Modelování chování složitých systémů pro podporu operátorů

Ivan Nagy[a,b], Petr Nedoma[b], Miroslav Kárný[b],
Lenka Pavelková[b], Pavel Ettler[c]

[a] Fakulta dopravní ČVUT,
No Florenci 25, 110 00 Praha 1, nagy@fd.cvut.cz
[b] Ústav teorie informace a automatizace AV ČR,
POB 18, 18208 Praha 8, {school,nedoma}@utia.cas.cz
[c] Compureg Plzeň, s.r.o., POB Plzeň, ettler@compureg.cz

1. července 2002

Abstrakt

Článek se zabývá problematikou modelování složitých systémů pomocí modelů tvořených směsí komponent, odhadováním jejich parametrů. Modely lze využít pro detekci aktivní komponenty a předpověď budoucích hodnot sledovaných veličin. Složitost modelovaného systému je jednak v jeho rozsahu (příkladem je rozsáhlá dopravní síť), jednak v jeho struktuře. Ta se připouští nelineární, popsatelná směsí lineárních modelů, odpovídajících konečnému počtu rozlišitelných stavů systému. Parametry modelu se odhadují na základě bayesovského přístupu. Je použita vhodná aproximace, aby bylo možné uchovávat pouze konečněrozměrnou statistiku odhadu. Teorie je ilustrována výsledky experimentů, které demonstrují široké možnosti využití zmíněných modelů pro popis mnoharozměrných reálných dat.

Klíčová slova

Modelování systémů, stochastické systémy, směsi dynamických komponent, bayesovská statistika, odhadování.

Tento výzkum byl částečně podporován EU grantem IST-1999-12058 a grantem GAČR č. 102/99/1564.

1 Úvod

Motivace

V mnoha případech se setkáváme se situací, kdy je určitý systém řízen zásahy operátora. Jako příklad lze uvést dopravního policistu, který na složité křižovatce přepíná světelná signalizační zařízení [1], operátora, který sleduje kvalitu vyráběného plechového pásu na válcovací stolici [2], případně papíru na papírenském stroji [3], nebo lékaře [4], který analýzou rentgenového snímku rozhoduje o léčbě pacientů. Všechny tyto případy jsou si podobné: na základě zjištěných (měřených) dat je činěno rozhodnutí. To má zlepšit stav systému z hlediska zvoleného kriteria. Operátor využívá kromě měřených dat také svou zkušenost, intuici a odhad. Tyto lidské schopnosti mají velikou váhu a způsobují, že zásahy zkušeného operátora jsou velmi dobré. Nevýhodou operátora, jako lidské bytosti, je to, že často z obrovského množství měřených dat dokáže sledovat a využít jen malou část. Uvádí se, že člověk dokáže sledovat nejvýše pět veličin. Tento počet také například odpovídá počtu linek v notové osnově. Na druhé straně, počítač s implementovanými algoritmy pro zpracování informace postrádá lidskou intuici, ale dokáže sledovat a zpracovávat téměř všechnu měřenou informaci. A právě tuto schopnost je žádoucí využít a podpořit pomocí ní činnost operátorů.

Zmíněnou podporou činnosti operátorů se zabývá EU projekt ProDaC-Tool [5], v jehož rámci je vyvíjen poradní systém pro operátory. Tento systém je založen na odhadu pravděpodobnostního shlukového modelu mnoharozměrných dat, měřených na uzavřené smyčce systém – operátor a obsahujících jak veličiny měřené na systému, tak i akční zásahy operátora. Datové shluky představují jednotlivé pracovní oblasti uzavřené smyčky systém – operátor, tj. pracovní režim systému a příslušnou akci operátora. Akce operátora mohou být dodatečně vyhodnoceny dosaženou kvalitou činnosti systému, a tím lze získat ohodnocení jednotlivých shluků. Tak dostáváme odhady "dobrých" a "špatných" pracovních oblastí. Průběžně odhadujeme, ve které oblasti se systém nachází. O tomto stavu buď operátora jen informujeme, nebo mu nabízíme směr, kterým se systém nejlépe dostane do jiné, lepší oblasti. Podrobnější informace o poradním systému je možno získat v [6].

Zde se budeme zabývat jen modelováním mnoharozměrných reálných dat a odhadem jejich modelu, který je prvním a velmi důležitým krokem směrem ke konstrukci zmíněného poradního systému. Kvalitu tohoto odhadu budeme poměřovat schopností modelu předpovídat budoucí hodnoty modelovaných veličin.

Prvá část článku je věnována nástinu teorie naznačeného způsobu modelování mnoharozměrných dat a odhadu tohoto modelu z datového vzorku. Ve druhé části článku jsou uvedeny výsledky odhadů na reálných datech. Pro experimenty jsou použita dopravní data a data měřená na válcovací stolici.

2 Teoretické základy pro model a jeho odhad

Model mnoharozměrných vícemodálních dat

V tomto článku se budeme se zabývat modely diskrétními v čase a spojitými v hodnotách. To znamená, že spojité měřené veličiny jsou v čase vzorkovány s určitou periodou a v modelu vystupují posloupnosti těchto vzorkovaných hodnot. Indexem t označujeme aktuální časový okamžik; výrazy $t-1$, $t-2$ atd. označují minulé časové okamžiky a hodnoty veličin, získané v těchto okamžicích, budeme nazývat historická data.

Protože modelujeme veličiny pod vlivem neurčitostí, nemůžeme od modelu požadovat popis modelované veličiny ve formě hodnot, ale jen jako její pravděpodobnostní rozdělení vyjádřené pomocí podmíněné hustoty pravděpodobnosti. V podmínce této hustoty jsou právě ty hodnoty historických veličin, na kterých rozdělení aktuální hodnoty modelované veličiny závisí, a parametry, které vyjadřují korespondenci modelu s modelovanými veličinami.

Model popisuje mnoharozměrnou modelovanou veličinu, jejíž hodnotu změřenou v čase t označíme

$$d_t = [d_{1;t}, d_{2;t}, \ldots, d_{\mathring{d};t}]', \quad t = 1, 2, \ldots, \mathring{t},$$

kde $\mathring{d}$ je dimenze modelované veličiny a $\mathring{t}$ je počet časových okamžiků, ve kterých je modelovaná veličina měřena (délka datového vzorku) a symbol $'$ je transpozice, která formálně označuje uvedený vektor jako sloupcový.

Posloupnost všech měřených hodnot veličiny d, až do časového okamžiku t budeme značit

$$d(t) = [d_1, d_2, \ldots, d_t]'.$$

Hodnota modelované veličiny d_t je chápána jako jeden bod v $\mathring{d}$ rozměrném vektorovém prostoru, který nazveme prostorem dat. Množina všech naměřených hodnot modelované veličiny $d(\mathring{t})$ tedy představuje množinu bodů v datovém prostoru. Pokud má modelovaný systém několik různých stavů, mezi nimiž přechází, mají měřené hodnoty modelované veličiny tendenci ke shlukování – každý shluk odráží určitý stav systému. Tyto shluky vyjadřují pravděpodobnostní vazby mezi složkami modelované veličiny. Vazby v jednotlivých stavech se předpokládají lineární, avšak přechody mezi stavy a jejich vazby mohou být nelineární.

Popis modelované veličiny je dán popisem shluků v prostoru dat pomocí podmíněné hustoty pravděpodobnosti

$$f(d_t|d(t-1)) = f(d_{1;t}, d_{2;t}, \ldots, d_{\mathring{d};t}|d(t-1)),$$

popisující rozdělení modelované veličiny v čase t v závislosti na jejích historických hodnotách.

Takový popis systému však většinou není přímo dostupný, protože systém neznáme. Proto zavádíme parametry modelu Θ, s jejichž pomocí vytvoříme model v tvaru

$$f(d_t|d(t-1), \Theta). \tag{1}$$

Neznámé parametry Θ je třeba odhadovat na základě měřených dat.

Cílem je modelovat shluky bodů, které reprezentují měřená data a vypovídají o stavech sledovaného systému. Shluk bodů je v každém místě dán určitou hustotou bodů, tedy z pravděpodobnostního hlediska určitou pravděpodobností výskytu bodu, tj. datového údaje.

