

# Machine Intelligence 2

## 4.2 Pairwise Clustering

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2017

# Pairwise Clustering

# Clustering problem

"how different are data points?"

- *observations*: set of  $p$  "objects"  
 $\alpha = 1, \dots, p$
- *distance matrix*  $\{d_{\alpha\alpha'}\}$

	1	2	3	$p$
1	0	1.7	0.99	3.0
2	1.7	0	0.3	...
3	0.9	0.3	0	0.2
⋮	⋮	⋮	⋮	⋮
$p$	3.0	0.1	0.2	...

relational representation  
"pairwise data"

## Common constraints

- distance to self is zero
- symmetry

## Examples cases

- distances derived from an underlying vector space representation

$$d : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}_0^+, \text{ e.g. } d_{\alpha\alpha'} = \frac{1}{2} (\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{x}}^{(\alpha')})^2$$

- elements derived via a "kernel trick" (*cf. chapter 2.3.3*)

$$\underline{\phi} : \underline{\mathbf{x}}^{(\alpha)} \rightarrow \underline{\phi}_{(\underline{\mathbf{x}}^{(\alpha)})} \equiv \underline{\phi}^{(\alpha)}$$

$$\begin{aligned} d_{\alpha\alpha'} &= \frac{1}{2} (\underline{\phi}^{(\alpha)} - \underline{\phi}^{(\alpha')})^2 \\ &= \frac{1}{2} \left\{ (\underline{\phi}^{(\alpha)})^2 - 2(\underline{\phi}^{(\alpha)})^T \underline{\phi}^{(\alpha')} + (\underline{\phi}^{(\alpha')})^2 \right\} \\ &= \frac{1}{2} \left\{ k_{(\underline{\mathbf{x}}^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha)})} + k_{(\underline{\mathbf{x}}^{(\alpha')}, \underline{\mathbf{x}}^{(\alpha')})} - 2k_{(\underline{\mathbf{x}}^{(\alpha)}, \underline{\mathbf{x}}^{(\alpha')})} \right\} \end{aligned}$$

## Example cases

- distances directly determined by measurements (e.g. dissimilarity judgments in a psychophysics experiment, e.g. confusion matrices)
- distances determined through algorithms (e.g. sequence alignment procedures, graph-similarity measures)

# Problem statement

- set of clusters (partitions):  $q = 1, \dots, M$
- binary assignment variable :



$$m_q^{(\alpha)} = \begin{cases} 1, & \text{if object } \alpha \text{ belongs to cluster } q \\ 0, & \text{else} \end{cases}$$

- distance matrix  $d_{\alpha\alpha'}$
- cost function & model selection:

av. distance between  
 $\alpha$  and all other objects  $\alpha'$   
 from the same cluster  $q$

$$E[\{m_q^{(\alpha)}\}] = \frac{1}{2p} \sum_q \sum_{\alpha} \frac{\overbrace{\sum_{\alpha'} m_q^{(\alpha)} m_q^{(\alpha')} d_{\alpha\alpha'}}^{\text{av. distance between } \alpha \text{ and all other objects } \alpha' \text{ from the same cluster } q}}{\sum_{\alpha'} m_q^{(\alpha')}} \stackrel{!}{=} \min$$

# Pairwise clustering with squared Euclidean distance

→ observations (feature vectors):  $\underline{\mathbf{x}}^{(\alpha)}, \alpha = 1, \dots, p; \underline{\mathbf{x}}^{(\alpha)} \in \mathbb{R}^N$

→ distance measure:

$$d_{\alpha\alpha'} = \frac{1}{2} (\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{x}}^{(\alpha')})^2$$

$$E[\{m_q^{(\alpha)}\}] = \frac{1}{2p} \sum_q \frac{\sum_{\alpha\alpha'} m_q^{(\alpha)} m_q^{(\alpha')} (\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{x}}^{(\alpha')})^2}{\sum_{\alpha} m_q^{(\alpha)}}$$

