Machine Intelligence 2

4.2 Pairwise Clustering

Prof. Dr. Klaus Obermayer

Fachgebiet Neuronale Informationsverarbeitung (NI)

SS 2018

Pairwise Clustering

Clustering problem

→ *observations:* set of p "objects"
$\alpha = 1, \ldots, p$

→ *distance matrix* $\{d_{\alpha\alpha'}\}$

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

relational representation
"pairwise data"

Common constraints

→ distance to self is zero

→ symmetry

Examples cases

- distances derived from an underlying vector space representation

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

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

$$\underline{\phi} : \underline{\mathbf{x}}^{(\alpha)} \to \underline{\phi}_{\left(\underline{\mathbf{x}}^{(\alpha)}\right)} \equiv \underline{\phi}^{(\alpha)}$$

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

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): $q = 1, \ldots, M$
→ 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:

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

Pairwise clustering with squared Euclidean distance

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

→ distance measure:

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

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

...see blackboard

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

$$= E_{\left[\left\{m_q^{(\alpha)}\right\},\left\{\underline{\mathbf{w}}_q\right\}\right]} \widehat{=} \text{ cost function of K-means clustering}$$

The mean-field approximation for pairwise clustering

Discrete (binary) optimization problem

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

$\Rightarrow$ gradient-based methods are not applicable

$\Rightarrow$ 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)

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

The mean-field approximation for pairwise clustering

Nomenclature ($\otimes \rightarrow$ *set-product*)

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

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

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

The mean-field approximation for pairwise clustering

assignment noise $\rightarrow$ Gibbs distribution

$$P_{\left(\left\{m_q^{(\alpha)}\right\}\right)} = \frac{1}{Z_p} \exp\left\{ -\beta E^p_{\left[\left\{m_q^{(\alpha)}\right\}\right]} \right\}$$

where

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

factorizing distribution

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

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

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

see blackboard

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

Intuition for above result

→ $\sum_r \langle m_r^{(\gamma)} \rangle = 1$ and $\left\langle m_q^{(\gamma)} \right\rangle_Q \in [0, 1]$ => assignment probabilities

→ $\beta \to \infty : \left\langle m_q^{(\gamma)} \right\rangle_Q \in \{0, 1\}$ => "hard assignments" (cmp. k-means)

Minimization of the KL-divergence

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

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

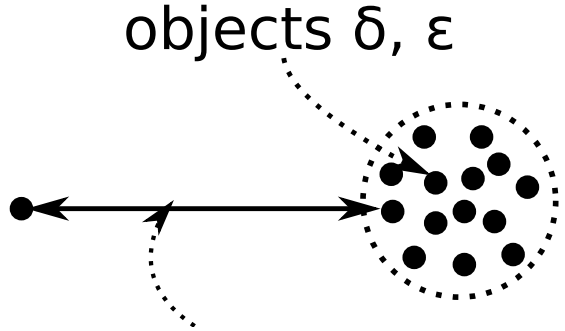
$$\frac{\partial\left\langle E^p \right\rangle_Q}{\partial e_q^{(\alpha)}} - \sum_{r,\gamma} \overbrace{\frac{\partial\left\langle m_r^{(\gamma)} \right\rangle_Q}{\partial e_q^{(\alpha)}}}^{\text{depends only on data point } \gamma} e_r^{(\gamma)} \stackrel{!}{=} 0$$

