkladem signálu na dílčí komponenty se zabývá následující kapitola.

2.3 Časově frekvenční rozklad signálu

Signály, které slouží ke zjištění informací o činnosti procesu, získáváme obvykle v technické praxi přirozeným způsobem ve formě reálných funkcí závislých na čase. Pojem signál zde bude použit v rozšířeném kontextu pro data obsahující kromě užitečné informace také náhodný šum. Naměřený signál jako reálná funkce času $x(t)$ představuje zobrazení $R \to R$, kde R představuje množinu reálných čísel. V diskrétní formě může být vyjádřen jako posloupnost čísel $\{x_k\}$. Reprezentace signálu ve formě číselné hodnoty nepředstavuje pro praktické použití významné omezení, např. lingvistické proměnné je možné snadno označit pomocí číselných indexů.) V teorii signálu se pracuje obvykle s obecnějšími komplexními signály. Naměřený reálný signál pak obvykle představuje reálnou složku komplexního signálu. [54],[55]

2.3.1 Spektrum signálu

Zkoumání signálu pouze v časové oblasti nemusí být vždy výhodné. Užitečná data mohou být překryta náhodným šumem nebo jinými, pro daný cíl neužitečnými, složkami signálu. Zejména v případě periodických dějů, například u motorů, čerpadel a jiných rotačních strojů, bývá generováno velké množství periodických složek, které jsou přítomny bez ohledu na aktuální výskyt poruchy. Proto se provádí také analýza signálu ve frekvenční oblasti.

Jako základní metoda frekvenční analýzy signálu slouží Fourierova transformace. Je založena na Fourierově nekonečné řadě

$$x(t) = \sum_{k=-\infty}^{\infty} F_k \cdot e^{\mathrm{j}\frac{2\pi}{T}kt}, \tag{2.3.1}$$

$$F_k = \frac{1}{T}\int_T x(t) \cdot e^{-\mathrm{j}\frac{2\pi}{T}kt} dt \;;\quad k = 0, \pm 1, \pm 2, \ldots, \pm\infty, \tag{2.3.2}$$

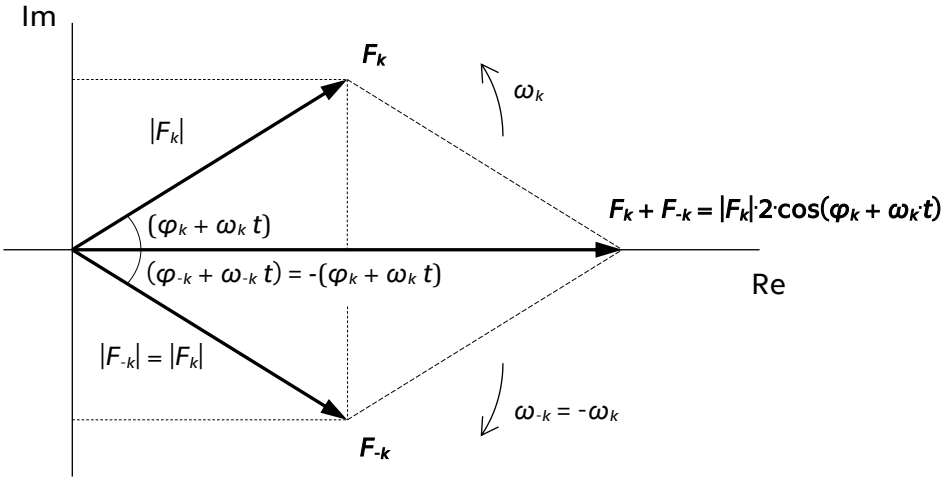
pomocí které je možné rozložit libovolnou periodickou funkci s periodou T na nekonečné množství harmonických složek za předpokladu, že tato funkce je alespoň po částech hladká [54]. Každý prvek řady (2.3.1) představuje *k*-tý vektor rotující v komplexní rovině s úhlovou frekvencí

$$\omega_k = \frac{2\pi}{T}k. \tag{2.3.3}$$

Modul $|F_k|$ komplexního koeficientu (2.3.2) určuje amplitudu a argument $\varphi_k = \arg(F_k)$ počáteční fázi rotace tohoto vektoru. Koeficient F_0 je vždy reálný a představuje střední hodnotu periodické funkce $x(t)$, protože přísluší prvku s nulovou frekvencí a tedy konstantnímu. Jak ukazuje vztah (2.3.3), frekvence harmonických složek periodického signálu nejsou libovolné. Vždy musí být celočíselným násobkem základní frekvence dané periodou T. Pokud je funkce $x(t)$ reálná (viz úvaha na začátku kapitoly), pak každá dvojice výrazů z (2.3.1) odpovídajících koeficientům $\pm k$ reprezentuje komplexně sdruženou dvojici vektorů (fázorů) rotujících v opačných směrech [54], [56], jak ukazuje obr. 2.3.1.

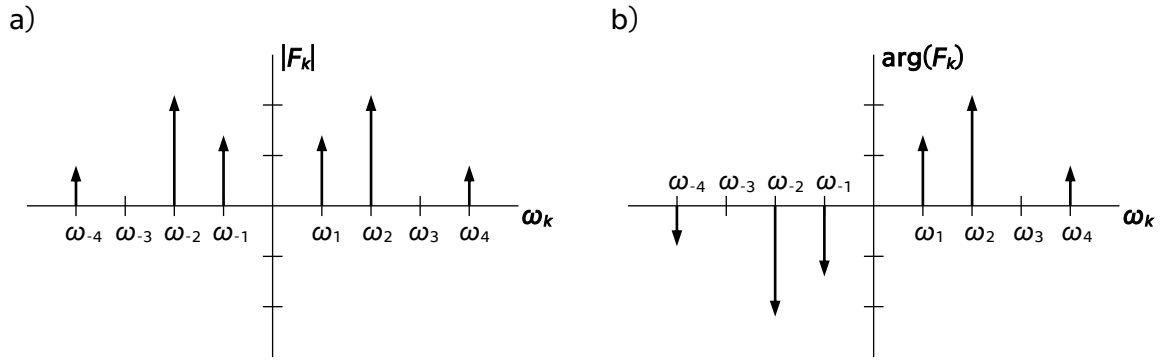
Jejich sečtením vznikne *k*-tá (reálná) harmonická složka funkce $x(t)$, jejíž kmitání lze popsat vztahem

$$2 \cdot |F_k| \cdot \cos(\omega_k \cdot t + \varphi_k). \tag{2.3.4}$$



obr. 2.3.1 - Rotace fázorů v komplexní rovině

Spektrum funkce $x(t)$ je komplexní funkcí frekvence a vyjádřením v polárních souřadnicích je možné ho zobrazit do grafu jako závislost amplitudy $|F_k|$ (amplitudové spektrum) a fáze φ_k (fázové spektrum) na frekvenci ω_k. Pro periodickou funkci je spektrum diskrétní s intervalem rovným základní frekvenci ω_1, viz (2.3.1) a (2.3.3). Amplitudové spektrum reálné funkce je sudé a nezáporné, fázové spektrum je liché, jak naznačuje obr. 2.3.2.



obr. 2.3.2 – Spektrum periodického signálu; a) amplitudové spektrum, b) fázové spektrum.

Fourierova transformace vznikne zobecněním Fourierovy řady (2.3.1) také na neperiodické funkce. Předpokládejme, že signál $x(t)$ je periodický s periodou T. Pokud periodu prodloužíme limitně k nekonečnu, základní frekvence ω_1 se přiblíží k nule a diskrétní složky spektra splynou do spojité funkce (viz [56]).

Fourierova transformace je popsaná vztahem

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt = F\{x(t)\} \ , \tag{2.3.5}$$

kde Fourierův obraz $X(\omega)$ představuje spojité spektrum (obecně neperiodického) časového signálu $x(t)$. Spektrum je opět komplexní funkce frekvence a je ho možno vyjádřit ve formě amplitudového a fázového spektra. Zpětná Fourierova transformace má tvar

$$x(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega) e^{j\omega t} d\omega = F^{-1}\{X(\omega)\} \ . \tag{2.3.6}$$

V praxi se častěji používá varianta Fourierovy transformace vhodná pro počítačové zpracování – Diskrétní Fourierova transformace (DFT). DFT pracuje se signálem, který je diskrétní v čase a obsahuje konečný počet naměřených vzorků. Obraz získaný pomocí DFT je diskrétní i ve frekvenční oblasti, protože je na něj možno pohlížet jako na periodický s volitelnou periodou odpovídající rozsahu naměřených dat anebo delší. Dopřednou a zpětnou DFT představují vztahy (2.3.7) a (2.3.8), kde N je počet vzorků (periodické) časové posloupnosti a zároveň počet prvků frekvenčního spektra. Pokud je zvolené N větší, než počet skutečně známých vzorků, je nutné prodloužit posloupnost $\{x_n\}$ o potřebný počet vzorků s nulovou hodnotou.

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-j\frac{2\pi}{N} \cdot k \cdot n} = DFT\{x_n\}, \quad k = 0,1,2,\ldots,N-1 \tag{2.3.7}$$

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \cdot e^{j\frac{2\pi}{N} \cdot k \cdot n} = DFT^{-1}\{X_k\}, \quad n = 0,1,2,\ldots,N-1 \tag{2.3.8}$$

Mezi dopřednou a zpětnou Fourierovou transformací, DFT nevyjímaje, existuje značná symetrie vlastností. Lze snadno odvodit, že jako periodický průběh v časové oblasti vede na diskrétní spektrum, tak i diskrétní charakter časového signálu znamená, že spektrum bude periodické s diskrétní periodou odpovídající N [56].

DFT je plně diskrétní a tedy nezávislá na velikosti vzorkovací periody Δt, stejně jako na velikosti kroku diskrétního spektra. Pro zohlednění časového měřítka do frekvenčního spektra je nutno upravit hodnotu Fourierova obrazu dle (2.3.9) a obdobně zpětné DFT dle (2.3.10).

$$X(\omega_k) = X\left(\frac{2\pi}{N \cdot \Delta t} k\right) = \Delta t \cdot X_k \approx \Delta t \cdot DFT\{x_n\} \tag{2.3.9}$$

$$x(t_n) = x(\Delta t \cdot n) = \frac{1}{\Delta t} x_n \approx \frac{1}{\Delta t} DFT^{-1}\{X_k\} \tag{2.3.10}$$

Samotná DFT je poměrně výpočetně náročná. Proto byla vytvořena skupina výrazně efektivnějších algoritmů souhrnně nazvaných Rychlá Fourierova transformace (FFT, Fast Fourier Transform). První a nejznámější algoritmus FFT byl zveřejněn v roce 1965 (viz [59]), i když později byl objeven již v Gaussově málo známé práci z roku 1805.

Je založen na rekurzivním rozdělování výpočtu DFT na menší části. Počet potřebných operací se tak z N^2 sníží u FFT pouze na $N \cdot \log(N)$. Blíže o různých variantách algoritmu FFT viz např. [54], [55], [58], [59].

2.3.2 Metody časově frekvenční analýzy signálu

Fourierova transformace vychází z předpokladu, že jednotlivé harmonické složky signálu jsou přítomny v celém jeho časovém průběhu, resp. na celém zaznamenaném intervalu. Jedná se tedy o analýzu signálu čistě ve frekvenční oblasti. To však přináší určité problémy, pokud nejsou data (alespoň po částech) stacionární nebo pokud nereprezentují lineární proces [46]. Dalším významným problémem Fourierovy transformace je předem zvolený tvar dílčích složek signálu, který nemusí nijak souviset se skutečnou fyzikální podstatou zkoumaného děje.

Metody, které zohledňují proměnlivost signálu v čase, jsou velmi důležité právě v oblasti detekce poruch. Průvodním znakem nástupu poruchy je velmi často prudká změna frekvenčního spektra signálů z procesu, např. posun dominantních frekvencí nebo vznik nových významných frekvencí navíc k existujícímu rozložení spektra. Zde patří mezi základní požadavky co nejpřesněji určit moment, kdy porucha vznikla.

Krátkodobá Fourierova transformace

Krátkodobá Fourierova transformace (STFT, Short Time FT) je vlastně klasická FT, ovšem původní (časový) signál je rozdělen na kratší intervaly, na kterých je ještě možné očekávat, že je stacionární. Toho je dosaženo přenásobením okénkovou funkcí, která zachová původní signál pouze do určité vzdálenosti kolem zvoleného časového okamžiku (v souvislosti s STFT je to obvykle jednoduchá obdélníková funkce s hodnotou 1 v blízkosti času t a 0 jinde). Protože okénko se posouvá společně se zkoumaným časovým bodem, tak i samotné spektrum, resp. amplitudy a počáteční fáze jednotlivých složek jsou funkcemi nejen frekvence, ale také času [57].

Šířka okénkové funkce podstatně ovlivňuje charakter výsledného spektra. Zvětšením šířky okénka se zlepší rozlišení ve frekvenční oblasti, ale prodlouží se interval, na kterém je signál považován za stacionární, resp. ergodický, a tím se zhorší rozlišení v oblasti časové. Tuto vzájemnou závislost postihuje Heisenbergův-Gáborův princip neurčitosti [46], [57] a platí obecně pro všechny metody založené na Fourierově transformaci.

Grafem STFT je spektrogram, který zobrazuje spektrum hustoty energie signálu v závislosti na čase. Na vodorovné ose spektrogramu je čas a na svislé ose frekvence. Hustota energie signálu pro určitou frekvenci a určitý čas je vyjádřena pomocí barevné škály. Frekvence je ve spektrogramu obvykle zobrazována v normovaném měřítku (hodnota 1 odpovídá vzorkovací frekvenci). Spektrum hustoty energie je popsáno vztahem

$$E(t,\omega) = \left|X_t(\omega)\right|^2 = \frac{1}{2\pi}\left|\int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-\mathrm{j}\omega\tau}d\tau\right|^2, \qquad (2.3.11)$$

kde $h(\cdot)$ je okénková funkce [57].

Vlnková transformace

Podobně jako Fourierova transformace, i vlnková transformace (WT, Wavelet transform) provádí rozklad signálu na frekvenční složky. Na rozdíl od FT však není vzorem pro rozklad harmonická funkce, ale takzvaná základní (mateřská) vlnka $\psi(\cdot)$, která musí splňovat jen poměrně velmi obecné podmínky [46]. WT je popsána vztahem (2.3.12),

$$W(a,b;x,\psi) = |a|^{-1/2} \cdot \int_{-\infty}^{\infty} x(t) \cdot \psi^* \left(\frac{t-b}{a} \right) dt \qquad (2.3.12)$$

kde parametr $1/a$ představuje frekvenční měřítko a parametr b časové posunutí (polohu) vlnky. Symbol $\psi^*(\cdot)$ značí komplexně sdruženou funkci k $\psi(\cdot)$.

Tvar vlnky je možné přizpůsobit specifickým požadavkům, stále však platí omezení jako u FT, že tvar funkce musí být předem zvolen a není tedy adaptivní. Ze stejného důvodu nebo v důsledku nelinearity je také možné, že rozklad signálu na složky nemusí postihnout fyzikální podstatu procesu. V praxi se pro mateřskou vlnku obvykle používá některá ze standartních dobře zdokumentovaných funkcí. Nejčastěji je to Morletova vlnka [46], tedy komplexní exponenciální funkce omezená Gaussovskou obálkou (typicky na 5,5 period).

Podobně jako je ve FT harmonická funkce modifikována změnou amplitudy, frekvence a počáteční fáze, tak i vlnka je při WT roztahována změnou parametru a a posouvána změnou parametru b. Výstupem WT je potom soustava ortonormálních složek původního signálu.

Z toho, že je vlnková funkce časově omezená, vyplývají dva důležité rozdíly mezi WT a FT: *a)* Jejím posouváním je možné se zaměřit pouze na vybraný krátký úsek zkoumaných dat. WT je proto už z principu metoda časo-frekvenční analýzy signálu na rozdíl od FT, která je primárně čistě frekvenční. *b)* Se změnou měřítka je svázána nejen šířka vlnky, ale i frekvence, takže počet jejích oscilací zůstává pořád stejný. (Oproti STFT, kde je šířka okénka vzhledem k frekvenci konstantní a mění se počet porovnávaných period) Proto je WT pro vyšší frekvence schopná dosáhnout lepšího rozlišení v časové oblasti za cenu horšího rozlišení v oblasti frekvenční. [46], [57]

Vztah (2.3.12) je možno interpretovat tak, že $W(a,b;x,\psi)$ představuje hustotu energie signálu na měřítku a a v čase b. Grafická reprezentace vlnkové transformace, kde na vodorovné ose je b a na svislé a, se nazývá škálogram. Je-li potřeba ho transformovat na spektrogram, stačí pouze převrátit svislé měřítko na hodnotu $1/a$, která již odpovídá frekvenci. Posunutí b přímo odpovídá pozici v čase t.

2.3.3 Okamžitá frekvence a amplituda

Analytický signál

Signál, který je nestacionární, resp. proměnlivý v čase, má proměnlivé také rozložení frekvenčního spektra. Proto by bylo užitečné znát jeho okamžitou amplitudu a frekvenci v každém časovém bodě. Okamžitá amplituda, resp. obálka kmitavého signálu je poměrně dobře akceptovatelná (stejně jako okamžitá energie signálu). Oproti tomu představa okamžité frekvence je problematická.

Klasická Fourierova transformace určuje frekvenci signálu jeho srovnáním s harmonickými funkcemi. Proto k rozpoznání momentální frekvence potřebuje alespoň jednu celou její periodu. Tento přístup je však pro nestacionární signály nepoužitelný, protože se u nich frekvence mění v každém okamžiku. Stejný problém se týká i dalších klasických metod časově frekvenční analýzy, kde je pouze vyhledáván jiný opakovaně se vyskytující průběh namísto harmonického.

Proto je třeba použít odlišného přístupu založeného na představě rotace fázoru v komplexní rovině, jak byla zavedena v kapitole 2.3.1. Pokud převedeme naměřený reálný signál na komplexní takovým způsobem, aby byly zachovány jeho frekvenční vlastnosti, můžeme určit jeho okamžitou fázi a amplitudu. Okamžitou úhlovou frekvenci potom získáme derivací okamžité fáze.

Vhodné vlastnosti má analytický signál, který vznikne použitím Hilbertovy transformace (viz [46], [54], [57]). Jak již bylo uvedeno, Fourierův obraz reálného signálu je komplexní funkce, jejíž amplituda je sudá a fáze lichá funkce frekvence. Z jejich symetrie vůči nulové frekvenci vyplývá, že odstraněním záporné části spektra (pro záporné frekvence) nedojde ke ztrátě informace, ale pouze ke změně měřítka. Jak je patrné ze vztahu (2.3.1) a z obr. 2.3.1, zpětnou Fourierovou transformací pouze kladné části spektra původního reálného signálu vznikne komplexní signál s poloviční amplitudou. Pro získání analytického signálu zachovávajícího velikost amplitudy reálné složky je proto nutné jeho spektrum vynásobit dvěma (viz [54]).

Hilbertova transformace

Analytický signál je možné vyjádřit v komplexních souřadnicích jako

$$z(t) = x(t) + \mathrm{j}\hat{x}(t) = a(t)e^{\mathrm{j}\varphi(t)}, \tag{2.3.13}$$

kde reálná složka $x(t)$ je původní naměřený signál a imaginární složka $\hat{x}(t)$ je jeho Hilbertova transformace

$$\hat{x}(t) = H\{x(t)\} = \frac{1}{\pi}\mathrm{P}\int_{-\infty}^{\infty}\frac{x(\tau)}{t-\tau}d\tau\,. \tag{2.3.14}$$

Symbol P označuje Cauchyho hlavní hodnotu integrálu. Hilbertova transformace (2.3.14) představuje konvoluci originálního signálu s funkcí $1/t$ a tím zdůrazňuje jeho lokální vlastnosti. Vyjádření analytického signálu (2.3.13) v polárních souřadnicích přímo ukazuje okamžitou amplitudu a okamžitou fázi

$$a(t) = \sqrt{x(t)^2 + \hat{x}(t)^2}\,,\ \varphi(t) = \arctan\left(\frac{\hat{x}(t)}{x(t)}\right). \tag{2.3.15}$$

Okamžitá frekvence, která se získá z okamžité fáze derivací

$$\omega(t) = \frac{\mathrm{d}\varphi(t)}{\mathrm{d}t}, \tag{2.3.16}$$

nabývá v každém okamžiku právě jedné hodnoty a dokáže tak interpretovat pouze tzv. monokomponentní signál [46], který obsahuje v každém okamžiku právě jednu frekvenční složku. Při pokusu použít výpočet na vícesložkový signál, jehož spektrum je tvořeno mnoha různými frekvencemi, je často výsledkem fyzikálně nesmyslná záporná

hodnota frekvence. Protože pojem monokomponentního signálu není přesně definován, uznává se úzkopásmový signál jako podmínka pro smysluplné použití okamžité frekvence. Přesnější definice úzkopásmového signálu viz např. [46], [55], [57].

I na jednosložkový (úzkopásmový) signál jsou však kladena další dodatečná omezení. Aby byla okamžitá amplituda a fáze (a v důsledku toho i okamžitá frekvence) správně interpretována, střední hodnota signálu musí být nulová.

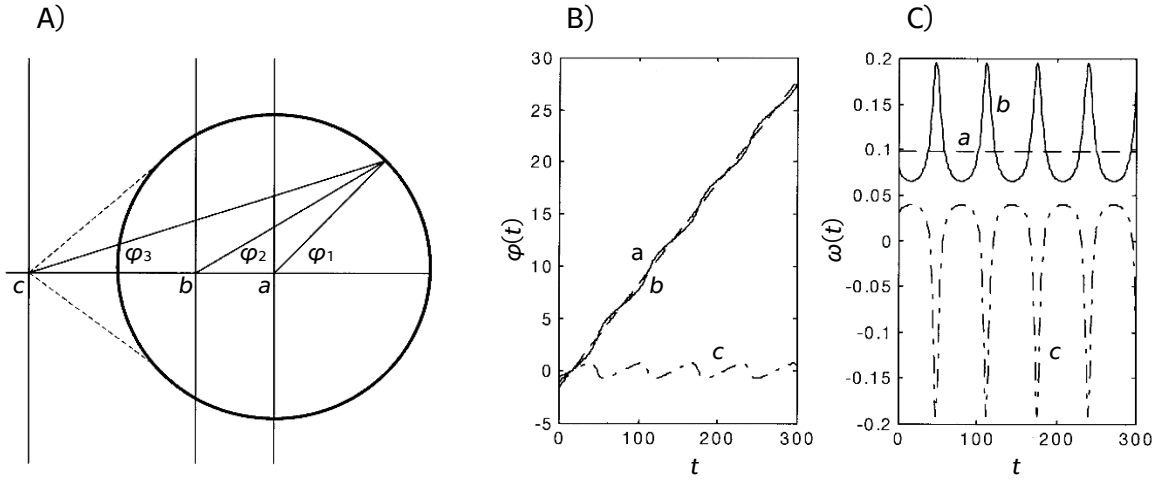
Problém nejlépe ilustruje obr. 2.3.3 na příkladu jednotkové harmonické funkce s posunutou střední hodnotou. Harmonická funkce $x(t)=\alpha+\sin(t)$ se zobrazí do fázové roviny, obr. 2.3.3 A), jako rotace konstantní frekvencí na jednotkové kružnici se středem ležícím na reálné ose posunutým o střední hodnotu α.

Obecně mohou nastat tři případy (v obr. 2.3.3 označeny malými písmeny *a*, *b*, *c*). Pouze v případě, že střední hodnota je nulová (na obrázku případ *a*), bude okamžitá fáze φ_1, která je vypočtena podle (2.3.15), odpovídat přesně skutečnosti. Fáze poroste lineárně, viz obr. 2.3.3 B), a okamžitá frekvence vypočtená dle (2.3.16) bude konstantní, viz obr. 2.3.3 C).

Bude-li střední hodnota nenulová, ale menší než velikost amplitudy (případ *b*), hodnota okamžité fáze φ_2 bude kolísat kolem skutečné hodnoty, obr. 2.3.3 B), ale stále poroste. Hodnota okamžité frekvence bude výrazně kolísat, ale stále zůstane kladná a její střední hodnota bude stále odpovídat skutečné frekvenci.