...see blackboard    [OneNote](#)

$$= \frac{1}{p} \sum_{q,\alpha} m_q^{(\alpha)} (\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_q)^2 \quad \text{with} \quad \underline{\mathbf{w}}_q = \frac{\sum_{\alpha} m_q^{(\alpha)} \underline{\mathbf{x}}^{(\alpha)}}{\sum_{\alpha} m_q^{(\alpha)}}$$

$$= E[\{m_q^{(\alpha)}\}, \{\underline{\mathbf{w}}_q\}] \stackrel{\hat{=}}{\sim} \text{cost function of K-means clustering}$$

# The mean-field approximation for pairwise clustering

Discrete (binary) optimization problem

$$E\left[\left\{m_q^{(\alpha)}\right\}\right] = \frac{1}{2p} \sum_q \frac{\sum_{\alpha\alpha'} m_q^{(\alpha)} m_q^{(\alpha')} d_{\alpha\alpha'}}{\sum_{\alpha'} m_q^{(\alpha')}} \stackrel{!}{=} \min$$



- ⇒ gradient-based methods are not applicable
- ⇒ methods from combinatorial optimization are needed

# The mean-field approximation for pairwise clustering

simulated annealing	vs.	mean-field annealing
straightforward but slow		approximation why? good and fast!

## Application of Mean-Field Annealing (c.f section Stochastic Optimization)

- ~ variables  $m_q^{(\alpha)}$  are *normalized* to  $\sum_q m_q^{(\alpha)} = 1$
- ~ calculation of moments and mean-fields must be adapted because marginalized variables are no longer independent

# The mean-field approximation for pairwise clustering

Nomenclature ( $\otimes \rightarrow \text{set-product}$ )

$\{\underline{\mathbf{m}}^{(\alpha)}\}$ : set of all  $M$ -dimensional binary vectors  $(m_1^{(\alpha)}, m_2^{(\alpha)}, \dots, m_M^{(\alpha)})^T$  which fulfill the normalization condition (exactly one element equals 1).

$\mathcal{M}$ :  $\{\underline{\mathbf{m}}^{(1)}\} \otimes \{\underline{\mathbf{m}}^{(2)}\} \otimes \dots \otimes \{\underline{\mathbf{m}}^{(p)}\}$   
set product between all possible binary assignment variables i.e. all possible valid assignments for the full dataset

$\mathcal{M}_\gamma$ :  $\{\underline{\mathbf{m}}^{(1)}\} \otimes \dots \otimes \{\underline{\mathbf{m}}^{(\gamma-1)}\} \otimes \{\underline{\mathbf{m}}^{(\gamma+1)}\} \otimes \dots \otimes \{\underline{\mathbf{m}}^{(p)}\}$   
set of all possible assignments for all data points except  $\gamma$

# The mean-field approximation for pairwise clustering

assignment noise  $\rightarrow$  Gibbs distribution

$$P(\{m_q^{(\alpha)}\}) = \frac{1}{Z_p} \exp \left\{ -\beta E_p^p [\{m_q^{(\alpha)}\}] \right\}$$

where

$$Z_p = \sum_{\mathcal{M}} \exp \left\{ -\beta E_p^p [\{m_q^{(\alpha)}\}] \right\}$$

factorizing distribution

$$Q[\{m_q^{(\alpha)}\}] = \frac{1}{Z_Q} \exp \left\{ -\beta \sum_{p,\gamma} m_p^{(\gamma)} \underbrace{e_p^{(\gamma)}}_{\text{mean-fields}} \right\}$$

where:

$$Z_Q = \sum_{\mathcal{M}} \exp \left\{ -\beta \sum_{p,\gamma} m_p^{(\gamma)} e_p^{(\gamma)} \right\}$$

