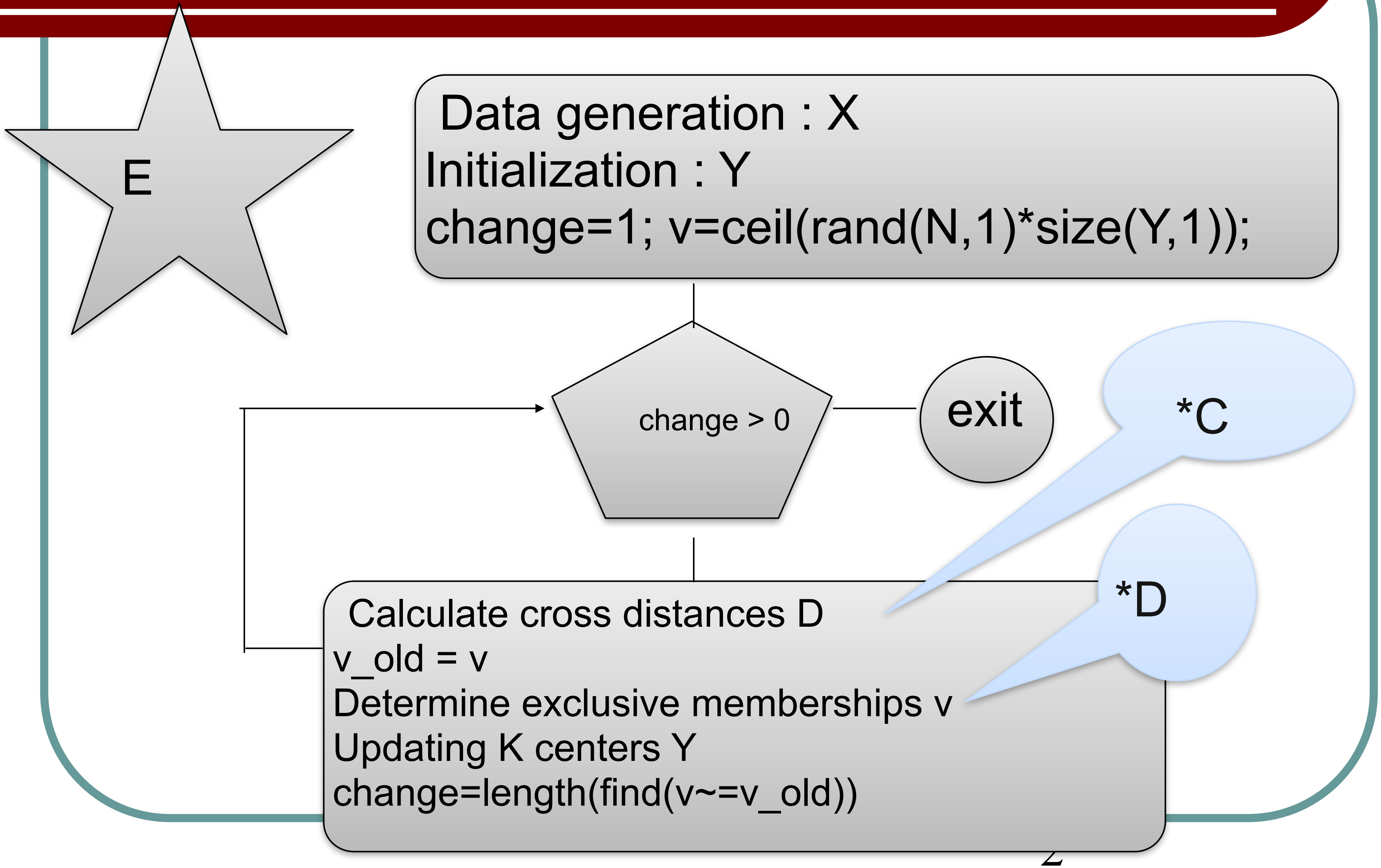
Lecture 9 Advanced Clustering

- Clustering
  - Exclusive memberships
  - Quantization & maximization

  - Overlapping memberships
  - Expectation & maximization

  - From expectation to quantization

An iterative approach for K-means



E

Data generation : X
Initialization : Y
change=1; v=ceil(rand(N,1)*size(Y,1));