Pokud bude posunutí středu větší než velikost amplitudy (případ *c*), okamžitá fáze φ_3 přestane zcela odpovídat skutečnému průběhu. Nebude narůstat a jak okamžitá fáze, tak okamžitá frekvence začnou kolísat kolem nuly a nabývat tak i záporných hodnot, které neodpovídají fyzikální představě o frekvenci harmonického signálu.

Aby bylo možné získat smysluplné okamžité frekvence z jednoduchého signálu, musí být tedy signál lokálně symetrický kolem nulové střední hodnoty.



obr. 2.3.3 – Fyzikální interpretace okamžité frekvence (převzato z [46]).
A) Fázová rovina funkce x(t) = α+sin(t), a: α=0, b: 0<|α|<1, c: |α|>1. B) Rozbalená fáze modelové funkce. C) Okamžitá frekvence modelové funkce vypočtená podle (2.3.16).

Pro obecné složitější signály platí, že případu $\alpha>1$ odpovídají místa, kde mezi dvěma sousedními lokálními extrémy signál neprochází nulou a případu $\alpha<1$

odpovídají úseky, kde je průběh signálu lokálně nesymetrický kolem nulové střední hodnoty. [46], [57]

2.3.4 Empirická modální dekompozice

Hilbert-Huangova transformace

Hilbert-Huangova transformace (HHT) je metoda rozkladu signálu v časově frekvenční oblasti. Při aplikaci HHT se nejprve signál rozloží pomocí algoritmu empirické modální dekompozice (EMD) na složky reprezentující módy jeho kmitání a následně se na tyto složky aplikuje Hilbertova transformace. Výstupem HHT je potom sada okamžitých frekvencí a amplitud reprezentující vlastnosti signálu v určitém časovém bodě. HHT publikoval Huang se svými kolegy v roce 1998 jako novou metodu pro analýzu nelineárních a nestacionárních dat [46].

Empirická modální dekompozice

Empirická modální dekompozice (EMD, Empirical Mode Decomposition) je metoda rozkladu číselných řad na složky, které jsou díky svým vlastnostem vhodné pro převedení na analytický signál pomocí Hilbertovy transformace a pro následné určení okamžité frekvence a amplitudy.

Dílčí složky se nazývají vlastní modální funkce (IMF, Intrinsic Modal Function) a ve své podstatě reprezentují módy kmitání původního signálu. Původní off-line algoritmus vyvinul N. E. Huang pro NASA [46] jako nástroj na analýzu vlnění mořské hladiny [47].

Algoritmus EMD provádí rozklad naměřené časové posloupnosti $x(t)$ na sadu vlastních modálních funkcí $c_i(t)$, $i = 1, 2, \ldots, n$ a zbytkové reziduum $r(t)$:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r(t), \tag{2.3.17}$$

kde n je počet vlastních modálních funkcí. Reziduum $r(t)$ je monotónní funkce, která odráží průměrný trend signálu $x(t)$ na celém zkoumaném rozsahu.

Vlastní modální funkce je definována dvěma vlastnostmi (podmínkami) [46] vyvozenými v předchozí kapitole:

- Počet lokálních extrémů (minim a maxim) a počet průchodů signálu nulou musí být shodný nebo se může lišit maximálně o jeden. (Neboli mezi každými dvěma sousedními lokálními extrémy musí funkce protnout nulu.)
- Střední hodnota definovaná mezi obálkou lokálních maxim a obálkou lokálních minim musí být rovna nule v každém bodě funkce. (Neboli obálky lokálních extrémů musí být symetrické kolem nulové střední hodnoty signálu.)

První podmínka představuje obdobu požadavku na úzkou šířku pásma pro stacionární Gaussovský proces. Druhá podmínka zavádí lokální ochranu před nežádoucími fluktuacemi okamžité frekvence z důvodu nesymetrických výkyvů signálu. Ideální požadavek na nulovou střední hodnotu signálu v každém okamžiku není v praxi

realizovatelný, neboť pro nestacionární data by bylo nutné definovat lokální střední hodnotu na lokálním časovém intervalu, což je nesmysl. Proto je tato podmínka zmírněna pouze na symetrii obálek lokálních extrémů. Blíže o obálkách lokálních extrémů bude pojednáno později.

Algoritmus prosévání

Rozklad signálu na IMF provádí rekurzivní algoritmus, který autoři nazvali příhodně „Prosévání" (Sifting), jak je uvedeno např. v [49], [50], [A4].

Je postaven na následujících předpokladech:

1. Signál obsahuje alespoň dva lokální extrémy (jedno maximum a jedno minimum).
2. Charakteristické časové měřítko je definováno intervalem mezi extrémy.
3. Pokud data neobsahují žádné lokální extrémy, ale obsahují inflexní body, je možné extrémy získat derivací (i vícenásobnou). Konečný výsledek se získá opětovnou integrací jednotlivých složek.

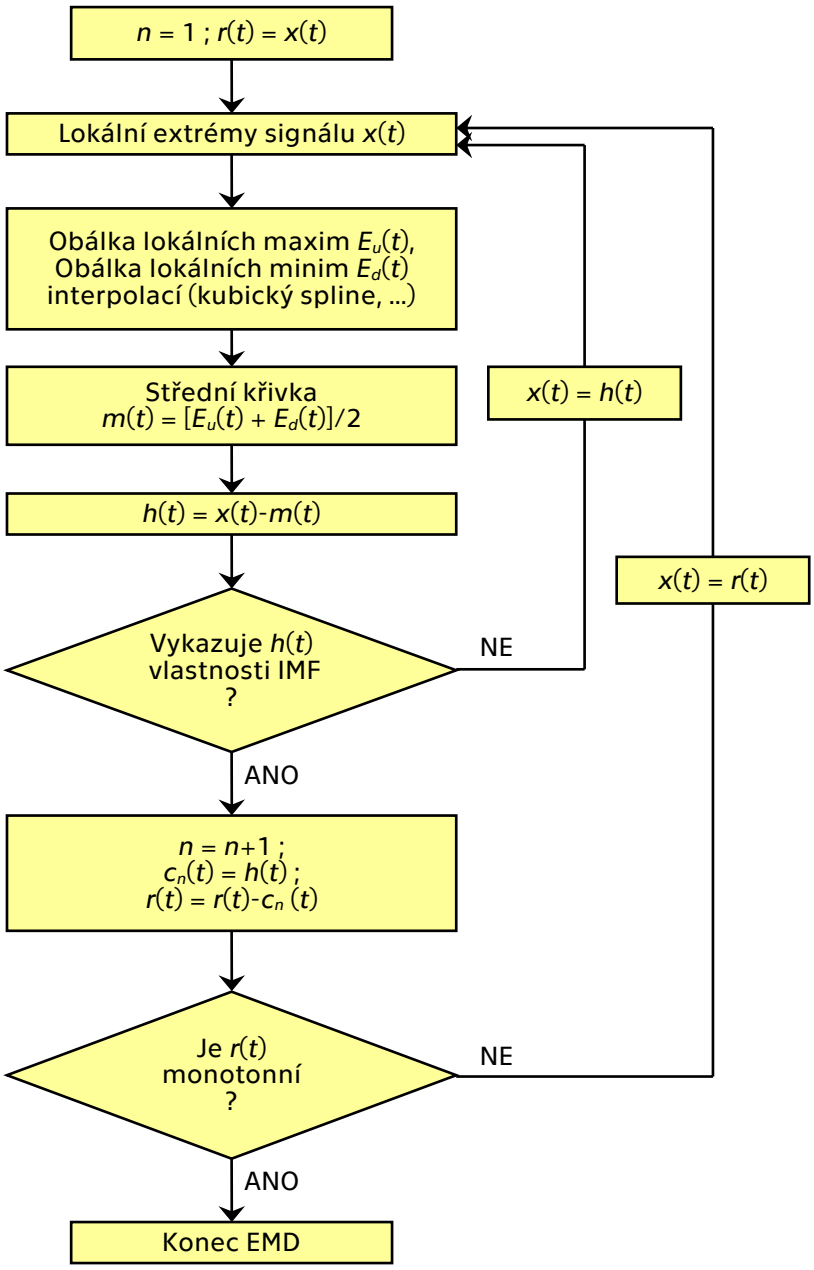
Algoritmus prosévání ukazuje vývojový diagram na obr. 2.3.4. Jedná se o dvouúrovňový iterační proces, který má za cíl rozdělit signál na jednotlivé IMF podle jejich charakteristického časového měřítka. (To se určuje na základě vzdáleností mezi extrémy a pro nestacionární signál ho lze chápat jako obdobu periody stacionárních periodických signálů).

Vnější iterační cyklus odděluje postupně z původního signálu jednotlivé komponenty (IMF) za vzniku průběžného rezidua a pokračuje tak dlouho, dokud není splněna jedna ze dvou ukončovacích podmínek:

- Další rozklad již není možný, protože zbytkové reziduum neobsahuje žádné extrémy (je tedy monotónní nebo konstantní).
- Reziduum nebo poslední nalezená IMF jsou již menší, než určuje předem stanovená hranice. Tato podmínka má zabránit vytváření „falešných" modů, které již jen kompenzují zaokrouhlovací chyby apod. V praxi se tato podmínka často doplňuje ještě omezením na maximální počet komponent.

Vnitřní iterační cyklus slouží k samotnému nalezení jednotlivé modální funkce IMF a tvoří samotné jádro algoritmu prosévání.

Zpracovává průběžné reziduum (získané z předchozí iterace vnějšího cyklu) postupným odečítáním jeho střední křivky, dokud výsledný průběh neodpovídá definici IMF. Střední křivka signálu (průběžného rezidua) se získá jako střední hodnota mezi obálkou lokálních maxim a obálkou lokálních minim. Protože obálky extrémů (v původním Huangově algoritmu vytvořené pomocí kubických spline funkcí) neodpovídají zcela přesně skutečným obálkám signálu, mohou vznikat vlivem jejich překmitů a podkmitů nové lokální extrémy. Z toho důvodu IMF nemusí vzniknout rovnou v prvním kroku, ale až po několika iteracích vnitřního cyklu.



obr. 2.3.4 – Vývojový diagram algoritmu Empirické modální dekompozice (EMD)

Aby zůstal zachován fyzikální význam IMF jako módu signálu (tj. aby nedošlo k přeiterování či zacyklení), zavádí se ještě dodatečné kritérium pro ukončení vnitřního cyklu. Huang ve [46] navrhuje jako kriterium ukončení hodnotu směrodatné odchylky mezi dvěma po sobě následujícími iteracemi odhadu IMF a jako typickou hodnotu uvádí $\sigma \leq 0{,}2 \div 0{,}3$.

Celý algoritmus prosévání probíhá v následujících hlavních krocích, kde kroky 2 až 5 reprezentují vnitřní iterační cyklus a kroky 6, 7 uzavírají vnější iterační cyklus:

1. Založit výchozí průběžné reziduum $r(t) = x(t)$ a výchozí odhad komponenty $h_1(t) = x(t); k = 1.$
2. Ověřit, jestli křivka $h_k(t)$ splňuje podmínky pro IMF (nebo je dosaženo ukončovacího kritéria). Pokud ano, přejít na krok 6. Jinak a pokračovat krokem 3.
3. Vytvořit obálku lokálních maxim $E_u(t)$ a obálku lokálních minim $E_d(t)$ vyšetřovaného signálu.
4. Vypočítat křivku středních hodnot mezi obálkami:
$$m(t) = \frac{E_u(t) + E_d(t)}{2} \tag{2.3.18}$$
5. Označit $k = k + 1$ a provést nový odhad modální funkce odečtením střední křivky od odhadu z předchozí iterace:
$$h_k(t) = h_{k-1}(t) - m(t). \tag{2.3.19}$$
6. Vytvořit IMF a přepočítat průběžné reziduum: $c_i(t) = h_k(t)$, $r(t) = r(t) - c_i(t)$.
7. Ověřit, jestli je vzniklé reziduum monotónní, případně zda je splněna jiná podmínka pro ukončení algoritmu. Pokud ano, ukončit algoritmus. V opačném případě nahradit $h_1(t) = r(t); k = 1$ a pokračovat opět od kroku 2.

Z průběhu algoritmu prosévání je patrné, že první nalezená IMF bude obsahovat kmitání o nejvyšších frekvencích zastoupených v signálu, které odpovídají nejkratším charakteristickým časovým měřítkům. Další IMF budou pokrývat postupně stále pomalejší frekvence.

Hilbertovo *spektrum*

Po aplikaci EMD a rozložení signálu na komponenty v souladu s (2.3.17) je již možné použít na každou IMF Hilbertovu transformaci (2.3.14), převést je na analytické signály podle (2.3.13) a vypočítat pro ně okamžitou amplitudu a okamžitou frekvenci.

Celkový signál převedený na analytickou funkci je pak možno vyjádřit ve formě

$$z(t) = \sum_{k=1}^{n} a_k(t) \cdot e^{\mathrm{j}\int \omega_k(t)dt}. \tag{2.3.20}$$

Přitom reziduum není do vztahu zahrnuto, protože jako monotónní funkce není z frekvenčního hlediska zajímavé a naopak představuje složku signálu, která se nehodí pro Hilbertovu transformaci.

Z vyjádření signálu ve formě (2.3.20) je již možné zobrazit Hilbertovo amplitudové spektrum $H(\omega,t)$, které vyjadřuje amplitudu jako funkci frekvence a času. Obdobně vyjádřením hustoty energie signálu ve formě druhých mocnin amplitud je možno získat Hilbertovo energetické spektrum. [46], [57]

Graficky je možné Hilbertovo spektrum vyjádřit podobně jako spektrogram STFT (viz kapitola 2.3.2).

2.3.5 EMD v reálném čase

Původní algoritmus EMD (viz [46]) je primárně navržen pro zpracování značně rozsáhlých, předem naměřených datových posloupností. Jejich vyhodnocování probíhá naráz až po skončení měření. V takové situaci nehraje významnou roli ani rychlost zpracování, ani kvalita dekompozice na začátku a na konci posloupnosti. Oproti tomu pro účely diagnostiky poruch, která je založena na on-line zpracování dat v reálném čase, jsou obě tyto vlastnosti důležité, pokud ne přímo kritické.

Problémy, se kterými se potýká původní metoda EMD, převážně souvisí s určením průběhu křivky středních hodnot signálu, resp. s konstrukcí obálek lokálních minim a maxim. Lze je rozdělit do dvou hlavních oblastí, které se však do určité míry překrývají.

První problém souvisí s překmity obálek lokálních extrémů a do značné míry je svázán s volbou kubických spline pro aproximaci obálek. Kubický spline má totiž tendenci se rozkmitávat a vytvářet tak mezi uzlovými body značné výkyvy. Stává se, že obálka překříží původní funkci – v takovém případě se hovoří o nekompletní obálce. To vede následně na „poškození" vytvořené IMF, na které se vytvoří nové lokální extrémy. Ty samozřejmě nereprezentují fyzikální podstatu procesu generujícího zkoumaný signál. Jsou jen projevem nekvalitní aproximace obálky signálu. Protože je celý algoritmus rekurzivní, každá deformace modální funkce se nepříznivě projeví také na všech následujících IMF. Za nepříznivých okolností mohou dokonce vzniknout výkyvy IMF výrazně větší, než jaké se vyskytují v původním signálu. Jejich jediným významem je přitom kompenzovat překmity nasbírané v předchozích IMF.

druhý problém se týká rozkladu signálu na jeho okrajích. Pokud je krajní lokální extrém příliš daleko od okraje datové řady, není možné tvar obálky na tomto okraji korektně určit a získaná IMF může být díky tomu zdeformovaná. I tento problém je nepříznivě ovlivněn použitím kubických spline. Ty nejsou příliš vhodné pro extrapolaci průběhu za krajním uzlovým bodem a mají tendenci zde výrazně zvětšovat svoji odchylku.

Oběma problémy se v posledních letech zabývalo mnoho autorů. Zajímavý souhrn modifikací a vylepšení původní metody je v [51]. Problém s okraji je možné řešit například ořízntutím dat v každém kroku až k nejbližšímu lokálnímu extrému. Tato metoda je však nepraktická pro krátké datové řady nebo v případě, že nás zajímá právě okraj rozsahu. Mnoho autorů navrhuje řešit uvedený problém nějakou formou aproximace (extrapolace) dat. Tímto způsobem jsem také zpočátku postupoval [A4], experimenty s reálnými daty však nevedly vždycky k dobrým výsledkům. Nutnou podmínkou tohoto řešení je totiž stacionarita sledovaného procesu, což je většinou v rozporu s aplikační oblastí (hlavní předností HHT je právě schopnost pracovat s nestacionárními a nelineárními daty a jeho adaptabilita).

Problém překmitů obálek lokálních extrémů je možné řešit změnou funkce popisující křivku obálky. Mezi úspěšnější řešení – alespoň při testování na simulovaných datech – patří nedávno publikovaná Improved EMD (IEMD) [53], kde problém okrajů byl částečně ošetřen přidáním pomocného krajního bodu získaného lineární extrapolací přes dva krajní lokální extrémy a problém překmitů obálek byl řešen nahrazením kubických spline zobecněnými Bézierovými křivkami NURBS (nonuniform rational B-spline).

On-line EMD v diagnostice poruch

Algoritmus HHT (resp. EMD) je velmi užitečný nástroj analýzy signálu v časově-frekvenční oblasti a v diagnostice poruch má velký potenciál, který dosud není plně využit. Stávající publikované metody diagnostiky poruch pomocí HHT jsou většinou zaměřeny na jeden úzce specifikovaný problém (viz např. [39]). Přímé využití kombinace markovského modelu s EMD patrně ještě nebylo publikováno. Nejblíže se podobnému využití přiblížila aplikace HHT a skrytých markovských modelů při analýze srdečních arytmií [60].

3 Cíle disertační práce

Motivace a obecný cíl

Diagnostika poruch má trvalý význam ve všech oblastech lidského konání. Diagnostika poruch pomocí markovských modelů se obecně řadí mezi pravděpodobnostní modelovací techniky založené na Bayesovském přístupu. Díky svým vlastnostem se Bayesovský klasifikátor využívající řízený markovský model jeví jako silný nástroj použitelný pro širokou škálu průmyslových i jiných aplikací. Proto byl zvolen při návrhu a realizaci systému diagnostiky poruch (viz [44], [45]), do jehož dalšího rozvoje jsem se zapojil na začátku svého doktorského studia. Během experimentů s diagnostickým systémem jsem zjistil, že v některých případech dochází k situaci, kdy není porucha správně rozpoznána. Diagnostický systém přítomnost poruchy buď vůbec nezaregistruje, nebo ji mylně vyhodnotí jako poruchu jinou.

Obecným cílem mé disertační práce je navržení postupů či metod, které povedou ke zvýšení úspěšnosti a obecně ke zlepšení detekce a identifikace poruch diagnostickým systémem založeným na markovském modelu.

Konkrétní cíle

Stochastický diskrétní model založený na markovském modelu je vhodný především pro rozpoznávání poruch, jejichž nástup je postupný a relativně velmi pomalý vzhledem ke zvolené vzorkovací periodě (viz [45]). V praxi však dochází velice často k situaci, že porucha nastane velmi rychle až zlomově. Například pokud je příčinou poruchy chybný zásah operátora, náhlé vniknutí nečistot do aparatury apod. Právě na tento typ poruch jsem se zaměřil ve své disertační práci.

Primárním cílem mé disertační práce je nalezení postupů nebo modifikace diagnostického systému, které povedou ke zlepšení schopnosti rozpoznat poruchy s rychlým, krátkodobým nástupem.

K dosažení primárního cíle je třeba splnit následující dílčí cíle:

1) Navrhnout metodu, jak modifikovat způsob, jakým diagnostický stochastický model založený na markovském řetězci interpretuje naměřená data z procesu s cílem eliminovat nežádoucí vliv rozdílných délek trénovacích množin pro různé provozní režimy. (Trénovací množina představuje konečnou posloupnost diskrétní v čase vstupních a výstupních veličin naměřených na technologickém procesu určenou k naučení statistik stochastického modelu na rozpoznání určitého provozního režimu. Délka trénovací množiny je celkový počet diskrétních okamžiků, na kterých byla data sebrána.)
2) Navrhnout metodu jak zajistit, aby diagnostický stochastický model založený na markovském řetězci nepotlačoval vliv přechodových dějů při změně provozního režimu, které jsou tak rychlé (krátkodobé), že jejich zastoupení v trénovací množině se jeví jako statisticky málo významné.
3) Navrhnout vlastní modifikaci algoritmu empirické modální dekompozice tak, aby byl použitelný pro výpočet vlastních modálních funkcí v reálném čase se zaměřením na využití v diagnostice poruch.
4) Experimentálně ověřit navržené postupy a metody.

4 Řešení cílů

V této části představím a zdůvodním postupy k dosažení cílů specifikovaných v kapitole 3. Následně doložím na vybraných reálných příkladech použitelnost navrženého řešení.

4.1 Metody řešení cílů

Pro účely detekce a lokalizace poruch (FDI) využívám stochastického modelu založeného na bayesovském přístupu s využitím markovského modelu, který byl navržen a implementován na Ústavu přístrojové a řídicí techniky, FS, ČVUT v Praze, viz Hofreiter, Garayaewa [44], [45]. Regresní vektor modelu je sestaven ze vstupních a výstupních veličin sledované soustavy a jeho výstupem je veličina, která klasifikuje poruchový stav soustavy [A10]. Princip uvedeného stochastického modelu jsem popsal v první polovině kapitoly 2. Tento model dokáže rozpoznávat především pomalu vznikající poruchy, které se vyznačují dlouhým plynulým přechodem z bezporuchového stavu.

V rámci řešení své disertační práce jsem navrhl modifikované postupy zpracování naměřených dat ze sledované soustavy. Modifikace vedou k zásadní změně interpretace dat stochastickým modelem, díky které jsem dosáhl značného urychlení a v některých případech i zpřesnění rozpoznání poruch, které se vyznačují prudkým až zlomovým nástupem s velice krátkým přechodovým dějem. Takové přechodové děje budu dále v textu označovat jako děje s rychlou dynamikou.

4.1.1 Dynamický model poruchových stavů

V této kapitole se budu zabývat řešením primárního cíle disertační práce, konkrétně body 1) a 2).

Poznámka k terminologii

V následujícím textu používám výrazy „stav“ a „poruchový stav“, které by neměly být zaměňovány. Výraz „stav“ bez přívlastků znamená obvykle úzce definovaný vnitřní stav procesu, který se vždy vztahuje právě k jedné konkrétní realizaci regresního vektoru (RV) a je tímto RV jednoznačně určen. Výraz „poruchový stav“ či „provozní stav“ se naopak vztahuje ke globálně definovanému provoznímu režimu procesu (může být také označen porucha, porucha 1, nominální stav apod.) a obecně pokrývá množinu mnoha hodnot regresního vektoru (a „stavů“), přičemž množiny RV různých poruch se mohou částečně překrývat. Diagnostickým systémem je odhadován „poruchový stav“ procesu.

Z kontextu by mělo být vždy jasné, o jaké použití slova stav se jedná.

Kategorizace stavů v markovském modelu FDI

Z hlediska diagnostiky poruch postačuje poměrně hrubé rozlišení stavů sledovaného systému. V nejjednodušší formě je to pouze dvoustavové rozlišení na „normální (bezporuchové) chování“ vs. „jiné chování“. Pro podrobnější určení je možné definovat kromě uvedených dvou základních stavů ještě řadu několika specifických poruchových stavů („Porucha 1“, „Porucha 2“, …). Kategorie „jiné chování“ potom převezme význam nespecifikovaného chování, které neodpovídá žádnému ze známých provozních či poruchových stavů.

Stavy sledovaného systému z hlediska diagnostiky poruch je tedy možné rozdělit do tří kategorií:

- bezporuchové chování,
- chování odpovídající některé ze známých poruch,
- chování jakékoliv jiné, které diagnostika v daném okamžiku neumí zařadit.