# Calculation of moments

→ moments of  $Q$  can be computed

$$\langle m_q^{(\gamma)} \rangle_Q = \langle m_q^{(\gamma)} \rangle_Q = \frac{1}{Z_Q} \sum_{\mathcal{M}} m_q^{(\gamma)} \exp \left\{ -\beta \sum_{r,\delta} m_r^{(\delta)} e_r^{(\delta)} \right\}$$

see blackboard OneNote

$$= \frac{\exp \left\{ -\beta m_q^{(\gamma)} e_q^{(\gamma)} \right\}}{\sum_r \exp \left\{ -\beta m_r^{(\gamma)} e_r^{(\gamma)} \right\}} = \underbrace{\frac{\exp \left\{ -\beta e_q^{(\gamma)} \right\}}{\sum_r \exp \left\{ -\beta e_r^{(\gamma)} \right\}}}_{\text{soft-max of the mean-fields}}$$

## Intuition for above result

- $\sum_r \langle m_r^{(\gamma)} \rangle = 1$  and  $\langle m_q^{(\gamma)} \rangle_Q \in [0, 1] \Rightarrow$  assignment probabilities 
- $\beta \rightarrow \infty : \langle m_q^{(\gamma)} \rangle_Q \in \{0, 1\} \Rightarrow$  "hard assignments" (cmp. k-means)

# Minimization of the KL-divergence

Mean Field equation (c.f. section on Stochastic Optimization)

$$\frac{\partial}{\partial e_l} \langle E_p \rangle_Q - \sum_k e_k \frac{\partial}{\partial e_l} \langle s_k \rangle_Q = 0$$

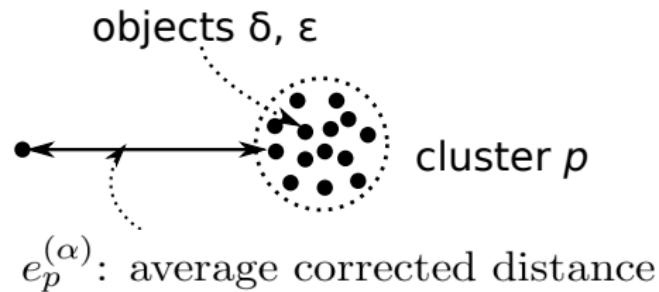
$$\frac{\partial \langle E^p \rangle_Q}{\partial e_q^{(\alpha)}} - \sum_{r,\gamma} \underbrace{\frac{\partial \langle m_r^{(\gamma)} \rangle_Q}{\partial e_q^{(\alpha)}}}_{\text{depends only on data point } \gamma} e_r^{(\gamma)} \stackrel{!}{=} 0$$

$$\frac{\partial \langle E^p \rangle_Q}{\partial e_q^{(\alpha)}} - \sum_r \frac{\partial \langle m_r^{(\alpha)} \rangle_Q}{\partial e_q^{(\alpha)}} e_r^{(\alpha)} \stackrel{!}{=} 0$$

# Solution of the mean field equation

Assumptions applied to distance matrix

- ~ symmetric:  $d_{\alpha\alpha'} = d_{\alpha'\alpha}$
- ~ diagonal elements  $d_{\alpha\alpha'} \stackrel{!}{=} 0$



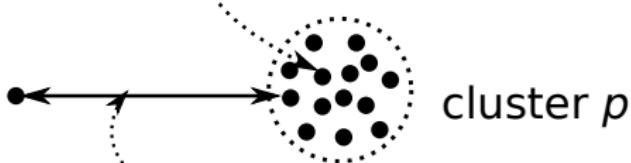
$$e_q^{(\alpha)} = \frac{2}{p} \underbrace{\frac{1}{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q} \sum_{\delta} \langle m_q^{(\delta)} \rangle_Q}_{\text{average corrected distance between data objects } \alpha \text{ and all objects } \delta \text{ of cluster } q} \underbrace{\left\{ d_{\delta\alpha} - \frac{1}{2} \frac{1}{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q} \sum_{\varepsilon} \langle m_q^{(\varepsilon)} \rangle_Q d_{\varepsilon\delta} \right\}}_{\text{distance between data objects } \alpha \text{ and } \delta, \text{ corrected by the average distance between objects of the cluster, to which } \delta \text{ belongs (here: } q\text{)}}$$

average corrected distance between data objects  
 $\alpha$  and all objects  $\delta$  of cluster  $q$

