

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

pro **s. Miroslava Dvořáka**
obor **23-40-8 ASŘ výrobních procesů ve strojírenství**

Vedoucí katedry Vám ve smyslu nařízení vlády ČSSR č. 90/1980 Sb., o státních závěrečných zkouškách a státních rigorózních zkouškách, určuje tuto diplomovou práci:

Název tématu: **Gnostický přístup k zpracování dat**

Zásady pro vypracování:

1. Studium teoretických materiálů.
2. Zpracování programu pro estimaci polohy a rozptylů datového seuboru.
3. Odzkoušení programu na medeleových resp. reálných datech.

V 33/88 5
Data - zpracování -
Gnostika -
statistika mat-
matika

VYSOKÁ ŠKOLA STROJNÍ A TEXTILNÍ
Ústřední knihovna
LIBEREC 1, STUDENTSKÁ 8
PSČ 461 17

Rozsah grafických prací:

Rozsah průvodní zprávy: 50 stran

Seznam odborné literatury:

1. Kevanie, P.: Gnostika jako zobecnění statistiky. Výzkumná zpráva 1188 ÚTIA Praha.
2. Kevanie, P.: Gnostical Theory of Individual Data. Prog. of Centr. and Inform. Theory 13(4) 1984.
3. Kevanie, P.: Gnostical Theory of Small Samples. Prog. of Centr. and Inform. Theory 13(5) 1984.

Vedoucí diplomové práce: Ing. Vladimír Kracík, CSc.

Datum zadání diplomové práce: 28.9.1987

Termín odevzdání diplomové práce: 10.5.1988




Doc. Ing. Vladimír Věchet, CSc.

Vedoucí katedry


Prof. Ing. Vl. Prášil, DrSc.

Děkan

v Liberci dne 24.9. 1987

VYSOKÁ ŠKOLA STROJNÍ A TEXILNÍ V LIBERCI

nositelka Řádu práce

Fakulta strojní

Katedra technické kybernetiky

obor: 23 - 40 - 8

ASŘ výrobních procesů ve strojírenství

GNOSTICKÝ PŘÍSTUP K ZPRACOVÁNÍ DAT

KTK ASŘ - SF - 158

Miroslav DVOŘÁK

Vedoucí práce: Ing. Vladimír Krocík, CSc.

Počet stran	:	69
Počet příloh	:	4
Počet tabulek	:	1
Počet obrázků	:	12

Datum: 5. 5. 1986

Místopředsedně smel ležeti, že jsem diplomatem učení
vypracoval smectotně a rozšířil avolební literatury.

miroslav Dvořák

V Jihlavě, dne 5. 5. 1968

THE 11th + 12th of 1944. The 11th of 1944 was
Wednesday, 11th of 1944. The 12th of 1944 was
Thursday, 12th of 1944. The 13th of 1944 was
Friday, 13th of 1944. The 14th of 1944 was
Saturday, 14th of 1944. The 15th of 1944 was
Sunday, 15th of 1944.

Děkují této vědecké diplomové práci Ing. Vladimírovi
Kracíkovi, CSc. za odborné a pečlivé vedení diplomové práce
a Ing. Pavlovi Kovanicovi, CSc. za laskavé zaslání uvedené
literatury.

Děkuji tímto vedoucímu diplomové práce Ing. Vladimírovi Kracíkovi, CSc za odborné a pečlivé vedení diplomové práce a Ing. Pavlovi Kovanicovi, CSc za laskavé zaslání uvedené literatury.

O B S A H

	str.
1. Úvod	8
2. Teoretická část	10
2.1 Reálná část	10
2.2 Kvantifikační transformace	13
2.3 Souvislosti gnostické teorie a speciální teorie relativity	14
2.4 Druhý axiom gnostické teorie	18
2.5 Věrnost a irrelevance	21
2.6 Estimace	23
2.7 Distribuční funkce, entropie a informace v pojetí Gnostické teorie dat	28
3. Praktická část	30
3.1 Aplikace Gnostické teorie dat k určení parametru polohy datového souboru	30
3.1.1 Gnostický parametr polohy	30
3.1.2 Některé nové možnosti využití parametru polohy	35
3.1.3 Popis programu EKON	37
3.1.4 Příklad výpočtu - program EKON	45
3.2 Aplikace Gnostické teorie k výpočtu regresních koeficientů gnostickou metodou	47
3.2.1 Odhad měřítka s	47
3.2.2 Aplikace Gnostické teorie pro identifikaci regresního modelu	49
3.2.3 Popis programu REGR	55
3.2.4 Příklad výpočtu - program REGR	65
4. Závěr	67
5. Literatura	68

Seznam příloh

1/0	Výpis programů EKON
1/1 - 1/5	Příklady výpočtu programu EKON
2/0	Výpis programu REGR
2/1 - 2/5	Příklady výpočtu programu REGR

Seznam nejčastěji používaných symbolů

z	- datová položka
z_0	- ideální hodnota
ε	- aditivní chyba
ξ	- multiplikativní chyba
Ω, ω	- parametry charakterizující neurčitost
x, y	- kartézské souřadnice hodnoty z
$\underline{K}_q$	- kvantifikační vektor skládání
$\underline{K}_e$	- estimační vektor skládání
$\underline{g}$	- minkovského metrický tenzor
c	- rychlost světla
t	- čas
v	- rychlost částice
τ^0	- jednotkový "tečný" vektor
w_q	- kvantifikační váha
w_e	- estimační váha
f	- věrnost
h_q	- kvantifikační irrelevance
h_e	- estimační irrelevance
p_e	- distribuční funkce
$H(p_e)$	- entropie
I_e	- estimační změna informace
$\bar{f}$	- označení aritmetického průměru
z_{qI}, z_{es}	- konkrétní varianty odhadu z_0
z_n, z'_n	- meze intervalu normality dat
S	- parametr měřítka
F_{j-}, F_{j+}	- empirická distribuční funkce
$\varphi(h_e)$	- kritériální funkce
$D(h_e)$	- dvakrát diferencovatelná funkce h_e
$\underline{e}$	- vektor normalizovaných nezávislých proměnných
$\underline{c}$	- vektor koeficientů
$\underline{x}$	- matice pravých stran

1. Ú V O D

Člověk od "nepaměti" poznával svět, aby si jej mohl co nejlépe přizpůsobit svým potřebám. Svět je možné poznávat různými způsoby a metodami. Velice důležitým způsobem je kvantitativní poznávání světa.

Kvantitativní poznávání světa zahrnuje měření veličin nebo čítání stejnorodých objektů /"kvantifikaci"/ a zpracování výsledků. Výsledky kvantifikace /"data"/ jsou proto jen nepřesnými obrazy skutečné kvantity. Pomocí matematického zpracování dat se proto snažíme získat co nejlepší odhad /"estimaci"/ hledané veličiny. Statistická estimace je založena na zákonu velkých čísel, v jehož důsledku se pozorované četnost náhodných čísel ustaluje na konstantní hodnotě, zvané pravděpodobnost. Praxe mnoha desetiletí přinesla mnoho potvrzení oprávněnosti statistického přístupu k řešení důležitých problémů.

Rychlý rozvoj aplikací statistických metod s využitím počítačů však zvýraznil i výrazná omezení použitelnosti statistického přístupu: reálné procesy jsou mnohdy nestacionární, pořizování velkých datových souborů je nákladné a často z různých důvodů neuskutečnitelné. Nedostatek dat a z něho vyplývající nepřesnost statistických modelů spolu s proměnlivostí objektů i podmínek pozorování tak vedou ke značné nespolehlivosti výsledků statistického zpracování dat.

K odstranění nespolehlivosti výsledků byla vyvinuta celá řada robustních statistických metod, které jsou necitlivé vzhledem ke změnám v modelech i k silným poruchám v datech. Z hlediska praktické použitelnosti tyto metody nemohou uspokojit. Jak vybrat jedinou, když jich existuje celá řada a každá z nich dává různé výsledky. Pro jeden soubor dat se jeví jako vhodná, pro jiný méně vhodná. Robustní statistické metody jsou odvozeny na základě určitých předpokladů, pro určitou třídu matematických modelů. U reálných dat málokdy zajistíme, aby daná data odpovídala nějaké třídě modelů. Snad právě proto jsou v praxi tyto metody tak málo užívané. Tyto obtíže motivují intenzivní hledání účinných alternativ statistických metod.

Překonání zmíněných obtíží by měla umožňovat gnostická teorie dat. Gnostická teorie, jakožto alternativa matematicko-statistické metody zpracování dat, byla vypracována v Ústavu teorie informace a automatizace ČSAV Ing. Pavlem Kovanicem, CSc. Formálně vychází z matematického aparátu speciální teorie relativity. Ve svém důsledku potom vytváří robustní odhady charakteristik /parametr polohy, rozptyl atd./, tj. odhady málo citlivé na odlehlá pozorování. Teorie nepracuje se zákonem velkých čísel, ale vychází z analýzy neurčitosti jednoho data. Popis neurčitosti jednoho data je pak zavedením vhodného axiomu rozšířen na popis neurčitosti souboru. Odhady charakteristik potom jsou počítány na základě optimalizace neurčitosti souboru, při různých specifických kritériích.

Rozbor gnostické teorie, aplikace a příklady využití je tématem mé diplomové práce.

2. TEORETICKÁ ČÁST

2.1 REÁLNÁ DATA

Za reálná data budeme považovat výsledky skutečné kvantifikace. Předpokládejme, že měříme nějakou fyzikální veličinu, která reálně nabývá nějaké konkrétní hodnoty z_0 /měřeno v předem dané jednotce/. Výsledek našeho měření - jak víme z tisíciletí praxe - není ovšem zpravidla přesně z_0 , ale nějaká hodnota $z \neq z_0$, neboli

$$z = z_0 + \varepsilon, \quad /1/$$

kde ε je chyba ve smyslu aditivním. Pro použití u Gnostické teorie je výhodnější použití multiplikativního tvaru

$$z = z_0 \cdot \xi \quad /2/$$

Rozdíl mezi /1/ a /2/ je nepodstatný, vztah /2/ lze převést na tvar /1/ logaritmováním.

Primární výsledek při měření nebo čítání je vždy kladné a konečné číslo. Můžeme tedy vyslovit první axiom gnostické teorie: Výsledky kvantifikace, reálná data i jejich ideální hodnota z_0 jsou striktně kladná konečná reálná čísla [viz např. 10].

Přijetí takovýchto předpokladů je **z** technického hlediska zcela realistické /např. místo $^{\circ}\text{C}$ použijeme K/.

Potom můžeme psát:

$$\xi = e^{\Omega} \quad /3/$$

a tedy

$$z = z_0 e^{\Omega} \quad /4/$$

kde

$z_0 > 0$ je skutečná hodnota fyzikální veličiny,

$z > 0$ je naměřená hodnota /neboli datum/

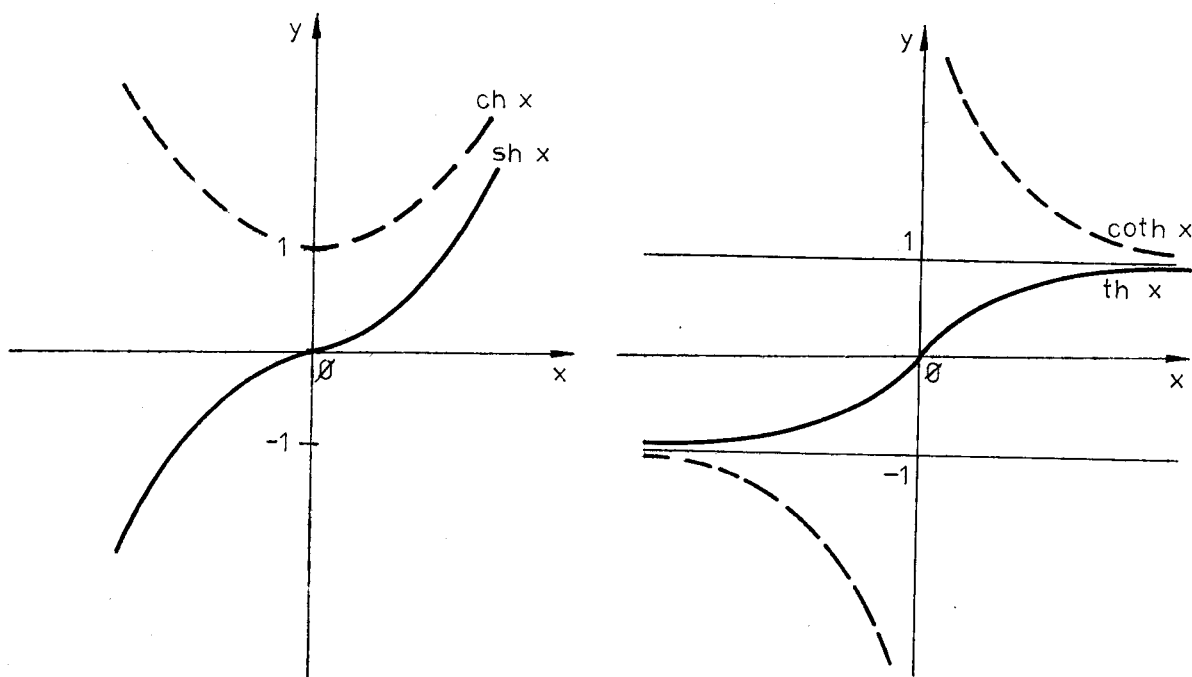
Ω - je veličina popisující nepřesnost /neurčitost/ naměřené hodnoty, pro $\Omega = 0$ dostáváme přesné měření, pro $\Omega > 0$ je naměřená hodnota větší než z_0 , pro $\Omega < 0$

je naměřená hodnota menší než z_0 . Ve vztahu /4/ figurují 3 veličiny z , z_0 , Ω , z nichž známe pouze z . Zdůrazněme, že teprve na základě dalších úvah budou zbývající veličiny odhadovány /tedy stačí odhadnout jedna, třetí již vyplývá ze vztahu/. Prozatím s nimi operujeme jako s danými.

Bylo již řečeno, že datum z se získá měření nebo čítáním. V případě, že získávám z čítáním /zjištění počtu prvků s danou vlastností/ je neurčitost dána neurčitostí při ověřování, zda jistý prvek má předepsanou vlastnost nebo nemá.

Připomeňme nyní některé vlastnosti hyperbolických funkcí

$$\operatorname{sh} x = \frac{e^x - e^{-x}}{2}, \quad \operatorname{ch} x = \frac{e^x + e^{-x}}{2}, \quad \operatorname{th} x = \frac{\operatorname{sh} x}{\operatorname{ch} x} \quad /5/$$



Obr. 1 Grafy hyperbolických funkcí

Z definice potom vyplývají vztahy

$$\operatorname{ch}^2 x - \operatorname{sh}^2 x = 1,$$

$$\operatorname{ch} 2x = \operatorname{ch}^2 x + \operatorname{sh}^2 x = 2\operatorname{ch}^2 x - 1 = 2\operatorname{sh}^2 x + 1,$$

$$\operatorname{sh} 2x = 2\operatorname{sh} x \operatorname{ch} x,$$

$$\operatorname{ch} /x_1 + x_2/ = \operatorname{ch} x_1 \operatorname{ch} x_2 + \operatorname{sh} x_1 \operatorname{sh} x_2,$$

$$\operatorname{sh} /x_1 + x_2/ = \operatorname{sh} x_1 \operatorname{ch} x_2 + \operatorname{sh} x_2 \operatorname{ch} x_1,$$

/6/

a pro derivace hyperbolických funkcí platí vztahy

$$\operatorname{sh}' x = \operatorname{ch} x$$

$$\operatorname{ch}' x = \operatorname{sh} x$$

/7/

$$\operatorname{arth} x' = \frac{1}{1 - x^2}$$

Pro další úvahy bude výhodné přejít od "polárních" souřadnic /viz vztah /4/ ke kartezským souřadnicím x a y :

$$x = z_0 \operatorname{ch} \Omega$$

$$y = z_0 \operatorname{sh} \Omega$$

/8/

Při využití vlastností hyperbolických funkcí /5/ dostaneme:

$$z = x + y$$

/9/

Tím jsme rozložili datovou položku z na dvě složky. První z nich x je při nulovém vlivu neurčitosti rovna ideální hodnotě z_0 , zatímco druhá, y při nulové hodnotě parametru Ω vynizí.