change > 0

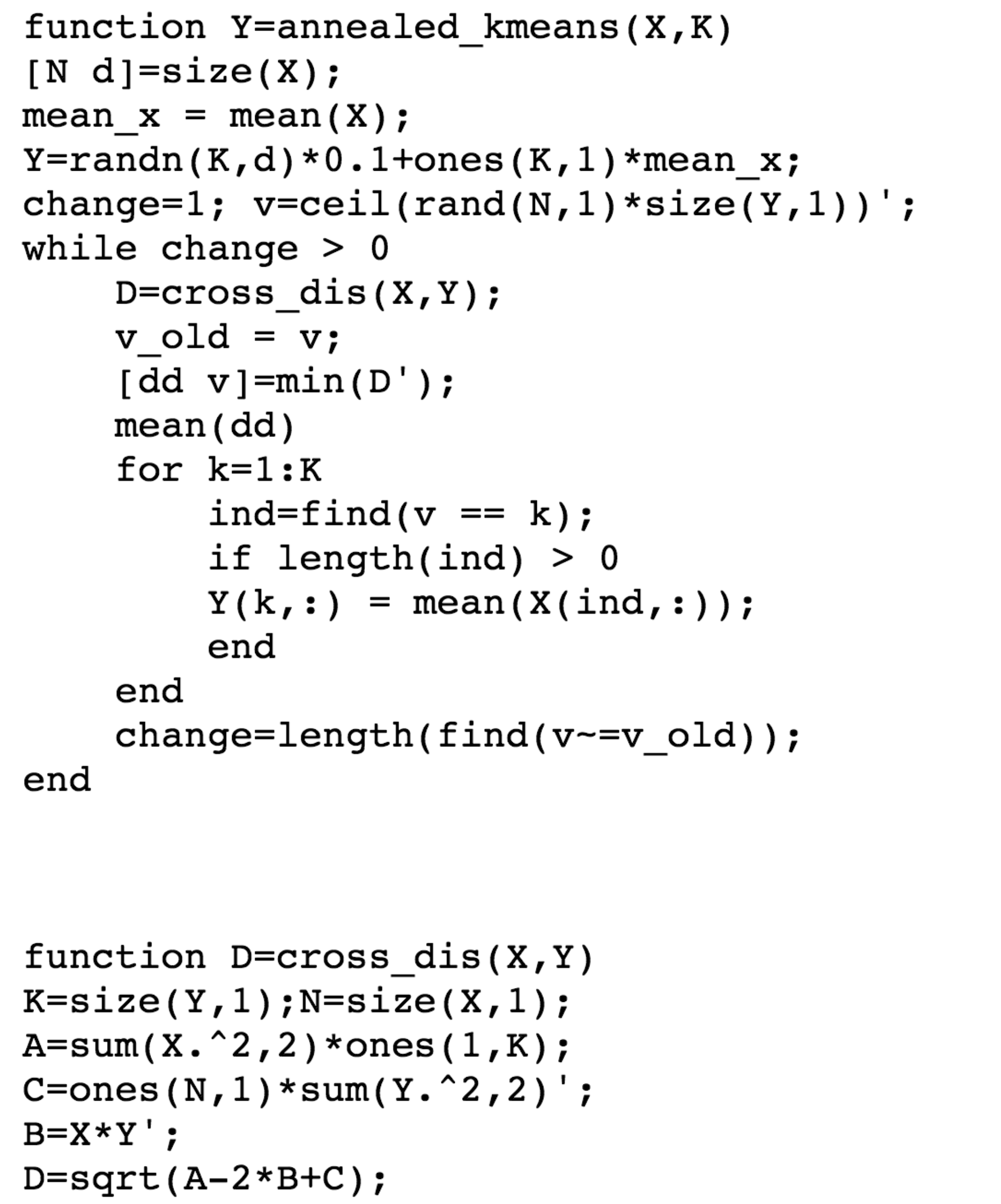
exit

*C

*D

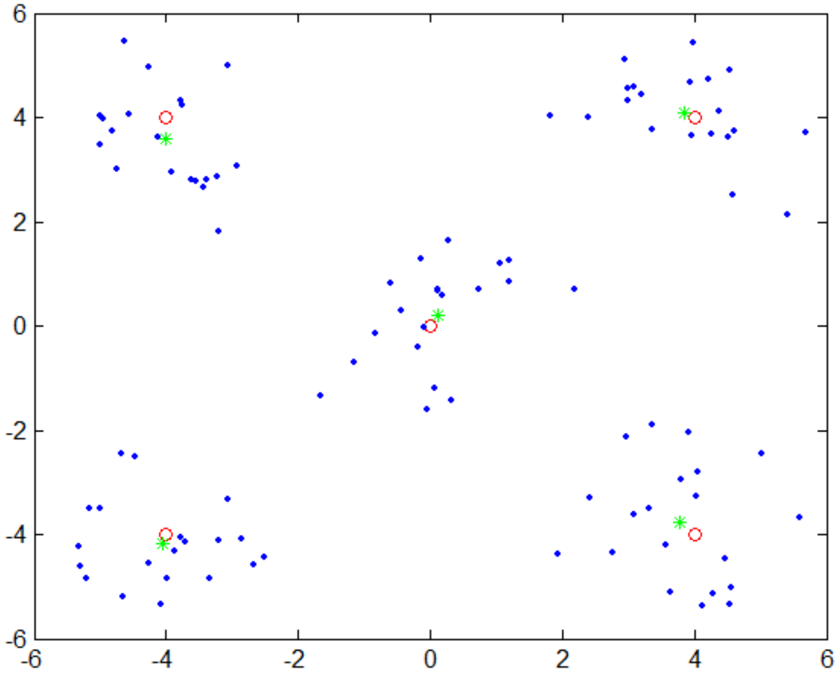
Calculate cross distances D
v_old = v
Determine exclusive memberships v
Updating K centers Y
change=length(find(v~=v_old))

```
y=[0 0; 4 4; -4 4; -4 -4;4 -4];
[K,d]=size(y);
X=[  ];
for i=1:K
    Xi=randn(20,2) + ones(20,1)*y(i,:);
    X=[X;Xi];
end
plot(y(:,1),y(:,2),'ro');
hold on; plot(X(:,1),X(:,2),'.');
```