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

Solution of the mean field equation

Assumptions applied to distance matrix

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

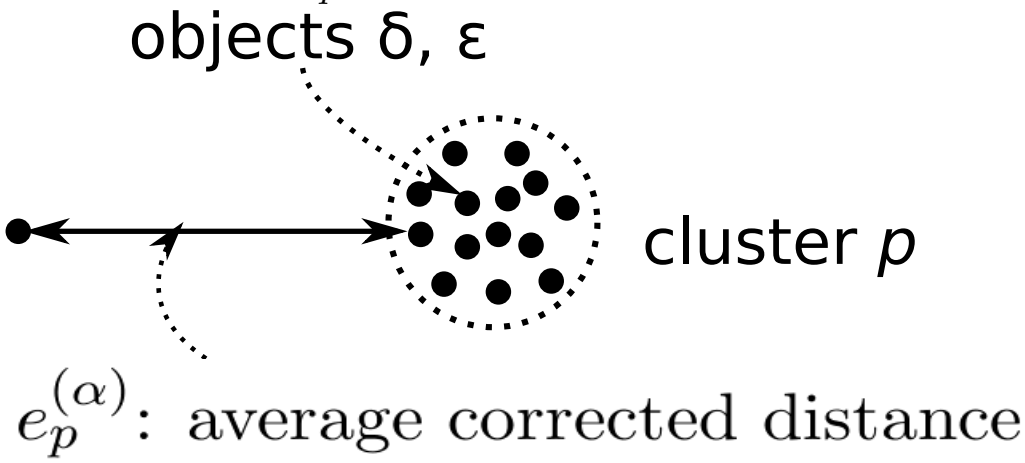


$e_p^{(\alpha)}$: average corrected distance

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

Remarks

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



$\Rightarrow$ mean-fields depend on assignment probabilities $\langle m_q^{(\alpha)} \rangle_Q$ rather than on the "hard" binary assignments $\rightarrow$ effect of the underlying stochastic optimization procedure

$\Rightarrow$ "fuzzy" memberships: objects contribute only weighted by their probability $\langle m_q^{(\alpha)} \rangle_Q$ of assignment

$\Rightarrow$ assignment probabilities $\langle m_1^{(\alpha)} \rangle_Q$ depend on the average normalized distance between the object α under consideration and the objects of cluster q (probability is high, if distance to α 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 (β_0) and final (β_f) values of the noise parameter, annealing factor η, convergence criterion θ
- 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: $\left\langle m_q^{(\alpha)} \right\rangle_Q = \frac{\exp\left\{-\beta\left(e_q^{(\alpha)}\right)_{\text{old}}\right\}}{\sum_r \exp\left\{-\beta\left(e_r^{(\alpha)}\right)_{\text{old}}\right\}} \quad \forall q, \alpha$

    compute new mean-fields:

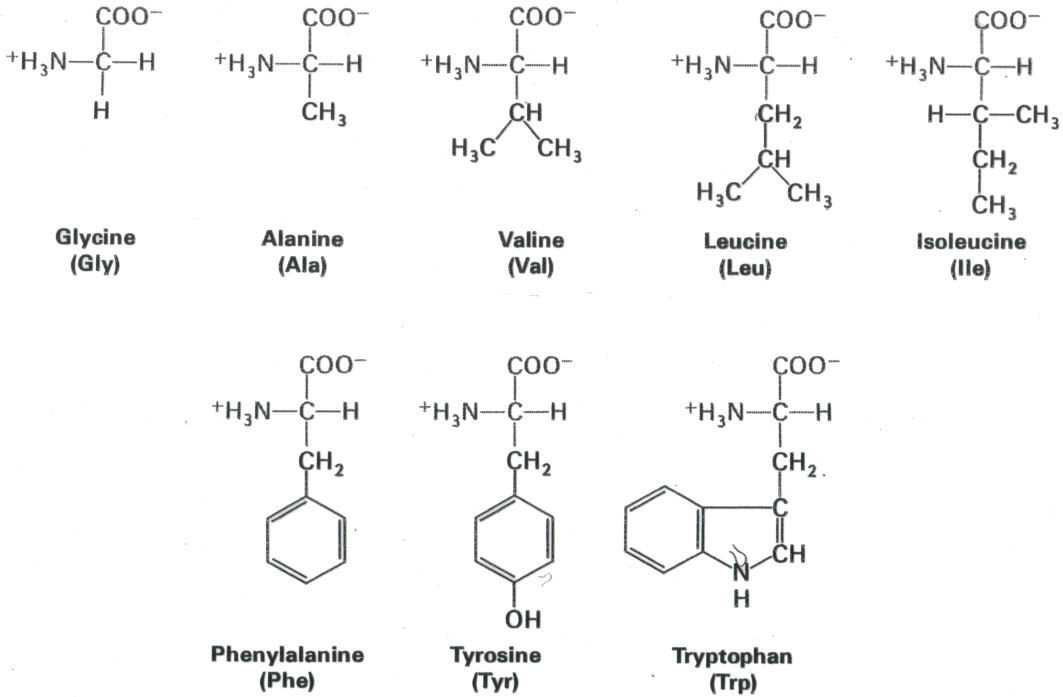
$$\left(e_q^{(\alpha)}\right)_{\text{new}} = \frac{2}{p} \frac{1}{\sum_\gamma \left\langle m_q^{(\gamma)} \right\rangle_Q} \sum_\delta \left\langle m_q^{(\delta)} \right\rangle_Q \cdot \left\{ d_{\delta\alpha} - \frac{1}{2} \frac{1}{\sum_\gamma \left\langle m_q^{(\gamma)} \right\rangle_Q} \sum_\varepsilon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