Ze vztahu /8/ plynou i obrácené vztahy

$$z_0^2 = x^2 - y^2$$

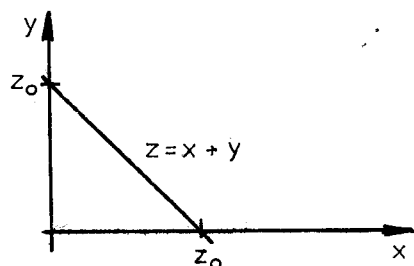
$$\operatorname{th} \Omega = \frac{y}{x}$$

/10/

$$\Omega = \operatorname{arth} \frac{y}{x}$$

Množinu vektorů $/x, y/$ nazveme jevové varianty, jednotlivé vektory nazveme gnostické jevy,

Je-li dáno z /tj. datum/, vyplňují možné body v rovině $/x, y/$ příčku



Obr. 2 Datová položka z

2.2 KVANTIFIKAČNÍ TRANSFORMACE

Kvantifikací rozumíme proces, při kterém "příroda" "zkazí" původní hodnotu z_0 a přetvoří ji na výstupní měřenou hodnotu $z = z(z_0, \Omega)$, přesněji řečeno je to proces, kterým je dvojice $(z_0, \emptyset)$ transformována na dvojici (z_0, Ω) , resp. vektor $\begin{pmatrix} x_0 \\ y_0 \end{pmatrix}$ transformován na vektor $\begin{pmatrix} x \\ y \end{pmatrix}$, kde

$$\begin{aligned} x_0 &= z_0 \cdot \operatorname{ch} \emptyset = z_0 & x &= z_0 \operatorname{ch} \Omega \\ y_0 &= z_0 \cdot \operatorname{sh} \emptyset = \emptyset & y &= z_0 \operatorname{sh} \Omega \end{aligned} \quad /11/$$

Taková transformace je popsána např. maticí

$$\underline{K}_q(\Omega) = \begin{pmatrix} \operatorname{ch} \Omega & \operatorname{sh} \Omega \\ \operatorname{sh} \Omega & \operatorname{ch} \Omega \end{pmatrix} \quad /12/$$

poněvadž platí

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} z_0 \operatorname{ch} \Omega \\ z_0 \operatorname{sh} \Omega \end{pmatrix} = \underline{K}_q(\Omega) \cdot \begin{pmatrix} z_0 \\ \emptyset \end{pmatrix} \quad /13/$$

Vektory v (x, y) budeme označovat zpravidla $\underline{u}$, a tedy

$$\underline{u} = \begin{pmatrix} x \\ y \end{pmatrix} = \underline{K}_q(\Omega) \cdot \underline{u}_0; \quad \underline{u}_0 = \begin{pmatrix} z_0 \\ \emptyset \end{pmatrix} \quad /14/$$

Transformaci /12/ nazveme kvantifikační transformace. Považujeme-li vektor $\underline{u} = \underline{K}_q(\Omega) \underline{u}_0$ za východisko další kvantifikace /např. výsledek měření z je zasílán nějakým poruchovým kanálem a přijatá hodnota je z_1 / dostaneme

$$\begin{aligned} \underline{u}_1 &= \underline{K}_q(\Omega_1) \underline{K}_q(\Omega) \underline{u}_0 = \underline{K}_q(\Omega_1) \underline{u} \\ &= \underline{K}_q(\Omega_1 + \Omega) \underline{u}_0 \end{aligned} \quad /15/$$

což snadno vyplývá z /6/.

/Množina operací $K_q(\Omega)$ /pro reální Ω / tvoří grupu vzhledem k operaci násobení/.

Kvantifikační transformace má důležitou vlastnost, totiž, že zachovává kvadratickou formu $x^2 - y^2$, neboli $x^2 - y^2$ je invariant kvantifikační transformace, poněvadž platí

$$x^2 - y^2 = (x, y) \cdot \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix} = \underline{u}^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \underline{u}$$

a je-li

$$\underline{u}' = K_q(\Omega) \cdot \underline{u}$$

$$\underline{u} = \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\underline{u}' = \begin{pmatrix} x' \\ y' \end{pmatrix}$$

pak také je

$$\underline{u}'^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \underline{u}' = \underline{u}^T K_q(\Omega)^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} K_q(\Omega) \underline{u},$$

ale po vynásobení matic s využitím vztahu /6/ dostaneme

$$\begin{aligned} K_q(\Omega)^T \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} K_q(\Omega) &= \begin{pmatrix} \operatorname{ch} \Omega & \operatorname{sh} \Omega \\ \operatorname{sh} \Omega & \operatorname{ch} \Omega \end{pmatrix} \cdot \begin{pmatrix} \operatorname{ch} \Omega & \operatorname{sh} \Omega \\ -\operatorname{sh} \Omega & -\operatorname{ch} \Omega \end{pmatrix} = \\ &= \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \end{aligned} \quad /16/$$

a tedy je zřejmé, že platí

$$x'^2 - y'^2 = x^2 - y^2$$

2.3 SOUVISLOSTI GNOSTICKÉ TEORIE A SPECIÁLNÍ TEORIE RELATIVITY

Zavedeme-li formálně metrický tenzor

$$\underline{g} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = (\varepsilon_{ij})$$

na jevové variantě dostáváme vektorový prostor, kde je definován skalární součin

$$(\underline{u}, \underline{v}) = \underline{u}^T \underline{g} \underline{v} = \underline{u}_1 \underline{v}_1 - \underline{u}_2 \underline{v}_2 ,$$

kde

$$\underline{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} , \quad \underline{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Takový prostor se nazývá pseudoeuklidovský-minkowského metrika. Speciálně tedy

$$(\underline{u}, \underline{u}) = (u_1^2 - u_2^2)$$

Tato metrika je tedy vzhledem ke kvantifikační transformaci invariantní / $\underline{K}_q(\Omega)$ představuje tzv. pseudoeuklidovskou rotaci/.

Tento matematický model, tj. dvojrozměrný pseudoeuklidovský prostor je ovšem, jak je známo, modelem časoprostoru ve speciální teorii relativity /prostor Minkowského/. V dvojrozměrném časoprostoru je invariantem $x^2 - c^2 t^2$, tj. jsou-li t, c souřadnice relativistického jevu v soustavě 1 a t', c' souřadnice téhož jevu v soustavě 2, která se rovnoměrně pohybuje rychlostí v vzhledem k soustavě 1, platí

$$x^2 - c^2 t^2 = x'^2 - c'^2 t'^2 \quad /17/$$

Píšeme-li $ct = \tau$ je tedy podle předchozího transformace, která převádí souřadnice nečárkované na čárkované tedy

$$\begin{pmatrix} \tau' \\ x' \end{pmatrix} = \underline{L} \begin{pmatrix} \tau \\ x \end{pmatrix}$$

$$\underline{L} = \begin{pmatrix} \text{ch } \Omega & \text{sh } \Omega \\ \text{sh } \Omega & \text{ch } \Omega \end{pmatrix} , \quad /18/$$

kde Ω je zatím neurčený parametr.

Počátek v soustavě 2 v čase t' má souřadnice $\begin{pmatrix} \tau' \\ \varnothing \end{pmatrix} = \begin{pmatrix} ct' \\ \varnothing \end{pmatrix}$
 a tentýž jev má v soustavě 1 souřadnice

$$\begin{pmatrix} \tau \\ vt \end{pmatrix} = \begin{pmatrix} \tau \\ \frac{v}{c} \tau \end{pmatrix} \quad \text{a tedy}$$

$$\begin{pmatrix} \tau' \\ \varnothing \end{pmatrix} = \begin{pmatrix} \text{ch } \Omega & \text{sh } \Omega \\ \text{sh } \Omega & \text{ch } \Omega \end{pmatrix} \cdot \begin{pmatrix} \tau \\ \frac{v}{c} \tau \end{pmatrix}$$

/19/

pro druhou složku dostaneme

$$\varnothing = \tau \text{ sh } \Omega + \frac{v}{c} \tau \text{ ch } \Omega$$

z čehož plyne

$$\text{th } \Omega = - \frac{v}{c}$$

$$\text{sh } \Omega = \frac{- \frac{v}{c}}{\sqrt{1 - \frac{v^2}{c^2}}}$$

$$\text{ch } \Omega = \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}}$$

a tedy transformace L /Lorentzova/ je

$$\begin{pmatrix} \tau' \\ x' \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} & \frac{- \frac{v}{c}}{\sqrt{1 - \frac{v^2}{c^2}}} \\ \frac{- \frac{v}{c}}{\sqrt{1 - \frac{v^2}{c^2}}} & \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \end{pmatrix} \begin{pmatrix} \tau \\ x \end{pmatrix}$$

/20/

nebo v polárních souřadnicích $\begin{pmatrix} x \\ t \end{pmatrix}$, $\begin{pmatrix} x' \\ t' \end{pmatrix}$

$$\begin{pmatrix} t' \\ x' \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} & \frac{-\frac{v}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}} \\ -u & \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \end{pmatrix} \begin{pmatrix} t \\ x \end{pmatrix} \quad /21/$$

Inverzní transformaci zřejmě získáme záměnou v za $-v$ a tedy platí

$$\begin{pmatrix} t \\ x \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} & \frac{\frac{v}{c^2}}{\sqrt{1 - \frac{v^2}{c^2}}} \\ v & \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \end{pmatrix} \begin{pmatrix} t' \\ x' \end{pmatrix} \quad /22/$$

respektive

$$\begin{pmatrix} \tau \\ x \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} & \frac{\frac{v}{c}}{\sqrt{1 - \frac{v^2}{c^2}}} \\ \frac{v}{c} & \frac{1}{\sqrt{1 - \frac{v^2}{c^2}}} \end{pmatrix} \begin{pmatrix} \tau' \\ x' \end{pmatrix} \quad /23/$$

neboli

$$\begin{pmatrix} \tau \\ x \end{pmatrix} = \begin{pmatrix} \text{ch } \Omega & -\text{sh } \Omega \\ -\text{sh } \Omega & \text{ch } \Omega \end{pmatrix} \begin{pmatrix} \tau' \\ x' \end{pmatrix} \quad /24/$$

Zaměníme-li $\Omega = -\Omega_1$ /tj. zaměníme orientaci rychlosti v/ máme

$$\begin{pmatrix} \tau \\ x \end{pmatrix} = \begin{pmatrix} \text{ch } \Omega_1 & \text{sh } \Omega_1 \\ \text{sh } \Omega_1 & \text{ch } \Omega_1 \end{pmatrix} \begin{pmatrix} \tau' \\ x' \end{pmatrix} \quad /25/$$

a speciálně pro počátek v soustavě 2

$$\begin{pmatrix} \tau \\ x \end{pmatrix} = \begin{pmatrix} \text{ch } \Omega_1 & \text{sh } \Omega_1 \\ \text{sh } \Omega_1 & \text{ch } \Omega_1 \end{pmatrix} \begin{pmatrix} \tau' \\ \emptyset \end{pmatrix} \quad /26/$$

Srovnáním výrazů /14/ a /26/ vidíme, že kvantifikační transformace je formálně shodná s Lorentzovou transformací. Z gnostického hlediska je ideální hodnota z_0 resp. jí odpovídající vektor $\underline{u}_0 = \begin{pmatrix} z_0 \\ \emptyset \end{pmatrix}$ transformován na $\underline{u} = \begin{pmatrix} x \\ y \end{pmatrix}$, přičemž parametr Ω specifikuje transformaci. V relativistické interpretaci je počátek v soustavě 2, tj. bod $\begin{pmatrix} \tau' \\ \emptyset \end{pmatrix}$ pozorování v soustavě 1. Toto pozorování vzniklé transformací vektoru $\begin{pmatrix} \tau' \\ \emptyset \end{pmatrix}$ je $\begin{pmatrix} \tau \\ x \end{pmatrix}$. Parametr transformace $\Omega_1 = \text{th } \frac{v}{c}$ je určen rychlostí v, s jakou se vzhledem k sobě pohybují soustavy 1, 2.

Při proceduře měření známe pouze výsledné datum $z = x + y$. To odpovídá v relativistické verzi situaci, kdy je k dispozici pouze měření $\tau + x = ct + x$. Takové pozorování se v dané teorii nazývá neúplné pozorování.

2.4 DRUHÝ AXIOM GNOSTICKÉ TEORIE

Pro zdůvodnění hlavního axiomu teorie, kterým je charakterizována neurčitost souboru dat, uveďme ještě další paralelu ze speciální teorie relativity.

Uvažujme nějakou trajektorii v časoprostoru (τ, x) a její tečný vektor $(d\tau, dx)$ a dále "jednotkový" tečný vektor

$$\underline{\tau}^0 = \left(\frac{d\tau}{\sqrt{d\tau^2 - dx^2}}, \frac{dx}{\sqrt{d\tau^2 - dx^2}} \right) \quad /27/$$

/tj. jednotkový ve smyslu Minkowského metriky/. Tenzor $\mu_0 c^2 \underline{\tau}^0 \otimes \underline{\tau}^0$, kde μ_0 je klidová hustota hmotného toku, jehož dráhu sledujeme je tzv. energie - impuls tenzor daného toku. Důležité je, že energie - impuls tenzor pro soustavu toků je součtem jednotlivých tenzorů.

Jak již víme z předchozího, lze psát

$$\underline{\tau}^0 = (\text{ch } \Omega_1, \text{sh } \Omega_1) \quad /28/$$

pro jisté Ω_1 , a tedy tenzor $\underline{\tau}^0 \otimes \underline{\tau}^0$ zapsaný ve složkách je

$$\begin{pmatrix} \text{ch}^2 \Omega_1 & \text{ch } \Omega_1 \text{ sh } \Omega_1 \\ \text{ch } \Omega_1 \text{ sh } \Omega_1 & \text{sh}^2 \Omega_1 \end{pmatrix} \quad /29/$$

S použitím Minkowského metrického tenzoru

$$\underline{g} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

je dále zřejmé, že

$$\begin{aligned} 2 (\underline{\tau}^0 \otimes \underline{\tau}^0) - \underline{g} &= \begin{pmatrix} 2\text{ch}^2 \Omega_1 & -1 & 2\text{ch } \Omega_1 \text{ sh } \Omega_1 \\ 2\text{sh } \Omega_1 \text{ ch } \Omega_1 & & 2\text{sh } \Omega_1 + 1 \end{pmatrix} = \\ &= \begin{pmatrix} \text{ch}^2 \Omega_1 & \text{sh}^2 \Omega_1 \\ \text{sh}^2 \Omega_1 & \text{ch}^2 \Omega_1 \end{pmatrix} = \end{aligned}$$

$$= \begin{pmatrix} \text{ch } \Omega_1 & \text{sh } \Omega_1 \\ \text{sh } \Omega_1 & \text{ch } \Omega_1 \end{pmatrix}^2 = K_q^2(\Omega_1)$$

Okolnost, že energie - impuls tenzory pro jednotlivé toky se sčítají je výchozím argumentem pro stanovení nejdůležitějšího axiomu teorie, axiomu, který určuje neurčitost souboru na základě neurčitosti jednotlivých dat.

Označme $Z(z_0, n)$ soubor n -měřené ideální hodnoty z_0 . Označme $u_{cq}^T = (z_0 \text{ch } \Omega_c, z_0 \text{sh } \Omega_c)$ složený vektor /composite vector/, tj. vektor, který reprezentuje daný soubor. Axiom 2 stanoví způsob výpočtu parametru Ω_c . Axiom skládání můžeme napsat např. ve tvaru

$$\underline{K}_q^2(\Omega_c) = \frac{1}{W_q} \sum_{i=1}^n \underline{K}_q^2(\Omega_i) \quad /30/$$

kde W_q je normalizační konstanta stanovená tak, aby

$$\det \underline{K}_q^2(\Omega_c) = 1, \quad \text{tedy}$$

$$W_q = \left[\det \sum_{i=1}^n \underline{K}_q^2(\Omega_i) \right]^{1/2} \quad /31/$$

S využitím vztahu /6/ a /12/ dostaneme

$$\underline{K}_q^2(\Omega_c) = \underline{K}_q(2\Omega_c) = \begin{pmatrix} \text{ch } 2\Omega_c & \text{sh } 2\Omega_c \\ \text{sh } 2\Omega_c & \text{ch } 2\Omega_c \end{pmatrix}$$

2.5 VĚRNOST A IRRELEVANCE

Zavedeme-li následující označení

$$\underline{K}_q^2(\Omega) = \begin{pmatrix} \text{ch} 2\Omega & \text{sh} 2\Omega \\ \text{sh} 2\Omega & \text{ch} 2\Omega \end{pmatrix} = \begin{pmatrix} 1/f & h_q \\ h_q & 1/f \end{pmatrix} \quad /32/$$

dostaneme definice následujících pojmů:

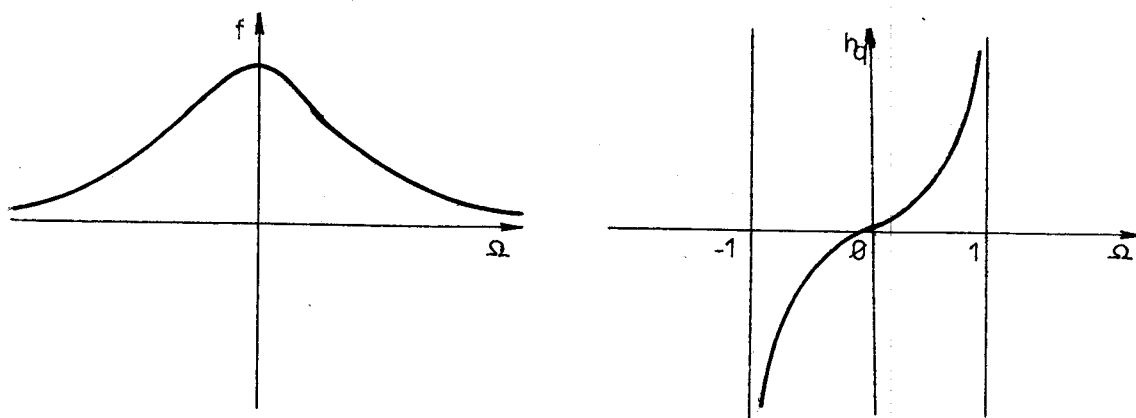
Prvním pojmem je věrnost /fidelity/. Vyjadřuje váhu jednotlivé datové položky. Její vyjádření dostaneme z /32/

$$f = \frac{1}{\text{ch} 2\Omega} \quad /33/$$

Druhým pojmem je kvantifikační irrelevance. Je charakteristikou nepodobnosti skutečného výsledku výsledku ideálnímu

$$h_q = \text{sh} 2\Omega \quad /34/$$

Grafické vyjádření těchto pojmů je na obr. 3.