Protože očekáváme, že shluků bude více, musíme uvažovat také několik modelů – viz např. [7, 8]. Shluky a jim odpovídající modely budeme označovat přirozenými čísly $c = 1, 2, \ldots, \mathring{c}$, kde $\mathring{c}$ označuje celkový počet modelů – komponent. Jednotlivé komponenty mají tvar

$$f(d_t|d(t-1), \Theta, c) = f_c(d_t|d(t-1), \Theta), \qquad c = 1, 2, \ldots, \mathring{c}. \tag{2}$$

Výchozí představa o složeném modelu tvořeném více komponentami je taková, že v každém časovém okamžiku je aktivní právě jedna komponenta. To odpovídá představě, že datové shluky vypovídají o stavech systému. V každém okamžiku je systém v jednom určitém stavu, jemuž odpovídá aktivní komponenta složeného modelu.

Z uvedeného je zřejmé, že pro popis složeného modelu nestačí definovat jen množinu komponent. Je třeba uvažovat ještě jednu veličinu, označující v každém časovém okamžiku aktivní komponentu. Tuto veličinu v časovém okamžiku t označíme symbolem

$$c_t \in \{1, 2, \ldots, \mathring{c}\}. \tag{3}$$

Obecně je třeba veličinu c_t považovat za neznámou a neměřenou.

Složený model je potom vyjádřen pomocí hustoty pravděpodobnosti

$$f(d_t, c_t | d(t-1), \Theta, \alpha), \qquad (4)$$

kde α je vektor parametrů, dále používaný pro modelování veličiny c_t.

Sdruženou hustotu pravděpodobnosti (4) lze podle řetězového pravidla rozložit následovně

$$f(d_t | d(t-1), c_t, \Theta, \alpha) f(c_t | d(t-1), \Theta, \alpha).$$

První činitel modeluje data v rámci komponenty c_t, druhý je modelem pro přepínání komponent. Tyto modely lze ještě zjednodušit pomocí následujících, dále přijímaných, podmínek:

Pro první hustotu pravděpodobnosti předpokládáme, že data d_t nezávisí na parametru α a závisí maximálně na n historických hodnotách měřených veličin, které ukládáme do regresního vektoru $\varphi_{t-1} = [d'_{t-1}, \ldots, d'_{t-n}, 1]'$. Se zavedeným označením pro model komponenty (2) můžeme tuto hustotu pravděpodobnosti vyjádřit ve tvaru

$$f(d_t | d(t-1), c_t, \Theta, \alpha) = f_{c_t}(d_t | \varphi_{t-1}, \Theta). \qquad (5)$$

Pro druhou hustotu pravděpodobnosti předpokládáme nezávislost c_t na datech $d(t-1)$ a parametrech Θ a vyjádříme ji přímo jako složku vektoru α

$$f(c_t | d(t-1), \Theta, \alpha) = f(c_t | \alpha) = \alpha_{c_t}. \qquad (6)$$

Po dosazení těchto předpokladů do složeného modelu dostáváme

$$f(d_t, c_t | d(t-1), \Theta, \alpha) = f_{c_t}(d_t | \varphi_{t-1}, \Theta) \alpha_{c_t}.$$

Tento popis současně modeluje jak data d_t, tak i veličinu c_t. Jeho rozklad říká, že popis dat d_t pomocí komponenty s indexem c_t získáme tak, že jej modelujeme pomocí aktivní komponenty c_t a násobíme ji pravděpodobností α_{c_t}, příslušnou této komponentě. Protože takový model závisí na neznámé hodnotě veličiny c_t, není přímo použitelný.

Model, který lze formálně využít, nesmí obsahovat veličinou c_t. Dostaneme jej jako marginální hustotu pravděpodobnosti

$$f(d_t | d(t-1), \Theta, \alpha) = \sum_{c=1}^{\mathring{c}} f(d_t, c | d(t-1), \Theta, \alpha) = \sum_{c=1}^{\mathring{c}} f_c(d_t | \varphi_{t-1}, \Theta) \alpha_c. \qquad (7)$$

Toto je však konečná směs hustot pravděpodobnosti, pro níž přesné odhadování parametrů není prakticky proveditelné.

Aproximace modelu

Řádný výpočet marginální pravděpodobnosti (7) vede k modelu, jehož parametry nelze korektně odhadovat. Spokojíme se proto s modelem přibližným, s prakticky odhadovatelnou parametrizací.

Složený model formálně vyjádříme pomocí Kroneckerovy funkce

$$\delta(c, c_t) = 1 \text{ pro } c = c_t; \quad \delta(c, c_t) = 0 \text{ pro } c \neq c_t$$

ve tvaru

$$f(d_t, c_t | d(t-1), \Theta, \alpha) = \prod_{c=1}^{\mathring{c}} \left[f_c(d_t | \varphi_{t-1}, \Theta) \alpha_c \right]^{\delta(c, c_t)}. \tag{8}$$

Tímto obratem se soustředila neznámá veličina c_t na jediné místo, a to do argumentu Kroneckerovy funkce v exponentu. Funkci $\delta(c, c_t)$ s neznámou hodnotou vyloučíme tím, že ji nahradíme její podmíněnou střední hodnotou, která je jejím bodovým odhadem

$$w_{c,t} = E[\delta(c, c_t) | d(t)] = \sum_{c=1}^{\mathring{c}} \delta(c, c_t) f(c_t | d(t)) = f(c_t = c | d(t)), \tag{9}$$

pro $c = 1, 2, \ldots, \mathring{c}$ a $t = 1, 2, \ldots, \mathring{t}$.

Potřebnou hustotu pravděpodobnosti $f(c_t = c | d(t))$ získáme marginalizací a užitím Bayesova vzorce

$$f(c_t = c | d(t)) \propto \int_{\Theta^*} \int_{\alpha^*} \alpha_c f_c(d_t | \varphi_{t-1}, \Theta) f(\Theta, \alpha | d(t-1)) d\alpha d\Theta,$$

kde

- Θ^* a α^* jsou obory hodnot veličin Θ a α,
- $\propto$ označuje úměrnost s tím, že koeficient úměrnosti plyne z normalizace hustoty pravděpodobnosti,
- $f(c_t = c | d(t))$ vyjadřuje pravděpodobnost, že $c_t = c$ za podmínky, že známe data $d(t)$,
- $f(\Theta, \alpha | d(t-1))$ popisuje neznámé parametry složeného modelu při znalosti dat $d(t-1)$.

POZNÁMKA: *Praktický význam aproximace je možno vyjádřit takto. Kroneckerovu funkci lze chápat jako hustotu pravděpodobnosti náhodné veličiny c_t v případě, že známe, kdy je která komponenta aktivní. Tato funkce* *přiřazuje jednotkovou pravděpodobnost aktivní komponentě, pravděpodobnosti ostatních komponent jsou nulové. Protože ale nevíme, kdy je která komponenta aktivní, tj. hodnoty veličiny c_t neznáme, nahradíme přibližně Kroneckerovu funkci v každém okamžiku vektorem $w_t = [w_{1,t}, w_{2,t}, \ldots, w_{\mathring{c},t}]'$. Tento vektor, na základě informace z dat až do současného okamžiku, přiřazuje každé komponentě obecně nenulovou pravděpodobnost její aktivity. Aproximace tedy "rozštěpí" Kroneckerovu funkci, přesně ukazující na aktivní komponentu. Tak vezme v každém časovém okamžiku v úvahu všechny komponenty, každou s pravděpodobností její aktivity.*

Odhadování parametrů

Pro práci se složeným modelem je třeba znát odhady jeho parametrů, vyjádřené pomocí hustoty pravděpodobnosti $f(\Theta, \alpha|d(t))$. Tuto hustotu pravděpodobnosti lze počítat rekurzivně, počínaje apriorní hustotou pravděpodobnosti $f(\Theta, \alpha|d(0))$ pomocí Bayesova vztahu

$$f(\Theta, \alpha|d(t)) \propto f(d_t|d(t-1), \Theta, \alpha) f(\Theta, \alpha|d(t-1)). \tag{10}$$

Ten přepočítává hustotu pravděpodobnosti parametrů z předchozího kroku na aposteriorní hustotu pravděpodobnosti s pomocí (7), kterou ovšem uvažujeme ve tvaru (8) a v aproximaci (9). Pro normální komponenty patřící do exponenciální třídy, viz odstavec 3, vede toto odhadování na algoritmus vážených nejmenších čtverců, s vahami dat w_t určených podle (9). Bayesův vzorec lze obohatit o zapomínání [9],[10], které dovolí sledovat pomalé změny parametrů jednotlivých komponent i parametru α, a tím rozšířit použitelnost studovaného modelu.

