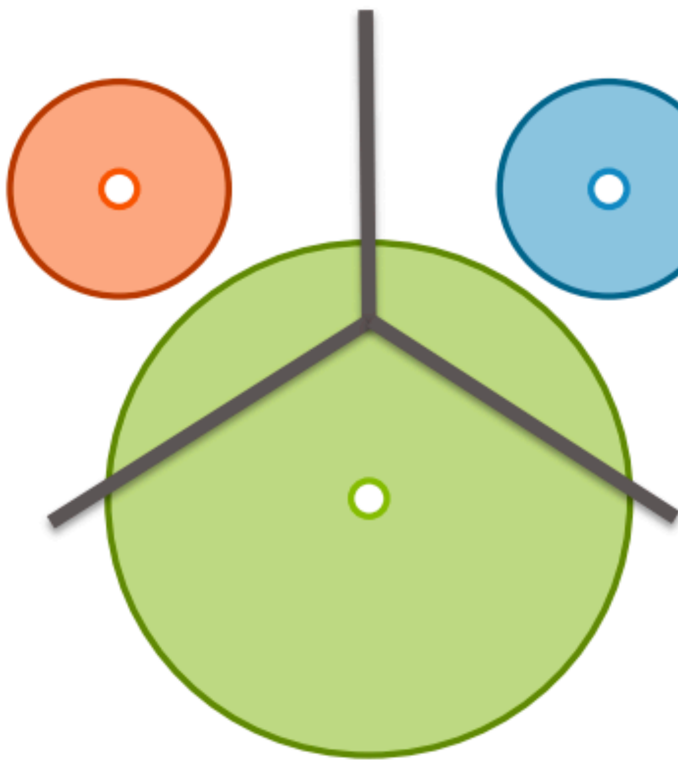
Expectation Maximization

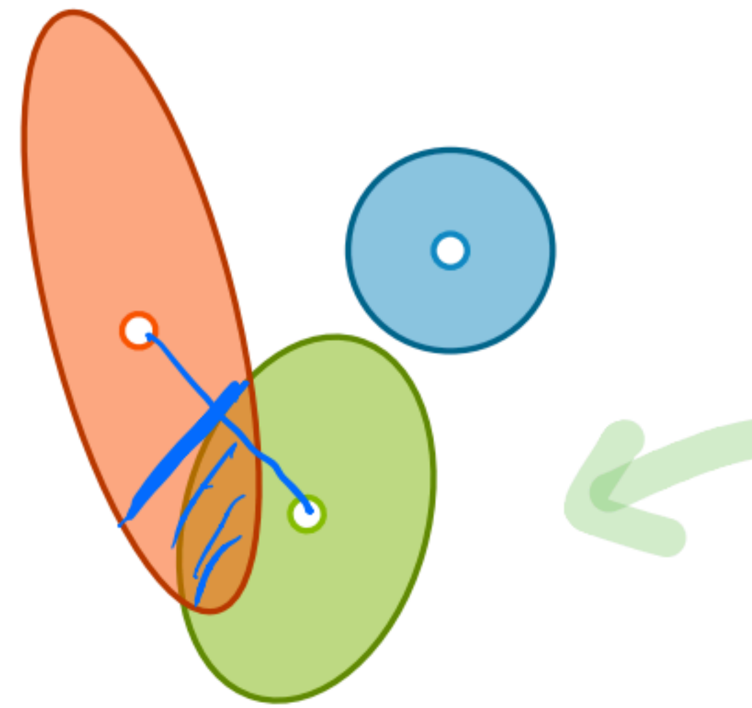
Sewoong Oh

CSE446
University of Washington

- K-means algorithm fails, when



disparate cluster sizes



different shaped/oriented clusters

- one way to capture such clustering is by training the parameters of a **Gaussian Mixture Model (GMM)** that best captures the data

demo: https://lukapopijac.github.io/gaussian-mixture-model/

Gaussian Mixture Model.

input: $\{X_i\}_{i=1}^{n}$, fix K: # of clusters

Parameters: $\pi = (\pi_1, ---, \pi_K) \in \mathbb{R}^K$: mixture weights

μ_j , $j \in \{1, ---, K\}$ $\in \mathbb{R}^d$: mean

$C_j \in \mathbb{R}^{d \times d}$, : Covariance.

$d = 1$, $K = 2$. Parameters. $\underbrace{\pi_1, \pi_2, \mu_1, \mu_2, C_1, C_2 \in \mathbb{R}}$

$$\mathbb{P}(X_i \mid \text{Parameters}) = \pi_1 \frac{1}{\sqrt{2\pi C_1}} e^{-\frac{(X_i - \mu_1)^2}{2C_1}} + \pi_2 \frac{1}{\sqrt{2\pi C_2}} e^{-\frac{(X_i - \mu_2)^2}{2C_2}}$$

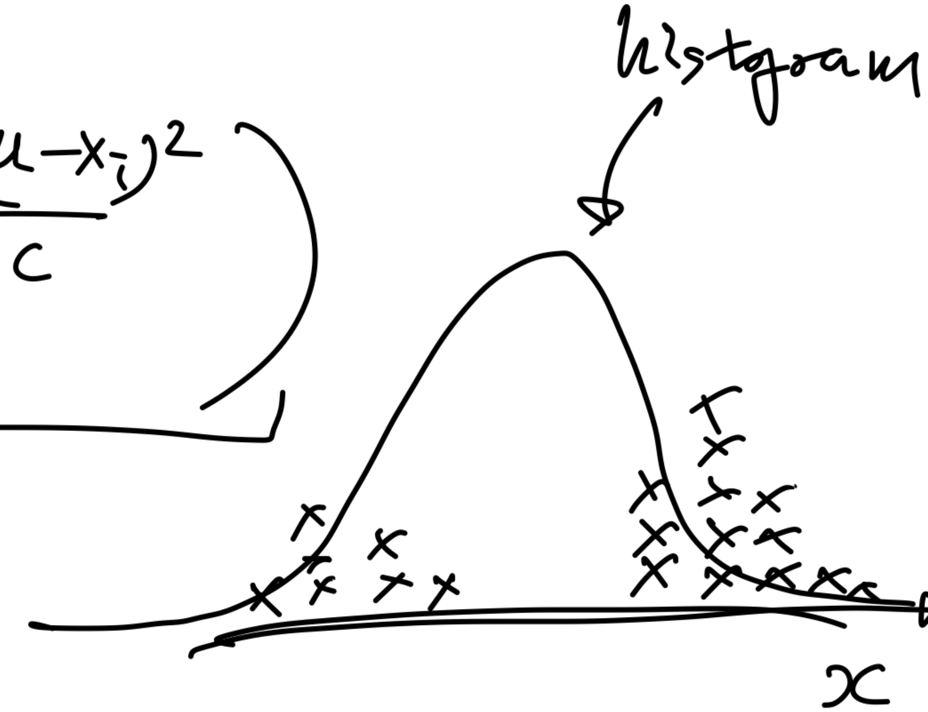
MLE:

$$\underset{\text{Parameters}}{\text{maximize}} \sum_{i=1}^{n} \log \mathbb{P}(X_i \mid \text{Parameters})$$

Toy problem: $N(\mu, C)$

$$\max_{\mu, C\in\mathbb{R}} \sum_{i=1}^{n} \underbrace{\log\left(\frac{1}{\sqrt{2\pi C}}\, e^{-\frac{(\mu-x_i)^2}{2C}}\right)}$$

$$\max_{\mu, C} \underbrace{\sum_{i=1}^{n} \left\{-\frac{(\mu-x_i)^2}{2C} - \frac{1}{2}\lg(2\pi C)\right\}}_{\mathcal{L}(\mu, C)}$$



$$\nabla_\mu \mathcal{L}(\mu, C) = \sum_{i=1}^{n} -\frac{2}{2C}\cdot(\mu - x_i) = 0 \longleftrightarrow n\mu = \sum x_i$$

$$\longleftrightarrow \boxed{\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i}$$

$$\nabla_C \mathcal{L}(\mu, C) = \sum_{i=1}^{n} \frac{(\mu-x_i)^2}{2C^2} - \frac{n}{2C} = 0 \longleftrightarrow \boxed{C^* = \frac{1}{n}\sum_{i=1}^{n}(\mu^* - x_i)^2}$$

MLE for GMM

Maximize $\sum_{i=1}^{n} \log \left(\pi_1 \boxed{\frac{1}{\sqrt{2\pi c_1}} e^{-\frac{(x_i-\mu_1)^2}{2c_1}}} + \pi_2 N(x_i \mid \mu_2, c_2) \right)$

$\pi_1, \pi_2, \mu_1, \mu_2, c_1, c_2$

$$\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{2\pi c_k}} e^{-\frac{(x_i-\mu_k)^2}{2c_k}}$$

---

define $r_i \triangleq \mathbb{P}(z_i = 1 \mid x_i) = \frac{\boxed{\mathbb{P}(z_i=1, x_i)}}{\mathbb{P}(z_i=1, x_i) + \mathbb{P}(z_i=2, x_i)}$ Bayes' Rule

$$= \frac{\pi_1 N(x_i \mid \mu_1, c_1)}{\pi_1 N(x_i \mid \mu_1, c_1) + \pi_2 N(x_i \mid \mu_2, c_2)}$$

$$1 - r_i = \mathbb{P}(z_i = 2 \mid x_i)$$

$N_1 = \sum_{i=1}^{n} r_i$, $N_2 = \sum_{i=1}^{n} (1 - r_i)$ $\longrightarrow$ $\pi_1 = \frac{N_1}{n}$, $\pi_2 = \frac{N_2}{n}$

$\mu_2 = \frac{1}{N_2} \sum x_i (1 - r_i)$, $\mu_1 = \frac{1}{N_1} \sum_{i=1}^{n} x_i \cdot r_i$

$c_2 = \frac{1}{N_2} \sum_i (1 - r_i)(x_i - \mu_2)^2$, $c_1 = \frac{1}{N_1} \sum_i r_i (x_i - \mu_1)^2$

Gaussian Mixture Model

- input: data $\{x_i\}_{i=1}^{n}$ in $\mathbb{R}^d$
- parameters of a **Gaussian Mixture Model**
  - mixing weights:
    - $\pi_j = \mathbf{P}(\text{cluster membership} = j)$ for $j \in \{1,\ldots,K\}$
  - means:
    - $\mu_j \in \mathbb{R}^d$ for $j \in \{1,\ldots,K\}$
  - covariance matrices:
    - $\mathbf{C}_j \in \mathbb{R}^{d\times d}$ for $j \in \{1,\ldots,K\}$
- we suppose that the given data has been generated from a GMM, and try to find the best GMM parameters (this naturally will define clustering of the training data)
- under the GMM, the i-th sample is drawn as follows
  - first sample a cluster $z_i \in \{1,\ldots,K\}$, from $\pi = [\pi_1, \ldots, \pi_K]$
  - conditioned on this cluster, x_i is sampled from

$$x_i \sim N(\mu_{z_i}, \mathbf{C}_{z_i})$$

Maximum likelihood estimation (MLE)

- we can find the best GMM, by MLE
- for simplicity, suppose $d = 1$ and $K = 2$
- Model parameters are $\pi_1, \pi_2, \mu_1, \mu_2, \mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}$
- the probability of observing a sample x_i can be written as

$$\mathbf{P}(x_i \,|\, \pi_1, \pi_2, \mu_1, \mu_2, \mathbf{C}_1, \mathbf{C}_2) \;=\; \pi_1 \underbrace{\frac{1}{\sqrt{2\pi \mathbf{C}_1}} e^{-\frac{(x_i-\mu_1)^2}{2\mathbf{C}_1}}}_{\triangleq\, N(x_i|\mu_1,\mathbf{C}_1)} + \pi_2 \underbrace{\frac{1}{\sqrt{2\pi \mathbf{C}_2}} e^{-\frac{(x_i-\mu_2)^2}{2\mathbf{C}_2}}}_{\triangleq\, N(x_i|\mu_2,\mathbf{C}_2)}$$