Obr. 3 Graf věrnosti a kvantifikační irrelevance

Pro "téměř přesné" datum, tj. $\Omega \sim 0$ je $f \sim 1$,
naopak $h_q \sim 0$, na proti tomu pro Ω velké je $f \sim 0$ a

h_q je velké.

Dále můžeme věrnost vyjádřit následovně:

Podle /8/ platí

$$x = z_0 \operatorname{ch} \Omega$$

$$y = z_0 \operatorname{sh} \Omega$$

po úpravě a dosazení s využitím vztahu /6/ dostaneme

$$\operatorname{ch} 2 \Omega = \operatorname{ch}^2 \Omega + \operatorname{sh}^2 \Omega = \frac{x^2 + y^2}{z_0^2}$$

s využitím vztahu /10/

$$z_0^2 = x^2 - y^2$$

nakonec dostaneme vyjádření věrnosti:

$$f = \frac{1}{\operatorname{ch} 2 \Omega} = \frac{x^2 - y^2}{x^2 + y^2} \quad /35/$$

Totéž můžeme provést i při vyjadřování kvantifikační irrelevance. Tedy

$$\operatorname{sh} 2 \Omega = 2 \operatorname{sh} \Omega \operatorname{ch} \Omega = \frac{2xy}{z_0^2}$$

tudíž je kvantifikační irrelevance

$$h_q = \operatorname{sh} 2 \Omega = \frac{2xy}{x^2 - y^2} \quad /36/$$

Využijeme-li vztah /3/

$$\xi = e^{\Omega}$$

dostaneme

$$\operatorname{ch} 2\Omega = \frac{1}{2} (e^{2\Omega} + e^{-2\Omega}) = \frac{1}{2} (\xi^2 + \xi^{-2}) \quad /37/$$

a tedy pro věrnost platí

$$f = \frac{2}{\xi^2 + \xi^{-2}} \quad /38/$$

Provedeme-li totéž s

$$\operatorname{sh} 2\Omega = \frac{1}{2} (e^{2\Omega} - e^{-2\Omega})$$

dostaneme vztah pro kvantifikační irrelevanci

$$h_Q = \frac{1}{2} (\xi^2 - \xi^{-2}) \quad /39/$$

2.6 ESTIMACE

Jak již bylo uvedeno, kvantifikační transformace představuje formálně proceduru pomocí níž "příroda" transformuje výchozí ideální hodnotu z_0 na naměřenou hodnotu z . Procedura estimace má za úkol naopak transformovat datum z a v nějakém smyslu se přiblížit ideální hodnotě z_0 .

Je-li dáno $z = x + y$, přičemž zašumění, tj. Ω je velké, jeví se přirozené, že estimační pohyb probíhá ve směru tečny k přímce $z = x + y$. Z dalších formálně matematických důvodů se jeví adekvátní definovat estimační transformaci jako obyčejné /euklidovské/ otočení daného bodu do osy x , tedy:

$$K_e(\omega) = \begin{pmatrix} \cos \omega & -\sin \omega \\ \sin \omega & \cos \omega \end{pmatrix} \quad /40/$$

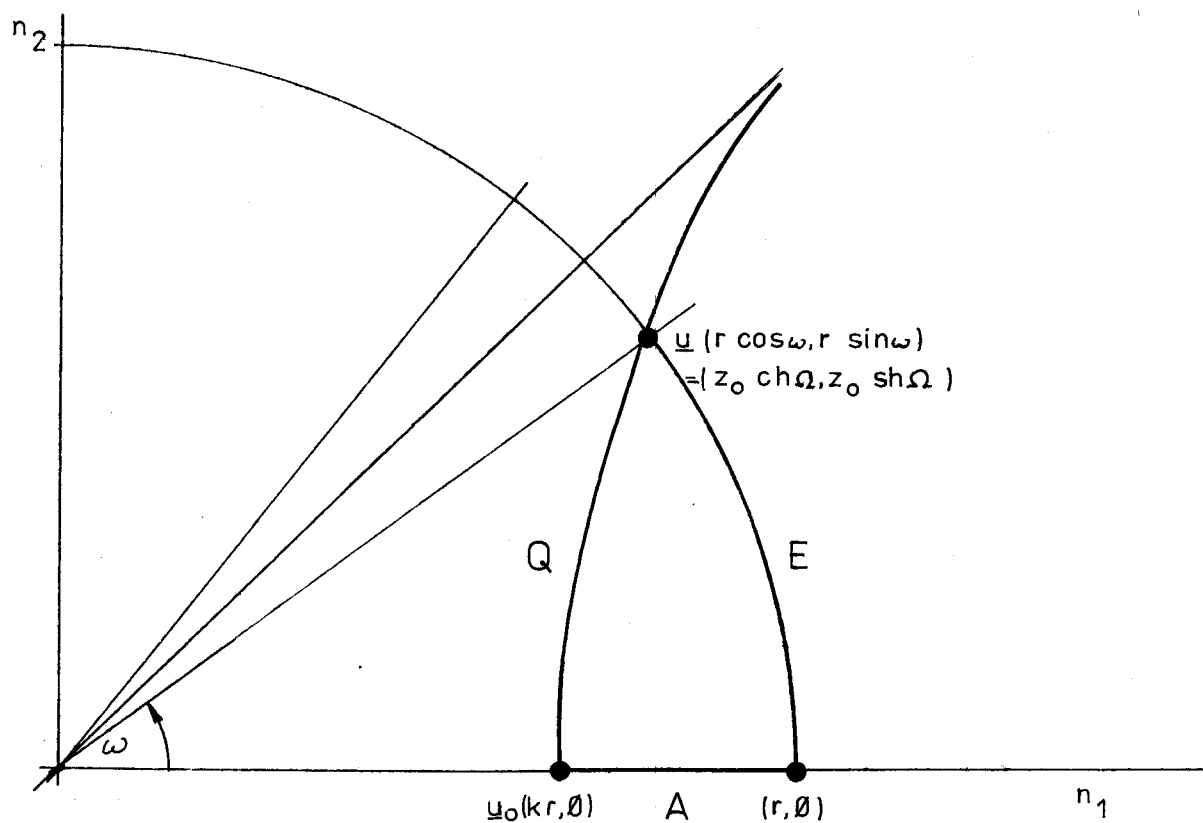
kde

$$\omega = - \operatorname{arctg} \operatorname{th} \Omega = - \operatorname{arc} \operatorname{th} \frac{y}{x}$$

Dále je zřejmé, že platí

$$K_e^2(\omega) = \begin{pmatrix} \cos 2\omega & -\sin 2\omega \\ \sin 2\omega & \cos 2\omega \end{pmatrix} \quad /41/$$

V teorii je zaveden pojem tzv. ideálního gnostického cyklu, který je dán podle obr. 4



Obr. 4 Ideální gnostický cyklus /Pramen [5] /

kde E - estimace

Q - kvantifikace

A - zmenšení

Tento cyklus, jak je dokázáno, vykazuje v jistém smyslu minimální ztrátu informace.

Používáme-li při estimaci veličiny z_0 transformace euklidovského typu, postupujeme nejlepším z možných způsobů a maximalizujeme informaci. Složky x a y v euklidovské rovinné geometrii lze pomocí polárních souřadnic zapsat takto

$$\begin{aligned} x &= r \cos \omega \\ y &= -r \sin \omega \end{aligned} \quad /42/$$

přičemž

$$r^2 = x^2 + y^2$$

odtud s použitím vztahu /10/ dostaneme

$$\operatorname{th} \Omega = \frac{y}{x} = -\operatorname{tg} \omega$$

Ze známých vlastností goniometrických funkcí můžeme odvodit

$$\cos 2\omega = \frac{1 - \operatorname{tg}^2 \omega}{1 + \operatorname{tg}^2 \omega}$$

a pokud totéž uděláme i pro hyperbolickou funkci

$$\operatorname{ch} 2\Omega = \frac{1 + \operatorname{th}^2 \Omega}{1 - \operatorname{th}^2 \Omega}$$

Srovnáním obou výrazů dostáváme

$$\cos 2\omega = \frac{1}{\operatorname{ch} 2\Omega} \quad /43/$$

Totéž můžeme udělat i pro

$$\sin 2\omega = \frac{2 \operatorname{tg} \omega}{1 + \operatorname{tg}^2 \omega}$$

a hyperbolickou funkcí

$$\operatorname{tg} 2 \Omega = \frac{\sin 2 \Omega}{\cos 2 \Omega} = \frac{\frac{2 \operatorname{th} \Omega}{1 - \operatorname{th}^2 \Omega}}{\frac{1 + \operatorname{th}^2 \Omega}{1 - \operatorname{th}^2 \Omega}} = \frac{2 \operatorname{th} \Omega}{1 + \operatorname{th}^2 \Omega}$$

odtud opět po srovnání dostáváme

$$\sin 2 \omega = - \operatorname{th} 2 \Omega \quad /44/$$

Analogicky bychom odvodili

$$\operatorname{sh} 2 \Omega = - \operatorname{tg} 2 \omega \quad /45/$$

Podle výrazů /43/, /44/, /45/ a /32/ lze tedy psát

$$\underline{\underline{\mathbf{K}}}_e^2 = \begin{pmatrix} f & - h_e \\ h_e & f \end{pmatrix} \quad /46/$$

kde

$$h_e = \sin 2 \omega = - \operatorname{th} 2 \Omega \quad /47/$$

se nazývá estimační irrelevance.

Jelikož podle vztahů /8/ s využitím vlastností hyperbolických funkcí /6/ dostáváme

$$\operatorname{sh} 2 \Omega = \frac{2 x y}{z_0^2}$$

$$\operatorname{ch} 2 \Omega = \frac{x^2 + y^2}{z_0^2}$$

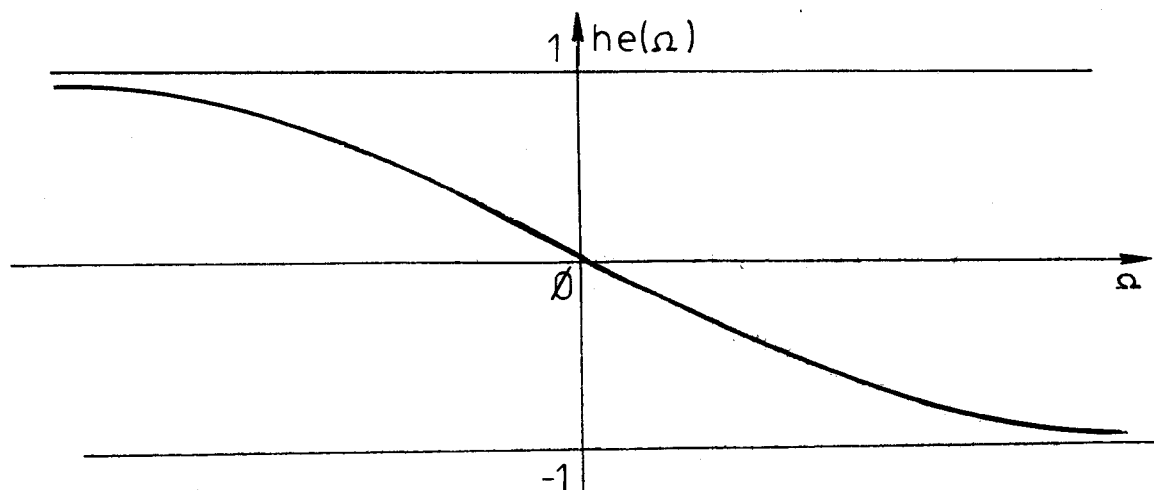
odtud

$$\operatorname{th} 2 \Omega = \frac{2 x y}{x^2 + y^2}$$

Ve vztahu /47/ pak můžeme pro estimační irrelevanci

nsát

$$h_e = - \operatorname{th} 2\Omega = - \frac{2xy}{x^2 + y^2} \quad /48/$$



Obr. 5 Graf estimační irrelevance h_e

Dále ještě podle

$$\xi = e^{\Omega}$$

je

$$\operatorname{th} 2\Omega = \frac{e^{2\Omega} - e^{-2\Omega}}{e^{2\Omega} + e^{-2\Omega}} = \frac{\xi^2 - \xi^{-2}}{\xi^2 + \xi^{-2}}$$

a tedy dostáváme

$$h_e = - \frac{\xi^2 - \xi^{-2}}{\xi^2 + \xi^{-2}} \quad /49/$$

a srovnáním vztahů /38/, /39/ a /49/ je snadno vidět, že platí

$$h_q = - \frac{h_e}{f}, \quad h_e = - h_q f$$

2.7 DISTRIBUČNÍ FUNKCE, ENTROPIE A INFORMACE V POJETÍ GNOSTICKÉ TEORIE DAT

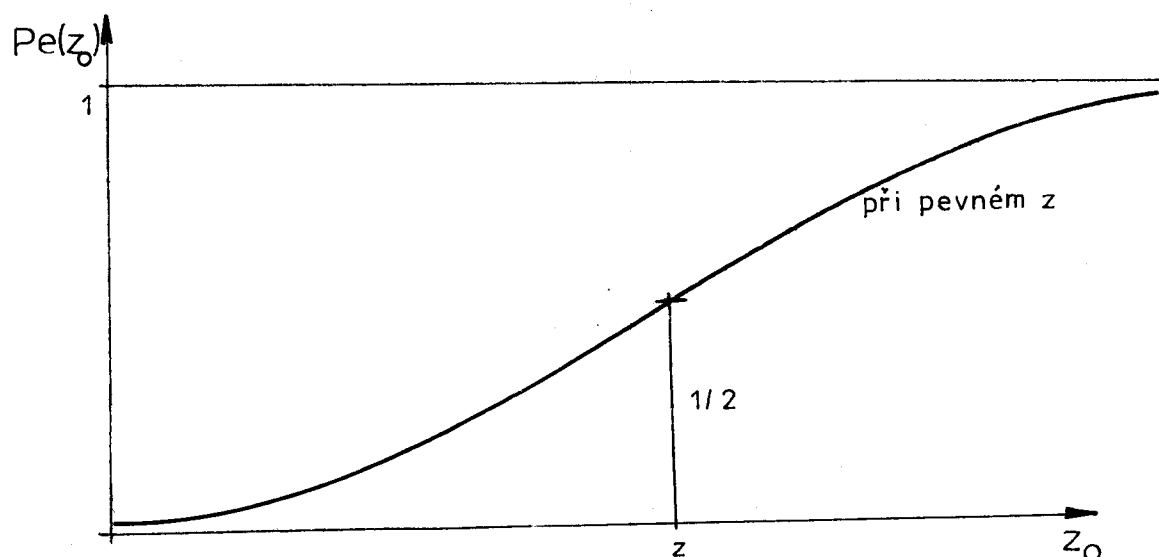
Jak plyne z /48/ probíhá h_e v intervalu $\langle -1, 1 \rangle$
a tedy veličina

$$p_e = (1 + h_e) / 2 \quad /50/$$

nabývá hodnot od 0 do 1

Podle /49/ a podle vztahu $\xi = \frac{z}{z_0}$ můžeme psát

$$\begin{aligned} p_e &= \left(1 - \frac{\xi^4 - 1}{\xi^4 + 1} \right) \frac{1}{2} = \left(1 - \frac{z^4 - z_0^4}{z^4 + z_0^4} \right) \frac{1}{2} = \\ &= \frac{z_0^4}{z_0^4 + z^4} \quad /51/ \end{aligned}$$



Obr. 6 Graf gnostické distribuční funkce

Je zřejmé, že pro $z = z_0$ je $p_e = 1/2$

p_e / z_0 má formální vlastnosti jako distribuční funkce
rozdělení pravděpodobnosti nějaké náhodné veličiny, lze ji

interpretovat jako míru důvěry o poloze ideální hodnoty z_0 . Tj. konkrétně pro nějaký interval (a, b) míra důvěry, že $z_0 \in (a, b)$ je $p_e(b) - p_e(a)$.

Speciálně pro $(-\infty, z)$ a (z, ∞) je míra důvěry $1/2$.

V [5] je zavedena Shanonova entropie pokusu

$$H = H(p_e) = -p_e \ln p_e - (1 - p_e) \ln (1 - p_e)$$

Odtud je v teorii zaveden pojem estimační změna informace

$$I_e = H\left(\frac{1}{2}\right) - H(p_e)$$

3. PRAKTICKÁ ČÁST

3.1 APLIKACE GNOSTICKÉ TEORIE DAT K URČENÍ PARAMETRU POLOHY DATOVÉHO SOUPORU

3.1.1 Gnostický parametr polohy

Ve statistice charakterizuje parametr polohy umístění distribuční funkce náhodné veličiny, nejjednodušším příkladem je střední hodnota náhodné veličiny, která se z dat odhaduje jako jejich průměr. V robustní statistice se staráme je parametr polohy pomocí nelineárních vzorců, každá datová položka je vážena vahou, které závisí na její hodnotě i na hodnotě hledaného parametru polohy. Gnostickým zobecněním parametru polohy tohoto typu je veličina z_0 .

