# Přednáška 3 – Model směsi distribucí

## Komponenty

- modely jednotlivých shluků

$$f_j(y_t|\Theta),\ j = 1, 2, \ldots, n_c$$

- $n_c$ – počet komponent
- statické vs. dynamické komponenty
- normální, kategorické, Poissonovy

Příklad: $f_1(y_t|\Theta)$ :

$$\begin{bmatrix} y_{1;t} \\ y_{2;t} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} e_{1;t} \\ e_{2;t} \end{bmatrix}$$

## Ukazovátko

- diskrétní náhodná veličina – ukazuje aktivní komponentu

$$c_t \in \{1, 2, \ldots, n_c\}$$

- aktivní komponenta v čase $t$ (aktuální shluk, kam patří data)
- model ukazovátka – přepínání komponent (kategorický, rovnoměrný)

| $c_t$ | 1 | 2 | ... | $n_c$ |
|---|---|---|---|---|
| $f(c_t\|\alpha)$ | $\alpha_1$ | $\alpha_2$ | ... | $\alpha_{n_c}$ |

- Shlukování – odhadujeme parametry komponent (identifikujeme shluky)
- Klasifikace – třídíme data do odhadnutých komponent (odhad ukazovátka)

# Příklad – model směsi statických normálních komponent

Komponenta 1: $f_1(y_t|\Theta)$

$$\begin{bmatrix} y_{1;t} \\ y_{2;t} \end{bmatrix} = \begin{bmatrix} 8 \\ 1 \end{bmatrix} + \begin{bmatrix} e_{1;t} \\ e_{2;t} \end{bmatrix}$$

Ukazovátko:

$c_t = 1$

Komponenta 2: $f_2(y_t|\Theta)$

$$\begin{bmatrix} y_{1;t} \\ y_{2;t} \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \end{bmatrix} + \begin{bmatrix} e_{1;t} \\ e_{2;t} \end{bmatrix}$$

Ukazovátko:

$c_t = 2$

Komponenta 3: $f_3(y_t|\Theta)$

$$\begin{bmatrix} y_{1;t} \\ y_{2;t} \end{bmatrix} = \begin{bmatrix} 4 \\ 12 \end{bmatrix} + \begin{bmatrix} e_{1;t} \\ e_{2;t} \end{bmatrix}$$

Ukazovátko:

$c_t = 3$

Model ukazovátka:

| $c_t$ | 1 | 2 | 3 |
|---|---|---|---|
| $f(c_t\|\alpha)$ | 0.3 | 0.4 | 0.3 |

# Program – generování multimodálních spojitých dat

```
clear, clc, close
nd=500; % počet dat
th=[8 1;0 5;4 12]'; % parametry komponent ve sloupcích
r = 1.1; % rozptyl
al=[.3 .4 0.3]; % parametry ukazovátka
for t=1:nd
    % generování ukazovátka
    c(t)=sum(rand(1,1)>cumsum(al))+1;
    y(:,t)=th(:,c(t))+sqrt(r)*randn(2,1); % výstup
end
% Výsledky
plot(y(1,:),y(2,:),'.')
xlabel('Veličina 1')
ylabel('Veličina 2')
```

- Co ovlivní vzdálenost mezi shluky?

# Rekurzivní Bayesovský odhad modelu směsi distribucí

Shlukování a klasifikace průběžně měřených dat – odhad:

- parametry komponent $\theta, r$, $\Theta$ nebo $\lambda$ (podle typu komponent)
- ukazovátko $c_t$ v čase $t$

## Základní princip – obecně:

- Inicializace komponent

V cyklu:

- Měříme data
- Vzdálenosti od jednotlivých komponent – proximity
- Normalizace proximit – váhy komponent
  (pravděpodobnost, že je komponenta aktivní)
- Maximální váha – bodový odhad ukazovátka
  (klasifikace)
- Update statistik s váženými daty
- Přepočet bodových odhadů podle typu komponent
- Jdeme na krok 1

Poznámka: shlukování a klasifikace dat online vs. offline

# Algoritmus rekurzivního odhadu modelu směsi normálních komponent:

Pro čas $t = 0$

1. Nastavíme počet komponent + jejich počáteční statistiky
2. Vypočteme počáteční bodové odhady parametrů $(\hat{\theta}_0)_j$, $(\hat{r}_0)_j$