## Remarks

⇒ "metric visualization" of the  $e_q^{(\alpha)}$ :

objects  $\delta, \varepsilon$



$e_p^{(\alpha)}$ : average corrected distance

- ⇒ mean-fields depend on assignment probabilities  $\langle m_q^{(\alpha)} \rangle_Q$  rather than on the "hard" binary assignments → effect of the underlying stochastic optimization procedure
- ⇒ "fuzzy" memberships: objects contribute only weighted by their probability  $\langle m_q^{(\alpha)} \rangle_Q$  of assignment
- ⇒ assignment probabilities  $\langle m_1^{(\alpha)} \rangle_Q$  depend on the average normalized distance between the object  $\alpha$  under consideration and the objects of cluster  $q$  (probability is high, if distance to  $\alpha$  is small compared to average distance within cluster).

# Mean-field annealing for pairwise clustering

---

**Algorithm 1:** Mean-field annealing for pairwise clustering
 

---

**Initialization:**

- max. number  $M$  of partitions, initial ( $\beta_0$ ) and final ( $\beta_f$ ) values of the noise parameter, annealing factor  $\eta$ , convergence criterion  $\theta$
- initialize mean-fields  $e_q^{(\alpha)}$  with random numbers  $\in [0, 1]$
- $\beta \leftarrow \beta_0$

**while**  $\beta < \beta_f$  **do** annealing

**repeat** EM (fixed point iteration)

compute assignment probabilities:  $\langle m_q^{(\alpha)} \rangle_Q = \frac{\exp\left\{-\beta(e_q^{(\alpha)})_{\text{old}}\right\}}{\sum_r \exp\left\{-\beta(e_r^{(\alpha)})_{\text{old}}\right\}} \quad \forall q, \alpha$

compute new mean-fields:

$$(e_q^{(\alpha)})_{\text{new}} = \frac{2}{p} \frac{1}{\sum_\gamma \langle m_q^{(\gamma)} \rangle_Q} \sum_\delta \langle m_q^{(\delta)} \rangle_Q \cdot \left\{ d_{\delta\alpha} - \frac{1}{2} \frac{1}{\sum_\gamma \langle m_q^{(\gamma)} \rangle_Q} \sum_\varepsilon \langle m_q^{(\varepsilon)} \rangle_Q d_{\varepsilon\delta} \right\} \quad \forall q, \alpha$$

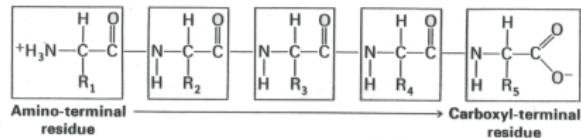
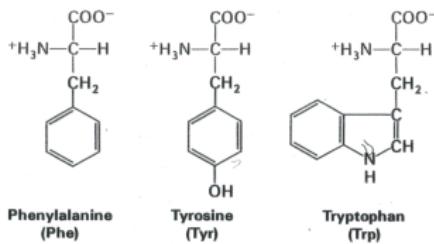
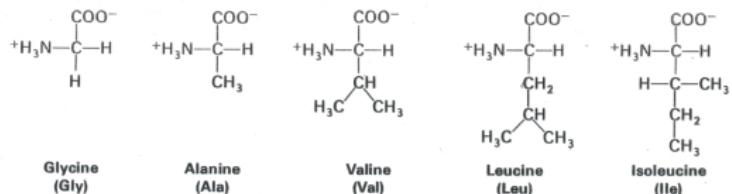
**until**  $|(e_q^{(\alpha)})_{\text{new}} - (e_q^{(\alpha)})_{\text{old}}| < \theta \quad \forall q, \alpha$