Přistupme nyní k popisu procedur estimace ideální hodnoty z_0 při daných pozorovaných hodnotách $z_1, z_2, \dots, z_n$ /data/. Sledované charakteristiky f, h_g, h_e jsou danými funkcemi argumentu $\xi_i = \frac{z_i}{z_0}$ a tedy funkcemi z_0 .

Označme $\bar{f} = \frac{1}{n} \sum f_i$, kde $f_i = f_i(z_0)$

pro datum z_i . Obdobně $\bar{h}_g, \bar{h}_e^2$ apod.

Např. v [4] se uvažují následující varianty odhadu z_0 :

$$z_{qi} = \arg \min_{z_0} \bar{h}_g^2$$

$$z_{gc} = \arg \min_{z_0} \left(\frac{1}{n} \bar{h}_g^2 - (\bar{h}_g)^2 \right) \quad /52/$$

$$z_{qf} = \arg \min_{z_0} \bar{f}^{-1}$$

$$z_{qs} = \text{řešení rovnice } \bar{h}_g(z_0) = \emptyset$$

z_{es} : řešení rovnice $\overline{h_e}(z_0) = \emptyset$

$$z_{ef} = \arg \max_{z_0} \overline{f}$$

/52/

$$z_{ec} = \arg \min_{z_0} \left(\frac{1}{n} \overline{h_e^2} - (\overline{h_e})^2 \right)$$

$$z_{ei} = \arg \min_{z_0} \overline{h_e}^2$$

Každý z odhadů charakterizuje z_0 v jiném smyslu. V příkladovém materiálu budou výsledky komentovány. Markantní je rozdíl mezi z_{eI} a z_{eI} ; je velmi citlivý na odchýlené hodnoty, naopak z_{eI} je robustní /vyplývá ze srovnání grafů funkcí $h_0(\Omega)$, $h_e(\Omega)$ /.

Ukažme např. výpočet $z_{eI} = \arg \min_{z_0} \overline{h_e^2}$

z_{eI} se spočítá z podmínky

$$\frac{d}{dz_0} \sum_{i=1}^n h_{ei}^2 = 0$$

a tedy po derivaci součtu

$$\sum_{i=1}^n h_{ei} \frac{d}{dz_0} h_{ei} = 0$$

neboli

$$\sum_{i=1}^n h_{ei} \frac{d}{d\xi_i} h_{ei} \frac{d\xi_i}{dz_0} = 0$$

/53/

Podle /47/ je

$$h_{ei} = - \frac{\xi_i^4 - 1}{\xi_i^4 + 1} ; \quad \xi_i = \frac{z_i}{z_0}$$

Z toho po derivování dostáváme

$$\frac{d}{d \xi_i} h_{ei} = - \frac{8 \xi_i^3}{(\xi_i^4 + 1)^2} \quad /54/$$

$$\frac{d}{d z_0} \xi_i = - \frac{z_i}{z_0^2} = - \xi_i \frac{1}{z_0} \quad /55/$$

Podle vztahu /36/ platí pro věrnost

$$f = \frac{2}{\xi^2 + \xi^{-2}} = \frac{2 \xi^2}{\xi^4 + 1}$$

S využitím tohoto vztahu a vztahů /54/ a /55/ dostáváme

$$\frac{d}{d z_0} h_{ei} = \frac{8 \xi_i^4}{(\xi_i^4 + 1)^2 z_0} = \frac{2 f_i^2}{z_0} \quad /56/$$

a tedy výraz /53/ můžeme upravit na tvar

$$\sum_{i=1}^n \frac{\xi_i^4 - 1}{\xi_i^4 + 1} \cdot \frac{\xi_i^4}{(\xi_i^4 + 1)^2} \cdot \frac{1}{z_0} = 0$$

tedy

$$\sum_{i=1}^n \frac{\xi_i^8}{(\xi_i^4 + 1)^3} = \sum_{i=1}^n \frac{\xi_i^4}{(\xi_i^4 + 1)^3}$$

z čehož plyne

$$\sum_{i=1}^n \xi_i^2 f_i^3 = \sum_{i=1}^n \xi_i^{-2} f_i^3$$

a dosazením za $\xi_i = \frac{z_i}{z_0}$ dostaneme

$$\frac{1}{z_0^2} \sum_{i=1}^n z_i^2 f_i^3 = z_0^2 \sum_{i=1}^n z_i^{-2} f_i^3$$

z čehož konečně dostáváme

$$z_0 = z_{ei} = 4 \sqrt{\frac{\sum_{i=1}^n f_i^3 z_i^2}{\sum_{i=1}^n f_i^3 z_i^{-2}}} \quad /58/$$

Pokud bychom odvozovali naznačeným způsobem také další odhady, dostávali bychom postupně různé mocniny u věrnosti. Jestliže toto zobecníme můžeme pro parametr polohy psát vztah:

$$A: z_{oj} = 4 \sqrt{\frac{\sum_{i=1}^n f_i^n z_i^2}{\sum_{i=1}^n f_i^n z_i^{-2}}} \quad /59/$$

nebo obecnějšího vztahu

$$B: z_{oj} = 4 \sqrt{\frac{\sum_{i \neq j}^n f_j^k f_i^m z_i^2}{\sum_{i \neq j}^n f_j^k f_i^m z_i^{-2}}} \quad /60/$$

Parametry m a k jsou uvedeny v Tab. 1

z _{oj} Typ odhadu	Podmínka optimality odhadu	Varianta A / B	Parametry odhadu	
			m	k
z _{qI}	$\frac{d}{dz_o} \overline{h_q^2} = 0$	A	-1	-
z _{qc}	$\frac{d}{dz_o} \overline{h_q^2} - \frac{1}{n} \overline{h_q^2} = 0$	B	0	-1
z _{qF}	$\frac{d}{dz_o} \overline{f^{-1}} = 0$	A	0	-
z _{qs}	$\overline{h_q} = 0$	A	0	-
z _{es}	$\overline{h_e} = 0$	A	1	-
z _{eF}	$\frac{d}{dz_o} \overline{h_e^2} - \frac{1}{n} \overline{h_e^2} = 0$	A	1	2
z _{ec}	$\frac{df}{dz_o} = 0$	B	2	-
z _{eI}	$\frac{d}{dz_o} \overline{h_e^2} = 0$	A	3	-

Tab. 1 Tabulka gnostických parametrů polohy datového souboru

Důsledkem volby exponentů m a k je stupeň = druh robustnosti odhadu. Nejvíce robustní je algoritmus pro z_{eI}, tedy pro m = 3 nejvíce citlivý je pak pro z_{qI}, tedy pro m = -1.

Je třeba připomenout, že výraz /59/ resp. /60/ je skutečně rovnice a nikoliv přímé řešení optimalizační úlohy, neboť věrnost f_i závisí na neznámém řešení této rovnice. Silná nelinearita rovnice /58/ také způsobuje, že parametrů polohy můžeme najít u jediného datového souboru více, nejen jediný. To je ale důležité a rozumná vlastnost těchto odhadů, neboť datový soubor může obsahovat dva či více shluků dat, každému z nich pak patří svůj parametr polohy. Takový soubor není účelné charakterizovat jediným celkovým parametrem polohy.

3.1.2 Některé nové možnosti využití parametru polohy

Již zmíněné zajímavé praktické důsledky nelinearity a robustnosti gnostických parametrů polohy dovolují formulovat a řešit některé důležité úlohy nového typu:

a/ Test příslušnosti dat k danému datovému souboru

Je dán datový soubor, jehož jednotlivé položky značíme z_i . Volme si vhodně hustou posloupnost poslední "nové" položky z_n a pro každou z nich vypočteme odhad z_{eI} . Charakter získané funkce je na obr. 7

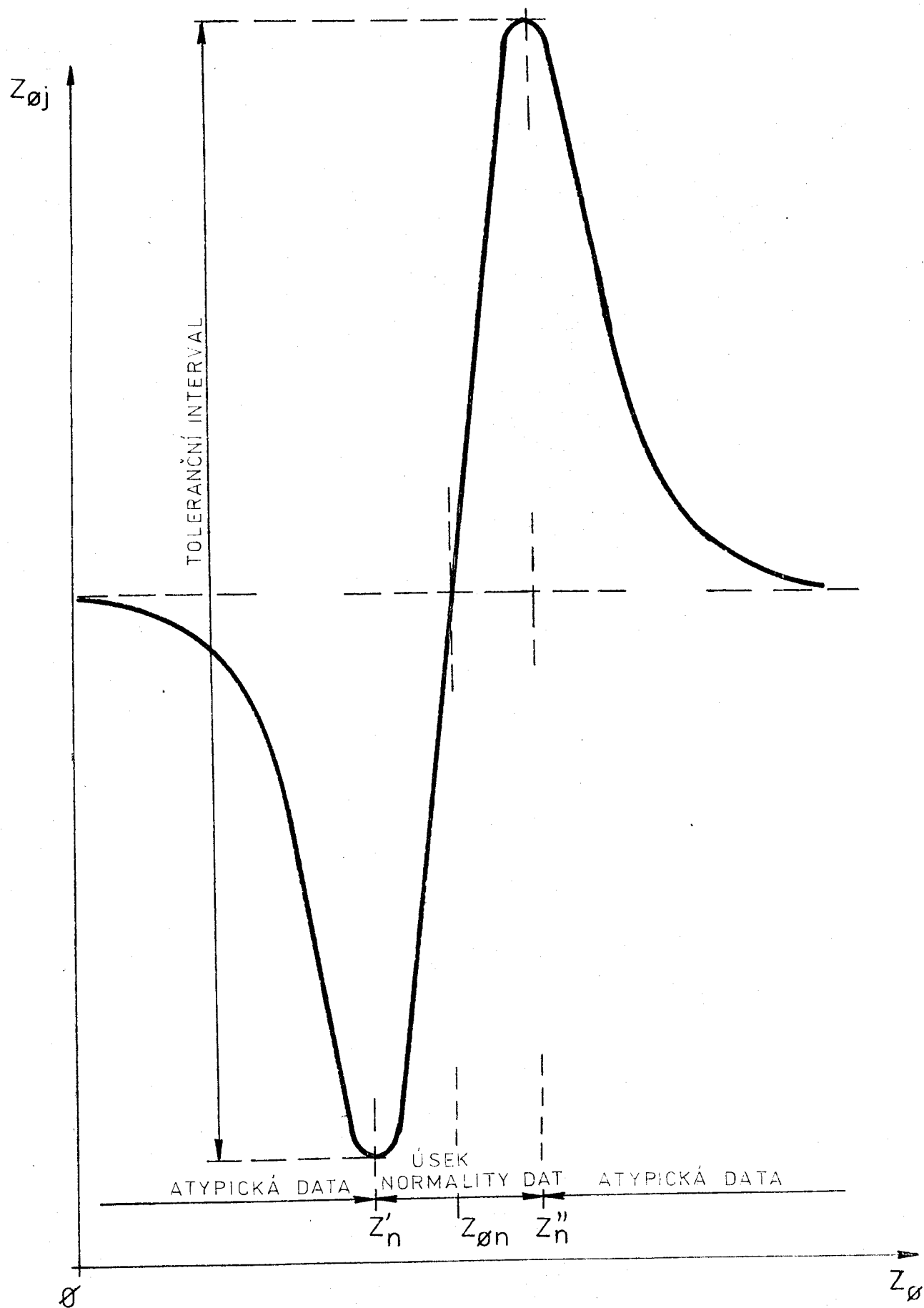
Pokud nová datová položka odpovídá vlastnostem $n - 1$ předchozích datových položek, leží-li v "intervalu normality dat", tedy platí

$$z'_n \leq z_n \leq z''_n$$

Pokud použijeme při výpočtu z' a z'' parametr $m = 2$ místo $m = 3$ dostaneme pak méně přísné meze. Tento fakt ukazuje, že parametr m silně ovlivňuje "interval normality dat".

b/ Testování homogenity datového souboru

Homogenní je ten datový soubor, který má jediný gnostický parametr polohy daného typu. Jeho výpočtem při různých počátečních hodnotách se o unicítě parametru polohy snadno přesvědčíme. Dosadíme-li např. do daného souboru jednou za po-



obr. 7. Graf funkce $z_{ei}(z_n)$ při $z_1, z_2, \dots, z_{n-1} = \text{konst.}$

čáteční hodnotu z_0 hodnotu maximální, podruhé minimální a dostaneme shodnou hodnotu parametru polohy, je soubor homogenní.

c/ Rozčlenění datového souboru na homogenní podsoubory

Zjistíme-li, že datový soubor není homogenní, můžeme se pokusit o rozčlenění datového souboru na homogenní podsoubory. Výpočtem křivky z_{eI} /podle obr. 7/ při počáteční hodnotě odhadu rovné prvnímu parametru polohy, určíme rozsah prvního podsouboru. Podobně přiřadíme ostatním parametrům polohy rozsahy dalších podsouborů. Data připadající do průniku takto určených intervalů vypustíme jako nerozhodnutelná nebo pro jejich rozčlenění použijeme distribuční funkce /viz vztahy /50/ a /51// určené daty a to tak, že přiřadíme datovou položku z průniku intervalů k tomu podsouboru, jehož distribuční funkce má v testovaném bodě větší hodnotu.

3.1.3 POpis programu EKON

Výpis programu je v příloze 1

a/ Funkce programu

Program řeší výpočet gnostického parametru polohy. Testuje homogenitu datového souboru a pokud soubor není homogenní, vypočte všechny parametry polohy a nehomogenní soubor rozčlení na homogenní podsoubory

b/ Seznam volaných podprogramů

POLY - výpočet parametrů polohy
ITER - iterační výpočet jednotlivého parametru polohy
ROZL - výpočet rozložení intervalu
NORM - výpočet intervalu normality dat
SCHE - třídění dat schellovou metodou

c/ Vstupní data

N - počet datových položek
Z(I) - jednotlivé datové položky
EPS - požadovaná přesnost
M - optimalizační parametr

d/ Výstupní proměnné a pole

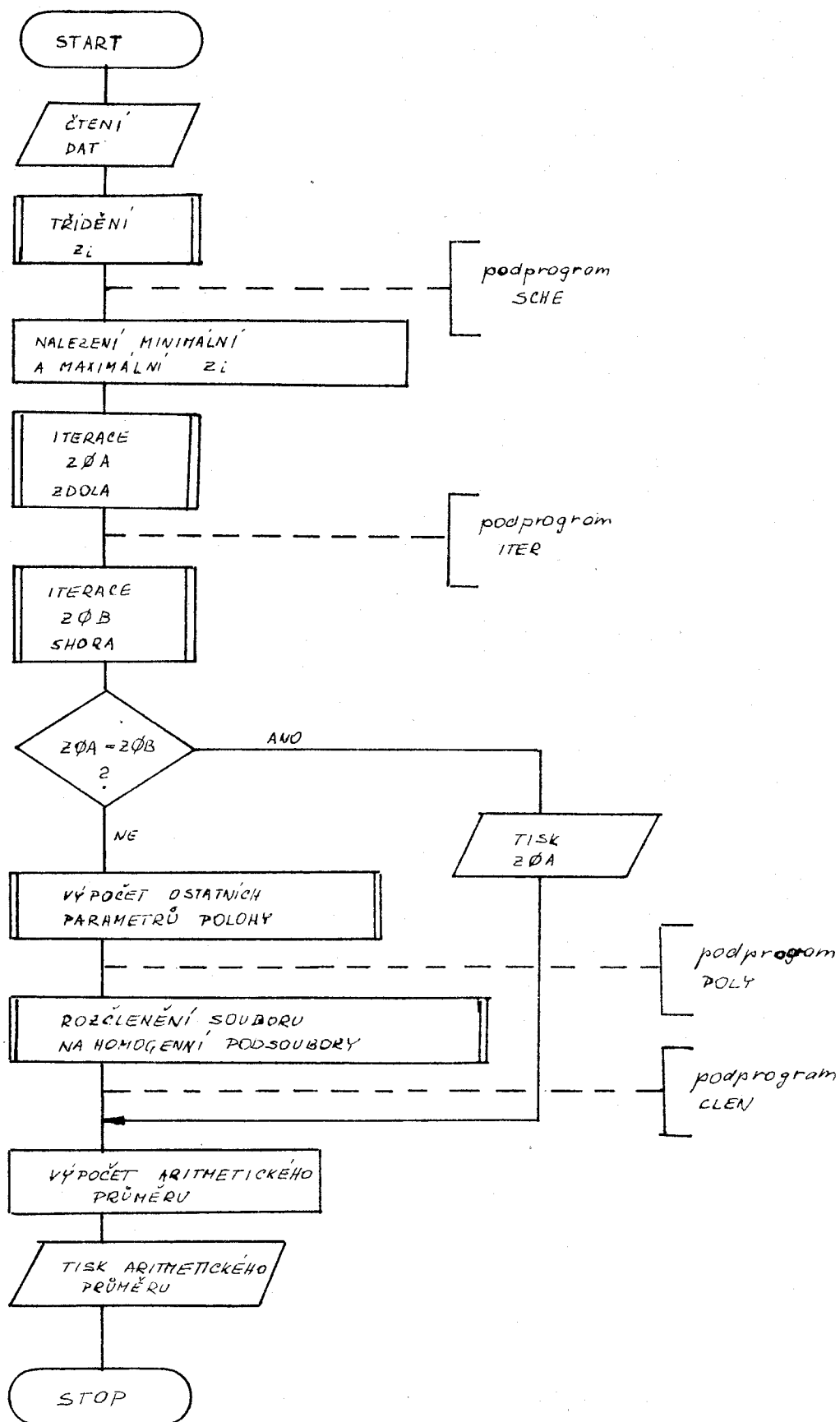
I - číslo datové položky
Z(I) - vstupní datová položka
A(I) - seříděný vstupní soubor datových položek
EPS - požadovaná přesnost
M - optimalizační parametr
v případě homogenního dat. souboru
ZØA - parametr polohy datového souboru
v případě nehomogenního dat. souboru
LV - číslo parametru polohy
B(I) - parametr polohy dat. souboru
DM(I) - dolní a horní mez intervalu normality dat
HM(I) -
ARP - aritmetický průměr

e/ Popis činnosti hlavního programu

Hlavní program načítá vstupní data. Dále setřídí hodnotu Z(I) od nejmenší hodnoty k největší. Poté se nalezne minimální a maximální datová položka. Tyto hodnoty volíme za počáteční podmínky iterace. Naleznou se tedy dva odhady parametru polohy a testuje se jejich shoda, tedy "homogenita" datového souboru. Je-li soubor homogenní, tedy oba odhady jsou shodné, vytiskne se statistický parametr polohy, aritmetický průměr a výpočet se ukončí. V opačném případě se předá výpočet podprogramu POLY, který vypočte všechny parametry polohy a podprogramu CIEN, který daný nehomogenní soubor rozdělí na homogenní podsoubory. Nakonec se opět provádí výpočet a tisk aritmetického průměru dat.