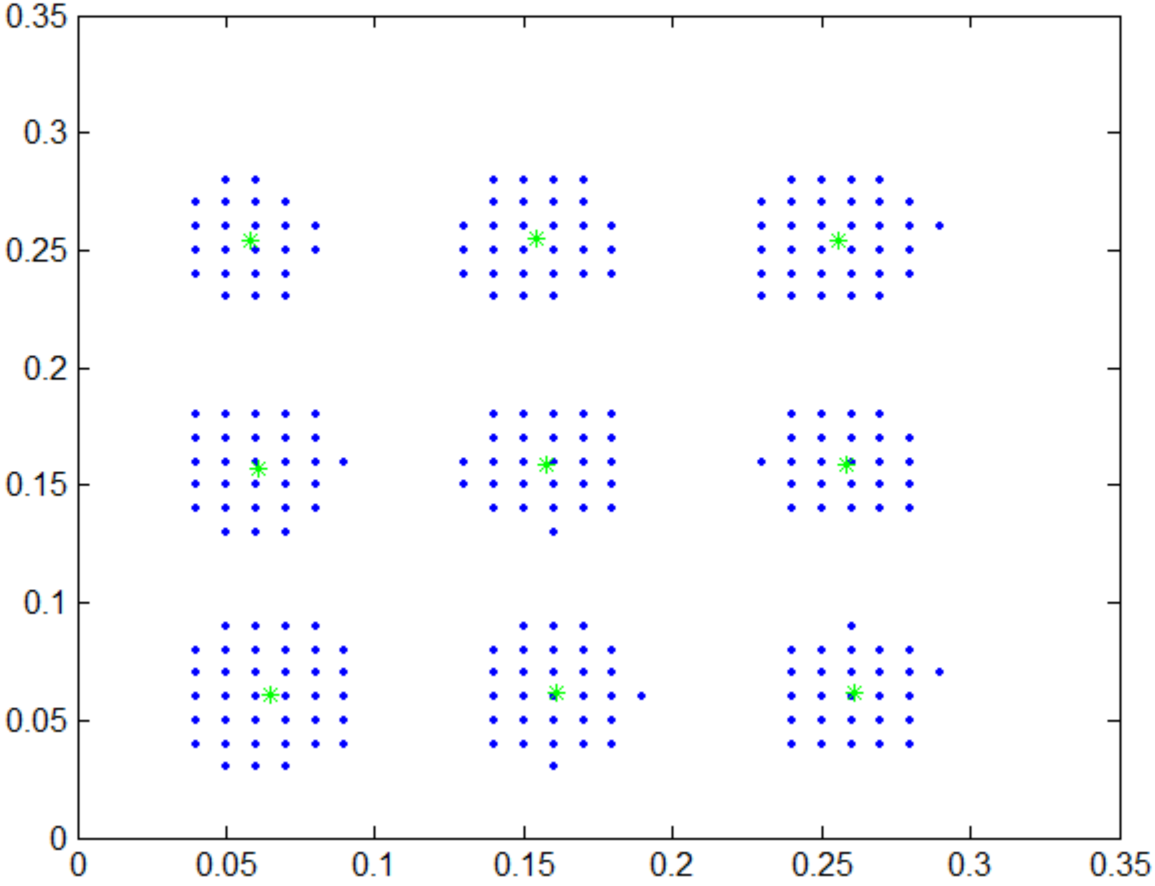


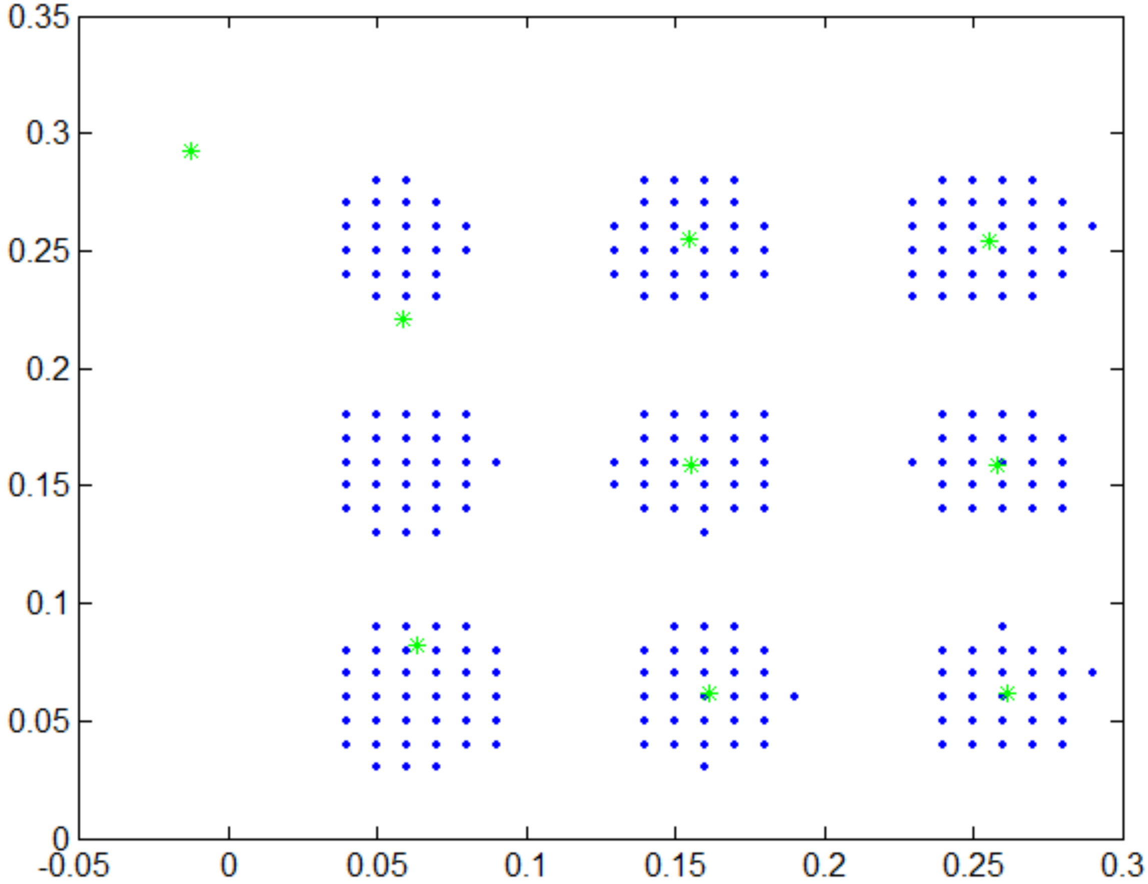
```
function Y=annealed_kmeans(X,K)
[N d]=size(X);
mean_x = mean(X);
Y=randn(K,d)*0.1+ones(K,1)*mean_x;
change=1; v=ceil(rand(N,1)*size(Y,1))';
while change > 0
    D=cross_dis(X,Y);
    v_old = v;
    [dd v]=min(D');
    mean(dd)
    for k=1:K
        ind=find(v == k);
        if length(ind) > 0
        Y(k,:) = mean(X(ind,:));
        end
    end
    change=length(find(v~=v_old));
end



function D=cross_dis(X,Y)
K=size(Y,1);N=size(X,1);
A=sum(X.^2,2)*ones(1,K);
C=ones(N,1)*sum(Y.^2,2)';
B=X*Y';
D=sqrt(A-2*B+C);
```

```
>> Y=annealed_kmeans(X,5)

ans =

    4.6041


ans =

    2.7560


ans =

    1.2952


ans =

    1.1428


Y =

    3.7748   -3.7540
    0.1161    0.1921
    3.8429    4.0855
   -4.0548   -4.1799
   -3.9969    3.5789
```