$\beta \leftarrow \eta \cdot \beta$

**end**



# Example: clustering of protein sequences



# Example: clustering of protein sequences

## Hemoglobin and Myoglobin



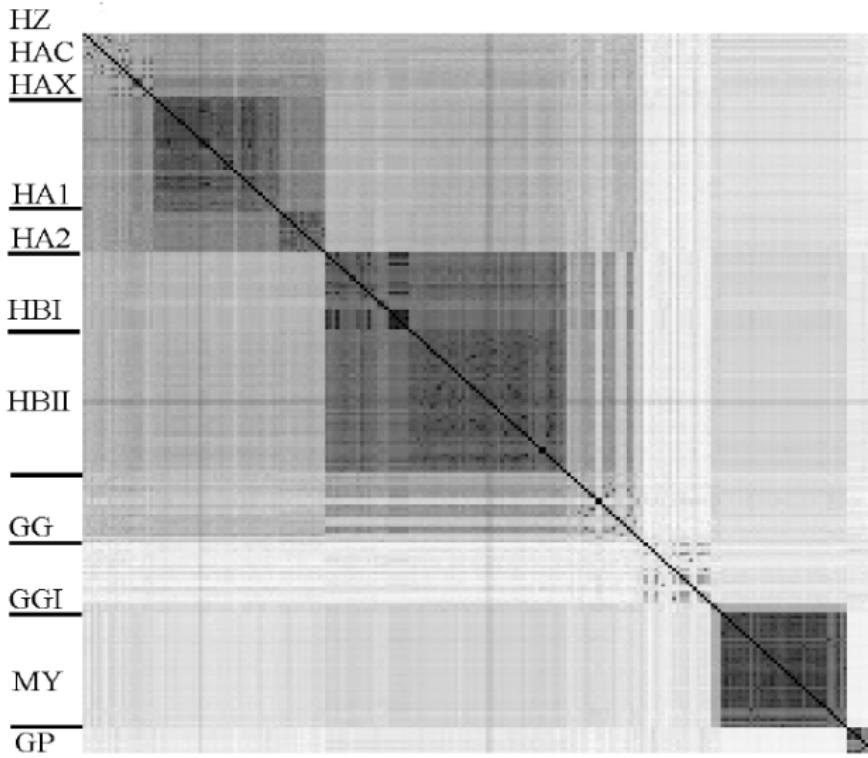
Cpe1\$Human	G.DLPAFHA	HRDRGII.FNNGPTW	KDIRRFSEYLRLNYGMGKQGNESRIQRE
Cpe1\$Rabbit	G.EIPAFRE	FKKDGGII.FNNGPTW	KDTRRFSLTTLRDYGMGKQGNEDRIQKE
Cpe1\$Rat51	G.DIPVFQE	YKNKGII.FNNGPTW	KDTRRFSLTTLRDYGMGKQGNESRIQRE
CpF1 Human	GDYPAAFFNF	TKGNGIA.FSSGDRW	KVLRQFSIQILRNFGMGKRSIEERILEE
Cpf2\$Chick	GNPLLFKEV	FKGTGIV.TSNGESW	RQMRRFALTLRDFGMGKKKSIEERIQEE
Cpg1\$Rat	GEMPTLEKN	FQGYGLA.LSNGERW	KILRRFSLTVLRNFGMGKRSIEERIQEE
CpG1\$Rabbit	GELASVERN	FQGHGVALAN.GERW	RILRRFSLTILRDFGMGKRSIEERIQEE
cph1\$Chick	GILPLIEKL	FKGTGIV.TSNGETW	RQLRRFALTLRDFGMGKKGIEERIQEE

# Example: clustering of protein sequences



Distance Matrix

Clustering Groups



# Missing data

	1	2	3	$p$
1	0	1.7	0.99	3.0
2	1.7	0	0.3	...
3	0.9	0.3	0	0.2
:	:		⋮⋯	⋮
$p$	3.0	0.1	0.2	...

relational representation  
"pairwise data"

- ~ number of matrix elements  $\sim p^2$
- ~ calculation or measurement of all distances may be computationally expensive or even unfeasible
- ~ **distance matrices** are often "**redundant**" ~ not all matrix entries are needed