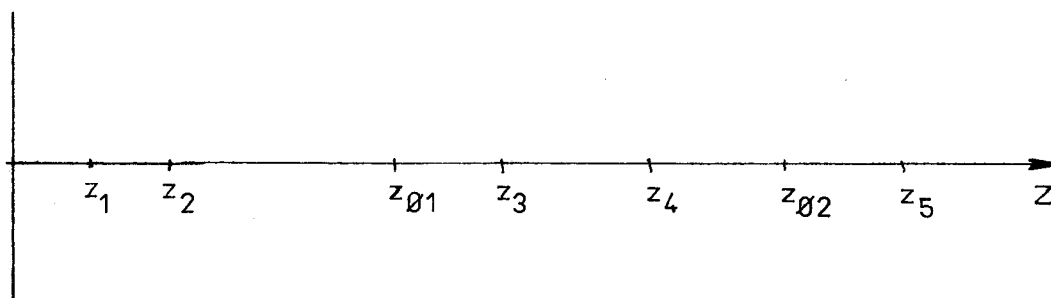
f/ Popis činnosti podprogramu ROZI

Podprogram provádí "rozložení" intervalu. Vstupem je zde dvojice odhadnutých parametrů polohy, $zØ1$, $zØ2$.



VD1 - Hrubý vývojový diagram
EKON - hlavní program

Podprogram nalézá indexy $I1$, $I2$ reálných dat $z(I)$, tak, aby platilo $z(I1) \leq z\emptyset1$ a zároveň $z(I2) \geq z\emptyset2$ a $z(I1)$ a $z(I2)$ byly od $z\emptyset1$ a $z\emptyset2$ nejbližší nižší resp. vyšší datové položky. To je nejlépe vidět na příkladu viz obr. 8.



Obr. 8 Příklad rozložení intervalu dat

V příkladě na obr. 8 budou tedy hledanými datovými položkami z_2 a z_5 a tedy výstupními indexy $I1 = 2$ a $I2 = 5$.

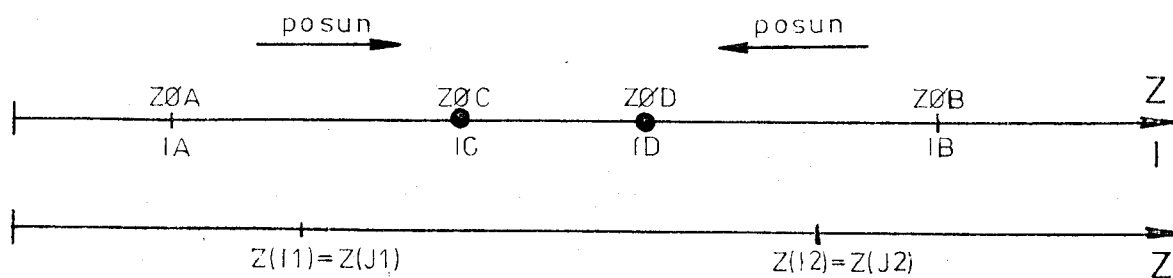
g/ Popis činnosti podprogramu POLY

Podprogram nalézá všechny parametry polohy datového souboru. Vstupem jsou mezní parametry polohy $z\emptyset A$ a $z\emptyset B$ odhadnuté z minimální a maximální počáteční podmínky. Po provedení rozložení intervalu /viz program ROZL/, získáme meze intervalu $z(IA)$ a $z(IB)$, mezi kterými mohou ležet další parametry polohy. Nyní procedura provádí dělení intervalu (IA, IB) metodou zlatého řezu. Metoda zlatého řezu je popsána např. v [2]. Získáme tím hodnoty IC a ID a tedy $z(IC)$ a $z(ID)$. Procedura dále nalézá odpovídající parametry polohy $z\emptyset C$ a $z\emptyset D$. Po provedení rozložení intervalu /tentokrát je vstupem $z\emptyset C$ a $z\emptyset D$ a výstupem $J1$ a $J2$ / můžeme testovat alternativy pokračování.

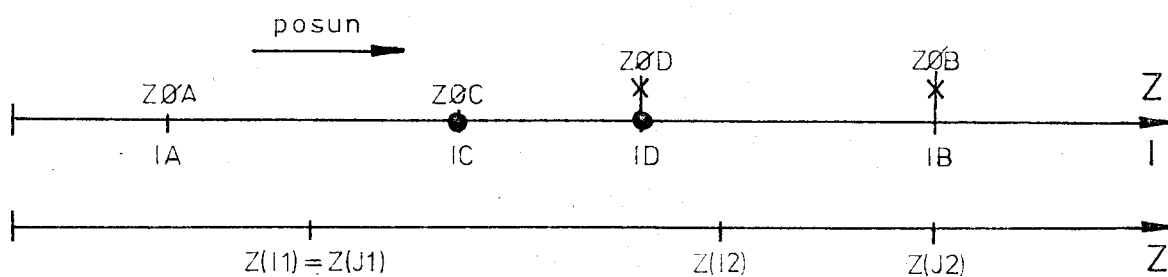
Indexy $IA - ID$ jsou indexy z dělení zlatým řezem, kdežto $I1, I2, J1, J2$ jsou indexy z rozložení intervalu.

Při testování mohou nastat celkem čtyři možnosti. To znázorňuje obr. 9a - d:

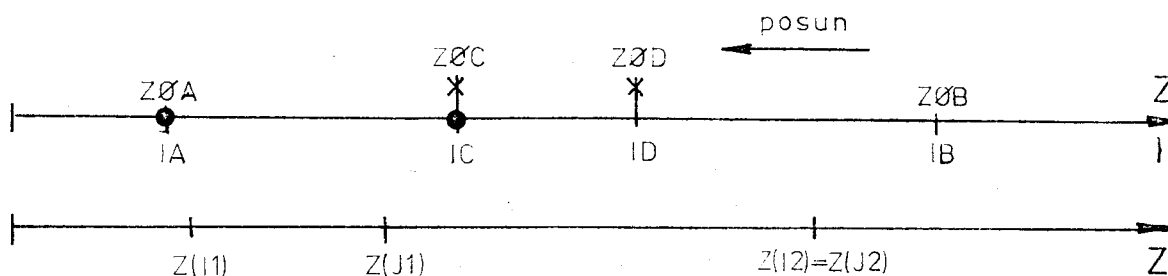
a/ dvouříd shoda



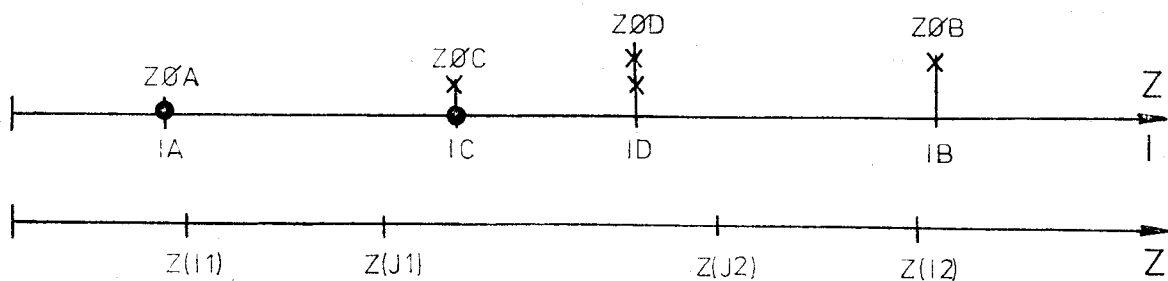
b/ posun doprava



c/ posun doleva



d/ bez posunu



obr. 9. Varianty posunu procedury POLY

● - označení intervalu, ve kterém se hledá v následujícím kroku

x - meze intervalu, které se ukládají

a/ dvojitá shoda

- platí $Z(I1) = Z(J1)$ a tedy $I1 = J1$ a zároveň $Z(I2) = Z(J2)$ a tedy $I2 = J2$. Z toho plyne, že je možné provést zmenšení intervalu oběma směry.

b/ posun doprava

- platí $I1 = J1$, ale $I2 \neq J2$ tedy není možné provést posun tímto směrem. Meze intervalu ID a IB se uloží do proměnné $MEZE$ a výpočet pokračuje v intervalu $\langle IC : ID \rangle$

c/ posun doleva

- je v podstatě shodný s variantou b/, jen posun se provádí vlevo, tedy $I1 \neq J1$ a $I2 = J2$

d/ bez posunu

- platí $I1 \neq J1$ a $I2 \neq J2$. Proto je zde nutné uložit obě dvojice mezí

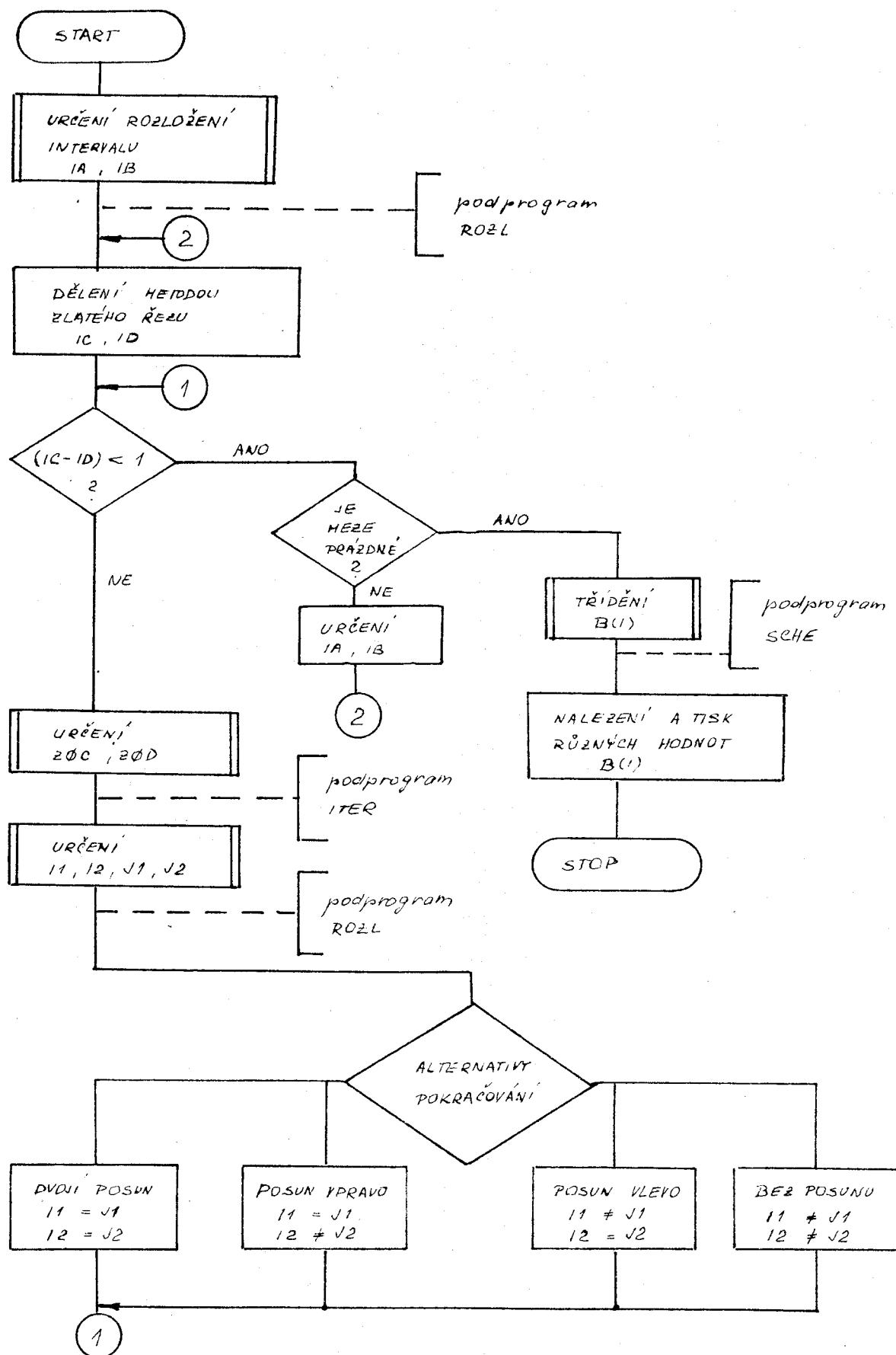
Po tomto se opět vracíme na dělení intervalu metodou zlatého řezu. Tímto postupem postupujeme stále až platí, že $(ID - IC) < 1$, tedy mezi těmito hodnotami již není žádné číslo.

Potom ukládám do mezí IA a IB hodnoty, které jsem dříve uložil do proměnné $MEZE$. Způsob ukládání a vybírání mezí intervalu připomíná sklípek. Tímto způsobem postupujeme až do toho okamžiku, kdy už je sklípek prázdný.

Při každém výpočtu odchodu $Z\emptyset J$ se tento ukládá do zvoleného pole $B(I)$. Nakonec je tedy nutné vyloučit všechny shodné parametry a vytisknout jen všechny rozdílné parametry polohy.

h/ Popis činnosti podprogramu NCRM

Podprogram provede nalezení minima a maxima funkce viz obr. 3, tedy hledá meze intervalu normality dat pro určitý parametr polohy. Vstupem je zde $Z\emptyset$ a výstupem jsou meze intervalu. Hledání obou extrémů se provádí metodou



VD2 - Hrubý vývojový diagram
procedura POLY

zlatého řezu. Tato metoda je popsána např. v [2]. Maximální počet iterací je zvolen 50. Počáteční meze intervalu pro minimum je 0 a $Z\emptyset$, pro maximum $Z\emptyset$ a $2 * Z\emptyset$. Přesnost je zde nutno brát ve velikosti $2 * EPS$, protože iterují zdola i zhora s přesností EPS .

i/ Popis činnosti podprogramu ITER

Podprogram provede iterační výpočet $Z\emptyset J$ podle vztahu /59/. Vstupem je zde počáteční hodnota iterace $Z\emptyset$ a výstupem je nalezený parametr polohy $Z\emptyset J$. Iterace se provádí modifikovanou Newtonovou metodou. Přesnost výpočtu je EPS a maximální počet iterací je 50.

j/ Popis činnosti podprogramu CLTN

Podprogram rozčlení nehomogenní datový soubor na homogenní podsoubory. Nejprve se vypočítají intervaly normality dat, pro každý parametr polohy. Poté se vypustí všechny prvky, které leží v průniku intervalů normality a ve třetí části se konečně rozčlení soubor na homogenní podsoubory. Tisknou se postupně všechny data, která leží v intervalu normality č. 1, intervalu normality 2, atd.

k/ Popis činnosti podprogramu SCHE

Podprogram provádí třídění upravenou Schellovou metodou. Tato metoda je popsána v [16]. Její úprava je provedena v [18]. Odtud je také v podstatě převzat daný program.

l/ Omezení programu EKON

omezení iterace

Jak již bylo řečeno, je maximální počet iterací volen 50. Toto je v programu pevně stanoveno. Při překročení počtu iterací se vypíše oznámení o překročení počtu iterace. Program však pokračuje dále s tím, že poslední nalezená hodnota se bere jako hodnota výstupní. Pokud bychom chtěli změnit počet maximálních iterací, je třeba změnit všechny rozhodovací příkazy tvaru IF (IP.GT.50)... na jiný.

omezení přesnosti

Přesnost je zadávána v datech, ale je pro všechny procedury stejná = figuruje v příkazu COMMON. Pokud chceme přesnost změnit jen v některé proceduře, je třeba upravit program násobením přesnosti v příkazech IF.

omezení velikosti dat

Data jsou zadávána ve tvaru F 7.3 a tedy jsou od - 99.999 do 999.999, pokud chceme pracovat s daty, které přesahují tento interval, je třeba změnit příslušné popisy FORMAT. Dále musíme vzít v úvahu platnost, ale i omezení 1. axiomu gnostické teorie. Všechna data musí být podle tohoto zákona kladná. Pokud pracujeme se zápornými daty, pak máme možnost tato data posunout do kladných hodnot, uskutečnit výpočet a přesunout výsledek zpět.

omezení univerzálnosti

Tento program - respektive jeho iterační procedura - je sestaven podle vztahu /59/. Univerzálnější vztah je však /60/. Tento vztah není výhodné při každém výpočtu používat, protože se čas výpočtu neúměrně prodlužuje.

omezení časové

Výpočty, jak je vidět z předchozího, jsou značně komplikované a tedy náročné na strojový čas počítače i pro poměrně malé soubory dat. /Výpočet pro 40 hodnot a 4 parametry polohy trvá na počítači EC 1033 asi 160 s CPU/.