- MLE tries to find

$$\arg \max_{\pi_1,\pi_2,\mu_1,\mu_2,\mathbf{C}_1,\mathbf{C}_2} \sum_{i=1}^{n} \log \mathbf{P}(x_i \,|\, \pi_1, \pi_2, \mu_1, \mu_2, \mathbf{C}_1, \mathbf{C}_2)$$

- however, unlike least squared or logistic regression, this is not a concave function of the parameters (thus hard to find the optimal solution)
- in general, MLE of a mixture model is not convex/concave optimization

exercise: fitting a single Gaussian model

- given $\{x_i\}_{i=1}^n \in \mathbb{R}$,fit the best Gaussian model with mean $\mu \in \mathbb{R}$ and variance $\mathbf{C} \in \mathbb{R}$
- using MLE we want to solve

$$\text{maximize}_{\mu,\mathbf{C}} \;\; \mathscr{L}(\mu, \mathbf{C}) \;=\; \sum_{i=1}^{n} \underbrace{\left(-\frac{(x_i-\mu)^2}{2\mathbf{C}} - \log\left(\sqrt{2\pi\mathbf{C}}\right) \right)}_{\log N(x_i|\mu,\mathbf{C})}$$

- we compute gradient and set it to zero:
  - $$\nabla_\mu \mathscr{L}(\mu, \mathbf{C}) \;=\; \frac{1}{\mathbf{C}} \sum_{i=1}^{n} (\mu - x_i)$$

    which is zero for $\boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i}$

    (which makes sense as it is the empirical mean)
  - $$\nabla_{\mathbf{C}} \mathscr{L}(\mu, \mathbf{C}) \;=\; \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\mathbf{C}^2} - \frac{n}{2\mathbf{C}}$$

    which is zero for $\boxed{\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n} (x_i-\mu)^2}$

    (which makes sense as it is the empirical variance)

MLE for GMM

- we want to fit a model by solving

$$\text{maximize}_{\pi_1,\pi_2,\mu_1,\mu_2,\mathbf{C}_1,\mathbf{C}_2} \sum_{i=1}^{n} \log\Big(\pi_1 \underbrace{\frac{1}{\sqrt{2\pi\mathbf{C}_1}} e^{-\frac{(x_i-\mu_1)^2}{2\mathbf{C}_1}}}_{\triangleq N(x_i|\mu_1,\mathbf{C}_1)} + \pi_2 \underbrace{\frac{1}{\sqrt{2\pi\mathbf{C}_2}} e^{-\frac{(x_i-\mu_2)^2}{2\mathbf{C}_2}}}_{\triangleq N(x_i|\mu_2,\mathbf{C}_2)} \Big)$$

- define $r_i \;=\; \mathbf{P}(z_i = 1 \,|\, x_i) \;=\; \dfrac{\mathbf{P}(z_i = 1, x_i)}{\mathbf{P}(z_i = 1, x_i) + \mathbf{P}(z_i = 2, x_i)}$

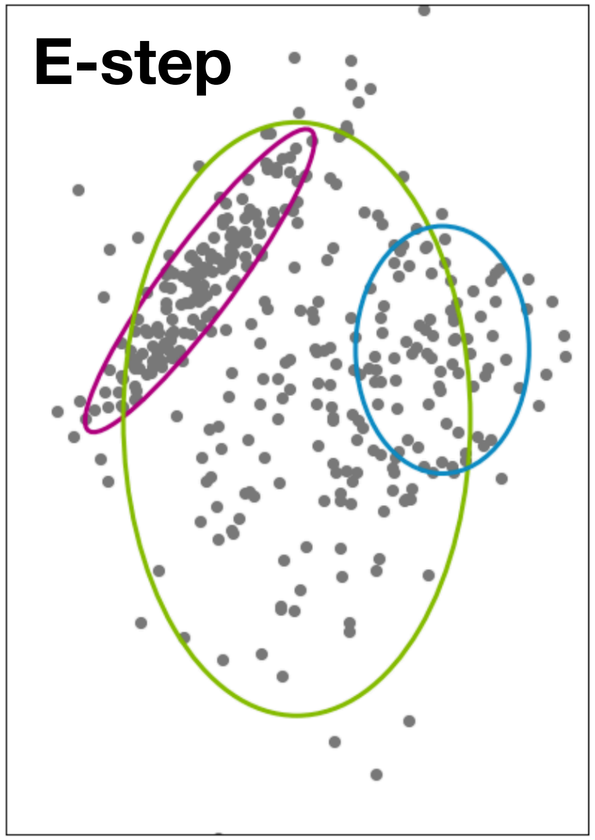
$$= \frac{\pi_1 N(x_i \,|\, \mu_1, \mathbf{C}_1)}{\pi_1 N(x_i \,|\, \mu_1, \mathbf{C}_1) + \pi_2 N(x_i \,|\, \mu_2, \mathbf{C}_2)}$$