Bayesovský klasifikátor stavů

Předpokládejme reálný dynamický systém (technologický proces) popsatelný (diskrétním) řízeným markovským řetězcem *m*-tého řádu tak, jak byl popsán v kapitole 2.2.2, se vstupním vektorem $\mathbf{v}_k$ a výstupním vektorem $\mathbf{y}_k$, které jsou formalizovány podle vztahů (2.2.29) až (2.2.32)

$$\mathbf{v}_k = \left[v_k[1], v_k[2], \ldots, v_k[\mu]\right]^T \in \varphi_v = \varphi_{v[1]} \times \varphi_{v[2]} \times \ldots \times \varphi_{v[\mu]}, \tag{4.1.1}$$

$$\mathbf{y}_k = \left[y_k[1], y_k[2], \ldots, y_k[\eta]\right]^T \in \varphi_y = \varphi_{y[1]} \times \varphi_{y[2]} \times \ldots \times \varphi_{y[\eta]}, \tag{4.1.2}$$

kde φ_v, φ_y jsou množiny všech možných hodnot vektorů $\mathbf{v}_k$, $\mathbf{y}_k$ a μ, η jsou počty vstupních a výstupních veličin

$$v_k[j] \in \varphi_{v[j]} = \left\{1,2,\ldots,N_{v[j]}\right\},\ N_{v[j]} < \infty,\ j = 1,2,\ldots,\mu,\ k = k_0+1, k_0+2, \ldots, k_K \tag{4.1.3}$$

$$y_k[j] \in \varphi_{y[j]} = \left\{1,2,\ldots,N_{y[j]}\right\},\ N_{y[j]} < \infty,\ j = 1,2,\ldots,\eta,\ k = k_0+1, k_0+2, \ldots, k_K \tag{4.1.4}$$

Stejným způsobem zavedeme doplňující diskrétní veličinu reprezentující režim činnosti technologického procesu, kterou nazveme poruchový stav

$$f_k \in \varphi_f = \left\{0,1,2,\ldots,N_f - 1\right\},\ N_f < \infty,\ k = k_0+1, k_0+2, \ldots, k_K \tag{4.1.5}$$

kde φ_f je konečná množina všech známých poruchových stavů;
N_f je celkový počet všech známých poruchových stavů a každý poruchový stav je označen indexem z množiny přirozených čísel rozšířené o nulu. Přitom pořadí indexů poruchových stavů není důležité s výjimkou bezporuchového stavu, který bude mít pro větší přehlednost vždycky přiřazen index 0 (nula).

Zavedeme stochastický model založený na řízeném markovském řetězci podle výše uvedených předpokladů a nazveme ho bayesovským klasifikátorem poruchového stavu. Bayesovský klasifikátor vyjadřuje rozdělení podmíněné pravděpodobnosti jevu, že se proces v diskrétním čase k nachází v poruchovém stavu f_k za předpokladu, že je v tomto okamžiku pozorován regresní vektor $\mathbf{z}_k$

$$p\left(f_k \middle| D^k\right) = p\left(f_k \middle| \mathbf{z}_k\right) \text{ pro } k = k_0+1, k_0+2, \ldots, k_K, \tag{4.1.6}$$

kde $\mathbf{z}_k$ je základní regresní vektor a D^k celá minulá historie dat naměřených na procesu až do diskrétního času k, viz (2.2.4). Vztah (4.1.6) ukazuje, že poruchový stav je podmíněně nezávislý na celé minulé historii procesu, jestliže známe regresní vektor definovaný zde jako

$$\mathbf{z}_k = \left\{ D_{k-m}^k \right\}; m \geq 1, \tag{4.1.7}$$

kde $m \geq 1$ udává maximální hloubku ukládaných dat v základním regresním vektoru s obecnou strukturou

$$\mathbf{z}_k = \left[y_k[1] \ldots y_k[\eta], v_k[1] \ldots v_k[\mu], y_{k-1}[1] \ldots y_{k-1}[\eta], v_{k-1}[1] \ldots v_{k-1}[\mu], \ldots \right]^T. \tag{4.1.8}$$

Počet jeho prvků se rovná

$$\rho_z = (\eta + \mu) \cdot (m+1) \tag{4.1.9}$$

Základní regresní vektor v bayesovském klasifikátoru stavů tedy zahrnuje vybraný úsek známé naměřené historie technologického procesu podle (4.1.7) včetně posledního známého vstupu a také poslední známé odezvy procesu na tento vstup. Regresní vektor naopak neobsahuje minulé hodnoty odhadů poruchového stavu f^{k-1}, neboť tato veličina nepřináší žádnou jinou informaci o technologickém procesu než tu, která je již obsažena v pozorovaných datech D^k.

Podmíněné pravděpodobnosti (4.1.6) nemáme přímo k dispozici, můžeme ale vytvořit množinu hypotéz ${}_iH$, $i = 1,2,\ldots,r$ o struktuře regresního vektoru ${}_i\mathbf{z}_k$ a k nim příslušné odhady parametrů stochastického modelu, tedy rozdělení aposteriorních pravděpodobností v maticích ${}_i\mathbf{K}$. Takto neúplně určený model bude potom určen aposteriorními pravděpodobnostmi

$$p\left(f_k \middle|\, {}_i\mathbf{z}_k, {}_i\mathbf{K}, {}_iH\right) \text{ pro } k = k_0 + 1, k_0 + 2, \ldots, k_K. \tag{4.1.10}$$

Regresní vektor se strukturou podle *i*-té hypotézy bude opět tvořen výběrem ze základního vektoru podle vztahu (2.2.38)

$${}_i\mathbf{z}_k = {}_i\mathbf{J} \cdot \mathbf{z}_k, \tag{4.1.11}$$

kde ${}_i\mathbf{J}$ je opět výběrová matice podle i-té hypotézy. Matici ${}_i\mathbf{K}$ v bayesovském klasifikátoru nazveme maticí klasifikace poruchových stavů a vzhledem k (4.1.10) má význam

$${}_i\mathbf{K} = \left[{}_iK_{{}_i\zeta,\psi} = p\left(f_k = \psi \middle|\, {}_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right) \right] \tag{4.1.12}$$

$$\text{pro } {}_i\zeta \in \varphi_{{}_iz},\ \psi \in \varphi_f,\ k = k_0 + 1, k_0 + 2, \ldots, k_K$$

kde:

${}_i\zeta$ je index jednoznačně přiřazený konkrétní realizaci regresního vektoru ${}_i\mathbf{z}_k$ a určuje, na jakém řádku matice ${}_i\mathbf{K}$ se prvek ${}_iK_{{}_i\zeta,\psi}$ nachází;

ψ je index jednoznačně přiřazený konkrétní hodnotě poruchového stavu f_k a určuje, v jakém sloupci matice ${}_i\mathbf{K}$ se prvek ${}_iK_{{}_i\zeta,\psi}$ nachází;

$\varphi_{{}_iz} = \varphi_{{}_iz[1]} \times \varphi_{{}_iz[2]} \times \ldots \times \varphi_{{}_iz[\rho_{{}_iz}]}$ je konečná množina všech možných realizací (hodnot) regresního vektoru ${}_i\mathbf{z}_k$ o délce $\rho_{{}_iz}$ sestaveného podle hypotézy ${}_iH$;

$\rho_{\varphi_{i z}}$ nazveme mohutnost množiny $\varphi_{i z}$, tedy celkový počet všech možných hodnot regresního vektoru ${}_i\mathbf{z}_k$;

φ_f je konečná množina všech možných hodnot poruchového stavu f_k;

ρ_{φ_f} nazveme mohutnost množiny φ_f, tedy celkový počet všech možných poruchových stavů f_k;

Ze srovnání vztahů (2.2.35), (2.2.39) a (4.1.10), (4.1.12) je zřejmá formální podobnost obou modelů, i když význam jednotlivých náhodných veličin ve vzorcích je poněkud odlišný. Že tato podobnost není nahodilá si snadno doložíme úvahou, že na rozdělení (4.1.10), resp. (4.1.6), můžeme obecně pohlížet jako na model určený k predikci jedné výstupní veličiny na základě pozorování regresního vektoru složeného z ostatních výstupů a vstupů. Taková úvaha nijak neomezuje obecnost použitého řešení.

Další odvozování potřebných aposteriorních hustot pravděpodobnosti $p\left({}_i\mathbf{K}\middle|f^k, D^k, {}_iH\right)$, $p\left(f_k\middle|{}_i\mathbf{z}_k, {}_i\mathbf{K}, {}_iH\right)$, $p\left({}_iH\middle|f^k, D^k\right)$ a případně $p\left(f_k\middle|{}_i\mathbf{z}_k, D^k\right)$ by tudíž pokračovalo stejně jako v kapitolách 2.2.1 a 2.2.2 a proto ho nemusíme znovu detailně rozepisovat.

Odvození statistik uvedeného bayesovského klasifikátoru opět vede na určení matic absolutních četností ${}_i\mathbf{n}(k)$, $i = 1, 2, \ldots, r$ pro jednotlivé hypotézy ${}_iH$

$$ {}_i\mathbf{n}(k) = {}_i\mathbf{n}(k_0) + {}_i\mathbf{n}^1(k) \text{ , } i = 1,2,\ldots,r, \tag{4.1.13} $$

kde ${}_i\mathbf{n}(k_0)$, $i = 1,2,\ldots,r$ jsou matice, jejichž prvky ${}_i n_{i\zeta,\psi}(k_0)$ nabývají apriorně zvolených nezáporných hodnot a vyjadřují naši subjektivní míru důvěry, že z naměřených dat $D^k_{k_0+1}$ získáme pomocí hypotézy ${}_iH$ sdruženou dvojici $\left\{{}_i\mathbf{z}_\kappa = {}_i\zeta, f_\kappa = \psi\right\}$ a můžeme je interpretovat jako počet takových událostí ještě před zahájením identifikace, tedy v diskrétních časech $\kappa \le k_0$.

${}_i\mathbf{n}^1(k)$, $i = 1,2,\ldots,r$ jsou matice, jejichž prvky ${}_i n^1_{i\zeta,\psi}(k)$ představují objektivně zjištěný počet výskytů sdružených dvojic $\left\{{}_i\mathbf{z}_\kappa = {}_i\zeta, f_\kappa = \psi\right\}$ získaných z naměřených dat $D^k_{k_0+1}$ pomocí hypotézy ${}_iH$ pro diskrétní časy $k_0 < \kappa \le k$ a spočítáme je pomocí jednorázového vztahu

$$ {}_i n^1_{i\zeta,\psi}(k) = \sum_{\kappa=k_0+1}^{k} \delta\left({}_i\zeta, {}_i\mathbf{z}_\kappa\right) \cdot \delta\left(\psi, f_\kappa\right) \text{ pro } \psi \in \varphi_f \text{ a } {}_i\zeta \in \varphi_{i z}, \tag{4.1.14} $$

nebo rekurzivního vztahu

$$ {}_i n^1_{i\zeta,\psi}(k) = {}_i n^1_{i\zeta,\psi}(k-1) + \delta\left({}_i\zeta, {}_i\mathbf{z}_k\right) \cdot \delta\left(\psi, f_k\right) \tag{4.1.15} $$

pro $i = 1, 2, \ldots, r$, $k = k_0 + 1, k_0 + 2, \ldots, k_K$, $\psi \in \varphi_f$ a ${}_i\zeta \in \varphi_{i z}$

kde Kroneckerův delta operátor $\delta(\alpha, \beta)$ je definován vztahem (2.2.67).

Také zde platí, že pokud se parametry stochastické soustavy pomalu průběžně mění, je možno zahrnout do výpočtu statistiky (4.1.14), (4.1.15) ještě zapomínání (například metodou exponenciálního zapomínání) pro zmenšení vlivu starších dat.

Praktické aspekty realizace bayesovského klasifikátoru

Statistiky výše popsaného bayesovského klasifikátoru jsou sice konečných rozměrů, ale přesto velice rozsáhlé. Vzhledem k tomu, že model musí být schopen zpracovávat data v reálném čase (alespoň ve fázi diagnostiky), jeví se v uvedené podobě téměř nepoužitelné. Naštěstí existují poměrně jednoduché postupy, jak rozměrnost zpracovávaných matic podstatně omezit.

První výrazné odlehčení přinese, když vybereme pouze jednu hypotézu o struktuře regresního vektoru a namísto statistiky $p\left(f_k \middle| \mathbf{z}_k\right)$ budeme počítat pouze $p\left(f_k \middle|_i\mathbf{z}_k, {}_i\mathbf{K}, {}_iH\right)$ pro jednu hodnotu i. Strukturu můžeme zvolit apriorně úvahou na základě důkladného obeznámení se s technologickým procesem nebo můžeme navrhnout více hypotéz a provézt předběžné experimentální měření a využít vztahu pro výpočet $p\left({}_iH \middle| f^k, D^k\right)$ k posouzení nejvhodnější struktury regresního vektoru.

Pro další zjednodušení využijeme skutečnosti, že statistiky v matici ${}_i\mathbf{K}$ je možné počítat nezávisle po řádcích a využitím úvahy, že aposteriorní pravděpodobnosti ${}_iK_{{}_i\zeta,\psi} = p\left(f_k = \psi \middle|_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right)$ jsou nenulové pouze v případě, že se ve fázi učení objevila v trénovacích datech alespoň jednou příslušná dvojice konkrétních hodnot $\left\{f_k = \psi, {}_i\mathbf{z}_k = {}_i\zeta\right\}$.

Algoritmus pak reálně pracuje pouze s redukovanými maticemi ${}_i\mathbf{K}^*$, ${}_i\mathbf{n}^*\left(k\right)$ vzniklými z matic ${}_i\mathbf{K}$, ${}_i\mathbf{n}\left(k\right)$ vynecháním všech prázdných řádků a sloupců.

Matice ${}_i\mathbf{n}^*\left(k\right)$ (a tím i ${}_i\mathbf{K}^*$) jsou stále ještě velice „řídké“ – obsahují mnoho prázdných (nulových) prvků. K další redukci rozměrnosti a zefektivnění algoritmu můžeme použít metod naznačených v části Redukce rozměrnosti matice přechodu kapitoly 2.2. Ty také řeší situace, kdy není aktuální RV nalezen v naučené statistice. [45]

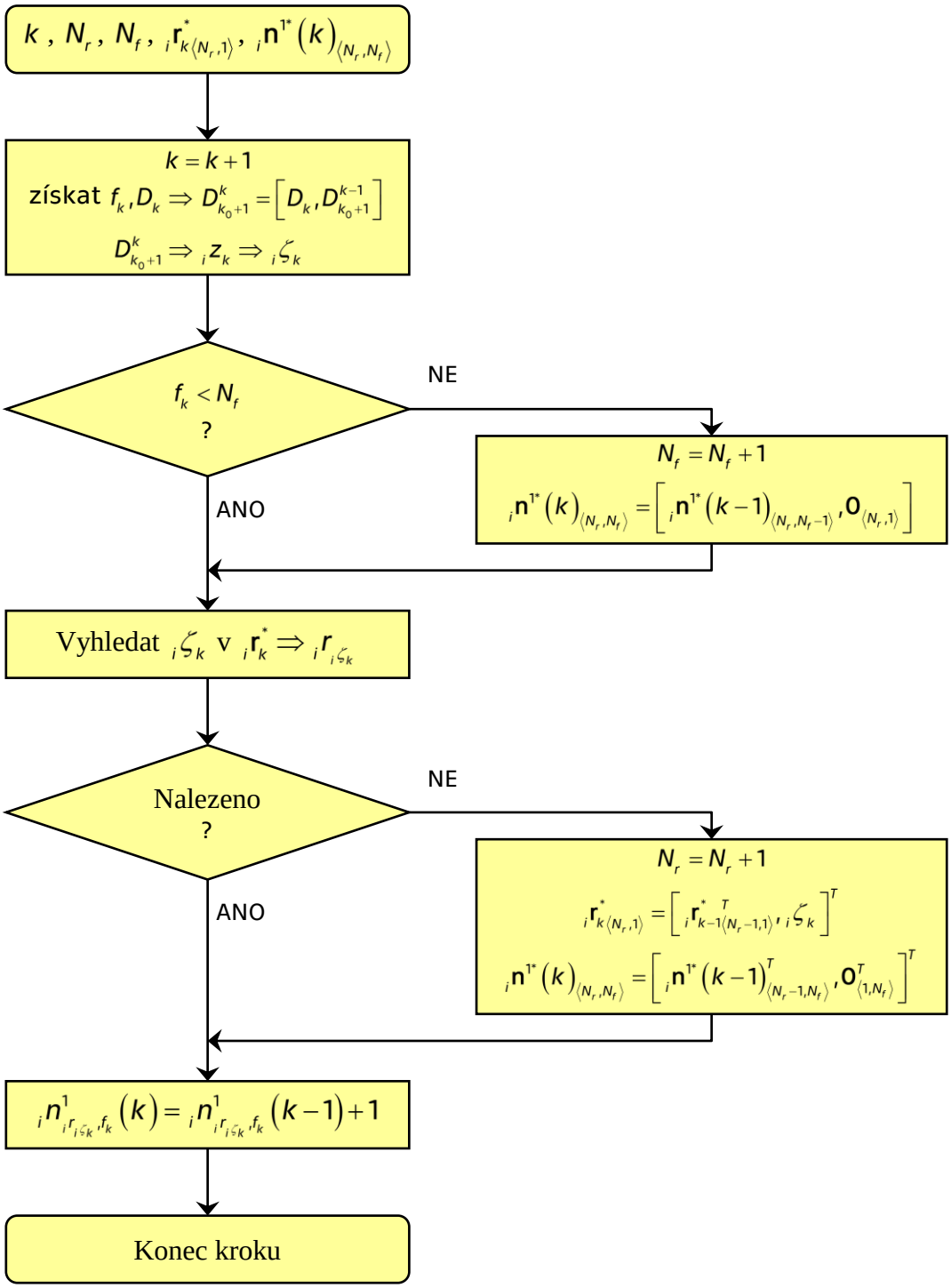
Fáze identifikace parametrů stochastického modelu probíhá podle algoritmu učení s učitelem na základě předběžně či průběžně získaných známých trénovacích dvojic $\left\{\mathbf{f}_{k_0+1}^{k_k}, D_{k_0+1}^{k_k}\right\}$. Pro trénovací množinu dat tedy předpokládáme znalost poruchového stavu (režimu), ve kterém se systém nacházel v diskrétních okamžicích $k = k_0+1, k_0+2, \ldots, k_k$.

Redukovaná matice četností ${}_i\mathbf{n}^{1*}\left(k\right)$,viz (4.1.13) , začíná jako prázdná matice (s 0 řádky a 0 sloupci). K ní přiřadíme redukovaný vektor všech známých indexů regresních vektorů (zpočátku také prázdný)

$$ {}_i\mathbf{r}_k^* = \left[{}_i\zeta\right]^T \text{ pro } {}_i\zeta \in \varphi_{{}_iz}{}^k, \tag{4.1.16}$$

kde $\varphi_{{}_iz}{}^k \subseteq \varphi_{{}_iz}$ je množina indexů všech realizací regresního vektoru ${}_i\mathbf{z}_k$, které byly pozorovány na trénovacích datech. Celočíselné indexy ${}_i\zeta$ nejsou regresním vektorům

přiřazeny náhodně, ale fungují jako kód, ze kterého je možné hodnotu regresního vektoru jednoznačně rekonstruovat. Možné způsoby kódování stavů viz např. [A10] nebo [45]. Každému řádku matice ${}_i\mathbf{n}^{1^*}(k)$ přísluší jeden řádek vektoru indexů ${}_i\mathbf{r}^*_k$.



obr. 4.1.1 – Jeden krok algoritmu identifikace parametrů bayesovského klasifikátoru poruch

Algoritmus identifikace probíhá v krocích, viz obr. 4.1.1, během kterých procházíme diskrétní časy $k = k_0 + 1, k_0 + 2, \ldots, k_k$ a postupně získáváme dvojice $\{f_k, {}_i z_k\}$.

V každém kroku nejprve ověříme, zda matice četností ${}_i\mathbf{n}^{1^*}(k)$ obsahuje sloupec přiřazený poruchovému stavu f_k. Pokud ne, jedná se o nový poruchový stav a je nutné rozšířit matici ${}_i\mathbf{n}^{1^*}(k)$ o jeden sloupec obsahující samé nuly a přiřadit tomuto sloupci číslo poruchy f_k. Dále z vektoru ${}_i z_k$ vygenerujeme index ${}_i\zeta_k$ a ten se pokusíme vyhledat ve vektoru indexů ${}_i\mathbf{r}_k^*$. Není-li index ${}_i\zeta_k$ nalezen, přidáme ho jako nový prvek do vektoru indexů ${}_i\mathbf{r}_k^*$ a matici četností ${}_i\mathbf{n}^1(k)$ rozšíříme na odpovídající pozici o řádek obsahující samé nulové prvky. Nakonec inkrementujeme prvek matice četností ${}_i\mathbf{n}^{1^*}(k)$ na souřadnicích daných hodnotami $\{f_k + 1, {}_i r_{i\zeta_k}\}$

$$ {}_i n^1_{{}_i r_{i\zeta_k}, f_k}(k) = {}_i n^1_{{}_i r_{i\zeta_k}, f_k}(k-1) + 1, \tag{4.1.17} $$

kde ${}_i r_{i\zeta_k}$ je číslo řádku ve vektoru ${}_i\mathbf{r}_k^*$, na kterém se nachází index regresního vektoru ${}_i\zeta_k$ a kde f_k je číslo poruchového stavu (pro jednoduchost budeme sloupce matice číslovat od nuly.

Uvedený postup opakujeme, dokud nedojdou trénovací data. Máme-li k dispozici ještě apriorní rozložení četností reprezentované maticí ${}_i\mathbf{n}(k_0)$, zahrneme je nejprve do redukované matice absolutních četností obdobným způsobem, kdy opět použijeme pouze řádky a sloupce, které obsahují nenulové hodnoty.

Výstupem fáze identifikace bude redukovaná matice četností ${}_i\mathbf{n}^*(k)$ s N_r řádky a N_f sloupci, viz (4.1.13), a k ní příslušný redukovaný vektor indexů ${}_i\mathbf{r}_k^*$.

Ve **fázi diagnostiky** získáváme průběžně z nově naměřených dat hodnoty regresního vektoru ${}_i\mathbf{z}_k = {}_i\zeta$ v diskrétních časech $k \geq k_k + 1$, přičemž ${}_i\mathbf{r}_k^*$ ani ${}_i\mathbf{n}^*(k)$ se již nemění. Model (4.1.10) generuje v každém diskrétním okamžiku k okamžitý odhad rozdělení aposteriorních pravděpodobností jednotlivých známých poruchových stavů ${}_i\hat{\mathbf{K}}^*_{{}_i\zeta}$, které přísluší aktuální hodnotě regresního vektoru ${}_i\mathbf{z}_k = {}_i\zeta$ podle vztahu

$$ {}_i\hat{K}_{{}_i\zeta,\psi} = \hat{p}\left(f_k = \psi \,\middle|\, {}_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_i H\right) = \frac{{}_i n_{{}_i\zeta,\psi}(k)}{\sum_{\psi_p=0}^{N_f} {}_i n_{{}_i\zeta,\psi_p}(k)} \text{ pro } \psi = 0,1,\ldots,N_f - 1 \tag{4.1.18} $$

Pro praktickou identifikaci je užitečnější sledovat trendy vývoje rozdělení pravděpodobností přes poněkud delší období. Nabízí se využití klouzavých průměrů metodou exponenciálního zapomínání

$$ {}_i\hat{K}_{{}_i\zeta,\psi}(k) = q \cdot {}_i\hat{K}_{{}_i\zeta,\psi} + (1-q) \cdot {}_i\hat{K}_{{}_i\zeta,\psi}(k-1);\ q \in \langle 0;1 \rangle;\ k = k_k + 1, k_k + 2, \ldots \tag{4.1.19} $$

kde volbou koeficientu zapomínání $q \in \langle 0;1 \rangle$ určujeme, jak konzervativní bude průběžný odhad rozdělení pravděpodobnosti ${}_i\hat{\mathbf{K}}_{{}_i\zeta}(k)$. Je zřejmé, že se zmenšujícím se q se bude

zmenšovat rozptyl, ale prodlouží se čas potřebný pro rozpoznání poruchy. Pro zde popisovaný diagnostický systém se osvědčila hodnota $q = 0{,}5$.

Dynamika přechodů mezi stavy soustavy

Bayesovský klasifikátor stavů (4.1.10), (4.1.12) předpokládá, že se chování sledovaného systému mění velmi zvolna, takže suficientní statistika obsažená v matici četnosti ${}_i\mathbf{n}(k)$ zahrnuje dostatečně obsáhlou informaci pro každý provozní režim a že případné přechodové děje jsou dostatečně významně zastoupeny v trénovací množině provozních dat. Pokud však nastane prudká změna provozního režimu vyznačující se rychlou dynamikou přechodového děje vzhledem k ustálenému chování procesu, může docházet ke zpomalení, zhoršení až ke znemožnění jeho rozpoznání.