3 Odhadování komponent exponenciální třídy

Jednotlivé kroky odhadování, naznačené v předchozí části, nyní shrneme do algoritmu.

Algoritmus odhadování složeného modelu

I. Inicializace

Volba struktury složeného modelu
- volba počtu komponent
- volba struktur modelů komponent

Konstrukce statistik pro apriorní hustotu pravděpodobnosti

Volba koeficientu zapomínání

II. Zpracování měřených dat (časová smyčka)

Přepočet apriorní hustoty pravděpodobnosti na aposteriorní

III. Závěrečné výpočty

Výpočet bodových odhadů parametrů z aposteriorní hustoty pravděpodobnosti

Algoritmus dále rozvedeme pro uvažovaný případ, kdy komponenty mají rozdělení z exponenciální třídy, tj. kdy platí

$$f_c(d_t|\varphi_{t-1},\Theta) = \exp\left\{b'_c(\Theta)s_c(d_t,\varphi_{t-1})\right\}, \quad \text{pro} \quad c = 1,2,\ldots,\mathring{c}, \tag{11}$$

kde

$b_c(\Theta)$ je vícerozměrná parametrická funkce c-té komponenty,
$s_c(d_t,\varphi_{t-1})$ je vícerozměrná statistika c-té komponenty modelu.

Odložíme-li otázku volby počtu komponent na později, můžeme zvolit reprodukující se tvar apriorní hustoty pravděpodobnosti pro zvolené rozdělení komponent a pro jejich váhy α. To má tvar

$$f(\Theta,\alpha|d(0)) \propto \prod_{c=1}^{\mathring{c}} \alpha_c^{\kappa_c(0)-1} \exp\left\{b'_c(\Theta)S_c(0)\right\}, \tag{12}$$

přičemž apriorní statistiky odhadu $\kappa_c(0)$, $S_c(0)$ umožňují respektovat dostupnou předběžnou informaci o odhadovaných parametrech.

Zpřesňování apriorní hustoty pravděpodobnosti na aposteriorní probíhá v každém časovém okamžiku $t = 1,2,\ldots,\mathring{t}$ podle Bayesova vzorce (10), kde za hustotu pravděpodobnosti parametrů dosadíme hustoty ve tvaru (12). Složený model vyjádříme v součinovém tvaru podle (8), přičemž za hustotu pravděpodobnosti dosadíme její exponenciální vyjádření (11). Bayesův vztah má potom tvar

$$\prod_{c=1}^{\mathring{c}} \exp\left\{b_c'(\Theta) S_c(t)\right\} \alpha_c^{\kappa_c(t)-1} \propto$$

$$\propto \prod_{c=1}^{\mathring{c}} \left[\exp\left\{b_c'(\Theta) s_c(d_t, \varphi_{t-1})\right\} \alpha_c\right]^{\delta(c,c_t)} \prod_{c=1}^{\mathring{c}} \exp\left\{b_c'(\Theta) S_c(t-1)\right\} \alpha_c^{\kappa_c(t-1)-1}.$$

Porovnáním exponentů u příslušných základů pravé a levé strany v předchozím vztahu dostáváme předpis pro přepočet statistik v hustotách pravděpodobnosti, majících neměnný funkční tvar. Předpis má jednoduchý tvar

$$S_c(t) = S_c(t-1) + \delta(c, c_t) s_c(d_t, \varphi_{t-1}), \quad (13)$$

$$\kappa_c(t) = \kappa_c(t-1) + \delta(c, c_t). \quad (14)$$

Odpovídající pravděpodobnost $f(c_t = c|d(t))$ určuje váhy $w_{c,t}$ (9)

$$w_{c,t} \propto \kappa_c(t) \frac{\int \exp\left\{b_c'(\Theta) S_c(t)\right\} \, d\Theta}{\int \exp\left\{b_c'(\Theta) S_c(t-1)\right\} \, d\Theta}. \quad (15)$$

Stojí za povšimnutí, že vyčíslení těchto vah vyžaduje výpočet integrálu jediného typu, který lze v prakticky zajímavých případech určit v analytickém tvaru.

Přepočet statistik podle (13), (14) nelze provést, protože neznáme argument c_t Kroneckerovy funkce δ. Použití diskutované aproximace vede na následující přepočet:

1. Statistiku každé komponenty přepočteme tak, jako by právě ona byla aktivní, tj. do (13), (14) dosadíme $\delta(c, c_t) \approx 1$.

2. Z takto pokusně přepočtených statistik vypočteme váhy $w_c(t)$ podle vztahu (15).

3. Provedeme skutečný přepočet statistik podle vztahů (13), (14), v nichž uplatníme přibližnou rovnost $\delta(c, c_t) \approx w_c(t)$.

Bodové odhady parametrů, pokud je potřebujeme znát, určíme jako podmíněné střední hodnoty

$$[\hat{\Theta}_t, \hat{\alpha}_t] = E[(\Theta, \alpha)|d(t)]. \quad (16)$$

4 Inicializace algoritmu odhadování

Cíle inicializace

Podle uvedeného algoritmu je třeba před započetím vlastního odhadování, tj. před vlastním zpracováním měřených dat, zvolit apriorní hustotu pravděpodobnosti parametrů složeného modelu. To předpokládá zvolit:

- vhodný počet komponent složeného modelu zajišťující dostatečné rozlišení všech shluků obsažených v měřených datech;
- apriorní hustotu pravděpodobnosti pro parametry jednotlivých komponent.

Určení apriorního rozdělení hustoty pravděpodobnosti parametrů pro obyčejný regresní model a dostatek informativních dat je jednoduché. Například lze rovnoměrným rozdělením simulovat počáteční neznalost parametrů. Naproti tomu kombinace obou uvedených podúloh inicializace složeného modelu je, i za stejně výhodné situace, úlohou poměrně složitou a přitom pro úspěch odhadování velmi důležitou.

Hlavní problém, který se ve spojitosti s inicializací objevuje, je následující. Jestliže je některá z komponent na počátku umístěna, nebo se během odhadování dostane do oblasti, kde nejsou žádná data, bude nadále její váha vzhledem k datům nulová a komponenta se dále již nebude upravovat. Nemá tak šanci se z této nevýhodné pozice dostat.

Štěpení komponent

Používaná metoda inicializace se opírá o postupné štěpení komponent a opakované použití buď apriorních dat nebo datového vzorku, určeného pro odhadování složeného modelu. Startuje s jedinou komponentou a použije všechna dostupná data pro její odhad. Výsledkem je jediná aposteriorní hustota pravděpodobnosti (pro lepší představu budeme uvažovat dvojrozměrnou normální hustotu pravděpodobnosti). Zjistíme, ve kterém směru je více protažena, tj. kde je podezření, že překrývá dva nebo více datových shluků, a napříč tohoto směru ji rozdělíme na dvě hustoty pravděpodobnosti. Opět provedeme na stejných datech odhadování. Zjistíme, která z obou komponent obsahuje více informace, a tuto komponentu znovu rozdělíme napříč směru, ve kterém je více protažena.

Uvedený postup opakujeme, pokud se zlepšuje kvalita modelu, nebo dokud není vyčerpán povolený počet iterací. Jako kriterium pro posouzení kvality odhadování se používá hodnota tzv. logaritmické věrohodnostní funkce variant (LVFV). To je hodnota věrohodnostní funkce pro daný model, použitý datový vzorek a nikoliv bodový, ale úplný pravděpodobnostní odhad parametrů daný aposteriorní hustotou pravděpodobnosti. Hodnota tohoto kritéria měří stupeň shody dané varianty odhadnutého modelu s použitým datovým vzorkem. Pod pojmem varianta modelu si lze představit např. řád modelu, použití nebo zakázání konstanty modelu a pod.

