Large Scale data clustering
Models and codes

2018

J.M. Wu

Large scaled data clustering

- Large scaled data
- Parallel and distributed processes
- Expectation Maximization
- K-means
- Hierarchical clustering models
- Codes : annealed K-Means, Annealed EM
- Numerical simulations

Images and Sounds

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

- Natural images
    - https://www.istockphoto.com/
    - http://deeplearning.net/datasets/
- Medical images
- Art Images https://www.pexels.com/search/art/

Gaussian pdf

MEAN

$$P_k(x) = P(x|y_k, A_k)$$

$$= \frac{1}{(2\pi)^{d/2}\sqrt{|A_k^{-1}|}} \exp\left(\frac{(x - y_k)^t A_k (x - y_k)}{2} \right)$$

WHY CLUSTERING

- Generative models Gaussian mixtures

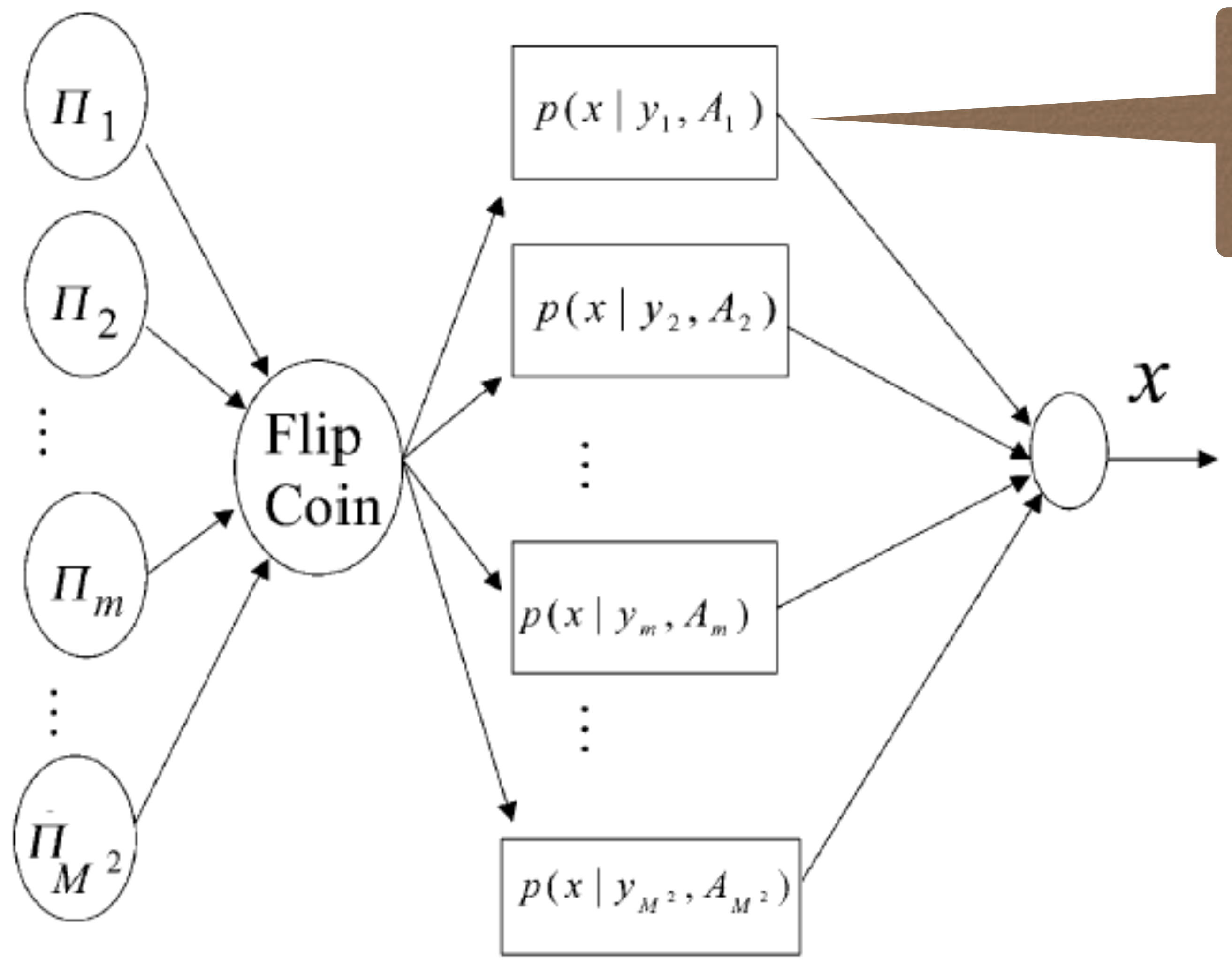
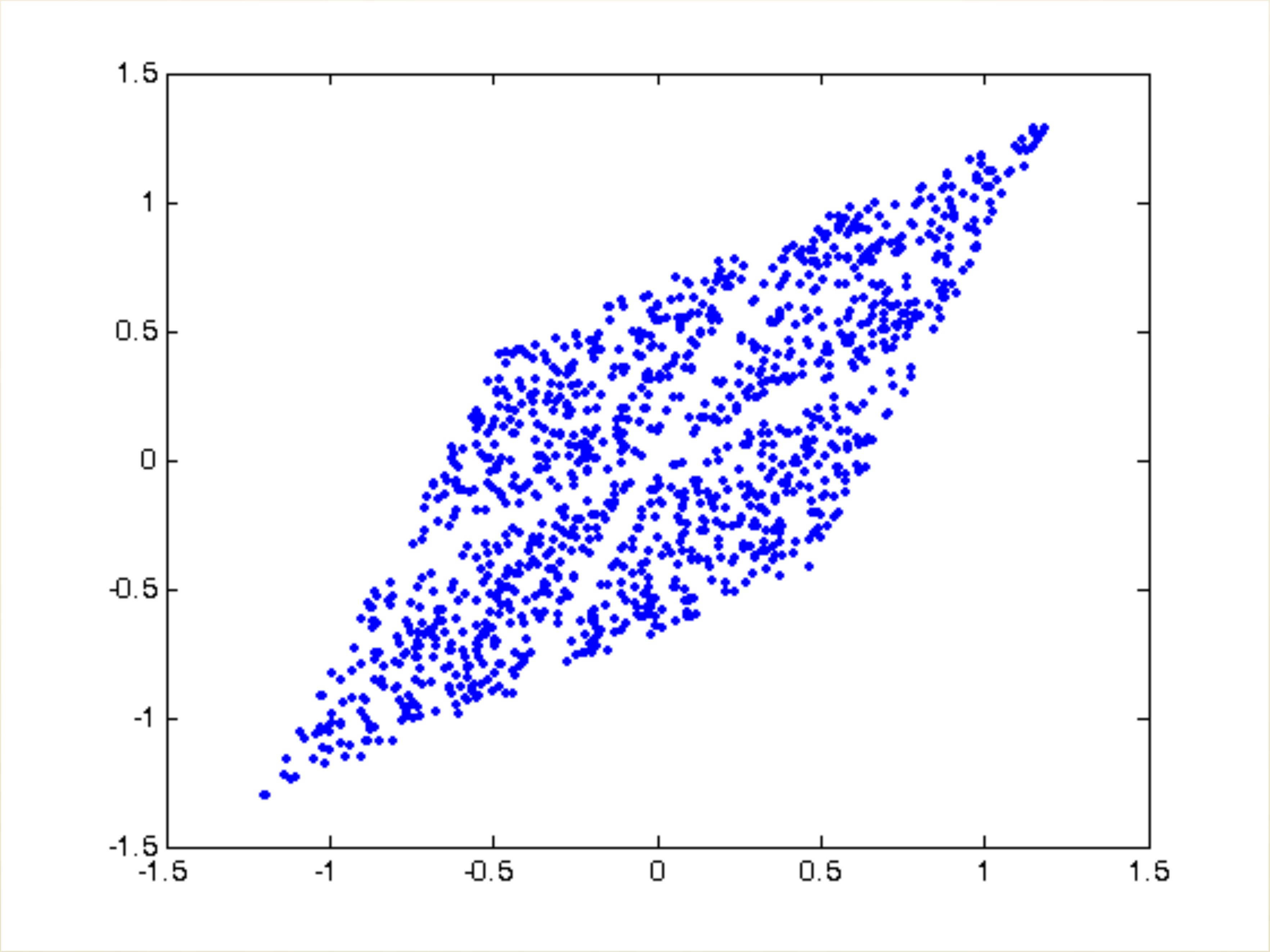
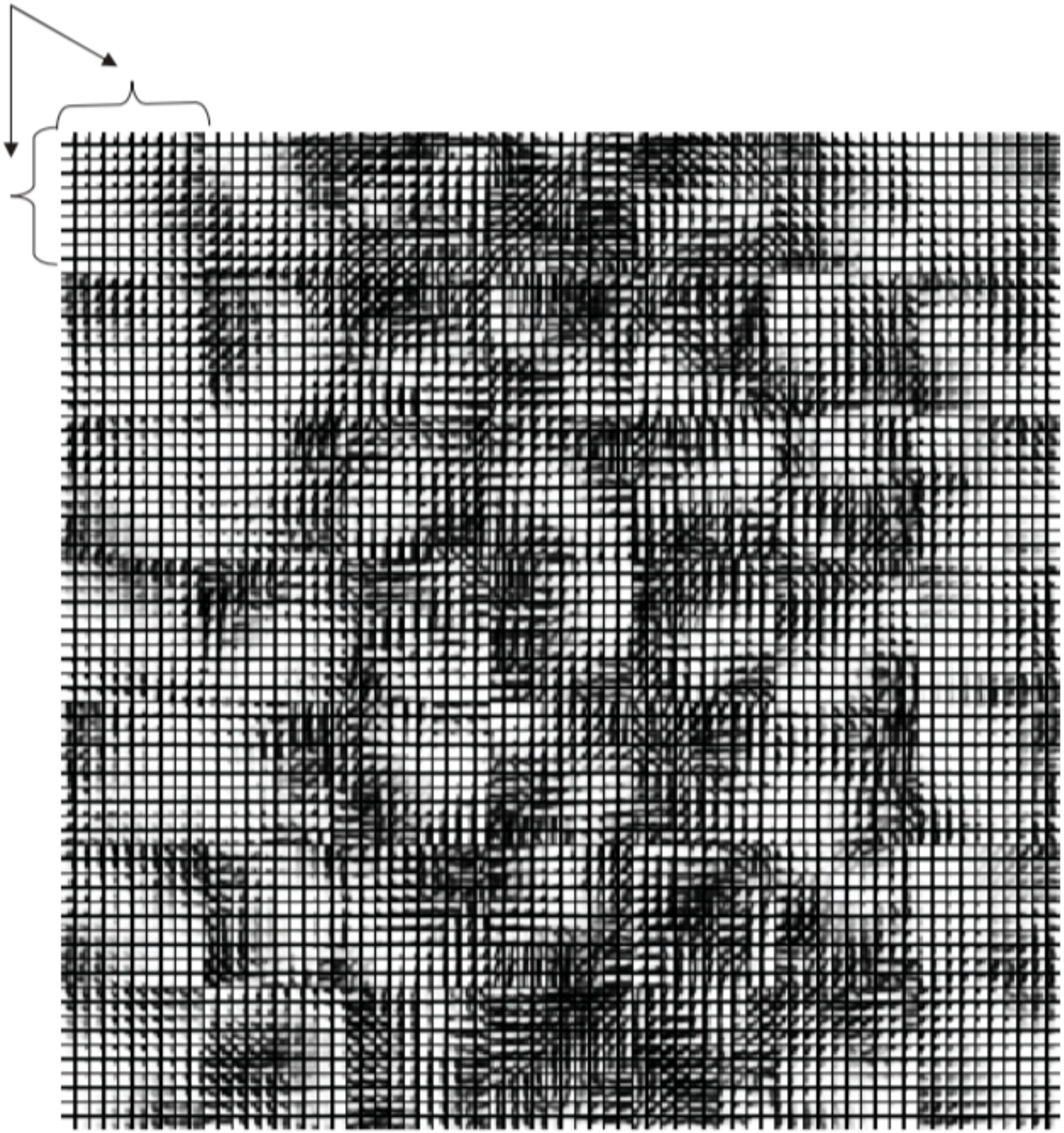