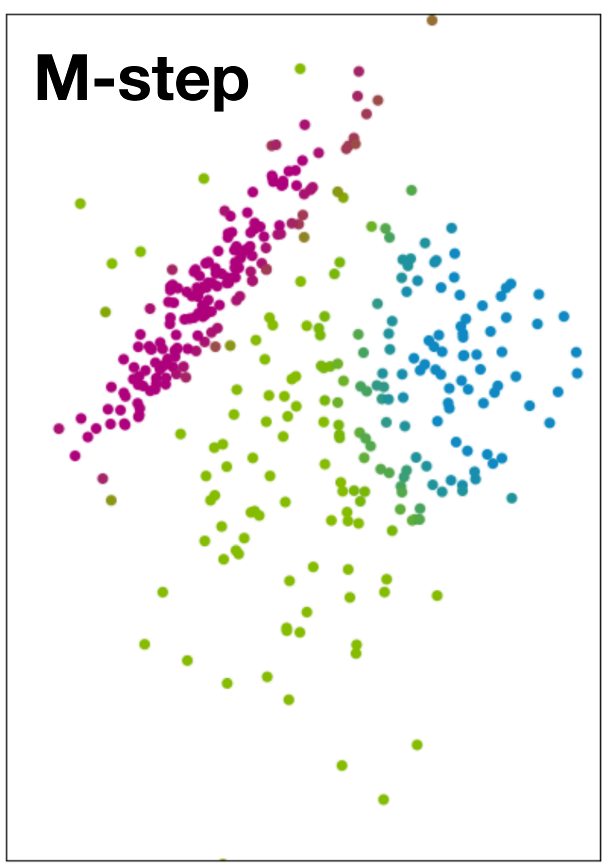
- setting the gradient to zero, we get
  - $\pi_1 = \dfrac{N_1}{n}$ where $N_1 = \displaystyle\sum_{i=1}^{n} r_i$, and $\pi_2 = \dfrac{N_2}{n}$ where $N_2 = \displaystyle\sum_{i=1}^{n} (1 - r_i)$
  - $\mu_1 = \dfrac{1}{N_1} \displaystyle\sum_{i=1}^{n} r_i x_i$ and $\mu_2 = \dfrac{1}{N_2} \displaystyle\sum_{i=1}^{n} (1 - r_i) x_i$
  - $\mathbf{C}_1 = \dfrac{1}{N_1} \displaystyle\sum_{i=1}^{n} r_i (x_i - \mu_1)^2$ and $\mathbf{C}_2 = \dfrac{1}{N_2} \displaystyle\sum_{i=1}^{n} (1 - r_i)(x_i - \mu_2)^2$
- both LHS and RHS depend on the parameters, and no closed form solution exists
- **note that if we know r_i's it is trivial to compute parameters, and vice versa**

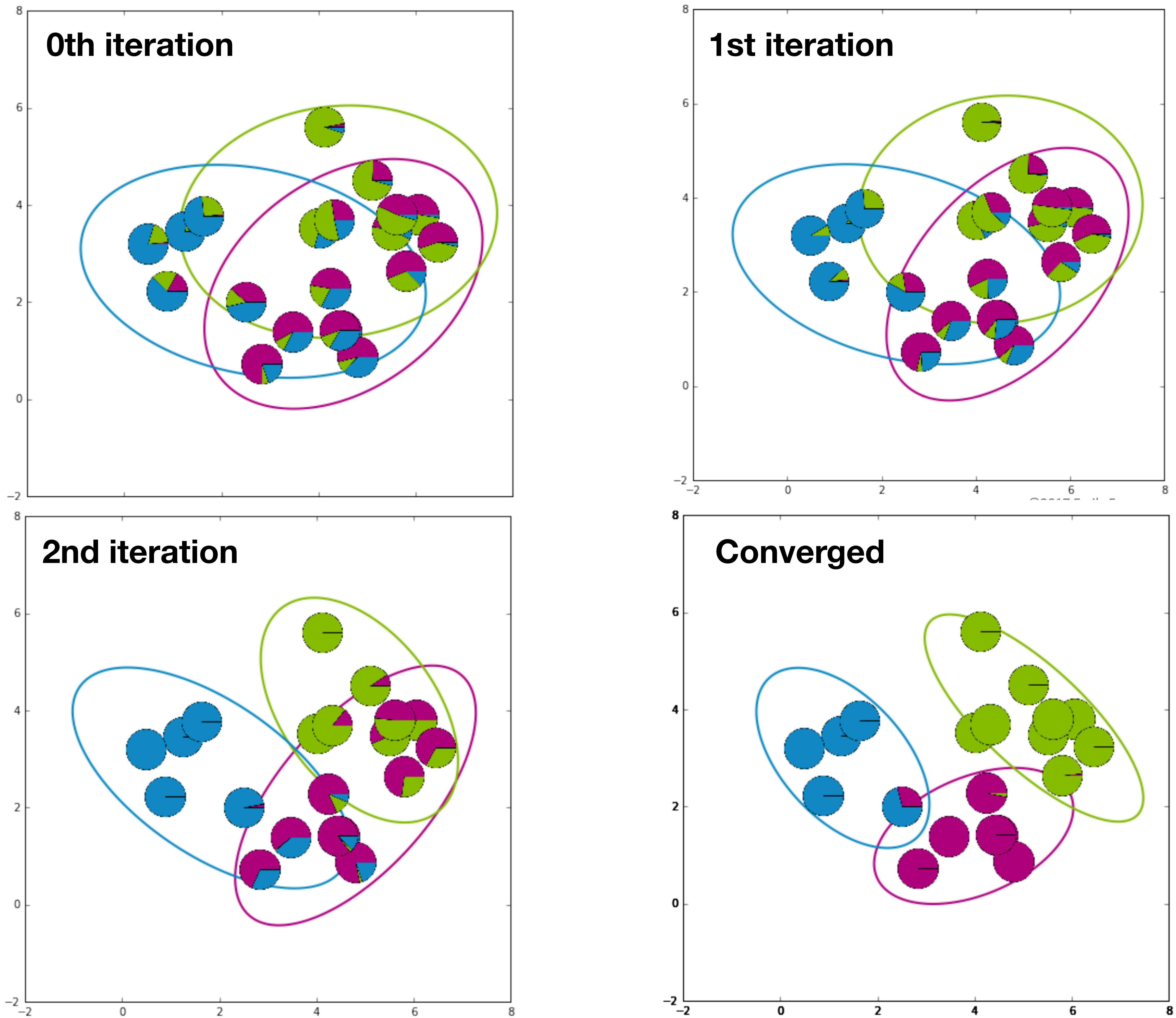
Expectation Maximization (EM) algorithm

- EM is a popular method to solve MLE for mixture models
- input: training data $\{x_i\}_{i=1}^n$
- output: $\pi_1, \pi_2, \mu_1, \mu_2, \mathbf{C}_1, \mathbf{C}_2 \in \mathbb{R}$