Zapomínání při opakovaném odhadování

Použití stejných dat pro opakované odhadování, ke kterému dochází při štěpení, je spojeno s dalším problémem. Jak známo, po zpracování datového vzorku a přepočtu apriorní hustoty pravděpodobnosti na aposteriorní, dojde k "utažení", tj. zmenšení šířky (kovariance) této hustoty pravděpodobnosti. Parametry této komponenty jsou fixovány a další užití stejných dat pak již nemá pro odhadování žádný efekt. Proto je na aposteriorní hustotu pravděpodobnosti z předchozího kroku štěpení nutno použít zapomínání, aby mohla být v dalším kroku dále použita jako apriorní. Požadavek zachování efektivního počtu zpracovaných dat určuje vhodnou míru zapomínání. Zapomínají se charakteristiky spojené s popisem neurčitosti hustoty pravděpodobnosti (kovariance, stupně volnosti). Charakteristiky polohy komponent (střední hodnoty) se při zapomínání zachovávají.

5 Příklady

Popsanou teorii modelování mnoharozměrných vícemodálních dat pomocí složeného modelu a odhad tohoto modelu budeme ilustrovat na dvou příkladech. První je odhadování modelu intenzity a hustoty dopravního proudu na jednom místě dopravní oblasti. Uvedené veličiny vystihují stav dopravy na tomto místě. Druhý příklad ukazuje odhadování modelu mnoharozměrných dat, měřených na válcovací stolici.

Cílem příkladů je demonstrovat schopnost složeného modelu detekovat shluky v prostoru dat. Ověření kvality této detekce není v případě mnoharozměrných dat triviální problém. V příkladech jsou uvedeny některé možnosti takové kontroly, opírající se o hodnotu kriteria LVFV (čím větší hodnota, tím lepší korespondence modelu s datovým vzorkem) a charakteristiky chyb predikce (požadujeme nulovou střední hodnotu a minimální rozptyl).

5.1 Model s dvourozměrnými dopravními daty

Popis systému

Uvažujeme dopravní komunikaci s vybraným místem, na kterém sledujeme dopravní proud. Základními charakteristikami dopravního proudu, určujícími stav dopravy, jsou *intenzita* q (počet vozidel / čas), *hustota* ρ (počet vozidel / dráhu) a *rychlost* v (dráha / čas). Mezi těmito veličinami platí zřejmý vztah $q = v.\rho$, tedy jedna z těchto veličin je nadbytečná a může být určena z ostatních. Zbylé dvě veličiny jsou ale pro jednoznačné určení stavu dopravního proudu potřebné. Například intenzita je nulová, jsou-li silnice prázdné, ale také když jsou přeplněné a kolony vozidel stojí. Pro správné určení stavu dopravního proudu je tedy zapotřebí sledovat např. dvojici [hustota, intenzita] v dvourozměrném prostoru dat.

Ideální průběh funkce $q = F(\rho)$ má přibližný tvar obrácené paraboly, vycházející z počátku (žádný provoz), postupně rostoucí k maximu (běžný provoz), obracející se v klesající funkci (velký provoz) a končící opět na ose ρ v bodě s maximální hustotou vozidel ve stojící koloně (dopravní kolaps). Běžně měřené body $[\rho, q]$ leží v okolí této křivky a jejich shluky označují určité stavy dopravního proudu, například výše jmenované.

Podle stavu dopravy v oblasti operátor přepíná tzv. řídící plány, tj. určité nastavení doby zelené a červené na světelných signalizačních zařízeních v řízené oblasti. Dobré poznání stavu v jednom místě je základem pro určení stavu celé oblasti, a tedy i jejího řízení.

Již samotné rozpoznání skutečného stavu dopravy, bez ohledu na šum obsažený v měřených hodnotách hustoty a intenzity dopravního proudu a předpověď jeho budoucích hodnot, jsou velmi potřebné a mohou být podporou pro známá dopravní hlášení o stupni provozu v určité oblasti.

Aplikace teorie

Pro popis prostoru dat, tvořeného dvojicemi $d_t = [\rho_t, q_t]'$, kde ρ_t a q_t jsou hustota a intenzita dopravního proudu v časovém okamžiku t, volíme složený model se statickými modely komponent

$$f_c(d_t|\theta_c) = f_c(\rho_t, q_t|\theta_{1c}, \theta_{2c}), \tag{17}$$

kde

$\theta_c = [\theta_{1c}, \theta_{2c}]'$ jsou konstanty určující střední hodnotu měřených dat v rámci c-té komponenty;

$c = 1, 2, \ldots, \mathring{c}$ odlišuje komponenty.

Jedna komponenta modelu je popsána dvourozměrnou hustotou pravděpodobnosti s maximem ve střední hodnotě $\theta_c = [\theta_{1c}, \theta_{2c}]'$ a s šířkou a tvarem vrstevnic určenými kovarianční maticí R_c

$$f_c(d_t|\Theta_c) \propto |R_c|^{-0.5} \exp\left\{-0.5(d_t - \theta_c)' R_c^{-1}(d_t - \theta_c)\right\}, \tag{18}$$

kde

$\Theta_c = [\theta_c, R_c]$ jsou neznámé parametry komponenty.

Uvažovaná hustota pravděpodobnosti patří do exponenciální třídy, a tedy přepočet statistik, na který se redukuje odhadování, probíhá podle rovnic (13) a (14). Pro tento speciální tvar modelu je můžeme upravit do tvaru

$$V_c(t) = V_c(t-1) + \delta(c, c_t) h_t, \tag{19}$$

$$\nu_c(t) = \nu_c(t-1) + \delta(c, c_t), \tag{20}$$

$$\kappa_c(t) = \kappa_c(t-1) + \delta(c, c_t), \tag{21}$$

kde

$c = 1, 2, \ldots, \mathring{c}$ a $t = 1, 2, \ldots, \mathring{t}$,

$V_c(t)$, $\nu_c(t)$ tvoří statistiku $S_c(t)$ c-té komponenty a rovnice (19), (20) jsou rozepsanou formou přepočtu (13),

$V_c(0)$, $\nu_c(0)$, $\kappa_c(0)$ jsou apriorní hodnoty statistik,

$h_t = [\rho_t, q_t, 1]'[\rho_t, q_t, 1]$ je datová matice, obsahující data změřená v okamžiku t a tvořící spolu s konstantou 1 statistiku $s_c(d_t, 1)$ v exponenciální třídě (11),

$\kappa_c(t)$ je statistika modelu přepínání komponent, vyvíjející se podle (14).

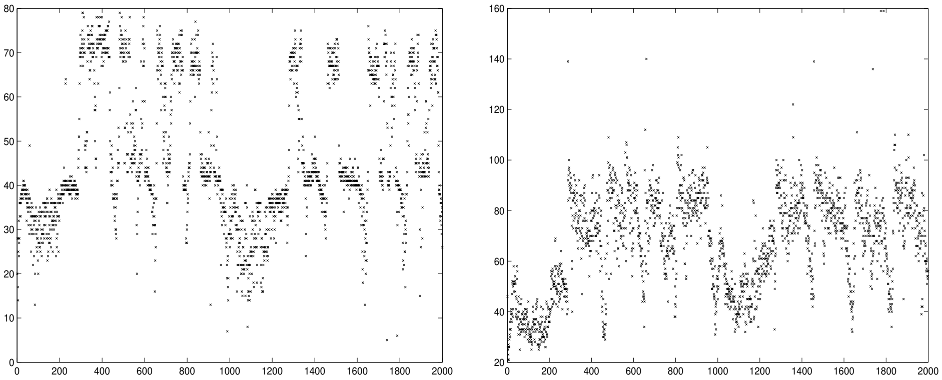
Váhy $w_c(t)$ použité v uvažované aproximaci se vypočtou podle vztahu (15). V uvažovaném případě má podíl integrálních výrazů tvar Studentova rozložení [10].

