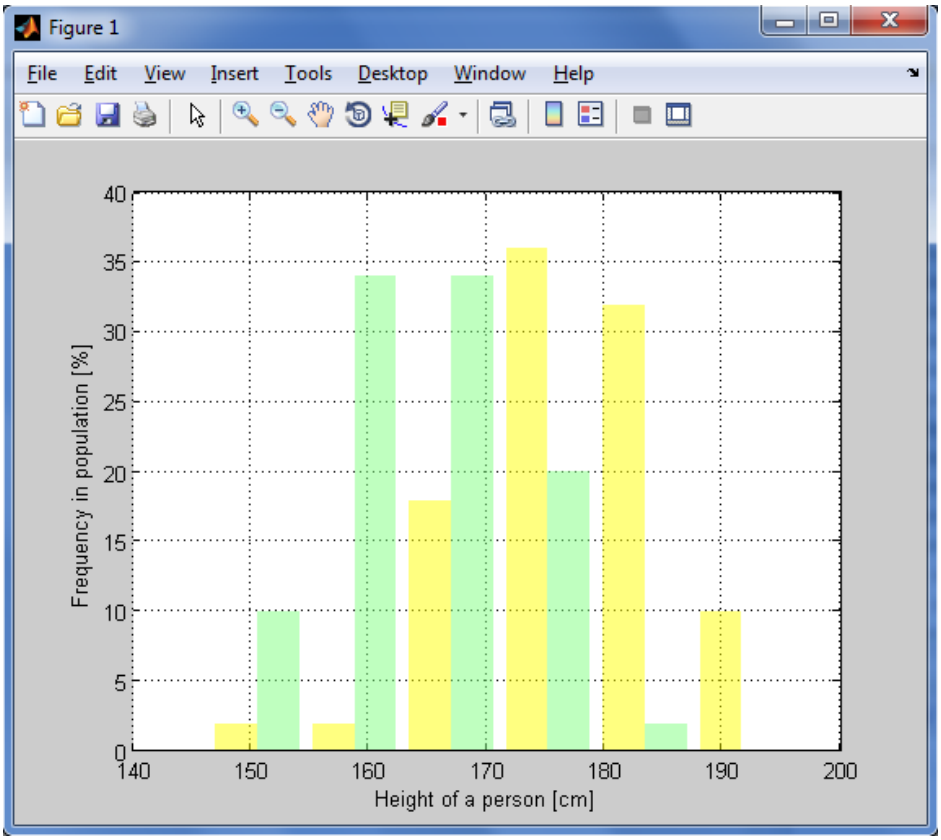
Úloha 3 – EM algoritmus

1. Seznamte se s daty v souboru **`height.csv`**, který obsahuje tělesnou výšku vzorku 100 lidí, Američanů ve věku mezi 20 a 29 lety. Kromě výšky lidí (1. sloupec) obsahují data i jejich pohlaví (2. sloupec). Každý záznam tvoří jeden řádek tabulky.
2. Prohlédněte si dokumentaci k přiložené funkci **`dataplot(data)`**, která načtená data vykreslí do grafu: **`>> data = csvread('height.csv.txt'); dataplot(data);`**
3. Implementujte EM algoritmus pro maximum-likelihood optimalizaci parametrů směsi dvou normálních rozdělení. Popis algoritmu naleznete ve třetí přednášce (str. 21-24).
   - Vstupem algorimu bude první sloupec načtených dat (druhý sloupec můžete použít pro zpětnou kontrolu). Vhodně zvolte počáteční parametry obou rozložení.
   - Pokud Váš algoritmus vrátí matici 2 *x* 2 ve formátu

     params = (μ ženy σ ženy
     μ muži σ muži)

     můžete pro vykreslení obou rozdělení použít příkaz **`>> dataplot(data, params);`**
4. Vyvořte protokol o rozsahu cca. 1 strany A4, která shrne Vaši práci a analyzuje výsledky. Doporučený obsah:
   - grafy obou gaussovských rozložení v několika počátečních iteracích algoritmu a stav po konvergenci
   - počet iterací algoritmu (dochází-li k velkému rozptylu hodnot pro různá počáteční nastaveni, spusťe algoritmus několikrat a výsledek vyhodnoťte statisticky)
   - diskuze o vlivu prvotního přiřazení parametrů na jejich výsledné hodnoty
   - rozbor, zda lze mezi výškou mužů a žen pozorovat statisticky významný rozdíl (využijte druhý sloupec vstupních dat a závěry z předchozích bodů)
   - poznámky k implementaci