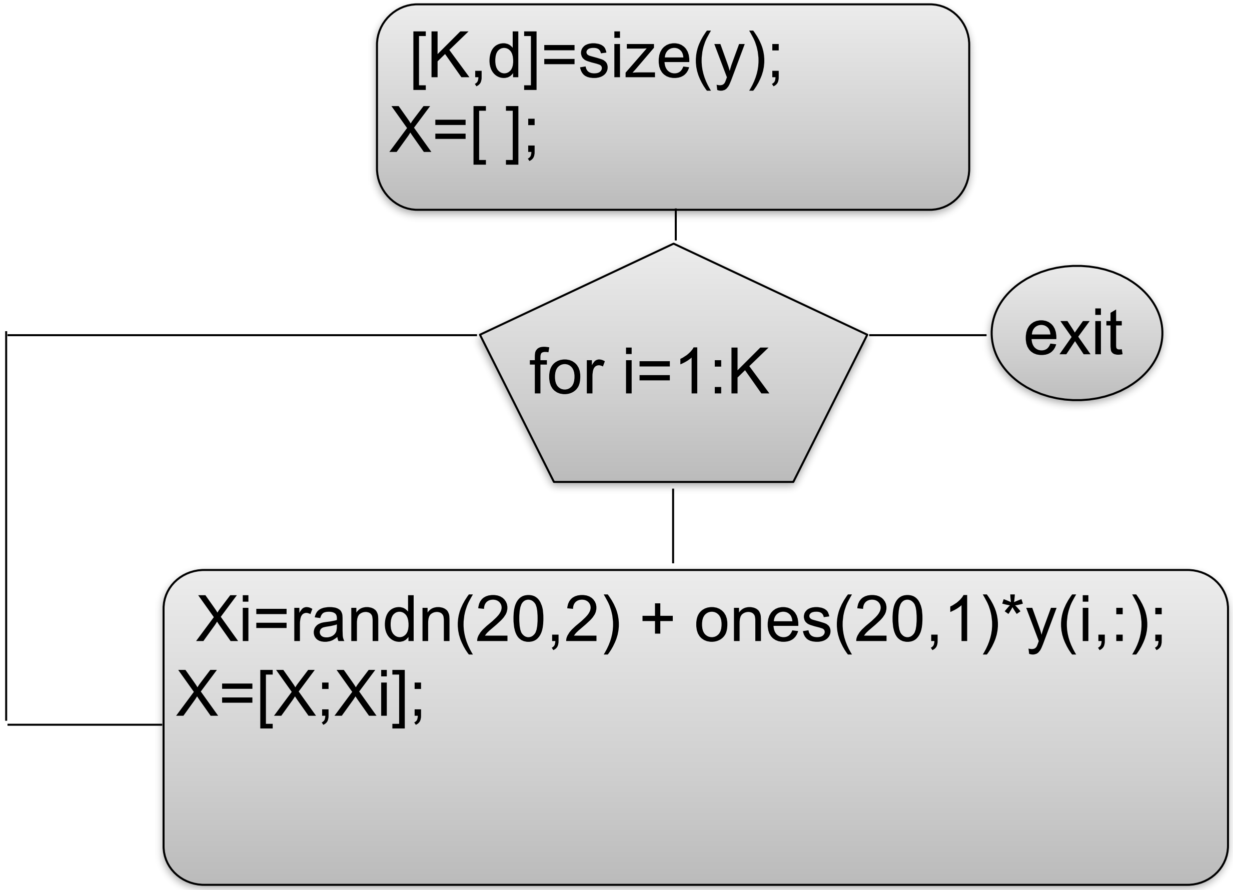
```
>> Y=annealed_kmeans(X,9);

ans =

   0.0593


ans =

   0.0315


ans =

   0.0258


ans =

   0.0212


ans =

   0.0205

>> plot(X(:,1),X(:,2),'.')
>> hold on;plot(Y(:,1),Y(:,2),'g*')
```




0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3

Data generation

A



```
 [K,d]=size(y);
X=[ ];
```

for i=1:K

exit

```
 Xi=randn(20,2) + ones(20,1)*y(i,:);
X=[X;Xi];
```

Initialization

B

```
mean_x = mean(X);
Y=randn(K,d)*0.01+ones(K,1)*mean_x;
```

$$D_{ij} = (\mathbf{x}_i - \mathbf{y}_j)(\mathbf{x}_i^T - \mathbf{y}_j^T)$$

$$= \mathbf{x}_i\mathbf{x}_i^T - 2\mathbf{x}_i\mathbf{y}_j^T + \mathbf{y}_j\mathbf{y}_j^T$$

$$= A_{ij} - 2B_{ij} + C_{ij}$$

C

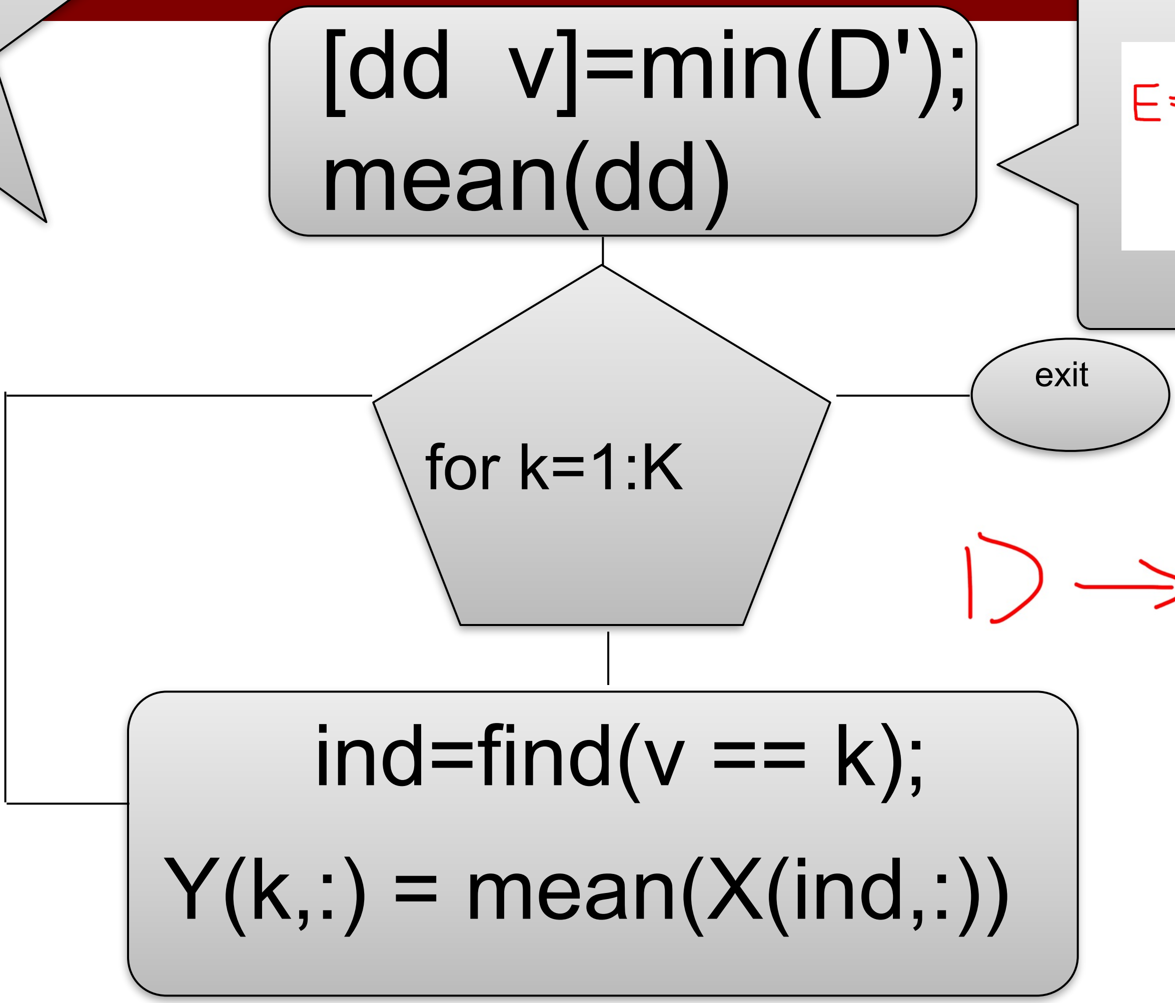
```
K=size(Y,1);N=size(X,1);
A=sum(X.^2,2)*ones(1,K);
C=ones(N,1)*sum(Y.^2,2)';
B=X*Y';
D=sqrt(A-2*B+C);
```

