Large Scale data clustering Parallel and distributed codes

2018

J.M. Wu

Large scaled data clustering

- Cross Distance
- Parallel and distributed codes of Cross distances
- Hierarchical clustering models
- Codes : annealed K-Means, Annealed EM
- Numerical simulations

```
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);
```

Case 1

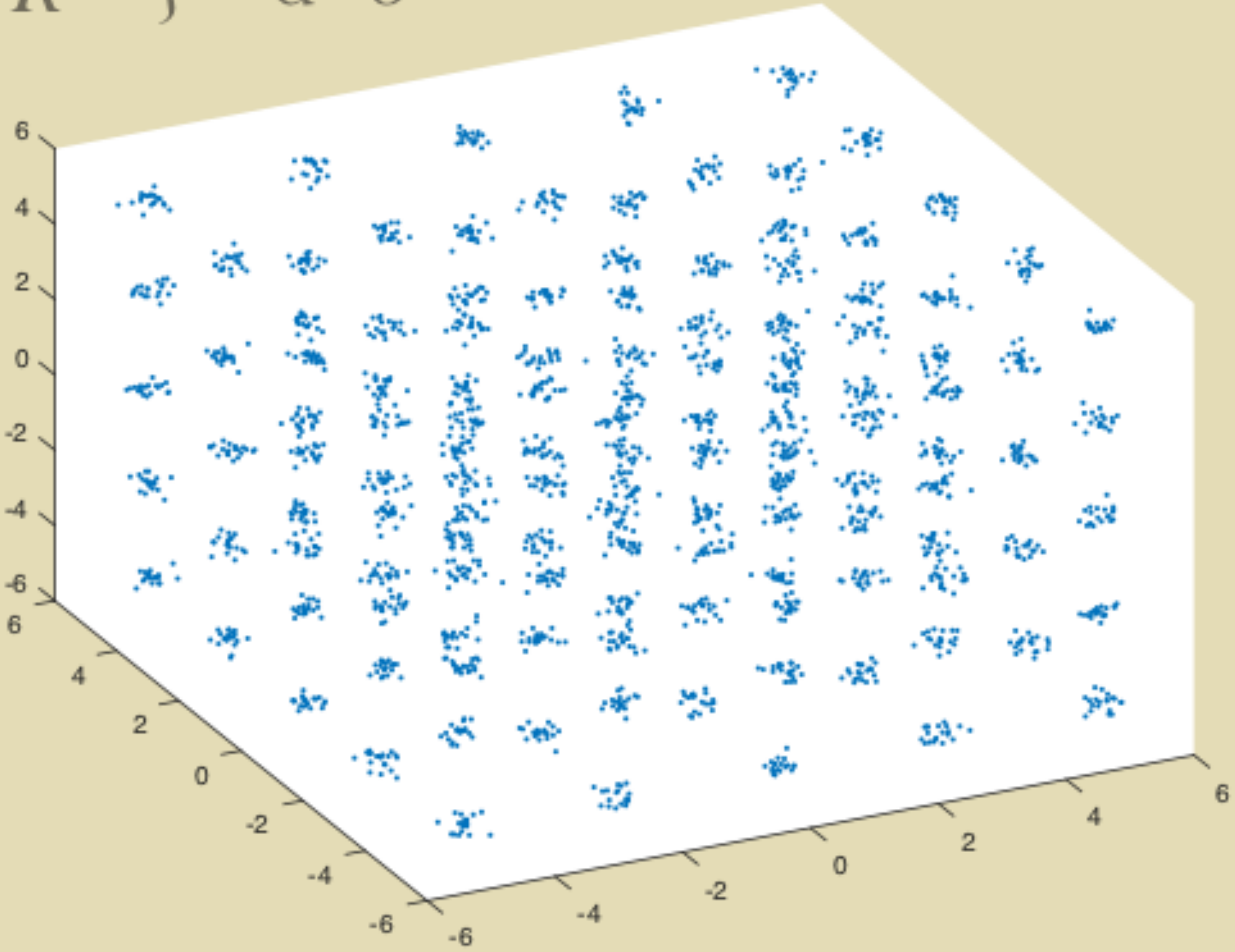
- dimension = 3
- 2500 data points
- 125 centers
- cross distances between data points and centers
- 2500x125