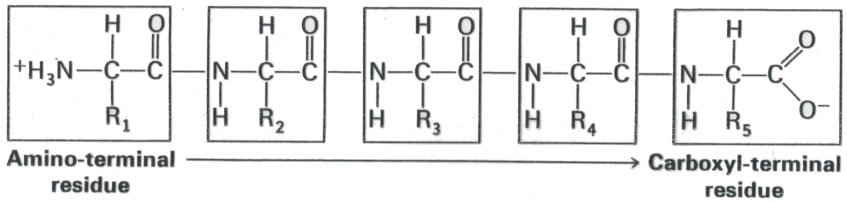
  until $\left|\left(e_q^{(\alpha)}\right)_{\text{new}} - \left(e_q^{(\alpha)}\right)_{\text{old}}\right| < \theta \quad \forall q, \alpha$

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

end

Example: clustering of protein sequences





Example: clustering of protein sequences

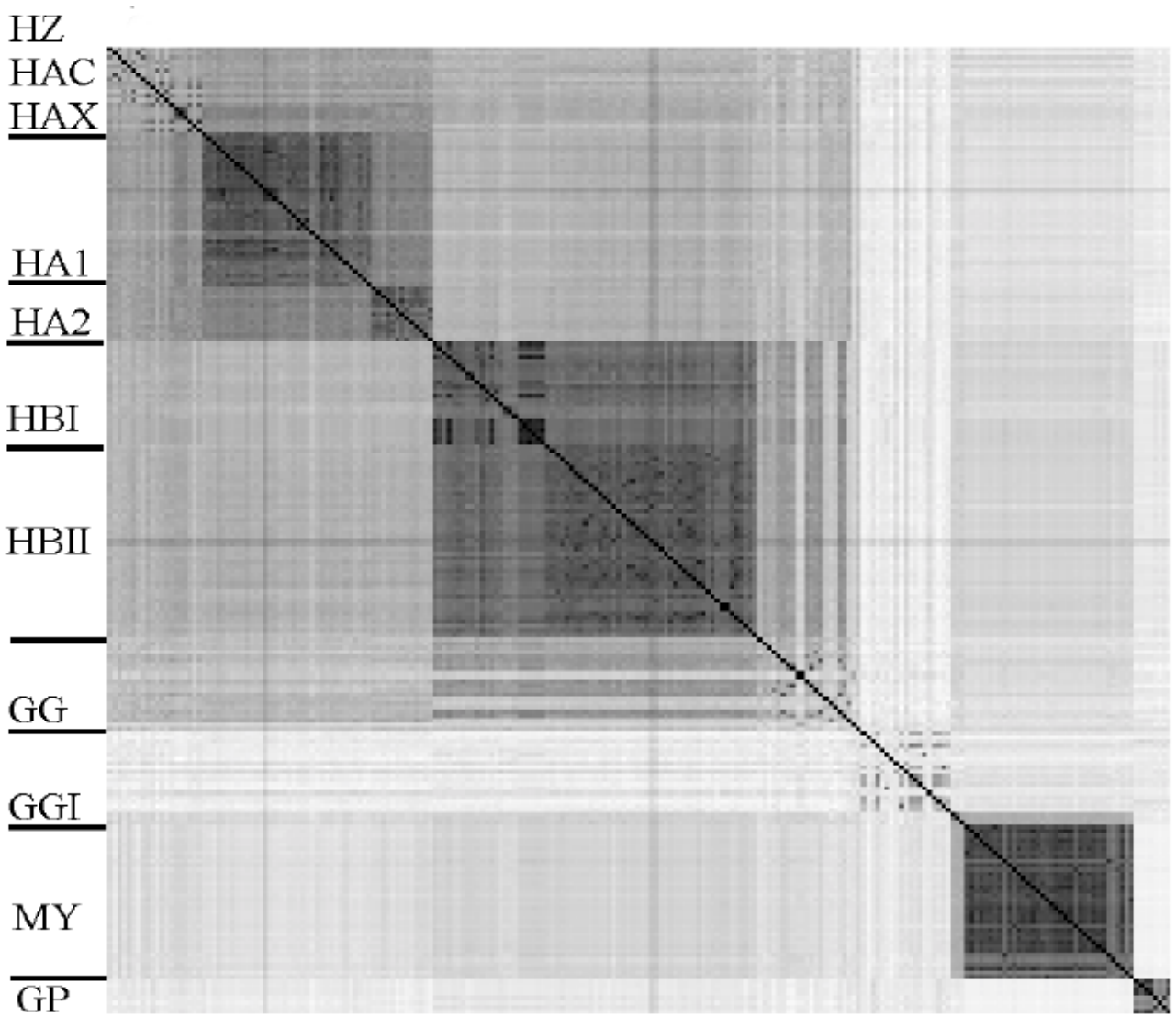
Hemoglobin and Myoglobin

```
Cpe1$Human      G.DLPAFHA    HRDRGII.FNNGPTW   KDIRRFSLTTLRNYGMGKQGNESRIQRE
Cpe1$Rabit      G.EIPAFRE    FKKDGII.FNNGPTW   KDTRRFSLTTLRDYGMGKQGNEDRIQKE
Cpe1$Rat51      G.DIPVFQE    YKNKGII.FNNGPTW   KDTRRFSLTTLRDYGMGKQGNESRIQRE

CpF1 Human      GDYPAFFNF    TKGNGIA.FSSGDRW   KVLRQFSIQILRNFGMGKRSIEERILEE
Cpf2$Chick      GNLPLFEKV    FKGTGIV.TSNGESW   RQMRRFALTTLRDFGMGKKSIEERIQEE

Cpg1$Rat        GEMPTLEKN    FQGYGLA.LSNGERW   KILRRFSLTVLRNFGMGKRSIEERIQEE
CpG1$Rabit      GELASVERN    FQGHGVALAN.GERW   RILRRFSLTILRDFGMGKRSIEERIQEE
cph1$Chick      GILPLIEKL    FKGTGIV.TSNGETW   RQLRRFALTTLRDFGMGKKGIEERIQEE
```