Posloupnost hodnot regresního vektoru použitá jako zdroj pro generování statistik, na jejichž základě následně probíhá rozpoznávání určitého provozního či poruchového stavu, totiž v největší míře obsahuje převážně ustálené údaje, kdy jednotlivé parametry sledovaného systému setrvávají v relativně úzkém rozmezí hodnot s malou amplitudou výchylek, nebo případně vykazují ustálené periodické chování.

Stejná statistika však zahrnuje zároveň velmi významné, i když relativně velice krátké přechodové děje, ke kterým dochází při změnách z jednoho provozního režimu (poruchového stavu) na jiný. Statistiky přiřazené k jednomu stavu jsou tedy z hlediska jeho dynamiky poměrně silně nevyvážené, protože vedle relativně malého množství dat naučených z přechodového děje se vyskytuje velké množství dat získaných v ustáleném stavu nebo v jeho blízkosti.

Tvar a dynamika přechodových dějů jsou přitom pro správnou diagnostiku velmi důležité. Proto je nutné zajistit, aby se v rozhodovacím procesu dostatečně výrazně projevily.

Interpretace matice absolutních četností

Způsob, jakým generujeme statistiky v matici četností ${}_i\mathbf{n}(k)$ (resp. v redukované matici ${}_i\mathbf{n}^*(k)$) vede při bližším prozkoumání k důležitým poznatkům, které přímo vyplývají z vlastností statistik popsaných v předchozích kapitolách, přesto však nemusí být zcela zjevné.

Každý prvek (redukované) matice absolutních četností ${}_i\mathbf{n}^*(k)$ představuje počet výskytů dvojic $\{{}_i z_k = {}_i\zeta, f_k' = \psi\}$ v trénovací množině a můžeme snadno odvodit odhad rozdělení aposteriorní pravděpodobnosti

$$\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta, f_k = \psi \,\middle|\, {}_i\mathbf{K}, {}_iH\right) = \frac{{}_i n_{{}_i\zeta,\psi}(k)}{\sum_{\psi_p \in \varphi_f^*} \sum_{{}_i\zeta_p \in {}_i\mathbf{r}_k^*} {}_i n_{{}_i\zeta_p,\psi_p}(k)} \quad \text{pro } {}_i\zeta \in {}_i\mathbf{r}_k^*,\ \psi \in \varphi_f^* \qquad (4.1.20)$$

Tento odhad nemusí být pro diagnostiku poruch zcela vhodný, neboť nepotlačuje výběrovou chybu trénovacích dat.

Z postupu učení popsaného algoritmem na obr. 4.1.1 je zřejmé, že při objevení nové poruchy přidáváme do matice četností nový sloupec, do kterého naplníme četnosti výskytu jednotlivých regresních vektorů z příslušné nově získané trénovací množiny. Učení tedy probíhá po sloupcích a každý sloupec reprezentuje rozdělení aposteriorní pravděpodobnosti

$$\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle| f_k = \psi, {}_i\mathbf{K}, {}_iH\right) = I_\psi\left(k\right)^{-1} \cdot {}_i n_{{}_i\zeta,\psi}\left(k\right) \text{ pro } {}_i\zeta \in {}_i\mathbf{r}_k^*, \tag{4.1.21}$$

kde normovací konstanta $I_\psi\left(k\right)$ se vzhledem k vlastnostem pravděpodobnosti určí jako součet celého sloupce

$$I_\psi\left(k\right) = \sum_{{}_i\zeta_p \in {}_i\mathbf{r}_k^*} {}_i n_{{}_i\zeta_p,\psi}\left(k\right). \tag{4.1.22}$$

Potom odhad rozdělení (4.1.21) nabude tvaru

$$\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle| f_k = \psi, {}_i\mathbf{K}, {}_iH\right) = \frac{{}_i n_{{}_i\zeta,\psi}\left(k\right)}{I_\psi\left(k\right)} = \frac{{}_i n_{{}_i\zeta,\psi}\left(k\right)}{\sum_{{}_i\zeta_p \in {}_i\mathbf{r}_k^*} {}_i n_{{}_i\zeta_p,\psi}\left(k\right)} \text{ pro } {}_i\zeta \in {}_i\mathbf{r}_k^*. \tag{4.1.23}$$

Normovací konstanty I_ψ příslušné jednotlivým sloupcům matice četností ${}_i\mathbf{n}^*\left(k\right)$ zřejmě odrážejí rozdělení pravděpodobnosti výskytu jednotlivých poruchových stavů

$$\hat{p}\left(f_k = \psi \middle|_i\mathbf{K}, {}_iH\right) \propto I_\psi\left(k\right) \text{ pro } \psi \in \varphi_f^*, \tag{4.1.24}$$

pro které je normovací konstantou součet všech prvků matice ${}_i\mathbf{n}^*\left(k\right)$

$$\hat{p}\left(f_k = \psi \middle|_i\mathbf{K}, {}_iH\right) = \frac{I_\psi\left(k\right)}{\sum_{\psi_p \in \varphi_f^*} I_\psi\left(k\right)} = \frac{\sum_{{}_i\zeta_p \in {}_i\mathbf{r}_k^*} {}_i n_{{}_i\zeta_p,\psi}\left(k\right)}{\sum_{\psi_p \in \varphi_f^*} \sum_{{}_i\zeta_p \in {}_i\mathbf{r}_k^*} {}_i n_{{}_i\zeta_p,\psi_p}\left(k\right)} \text{ pro } \psi \in \varphi_f^* \tag{4.1.25}$$

Spojením (4.1.21), (4.1.24) dostaneme

$${}_i n_{{}_i\zeta,\psi}\left(k\right) \propto \hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle| f_k = \psi, {}_i\mathbf{K}, {}_iH\right) \cdot \hat{p}\left(f_k = \psi \middle|_i\mathbf{K}, {}_iH\right) \text{ pro } {}_i\zeta \in {}_i\mathbf{r}_k^*, \tag{4.1.26}$$

odkud jasně vyplývá, že do statistik vnášíme apriorní rozdělení pravděpodobnosti jako **důsledek velikostí trénovacích množin.**

Vliv relativních délek trénovacích množin na statistiky modelu je nežádoucí, protože nepřináší žádnou užitečnou informaci. Je jasné, že v praxi bude vždy k dispozici nejvíce dat z bezporuchového stavu, zatímco data z poruch budou podstatně omezenější v závislosti na době, po kterou ponecháme technologický proces v daném poruchovém stavu. Přirozeně při běžném provozu je snahou poruchu odstranit ihned, jakmile je odhalena. I v případě, že připravujeme trénovací data a poruchu vyvoláme v modelových podmínkách uměle, nemusí být možné či přípustné setrvat v daném režimu příliš dlouho. Přitom je jasné, že si nepřejeme omezovat činnost diagnostického systému umělým potlačováním pravděpodobnosti vzniku poruchy kvůli nedostatku dat na straně jedné a zbytečným ořezáváním bohatých trénovacích dat pro bezporuchový stav na straně druhé. Proto je vhodné vliv rozdělení (4.1.24) v procesu detekce poruch potlačit.

Podle Bayesovy formule (2.2.3) můžeme vztah mezi podmíněnými pravděpodobnostmi $p\left({}_i\mathbf{z}_k \middle| f_k\, {}_i\mathbf{K}, {}_iH\right)$ a $p\left(f_k \middle|_i\mathbf{z}_k, {}_i\mathbf{K}, {}_iH\right)$ vyjádřit jako

$$p\left(f_k \middle|_i\mathbf{z}_k, {}_i\mathbf{K}, {}_iH\right) = \frac{p\left({}_i\mathbf{z}_k \middle| f_k\, {}_i\mathbf{K}, {}_iH\right)\cdot p\left(f_k \middle|_i\mathbf{K}, {}_iH\right)}{\sum\limits_{\varphi_f^*} p\left({}_i\mathbf{z}_k \middle| f_k\, {}_i\mathbf{K}, {}_iH\right)\cdot p\left(f_k \middle|_i\mathbf{K}, {}_iH\right)} \propto p\left({}_i\mathbf{z}_k \middle| f_k\, {}_i\mathbf{K}, {}_iH\right)\cdot p\left(f_k \middle|_i\mathbf{K}, {}_iH\right) \quad (4.1.27)$$

To znamená, že nejen (4.1.20), ale také odhad (4.1.18) bude objektivní pouze v případě, že potlačíme vliv délek množin trénovacích dat jednotlivých provozních režimů (poruchových stavů) procesu.

Odhadu nezávislého na délkách trénovacích množin dosáhneme nahrazením odhadů rozdělení aposteriorních pravděpodobností (4.1.18) vztahem

$${}_i\hat{K}_{{}_i\zeta,\psi} = \frac{\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle| f_k = \psi, {}_i\mathbf{K}, {}_iH\right)}{\sum\limits_{\psi_p\in\varphi_f^*}\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle| f_k = \psi_p, {}_i\mathbf{K}, {}_iH\right)} = \frac{\dfrac{{}_in_{{}_i\zeta,\psi}(k)}{\sum\limits_{{}_i\zeta_p\in{}_i\mathbf{r}_k^*}{}_in_{{}_i\zeta_p,\psi}(k)}}{\sum\limits_{\psi_p\in\varphi_f^*}\dfrac{{}_in_{{}_i\zeta,\psi_p}(k)}{\sum\limits_{{}_i\zeta_p\in{}_i\mathbf{r}_k^*}{}_in_{{}_i\zeta_p,\psi_p}(k)}} \text{ pro } \psi\in\varphi_f^*. \quad (4.1.28)$$

kde podle bayesovského přístupu

$$\frac{\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle| f_k = \psi, {}_i\mathbf{K}, {}_iH\right)}{\sum\limits_{\psi_p\in\varphi_f^*}\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle| f_k = \psi_p, {}_i\mathbf{K}, {}_iH\right)} = \frac{\dfrac{\hat{p}\left(f_k = \psi_p, {}_i\mathbf{z}_k = {}_i\zeta \middle|_i\mathbf{K}, {}_iH\right)}{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{K}, {}_iH\right)}}{\sum\limits_{\psi_p\in\varphi_f^*}\dfrac{\hat{p}\left(f_k = \psi_p, {}_i\mathbf{z}_k = {}_i\zeta \middle|_i\mathbf{K}, {}_iH\right)}{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{K}, {}_iH\right)}}. \quad (4.1.29)$$

Tento výraz můžeme upravit aplikací pravidla násobení (2.2.1) a pravidla marginalizace (2.2.2) na tvar

$$\frac{\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle|_i\mathbf{K}, {}_iH\right)\cdot\dfrac{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right)}{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{K}, {}_iH\right)}}{\hat{p}\left({}_i\mathbf{z}_k = {}_i\zeta \middle|_i\mathbf{K}, {}_iH\right)\cdot\sum\limits_{\psi_p\in\varphi_f^*}\dfrac{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right)}{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{K}, {}_iH\right)}} \quad (4.1.30)$$

a po vykrácení a dosazení do vztahu (4.1.28) obdržíme

$${}_i\hat{K}_{{}_i\zeta,\psi} = \frac{\dfrac{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right)}{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{K}, {}_iH\right)}}{\sum\limits_{\psi_p\in\varphi_f^*}\dfrac{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right)}{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{K}, {}_iH\right)}} \propto \frac{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{z}_k = {}_i\zeta, {}_i\mathbf{K}, {}_iH\right)}{\hat{p}\left(f_k = \psi_p \middle|_i\mathbf{K}, {}_iH\right)} \text{ pro } \psi\in\varphi_f^*. \quad (4.1.31)$$

Když srovnáme vztah (4.4.31) se vztahem (4.1.27), je zřejmé, že úpravou podle (4.1.28) jsme dosáhli vyváženého odhadu rozdělení aposteriorních pravděpodobností známých poruchových stavů.

Pokud bychom potřebovali předepsat jiné než rovnoměrné rozdělení poruchových stavů, můžeme k tomu využít vyvážený vztah (4.1.28) nezávislý na délkách trénovacích množin

$$ {}_i\hat{K}_{{}_i\zeta,\psi} = p\left(f_k = \psi \middle| {}_i\mathbf{K}, {}_iH\right) \cdot \frac{\dfrac{{}_i n_{{}_i\zeta,\psi}(k)}{\sum_{{}_i\zeta_p \in {}_i\mathbf{r}_k^*} {}_i n_{{}_i\zeta_p,\psi}(k)}}{\sum_{\psi_p \in \varphi_f^*} \dfrac{{}_i n_{{}_i\zeta,\psi_p}(k)}{\sum_{{}_i\zeta_p \in {}_i\mathbf{r}_k^*} {}_i n_{{}_i\zeta_p,\psi_p}(k)}} \quad \text{pro } \psi \in \varphi_f^* . \tag{4.1.32} $$

kde $p\left(f_k \middle| {}_i\mathbf{K}, {}_iH\right)$ je námi zvolené apriorní rozdělení pravděpodobností výskytu poruchových stavů.

Klasifikační matice s prvky počítanými podle (4.1.32) představuje sufficientní statistiku bayesovského klasifikátoru poruch, která je nezávislá na délce trénovacích množin, zachovává významy pravděpodobností v souladu s definicemi v kapitole 2.2 a přitom nám poskytuje volnost ve volbě apriorního rozdělení pravděpodobností jednotlivých provozních režimů (poruchových stavů). Je zřejmé, že potom vztah (4.1.28), resp. (4.1.31) chápeme jako zvláštní případ rozdělení (4.1.32) pro rovnoměrné apriorní rozdělení poruchových stavů $p\left(f_k \middle| {}_i\mathbf{K}, {}_iH\right)$. [A16]

4.1.2 Transformace stavového prostoru

Pro lineární i nelineární dynamické časově invariantní procesy obvykle platí, že se většinu času nachází v nominálním bezporuchovém stavu. Přitom se sledované provozní parametry pohybují v blízkosti pracovního bodu a nevykazují od něj výraznější výkyvy.

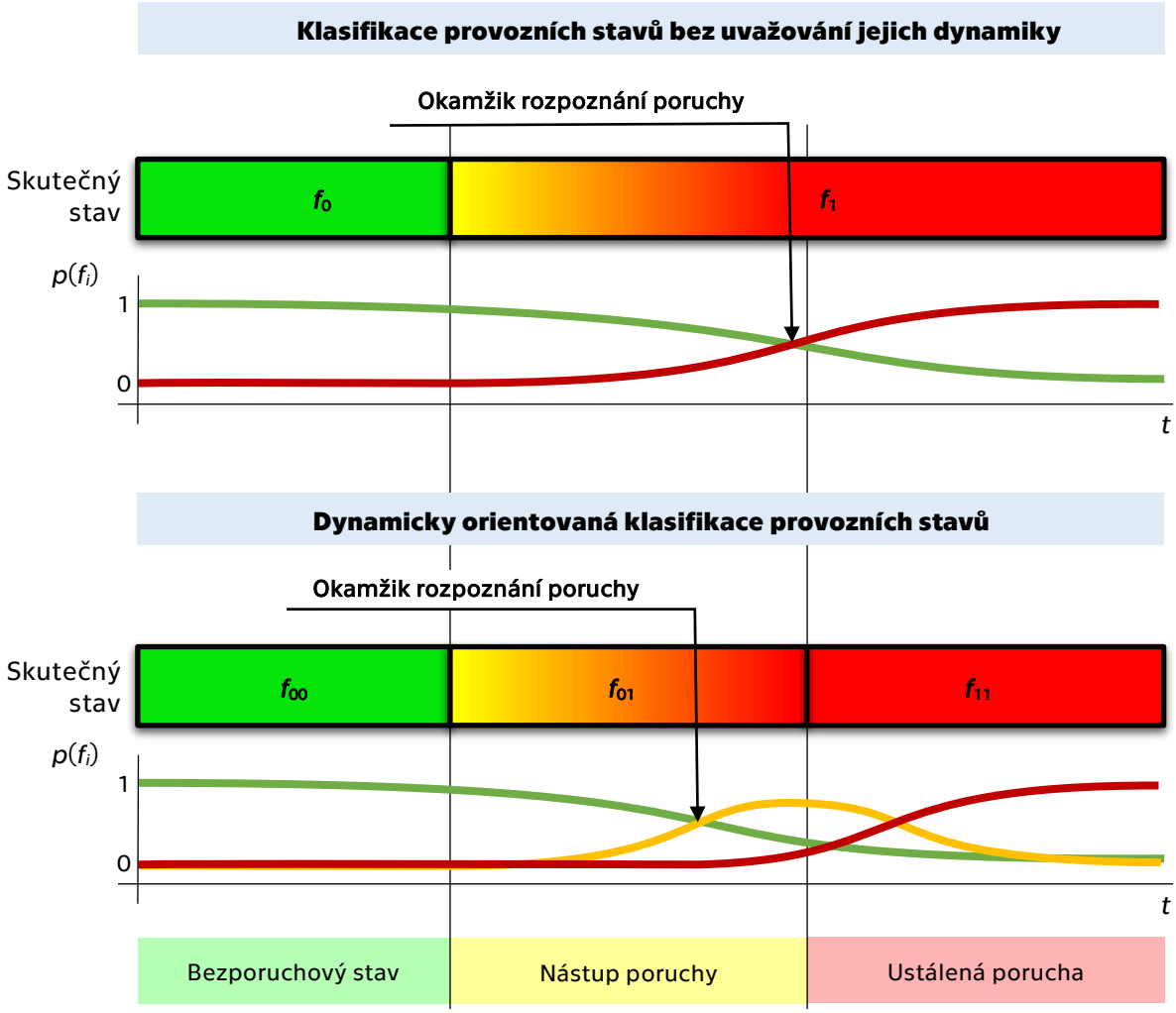
Dojde-li k poruše (kde poruchový stav nemusí být nutně havarijní, naopak často se také může jednat o přirozené stabilní chování soustavy, ovšem mimo požadovaný rámec kontrolovaných parametrů), nastane relativně velmi krátký, ale zároveň často poměrně výrazný přechodový děj, než se soustava opět ustálí v novém režimu. Přestože v poruše může být proces do určité míry destabilizován, obvyklejší je, že se pouze ustálí v jiném pracovním bodě. (Nemluvíme zde o kritických chybách. To už je funkce procesu výrazně narušena a rozpoznání poruchy není obtížné.). Přitom je jasné, že čím rychlejší je nástup poruchy (resp. čím strmější je gradient změny provozního/poruchového stavu), tím výraznější bude také přechodový děj.

Třídu poruch (provozních režimů) s rychlým nástupem, na které proces zareaguje rychlým (krátkodobým) přechodovým dějem, budu dále nazývat poruchy s rychlou dynamikou.

Přechodový děj má často průběh charakteristický pro určitou poruchu. V některých případech je dokonce jediným ukazatelem poruchy, neboť po jeho odeznění se hodnoty regresního vektoru vrátí téměř do stejného stavu, v jakém se nacházely před jejím vznikem.

Typickým příkladem může být technologický proces, jehož určitá část je vybavena vlastní zpětnovazebnou regulací, která pro svůj lokální charakter není či nemůže být sledována diagnostickým systémem. Po vzniku poruchy se chování procesu dočasně změní. Jakmile (pokud) se však vnitřní regulaci podaří vliv poruchy vykompenzovat, chování procesu jako celku se vrátí prakticky k nominálnímu chování, přestože porucha i nadále trvá. Charakteristickým projevem takového typu poruchy je mírné zakolísání statistik v diagnostickém systému, které nemusí být obsluhou zpozorováno, protože vyvolá varování buď krátkodobě, nebo vůbec.

Běžný mechanizmus (nejen) markovských systémů FDI přitom používá jako základní indikátor přítomnosti a/nebo typu poruchy právě to, že se sledované provozní parametry pohybují po dostatečně dlouhou dobu na určité kombinaci provozních rozsahů.



obr. 4.1.2- Princip dynamicky orientované klasifikace provozních a poruchových stavů.[A1]. Nahoře klasická definice stavu, dole dynamicky orientovaná. Barevný pás značí skutečný stav procesu, linie kumulativní odhad pravděpodobnosti generovaný systémem FDI. (zelená – bezporuchový stav, červená – poruchový stav, oranžová – přechodový stav)

V předchozí kapitole jsem ukázal, jak významně může být ovlivněn rozhodovací proces diagnostického systému v důsledku nevyváženosti délek trénovacích množin (posloupností trénovacích dat). Přechodový děj při nástupu poruchy s rychlou dynamikou je vzhledem k celkové délce trénovací množiny velmi krátký, takže četnost výskytu příslušných hodnot regresního vektoru je v rámci daného provozního stavu velmi malá a z hlediska stochastického klasifikátoru se falešně jeví jako statisticky málo významná. Zde se projevuje slabé místo pravděpodobnostních modelů, které nerozlišují, zda je malá četnost výskytu určitého regresoru důsledkem jeho malého významu, nebo důsledkem jeho nedostatečného zastoupení v trénovací množině z jiných příčin.

V tomto případě nepomůže ani použití vyváženého vyhodnocování provozních stavů podle (4.1.32), protože přechodový děj je součástí společné trénovací množiny s dlouhodobým ustáleným chováním dané poruchy.

Jako řešení problému dynamické nevyváženosti jsem použil **modifikovanou množinu provozních/poruchových stavů procesu**. Změna zachovává tvar regresního vektoru a rozsahy i rozlišení sledovaných veličin, avšak dojde k úpravě klasifikace provozních/poruchových stavů. Z hlediska markovského modelu FDI tedy změníme množinu hodnot výstupní veličiny a tím také klasifikační matici, avšak celý mechanizmus rozpoznávání zůstane zachován. [A9], [A2], [A1]

Každý provozní/poruchový stav rozdělíme na jednu stacionární část a jednu nebo více dynamických, přechodových částí.

Přechodový děj při nástupu poruchy (nebo obecně při změně mezi dvěma ustálenými stavy) tak bude markovským modelem interpretován jako nový samostatný poruchový stav. Ustálené chování soustavy v poruše po odeznění přechodového děje převezme formálně identitu původního nerozděleného poruchového stavu.

Při praktické realizaci můžeme ještě uvažovat předpoklad, že jednotlivé definované poruchy se navzájem nekombinují ani na sebe nenavazují. Je možné snadno očekávat, že porucha nastává vždy po delším období bezporuchového chování a že po odstranění příčin poruchy se soustava opět vrací do bezporuchového stavu.

Za uvedených podmínek se každý poruchový stav rozpadne přesně na dva stavy a bezporuchový stav se rozpadne na $n+1$ stavů, kde n je počet přechodových dějů z jednotlivých poruch zpět do bezporuchového stavu.

Princip metody je zobrazen schematicky na obr. 4.1.2. V horní části obrázku je situace se standartní klasifikací provozních stavů, v dolní části obrázku je naznačena nová, dynamicky orientovaná klasifikace provozních stavů.

Obě poloviny obrázku zachycují stejnou situaci. Barevný pás symbolizuje časový průběh skutečného stavu zkoumaného procesu a také symbolická označení provozních stavů, která jim byla přidělena během učení statistik. Jsou to: f_0: bezporuchový stav bez rozlišení dynamiky (zelené pole), f_1: poruchový stav bez rozlišení dynamiky (žluto-červené pole), f_{00}: bezporuchový ustálený stav (zelené pole), f_{01}: přechodový stav – nástup poruchy (žluto-červené pole), f_{11}: poruchový ustálený stav (červené pole).

Liniové grafy pod barevnými pásy představují schematicky kumulativní odhady pravděpodobností jednotlivých provozních stavů. Ty počítá diagnostický systém průběžně z okamžitých přechodových pravděpodobností (4.1.32) podle (4.1.19) (zelená linie – odhad pravděpodobnosti bezporuchového stavu, červená linie – odhad pravděpodobnosti ustálené poruchy, oranžová – odhad pravděpodobnosti přechodového stavu.). Součet kumulativních pravděpodobností všech známých stavů je v každém okamžiku roven jedné.

K rozpoznání poruchy dojde v okamžiku, kdy její kumulativní pravděpodobnost bude vyšší, než pravděpodobnost bezporuchového stavu.

Z principu metody vyplývá, že nejvýraznějšího zlepšení oproti původní metodě můžeme dosáhnout v případě skokového nástupu poruchy. Čím bude nástup poruchy pomalejší, tím více se průběhy kumulativních pravděpodobností přechodových a stacionárních stavů budou přibližovat, až nakonec prakticky splynou. Princip metody

však zaručuje, že výsledek rozpoznávání nemůže být nikdy horší, než při použití klasické klasifikace provozních stavů.