data_gen.m

data_gen.m

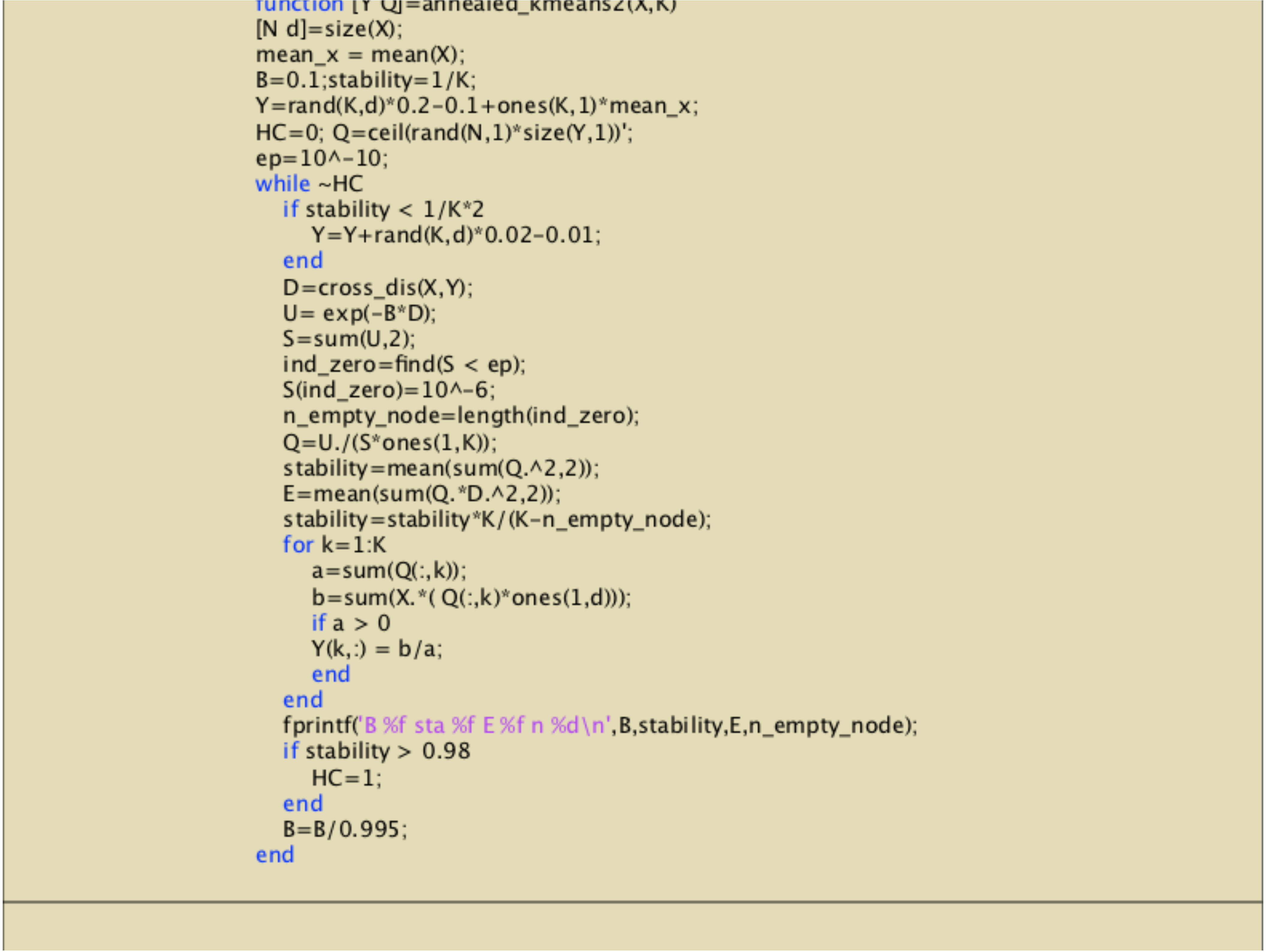
```
clear all
L=5;
a(1,:)=linspace(-5,5,L);
a(2,:)=linspace(-5,5,L);
a(3,:)=linspace(-5,5,L);
X=[]; Y=[];
for i=1:L
    for j=1:L
        for k=1:L
            center=[a(1,i) a(2,j) a(3,k)];
            Xi=randn(20,3)*0.15+ ones(20,1)*center;
            X=[X;Xi];
            Y=[Y;center];

        end
    end
end
plot3(X(:,1),X(:,2),X(:,3),'.');
```

$X=\{x[t] \in R^d \quad \} \quad d=3$



- D=cross_dis(X,Y);
- >> size(D)
- ans =
- 2500 125
- >> tic;D=cross_dis(X,Y);toc
- Elapsed time is 0.046284 seconds.

```
function [Y,Q]=annealed_kmeans2(X,K)
```