Graf načtených dat:
`>> data = csvread('height.csv.txt'); dataplot(data);`



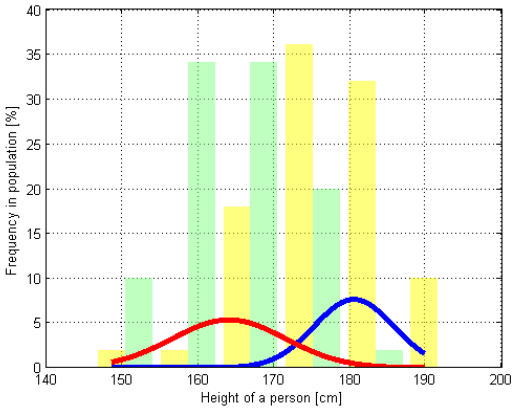
Zpracovala: **Radoslava Jandová** E-mail: **jandora1@fel.cvut.cz**
Obor: **STM, Softwarové inženýrství** Semestr: **zimní 2011/2012**

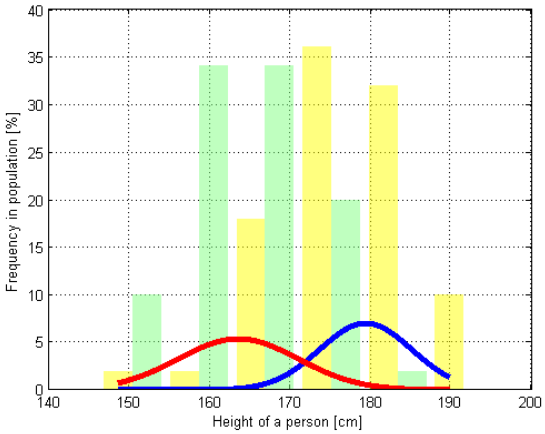
Implementace v aplikaci Matlab:

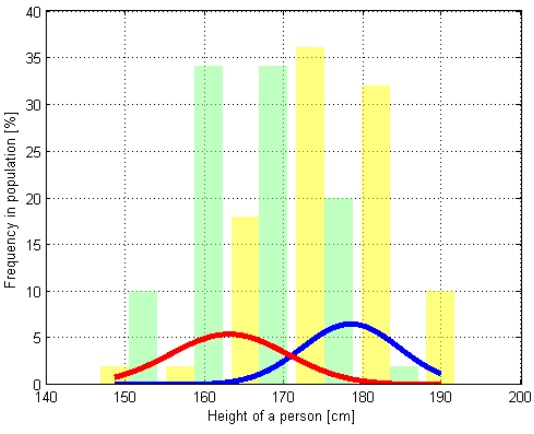
```
function [ vysledek ] = em( data )
% em(data(:,1)) = dle zadani pouzijeme prvni sloupec dat
% data = csvread('height.csv.txt'); = inicializace dat

% inicializace pocatecnich parametru podle zadani ulohy
uMuz = 200.0;
uZena = 150.0;
oMuz = 10.0;
oZena = 10.0;
convergence = 0.00001;          % convergence = presnost algoritmu
nextStep = 1;
pocetIteraci = 0;

while nextStep == 1
    hodnotaM = uMuz;
    hodnotaZ = uZena;
    pocetIteraci = pocetIteraci + 1;            % pocet iteraci
    [uMuz, uZena, oMuz, oZena] = emIteration(data, uMuz, uZena, oMuz, oZena);
    if((abs(hodnotaM - uMuz) < convergence) && (abs(hodnotaZ - uZena) < convergence))
        nextStep = 0;
    end

    % nastaveni iteraci
    % if (pocetIteraci == 10)
    % nextStep = 0;
    % end
end

fprintf('Number of iterations: %d\n', pocetIteraci);
vysledek = [uMuz oMuz; uZena oZena];
return
end

% em algoritmus
function [uMuz, uZena, oMuz, oZena] = emIteration(data, uMuzIt, uZenaIt,
oMuzIt, oZenaIt)
uMuz = uMuzIt;
uZena = uZenaIt;
oMuz = oMuzIt;
oZena = oZenaIt;

length = size(data, 1);                    % pocet hodnot
probabilities1 = zeros(1, length);
probabilities2 = zeros(1, length);

% expectation krok - viz vzorec z prednasky
for i = 1:length
    temp1 = expectationForDataPart1(data(i), uMuz, oMuz);
    temp2 = expectationForDataPart1(data(i), uZena, oZena);
    probabilities1(i) = temp1 / (temp1 + temp2);
    probabilities2(i) = (1.0 - probabilities1(1, i));
end

% maximization krok - viz vzorec z prednasky
uMuz = (probabilities1 * data) / sum(probabilities1);
uZena = (probabilities2 * data) / sum(probabilities2);
oMuz = sqrt((probabilities1 * ((data - uMuz).^2.0)) / sum(probabilities1));
oZena = sqrt((probabilities2 * ((data - uZena).^2.0)) /
sum(probabilities2));
```

```
return
end

function [ vysledek ] = expectationForDataPart1(data, mean, stdDev)
% pomocna fce pro vypocet expectation kroku
vysledek = (exp(-(((data - mean)^2.0) / (2.0 * stdDev^2.0))) * 0.5) /
sqrt(2.0 * pi * stdDev);
return
end
```