Calculation of Cross distances

- Given N points X: Nx2
- K centers Y: Kx2
- D: NxK
- D(i,j) denotes the distance between X(i,:) and Y(j,:)
- Given X and Y, find D

Exclusive memberships, Criterion, Refining K centers

D



```
[dd  v]=min(D');
mean(dd)
```

for k=1:K

exit

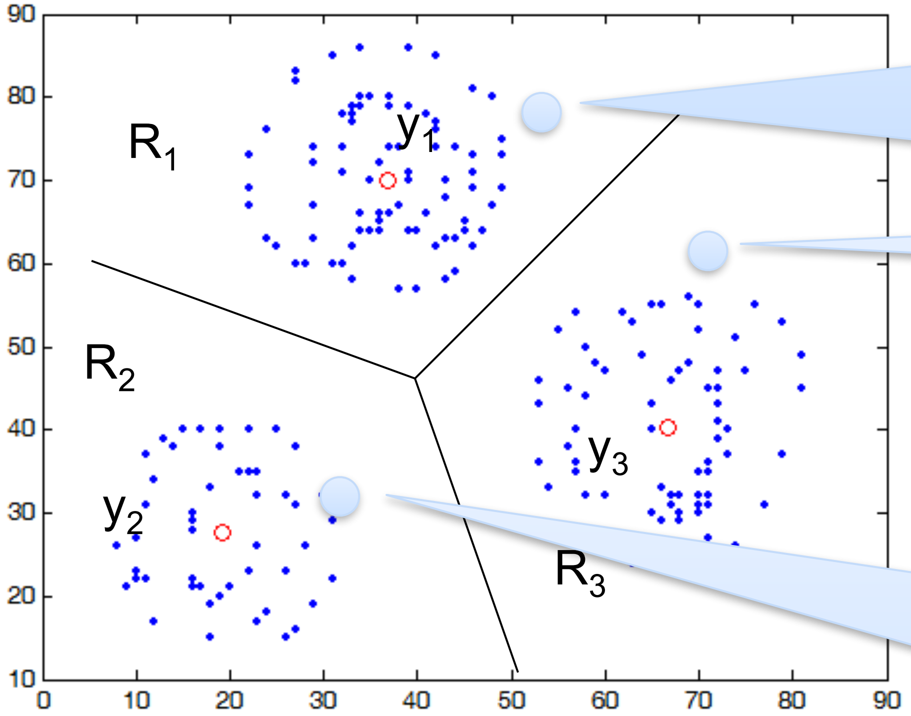
```
ind=find(v == k);
Y(k,:) = mean(X(ind,:))
```

$$E = \frac{1}{N}\sum_i \min_j \|x_i - y_j\|$$

mean(dd)

$$D \longrightarrow \overset{v}{dd} \rightarrow \overset{Y}{E}$$

Exclusive membership q[i]



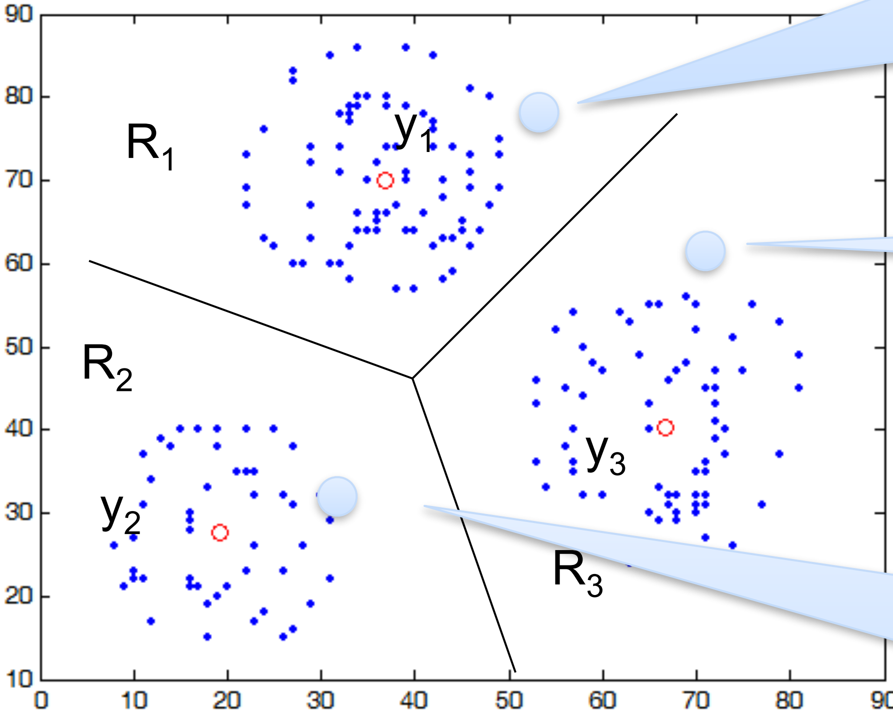
q[i] = (q1[i] q2[i] ... qk[i])
q[i] = (1 0 0)
One and only one active bit in q[i]

q[i] = (0 0 1)

q[i] = (0 1 0)