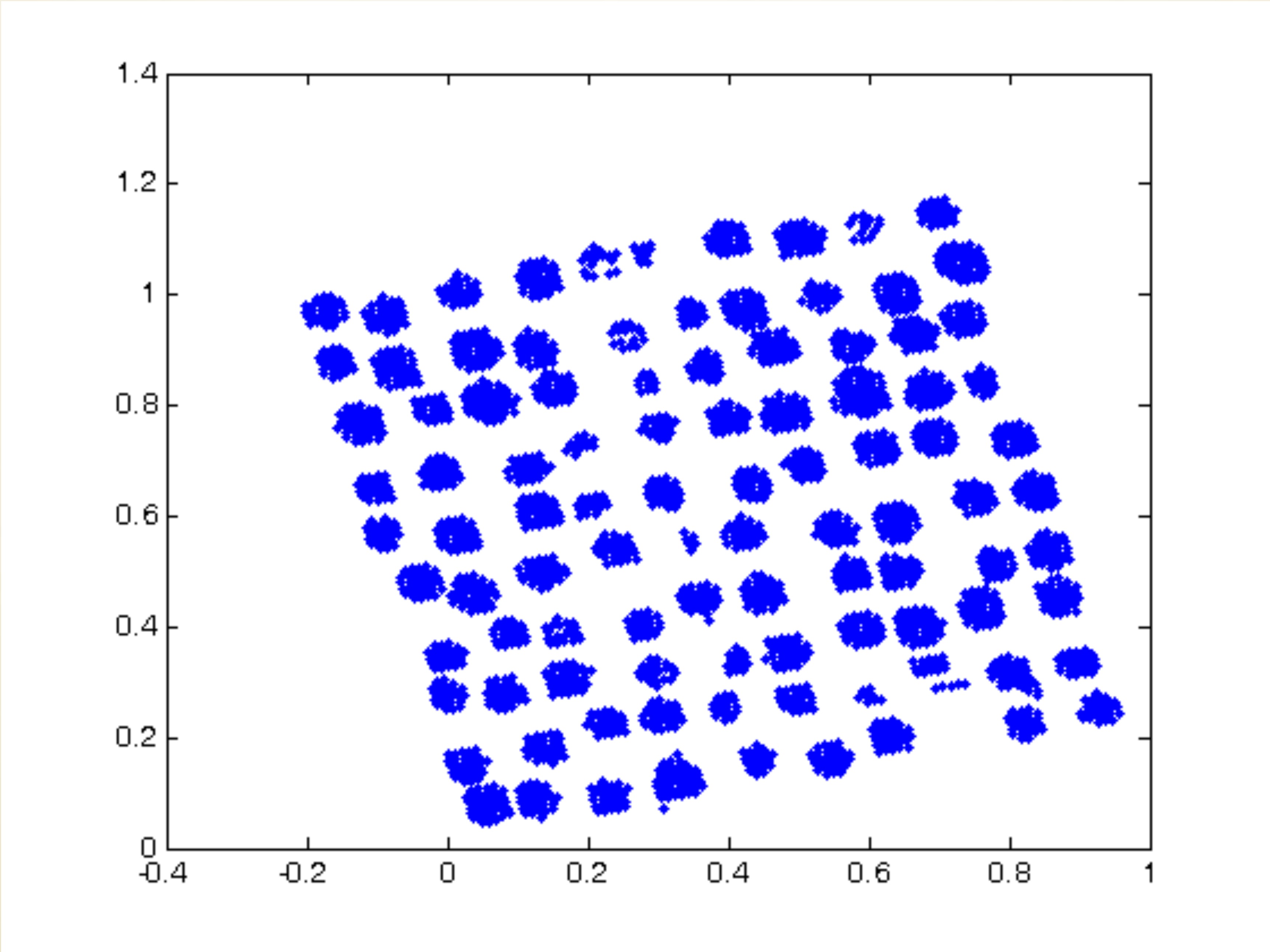
Fig. 1. The generative model.