Pro čas $t = 1, 2, \ldots$, pro každou komponentu

1. Měříme nová data $y_t$
2. Určíme proximity $m_j$ – dosadíme do komponent $(\hat{\theta}_{t-1})_j$, $(\hat{r}_{t-1})_j$ a $y_t$
3. Určíme váhy komponent $w_{j;t} = \frac{m_j}{\sum_{i=1}^{n_c} m_i}$
4. Bodový odhad ukazovátka: $\hat{c}_t = \arg\max_j w_t$
5. Update statistik s váženými daty:

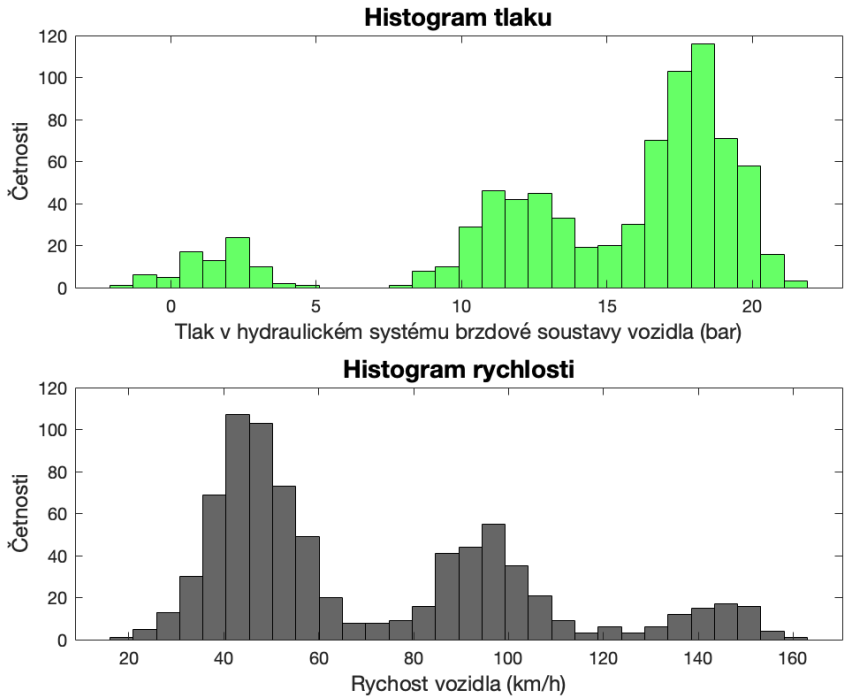
$$(V_t)_j = (V_{t-1})_j + w_{j;t} \begin{bmatrix} y_t \\ 1 \end{bmatrix} [y_t' \;\; 1], \quad (\kappa_t)_j = (\kappa_{t-1})_j + w_{j;t}$$

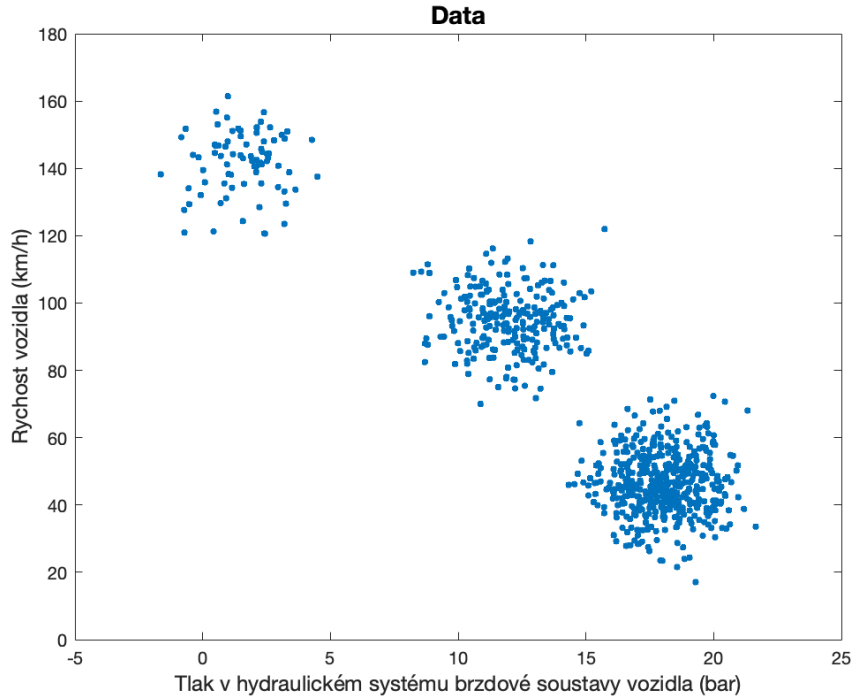
6. Přepočet bodových odhadů $(\hat{\theta}_t)_j$, $(\hat{r}_t)_j$ (viz přednáška 2)
7. Jdeme na krok 1