Example: clustering of protein sequences



Missing data

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

relational representation
"pairwise data"

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

Missing data

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

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

average distance of data object α all data objects of cluster q

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

mean fields in terms of $\bar{d}_{q\alpha}$

Missing data

Mean fields

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

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

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

Squared Euclidean distances

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

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

Mean fields:

$$e_q^{(\alpha)} = \frac{1}{2}\left(\underline{\mathbf{x}}^{(\alpha)} - \underline{\mathbf{w}}_q\right)^2 \qquad \underline{\mathbf{w}}_q = \frac{\sum_\gamma \left\langle m_q^{(\gamma)} \right\rangle_Q \underline{\mathbf{x}}^{(\gamma)}}{\sum_\gamma \left\langle m_q^{(\gamma)} \right\rangle_Q}$$

$\rightarrow$ 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 (β_0) and final (β_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 η, convergence criterion θ
- $\beta \leftarrow \beta_0$

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

  repeat EM

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

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

  until $\left|\underline{\mathbf{w}}_q^{\text{new}} - \underline{\mathbf{w}}_q^{\text{old}}\right| < \theta$ *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 (β_0) and final (β_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 η, convergence criterion θ
- $\beta \leftarrow \beta_0$

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

  repeat EM

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

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

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

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

  $\beta \leftarrow \eta\beta$

  change ε according to annealing schedule

end

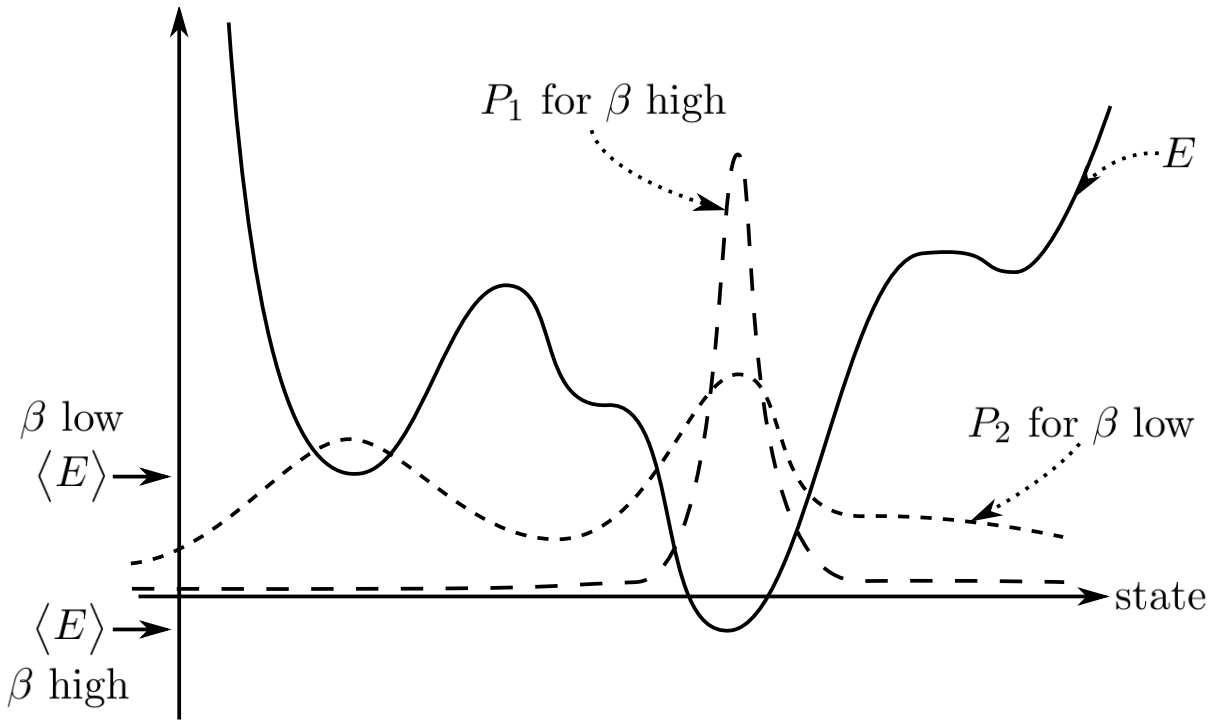
Phase transitions in clustering

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

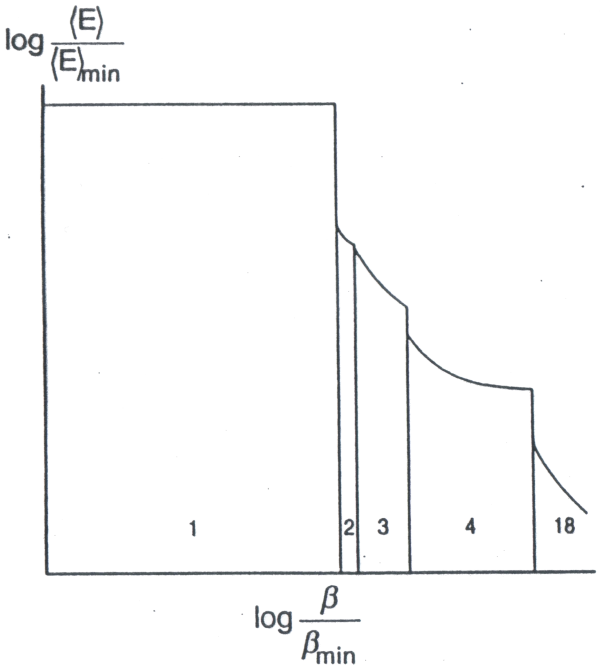
$$\langle E \rangle = \frac{1}{Z} \sum_{\{m_p^{(\alpha)}\}} E \exp\{-\beta E\}$$

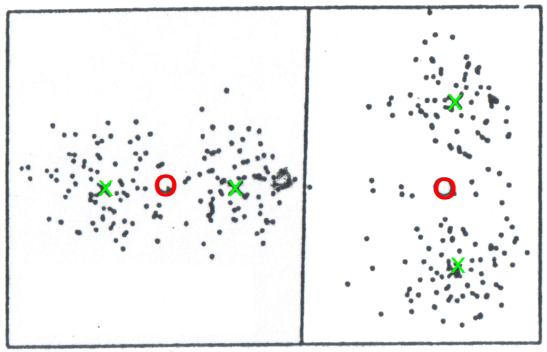
increase of β implies:

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

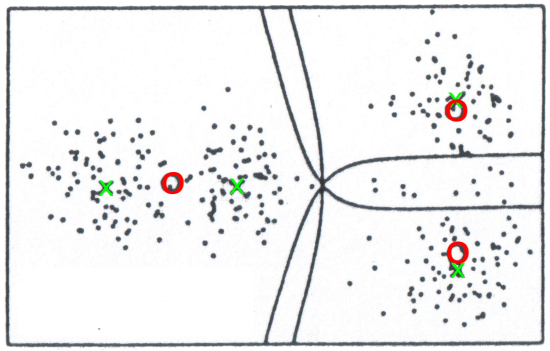


Phase transition in clustering

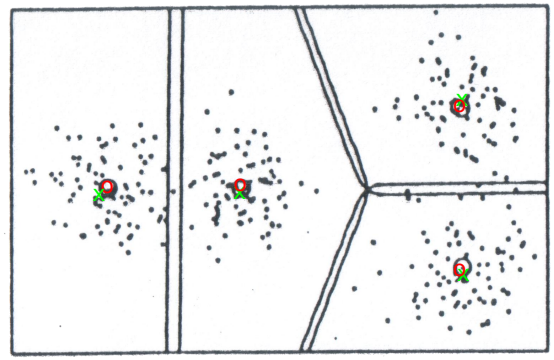




(β = 0.005)



(β = 0.01)

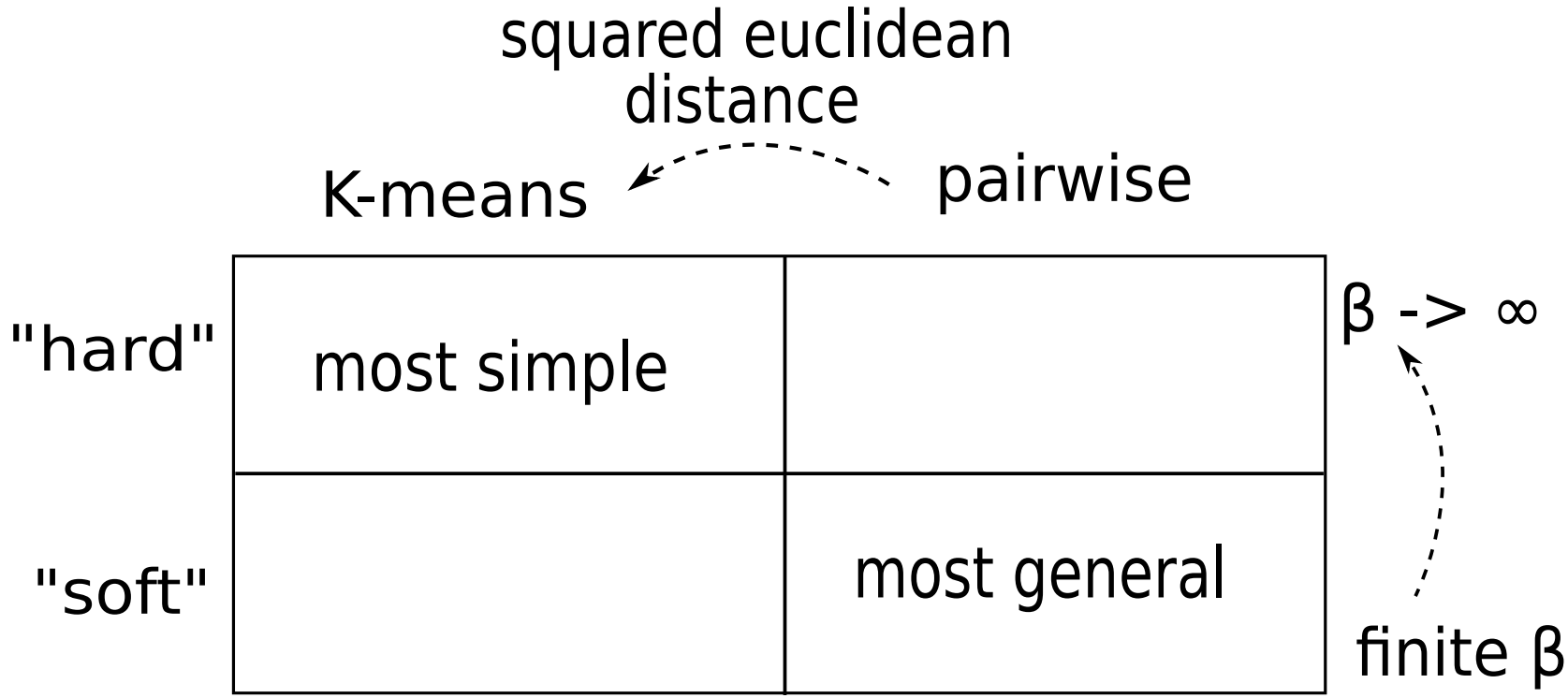


(β = 0.05)

(β = 0.1)

- 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