Many local means

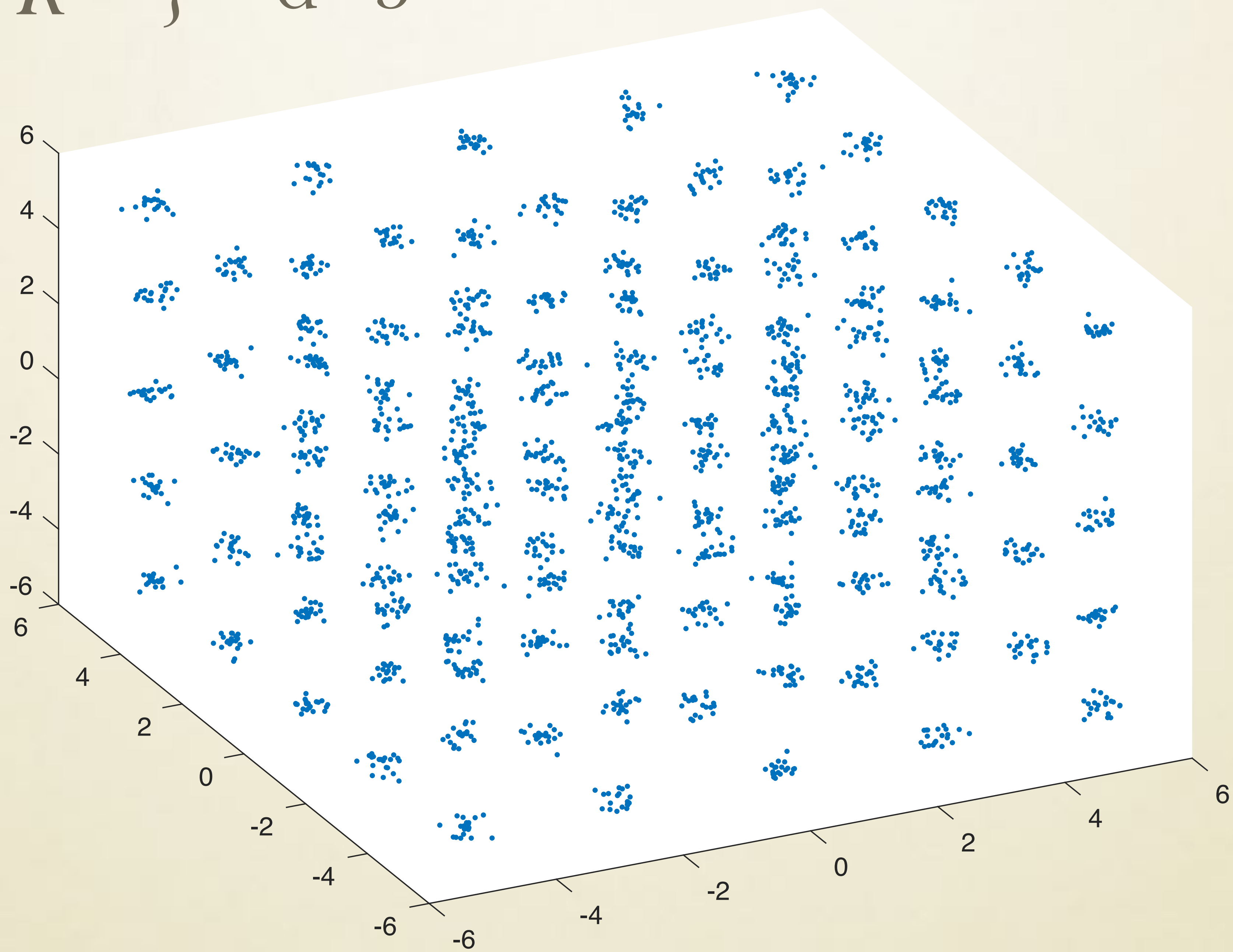

1.5
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1
1.5

A block for 10x10 cortical points of a natural elastic net




1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

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

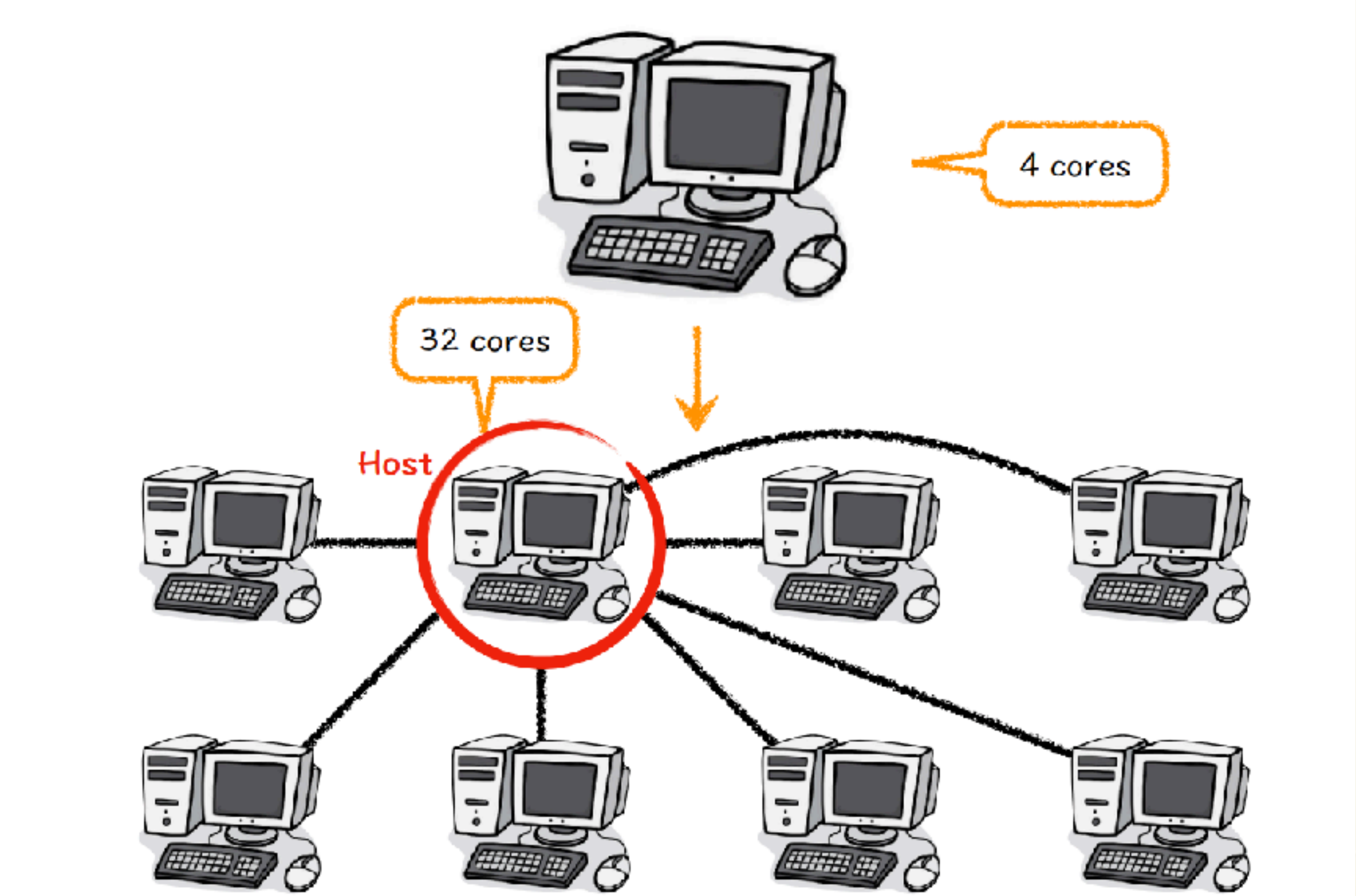


data_gen.m

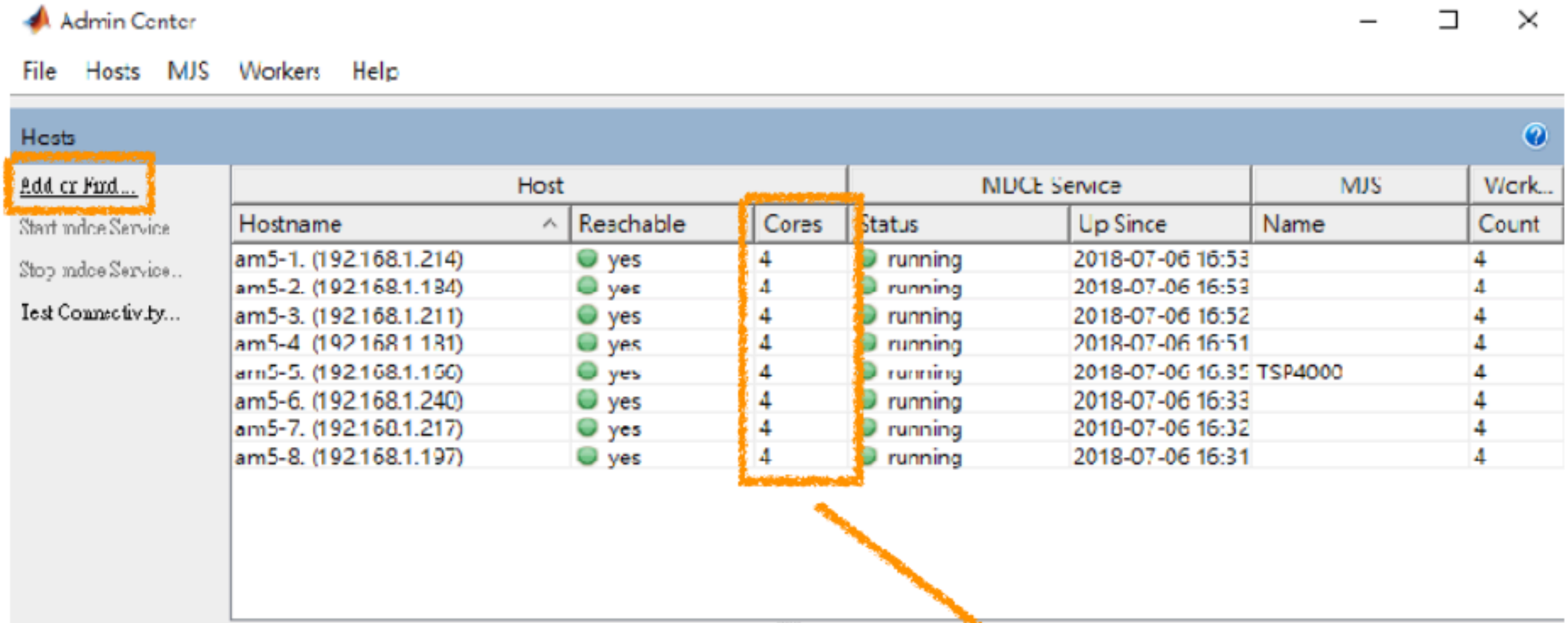
```
clear all
L=5;
a(1,:)=linspace(-5,5,L);
a(2,:)=linspace(-5,5,L);
a(3,:)=linspace(-5,5,L);
X=[];
for i=1:L
    for j=1:L
        for k=1:L
            center=[a(1,i) a(2,j) a(3,k)];
            Xi=randn(20,3)*0.15+ ones(20,1)*center;
            X=[X;Xi];

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

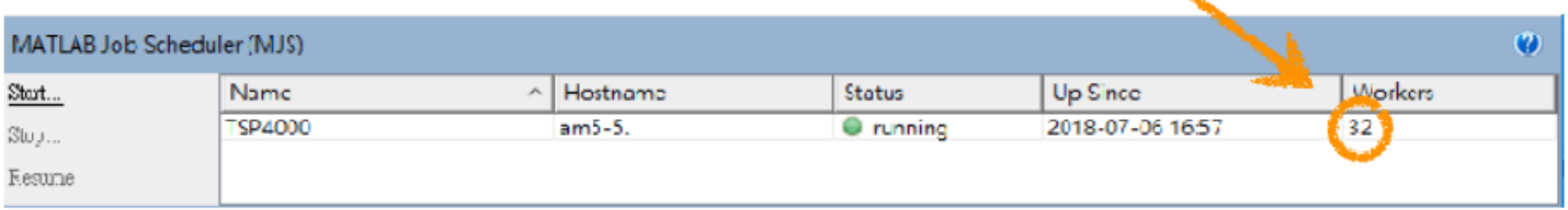
Parallel and Distributed Processes

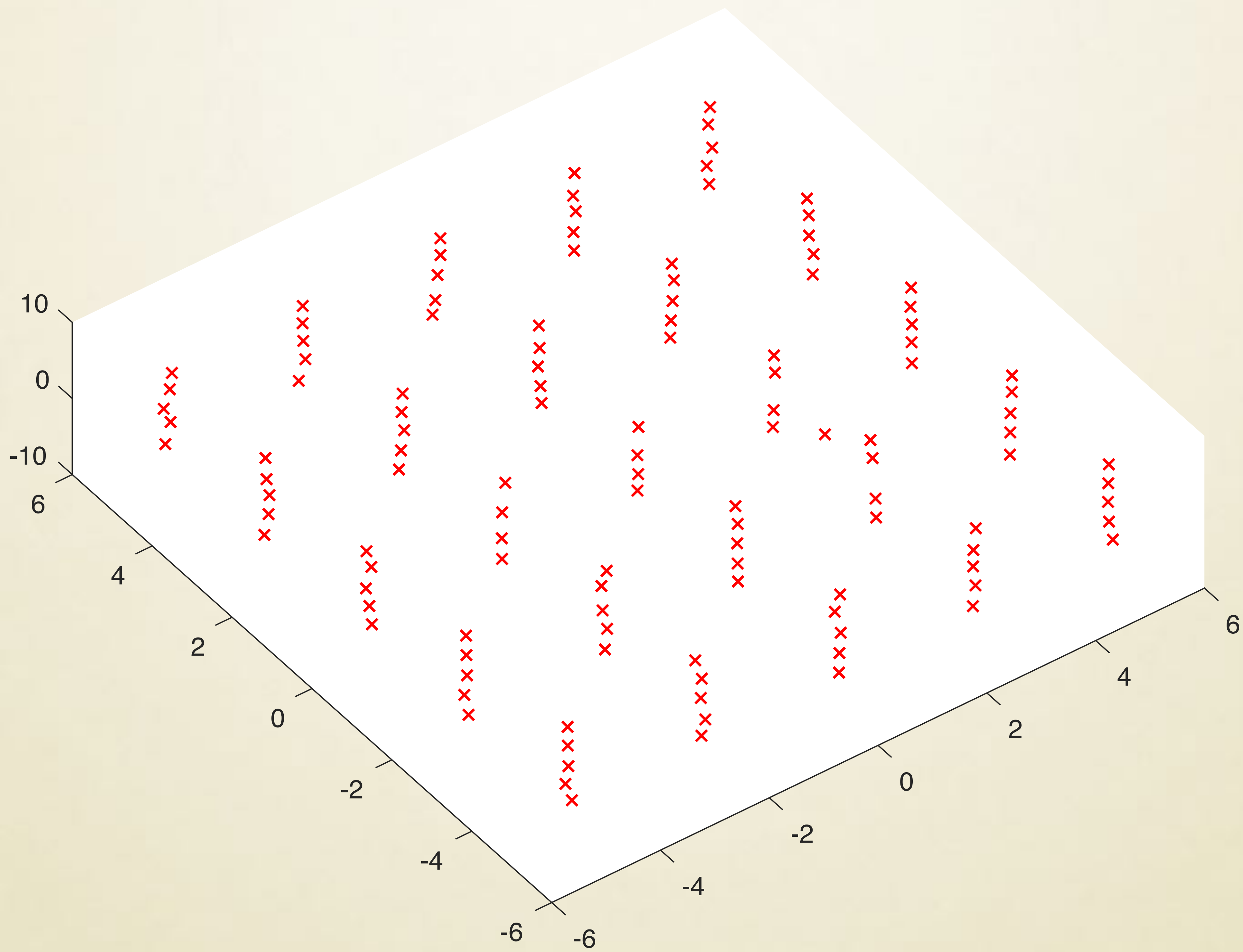


- 在其他台電腦輸入當台的 IP 與 作為 Host 的電腦 IP



- 8 台各有 4 核心的電腦 → 一台擁有 32 核心的電腦





10
0
-10
6
4
2
0
-2
-4
-6
-6
-4
-2
0
2
4
6

Criteria

- High-dimensional data: sub-images
- Large-scaled data: ten millions patterns
- High Speed: parallel and distributed processes
- Accuracy, High Quality

Exclusive Membership

$$e_k^K = [0, 0, \cdots, 0, 1, 0, \cdots, 0, 0]^T$$

$$\text{pos} \quad 1 \; 2 \cdots \text{k} - 1, \text{k}, \text{k} + 1, \cdots, \text{K}$$

- A vector of K binary values
- Only one active bit among K bits
- The kth bit is active and the remaining bits zeroes

Standard Basis

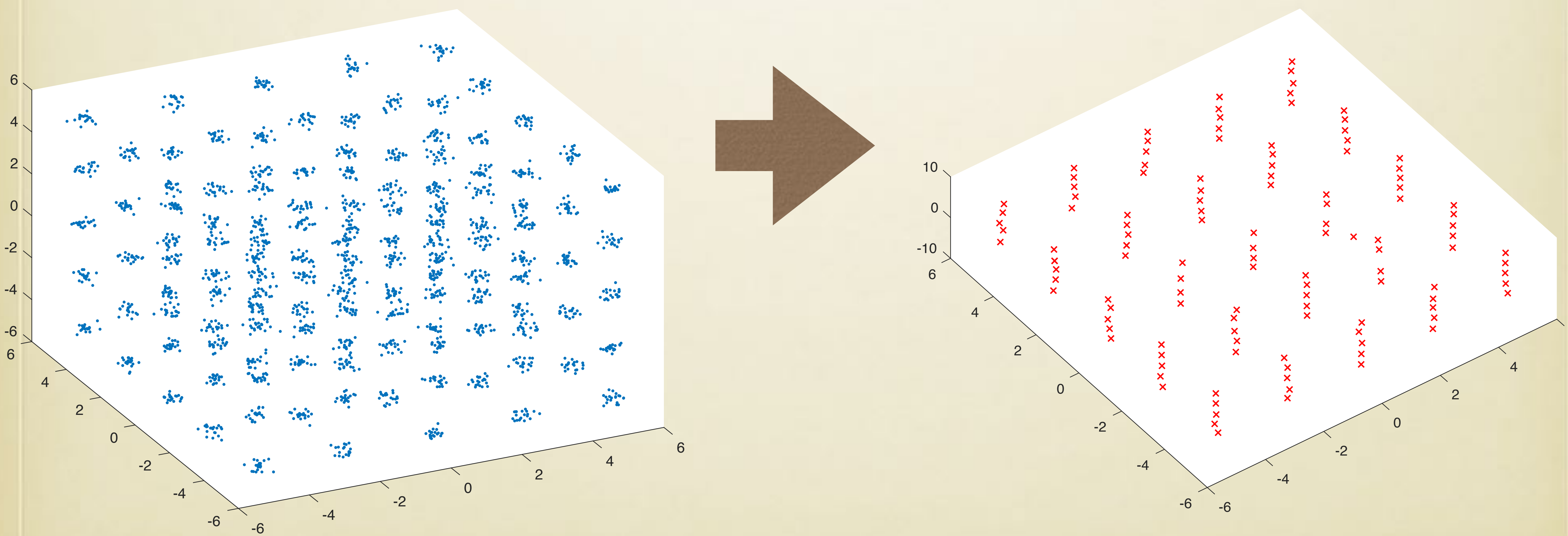
$$\boldsymbol{\Xi} = \{\mathbf{e}_1^K, \cdots, \mathbf{e}_k^K, \cdots, \mathbf{e}_K^K\}$$

$$\boldsymbol{\delta}[t] \in \Xi = \{e_1^K, \cdots, e_k^K, \cdots, e_K^K\}$$

$$\boldsymbol{\delta}[t] = e_k^K \Leftrightarrow \text{x[t] is generated by the kth pdf}$$

Clustering