Overlapping membership q[i]

q[i] = (q1[i] q2[i] ... qk[i])
q[i] = (0.6 0.1 0.3)



q[i] = (0.1 0.1 0.8)

q[i] = (0.15 0.7 0.15)

$q_j[i] \propto \exp(-\beta d_{ij})$

$d_{ij} \equiv D(i,j)$

$\sum_j q_j[i] = 1$

Overlapping memberships

$$g_j[i] \propto \exp(-\beta d_{ij})$$

$$\sum_j g_j[i] = 1$$

$$u_j[i] = \exp(-\beta d_{ij})$$

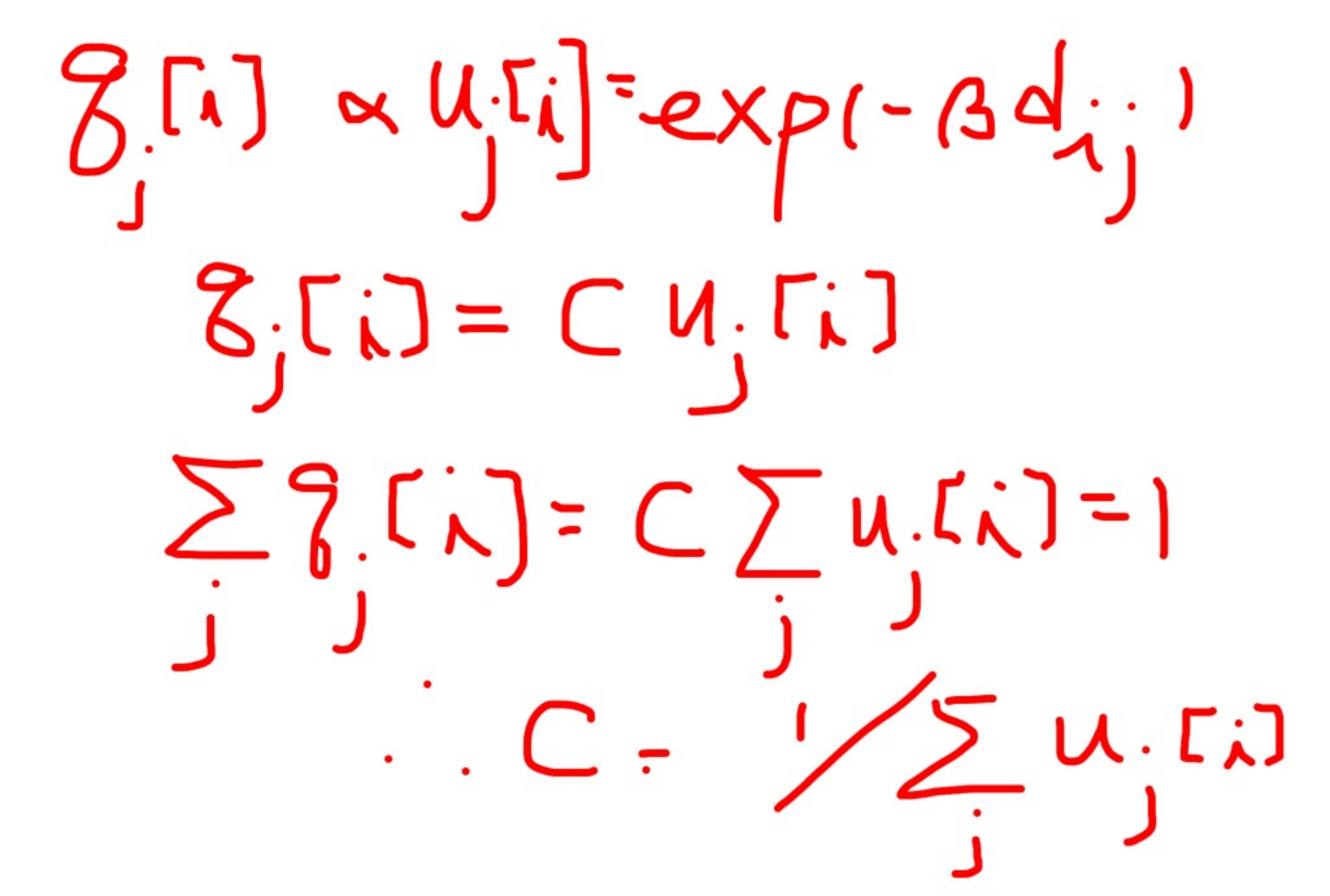
$$g_j[i] = \frac{u_j[i]}{\sum_k u_k[i]}$$

Small B : high degree of overlapping

Large B: exclusive membership

B modulate the overlapping degree inversely

Increase B from small to large values



$$g_j[i] \propto u_j[i] = \exp(-\beta d_{ij})$$

$$g_j[i] = C\, u_j[i]$$

$$\sum_j g_j[i] = C \sum_j u_j[i] = 1$$

$$\therefore C = 1 / \sum_j u_j[i]$$

Overlapping memberships, Criterion,

E

```
U= exp(-B*D);
S=sum(U,2);
Q=U./(S*ones(1,K));


stability=mean(sum(Q.^2,d))
E=mean(sum(Q.*D.^2,2))
```

$$\frac{1}{N}\sum_{i}\sum_{k}\left(g_k[i]\right)^2$$

Small B $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right),\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$ $\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right)$

Large B $(1,0,0),(0,1,0)$ $(0,0,1)$

$$E = \frac{1}{N}\sum_{i}\sum_{k} g_k[i]\,\|x_i - y_k\|^2$$

$$u_j[i] = \exp(-B\, d_{ij})$$

$$g_j[i] = \frac{u_j[i]}{\sum_k u_k[i]}$$

Maximization (minimization)

$$E = \frac{1}{N}\sum_{i}\sum_{k} g_k[i]\,\|x_i - y_k\|^2$$

Minimization of E with respect to all yk
E is in a quadratic form
Setting zero to the derivative of E with respect to yk