Cíle experimentu

Cílem experimentu je ukázat na jednoduchém a přehledném případě, že navrhovaný algoritmus identifikace je schopen se v procesu inicializace sám nastavit, tj. určit vhodný počet komponent a jejich apriorní statistiky tak, aby vlastní identifikace byla úspěšná. Úspěšnost identifikace je v tomto jednoduchém případě posuzována vizuálně. Shluky reálných dat se musí shodovat se shluky dat, generovaných z odhadnutého modelu. Toto kriterium sice nezaručuje absolutní posouzení úspěšnosti identifikace, ale o její kvalitě dosti vypovídá.

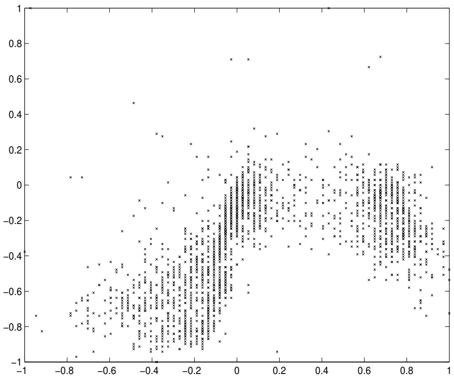
Použitá data

Při experimentu bylo použito 2000 dvojic $[\rho, q]$ získaných v centru Prahy. Tento vzorek obsahuje data měřená po deseti minutách, a tedy zahrnuje souvislé období asi dvou týdnů. Časové průběhy obou signálů a bodový diagram jejich uspořádaných dvojic je uveden na obrázcích 1 a 2.



Průběh hustoty ρ Průběh intenzity q

Obrázek 1: Časové průběhy modelovaných signálů



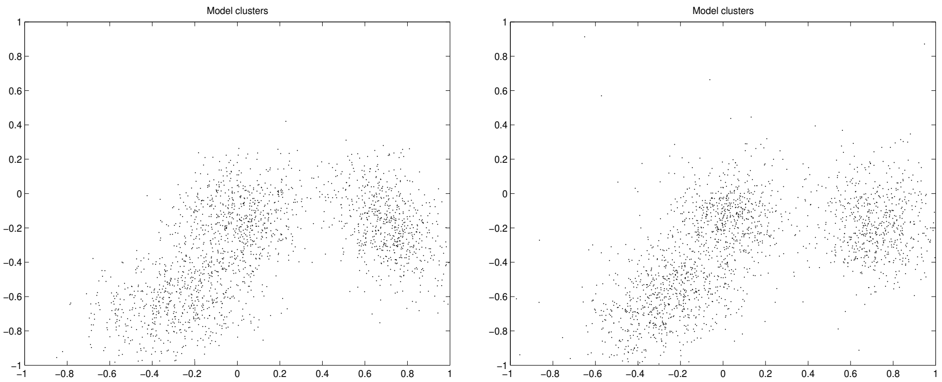
Body $q = F(\rho)$

Obrázek 2: Bodový diagram datových dvojic

Z časových průběhů je dobře patrná periodicita signálů, způsobená střídáním dopravní zátěže během dne a jejím útlumem přes noc, stejně jako střídání pracovních dní a víkendů.

Výsledky experimentu

Na obrázku 2 je uvedeno dvourozměrné zobrazení měřených dat, zobrazených jako body se souřadnicemi $[\rho, q]$, které vytvářejí shluky tvořící přibližně parabolický průběh závislosti dopravní intenzity a hustoty. Odhady shluků jsou na obrázcích č. 3. V levé části je výsledek odhadování se složeným modelem s apriorním nastavením využívajícím veškerou dostupnou informaci, vpravo je model nejprve automaticky inicializován (s výsledkem deseti komponent) a potom teprve odhadován. Výsledky experimentu jsou prakticky stejné, což svědčí o značném úspěchu automatické inicializace. Ruční nastavení modelu je poměrně složité a vyžaduje dosti hluboké znalosti nejen modelovaného procesu, ale i samotné teorie.



Manuální apriorní nastavení | Automatické apriorní nastavení

Obrázek 3: Odhad datových shluků závislosti dopravní intenzity a hustoty

5.2 Model s mnoharozměrnými daty z válcovací stolice

Popis systému

Data byla získána na válcovací stolici, na které se vyrábí plech s velmi vysokými požadavky na přesnost ve tloušťce. Na válcovací stolici je měřeno celkem 46 veličin, z nich 10 bylo označeno za nejvýznamnější vzhledem k ovlivnění kvality (stálosti tloušťky) vyráběného pasu. Operátor sleduje běžící proces a měřené veličiny a na tomto základě přestavuje řízené veličiny tak, aby dosáhl maximální kvality výroby. Přitom je zřejmé, a operátoři to nepopírají, že hlavním vodítkem pro jejich rozhodování je pohled na stroj a jejich zkušenost. Měřená data jsou pro ně jen doplňkovou informací. V každém případě, i kdyby operátor chtěl plně využít veškerou informaci z měřených dat, nebyl by toho schopen pro jejich velký počet.

Aplikace teorie

Pro popis tohoto systému uvažujeme mnoharozměrný normální dynamický model komponenty ve tvaru

$$f_c(d_t|\varphi_{t-1}, \Theta_c) \propto |R_c|^{-0.5} \exp\left\{-0.5(d_t - \theta_c'\varphi_{t-1})' R_c^{-1}(d_t - \theta_c'\varphi_{t-1})\right\}, \quad (22)$$

kde

- c označuje komponentu,
- $d_t = [d_{1;t}, d_{2;t}, \ldots, d_{\mathring{d};t}]'$ je modelovaná $\mathring{d}$rozměrná veličina,
- $\varphi_{t-1} = [d_{t-1}', d_{t-2}', \ldots, d_{t-n}', 1]'$ je regresní vektor modelu řádu n, jednička na konci modeluje konstantní člen,
- Θ_c značí všechny neznámé parametry $c$té komponenty, tj. regresní koeficienty θ_c a kovariance šumu R_c, $c = 1, 2, \ldots, \mathring{c}$.

Přepočet statistik probíhá obdobně jako v předchozím příkladě, ale s datovou maticí

$$h_t = [d_t', \varphi_{t-1}']'[d_t, \varphi_{t-1}],$$

tj. předchozí statický případ získáme pro $\varphi_{t-1} = 1$.

Poznamenejme, že v obou uvažovaných případech jsou přesné přírůstky statistik $s_c(d_t, \varphi_{t-1})$ stejné pro všechny komponenty, a tedy výsledek odhadování stojí i padá s volbou apriorní hustoty pravděpodobnosti. Z tohoto relativně častého případu je patrná důležitost inicializace pro dobré odhadování složeného modelu.

Cíle experimentu

Na rozdíl od předchozího příkladu, kdy jsme byli schopni výsledky experimentu graficky zobrazit, je v tomto případě s vyhodnocením výsledků značná potíž. Vyhodnocení experimentu s mnoharozměrnými daty je samostatný problém, vyžadující důkladné teoretické i programové řešení. Hlavním cílem tohoto experimentu je, kromě demonstrace schopnosti algoritmu odhadovat mnoharozměrný složený model, ukázat některé možnosti zobrazení výsledků z mnoharozměrného datového prostoru a posouzení kvality odhadnutého modelu.

Použitá data

Experimenty byly prováděny s datovými vzorky o rozměru 10. Časové průběhy těchto veličin jsou na obrázku 4. Každý z obrázků obsahuje 25000 vzorků jedné modelované veličiny. Jejich absolutní velikosti nejsou zajímavé, neboť veličiny jsou před použitím normalizovány. Tato normalizace spolu s odstraněním odlehlých pozorování jsou důležitým předpokladem k dalšímu použití dat pro odhad.

Uvedený datový vzorek vznikl spojením datových souborů, měřených pro několik kovových pásů. Dobře patrné skoky v jejich průbězích jsou způsobeny přechody mezi pásy. Pomalé změny dat jsou nejvíce ovlivněny narůstajícím množstvím navíjeného plechu na výstupu stolice.

Veličina 2 je výstupní a je na ní patrná vysoká kvalita válcování. Je to odchylka výsledné tloušťky plechu od její požadované hodnoty. Z obrázku je patrné, že je to prakticky bílý šum. Veličiny 3, 9 a 10 jsou ovlivňovány operátorem a vystupují tedy v roli akčních veličin. Ostatní jsou měřené veličiny, nesoucí informaci o procesu. Podle informací od operátorů a rovněž podle naší předběžné analýzy dat, nejvýznamnější z hlediska informace o kvalitě válcování a jejích kritických stavech (jako např. přetržení pásu), je veličina 8.