- input : x[t] for all t
- output : $\delta[t]$ for all t
- Representatives or local means : $\{y[m]\}_{m=1}^{K}$



$$\Pr(\boldsymbol{\xi}_i = e_k^K) \propto \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)$$

$$\sum_{k=1}^{K} \Pr(\boldsymbol{\xi}_i = e_k^K) = 1$$

$$\Pr(\boldsymbol{\xi}_i = e_k^K) = ?$$

Probabilistic Memberships

$$\Pr(\xi_i = e_k^K) \propto \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)$$

$$\sum_{k=1}^{K} \Pr(\xi_i = e_k^K) = 1$$

$$\Pr(\xi_i = e_k^K) = ?$$

$$\Pr(\xi_i = e_k^K) = C \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)$$

$$C \sum_{k=1}^{K} \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2) = 1$$

$$C = \frac{1}{\sum_{k=1}^{K} \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)}$$

$$\Pr(\xi_i = e_k^K) = \frac{\exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)}{\sum_{k=1}^{K} \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)}$$

Expectation

$$\Pr(\boldsymbol{\xi}_i = e_k^K) \propto \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)$$

- Consider an exclusive membership as a random vector
- Assumption of probability
- Expectation $\langle \xi_i \rangle = ?$

Expectation

$$\langle \xi_i \rangle = \sum_{k=1}^{K} \Pr(\xi_i = \mathbf{e}_k^K)\mathbf{e}_k^K = \sum_{k=1}^{K} \frac{\exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_\mathrm{k} \|^2)}{\sum_{h=1}^{\mathrm{K}} \exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_\mathrm{h} \|^2)} \mathbf{e}_k^K$$

$$= \left(\frac{\exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_1 \|^2)}{\sum_{\mathrm{k}=1}^{\mathrm{K}} \exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_\mathrm{k} \|^2)}, \frac{\exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_2 \|^2)}{\sum_{\mathrm{k}=1}^{\mathrm{K}} \exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_\mathrm{k} \|^2)}, \cdots, \frac{\exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_\mathrm{K} \|^2)}{\sum_{\mathrm{k}=1}^{\mathrm{K}} \exp(-\beta \| \mathbf{x}_\mathrm{i} - \mathbf{y}_\mathrm{k} \|^2)} \right)$$

Expectation Equation

$$v_{ik} \equiv \langle \xi_{\mathrm{ik}} \rangle = \Pr(\boldsymbol{\xi}_{\mathrm{i}} = \mathrm{e}_{\mathrm{k}}^{\mathrm{K}}) = \frac{\exp(-\beta \| \mathbf{x}_{\mathrm{i}} - \mathbf{y}_{\mathrm{k}} \|^2)}{\sum_{\mathrm{k=1}}^{\mathrm{K}} \exp(-\beta \| \mathbf{x}_{\mathrm{i}} - \mathbf{y}_{\mathrm{k}} \|^2)} \quad \text{(E1)}$$

Expectation Maximization

$$\mathrm{E}(\xi, Y) = \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{ik} (\mathbf{x}_i - \mathbf{y}_k)^T (\mathbf{x}_i - \mathbf{y}_k)$$

- Mathematical modeling
- The distance between x[t] and its representative is minimized.

One and only one active bit in $[\xi_{i1}, \ldots, \xi_{ik}, \ldots, \xi_{iK}]$
inner summation contains one non-zero term at most

$$\mathrm{E}(\boldsymbol{\xi}, Y) = \sum_{i=1}^{N} \sum_{k=1}^{K} \xi_{ik} (\mathbf{x}_i - \mathbf{y}_k)^T (\mathbf{x}_i - \mathbf{y}_k)$$

- EM minimizes $E(\langle \xi \rangle, Y)$ directly with respect to all y_k

$$\boxed{\mathrm{E}(\langle \boldsymbol{\xi} \rangle, Y) = \sum_{i=1}^{N} \sum_{k=1}^{K} \langle \xi_{ik} \rangle (\mathbf{x}_i - \mathbf{y}_k)^T (\mathbf{x}_i - \mathbf{y}_k)}$$

Maximization (minimization)

$$\mathrm{E}(\langle\boldsymbol{\xi}\rangle, Y) = \sum_{i=1}^{N}\sum_{k=1}^{K}\langle\xi_{ik}\rangle(\mathbf{x}_i - \mathbf{y}_k)^T(\mathbf{x}_i - \mathbf{y}_k)$$

$$\frac{\partial \mathrm{E}(\langle\boldsymbol{\xi}\rangle, Y)}{\partial \mathbf{y}_k} = -2\sum_{i=1}^{N}\langle\xi_{ik}\rangle(\mathbf{x}_i - \mathbf{y}_k) = 0$$