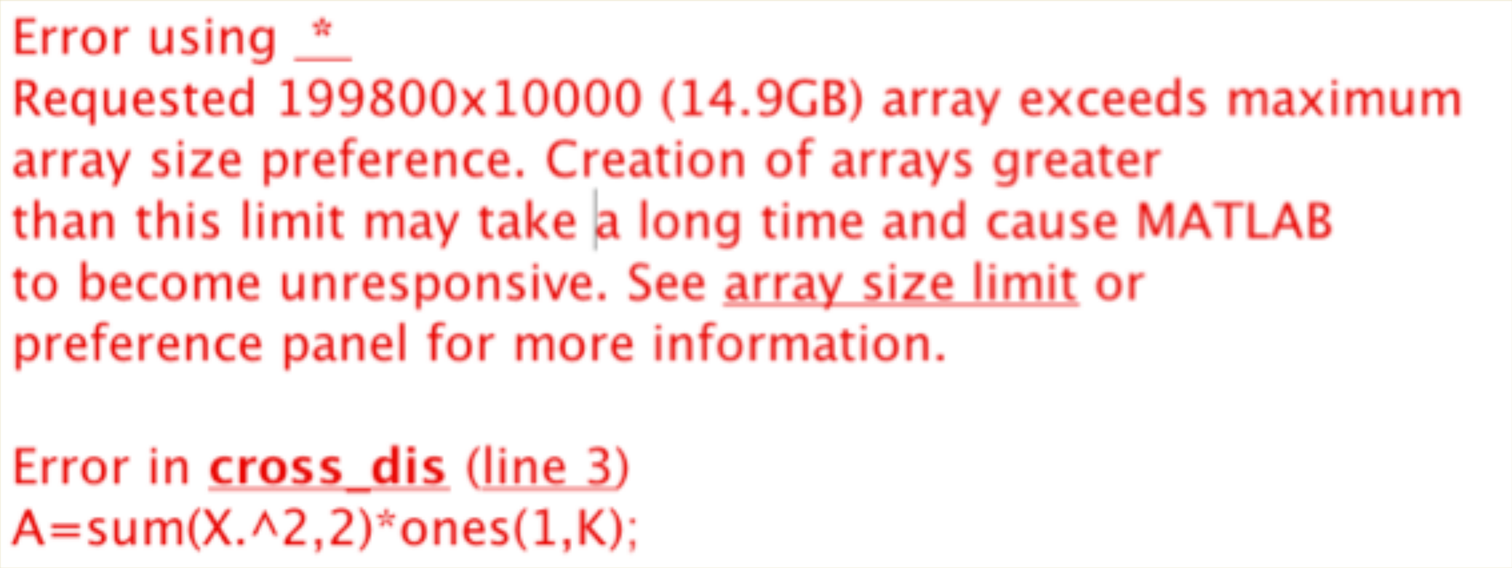


```
function [Y Q]=annealed_kmeans2(X,K)
[N d]=size(X);
mean_x = mean(X);
B=0.1;stability=1/K;
Y=rand(K,d)*0.2-0.1+ones(K,1)*mean_x;
HC=0; Q=ceil(rand(N,1)*size(Y,1))';
ep=10^-10;
while ~HC
    if stability < 1/K*2
        Y=Y+rand(K,d)*0.02-0.01;
    end
    D=cross_dis(X,Y);
    U= exp(-B*D);
    S=sum(U,2);
    ind_zero=find(S < ep);
    S(ind_zero)=10^-6;
    n_empty_node=length(ind_zero);
    Q=U./(S*ones(1,K));
    stability=mean(sum(Q.^2,2));
    E=mean(sum(Q.*D.^2,2));
    stability=stability*K/(K-n_empty_node);
    for k=1:K
        a=sum(Q(:,k));
        b=sum(X.*( Q(:,k)*ones(1,d)));
        if a > 0
        Y(k,:) = b/a;
        end
    end
    fprintf('B %f sta %f E %f n %d\n',B,stability,E,n_empty_node);
    if stability > 0.98
        HC=1;
    end
    B=B/0.995;
end
```

CASE 2

- dimension = 13
- data points 1998000
- centers 12000
- X= rand(1998000,13); Y=rand(12000,13);

- X= rand(199800,13); Y=rand(12000,13);
- tic;D=cross_dis(X,Y);toc



```
Error using _*_
Requested 199800x10000 (14.9GB) array exceeds maximum
array size preference. Creation of arrays greater
than this limit may take a long time and cause MATLAB
to become unresponsive. See array size limit or
preference panel for more information.

Error in cross_dis (line 3)
A=sum(X.^2,2)*ones(1,K);
```

BATCHES

BATCHES

- X= rand(100*4000,13); Y=rand(12000,13);
- x_batch =4000; y_batch=4000;
- x_batch_num = 100*4000/x_batch;
- y_batch_num =12000/y_batch;
- for i=1:x_batch_num
- XX{i}=X(1+(i-1)*x_batch:i*x_batch);
- end
- for j=1:y_batch_num
- YY{j}=Y(1+(j-1)*x_batch:j*x_batch);
- end

```
 for i=1:x_batch_num
 for j=1:y_batch_num
        D{i}{j}=zeros(x_batch,y_batch);
end
end
parfor i=1:x_batch_num
        D{i}{1}=cross_dis(XX{i},YY{1});
end
```

Images and Sounds

- Facial images
  - http://www.face-rec.org/databases/
- Hand-writing character images
- MFCC features of speeches
  - https://sounds.bl.uk/
-