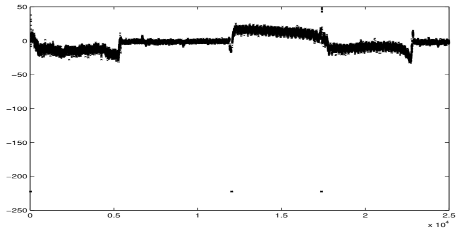
Výsledky experimentu

Budeme demonstrovat dva základní přístupy ke zobrazení výsledků odhadu pro mnoharozměrná data. Prvním z nich je pouhé rozšíření metody použité pro dvourozměrná dopravní data, vizuální posouzení dat a jejich předpovědí na základě odhadnutého modelu.

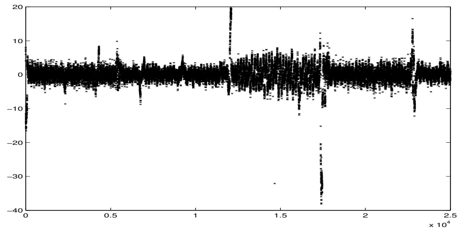
Hlavním měřítkem při posouzení kvality odhadnuté směsi je ale hodnota logaritmické věrohodnostní funkce variant, kterou jsme označili LVFV.

Grafické zobrazení řezů: Protože dvourozměrná data lze graficky zobrazit, nabízí se zobrazovat stavový prostor s vytvořenými shluky ve dvourozměrných řezech. Je však třeba si uvědomit, že v $\mathring{d}$ rozměrném prostoru bude takových řezů $\mathring{d}(\mathring{d}-1)/2$, což pro $\mathring{d}=10$ je čtyřicet pět obrázků. Tyto obrázky je třeba srovnat a pomyslně rekonstruovat shluky v celém základním prostoru. To je skutečně jen orientační pohled.

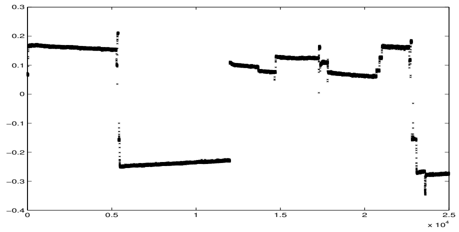
Prakticky jej lze realizovat tak, že srovnáme řezy původního datového prostoru s prostorem simulovaných dat, která získáme generováním z odhadovaného modelu směsi. Obrázek 5 ukazuje dvojice řezů pro první tři proměnné z uvedených veličin. Na obrázcích je patrná zásadní shoda



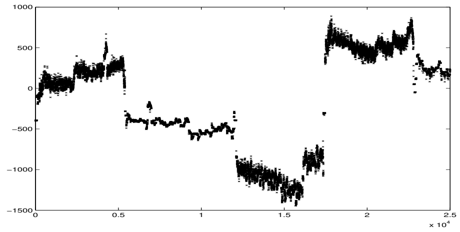
1. Vstupní odchylka tloušťky



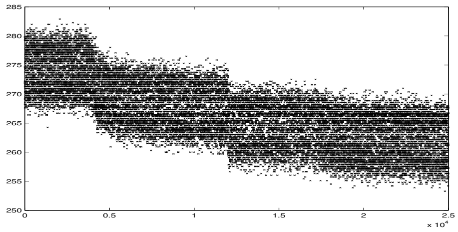
2. Výstupní odchylka tloušťky



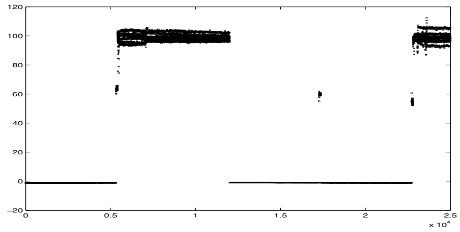
3. Rychlost



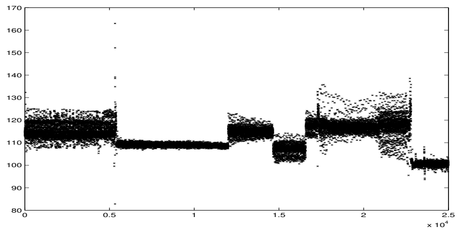
4. Poloha stavění



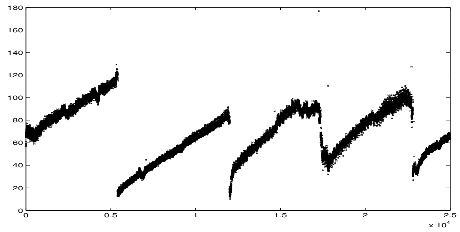
5. Potenciometr stavění



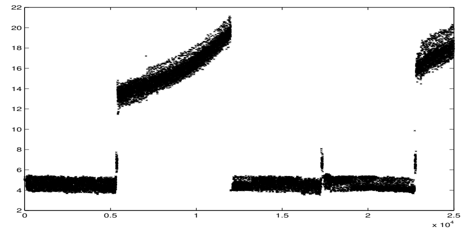
6. Vstupní navíjecí proud



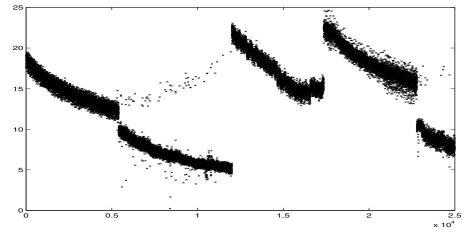
7. Výstupní navíjecí proud



8. Proud pohonu stolice



9. Vstupní tah



10. Výstupní tah

Obrázek 4: Časové průběhy shlukovaných veličin umístění (středních hodnot) jednotlivých datových a simulovaných shluků, vypovídací schopnost o velikosti a tvaru shluků (kovariancí) je již slabší. Navíc, posuzujeme jen prvé tři veličiny, a to po dvojicích. Posouzení všech dvojic by si vyžádalo 45 obrázků a i v tomto případě bychom shodu skutečných desetirozměrných shluků museli posuzovat jen na základě dvourozměrných řezů.

Uvedené srovnání datových a simulovaných shluků je možné jen v případě statického modelu. V povaze dynamických modelů je, že okamžitá pozice shluků je dána minulými daty, a tedy, simulované shluky se uvnitř prostoru dat v čase pohybují. Podobného efektu dosáhneme porovnáním shluků dat a shluků dat předpovídaných z odhadnutého modelu. Příklad tohoto postupu je pro stejná data jako ve statickém případě a pro model 1. řádu ilustrován na obrázku 6.

Z obrázků je patrné, že v tomto případě je korespondence dokonalejší. Zahrnuje jak polohu, tak i tvar odpovídajících si shluků. Zásadní nedostatky tohoto postupu, tj. množství obrázků a vazby pouze dvou veličin, však zůstávají stejné.

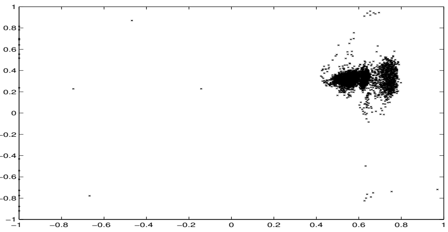
Srovnání podle hodnot LVFV: Toto kritérium hodnotí úspěch odhadu pomocí hodnoty věrohodnostní funkce variant. Přitom je však třeba dbát na skutečnost, že kritérium LVFV je počítáno pro konkrétní datový vzorek, a lze jej tedy použít pouze na srovnání experimentů, při nichž byla použita stejná data. Nejperspektivnějším z celé řady pokusů se zdá být následující postup.