- initialization: randomly initialize the parameters
- repeat
  - **E-step** (Expectation): parameters → soft membership
    - $r_i = \dfrac{\pi_1 N(x_i \mid \mu_1, \mathbf{C}_1)}{\pi_1 N(x_i \mid \mu_1, \mathbf{C}_1) + \pi_2 N(x_i \mid \mu_2, \mathbf{C}_2)}$
  - **M-step** (Maximization): soft membership → parameters
    - $\pi_1 = \dfrac{N_1}{n}$ where $N_1 = \sum_{i=1}^n r_i$, and $\pi_2 = \dfrac{N_2}{n}$ where $N_2 = \sum_{i=1}^n (1 - r_i)$
    - $\mu_1 = \dfrac{1}{N_1}\sum_{i=1}^n r_i x_i$ and $\mu_2 = \dfrac{1}{N_2}\sum_{i=1}^n (1 - r_i) x_i$
    - $\mathbf{C}_1 = \dfrac{1}{N_1}\sum_{i=1}^n r_i (x_i - \mu_1)^2$ and $\mathbf{C}_2 = \dfrac{1}{N_2}\sum_{i=1}^n (1 - r_i)(x_i - \mu_2)^2$







0th iteration
1st iteration
2nd iteration
Converged

For general number of clusters K and dimension d

- we can derive EM for general case, in an analogous way
- Initialize parameters: $\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \mathbf{C}_1, \ldots, \mathbf{C}_K$

- **E-step**:
  - For k=1,…,K

$$r_{i,k} = \frac{\pi_k \, N(x_i \,|\, \mu_k, \mathbf{C}_k)}{\sum_{j=1}^{K} \pi_j \, N(x_i \,|\, \mu_j, \mathbf{C}_j)}$$

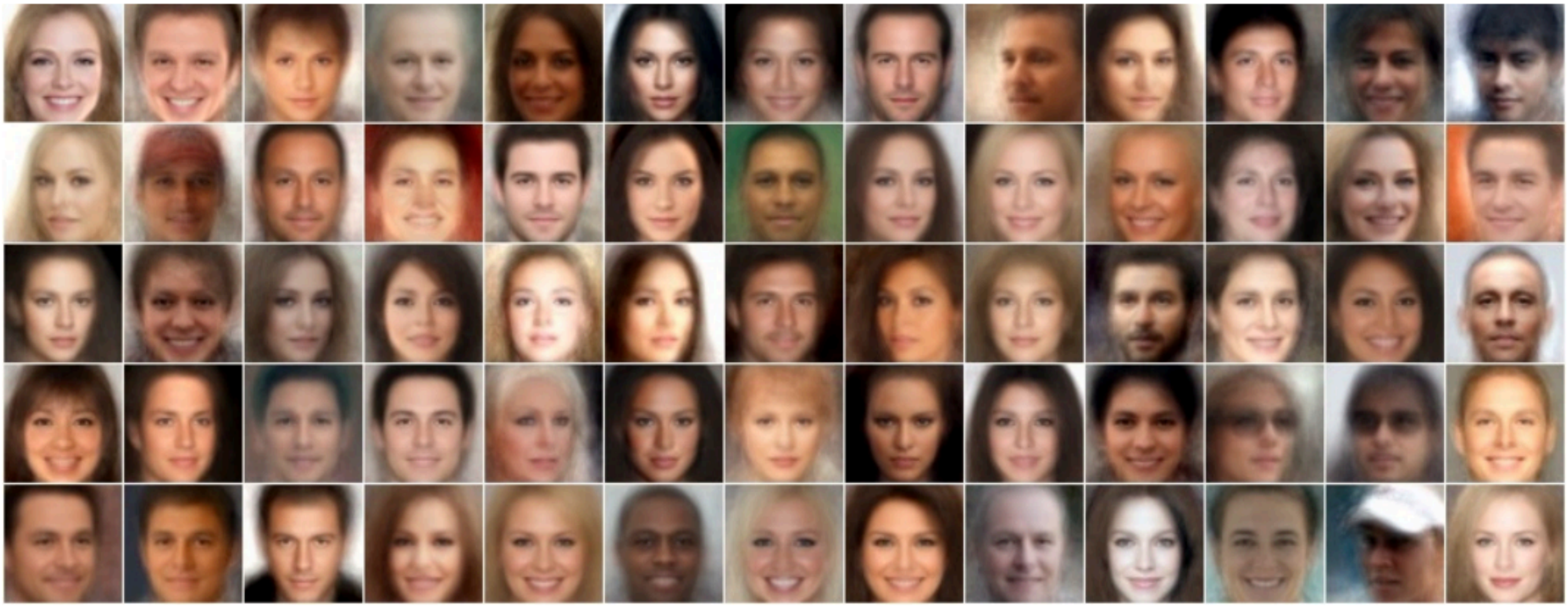
- **M-step**:
  - For k=1,..,K

$$\pi_k = \frac{N_k}{n} \quad \text{where} \quad N_k = \frac{\sum_{i=1}^{n} r_{i,k}}{n}$$

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{n} r_{i,k} x_i \quad \text{and} \quad \mathbf{C}_k = \frac{1}{N_k} \sum_{i=1}^{n} r_{i,k}(x_i - \mu_k)(x_i - \mu_k)^T$$