# Missing data

$$e_q^{(\alpha)} = \frac{2}{p} \frac{1}{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q} \sum_{\delta} \langle m_q^{(\delta)} \rangle_Q \left\{ d_{\delta\alpha} - \frac{1}{2} \frac{1}{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q} \sum_{\varepsilon} \langle m_q^{(\varepsilon)} \rangle_Q d_{\varepsilon\delta} \right\}$$

  $\bar{d}_{q\alpha} = \frac{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q d_{\gamma\alpha}}{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q}$

$\underbrace{\phantom{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q}}$

average distance of data object  $\alpha$   
all data objects of cluster  $q$

$$e_q^{(\alpha)} = \underbrace{\frac{2}{p} \left\{ \bar{d}_{q\alpha} - \frac{1}{2} \sum_{\delta} \langle m_q^{(\delta)} \rangle_Q \bar{d}_{q\delta} \right\}}_{\text{mean fields in terms of } \bar{d}_{q\alpha}}$$

# Missing data

## Mean fields

$$e_q^{(\alpha)} = \frac{2}{p} \left\{ \bar{d}_{q\alpha} - \frac{1}{2} \sum_{\delta} \langle m_q^{(\delta)} \rangle_Q \bar{d}_{q\delta} \right\}$$

computations depend only on the *average* distances and therefore enable the following *missing data heuristics*:

- ⇒ estimate average values  $\bar{d}_{p\alpha}$  using the measured distances only
- ⇒ perform summations within  $\bar{d}_{p\alpha}$  only over the available distances

# Squared Euclidean distances

Vectorial data:  $\underline{x}^{(\alpha)}$ ;  $\alpha = 1, \dots, p$ ;  $\underline{x}^{(\alpha)} \in \mathbb{R}^N$ , euclidean squared distances,

$$d_{\alpha\alpha'} = \frac{1}{2} (\underline{x}^{(\alpha)} - \underline{x}^{(\alpha')})^2$$

Mean fields:

$$e_q^{(\alpha)} = \frac{1}{2} (\underline{x}^{(\alpha)} - \underline{\mathbf{w}}_q)^2$$



$$\underline{\mathbf{w}}_q = \frac{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q \underline{x}^{(\gamma)}}{\sum_{\gamma} \langle m_q^{(\gamma)} \rangle_Q}$$

→ center of mass of all objects weighted by their probability of assignment to cluster  $q$

**Algorithm 2:** Soft K-means clustering for Euclidean distances**Initialization:**

- choose maximum number  $M$  of partitions
- choose initial ( $\beta_0$ ) and final ( $\beta_f$ ) values of the noise parameter
- initialize prototypes:  $\underline{\mathbf{w}}_q = \frac{1}{p} \sum_{\alpha} \underline{\mathbf{x}}^{(\alpha)} + \underline{\eta}_q$  (small random vector)
- choose annealing factor  $\eta$ , convergence criterion  $\theta$
- $\beta \leftarrow \beta_0$

**while**  $\beta < \beta_f$  (*annealing*) **do**

**repeat** EM

$$\text{compute assignment probabilities: } \langle m_q^{(\alpha)} \rangle_Q = \frac{\exp \left\{ -\frac{\beta}{2} (\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_q^{\text{old}})^2 \right\}}{\sum_r \exp \left\{ -\frac{\beta}{2} (\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_r^{\text{old}})^2 \right\}} \quad \forall \alpha, q$$

$$\text{compute new prototypes: } \underline{\mathbf{w}}_q^{\text{new}} = \frac{\sum_{\alpha} \langle m_q^{(\alpha)} \rangle_Q \underline{\mathbf{x}}^{(\alpha)}}{\sum_{\alpha} \langle m_q^{(\alpha)} \rangle_Q} \quad \forall q$$

**until**  $|\underline{\mathbf{w}}_q^{\text{new}} - \underline{\mathbf{w}}_q^{\text{old}}| < \theta$  *for all*  $q$

$$\beta \leftarrow \eta \beta$$

**end**

**Interpretation**

- ~ probabilistic cluster assignments.
- ~ natural extension of K-means clustering to the case of "fuzzy" assignments