$$\sum_{i=1}^{N}\langle\xi_{ik}\rangle\mathbf{y}_k = \sum_{i=1}^{N}\langle\xi_{ik}\rangle\mathbf{x}_i \Rightarrow \mathbf{y}_k = \frac{\sum_{i=1}^{N}\langle\xi_{ik}\rangle\mathbf{x}_i}{\sum_{i=1}^{N}\langle\xi_{ik}\rangle}$$

•

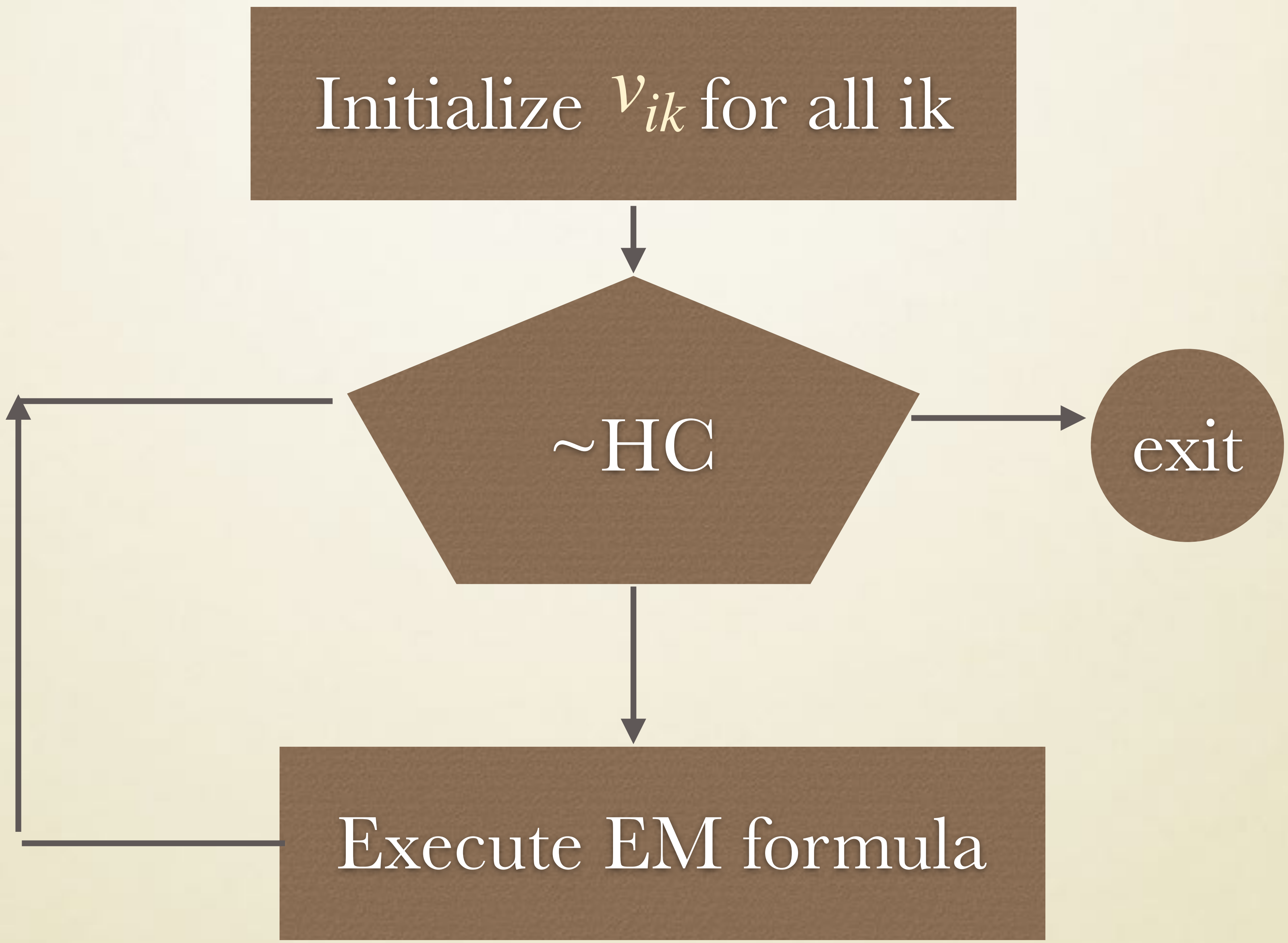
EM Formula

$$v_{ik} \equiv \langle \xi_{ik} \rangle = \Pr(\xi_i = e_k^K) = \frac{\exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)}{\sum_{k=1}^{K} \exp(-\beta \| \mathbf{x}_i - \mathbf{y}_k \|^2)} \quad \text{(E1)}$$

$\beta = 1$

$$\mathbf{y}_k = \frac{\sum_{i=1}^{N} \langle \xi_{ik} \rangle \mathbf{x}_i}{\sum_{i=1}^{N} \langle \xi_{ik} \rangle}$$

While-looping



K-means

- Consider fixed local means
- Determine exclusive memberships for each data
- Minimize

$$E_i = \sum_k \xi_{ik} \parallel x_i - y_k \parallel^2$$

ASSIGNMENT

$$E_i = \sum_k \xi_{ik} \parallel x_i - y_k \parallel^2$$

- is minimized by simply assigning one to ξ_{ik*}

$$x_i \quad is \quad closest \quad to \quad y_{k*}$$

Assignment

$$x_i \quad is \quad closest \quad to \quad y_{k*}$$

- is equivalent to

$$\| x_i - y_{k*} \| = min_k \| x_i - y_k \|$$

ASSIGNMENT

- x_i is simply assigned to a cluster whose representative is closest to x_i

$$\| x_i - y_{k*} \| = min_k \| x_i - y_k \|$$

- is equivalent to

$$k* = argmin_k \| x_i - y_k \|$$

Partition and Updating K-means

$$S_k = \{x_i \,|\, \xi_i = e_k \,|\, \xi_{ik} = 1\}$$

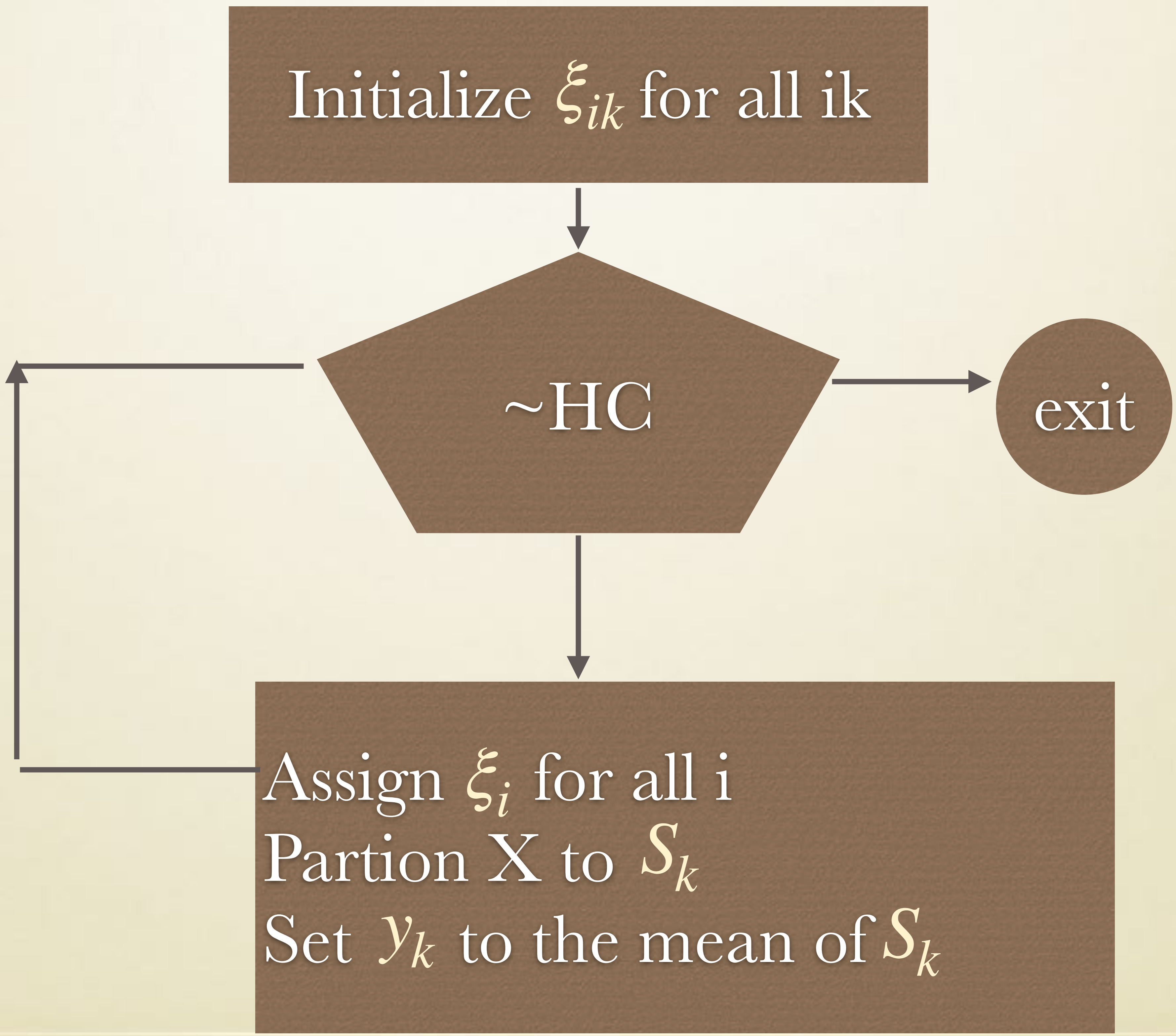
- Partition the whole data set into K non-overlapping subsets
- x_i is partitioned to S_k if $\xi_i = e_k$

Updating K-means

$$y_k = \frac{1}{|S_k|} \sum_{i \in S_k} x_i$$

- Recalculate the mean of elements in S_k

While-looping