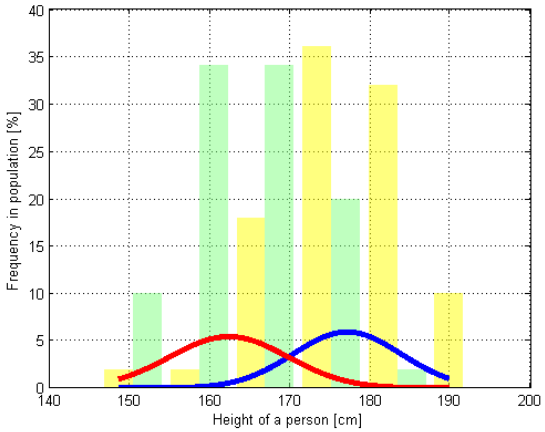
Přehled změn podle počtu iterací:

`>> em(data(:,1))`
`Number of iterations: 1`
 	
`ans = 180.4965  5.2399`
`  164.0837  7.4728`
 	
`>> dataplot(data, em(data(:,1)));`
`Number of iterations: 1`
	

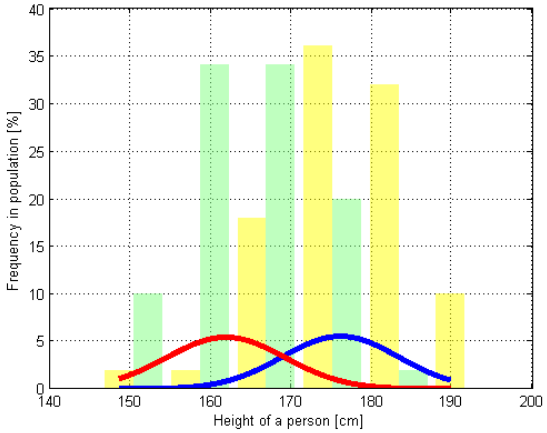
`>> em(data(:,1))`
`Number of iterations: 2`
 	
`ans = 179.3628  5.7099`
`  163.5752  7.4414`
 	
`>> dataplot(data, em(data(:,1)));`
`Number of iterations: 2`
	

`>> em(data(:,1))`
`Number of iterations: 3`
 	
`ans = 178.4129  6.1487`
`  163.0996  7.3916`
 	
`>> dataplot(data, em(data(:,1)));`
`Number of iterations: 3`
	

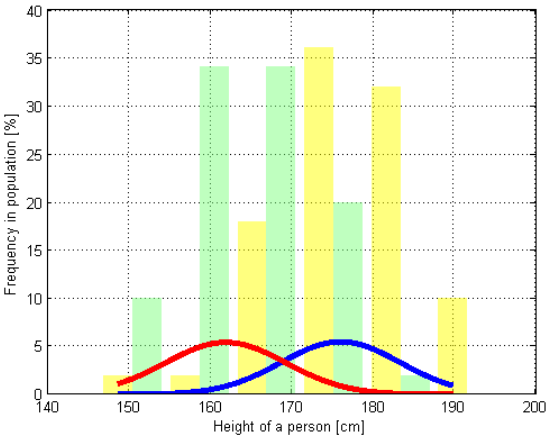
```
>> em(data(:,1))
Number of iterations: 5

ans = 177.0906    6.7379
      162.3929    7.3481

>> dataplot(data, em(data(:,1)));
Number of iterations: 5
```




```
>> em(data(:,1))
Number of iterations: 10

ans = 176.1025    7.2212
      161.9127    7.3626

>> dataplot(data, em(data(:,1)));
Number of iterations: 10
```