Rozdělením poruchového stavu na přechodový a ustálený se jejich statistiky oddělí. Relativní míra očekávání (4.1.32) realizací regresního vektoru, které souvisejí s přechodovým dějem poruchy, bude v samostatně definovaném poruchovém provozním stavu výrazně větší. Tím vzroste i míra očekávání přechodového stavu jako celku, kterou můžeme přímo ovlivnit volbou apriorního rozdělení $p\left(f_k \left| {}_i\mathbf{K}, {}_iH\right.\right)$.

Když potom dojde k dané poruše, nárůst odhadu pravděpodobnosti příslušného přechodového stavu bude výraznější. Vysoká pravděpodobnost přechodového stavu bude trvat pouze relativně krátce, ale mezitím již začne postupně narůstat pravděpodobnost ustáleného poruchového stavu. V důsledku vyššího počtu poruchových stavů s nenulovou pravděpodobností se navíc během přechodového děje poněkud sníží pravděpodobnost bezporuchového stavu, což rozpoznání poruchy dále urychlí.

Rozšířená logika

Vhodná volba struktury stavového slova (resp. regresního vektoru) a rozložení oborů hodnot měřených vstupních a výstupních veličin tvoří významnou apriorní informaci o statických a dynamických vlastnostech sledovaného procesu. Na vhodně zvolené struktuře závisí kvalita a úspěšnost systému detekce a lokalizace poruch.

Může být však výhodné zohlednit i další doplňkové apriorní znalosti, které o procesu máme. Tyto informace nemusí přímo souviset pouze s dynamikou procesu. Spadají sem především strukturální informace, které není dost dobře možné vyjádřit matematickým aparátem nebo rozumně začlenit do modelu dynamiky (do struktury regresního vektoru, do apriorního počátečního rozložení pravděpodobností v matici přechodu pod.).

Významnou skupinu takových podmínek tvoří informace obsažené ve stavovém diagramu procesu.

Příkladem může být znalost o nemožnosti nebo velmi nízké pravděpodobnosti přechodu mezi některými provozními či poruchovými stavy, případně naopak znalost, že po některém stavu může nastat pouze jediný jiný stav apod. Obecně lze říci, že se jedná o jistou omezenou znalost pravděpodobností přechodů mezi jednotlivými poruchovými stavy soustavy, která je však spíše kvalitativní než kvantitativní.

Pro podobné účely jsem sestavil jednoduchý modul rozšířené logiky. Ve své podstatě se jedná o primitivní pravidlově řízený systém. Tento nástroj nemá sám o sobě sloužit k diagnostikování poruch, pomáhá však rozhodnout v případech, kdy primární výsledky systému diagnostiky poruch nejsou dostatečně průkazné.

Hlavní motivací pro zavedení modulu rozšířené logiky do systému diagnostiky poruch byla potřeba sdružovat související poruchové stavy a potřeba zajistit logickou návaznost některých poruchových stavů.

Sdružování stavů

Umožňuje zastřešit více dílčích provozních stavů pod jeden nadřazený společný stav.

V případě realizace dynamicky orientovaného přístupu FDI umožní tento mechanismus logicky provázat přechodový a ustálený poruchový stav, takže výstupem diagnostického systému bude opět jediný zastřešující poruchový stav pro každý typ poruchy. Protože je však vazba mezi přechodovým a ustáleným stavem pouze kvalitativní a ne kvantitativní, zachová se zesílený význam přechodového děje.

Mechanismus definující sdružování stavů můžeme chápat jako zvláštní případ metody vázání stavů, která je popsána v následující podkapitole.

Vázání stavů

Z vlastností dynamického procesu je většinou možné snadno odvodit jednoduchá pravidla, podle kterých lze výrazně omezit množství reálně možných přechodů mezi jeho jednotlivými stavy. To platí i v případě souboru poruchových/provozních stavů.

Předpokládejme například ustálené stavy A, B, C a přechodové stavy A→B, B→A. Potom je zřejmé, že např. sekvence A, A→B, B nebo A, A→B, B→A může nastat s vysokou mírou očekávání. Oproti tomu např. posloupnost A, A→B, C je téměř nemožná, neboť nelze očekávat, že na přechod z A do B naváže rovnou ustálený stav C bez předchozího přechodu A→C nebo B→C [A9], [A2].

Apriorně nastavená pravidla, která definují míry pravděpodobností přechodů mezi známými poruchovými stavy, neslouží přímo k odhadu aktuálního poruchového stavu procesu. Pomáhají pouze rozhodnout v případě nejednoznačného výsledku odhadu z markovského modelu, kdy pravděpodobnosti několika poruch jsou na podobné úrovni.

Tento princip je velmi důležitý, protože není žádoucí zasahovat do nezávislého vyhodnocování statistik bayesovského klasifikátoru doplňujícími podmínkami, které jsou pouze důsledkem odhadů v předchozích diskrétních krocích.

Velmi výhodně můžeme uplatnit pravidla vázání stavů vytvořením speciálního jednoduchého markovského modelu s předdefinovanou pevnou maticí přechodu. Výstupem jsou stejně jako v hlavním modelu míry pravděpodobnosti jednotlivých poruchových stavů, vstupní regresní vektor však netvoří veličiny měřené ze soustavy, ale přímo jeden či více předchozích odhadů poruchových stavů. V nejjednodušším případě je pak možno takto definovanou matici přechodu naplnit pouze dvoustavovými hodnotami $p\left(f_k \middle| f_{k-1}\right) \in \{0;1\}$ určujícími, zda lze či nelze příslušný přechod očekávat.

Myšlenku dynamického přístupu v kategorizaci stavů a její implementace do algoritmu diagnostiky poruch (FDI) společně s metodou vázání stavů jsem publikoval v souhrnném článku [A1]. Experimentům s upraveným diagnostickým systémem na reálné soustavě se budu věnovat v kapitole 4.2. Jejich souhrn jsem nastínil také v [A9].

Přínosem metody je zejména zrychlení rozpoznání zlomově vznikajících poruch.

4.1.3 Návrh struktury regresního vektoru s využitím EMD

V předchozích kapitolách jsem několikrát zdůraznil význam vhodné volby struktury regresního vektoru pro správnou funkci markovského modelu a potažmo také celého systému FDI.

Obecná struktura regresního vektoru má tvar viz (4.1.8), kde pro řízený markovský model m-tého řádu jsou teoreticky v diskrétním čase k k dispozici minulé hodnoty vstupních a výstupních veličin až do hloubky $k-m$. V praxi však není účelné a ani dost dobře možné zahrnout do RV opravdu všechny dostupné vzorky. Proto sestavujeme množinu hypotéz s cílem vybrat co nejefektivnější strukturu regresního vektoru, která bude pouze natolik složitá, aby bezpečně zařadila aktuální stav procesu. Jak napovídá už název, regresní vektor plní do určité míry roli regresoru v ARX modelu [41], [44], [A9]. Volbou toho, které veličiny a s jakým dopravním zpožděním (o kolik kroků posunuté v diskrétním čase) zahrneme do regresního vektoru a také tím, které naopak vynecháme, přímo určujeme, jaké vlastnosti dynamického systému zdůrazníme a jaké potlačíme. Struktura regresního vektoru tak představuje významnou apriorní informaci o dynamických vlastnostech zkoumaného procesu. Úkol nalézt vyhovující regresní vektor je tedy značně obtížný. Proto patří mezi poměrně běžné úlohy srovnání více hypotéz o struktuře modelu.

RV se snažíme sestavit tak, aby umožnil zřetelně rozlišit všechny známé provozní a poruchové stavy procesu. Během provozu procesu (a systému FDI) se však průběžně objevují a doučují nové poruchy, které ještě v době návrhu nebyly známé. Může nastat situace, že stávající kombinace měřených veličin neumožňuje spolehlivě novou poruchu rozpoznat. Jedním řešením by bylo osadit nové senzory a regresní vektor modifikovat. To je však nákladné a představuje i technické komplikace.

Poněkud odlišným řešením může být pokus použít sice pouze stávající dostupná data, ovšem snažit se z nich získat nějakou dodatečnou, doposud ukrytou informaci. Poměrně běžné je například využití diskrétní varianty aproximace derivace a sestavení regresního vektoru z kombinace původních a derivovaných (diferencovaných) signálů.

S problematikou struktury regresního vektoru souvisí také poměrně slabá zobecňovací schopnost markovského modelu. Markovský diagnostický systém je schopen velmi dobře rozlišit stavy, se kterými se již dříve setkal nebo stavy jim velmi blízké, avšak schopnost správně zařadit stavy méně podobné je omezená.

Ve snaze o nalezení skrytých užitečných informací v měřených signálech a o vylepšení zobecňovací schopnosti markovského modelu jsem se začal zabývat rozkladem měřeného signálu na jednodušší složky takové, aby poskytly vhodnější stavové proměnné, které by lépe postihovaly charakteristické vlastnosti zkoumaného procesu.

Jako velmi perspektivní se mi jeví Hilbert-Huangova transformace (HHT), metoda rozkladu signálu v časově frekvenční oblasti [46]. Zevrubně jsem HHT pojednal v kapitole 2.3. Při aplikaci HHT se nejprve signál rozloží pomocí algoritmu empirické modální dekompozice (EMD) na složky reprezentující módy jeho kmitání, tzv. vlastní modální funkce, a následně se na tyto složky aplikuje Hilbertova transformace. Výstupem HHT je potom sada okamžitých frekvencí a amplitud reprezentující vlastnosti signálu v určitém časovém bodě. Určitou nevýhodou HHT pro zpracování v reálném čase je její poměrně značná výpočetní náročnost. Tu je však možné poněkud zmírnit.

Markovský model, na kterém je založen bayesovský klasifikátor poruch, využívá kategorizace měřených veličin ve formě celočíselých indexů, viz (4.1.3), (4.1.4), (4.1.5). Obdobné kvantování přirozeně použijeme také pro veličiny generované pomocí HHT.

Potom vzhledem k definici HHT není pro regresní vektor výrazný rozdíl mezi průběhem IMF a příslušným průběhem okamžitých amplitud.

Proto je nasnadě, že při hledání skrytých kvalit v signálu bude výhodnější, pokud si vystačíme pouze s první částí HHT – empirickou modální dekompozicí – a k sestavení regresního vektoru použijeme přímo vlastních módů signálu.

On-line empirická modální dekompozice

Pro účely detekce poruch je nutné použít on-line variantu dekompozice, která by byla vhodná pro zpracování signálu v reálném čase. V počáteční fázi práce s EMD jsem proto navrhl jednoduchou on-line modifikaci založenou na použití plovoucího časového okna o pevné nebo průběžně rostoucí délce [A4], [A6], [A8].

Plovoucí okno s pevnou šířkou

Původní Huangův algoritmus EMD zpracovává celý zaznamenaný signál naráz po skončení měření. To má za následek velkou výpočetní náročnost, která ale při off-line zpracování nepředstavuje zásadní problém, neboť zde není kladeno tak velké omezení na čas výpočtu. Navíc je možné konce intervalu, kde dochází ke zkreslení, vynechat po provedení EMD z analýzy. Avšak právě tyto dva problémy se stávají závažnými v případě, kdy chceme EMD algoritmus využít pro on-line analýzu.

Pro překonání uvedených problémů jsem použil následující postup.

Pro potřeby on-line zpracování signálu definujeme plovoucí časové okno na intervalu

$$\tau \in \left\langle \left(t - T_w\right); t \right\rangle, \tag{4.1.32}$$

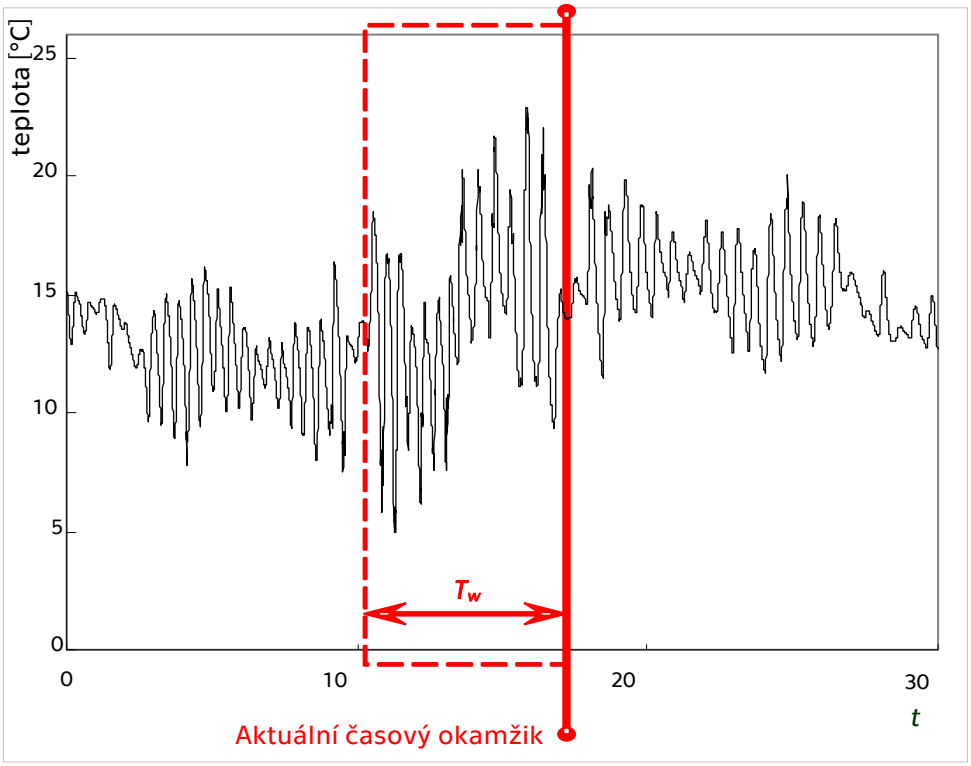
kde t je aktuální okamžik měření a T_w je velikost časového okna. EMD rozklad probíhá pouze na tomto intervalu. Protože v průběhu měření neroste počet zpracovávaných vzorků, nenarůstá ani výpočetní náročnost algoritmu. Modifikovaný algoritmus a jeho slabiny jsem popsal v [A5], [A7].

EMD rozkládá signál na vlastní modální funkce IMF a zbytkové reziduum. Superpozicí modálních funkcí a rezidua můžeme beze zbytku rekonstruovat původní signál. V průběhu rozkladu dochází ke značným deformacím modálních funkcí v blízkosti jejich okrajů v důsledku ne příliš dobré schopnosti extrapolace kubickými spline. Rekonstruovaného signálu se tyto deformace nedotknou, neboť se superpozicí vzájemně vyruší, ale okrajové deformace dílčích IMF mohou výrazně překročit průměrnou amplitudu nezkreslené části průběhu.

V off-line aplikacích metody EMD nehrají tyto deformace významnou roli, protože vyhodnocujeme vždy celý naměřený průběh za delší časové období a deformované okraje, v rámci celého rozsahu relativně úzké, po provedení dekompozice prostě odřízneme [A3]. Ovšem při on-line analýze v úlohách predikce, kdy postupně vyhodnocujeme data v plovoucím časovém okně, má největší význam právě tento nejvíce zdeformovaný konec sledovaného průběhu.

Z důvodu omezení dopadů rozkladu na konci intervalu měření signálu jsou z aktuálně známých průběhů vlastních modálních funkcí predikovány jejich budoucí průběhy, a odhadnuty i budoucí lokální extrémy využívané pro EMD algoritmus.

Uvedená modifikace EMD algoritmu umožňuje získávat vlastní modální funkce v reálném čase, což jsem ověřil na analýze environmentálních údajů a analýze stavu ekosystému v reálném čase (viz [A6]). Postup a výsledky uvedeného experimentu představím v kapitole 4.2.



obr. 4.1.3- Symbolická reprezentace plovoucího okna pro on-line EMD

4.2 Ověření řešení cílů

V následující části představím dva experimenty, pomocí kterých jsem prakticky ověřoval na reálných procesech funkčnost řešení popraných v kapitole 4.1.

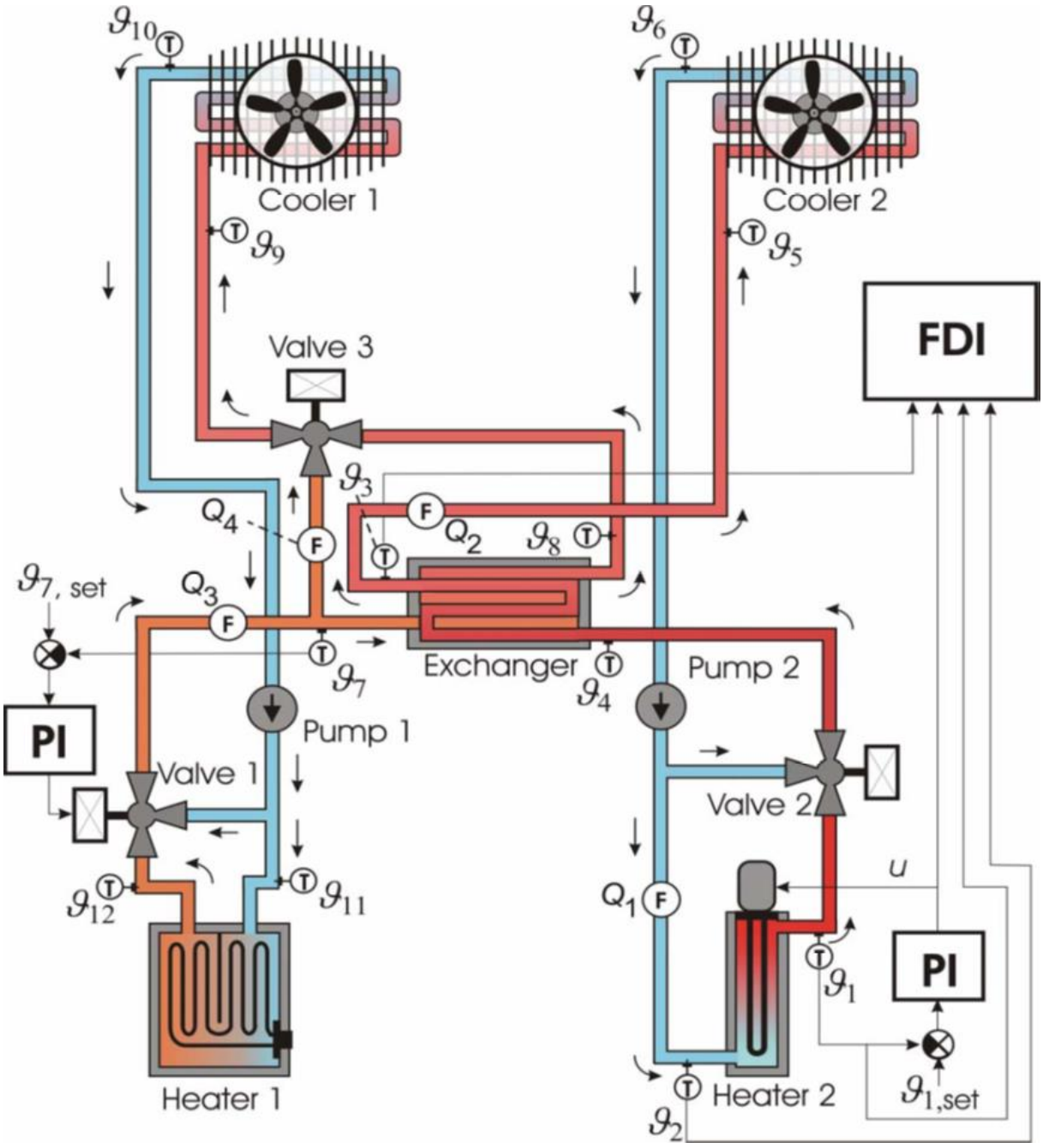
4.2.1 Realizace dynamického modelu

Pro praktické experimenty s popisovanou metodou jsem využil laboratorní model Tepelná soustava (obr. 4.2.1) navržený a postavený Centrem Aplikované Kybernetiky na ČVUT v Praze.

obr. 4.2.1 – Laboratorní model Tepelná soustava. (převzato z [45])

Experimentální model

Soustavu tvoří dva nezávislé uzavřené okruhy s nuceným oběhem ohřívané vody, které se navzájem ovlivňují pouze prostřednictvím deskového tepelného výměníku. Voda v okruhu „1" (vlevo) je zahřívána akumulačním ohřívačem **Heater1** a ochlazována vzduchovým aktivním chladičem **Cooler1**, cirkulace je zajištěna oběžným čerpadlem **Pump1**. Voda v okruhu „2" je zahřívána průtokovým ohřívačem **Heater2**. Chlazení zajišťuje aktivní vzduchový chladič **Cooler2** a cirkulaci vody oběžné čerpadlo **Pump2**. Vedení potrubí a umístění jednotlivých zařízení je schematicky znázorněno na obr. 4.2.2.



obr. 4.2.2 - Schéma laboratorního modelu Tepelná soustava. (převzato z [45])

Ohřev vody je v každém okruhu zpětnovazebně řízen softwarovým PI regulátorem. Teplotu vody a směr průtoku v obou okruzích je navíc možné ovlivnit trojicí směšovacích ventilů **Valve1**, **Valve2** a **Valve3**.

Soustava je vybavena dvanácti snímači **T** pro měření teplot $\vartheta_1, \vartheta_2, \ldots \vartheta_{12}$ a čtyřmi průtokoměry **F** pro měření průtoků $Q_1, \ldots Q_4$ (viz schéma obr. 4.2.2).

Řídicí a diagnostický software

Ovládání a sledování tepelné soustavy je realizováno prostřednictvím osobního počítače vybaveného softwarem Matlab – Simulink.

Jako základ diagnostického systému jsem použil software (viz [45]), který jsem následně modifikoval. Software je schopen diagnostikovat až deset provozních režimů (jeden bezporuchový stav a až devět různých poruch).

Software jsem doplnil o modul rozšířené logiky. Dále jsem přizpůsobil uživatelské rozhraní, aby bylo možné porovnávat původní a nové diagnostické metody.[A2]

Experimentální diagnostika poruch na modelu Tepelná soustava

Abych mohl porovnat výsledky diagnostiky s použitím navržené metodiky, provedl jsem nejprve kompletní měření podle standardního postupu popsaného v [43], kdy jsem v průběhu simulace postupně učil diagnostický model na bezporuchový stav a jednotlivé poruchy. Poté jsem provedl další měření, již bez učení, abych prověřil kvalitu rozpoznávání jednotlivých poruchových stavů.

Potom jsem využil zaznamenané průběhy všech sledovaných veličin. Tyto časové průběhy jsem potom spustil v reálném čase v simulačním prostředí Matlab-Simulink, takže pro diagnostický systém se simulace jevila zcela stejně jako reálný proces. Provedl jsem učení a následně i diagnostiku na upravené, dynamicky orientované sadě poruchových stavů.

Tím jsem dosáhl naprosto stejných podmínek experimentu pro obě verze diagnostického systému a maximální možnou míru objektivity při porovnání klasického a dynamického přístupu v diagnostice poruch.

Průběh experimentu

Ve všech prováděných experimentech jsem zachoval předpoklad, že jednotlivé definované poruchy se navzájem nekombinují ani na sebe přímo nenavazují. Tedy že porucha nastává vždy izolovaně po delším období bezporuchového provozu.

Také jsem dodržel pravidlo, že nástup každé poruchy a také přechod z poruchy zpět do bezporuchového stavu probíhá skokově nebo se alespoň skokovému průběhu co nejvíce blíží.

Vlastní experiment probíhal v následujících krocích:

- Volba měřených veličin a definice regresního vektoru.
- Návrh seřízení soustavy v různých režimech (bezporuchový stav NB a předdefinované poruchy).
- Naměření dostatečného množství dat a naplnění matice přechodu.
- Naměření dostatečného množství dat a prověření kvality diagnostiky.

Z každého experimentu jsem archivoval kompletní záznam naměřených dat včetně veličin, které nebyly součástí regresního vektoru. Také jsem vedl deník o průběhu experimentu obsahující podrobný protokol o provedených nastaveních všech stavitelných parametrů soustavy včetně přesných časů jejich změn a včetně poznámek o vnějších podmínkách měření.