Maximization (minimization)

$$\frac{dE}{dy_k} = \frac{1}{N}\sum_i \frac{d}{dy_k} g_k[i]\,\|x_i - y_k\|^2$$

$$= -\frac{2}{N}\sum_i g_k[i]\,(x_i - y_k) = 0$$

$$y_k \sum_i g_k[i] = \sum_i g_k[i]\,x_i$$

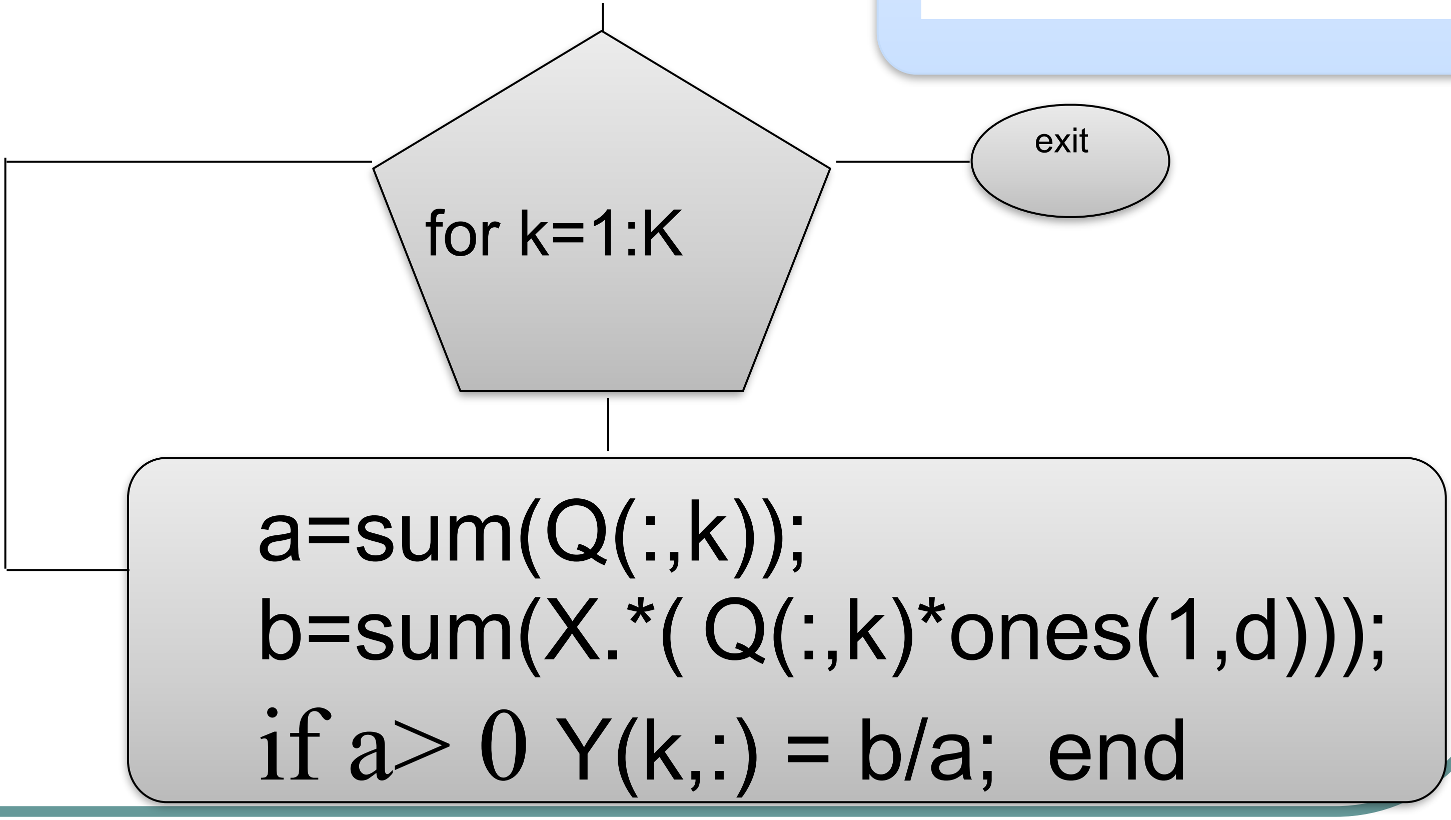
$$y_k \sum_i g_k[i] = \sum_i g_k[i]\, x_i$$