Annealed Expectation Maximization

1. Set β to a sufficiently low value

$$\mathrm{A} = 0.01 \times \mathrm{I}$$

$$\mathbf{y}_k \approx \frac{1}{N}\sum_t \mathbf{x}[t],\ v_k[t] \approx \frac{1}{K}$$

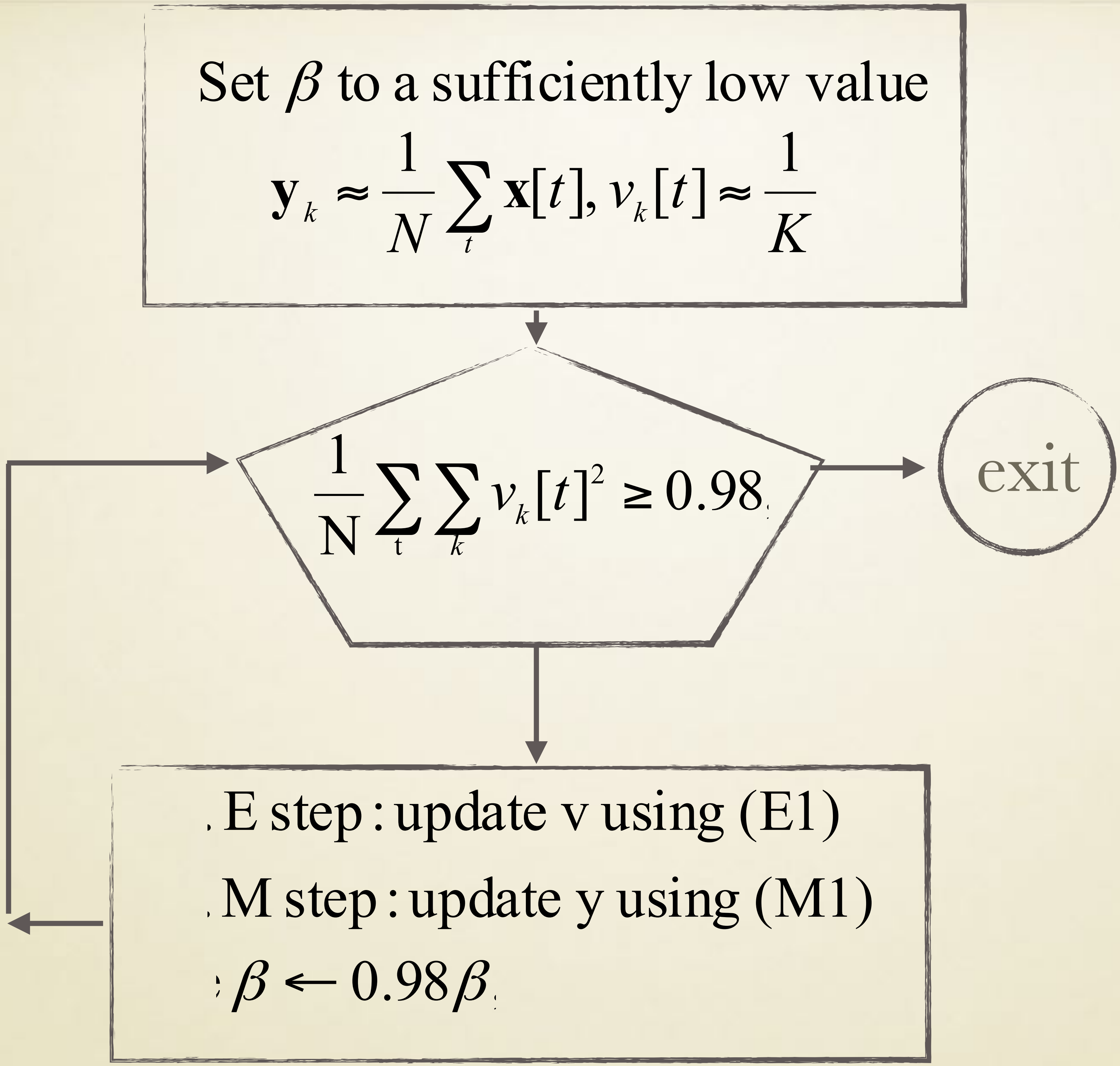
2. E step: update v using (E1)

3. M step: update y using (M1)

   update A using (M2)

4. If $\frac{1}{\mathrm{N}}\sum_{\mathrm{t}}\sum_k v_k[t]^2 \geq 0.98$, halt

   else $\beta \leftarrow 0.98\beta$, goto step 2



Set β to a sufficiently low value
$\mathbf{y}_k \approx \frac{1}{N}\sum_t \mathbf{x}[t], v_k[t] \approx \frac{1}{K}$
$\frac{1}{\mathrm{N}}\sum_t \sum_k v_k[t]^2 \geq 0.98$
exit
E step : update v using (E1)
M step : update y using (M1)
$\beta \leftarrow 0.98\beta$

EM Formula

$$v_{ik} \equiv \langle \xi_{\mathrm{ik}} \rangle = \Pr(\boldsymbol{\xi}_{\mathrm{i}} = \mathrm{e}_{\mathrm{k}}^{\mathrm{K}}) = \frac{\exp(-\beta \| \mathbf{x}_{\mathrm{i}} - \mathbf{y}_{\mathrm{k}} \|^2)}{\sum_{\mathrm{k}=1}^{\mathrm{K}} \exp(-\beta \| \mathbf{x}_{\mathrm{i}} - \mathbf{y}_{\mathrm{k}} \|^2)} \quad \text{(E1)}$$

$$\mathbf{y}_k = \frac{\sum_{i=1}^{N} \langle \xi_{ik} \rangle \mathbf{x}_i}{\sum_{i=1}^{N} \langle \xi_{ik} \rangle}$$

Derivation from Free Energy

$$\Pr(\boldsymbol{\delta}[t] = \mathbf{e}_k^K) = \frac{\exp(\beta u_k[t])}{\sum_{j=1}^{K} \exp(\beta u_j[t])}$$

$$\text{Expectation of } \boldsymbol{\delta}[t] = \sum_{k=1}^{K} \mathbf{e}_k^K \Pr(\boldsymbol{\delta}[t] = \mathbf{e}_k^K)$$

$$\text{Entropy of } \boldsymbol{\delta}[t] = -\sum_{k=1}^{K} \Pr(\boldsymbol{\delta}[t] = \mathbf{e}_k^K) \ln \Pr(\boldsymbol{\delta}[t] = \mathbf{e}_k^K)$$

$$\Pr(\boldsymbol{\delta}[t] = \mathbf{e}_k^K) = \frac{\exp(\beta u_k[t])}{\sum_{j=1}^{K} \exp(\beta u_j[t])} \equiv v_k[t] = \left\langle \delta_k[t] \right\rangle$$