3.1.4 Příklad výpočtu - program EKON

Data jsou volena podle [3] s úpravou na stávající formát F 7.3. Data v příloze 1/1 představují hustotu produktivity práce /vypočtenou z upravených vlastních výkonů na 1 pracovníka/ zjištěné v Severočeských dřevařských závodech za 1 měsíc roku 1985. Data mají rozměr tis. Kčs/ 1 pracovníka. Pořadí dat je dáno organizační strukturou podniku.

V dalších přílohách jsou provedeny změny dat se záměrem demonstrovat na příkladech citlivost srovnávaných odhadů k různým vlivům.

Původní data

V příloze 1/1 se hodnotí skutečná výchozí data. Všimněme si zde toho, že gnostický odhad má dvě hodnoty. To signalizuje skutečnost, že hodnota 9,722 tis. Kčs/1 prac. je odlehlá natolik, že je třeba analyzovat její kvalitativní odlišnost. Takováto "separace" uvedeně položky je v tomto případě nápadná i při pouhé prohlídce dat. Ukázalo se, že by separace nejmenší položky daného datového souboru nastávala i při hodnotách vyšších, až do 17,000 tis. Kčs/1 prac. V případě této hodnoty bychom se už těžko pouhou subjektivní úvahou dopracovali k rozhodnutí, že jde o odlehlou položku, zatímco položku 38,579 tis. Kčs/1 prac. bychom akceptovali jako neodlehlou od ostatních, rozprostřených až po 186,290 tis. Kčs/1 prac. My ovšem navíc usilujeme o automatizaci zpracování. To co může být evidentní pro zkušeného statistika, nemusí být evidentní pro počítač, dokud není adekvátní program.

Separace

V příloze 1/2 jsme vypustili položku 9,722 tis. Kčs/1 prac. z datového souboru. Dvojznačnost gnostického odhadu parametru polohy mizí a zůstává jediná hodnota, což signalizuje, že upravený soubor je homogenní, tedy můžeme ho považovat za 22 nepřesných číselných obrazů jediné "správné" charakteristické hodnoty odhadnutí jako 78,065 tis. Kčs/1 prac.

Poškození

Příloha 1/3 demonstruje situaci, kdy jsou původní data porušena tím, že se v posledních dvou položkách objeví místo jedničky čtyřka na nejvyšším řádu. Taková chyba je možná jak při ručním, tak i při automatizovaném zpracování dat. Takto odlehlé položky ovšem signalizuje separací další hodnoty odhadu parametru polohy, jejich počet stoupne na tři. Zajímavý

je však malý vliv této chyby na hodnotu prostředního parametru polohy.

Robustnost - citlivost

Vypustíme-li tedy poslední dvě poškozené hodnoty /viz příloha 1/4/ dostáváme případ, který demonstruje robustnost gnostického odhadu. Vypuštění těchto hodnot se ve srovnání s původními daty u gnostického parametru polohy téměř neprojeví. Jeho změna je o 0,20 %, kdežto aritmetický průměr se změní o 11,23 %. Ještě více se to projeví při vypuštění i třetí položky - viz příloha 1/5, kde vypouštím položku v řádku 1. Změna gnostického parametru polohy je o 0,48 %, zatímco aritmetický průměr se změní o 16,48 % vzhledem k původním datům.

Je třeba podotknout, že jsme volili odhad z_{eI} , tedy pro $M = 3$. Tento odhad je maximálně necitlivý. Zvolíme-li odhad s $M = 2$, nebo $M = 1$, pak dostaneme méně přísné kritérie - viz [20].

3.2 APLIKACE GNOSTICKÉ TEORIE K VÝPOČTU REGRESNÍCH KOEFICIENTU GNOSTICKOU METODOU

3.2.1 Odhad měřítka s

Dříve než přikročíme k odvození algoritmu, řešící výpočet regresních koeficientů regresní metodou, musíme si uvést modifikaci gnostické teorie. Tato modifikace je uvedena např. v [9], [11], [12], [13] a [15] v následujícím smyslu:

jedno datum z_i je dáno

$$z_i = z_0 e^{s \cdot \Omega_i} \quad /61/$$

kde s /scale parametr/ je měřítko společné pro všechna pozorování v daném souboru.

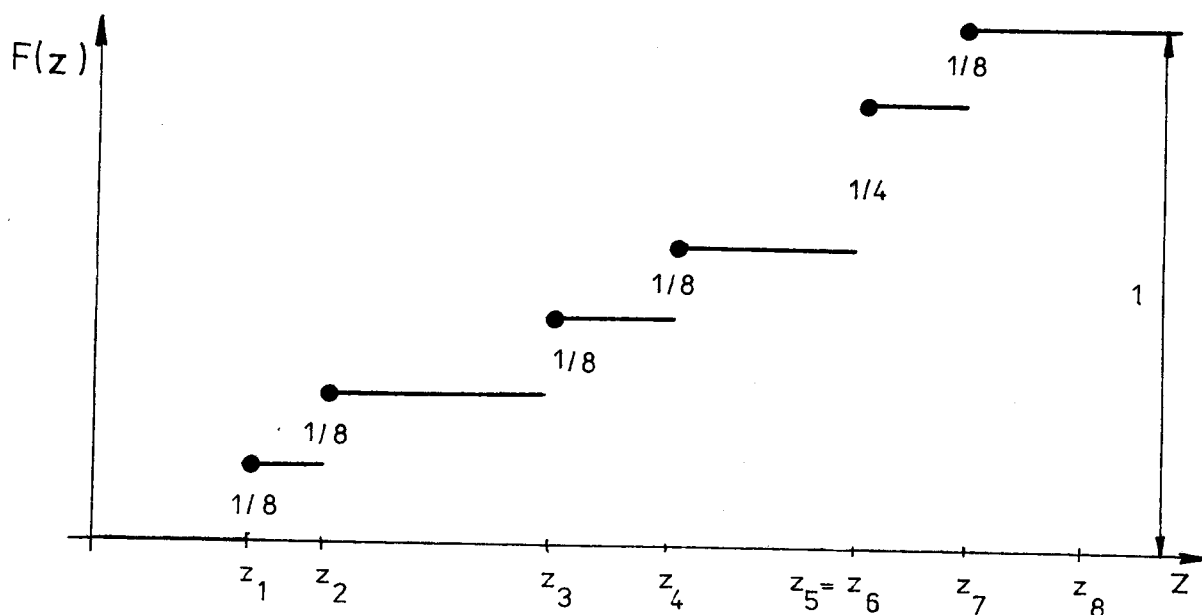
Rovnost /61/ lze zapsat

$$z_i^{\frac{1}{s}} = z_0^{\frac{1}{s}} e^{\Omega_i} \quad /62/$$

neboli místo veličin z_0, z_1 figurují $z_0^{\frac{1}{s}}, z_1^{\frac{1}{s}}$. Vhodnou volbou parametru s lze např. přiblížit gnostickou distribuční funkci s empirickou distribuční funkcí. Konkrétně takový odhad parametru s můžeme zapsat.

$$s = \operatorname{argmin}_s \max_j \max \{ |P_{ec} - F_{j-}|, |P_{ec} - F_{j+}| \} \quad /63/$$

kde F_{j-}, F_{j+} je empirická distribuční funkce. Empirickou distribuční funkcí se rozumí funkce, která je uvedena na příkladě viz obr. 10. Zde je vidět, že např. pro z_3 je empirická distribuční funkce $F_{3-} = \frac{1}{4}$ a $F_{3+} = \frac{3}{8}$



Obr. 10 Empirická distribuční funkce

P_{ec} je gnostická distribuční funkce viz /50/. Tedy můžeme psát pro určité datum z_j /pro $j = 1, \dots, n$ /

$$P_{ecj} = (1 + h_{ecj})/2 \quad \text{viz /50/}$$

dále odvodíme z kompozičního zákona /30/

$$h_{ecj} = W_{ej}^{-1} \sum_{i=1}^n h_{eij}$$

a pro $i = 1, \dots, n$ odvodíme

$$h_{eij} = (c_{ij}^{-2} - c_{ij}^2) / (c_{ij}^2 + c_{ij}^{-2})$$

kde

$$c_{ij} = (z_i / z_j)^{1/S}$$

S využitím kompozičního zákona viz /31/ a vztahu /46/ dostaneme

$$W_{ej}^2 = \left(\sum_i^n f_{ij} \right)^2 + \left(\sum_i^n h_{eij} \right)^2$$

kde, jak víme

$$f_{ij} = 2 / (c_{ij}^2 + c_{ij}^{-2})$$

Tímto je vlastně popsán algoritmus odhadu parametru měřítka s.

Rovnice /63/ je podmínkou nejlepššího přiblížení empirické distribuční funkce a gnostické distribuční funkce.

3.2.2 Aplikace gnostické metody pro identifikaci regresního modelu

Volme prozatím pro jednoduchost lineární regresní model; ideální hodnota z_{oi} je

$$z_{oi} = c_1 x_{i1} + c_2 x_{i2} + \dots + c_r x_{ir} \quad /64/$$

kde $x_{i1}, x_{i2} \dots$ jsou nějaké známé konstanty.

Toto platí pro

$$i = 1, 2, \dots, n \quad \text{kde } n > r$$

Naším úkolem je odhadnout $c_1, \dots, c_r$. Jako v případě odhadu parametru polohy hledáme extrém vzhledem k $c_1, \dots, c_r$ pro kritériální funkci

$$\varphi = \sum D(h_{ei}) \quad /65/$$

kde $D(h_{ei})$ je dvakrát diferenciovatelná funkce argumentu h_{ei} . Podle [13] volme

$$D(h_{ei}) = (1 - h_{ei}^2) / 2 \quad /66/$$

Z předchozího víme, že

$$h_{ei} = - \frac{\xi_i^4 - 1}{\xi_i^4 + 1} \quad /67/$$

a přitom

$$\xi_i = \left(\frac{z_i}{z_{oi}} \right)^{1/s} = \xi_i(c_1, \dots, c_r) \quad /68/$$

Navrhujeme se řešit úlohu iterací Newton-Raphsonovou tedy:

$$(d\varphi + d^2\varphi)_{\underline{c}} = 0 \quad /69/$$

/tj. φ se aproximuje Taylorovým rozvojem 2. řádu/
Označme

$$D'_i = \frac{d}{dh_{ei}} D(h_{ei}), \quad D''_i = \frac{d^2}{dh_{ei}^2} D(h_{ei}) \quad /70/$$

Potom je

$$d \gamma = D_i' d(h_{ei}) \quad /71/$$

$$d \gamma^2 = D_i'' (d h_{ei})^2 + D_i' d^2 h_{ei} \quad /72/$$

Diferenciál druhého řádu v rovnici /72/ lze zanedbat. /viz graf h_{ei} - obr. 5/. V oblasti počátku je inflexe a jak víme, se v místě inflexe druhá derivace blíží nule. Tím nedostávám sice čistou Newtonovu iterační metodu, ale toto zanedbání má vliv pouze na rychlost konvergence.

Derivujeme-li dále

$$d h_{ei} = \frac{d}{d \xi_i} h_{ei} \frac{d}{d z_{oi}} \xi_i \sum_{k=1}^r \frac{d}{d c_k} z_{oi} dc_k \quad /73/$$

$$\text{tedy} \quad h_{ei} = - \frac{\xi_i^4 - 1}{\xi_i^4 + 1}$$

$$\frac{d h_{ei}}{d \xi_i} = \frac{-8 \xi_i^3}{(\xi_i^4 + 1)^2} \quad /74/$$

a dále

$$\xi_i = \left(\frac{z_i}{z_{io}} \right)^{1/s}$$

$$\frac{d \xi_i}{d z_{io}} = \frac{1}{s_i} \left(\frac{z_i}{z_{io}} \right)^{1/s-1} \left(- \frac{z_i}{z_{io}^2} \right) = - \frac{1}{s} \frac{\xi_i}{z_{io}} \quad /75/$$

a konečně po dosazení dostaneme

$$d h_{ei} = - \frac{8 \xi_i^3}{(\xi_i^4 + 1)^2} \left(- \frac{1}{s} \frac{\xi_i}{z_{io}} \right) \sum_{k=1}^r x_{ik} dc_k =$$

$$= \frac{1}{s z_{is}} \frac{8 \xi_i^4}{(\xi_i^4 + 1)^2} \sum_{k=1}^r x_{ik} dc_k$$

a s využitím vztahu pro

$$f = \frac{2 \xi^2}{\xi^4 + 1}$$

dostaneme

$$d h_{ei} = \frac{1}{s z_{oi}} 2 f_i^2 \sum_{k=1}^r x_{ik} dc_k \quad /76/$$

Označíme-li $\underline{g}_i = \frac{1}{z_{oi}} (x_{i1}, x_{i2}, \dots, x_{ir})$ je

$$d h_{ei} = \frac{1}{s} 2 f_i^2 \underline{g}_i^T d\underline{c} \quad /77/$$

Potom dostaneme

$$d h_{ei}^2 = \frac{1}{s^2} 4 f_i^4 d\underline{c}^T \underline{g}_i \underline{g}_i^T d\underline{c}$$

Tyto vztahy dosadíme do /71/ a /72/ a můžeme psát:

$$d\varphi = \sum_i^n D_i' \frac{1}{S} 2 f_i^2 \underline{g}_i^T d\underline{c} \quad /78/$$

$$d\varphi^2 = \sum_i^n D_i'' \frac{1}{S} 4 f_i^4 d\underline{c}^T \underline{g}_i \underline{g}_i^T d\underline{c} \quad /79/$$

dosadíme-li tyto vztahy do /69/ dostaneme

$$\sum_i^n D_i'' 4 \frac{1}{S^2} f_i^4 d\underline{c}^T \underline{g}_i \underline{g}_i^T d\underline{c} = - \sum_i^n D_i' \frac{1}{S} 2 f_i^2 \underline{g}_i^T d\underline{c} \quad /80/$$

Jelikož /80/ má platit pro všechny $d\underline{c}$, musí platit

$$\sum_i^n D_i'' 2 f_i^4 d\underline{c}^T \underline{g}_i \underline{g}_i^T = \sum_i^n - S D_i' f_i^2 \underline{g}_i^T \quad /81/$$

a po transpozici

$$\sum_i^n D_i'' 2 f_i^4 \underline{g}_i \underline{g}_i^T d\underline{c} = - \sum_i^n S D_i' f_i^2 \underline{g}_i \quad /82/$$

a tedy konečně

$$d\underline{c} = - S \left(\sum_i^n 2 D_i'' f_i^4 \underline{g}_i \underline{g}_i^T \right)^{-1} \sum_i^n D_i' f_i^2 \underline{g}_i$$

a po vyjádření diferenciálu

$$\underline{c}_{k+1} = \underline{c}_k - S \left(\sum_i^n 2 f_i^4 D_i'' \underline{g}_i \underline{g}_i^T \right)^{-1} \cdot \left(\sum_i^n f_i^2 D_i' \underline{g}_i \right) / 83/$$

což je výsledný vztah pro iteraci koeficientů
 $\underline{c}_{k+1}$

Pokud zmíněné poznatky zobecníme, můžeme algoritmus hledání koeficientů modifikovat následujícím způsobem. Nechť vstupní regresní vektor je

$$z = F(\underline{c}, \underline{x}) \quad /84/$$

a jeho ideální hodnota

$$z_{oi} = F(\underline{c}, \underline{x}_{oi}) \quad /85/$$

Nechť tedy máme k-tou iteraci $\underline{c}_k$ odhadu $\tilde{\underline{c}}$ koeficientu $\underline{c}$ pak

$$h_{ei} = \left((\tilde{z}_{oi}/z_i)^{2/\tilde{S}} - (z_i/\tilde{z}_{oi})^{2/\tilde{S}} \right) / \left((\tilde{z}_{oi}/z_i)^{2/\tilde{S}} + \right. \\ \left. + (z_i/\tilde{z}_{oi})^{2/\tilde{S}} \right) \quad /86/$$

kde

$$\tilde{z}_{oi} = F(\underline{c}_k, \underline{x}_i)$$

a $\tilde{S}$ je odhad parametru $\tilde{S}$ podle /63/. Vstupem zde je $\tilde{z}_{oi}$.