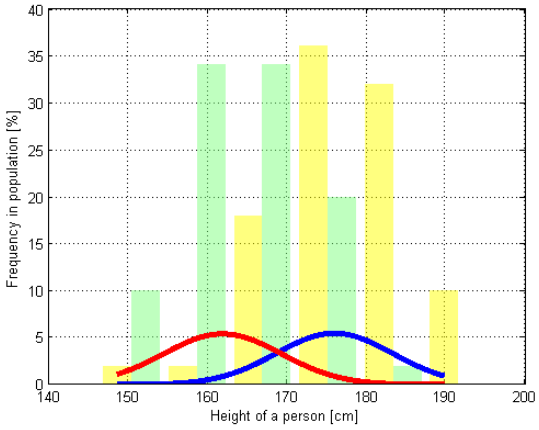



```
>> em(data(:,1))
Number of iterations: 15

ans = 175.9800    7.2977
      161.8967    7.3900

>> dataplot(data, em(data(:,1)));
Number of iterations: 15
```

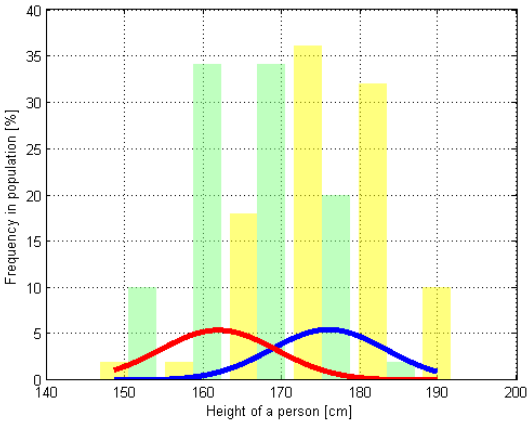


Zpracovala**: Radoslava Jandová**
Obor: **STM, Softwarové inženýrství**
E-mail: **jandora1@fel.cvut.cz**
Semestr: **zimní 2011/2012**

```
>> em(data(:,1))
Number of iterations: 25

ans = 175.9579    7.3153
      161.9070    7.4034

>> dataplot(data, em(data(:,1)));
Number of iterations: 25
```

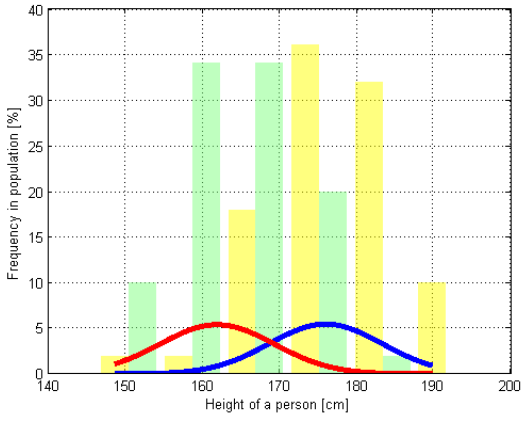



```
>> em(data(:,1))
Number of iterations: 35

ans = 175.9569    7.3163
      161.9081    7.4045

>> dataplot(data, em(data(:,1)));
Number of iterations: 35
```



Stav po konvergenci:

```
>> em(data(:,1))
Number of iterations: 39

ans = 175.9568    7.3164
      161.9082    7.4045

>> dataplot(data, em(data(:,1)));
Number of iterations: 39
```



Pokud algoritmus spustíme podle počtu iterací, dochází k výrazným změnám u iterací 1 - 5. Od 10té iterace již nejsou markatní rozdíly ve výsledných křivkách. Z toho vyplývá, že čím více iterací provedeme, tím přesnější výsledek získáme.

Při zadání vstupních dat muži = 200 a ženy = 150 zjistíme po první iteraci, že max výška u mužů dosahuje 180,5 a u žen 164. Průměrná výška mužů a žen pak činí 172,25.
Po konvergenci činí max výška u mužů cca 176 a u žen na 161,9. Tzn. průměrná výška činí 168,95. Z toho vyplývá, že čím více dat je zpracováno, tím více dochází ke snížení rozdílu ve výšce mužů a žen, řičemž větší rozdíl je v max. výšce mužů, který se snížil ze 180,5 na 176, zatímco u žen dochází k poklesu ze 164 na 161,9.

Zpracovala**: Radoslava Jandová** E-mail: **jandora1@fel.cvut.cz**
Obor: **STM, Softwarové inženýrství** Semestr: **zimní 2011/2012**