# Soft K-means clustering on-line version

---

**Algorithm 3:** Soft K-means clustering on-line version
 

---

**Initialization:**

- choose maximum number  $M$  of partitions
- choose initial ( $\beta_0$ ) and final ( $\beta_f$ ) values of the noise parameter
- initialize prototypes:  $\underline{\mathbf{w}}_q = \underline{\mathbf{x}}^{(0)} + \underline{\eta}_q$  (small random vector)
- choose annealing factor  $\eta$ , convergence criterion  $\theta$
- $\beta \leftarrow \beta_0$

**while**  $\beta < \beta_f$  (*annealing*) **do**

**repeat** EM

choose observation  $\underline{\mathbf{x}}^\alpha$

compute assignment probabilities:  $\langle m_q^{(\alpha)} \rangle_Q = \frac{\exp\left\{-\frac{\beta}{2}(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_q^{\text{old}})^2\right\}}{\sum_r \exp\left\{-\frac{\beta}{2}(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_r^{\text{old}})^2\right\}}$   $\forall q$

compute new prototypes:  $\underline{\mathbf{w}}_q^{\text{new}} = \underline{\mathbf{w}}_q^{\text{old}} + \varepsilon \langle m_q^{(\alpha)} \rangle_Q (\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_q) \forall q$

**until**  $|\underline{\mathbf{w}}_q^{\text{new}} - \underline{\mathbf{w}}_q^{\text{old}}| < \theta \forall q$

$\beta \leftarrow \eta \beta$

change  $\varepsilon$  according to annealing schedule

**end**

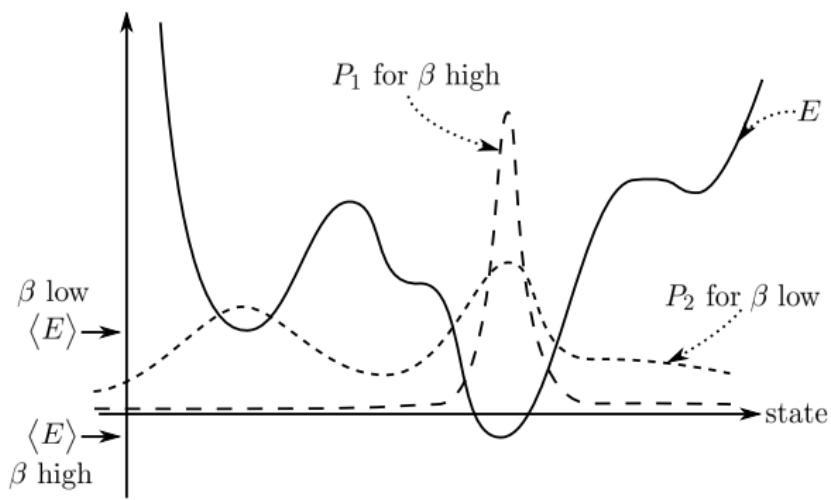
# Phase transitions in clustering

average cost (Gibb's distribution of simulated annealing):