Pro sloupcový vektor $\underline{g}_i$ můžeme psát

$$\underline{g}_i = \frac{1}{\tilde{z}_{oi}} (F'_{i1}, \dots, F'_{ir})^T \quad /87/$$

kde

$$F'_{ij} = \left(\frac{\partial F}{\partial c_j} \right)_{\underline{c}_k, \underline{x}_i} \quad /88/$$

a pro derivace kritériálních funkcí můžeme psát

$$D_i' = \left(\frac{\partial D}{\partial h_e} \right)_{z_o = z_{oi}} \quad D_i'' = \left(\frac{\partial^2 D}{\partial h_e^2} \right)_{z_o = z_{oi}} \quad /89/$$

a potom výsledný vzorec pro iteraci je /83/

3.2.3 Popis programu REGR

Výpis programu je v příloze 2/Ø.

a/ Funkce programu

Program řeší výpočet koeficientů regresního modulu nejprve metodou nejmenších čtverců, potom provádí odhad těchto koeficientů mostickou regresní metodou.

b/ Seznam volených podprogramů

SOBF - procedura odhadu měřítka s metodou zlatého řezu
 ITTX - iterační krok výpočtu koeficientů c
 INVM - inverze matice
 ODFS - iterační krok odhadu s, výpočet z rozdílu distribučních funkcí
 SCHE - třídění Schellovou metodou
 NASM - násobení matice

c/ Vsturní data

N, M - rozměry matice
 X (N, M) - matice hodnot nezávislých proměnných
 Z (N) - vektor pravých stran
 EPS - přesnost výpočtu při iteraci

d/ Výstupní proměnné a hole

I ... číslo řádku proměnné z_i
 Z (I) ... vsturní vektor pravých stran

$X(I, J)$... vstupní matice hodnot nezávislých proměnných
 $C(I)$... vektor koeficientů určených metodou nejmenších
 čtverců a gnostickou regresní metodou

e/ Popis činnosti hlavního programu

Hlavní program načítá vstupní data. Potom vypočítává a tiskne koeficienty metodou nejmenších čtverců podle známého vztahu

$$\underline{c} = (\underline{x}^T \cdot \underline{x})^{-1} \underline{x}^T \cdot \underline{z}$$

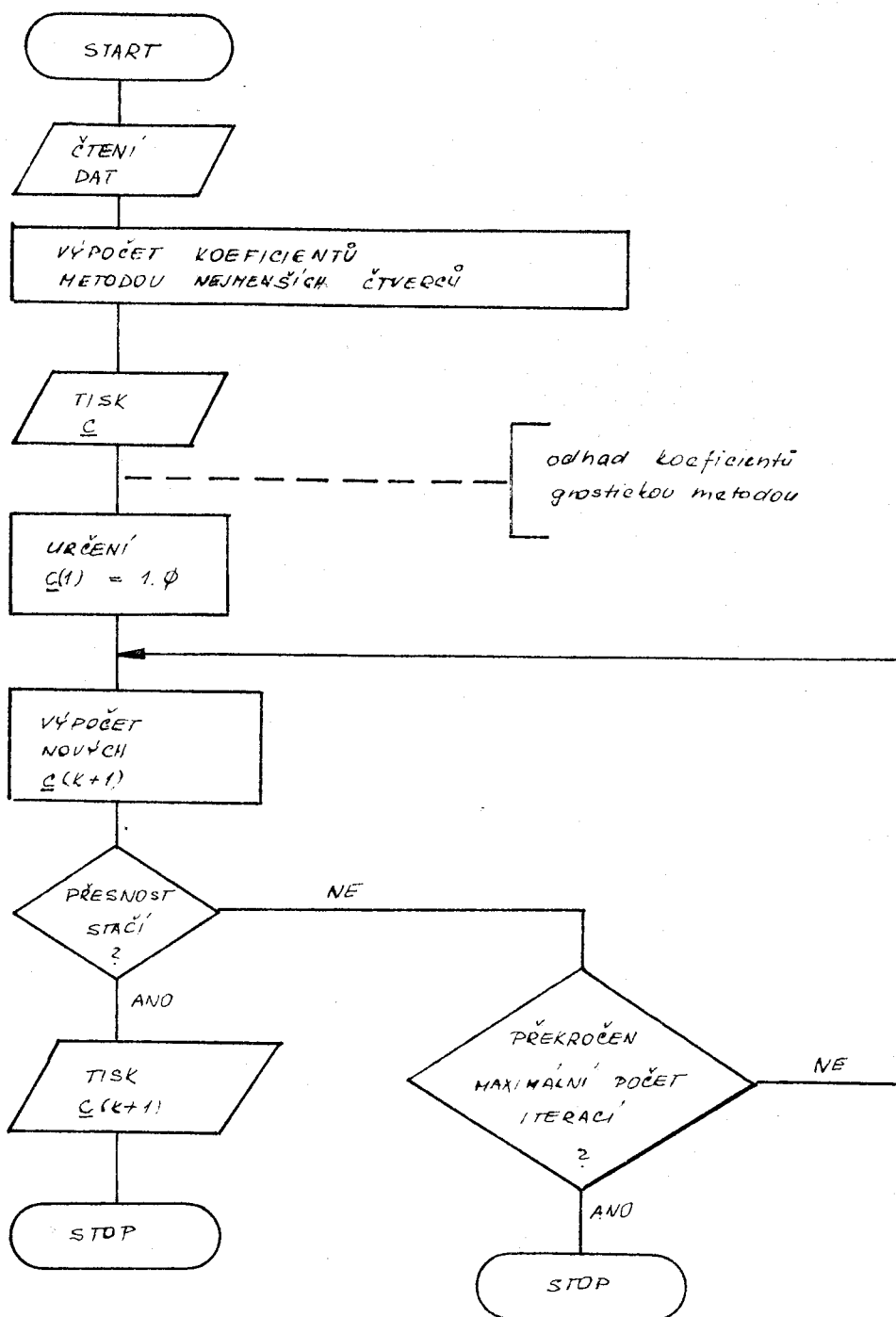
V další části se vypočítávají koeficienty gnostickou regresní metodou. Počáteční hodnotu koeficientu $\underline{c}$, bylo volena 1.0. Poté se provede odhad koeficientů $\underline{c}$, tyto koeficienty se porovnávají s hodnotou koeficientů, které byly nalezeny v minulém iteračním kroku. Výpočet se ukončí, až všechny koeficienty $\underline{c}$ dosáhnou požadované přesnosti EPS. Další možnost ukončení výpočtu je po překročení maximálního počtu iterací IPP. Maximální počet iterací byl v programu volen 50.

f/ Popis činnosti procedury odhadu měřítka S

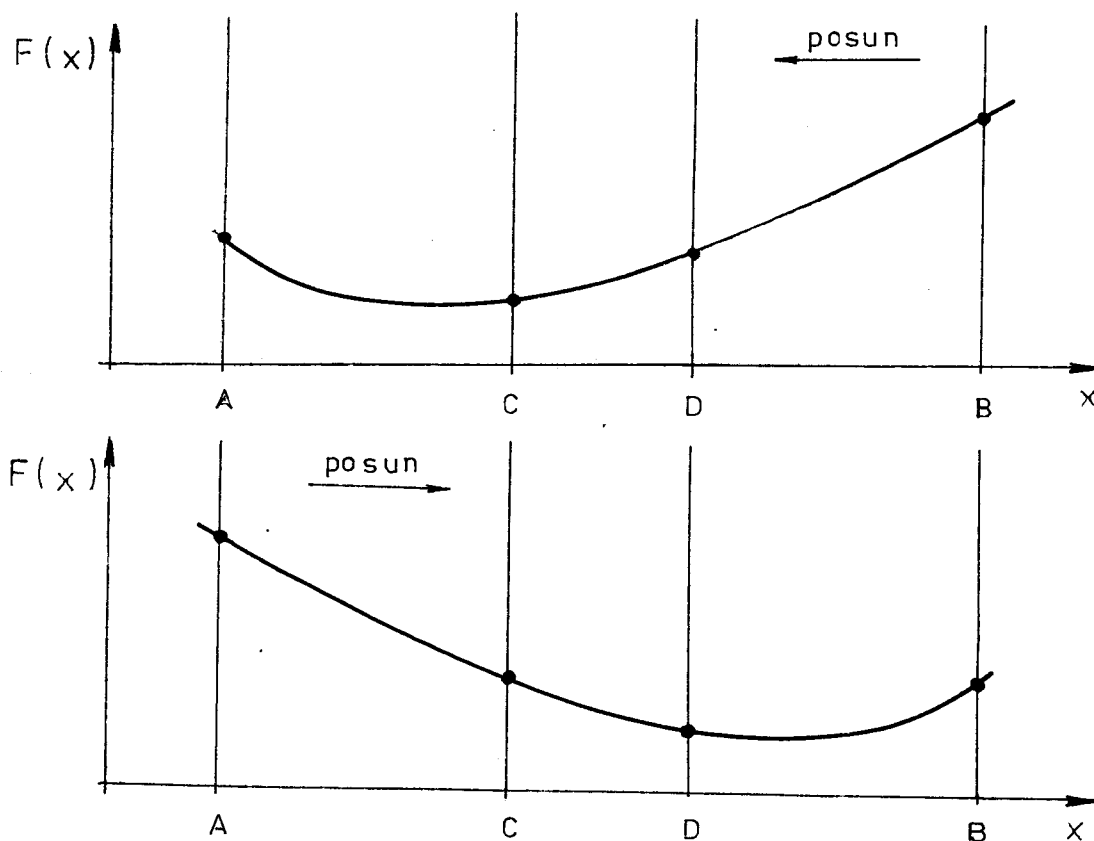
Vstupem do této procedury je vektor $\tilde{Z}_{oi}$, tedy $ZV(I)$. Tento vektor procedura nejprve vzestupně setřídí. Dostaneme setříděný vektor $ZSORT(I)$. Potom určíme empirickou distribuční funkci $F_+(I)$ tedy $FP(I)$ a $F_-(I)$ tedy $FM(I)$. Potom již hledáme měřítko S podle vztahu /57/. Určíme nejprve počáteční interval, ve kterém hledám minimum. Mohou nastat tři možnosti, které jsou nejlépe vidět na obr. 12.

Rozšiřování provádíme metodou zdvojnásobení kroku. Tato metoda je uvedena např. v [2]. Vlevo je interval omezen, neboť mezní měřítko je $S = \emptyset$. Toto rozšiřování provádíme tak dlouho, až platí varianta a/. Od tohoto okamžiku máme interval, ve kterém leží minimum.

Procedura dále provádí zpětný chod hledání extrému metodou zlatého řezu. Tato metoda je popsána opět v [2]. Při řešení nám mohou nastat dvě možnosti /viz obr. 11/.



VD3. Hrubý vývojový diagram
REGR - hlavní program



Obr. 11 Zmenšování intervalu metodou zlatého řezu

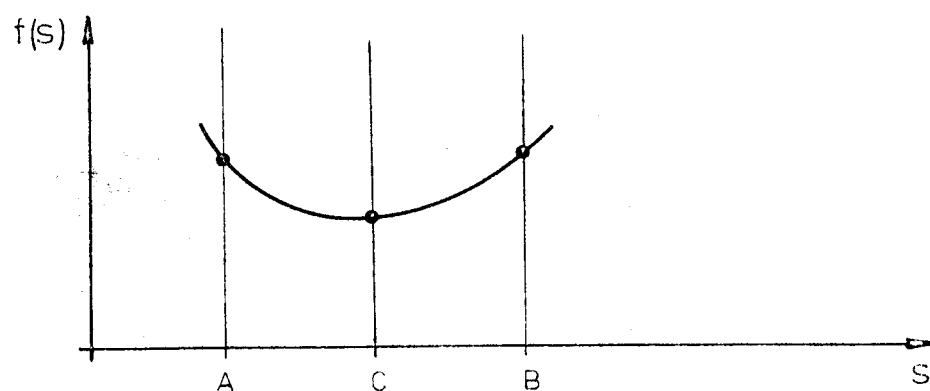
Zmenšování intervalu provádíme tak dlouho, dokud není interval $(C - D)$ menší než přesnost a zároveň rozdíl funkcí $S(C) - S(D)$ musí být menší než zadávaná přesnost.

Iterace se též ukončí pokud se přesáhne maximální počet iterací $IP > 50$. Pokud jsou splněny přesnosti nebo překročen maximální počet iterací, provede se přiřazení hodnoty parametru poloha S .

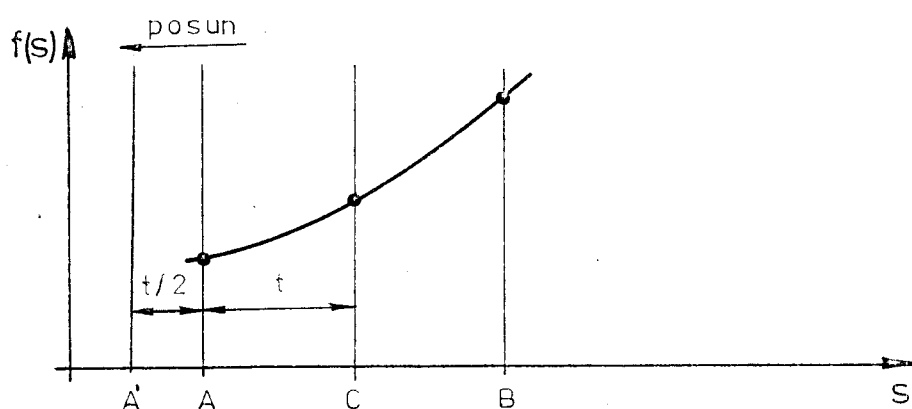
g/ Popis činnosti procedury ITTX

Vstupem této procedury jsou všechny vstupní data a dříve vypočtené koeficienty c_k . V první části procedury se definují derivace kritériální funkce $D(h_{ei})$. Kritériální funkce, která musí být dvakrát diferenciovatelná podle h_{ei} , jsem volil ve tvaru