# Příklad: klasifikace stylu jízdy online

$y_{1;t}$ – tlak v hydraulickém systému brzdové soustavy vozidla (bar)
$y_{2;t}$ – rychlost vozidla (km/h)
$t$ – sekundy
$c_t$ – styl jízdy

**Inicializace:** počet komponent, počáteční odhady



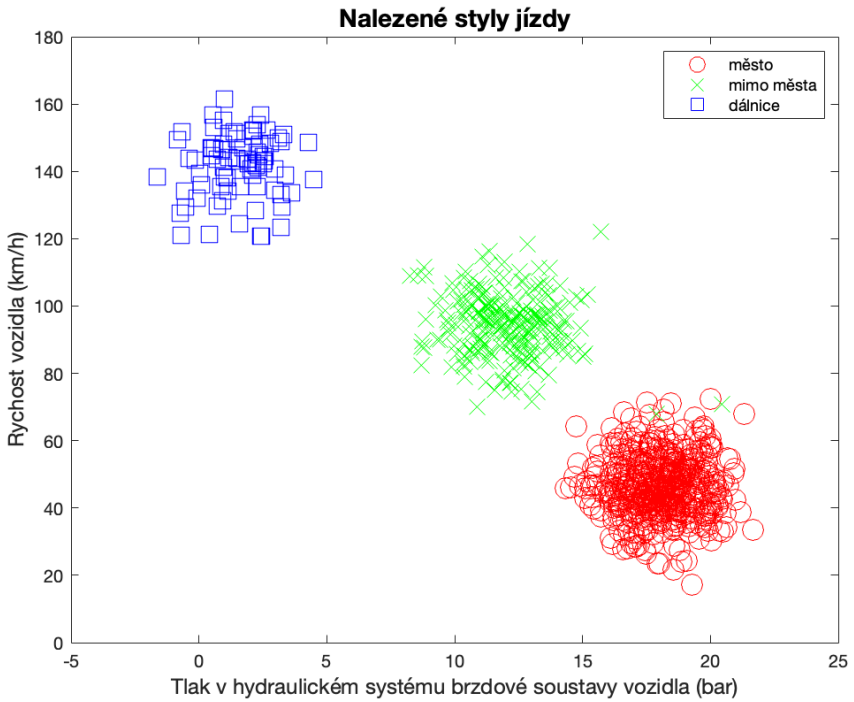
# Program – inicializace

```
% Initializace z histogramu
nc=3; % Počet komponent

% Počáteční odhad komponent
theta_odhad{1}=[3; 40];
theta_odhad{2}=[14; 90];
theta_odhad{3}=[18; 150];

ka=[1 1 1]; % počáteční počítadlo

for j=1:nc
  % Počáteční informační matice komponent
  V{j}=[theta_odhad{j};1]*[theta_odhad{j};1]';

  r_odhad{j}=5*eye(2,2);% Počáteční kovarianční matice
end
```

# Program – shlukování a klasifikace

```
% Cyklus shlukování a klasifikace
for t=1:nd
    for j=1:nc
        q(j)=GaussN(y(:,t),theta_odhad{j},r_odhad{j}); % proximity
    end
    w=q/sum(q); % váhy

    [nill,cp(t)]=max(w); % klasifikace (bodový odhad ukazovátka)


    for j=1:nc
        V{j}=V{j}+w(j)*[y(:,t);1]*[y(:,t);1]'; % Update statistik
        ka(j)=ka(j)+w(j);

        % Rozklad informační matice
        Vnn = V{j}(1:ny, 1:ny);
        Vy = V{j}(end, 1:ny);
        Vn = V{j}(end, end);

        % Bodové odhady regresních koeficientů
        theta_odhad{j} = Vy / Vn; % nebo inv(Vn)*Vy

        % Bodové odhady kovarianční matice
        if t>50
            r_odhad{j} = (Vnn - Vy' * inv(Vn) * Vy) / ka(j);
        end
    end
```

# Výsledky





## Poznámky:

- multimodalita dat
- inicializace
- překrývající se shluky
- modely ukazovátka
- online = trénovací a ~~testovací~~ data
- přesnost klasifikace ?

# Program – validace klasifikace z predikce z komponent

```
    % validace - predikce z komponent
    bb=[0;0];
    for j=1:nc
        for i=1:ny
            bb(i)=bb(i)+w(j)*theta_odhad{j}(i);   % vážený průměr komponent
        end
    end
    yp(:,t)=bb; % predikce
end
```

# Výsledky – predikce z komponent