Aplikujeme proces inicializace a odhadování na celý datový vzorek, který máme k dispozici. O výsledku chceme prohlásit, že je to ten nejlepší popis dat, jaký lze získat, a toto tvrzení testujeme. Vyhodnotíme, jak závisí hodnota kriteria LVFV, získaná odhadem na celém datovém vzorku či jeho části na několika různých hodnotách koeficientu alternativního zapomínání [9]. Přitom odhadnutý složený model, který je testován, slouží jako alternativa ve všech případech. Pokud byl testovaný složený model odhadnut skutečně dobře, měli bychom nejlepší výsledek, tj. největší hodnotu věrohodnostní funkce, dostat pro nejnižší hodnotu koeficientu zapomínání.

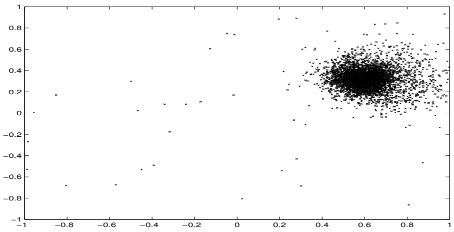
Obrázky 7 a 8 ilustrují tento případ.

Pro posouzení přínosu složeného modelu při popisu reálných dat srovnáme kvalitu odhadu modelu s jednou komponentou, tj. obyčejného regresního modelu, a složeného modelu, s automatickým apriorním nastavením.

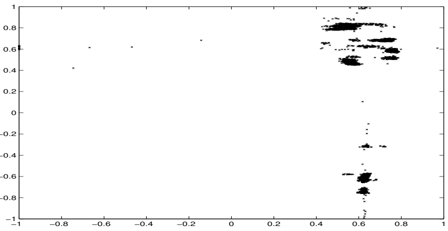
a) *Regresní model*



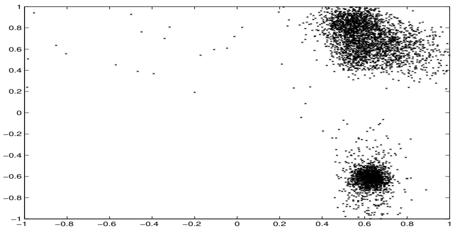
Datové shluky: veličiny 1 a 2



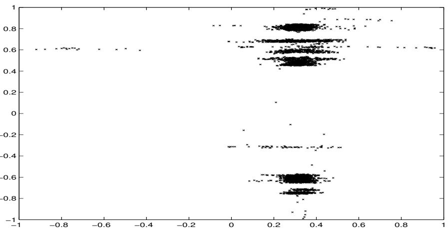
Simulované shluky



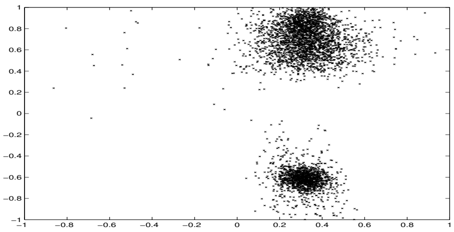
Datové shluky: veličiny 1 a 3



Simulované shluky

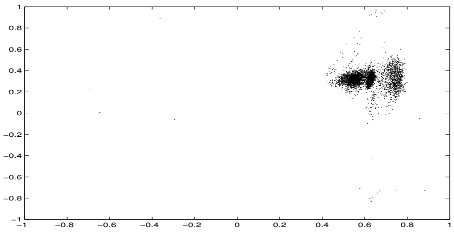


Datové shluky: veličiny 2 a 3

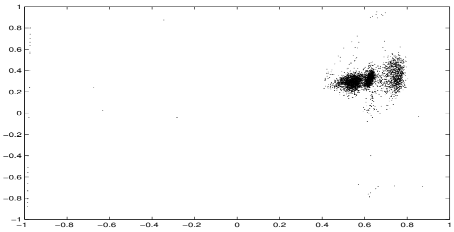


Simulované shluky

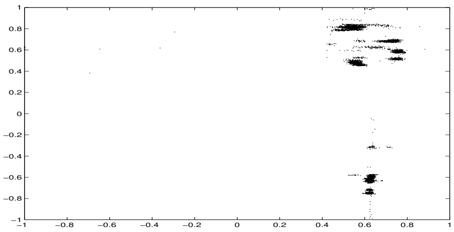
Obrázek 5: Datové a simulované shluky



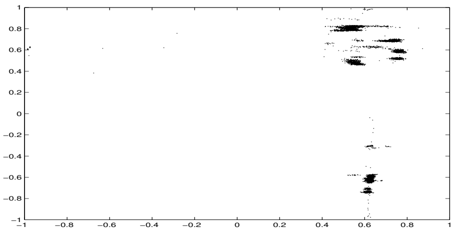
Datové shluky: veličiny 1 a 2



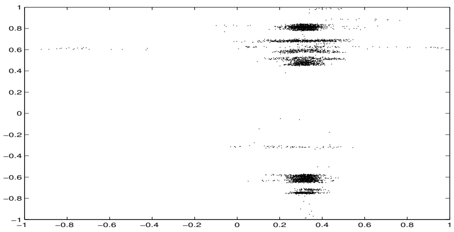
Shluky předpovědí



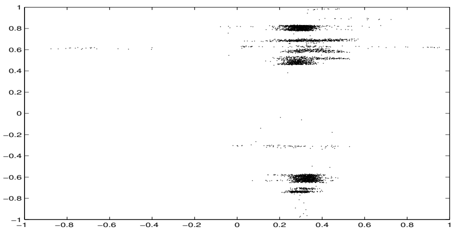
Datové shluky: veličiny 1 a 3



Shluky předpovědí

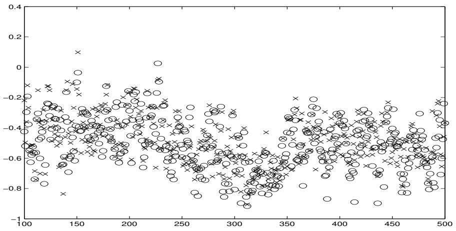


Datové shluky: veličiny 2 a 3

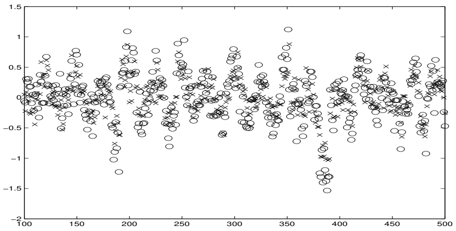


Shluky předpovědí

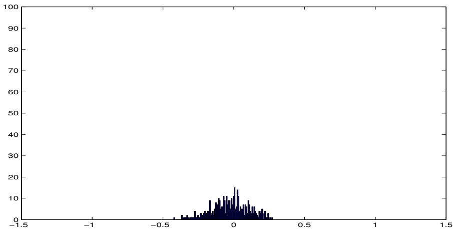
Obrázek 6: Datové shluky a shluky předpovědí



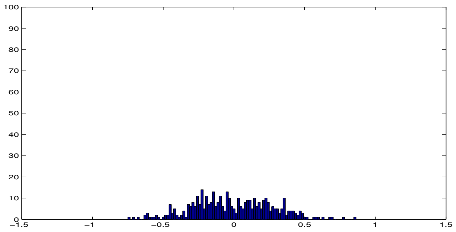
Data (o) a předpovědi (x) 1. veličina



Data (o) a předpovědi (x) 2. veličina



Histogram chyb předpovědí: 1. veličina



Histogram chyb předpovědí: 2. veličina

Obrázek 7: Předpovědi a jejich chyby pro regresní model

Obrázek 7 ukazuje výsledky odhadování dvou modelovaných veličin pomocí regresního modelu. Porovnání dat a jejich předpovědí je na prvních dvou podobrázcích, histogramy chyb předpovědí na dalších dvou obrázcích. Pro dobrý odhad by chyby předpovědí měly mít nulovou střední hodnotu a co nejmenší rozptyl. Vhodným měřítkem pro oba tyto požadavky je odmocnina z druhého obecného momentu chyby predikce, vztažený k rozptylu modelované veličiny. Tento koeficient označíme SE.

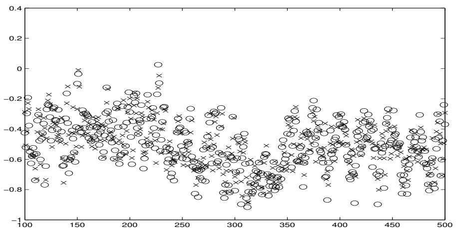
Pro doplnění výsledků na obrázku 7 uveďme numerické hodnocení pomocí SE. Jeho hodnoty v tomto případě jsou 0.7661 a 0.6745.