Regresní vektor jsem zvolil na základě předběžných experimentů. Je tvořen veličinami (viz schéma – obr. 4.2.2):

$$r_k = \left[\vartheta_1(k-12), \vartheta_2(k-1), \vartheta_3(k-1), u(k-12) \right]^T, \qquad (4.2.1)$$

kde ϑ_1, ϑ_2, ϑ_3 jsou měřené teploty a u je napětí topného tělesa Heater 2.

Úmyslně jsem do regresního vektoru nezahrnul žádná čidla na okruhu „1“, který tak představoval nepřístupnou část sledovaného technologického procesu.

Definice poruchových stavů soustavy jsou v tabulce 4.2.1.

Pozn.: Výtlak čerpadla je možné nastavit pouze manuálně pomocí ovládacího prvku na jeho svorkovnici. Dále uváděné hodnoty jsou v metrech výtlačné výšky a odpovídají značkám vyraženým na ovladači. Ostatní akční členy jsou ovládány přímo z řídícího počítače.

Tab. 4.2.1 – Definice známých poruchových stavů soustavy:

Stav soustavy	Popis stavu
Bezporuchový stav NB	Výkon chladičů: *P*1 = *P*2 = 50% Směšovací poměry ventilů Valve2 a Valve3: 100%:0% Výtlak čerpadel: Pump2 = 2 m; Pump1 = 1 m
Porucha F1	Zvýšení výkonu chladiče Cooler2: *P*2 = 90%
Porucha F2	Zvýšení výtlaku čerpadla Pump2 = 3,5 *m*
Porucha F3	Zvýšení výtlaku čerpadla Pump1 = 3 *m*
Porucha F4	Změna směšovacího poměru ve ventilu Valve3: 50:50%
Porucha F5	Změna směšovacího poměru ve ventilu Valve2: 70:30%
Porucha F6	Snížení výkonu chladiče Cooler2: *P*2 = 25%
Porucha F7	Snížení výtlaku čerpadla Pump2 = 1.5*m*

Rekonfigurace množiny poruchových stavů

Použitý software umožňuje definovat maximálně deset stavů včetně bezporuchového. Proto jsem diagnostikoval pouze tři chyby a jejich přechodové stavy. Původní stavy označené NB, F1, F2 až F7, jsem nahradil následující strukturou (Tab. 4.2.2).

Tab. 4.2.2 – Přemapování stavů při přechodu na dynamicky orientovaný přístup

Původní stavy	Nové stavy	Označení v programu
NB	F1 → NB F2 → NB F3 → NB **NB → NB**	F3 F6 F9 **NB**
F1	NB → F1 **F1 → F1**	F1 **F2**
F2	NB → F2 **F2 → F2**	F4 **F5**
F3	NB → F3 **F3 → F3**	F7 **F8**

Cíleně jsem vybral poruchy, u kterých se v případě použití přístupu bez zohlednění dynamických přechodových dějů vyskytovaly problémy s jejich rozpoznáváním nebo u kterých byla dosažena nízká věrohodnost odhadu (menší pravděpodobnost, že se jedná právě o tuto poruchu).

Modul rozšířené logiky

Modul jsem realizoval ve formě souboru jednoduchých binárních pravidel, jejichž výsledky jsem použil jako časově proměnlivé váhy přidávané k okamžitým odhadům pravděpodobnosti jednotlivých stavů.

Nadefinoval jsem následující pravidla:

1. Mezi dvěma ustálenými stavy musí být vždy příslušný přechodový děj.
2. Mezi dvěma poruchami musí vždy nejméně jednou nastat (ustálený) bezporuchový stav.
3. Přechodové děje se mohou řetězit, musí však logicky navazovat.
4. Dostane-li se soustava do neznámého stavu, může následovat libovolný ustálený nebo přechodový stav bez omezení.
5. Libovolný stav z druhého sloupce tabulky 4.2.2 je logicky přidružen k příslušnému nadřazenému stavu v prvním sloupci.

Zjednodušeně řečeno, pravidla 1. až 3. předepisují, že stavy definované ve druhém sloupci tabulky 4.2.2 na sebe musí navzájem navazovat „jako kostky domina“, pravidlo 4. tento požadavek ve speciálních případech poněkud zmírňuje, aby nemohlo dojít k zablokování běhu diagnostiky.

Pravidlo 5. slouží k vytvoření struktury kompatibilní s klasickým přístupem a je klíčové pro maximální využití výhod dynamického přístupu kategorizace stavů.

Srovnání výsledků experimentů

obr. 4.2.3 a), b) ilustruje příklad rozpoznání poruchy. Na svislé ose obou grafů je kumulativní míra pravděpodobnosti, že se soustava v daný okamžik nachází v určitém poruchovém stavu. Na vodorovné ose je čas měření v sekundách.

V horním grafu jsou zobrazeny průběhy odhadů pravděpodobností poruchových stavů definovaných podle klasické metody bez rozlišení přechodových dějů, ve spodním grafu jsou průběhy odhadů stavů definovaných podle nové metody zohledňující dynamický charakter procesu.

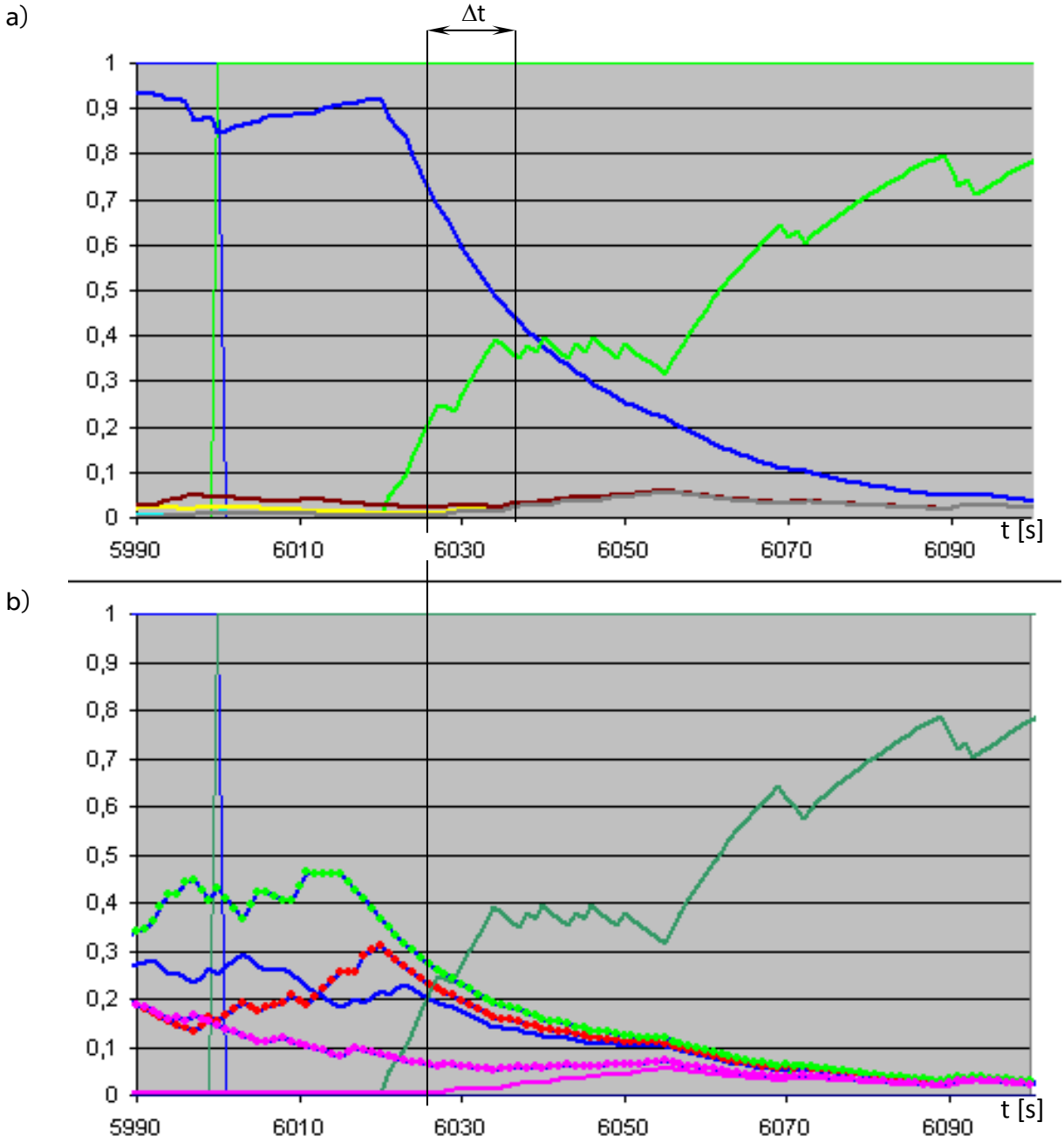
V tomto příkladu bylo cílem co nejrychlejší rozpoznání poruchy F2, která nastala skokově v čase $t = 6000\,s$. Klasickým přístupem byla porucha jednoznačně identifikována v čase $t = 6040\,s$, dynamicky orientovaným přístupem pak v čase $t = 6029\,s$. Znamená to o 11 s, tedy 11 vzorkovacích intervalů, rychlejší rozpoznání poruchy.

Metoda FDI zohledňující dynamický charakter sledovaných dějů je úspěšná a v prováděných experimentech jsem s ní dosahoval téměř bez výjimky prokazatelně lepších výsledků, než s klasickou metodou.

Pouze v jednom případě (u poruchy F3) se mi nepodařilo docílit úspěšného rozpoznání poruchy. V tomto případě však nebyla úspěšná ani klasická metoda. Příčinou je samotná definice poruchy, protože ze své podstaty ovlivní prakticky pouze primární okruh tepelné soustavy, kde není využitý žádný senzor. Porucha totiž ovlivní pouze velmi okrajově tepelný výměník a to ještě jenom při výraznější změně teploty. Lze tedy říci, že tato porucha je z objektivních příčin téměř nedetekovatelná.

Ve všech ostatních případech se mi podařilo docílit s dynamicky orientovaným přístupem znatelného urychlení rozpoznání nástupu poruchy. Doba potřebná na rozpoznání poruchy se v různých případech zkrátila v rozmezí přibližně o 5 % až 50 % původního trvání.

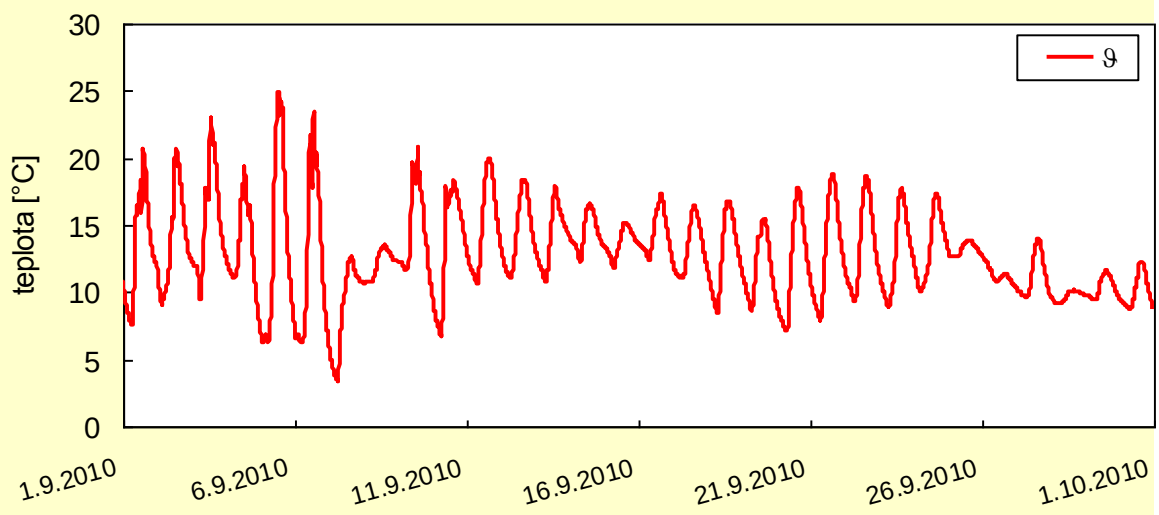
Obecně lze říci, že čím bylo obtížnější poruchu rozpoznat, tím zřetelněji byl dynamicky orientovaný přístup rychlejší oproti klasickému.



obr. 4.2.3 – Časový průběh kumulativních pravděpodobností poruch. Tenkými čarami jsou zobrazeny skutečné stavy systému, silnými čarami jejich kumulativní pravděpodobnosti odhadnuté diagnostickým systémem.
a) standardní klasifikace stavů – (modrá: bezporuchový stav NB, světle zelená: porucha F2; poruchy zobrazené dalšími barvami mají téměř nulovou pravděpodobnost)
b) dynamicky orientovaná klasifikace stavů (před aplikací rozšířené logiky) – (modrá: bezporuchový ustálený stav NB-NB, tmavě zelená: nástup poruchy NB-F2, modrá se sv. zelenou: nástup bezporuchového stavu F2-NB, modrá s červenou: nástup bezporuchového stavu F3-NB)

4.2.2 Realizace on-line EMD

Algoritmus on-line EMD rozkladu v reálném čase jsem experimentálně ověřil na reálných datech naměřených v rámci projektu TOKENELEK (obr. 4.2.4) [52], [A11], [A12], [A13], [A14], [A15]. Jedná se o průběh teploty půdy naměřené v hloubce 30 cm na meteorologické stanici označené Vrt_Domanín v průběhu září 2010. Vzorkovací perioda je 10 min [A7], [A8].



obr. 4.2.4 - Časový průběh původního signálu – teplota půdy.

Experimentální analýza probíhala ve dvou fázích.

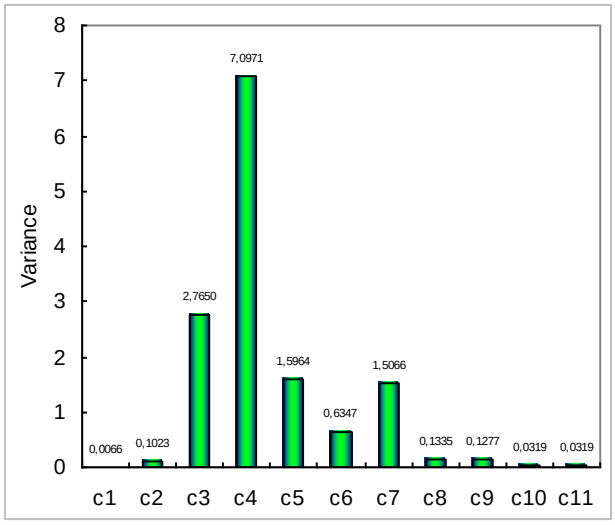
Nejprve jsem na měřená data aplikoval off-line empirickou modální dekompozici. Výsledné funkce jsou zobrazeny v grafech obr. 4.2.7 a) až l) a slouží jako referenční vzorky vlastních modálních funkcí IMF.

Pro každou IMF jsem vypočetl její rozptyl, který jsem použil jako jednoduchý indikátor pro přibližné posouzení, jak významnou složku původního signálu představuje daná modální funkce. Jak je patrné z obr. 4.2.5 a z obr. 4.2.7, čím větší má určitá komponenta rozptyl, tím větší má také průměrnou amplitudu a proto je vyhodnocena jako významnější. V tomto případě je nejvýznamnější složkou teploty půdy empirická modální funkce c_4, která pokrývá velkou část dějů s periodicitou odpovídající přibližně jednomu dni.

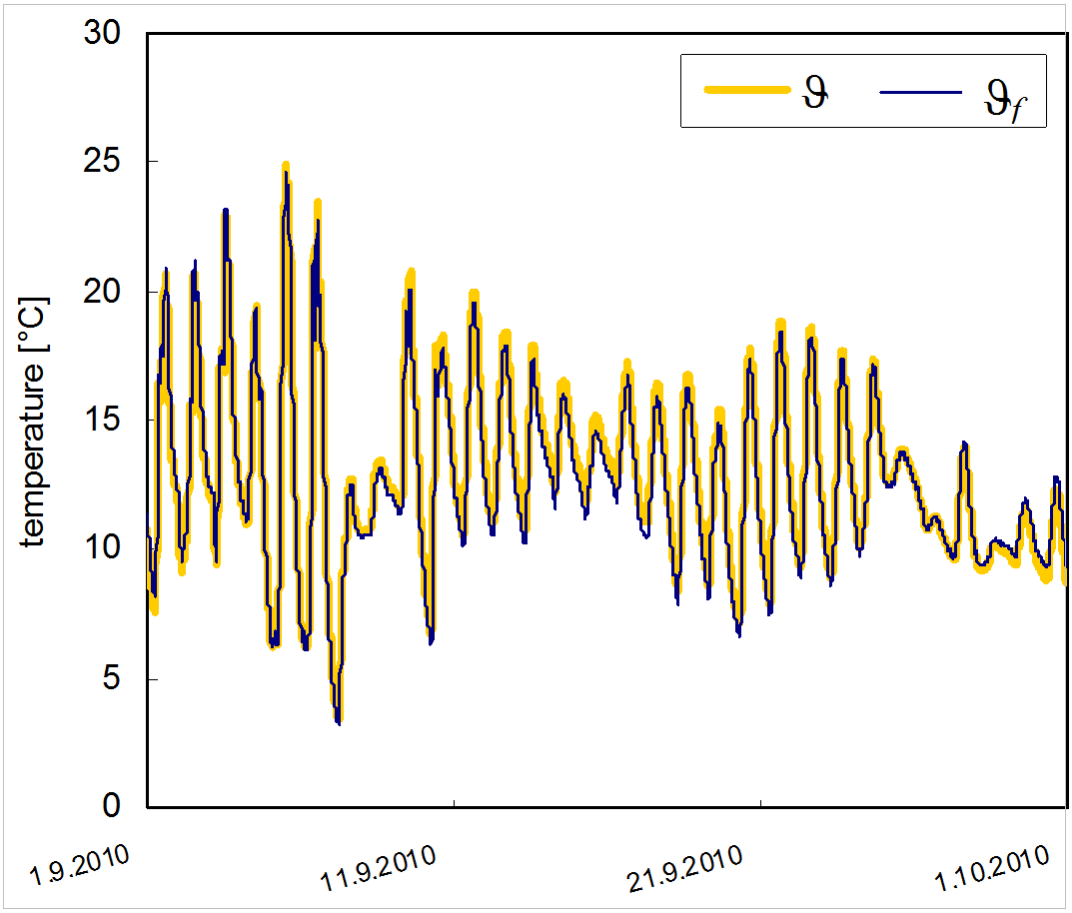
obr. 4.2.6 představuje srovnání původního signálu $\vartheta(t)$ s filtrovaným signálem $\vartheta_f(t)$, který je vyjádřen vztahem

$$\vartheta_f(t) = \sum_{i=2}^{9} c_i(t) + r(t). \tag{4.2.2}$$

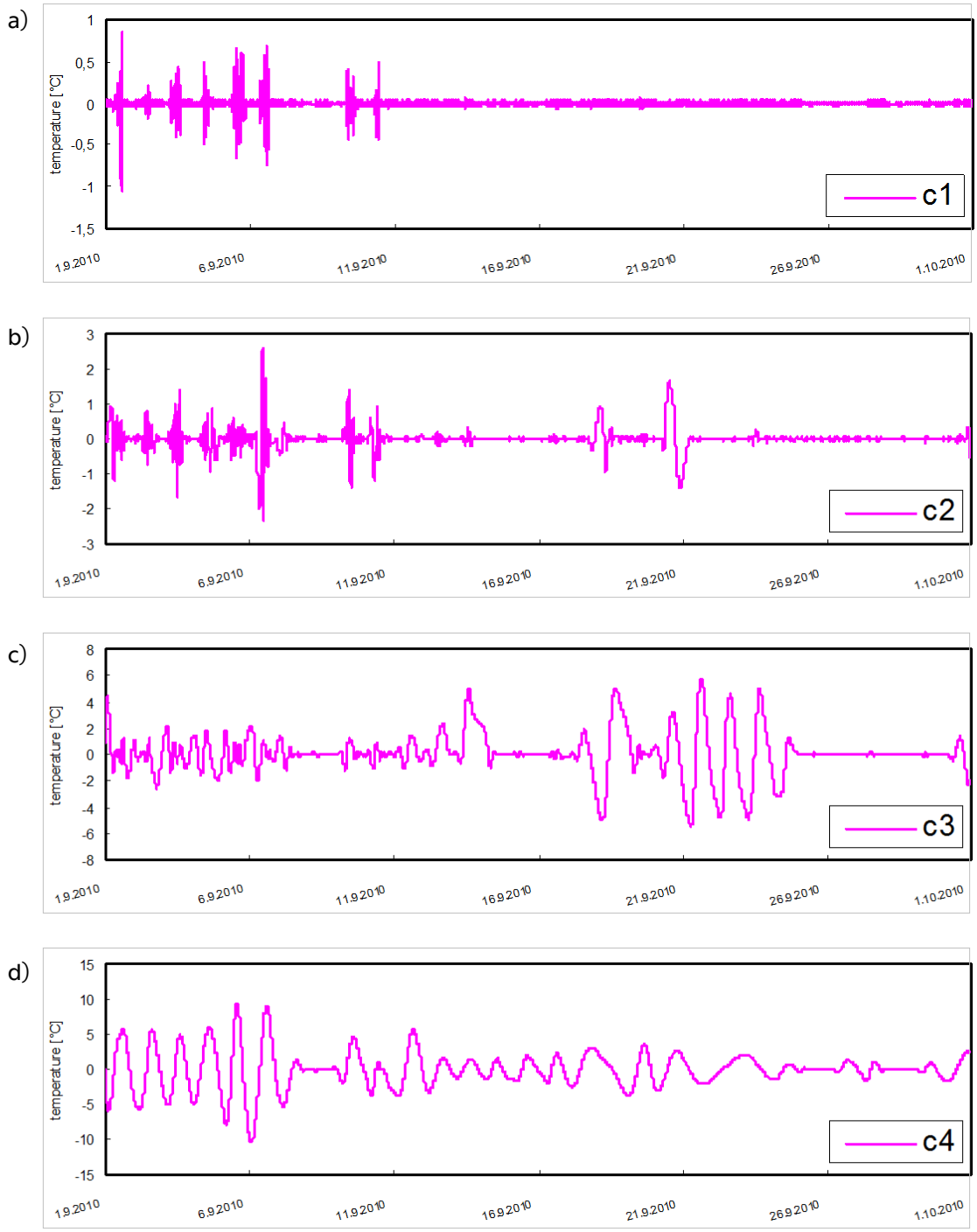
Do filtrovaného signálu jsem kromě rezidua zahrnul pouze jeho podstatné složky. Jako filtrovací kritérium jsem použil minimální hodnotu rozptylu. Limitní hodnotu jsem stanovil na 0,1.