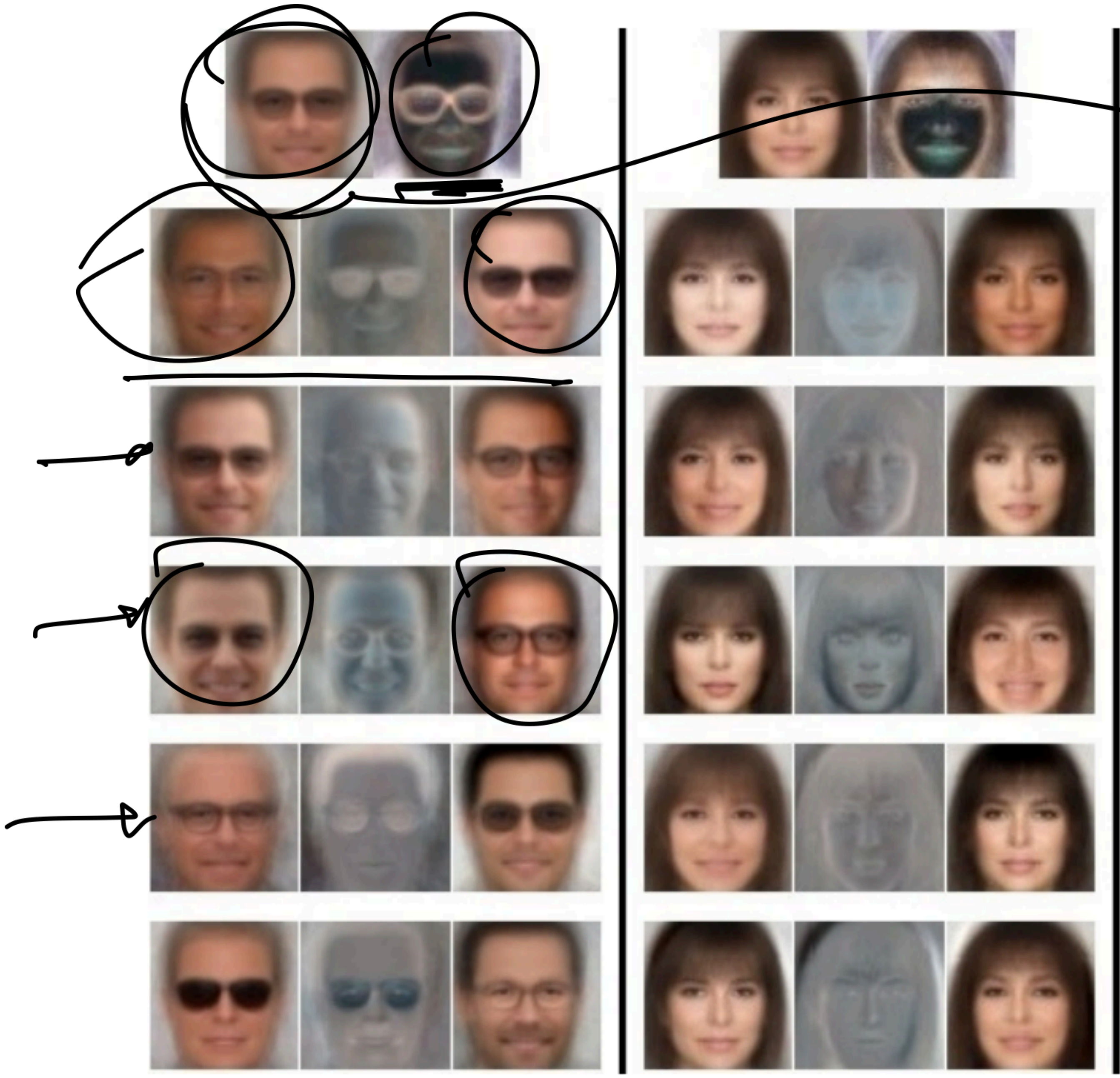
- **once GMM is learned, clustering is straight forward: cluster according to the $r_{i,k}$'s**

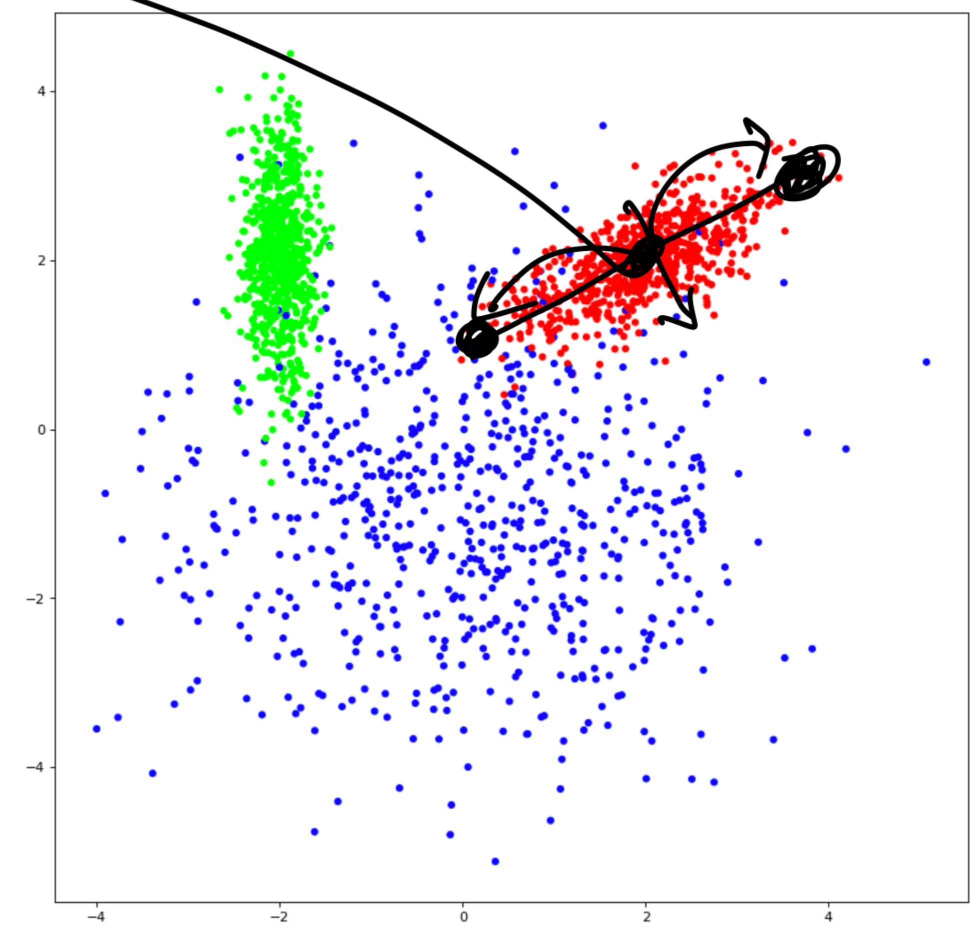
GMM for real data



- these are generated samples, from GMM trained on CelebA dataset
- image: 64*64*3=288 dimension
- covariance: restricted to rank-10 matrices
- mixture: K=1,000