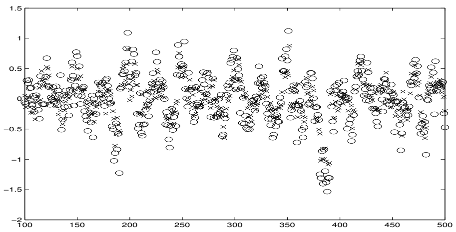
Nový odhad pro různé hodnoty koeficientu alternativního zapomínání dal tyto výsledky:

Hodnota koeficientu zapomínání	1	0.9	0.5	0.1	0.01
Hodnota LVFV	0.9985	0.7991	0.6406	0.6361	0.6353

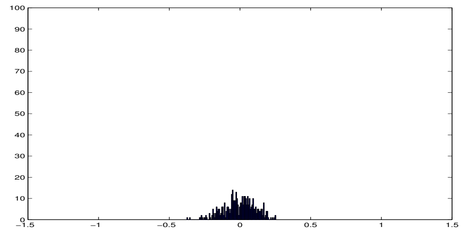
Hodnoty kriteria LVFV klesají s klesajícími hodnotami koeficientu zapomínání, což svědčí o tom, že model není správným reprezentantem dat.

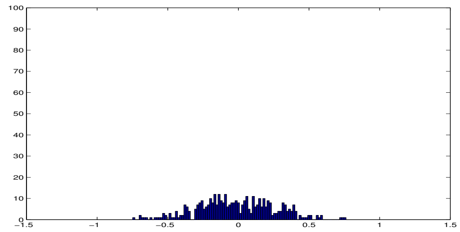
b) *Složený model*
Pro srovnání použijeme pro odhadování stejná data, ale s automatickou ini-





Data (o) a předpovědi (x) 1. veličina   Data (o) a předpovědi (x) 2. veličina





Histogram chyb předpovědí: 1. veličina   Histogram chyb předpovědí: 2. veličina

Obrázek 8: Předpovědi a jejich chyby pro složený model se sedmi komponentami

cializací, která našla sedm komponent. Výsledky s tímto složeným modelem jsou na obrázku 8.

Hodnoty koeficientu SE jsou nyní 0.6454 a 0.6465, a hodnoty kritéria LVFV pro různé koeficienty zapomínání jsou

Hodnota koeficientu zapomínání	1	0.9	0.5	0.1	0.01
Hodnota LVFV	3.1084	3.1863	3.1882	3.1887	3.1888

V tomto případě je hodnota kritéria LVFV nejvyšší pro nejmenší hodnotu koeficientu zapomínání. To potvrzuje správnost modelu vzhledem k použitým datům. Doplňující testy – vizuální kontrola grafů a poměry SE chyb předpovědí – tento výsledek potvrzují. Test, používající kriterium LVFV je ale obecnější, protože posuzuje odhadnutou směs jako celek a nikoliv jen po jednotlivých modelovaných veličinách.

6 Závěr

Pravděpodobnostní modelování a identifikace založené na bayesovském přístupu jsou teoreticky dosti složité, a snad proto ne příliš rozšířené v oblasti automatických systémů. Cílem prvé části článku je naznačit teoretický základ popisu a odhadu vícemodálních dat pomocí složeného modelu, tj. modelu tvořeného množinou komponent, reprezentovaných většinou pomocí regresního modelu, a náhodnou funkcí, která v čase označuje tu komponentu, která je právě aktivní.

Složený model je založen na principu bayesovství a odtud má dvě důležité přednosti: lze u něho specifikovat apriorní nastavení a jako komponenty je možno použít statické i dynamické regresní modely. Apriorní nastavení modelu je důležité zejména pro práci s reálnými daty. Zde se často můžeme setkat s kritickými situacemi, jako např. slabým informačním obsahem dat, nebo prostě nedostatkem měřených dat. Apriorní informace může v takové situaci významně pomoci.

Z výsledků experimentů s reálnými daty vyplývají následující skutečnosti:

- Složený model, i přes svojí značnou teoretickou složitost, je vhodným nástrojem pro popis vícemodálních dat, tj. dat, které mají tendenci vytvářet shluky reprezentující odlišné stavy systému.

- Navrhovaný algoritmus automatického apriorního nastavení je schopen vytvořit kvalitativně stejné podmínky pro start odhadování, jako vytvoří expert znalý teorie a odhadovaného systému.

- Identifikace mnoharozměrných dat přináší problémy nejen při vlastním odhadování, ale i při zobrazení výsledků a ověření kvality identifikace. Při uvedených experimentech je demonstrováno několik možných cest k řešení tohoto problému. Jsou to metody, při nichž se porovnávají skutečná data a předpovídané hodnoty dat ve dvourozměrných řezech prostorou dat. Tyto metody však lze označit jen jako orientační. Důkladnějším měřítkem kvality odhadu je hodnota logaritmické věrohodnostní funkce variant, popsaná v článku. Čím vyšší je hodnota této funkce, tím lepší je korespondence odhadnutého modelu a dat, ze kterých byl odhadnut.

- Nevýhodou kriteria logaritmické věrohodnostní funkce variant je skutečnost, že její hodnoty nejsou normovány, a tedy nevypovídají o absolutní kvalitě odhadu. Za tím účelem byla vypracována a v experimentech použita metoda ověření "dokončenosti" procesu identifikace, založená na porovnání uvedeného kriteria pro opakovaný odhad se stejnými daty a měnícím se koeficientem alternativního zapomínání.

Celkově lze konstatovat, že složený model v oblasti odhadování dobře plní funkce, které se od něho očekávaly. Jeho skutečný význam se však ukáže až v zamýšleném poradním systému, a to ve zpětné vazbě s reálným systémem řízeným operátorem. Tato realizace je již připravena a bude popsána v dalším, nezávislém článku.

Reference

[1] P. Přibyl, "Metody klasifikace dopravy a řízení dopravních sítí v městských aglomeracích", *Docentská habilitační práce, FD ČVUT*, 1998.

[2] P. Ettler and J. Štika, "Regulace elektrohydraulického stavění v rámci řídícího systémuválcovací tratě", *Hydraulika a pneumatika, Žilina*, 1998.

[3] M. Kárný, A. Halousková, and I. Nagy, "Modelling, identification and adaptive control of cross-direction basis weight of paper sheets", in *Proceedings IEE Int. Conf. CONTROL 88*, pp. 159–164. The Chameleon Press Limited, Wandsworth, London, Oxford, 1988.

[4] L. Jirsa and A. Quinn, "Mixture analysis of nuclear medicine data: Medical decision support", in *Irish Signals and Systems Conference 2001. Proceedings*, R. Shorten, T. Ward, and T. Lysaght, Eds., Maynooth, June 2001, pp. 393–398, NUI Maynooth.

[5] Project, "Ist-1999-12058, decision support tool for complex industrial processes based on probabilistic data clustering (prodactool)", 1999-2002.

[6] M. Kárný, "Probabilistic support of operators", *ERCIM News*, , no. 40, pp. 25–26, 2000.

[7] D.M. Titterington, A.F.M. Smith, and U.E. Makov, *Statistical Analysis of Finite Mixtures*, John Wiley & Sons, Chichester, New York, Brisbane, Toronto, Singapore, 1985, ISBN 0 471 90763 4.

[8] M. Kárný, J. Kadlec, and E. L. Sutanto, "Quasi-Bayes estimation applied to normal mixture", in *Preprints of the 3rd European IEEE Workshop on Computer-Intensive Methods in Control and Data Processing*,

J. Rojíček, M. Valečková, M. Kárný, and K. Warwick, Eds., Praha, September 1998, pp. 77–82, ÚTIA AV ČR.

[9] R. Kulhavý and M. B. Zarrop, "On general concept of forgetting", *International Journal of Control*, vol. 58, no. 4, pp. 905–924, 1993.

[10] V. Peterka, "Bayesian approach to system identification", in *Trends and Progress in System Identification*, P. Eykhoff, Ed., pp. 239–304. Pergamon Press, Oxford, 1981.