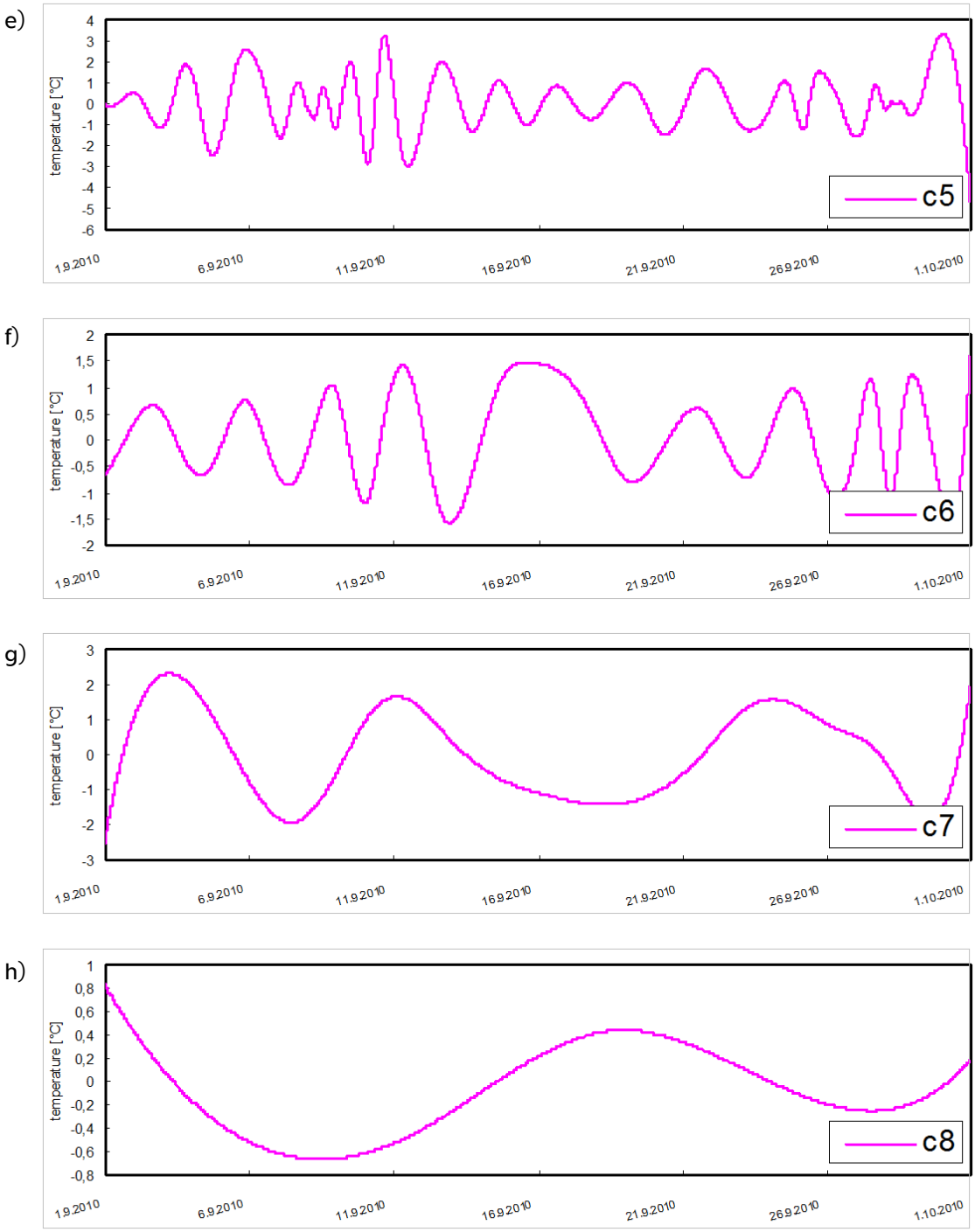
obr. 4.2.5 - Porovnání rozptylů vlastních modálních funkcí. Čím větší je rozptyl, tím významnější je příslušná složka signálu.



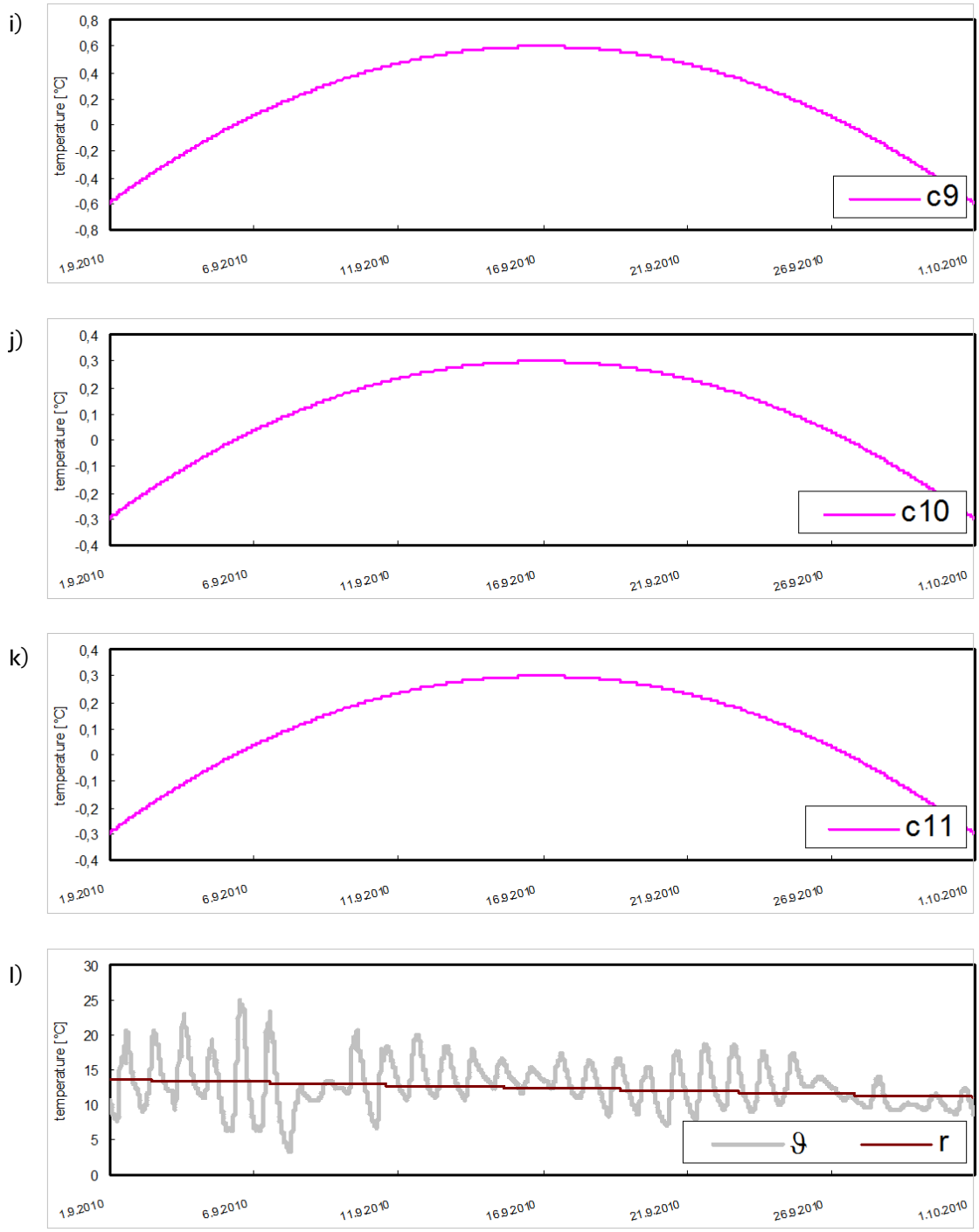
obr. 4.2.6 –Srovnání teplotních trendů ϑ a ϑ_f. Silná linka reprezentuje originální naměřená data, tenká linka filtrovanou funkci získanou kombinací vybraných nejvýznamnějších (off-line) modálních funkcí a rezidua.



obr. 4.2.7 a) až l) - Vlastní modální funkce (IMF) a reziduum získané jednorázovou off-line dekompozicí. V grafech je důležité nepřehlédnout měnící se měřítko na svislé ose. Reziduum v posledním grafu je zobrazeno společně s původními daty.



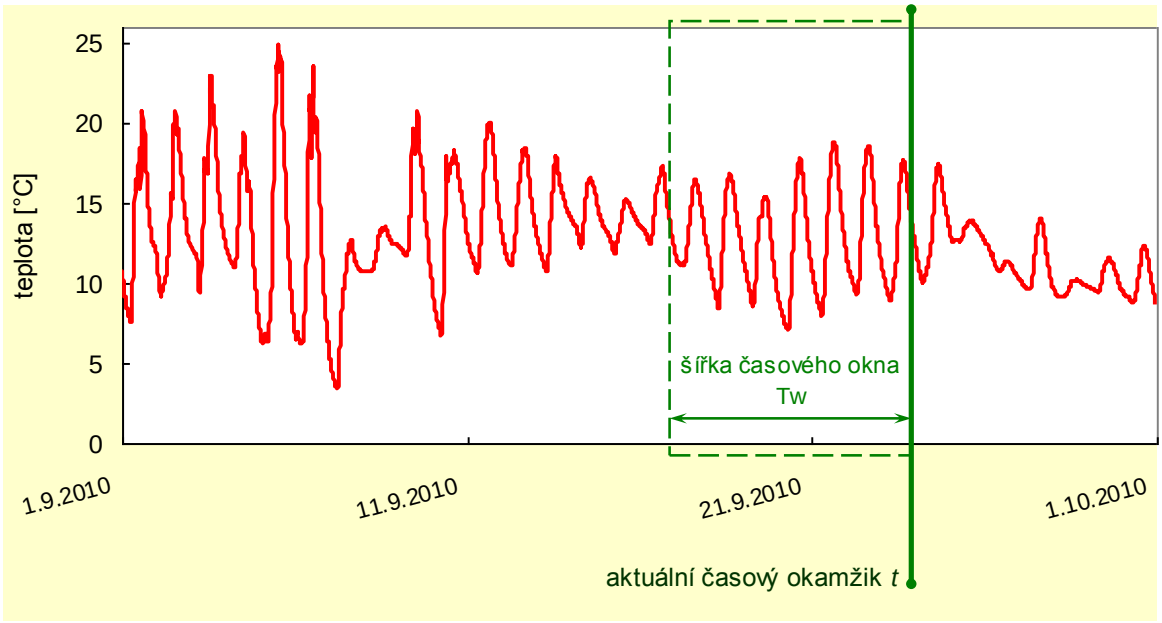
obr. 4.2.7 a) až l) - Vlastní modální funkce (IMF) a reziduum získané jednorázovou off-line dekompozicí. V grafech je důležité nepřehlédnout měnící se měřítko na svislé ose. Reziduum v posledním grafu je zobrazeno společně s původními daty.



obr. 4.2.7 a) až l) - Vlastní modální funkce (IMF) a reziduum získané jednorázovou off-line dekompozicí. V grafech je důležité nepřehlédnout měnící se měřítko na svislé ose. Reziduum v posledním grafu je zobrazeno společně s původními daty.

Ve druhé fázi experimentu Jsem zpracoval stejná data pomocí on-line modální dekompozice. Šířku časového okna jsem zvolil na1008 vzorků, což při dané vzorkovací periodě představuje sedm dní. Interval jsem volil tak, aby dostatečně pokryl nejpodstatnější složky signálu, viz obr. 4.2.5.

Z obr. 4.2.8 je patrný průběh naměřené teploty $\vartheta(t)$ s naznačenou jednou realizací časového okna v obecném okamžiku t. Plovoucí časové okno v průběhu analýzy je označeno zeleným obdélníkem.

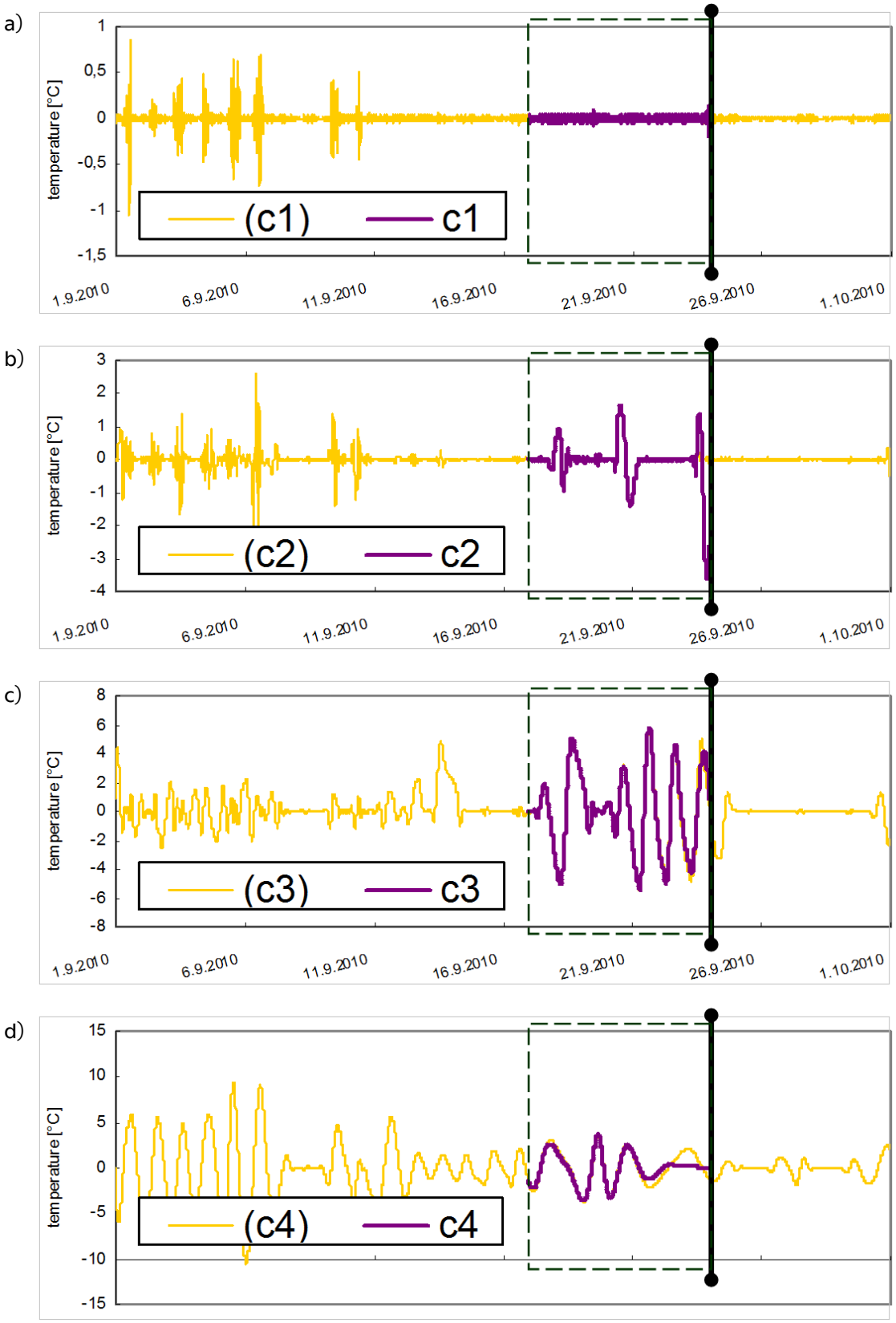


obr. 4.2.8 – Průběh teploty se symbolicky naznačeným časovým oknem.

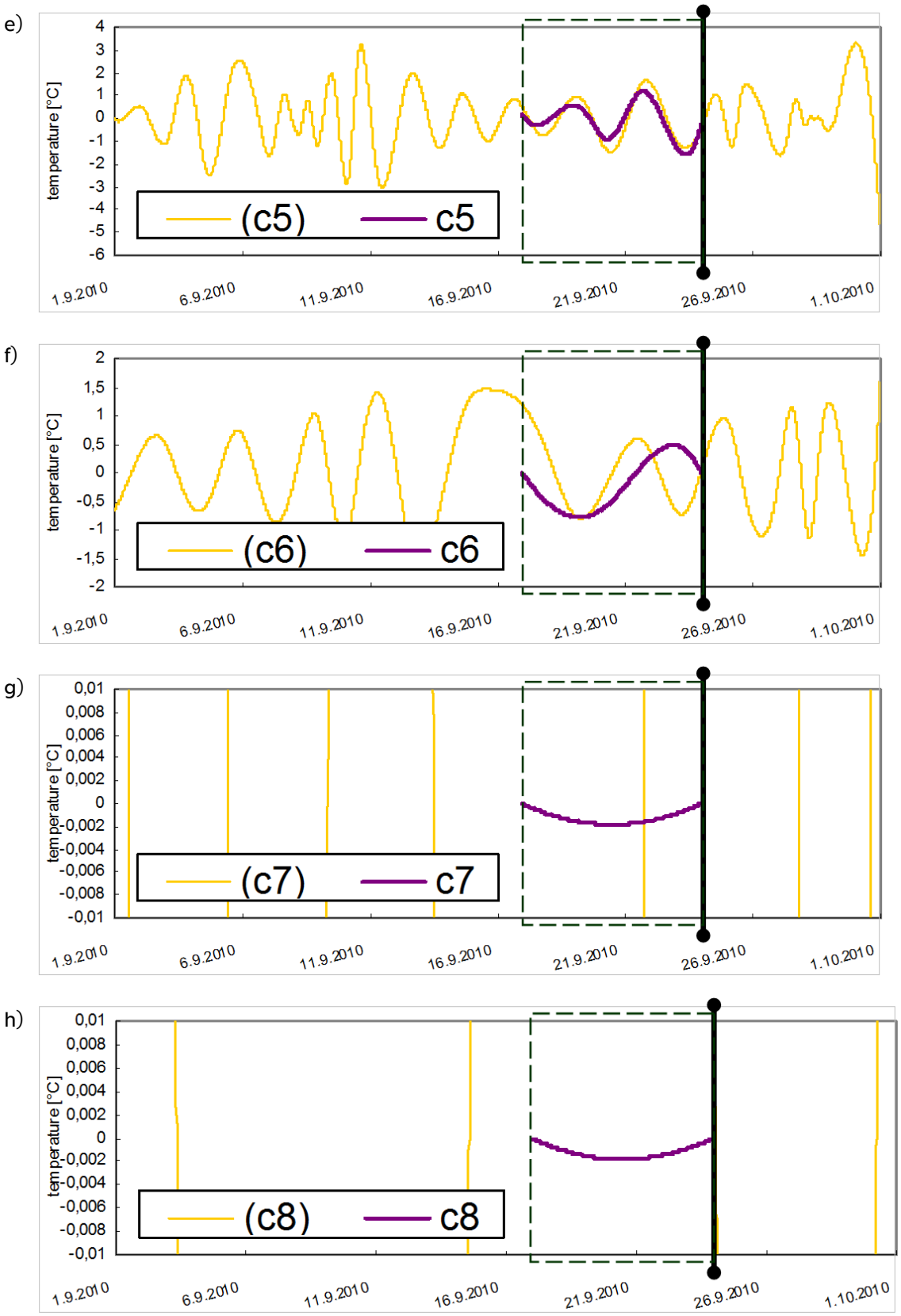
Na obr. 4.2.9 a) až i) je zobrazena jedna sada modálních funkcí získaných on-line metodou v plovoucím okně. Je patrné, že reziduum je zakřivenější oproti off-line zpracování. Stejně tak některé modální funkce mají větší rozptyl. Tento fenomén je nezbytným a očekávaným důsledkem zkráceného časového rozsahu analyzovaných dat.

Zároveň s vyhledáváním modálních funkcí jsem průběžně dopočítával také odhady jejich rozptylů. Na základě rozptylů jsem následně rozhodoval, které složky budou zahrnuty do po částech rekonstruovaného signálu $\vartheta_f(t)$ a které z něj budou odfiltrovány.

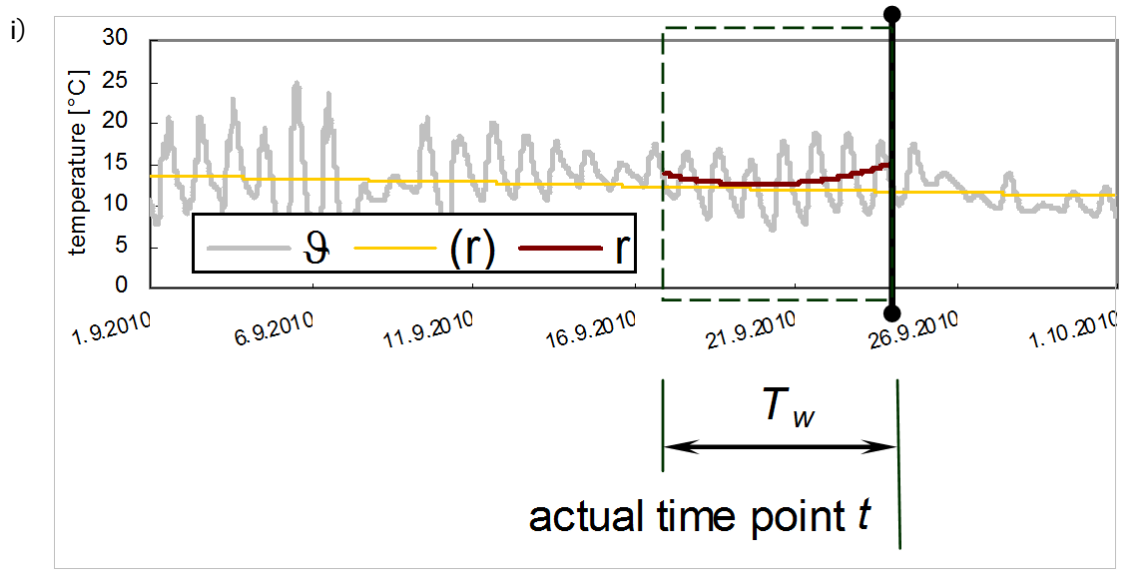
obr. 4.2.10 ukazuje průběh několika dílčích filtrovaných funkcí ve srovnání s originálním signálem. Silná čára představuje originální data, tenké barevné čáry představují dílčí filtrované signály pro vybraná časová okna.



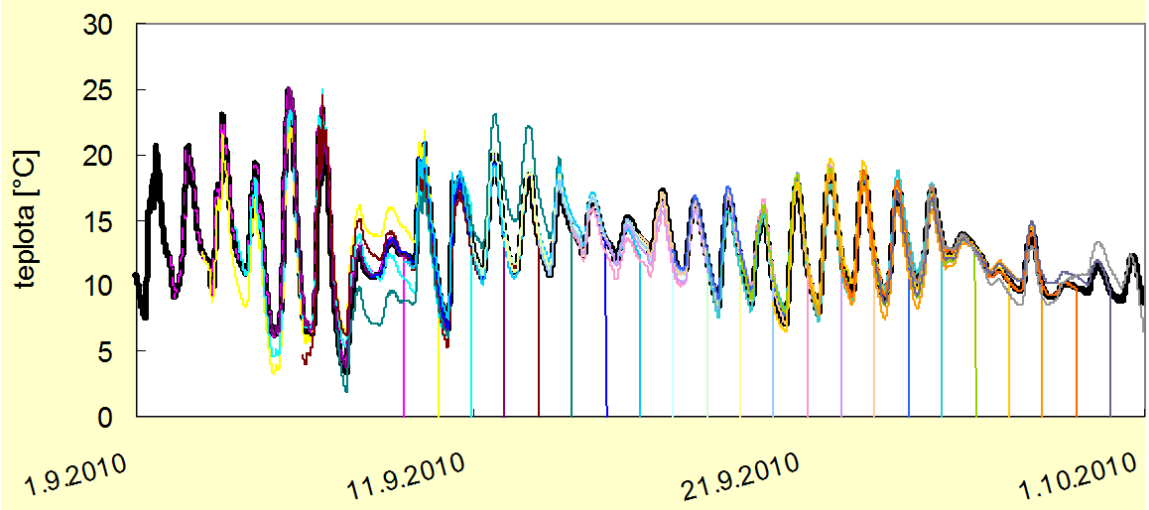
obr. 4.2.9 a) až i) – Porovnání IMF funkcí a reziduí získaných metodou off-line EMD (Žlutě) a metodou on-line EMD (Fialově). Počty nalezených vlastních modálních funkcí se liší díky rozdílné délce analyzovaných dat. Srovnání ukazuje především vliv velikosti časového okna na výsledek EMD algoritmu.



obr. 4.2.9 a) až i) – Porovnání IMF funkcí a reziduí získaných metodou off-line EMD (Žlutě) a metodou on-line EMD (Fialově). Počty nalezených vlastních modálních funkcí se liší díky rozdílné délce analyzovaných dat. Srovnání ukazuje především vliv velikosti časového okna na výsledek EMD algoritmu.



obr. 4.2.9 a) až i) – Porovnání IMF funkcí a reziduí získaných metodou off-line EMD (Žlutě) a metodou on-line EMD (Fialově). Počty nalezených vlastních modálních funkcí se liší díky rozdílné délce analyzovaných dat. Srovnání ukazuje především vliv velikosti časového okna na výsledek EMD algoritmu.



obr. 4.2.10 – Srovnání naměřeného průběhu teploty a (on-line) dílčích filtrovaných funkcí $\vartheta_f(t)$. Originální signál reprezentuje silná černá čára. Tenké barevné čáry reprezentují dílčí filtrované signály pro vybraná časová okna. Interval mezi zobrazenými časovými okny je 1 den, šířka oken 7 dní.

Z výsledků experimentů (viz obr. 4.2.9, obr. 4.2.10) je patrné, že on-line empirická modální dekompozice poskytuje výsledky kvalitativně srovnatelné s klasickou off-line analýzou. Srovnání obou metod ukazuje, že on-line analýza s plovoucím oknem zvládá dekompozici podstatně rychleji a s výrazně menšími paměťovými nároky.