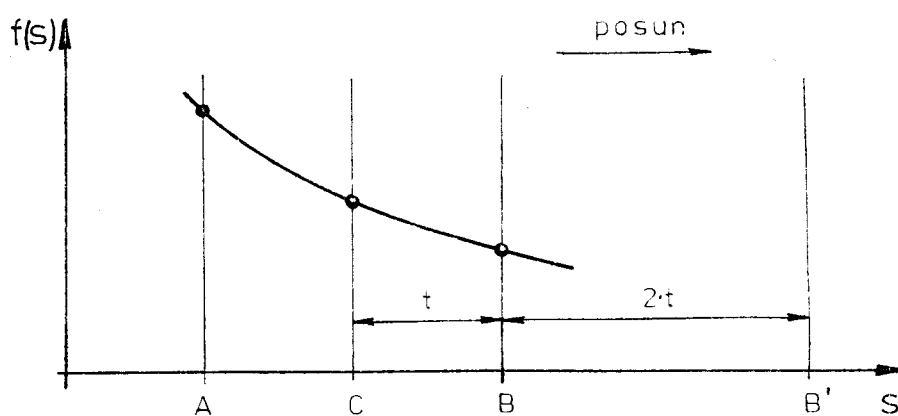
a/ minimum leží uvnitř intervalu



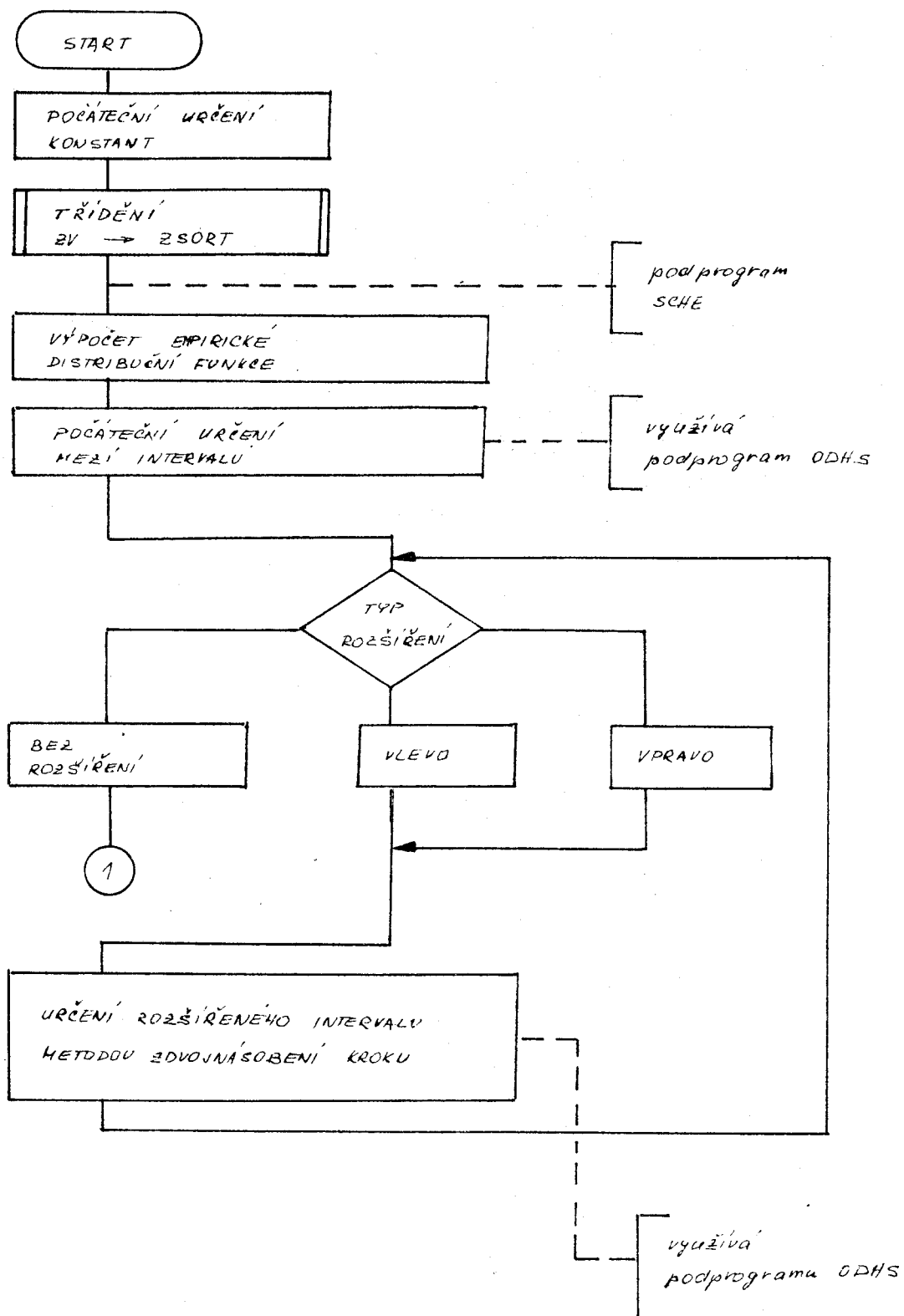
b/ minimum leží vlevo



c/ minimum leží vpravo



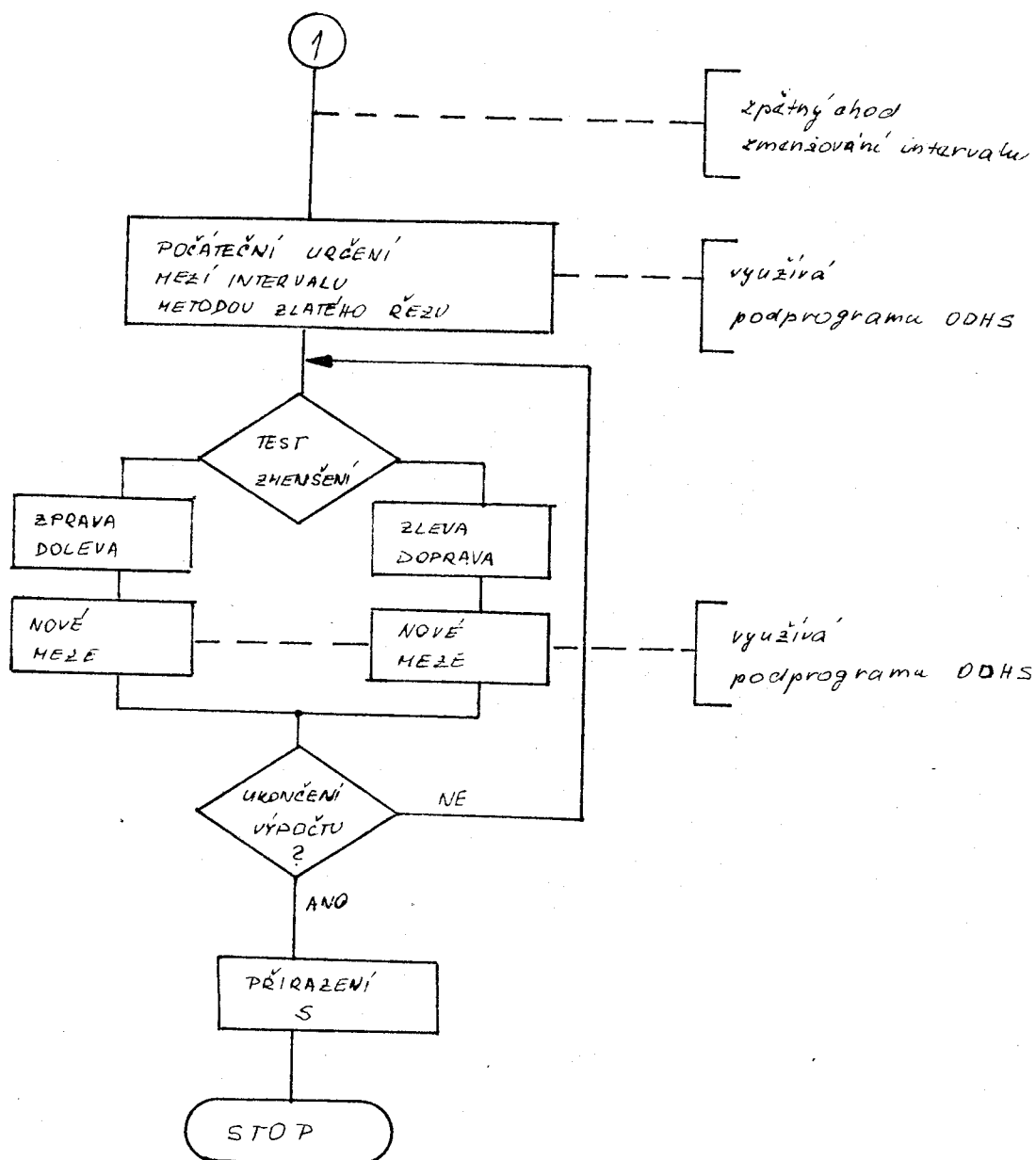
obr. 12. Rozšiřování intervalu



VD4 - Hrubý vývojový diagram

SODH - podprogram

/ 2 strany /



$$D(h_e) = (1 - h_e^2) / 2$$

a tedy derivace

$$D(h_e) = DC(HE) = -HE$$

$$D'(h_e) = DCC(HE) = -1$$

Pokud bychom chtěli použít jinou funkci, je třeba provést zásahy do této programové části. V další části se nulují sumační matice a vektor. Potom se vypočítávají pravé strany rovnice

$$\tilde{z}_{oi} = ZV(I) = FCE(\underline{c}_k, \underline{z}_i)$$

Z této definice pravých stran výpočtu odhad mřížka s. Definuje dále vektor $\underline{G}(J)$ a provádí výpočet sumační matice a vektoru podle vztahu /74/. Po vypočtení sumační matice provedu její inverzi a vynásobení sumačním vektorem. V poslední části procedury provedu určení nových koeficientů $\underline{c}_{k+1}$.

h/ Popis činnosti procedury ODHS

Vstupem je zde empirická distribuční funkce podle vztahů /49/ ... pro každé ZSORT(J), dále nachází maximální rozdíl empirické distribuční funkce a gnostické distribuční funkce ze všech hodnot ZSORT(J), pro $j = 1, 2, \dots, n$. Procedura vlastně počítá funkci

$$F_{\max}(S) = \max_j \max \{ |Pec - F_{j-}|, |Pec - F_{j+}| \}$$

Výstupem tedy je hodnota funkce $F_{\max}(S)$, pro určité S.

i/ Popis činnosti procedury INVM

Procedura řeší inverzi matice Jordanovou metodou. Tato metoda je popsána např. v [17]. V této metodě byly prove-

deny následující úpravy. Ve sloupci hledám maximální prvek a ten pak určuji na místo "pivota". Dělení pivotem provádím až na konci úprav. Vstupní matice je SUM (M, M) a výstupní inverzní matice je ve stejném poli SUM (M, M). Dodatková jednotková matice je v poli SUM (M, M + 1 + 2 M). Procedura předpokládá, že matice SUM (M, M) má hodnotu M a tedy je regulární.

j/ Popis činnosti procedury NASM

Podprogram provádí násobení matic. Tato procedura je ale převzata z [1].

k/ Popis činnosti procedury SCHE

Popis tohoto podprogramu je proveden v předcházející kapitole.

l/ Definice funkcí

Procedury - vnější funkce FCE a FCEC definují regresní model. V našem případě se jedná o lineární regresi, tedy

$$FCE J = \sum_{i=1}^n c_i x_{ij} \quad \text{viz vztah /75/}$$

a derivaci této funkce

$$FCEC (I, J) = \frac{\partial FCE (J)}{\partial c (I)} = x (I, J) \quad \text{viz vztah /79/}$$

Pokud bychom chtěli použít jinou regresi než lineární, pak je možné buď tuto funkci linearizovat - změnou matice SUM, nebo přeprogramovat dané definice funkcí. Pak si je třeba uvědomit, že neplatí metoda nejmenších čtverců.

m/ Omezení programu REGR

Omezení iterace

Stejně jako u programu EKON je i zde zadán mezní počet iteračních kroků IP = 50. Pokud je tento počet překročen, bere se právě nalezená hodnota /platí pro hledání parametru S/ nebo je výpočet ukončen /v hlavním programu/.

Omezení přesnosti

Přesnost je podobně jako u programu EKON zadávána v datech, ale je pro všechny procedury stejná. Pokud je chceme v různých procedurách různé, lze toto odstranit pře-programováním příslušných příkazů IF.

Omezení velikosti dat

Data jsou opět ve formátu F 7.3, viz program EKON a platí tedy tudíž skutečnosti, které byly uvedeny v předcházející kapitole.

Omezení rozměru dat

Protože jsou v programu poměrně často užívané matice je možné zpracování jen omezeného počtu dat. Volili jsme maximální rozměr modelu 20 x 20 s tím, že pokud by bylo třeba většího rozměru, je možné toto provést zvětšením rozsahu polí a buď zpracováním programu ze štítků s příslušným požadavkem na paměť, nebo vyřazením metody nejmenších čtverců s maticemi.

Omezení univerzálnosti

Jak bylo uvedeno výše, byla použita lineární regrese a určitá kritériální funkce D. V popisu procedur bylo uvedeno jakým způsobem lze použít zpracování i jiných regresních funkcí.

Omezení prvním axiomem

Je nutné mít na zřeteli, že metoda je funkční jen pro kladná data, jak plyne z prvního axiomu teorie. Pokud bychom chtěli pracovat i se zápornými daty, je nutné celou soustavu posunout směrem do kladných hodnot.

3.2.4 Příklady výpočtu - program REGR

Data byla vybrána záměrně tak, aby ukazovala "dobré" vlastnosti gnostické regresní metody.

Ideální model

V příloze 2/1 vidíme, že pro ideální model dat pracuje metoda nejmenších čtverců a gnostické naprosto stejně. To se ovšem u reálných dat nevyskytuje. Vždy se najde nějaké zašumění dat, v tomto případě vektoru z .

Drobná porucha

Pokud tedy ideální závislou proměnnou poškodíme malou poruchou - viz příloha 2/2 - dostáváme již rozdílné hodnocení metodou nejmenších čtverců a gnostickou metodou. Přesto jsou výsledky obou metod prakticky srovnatelné.

Mírná porucha

Zkusme nyní zašumět závislou proměnnou větší poruchou - viz příloha 2/3, ale ne zase rozhodující. Zde je vidět, že výsledky jsou sice ještě srovnatelné, ale odchylka metody nejmenších čtverců od ideální hodnoty je větší než u gnostické metody.

Střední porucha

Dále jsme vzali dvě závislé proměnné a způsobili poměrně velkou poruchu - viz příloha 2/4 - v řádku 5 a 6. Tím dostáváme naprosto chybné řešení metodou nejmenších čtverců. Gnostická metoda naproti tomu určí koeficienty, které odpovídají ideálnímu modelu dat.

Velká porucha

Vezměme opět proměnné x v řádcích 5 a 6 a zvětšeme tuto poruchu velmi razantně. Porucha je na první pohled vidět, neboť převyšuje ostatní hodnoty o celý jeden řád - viz příloha 2/5. Metoda nejmenších čtverců dává nyní zcela chybné výsledky, ale metoda gnostická je těmito "chybnými" daty téměř neovlivněna. Pokud porovnáme výsledky z přílohy 2/4 a 2/5 vidíme, že ovlivnění vzdálenějšími daty je mnohem menší než daty blízkými.

4. Z Á V Ě R

V práci jsou užity teoretické podklady autora metody Ing. P. Kovanice, CSc. Jde o metodu novou, která dosud není v povědomí odborné veřejnosti. Teorie se opírá o formální aparát speciální teorie relativity a termodynamiky. Pomocí teorie jsou vytvářeny robustní odhady ideální hodnoty polohy a jiné charakteristiky datového souboru. Ze statistického hlediska tyto odhady odpovídají v podstatě vývěrovému modu, tj. bodu největšího zhuštění.

Provedli jsme praktickou aplikaci této teorie na dvou případech. V prvním případě je provedeno odvození a výpočet parametru polohy a intervalu normality dat. Jak ukazují výsledky, je možné metodu použít při diagnostice chyb /interval normality dat/, při robustním výpočtu parametru polohy i odhalení více shluků v datovém souboru apod.

V druhém případě je řešen gnostický odhad regresních koeficientů. Podle dosahovaných výsledků můžeme soudit, že se tyto robustní regrese uplatní zejména při identifikaci a vyhodnocování výsledků měření. Použití teorie může být však daleko širší. Je možné tuto metodu použít jako robustního filtru, pro výpočet rozptylů dat apod.

Z výsledků obou příkladů vidíme, že metoda je vhodná pro malá soubory dat, které jsou méně či více poškozeny, všude tam, kde selhávají klasické statistické metody.

Další výhodou je, že gnostická teorie má dobré předpoklady automatizovat celý průběh výpočtů. To se samozřejmě může uplatnit při automatické identifikaci, zpracování dat, a nebo při automatickém řízení výrobních procesů, zejména nestabilních.

Naproti tomu je metoda nevhodná pro větší datové soubory. Důvodem je poměrná složitost výpočtů a tedy časová náročnost na počítač.

Vzhledem k vývojovým tendencím současných počítačů, kdy se zrychluje jak technické tak i programové vybavení a vzhledem k dobrým vlastnostem gnostické teorie se dá předpokládat, že se této metodě dostane širokého praktického použití.

5. L I T E R A T U R A

- 1 Vogel J., Programování v jazyku Fortran, SNTL, Praha, 1976
- 2 Hanuš B. a kol., Teorie automatického řízení II, skripta VŠST, Liberec, 1985
- 3 Hrdinová D., Kovanic P., Důvěřujete průřezu?, Mechanizace a automatizace, Praha, 1986
- 4 Kovanic P., Gnostical theory of small samples of real data, Problems of Control and Information Theory, Vol. 13, No. 4, 259 - 274, Praha, 1984
- 5 Kovanic P., Gnostical theory of individual data, Problems of Control and Information Theory, Vol. 13., No. 5, 304 - 319, Praha, 1984
- 6 Kovanic P., On relations between information and physics, Problem of Control and Information Theory, Vol. 13, No. 6, 383 - 399, Praha, 1984
- 7 Kovanic P., Novičová J., Srovnání statistických a gnostických odhadů parametru polohy na reálných datech, UTIA ČSAV, Praha, 1986
- 8 Kovanic P., Současný stav robustních programů pro gnostickou analýzu dat, UTIA ČSAV, Praha, 1986
- 9 Kovanic P., Základy gnostické teorie dat, Sborník ROBUST '84, MFF - UK, Praha, 1984
- 10 Kovanic P., Gnostika jako zobecnění statistiky, UTIA ČSAV, Praha, 1983
- 11 Kovanic P., Novičová J., On robust estimators worth to be applied to real data, UTIA ČSAV, Praha, 1986
- 12 Kovanic P., Pacovský J., Robust filtering and fault diagnosis by gnostical method, referát pro kongres IFAC, Mnichov, 1987
- 13 Kovanic P., A new theoretical and algorithmical basis for estimation, Identification and Control, Automatica - IFAC, 1986
- 14 Kovanic P., Gnostická teorie dat a její aplikace, ČSVTS, Ostrava, 1985

- 15 Kovanec P., Vlachý J., Gnostical analysis of international activities in physics, Czechoslovak Journal of Physics, Vol. 336, Praha, 1986
- 16 Olehle M., Král F., Švorc I., Tišer J., Programování, programovací jazyky a operační systémy, skripty VŠST, Liberec 1985
- 17 Olehle M., Tišer J., Praktické použití Fortran, NADAS, Praha, 1979
- 18 VTM - magazín - Proč a nač je počítač, Mladá fronta - Svoboda, Praha, 1987
- 19 Dvořák M., Gnostické metody vyhodnocování dat, SVOČ, KTK VŠST Liberec, 1986
- 20 Dvořák M., Vyhodnocování dat podle gnostické teorie, SVOČ, KTK VŠST Liberec, 1987