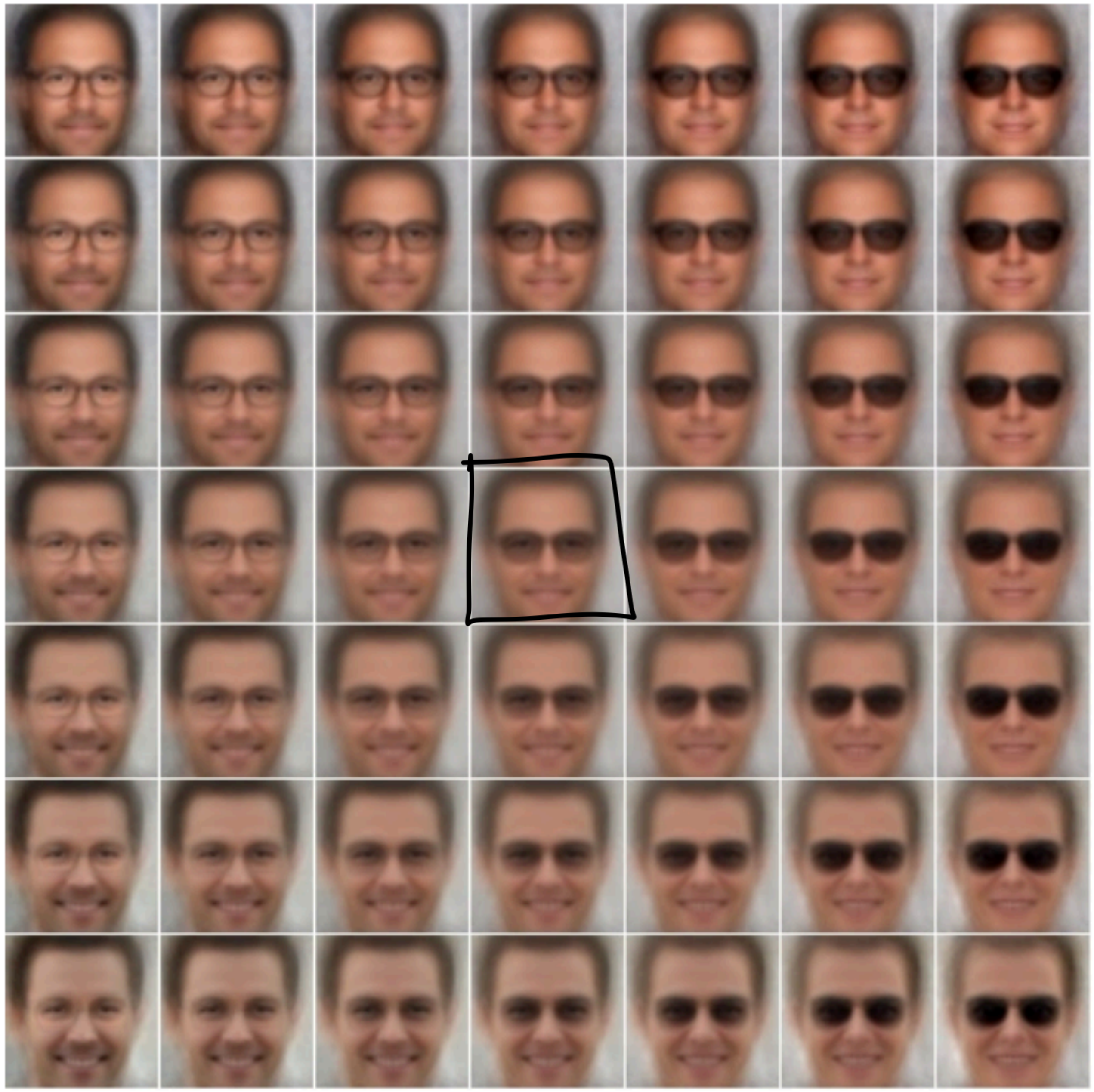
Images from "on GANs and GMMs", 2018, Richardson &Weiss

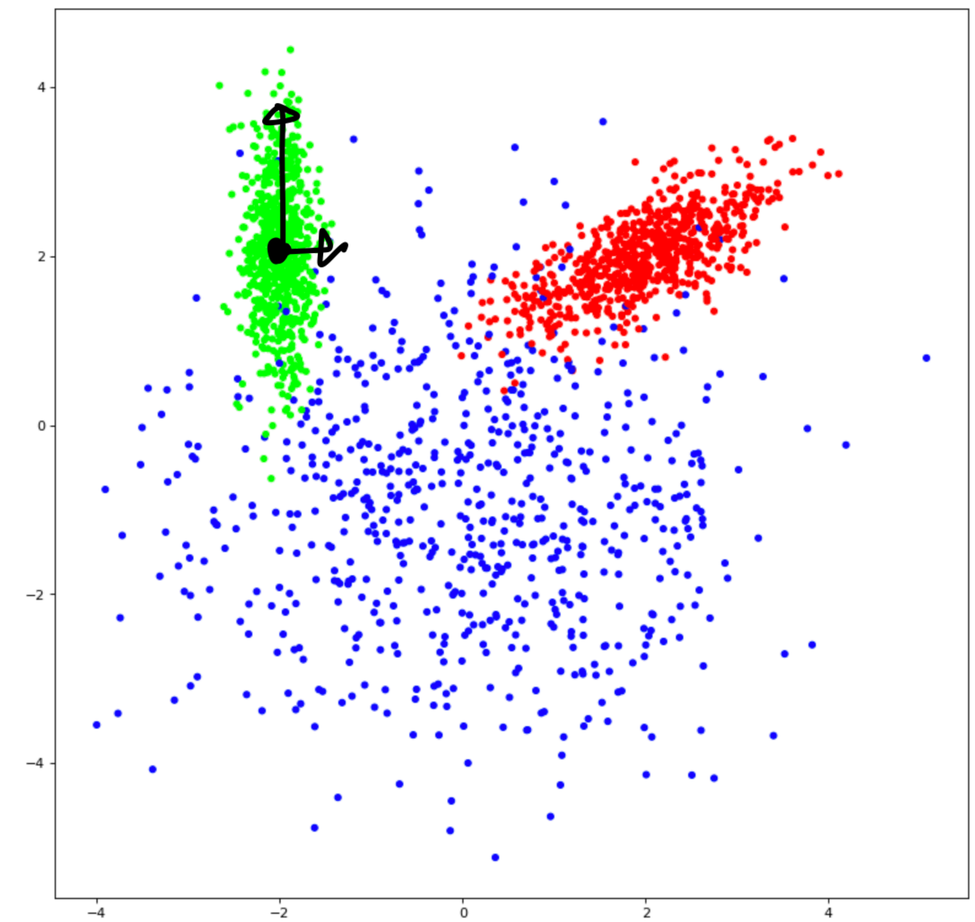




- top: center of a cluster μ_k and
  the diagonal entries of the covariance matrix $\mathbf{C}_k$
- note that we have trained 10-dimensional covariance matrix $\mathbf{C}_k = AA^T$, with $A \in \mathbb{R}^{288 \times 10}$, and let $A^{(j)}$ be the j-th column
- bottom: each row corresponds to different j, and we show
  $\mu_k + A^{(j)}, 0.5 + A^{(j)}, \mu_k - A^{(j)}$