$$\mathrm{H_t} \equiv \text{Entropy of } \boldsymbol{\delta}[t] = -\sum_{k=1}^{K} \Pr(\boldsymbol{\delta}[t] = \mathbf{e}_k^K) \ln \Pr(\boldsymbol{\delta}[t] = \mathbf{e}_k^K)$$

$$= -\sum_{k=1}^{K} v_k[t](\beta u_k[t] - \ln \sum_{j=1}^{K} \exp(\beta u_j[t]))$$

$$= -\beta \sum_{k=1}^{K} v_k[t] u_k[t] + \sum_{k=1}^{K} v_k[t] \ln \sum_{j=1}^{K} \exp(\beta u_j[t])$$

Free Energy

- A combination of Mean Energy and Negative Entropy

$$\mathrm{F} = \langle \mathrm{E}(\boldsymbol{\delta}) \rangle - \frac{1}{\beta} H(\boldsymbol{\delta})$$

$$\approx \mathrm{E}(\{\langle \boldsymbol{\delta}[t] \rangle\}) - \frac{1}{\beta} \sum_t H(\boldsymbol{\delta}[t])$$

Derived based on

Kullback - Leiberg(KL) divergence

Mean Field Equations

$$\frac{\partial \mathrm{F}}{\partial v_k[t]} = 0, \frac{\partial \mathrm{F}}{\partial u_k[t]} = 0, \forall k, t$$

$$u_k[t] = -\frac{\partial \mathrm{E}(\mathbf{v})}{\partial v_k[t]},$$

$$v_k[t] = \frac{\exp(\beta u_k[t])}{\sum_j \exp(\beta u_j[t])}$$

Free Energy

$$L(\boldsymbol{\delta}, \mathbf{y}, \mathbf{A}) = \sum_k L_k$$

$$= \frac{1}{2} \sum_t \sum_k \delta_k[t](\mathbf{x}[t] - \mathbf{y}_k)^T \mathbf{A}(\mathbf{x}[t] - \mathbf{y}_k) - \frac{N}{2} \log|\mathbf{A}|$$

$$\mathrm{F}(\mathbf{v}, \mathbf{u}, \mathbf{y}, \mathbf{A})$$

$$= \mathrm{E}(\mathbf{v}, \mathbf{y}, \mathbf{A}) + \sum_t \sum_k v_k[t] u_k[t]$$

$$- \frac{1}{\beta} \sum_t \ln \sum_j \exp(\beta u_j[t])$$

$$\frac{\partial \mathrm{F}}{\partial v_k[t]} = 0, \frac{\partial \mathrm{F}}{\partial u_k[t]} = 0, \forall k, t$$

$$\frac{\partial \mathrm{F}}{\partial \mathbf{y}_k} = \frac{dL(\mathbf{y} \mid \mathbf{v}, \mathbf{A})}{d\mathbf{y}_k} = 0, \quad (\mathrm{M1})$$