Na druhou stranu off-line analýza poskytuje o něco přesnější rozklad vzhledem k tomu, že dokáže zpracovat složky s delší periodicitou a tedy s nižšími frekvencemi. Hodnotu minimální zjistitelné frekvence je možné určit vhodnou volbou velikosti časového okna.

IMF pro vyšší frekvence jsou v obou případech prakticky totožné, s klesající frekvencí se výsledky stále více navzájem odchylují.

5 Důsledky pro vědu a praxi

5.1 Důsledky pro vědu

Metodika dynamicky orientované klasifikace provozních stavů s využitím vyvážených statistik vzhledem k délkám trénovacích množin především otevírá cestu k poněkud odlišnému pohledu na bayesovský stochastický systém založený na markovském modelu ve vztahu k jeho dynamickým vlastnostem a na způsob, jakým model interpretuje zpracovávaná data.

Mezi základní podmínky snadného použití markovského modelu s využitím matice aposteriorních přechodových pravděpodobností patří požadavek na stacionaritu, resp. ergodicitu modelovaného markovského procesu. Tato podmínka je nutná, aby bylo možné využívat časově invariantní rozdělení pravděpodobností v matici přechodu a tím využít velkou výhodu markovského modelu v porovnání např. s neuronovými sítěmi – jeho přehlednost a srozumitelnost. Můžeme říci, že markovský model je stacionární „ve velkém“ v tom smyslu, že pokud máme dostatečně kvalitní trénovací data, můžeme matici přechodu jednou sestavit a pak ji už beze změn používat.

Navržený postup rozdělení provozních režimů na přechodové a ustálené poskytuje názornou představu o principech funkce stochastického bayesovského modelu. Zároveň zmírňuje omezení vyvolaná požadavkem na časově invariantní statistiky. Na rozdíl od klasického přístupu nevyžaduje pomalé dlouhodobé přechodové děje, protože nevyžaduje srovnatelnou dobu trvání přechodového a ustáleného děje.

5.2 Důsledky pro praxi

Změna kategorizace stavů představuje velice praktickou modifikaci. Velkým kladem této metody je jednoduchost nasazení v reálných aplikacích.

Není potřeba provádět prakticky žádné úpravy řídicího algoritmu s vyjímkou doplnění modulu rozšířené logiky. Ten je však možno začlenit jako dodatečný a zcela samostatný prvek. Navíc není kriticky nezbytný pro fungování metody. Za cenu určitého nepohodlí obsluhy je dokonce možné modul rozšířené logiky zcela vynechat.

Modifikace výpočtu statistik tak, aby byly nezávislé na vzájemných délkách trénovacích množin, poskytuje projektantovi stochastického modelu podstatně větší svobodu nejen při přípravě trénovacích dat, ale také při návrhu vzájemných vztahů mezi provozními režimy, například při potlačování či zdůrazňování určité poruchy.

Určitou komplikací aplikace metody v praxi je tedy nutnost přegenerovat matici přechodu. To však patří mezi rutinní činnosti spojené s provozem bayesovského klasifikátoru FDI založeného na markovském modelu. Za přirozeného předpokladu, že jsou trénovací data průběžně archivována, nepředstavuje zásadní problém.

6 Závěr

V rámci disertační práce jsem popsal a prakticky ověřil inovativní přístup ke zpracování dat pravděpodobnostním systémem diagnostiky poruch založeným na bayesovském přístupu a využívajícím markovského modelu sledovaného procesu.

Splnění primárního cíle

Primárním cílem mé disertační práce bylo nalezení postupů nebo modifikace diagnostického systému, které by vedly ke zlepšení schopnosti rozpoznat poruchy s rychlým, krátkodobým nástupem. Tento cíl byl splněn.

Splnění dílčích cílů

1) Navrhnout metodu pro eliminaci nežádoucího vlivu rozdílných délek trénovacích množin pro různé provozní režimy.

V kapitole 2.2 jsem představil markovský model dynamického systému a ukázal jeho nejdůležitější vlastnosti. Kapitolu 2.2 jsem použil jako teoretické východisko k vytvoření vlastní metody pro získání vyvážených statistik, která je popsána v kapitole 4.1.1. V první části jsem na základě markovského modelu popsal bayesovský klasifikátor poruchových stavů. Na bayesovském klasifikátoru jsem následně formuloval metodu pro získání vyvážených statistik. Ta je vyjádřena ve stěžejní části „Dynamika přechodů mezi stavy soustavy", kde jsem odvodil vztah umožňující vypočítat vyvážené statistiky, které nejsou ovlivněny délkami dílčích trénovacích množin a které umožňují nezávisle definovat vlastní relativní váhy (míry očekávání) jednotlivých provozních režimů (poruch).

2) Navrhnout metodu, jak zabránit potlačování vlivu krátkodobých přechodových dějů při změně provozního režimu vlivem jejich malého zastoupení v trénovací množině.

Metodu rozšířené klasifikace provozních a poruchových stavů jsem popsal v kapitole 4.1.2, ve které jsem formuloval stěžejní myšlenku rozdělit trénovací množinu jednoho provozního režimu na přechodovou a ustálenou část. Tento postup umožnil při použití vyvážených statistik respektovat výrazně odlišné režimy chování stochastického procesu v ustálených stavech a během přechodových dějů. V rámci implementace metody jsem v části „Rozšířená logika" navrhl a realizoval jednoduchý pravidlový systém rozšířené logiky pro řetězení a vázání dílčích provozních režimů (poruch).

3) Navrhnout vlastní modifikaci algoritmu empirické modální dekompozice tak, aby byl použitelný pro výpočet vlastních modálních funkcí v reálném čase se zaměřením na využití v diagnostice poruch.

V kapitole 4.1.3 jsem navrhl inovativní přístup k sestavení regresního vektoru diagnostického systému spočívající ve využití empirické modální dekompozice (EMD). Modifikovaná struktura regresního vektoru obsahuje kromě dat získaných přímo z procesu také jejich hlavní (či podstatné) kmitavé mody nazvané vlastní modální fukce. Uvedená úprava vede k detailnějšímu rozlišení vnitřních stavů procesu a tím také k potenciálně úspěšnější diagnostice. Část „On-line empirická modální dekompozice" popisuje jednoduchý algoritmus on-line EMD spočívající v provádění dekompozice v plovoucím časovém okně.

4) Experimentálně ověřit navržené postupy a metody.

Implementaci uvedených metod jsem realizoval v prostředí Matlab/Simulink. Experimentální ověření cílů 1), 2) je popsáno v kapitole 4.2.1. Ve většině provedených experimentů se výsledky diagnostického systému prokazatelně zlepšily, v ostatních případech zůstaly prakticky shodné.

Experimentální ověření cíle 3) je popsáno v kapitole 4.2.2. Algoritmus on-line EMD je podstatně rychlejší a má výrazně menší paměťové nároky, poskytuje však méně přesný rozklad v porovnání s klasickou off-line metodou.

Poděkování

Děkuji svému školiteli prof. Ing. Milanu Hofreiterovi, CSc. za odborné vedení a konzultace v průběhu studia doktorského studijního programu, za přínosné rady a připomínky k disertační práci.

V neposlední řadě chci poděkovat celé mojí rodině za všemožnou podporu a trpělivost po celou dobu mého studia. Jmenovitě pak mamince, své partnerce Jitce, sourozencům Kamile a Ondrovi.

Práci věnuji svému otci Pavlu Trnkovi *in memoriam*.

7 Literatura

7.1 Cizí prameny

[1] Chen, J., Patton, R. J.: *Robust model-based fault diagnosis for dynamic systems.* Norwel (Massachusetts): Kluwer Academic Publishers, 1999, 356 s., ISBN 0-7923-8411-3

[2] Kárný, Miroslav (Ed.): *Optimized Bayesian Dynamic Advising. Theory and Algorithms*. Springer-Verlag, London, 2006, 529 s., ISBN: 978-1-85233-928-9

[3] Schlesinger, M. I., Hlaváč, V.: *Deset přednášek z teorie statistického a strukturního rozpoznávání*. Praha, Vydavatelství ČVUT 1999

[4] Dynkin, E. B.; Yushkevich, A. A.: *Controlled Markov processes*. New York, USA, Springer-Verlag New York Inc., 1979

[5] Kvišťák, M.: *Základy teórie stochastických procesov.* Bratislava, AX INZERT 1998

[6] Ding, X., Guo, L.: *An approach to time domain optimalisation of observer-based fault detection systems.* Int. J. Contr. 69(3), p. 419-442, 1998

[7] Gertler, J.: *Fault Detection and Diagnosis in Engineering in Engineering Systems.* Marcel Dekker, New York, 1998

[8] Chung, W. H., Speyer, J. L.: *A game theoretic fault detection filter.* IEEE Trans. Atomat. Contr. 43(2) p143-161, 1998

[9] Frank, P. M., Köppen-Selinger, B.: *New developments using AI in fault diagnosis.* Eng. Apl. of AI, 10(1), p. 3-14, 1997

[10] Gertler, J., DiPierro, G.: *On the relationship between parity relations and parameter estimation.* Proc. of the IFAC Sympo. on Fault Detection, Supervision and Safety for Technical Processes: SAFEPROCESS'97, Pergamon, Univ. of Hull, UK, p. 453-458, 1998

[11] Isermann, R., Ballé, P.: *Trends in the application of model-based fault detection and diagnosis of technical processes.* Contr. Eng. Practice 5(5), p. 709-719, 1997

[12] Patton, R. J., Chen, J.: *Observer-based fault detection and isolation: robustness and applications.* Contr. Eng. Practice 5(5), p. 671-682, 1997

[13] Zhang, J., Martin, E. B., Morris, A. J.: *Process monitoring using non-linear statistical techniques.* Chemical Eng. J. 67(3), p. 181-189, 1997

[14] Keller, J. Y., Summerer, L., Boutayeb, M., Darouach, M.: *Generalized likelihood ratio approach for fault detection in linear dynamic stochastic systems with unknown inputs.* Int. J. Sys. Sci. 27(12), p. 1231-1241, 1996

[15] Kinnaert, M., Peng, Y. B.: *Residual generator for sensor and actuator fault-detection and isolation - a frequency-domain approach.* Int J. Contr. 61(6), p. 1423-1435, 1995

[16] Frank, P. M., Ding, X.: *Frequency domain approach to optimally robust residual generation and evaluation for model-based fault diagnosis.* In Automatica 30(4), p. 789-804, 1994

[17] Isermann, R.: *Integration of fault detection and diagnosis methods.* Preprints of the IFAC Sympo. on Fault Detection, Supervision and Safety for Technical Processes: SAFEPROCESS'94, Espoo, Finland, p. 597-612 (Vol. 2), 1994

[18] Patton, R. J., Chen, J., Nielsen, S. B.: *Model-based methods for fault diagnosis: Some guidelines.* Inst. M. C. Colloquium on "Quantitative & Qualitative Methods for Fault Diagnosis in Process Control", London, 1994

[19] Gertler., J., Kunwer, M. K.: *Optimal residual decoupling for robust fault diagnosis.* Proc. of Int. Conf on Fault Diagnosis. TOOLDIAG'93, 1993

[20] Frank, P. M.: *Enhancement of robustness in observer-based fault detection.* Preprints of IFAC/IMACS Sympo. SAFEPROCESS'91, Baden-Baden, p. 275-287 (vol. 1), 1991

[21] Gertler, J.: *Analytical redundancy methods in ailure detection and isolation.* Preprints of IFAC/IMACS Sympo: SAFEPROCESS'91, Baden-Baden, p. 9-21, 1991

[22] Chen, J., Zhang, H. Y.: *Robust detection of faulty actuators via unknown input observers.* Int. J. Sys, Sci. 22(10), p. 1829-1839, 1991

[23] Isermann, R.: *Fault diagnosis of machine via parameter estimation and knowledge processing - tutorial paper.* Preprints of IFAC/IMACS Sympo: SAFEPROCESS'91, Baden-Baden, p. 121-133, 1991

[24] Patton, R. J., Chen, J.: *A review of parity space approaches to fault diagnosis.* Preprints of IFAC/IMACS Sympo: SAFEPROCESS'91, Baden-Baden, p. 239-255 (vol. 1), 1991

[25] Patton, R. J., Chen, J.: *A robust parity space approach to fault diagnosis based on optimal eigenstructure assignment.* Proc. of the IEE Int. Con.: Control'91, Peregrinus Press, IEE Conf. Pub. No. 332, Edinburgh, p. 1056-1061, 1991

[26] Frank, P. M.: *Fault diagnosis in dynamic system using analytical and knowledge based redundancy - a survey and some new results.* Automatica 26(3), 459-474, 1990

[27] Isermann, R., Freyermuth, B.: *Process fault diagnosis based on process model knowledge*. Journal A 31(4), p. 58-65, 1990

[28] Tzafestas, S. G., Watanabe, K.: *Modern approaches to system/sensor fault detection and diagnosis*, Journal A 31(4), p. 42-57, 1990

[29] Ge, W., Fang, C. Z.: *Extended robust observation approachfor failure isolation.* Int. J. Contr. 49(5), p. 1537-1553, 1989

[30] Patton, R. J., Frank, P. M., Clark, R. N.: *Fault Diagnosis in Dynamic Systems, Theory and Application.* Control Engineering Series, Prentice Hall, New York, 1989

[31] Gertler, J.: *Survey of model-based failure detection and isolation in complex plants.* IEEE Cotr. Syst. Mag. 8(6), 3-11, 1988

[32] Phatak, M. S., Viswanadham, N.: *Actuator fault detection and isolation in linear systems.* Int. J. Sys. Sci. 19(12), p. 2593-2603, 1988

[33] Frank, P. M.: *Fault diagnosis in dynamic system via state estimation - a survey.* in Tzafestas, Singh and Schmidt (eds), system fault diagnostics, Reliability & Related Knowledge-based Approaches, D. Reidel Press, Dordrecht (vol. 1), p. 35-98, 1987

[34] Isermann, R.: *Experiences with process fault detection via parameter estimation.* In S.G.Tzafestas, M. G. Singh, G. Schmidt (eds), System Fault Diagnostics, Reliability & Related Knowledge-based Approaches, D. Reidel Press, Dordrecht p. 3-33, 1987

[35] Chow, E. Y., Willsky, A. S.: *Analytical redundancy and the design of robust detection systems.* IEEE Trans. Automat. Contr. AC-29(7), 603-614, 1984

[36] Isermann, R.: *Process fault detection based on modelling and estimation methods: A survey.* Automatica 20(4), p. 387-404, 1984

[37] Frank, P. M., Keller, L.: *Sensitive discriminating observer design for instrument failure detection.* IEEE Trans. Aero. & Electron. Syst., AES-16, p. 460-467, 1981

[38] Leininger, G. G.: *Model degradation effects on sensor failure detection.* Proc. of the 1981 joint Amer. Control. Conf., Charlottesville, VA, paper FP-3A (Vol. 3), 1981

[39] Bakiotis, C., Raymond, J., Rault, A.: *Parameter and discrimiant analysis for jet engine mechanical state diagnosis.* Proc. of The 1979 IEEE Conf. on Decision & Control, Fort Lauderdale, USA, 1979

[40] Willsky, A. S.: *A survey of design methods for failure detection in dynamic systems.* in Automatica 12(6), p. 601-611, 1976

[41] Peterka, V.: *Bayesian approach to system identification, Trends and Progress in System Identification, Eykhoff P.* (Ed.). Pergamon Press, Oxford, 1981, pp. 239-304.

[42] Peterka, V.: *Bayesian system identification.* in Automatica. 1981, vol. 17, no. 1, p. 41-53.

[43] Kořenář, V.: *Stochastické procesy.* [s. n.], Praha, 1998, ISBN 80-7079-813-0

[44] Hofreiter, M.: *Pravděpodobnostní identifikace modelu technologického procesu pro syntézu řízení.* Praha, České Vysoké Učení Technické, 2004, ISBN 80-01-03068-7

[45] Garajayewa, G.: *Bayesian approach to real-time fault detection and isolation with supervised training.* Praha, 2005, ČVUT v Praze, vedoucí disertační práce Prof. Ing. Milan Hofreiter, CSc.

[46] Huang, et al.: *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*. Proc. R. Soc. Lond. A (1998) 454, p. 903–995, [cit. 2010-10-06], Online: <http://keck.ucsf.edu/~schenk/Huang_etal98.pdf>

[47] Huang, N. E., Shen, Z., Long, R. S.: *A New View of Nonlinear Water Waves—The Hilbert Spektrum.* In: Ann. Rev. Fluid Mech. 31, p. 417-457, 1999.

[48] Kokeš, Josef: *Hilbert-Huangova transformace aplikovaná v expertním systému.* In: Nové metody a postupy v oblasti přístrojové techniky, automatického řízení a informatiky: odborný seminář. Jindřichův Hradec: Ústav přístrojové a řídicí techniky ČVUT v Praze, 2009.

[49] Zhaohua, W., Huang, N. E.: *Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method.* In: Advances in Adaptive Data Analysis, Vol. 1, No. 1, p. 1–41, 2009, World Scientific Publishing Copany.

[50] Zhaohua, Wu: *HHT MATLAB Program*. [cit. 2010-10-06], Online: <http://rcada.ncu.edu.tw/research1_clip_program.htm>.

[51] Long, Si-sheng, Tie-bao Zhang, Feng Long. *Causes and Solutions of Overshoot and Undershoot and End Swing in Hilbert-Huang Transform*. Acta Seismologica Sinica 18, no. 5 (September 2005): 602–10. doi:10.1007/s11589-005-0039-3.

[52] Hofreiter, M.: *The Application of Hilbert-Huang Transform to Non-Stationary Environmental Data Sets.* In: TMT 2010. Zenica: Faculty of Mechanical Engineering in Zenica, 2010, p. 309-312. ISSN 1840-4944.

[53] Xianzhao, Gengguo, Huikang. *Improved Empirical Mode Decomposition Algorithm of Processing Complex Signal for IoT Application.* International Journal of Distributed Sensor Networks, Hindawi Publishing Corporation, 2014. Article ID 862807.

[54] Tůma, J.: *Zpracování signálů získaných z mechanických systémů užitím FFT*. Sdělovací technika, Praha, 1997, ISBN 80-901936-1-7

[55] Tůma, J.: *Vehicle gearbox noise and vibration: measurement, signal analysis, signal processing and noise reduction measures.* Wiley: Automotive series, Chichster, West Sussex, 2014, ISBN 978-1-118-35941-9

[56] Hofreiter, M.: *Idetifikace systémů I.* Česká technika – nakladatelství ČVUT, Praha, 2009. ISBN 978-80-01-04228-1

[57] Liška, J.: *Zpracování signálů pro diagnostiku a jeho aplikace.* [online] VUT v Brně, Brno, 2010. [cit. 14.2.201]
Dostupný na www: <http://www.crr.vutbr.cz/system/files/brozura__08__1012.pdf>

[58] Vejražka, F.: Signály a soustavy. ČVUT, Praha, 1995.

[59] Cooley, J. W., Tukey, J. W.: *An algorithm for the machine calculation of complex Fourier series.* Math. Comput. 19: 297−301, 1965.

[60] Becerra M.A., Orrego D.A., Mejia C., Delgado-Trejos E.: *Stochastic Analysis and Classification of 4-Area Cardiac Auscultation Signals Using Empirical Mode Decomposition and Acoustic Features.* In: Computing in Cardiology. New York, IEEE, 2012. ISBN 978-1-4673-2077-1.

[61] Hofreiter, M.: *Bayesovská identifikace technologických procesů.* Habilitační práce. ČVUT v Praze, Praha, 1998.

7.2 Vlastní publikace

[A1] Trnka, P., Hofreiter, M., Sova, J.
Combination of techniques for the fault diagnostics.
In: Proceedings of the 2017 18th International Carpathian Control Conference (ICCC). 18th International Carpathian Control Conference (ICCC), Sinaia. New York: IEEE, 2017, s. 499-502. ISBN 978-1-5090-5825-9.

[A2] Trnka, P.
Modified fault diagnosis system.
In: Konference Studentské Tvůrčí Činnosti STČ 2016, sborník konference. Praha: Fakulta strojní, ČVUT v Praze, 2016, jednací sekce D3. ISBN 978-80-01-05929-6.

[A3] Hofreiter, M., Trnka, P.
Analysis of Temperature Time Series Measured in Ecosystems. [online].
In: Journal of Trends in the Development of Machinery and Associated Technology. 2012, 16(1), s. 147-150. ISSN 2303-4009. Dostupné z: <http://tmt.unze.ba/zbornik/TMT2012Journal/32.pdf>

[A4] Trnka, P., Hofreiter, M.
The Empirical Mode Decomposition in Real-Time.
In: Proceedings of the 18th International Conference on Process Control. Bratislava: Slovenská technická univerzita, 2011, p. 284-289.
ISBN 978-80-227-3517-9.

[A5] Trnka, P., Hofreiter, M.
On-line empririckà modální dekompozice.
In: Sborník odborného semináře Nové metody a postupy v oblasti přístrojové techniky, automatického řízení a informatiky 2011. Praha: Ústav přístrojové a řídicí techniky FS ČVUT, 2011, s. 93-96. ISBN 978-80-01-05041-5.

[A6] Hofreiter, M., Trnka, P.
On-line Empirical Mode Decomposition of Environmental Data.
In: Journal Of Trends In The Development Of Machinery And Associated Technology, 15th International Research/Expert Conference TMT 2011. Dubai: UAE, 2011, roč. 15, s. 825-828, ISSN 1840-4944.

[A7] Trnka, P., Hofreiter, M.
Empirická modální dekompozice environmentálních časových řad v reálném čase.
In: Automatizácia a riadenie v teórii a praxi 2011. Košice: Technická univerzita v Košiciach, 2011, s. 52-1-52-9. ISBN 978-80-553-0606-3.

[A8] Trnka, P., Hofreiter, M.
Empirická modální dekompozice environmentálních časových řad v reálném čase.
In: Strojárstvo. 2011, roč. XV., č. 5, s. 10/1-10/4. ISSN 1335-2938.

[A9] Hofreiter, M., Trnka, P.
Fault Diagnosis for Nonlinear Stochastic Dynamic Systems.
In: Proceedings of The 9th International Conference Process Control 2010. Pardubice: Universita Pardubice, 2010, p. C009b - 1-9. ISBN 978-80-7399-951-3.

[A10] Trnka, P.
Diagnostika poruch neurčitých systémů.
Diplomová práce. Praha, 2002, České vysoké učení technické v Praze, Fakulta strojní. Vedoucí práce M. Hofreiter.

[A11] Hofreiter, M. - Trnka, P.
Assessment of Evapotranspiration in Ecosystems.
In: International Journal of Geology [online]. 2013, vol. 7, no. 2, p. 58-62. Internet: http://www.naun.org/main/NAUN/geology/b032004-106.pdf. ISSN 1998-4499.

[A12] Hofreiter, M. - Trnka, P.
Estimation of Evapotranspiration from Measured Climatic Data.
In: Proceedings of the International Conferences. Athens: World Scientific and Engineering Academy and Society, 2013, art. no. ENVIR-11, p. 72-75. ISBN 978-960-474-315-5.

[A13] Hofreiter, M. - Trnka, P.
Estimation of Evapotranspiration from Measured Climatic Data.
In: Recent Advances in Energy and Environmental Management. Athens: WSEAS Press, 2013, p. 72-75. ISSN 2227-4359. ISBN 978-960-474-312-4.

[A14] Hofreiter, M. - Trnka, P.
Analysis of Temperature Time Series Measured in Ecosystems.
In: 16th International Research/Expert Conference "Trends in the Development of Machinery and Associated Technology" TMT 2012 Proceedings. Barcelona: EA4EPQ, 2012, p. 359-362. ISSN 1840-4944.

[A15] Hofreiter, M. - Trnka, P.
Analysis of Temperature Time Series Measured in Ecosystems.
In: Journal of Trends in the Development of Machinery and Associated Technology [online]. 2012, vol. 16, no. 1, p. 147-150. Internet: http://tmt.unze.ba/zbornik/TMT2012Journal/32.pdf. ISSN 2303-4009.

[A16] Trnka, P.
Balanced Bayes classifier.
In: Sborník odborného semináře Nové metody a postupy v oblasti přístrojové techniky, automatického řízení a informatiky 2018. [v tisku]. Praha: České vysoké učení technické v Praze, 2018.