$$y_k = \frac{\sum_i g_k[i]\, x_i}{\sum_i g_k[i]}$$

Refining K centers

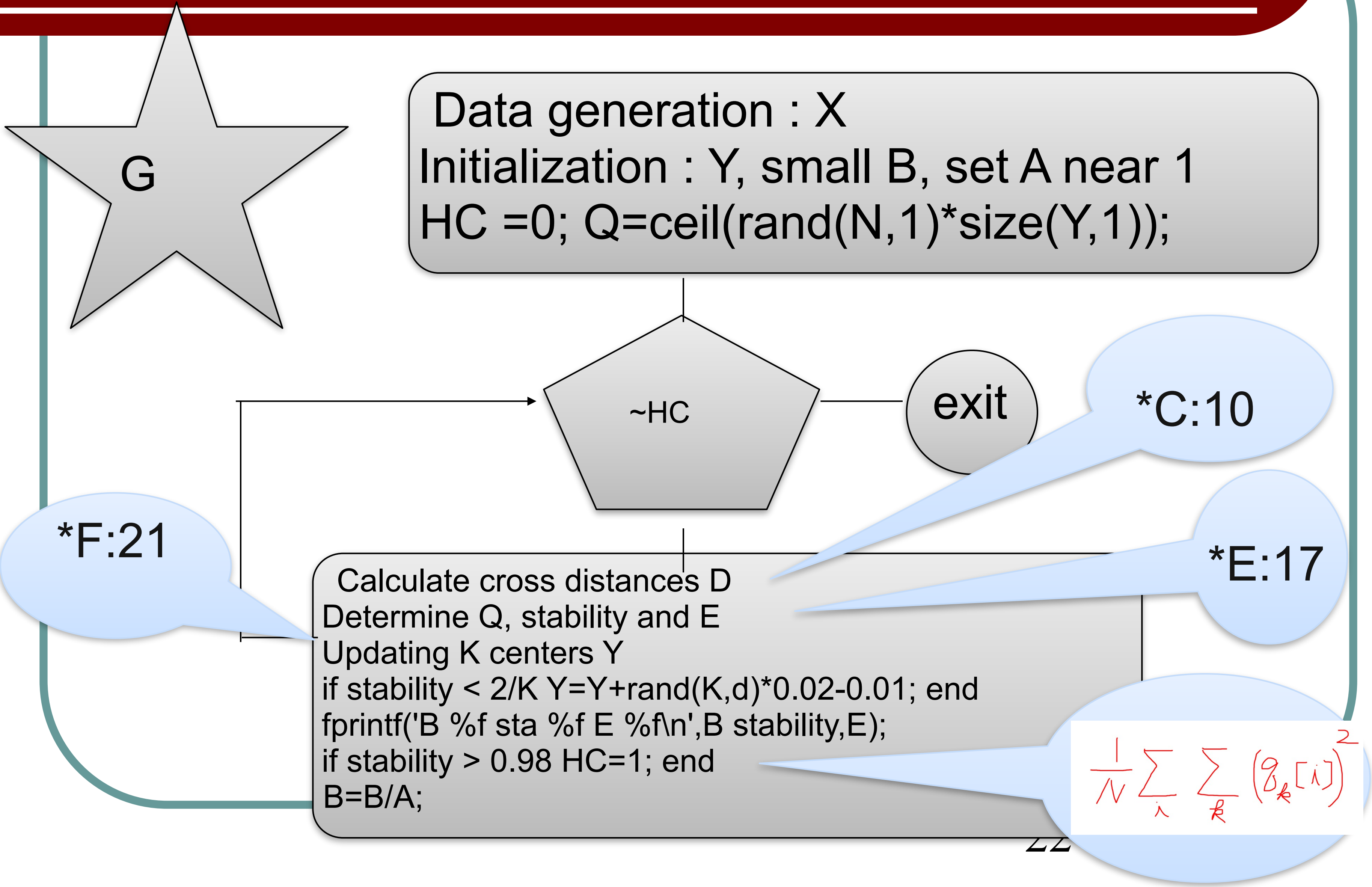
F

$$y_k = \frac{\sum_i q_k[i] X_i}{\sum_i q_k[i]}$$



```
a=sum(Q(:,k));
b=sum(X.*( Q(:,k)*ones(1,d)));
if a> 0 Y(k,:) = b/a;  end
```

Expectation maximization to Quantization maximization



G

Data generation : X
Initialization : Y, small B, set A near 1
HC =0; Q=ceil(rand(N,1)*size(Y,1));

~HC

exit

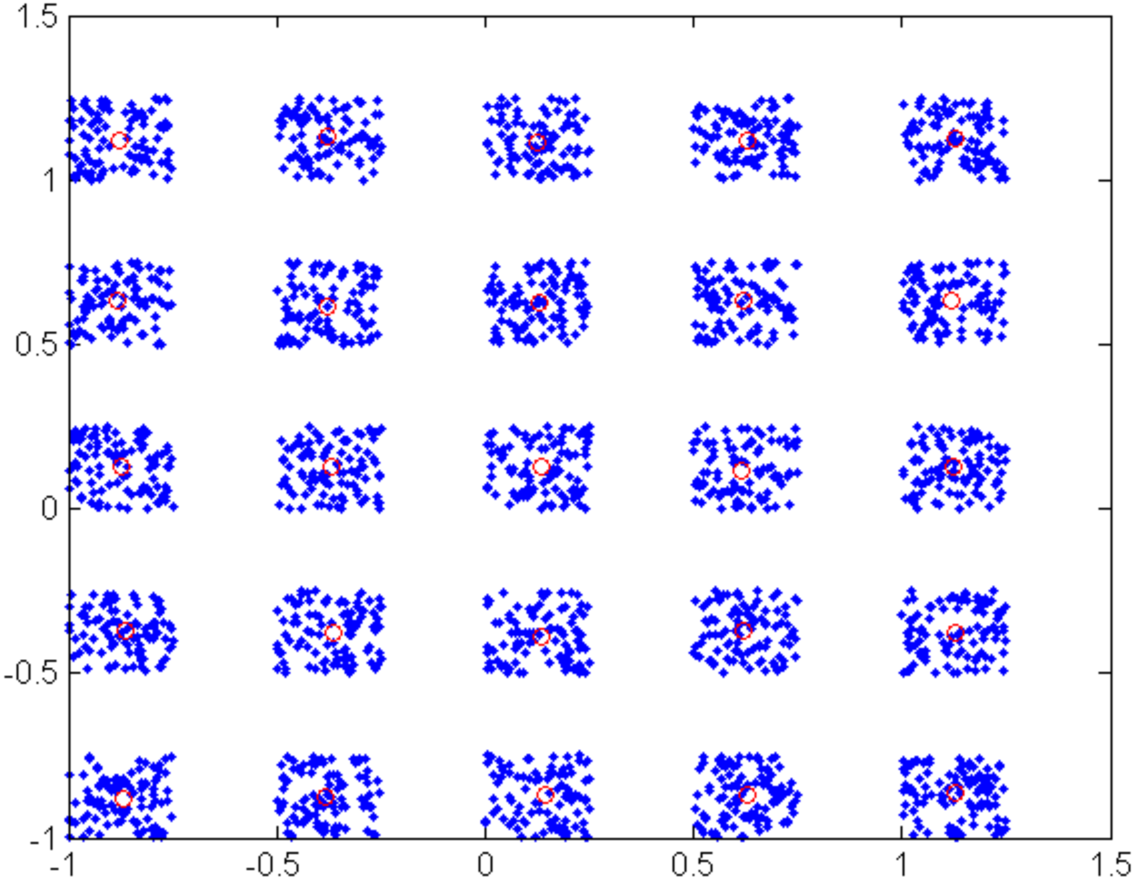
```
Calculate cross distances D
Determine Q, stability and E
Updating K centers Y
if stability < 2/K Y=Y+rand(K,d)*0.02-0.01; end
fprintf('B %f sta %f E %f\n',B stability,E);
if stability > 0.98 HC=1; end
B=B/A;
```

*C:10

*E:17

*F:21

$\frac{1}{N}\sum_{i}\sum_{k}(q_k[i])^2$


1.5
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5
1
1.5

```
x1=linspace(-1,1,5);
x2=linspace(-1,1,5);
X=[];
for i=1:5
    for j=1:5
        X=[X;rand(100,2)*0.25+[ones(100,1)*x1(i) ones(100,1)*x2(j)]];
    end
end

plot(X(:,1),X(:,2),'.');
```