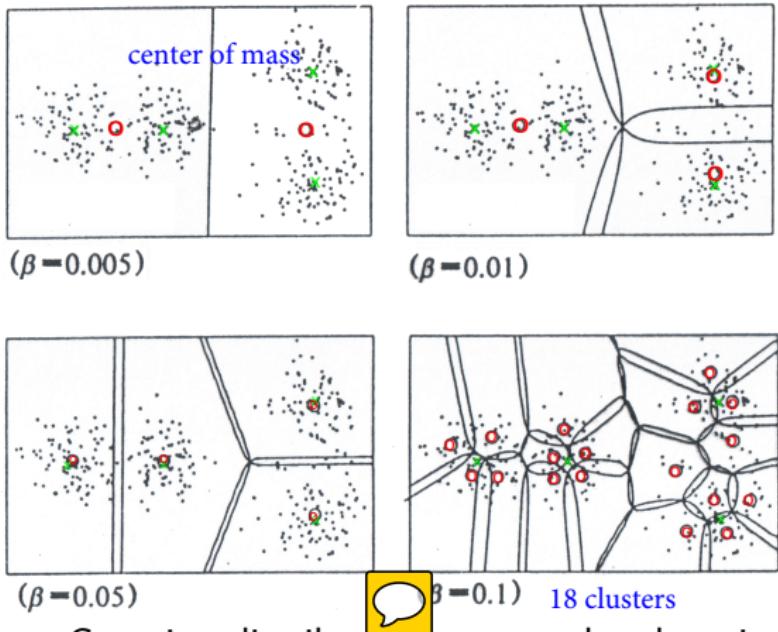
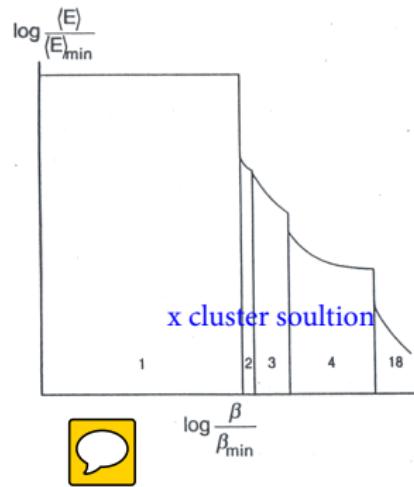
$$\langle E \rangle = \frac{1}{Z} \sum_{\{m_p^{(\alpha)}\}} E \exp \{ -\beta E \}$$


increase of  $\beta$  implies:

- decrease of average cost
- decrease of cluster size
- increase in spatial resolution (hierarchical clustering)

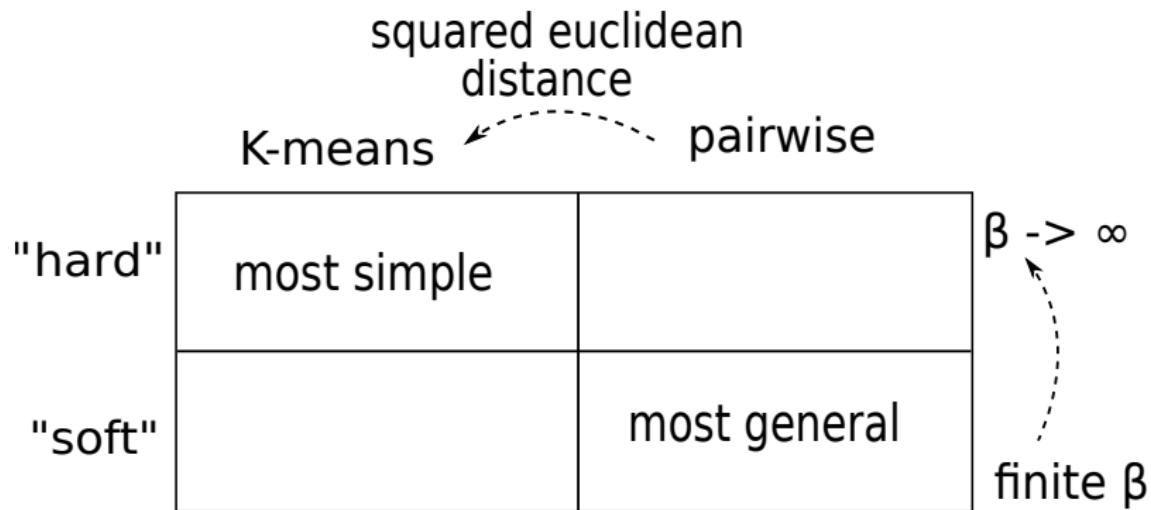


# Phase transition in clustering



- data generated from four Gaussian distributions centered at locations "x"
- calculated cluster centers at location "o"

## Comments



- principled alternative to fuzzy clustering methods
  - mean-field annealing is robust against convergence to local optima
  - choice of terminal value of noise parameter  $\beta \Rightarrow$  "resolution" of cluster analysis