Images from “on GANs and GMMs”, 2018, Richardson &Weiss





- middel: μ_k
- Each row: middel + $c \times A^{(1)}$
- Each column: middle + $c \times A^{(2)}$

Images from "on GANs and GMMs", 2018, Richardson &Weiss

Mixture model for documents

- Input: n documents $\{x_i\}_{i=1}^n$
- Each document is a sequence of words of length T
  $x_i = (w_1, w_2, \ldots, w_T)$
- Bag-of-words model:
  - parameters:
    - mixing weights: $\pi_k = \mathbf{P}(\text{topic} = k)$ for $k \in \{1,\ldots,K\}$
    - word probability: $b_{wk} = \mathbf{P}(\text{word} = w \,|\, \text{topic} = k)$
  - the generative model
    - first sample topic from $\pi = (\pi_1, \ldots, \pi_K)$
    - next sample T words i.i.d. from $b_k = (b_{w_1 k}, \ldots, b_{w_{200,000} k})$
  - to make the problem tractable, this completely ignores the order of the words in the document (but still very successful in document clustering)

$$\mathbf{P}(\text{topic } z_i = k, x_i = (w_1, \ldots, w_T)) = \pi_k b_{w_1 k} \cdots b_{w_T k}$$

Topic modeling

- to fit a topic model, we solve the following

$$\text{maximize}_{b \in \mathbb{R}^{K \times T}, \pi \in \mathbb{R}^K} \sum_{i=1}^{n} \log \mathbf{P}(x_i \mid b, \pi)$$

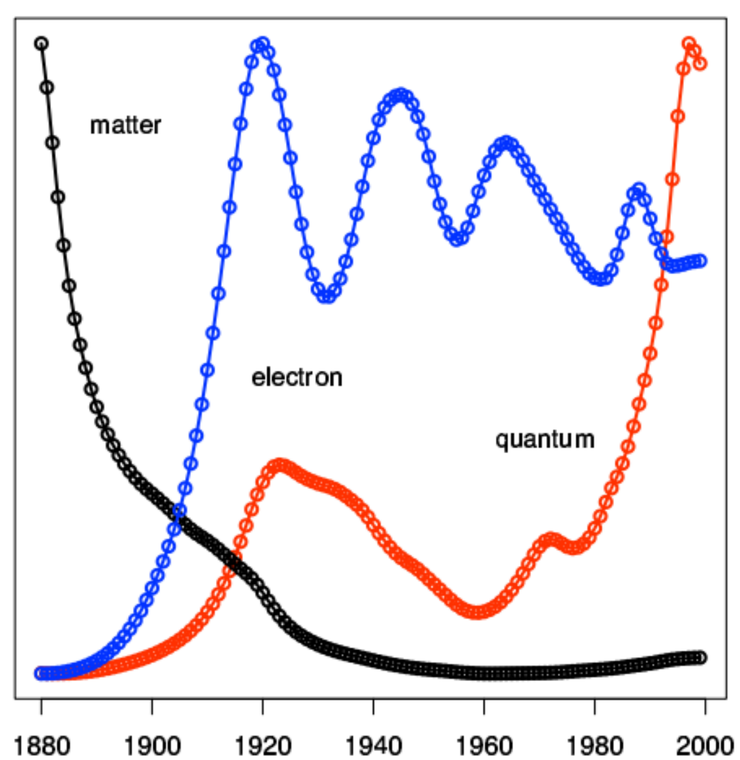
- we can apply EM algorithm
- initialize b, π
- **E-step**: parameters → soft assignments
   - $r_{ik} \;=\; \mathbf{P}(\text{topic } z_i = k \mid x_i) = \dfrac{\pi_k b_{w_1 k} \cdots b_{w_T k}}{\sum_{k'=1}^{K} \pi_{k'} b_{w_1 k'} \cdots b_{w_T k'}}$

- **M-step**: soft assignments → parameters
   - $\pi_k \;=\; \dfrac{N_k}{n}$ where $N_k \;=\; \displaystyle\sum_{i=1}^{n} r_{ik}$
   - $b_{wk} \;=\; \dfrac{1}{N_k} \displaystyle\sum_{i=1}^{n} r_{ik} \dfrac{\text{Count}(w \text{ in } x_i)}{T}$

Dynamic topic modeling (over time)

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
force	motion	magnet	force	atom	ray	energy	energy	radiat	electron	electron	electron	state
energy	force	electric	magnet	theory	measure	measure	radiat	energy	energy	energy	atom	energy
motion	magnet	measure	theory	electron	energy	electron	ray	electron	atom	particle	energy	electron
differ	energy	force	electric	energy	theory	light	electron	measure	measure	field	structur	magnet
light	measure	theory	atom	measure	light	atom	measure	ray	radiat	radiat	field	field
measure	differ	system	system	ray	wave	particle	atom	atom	field	model	model	atom
magnet	direct	motion	measure	electr	radiat	ray	particle	field	ray	atom	state	system
direct	line	line	line	line	atom	radiat	two	two	model	two	two	two
matter	result	point	energy	force	electric	point	light	particle	particle	ray	magnet	quantum
result	light	differ	body	value	value	theory	absorpt	observe	magnet	measure	ray	physic

"Atomic Physics"



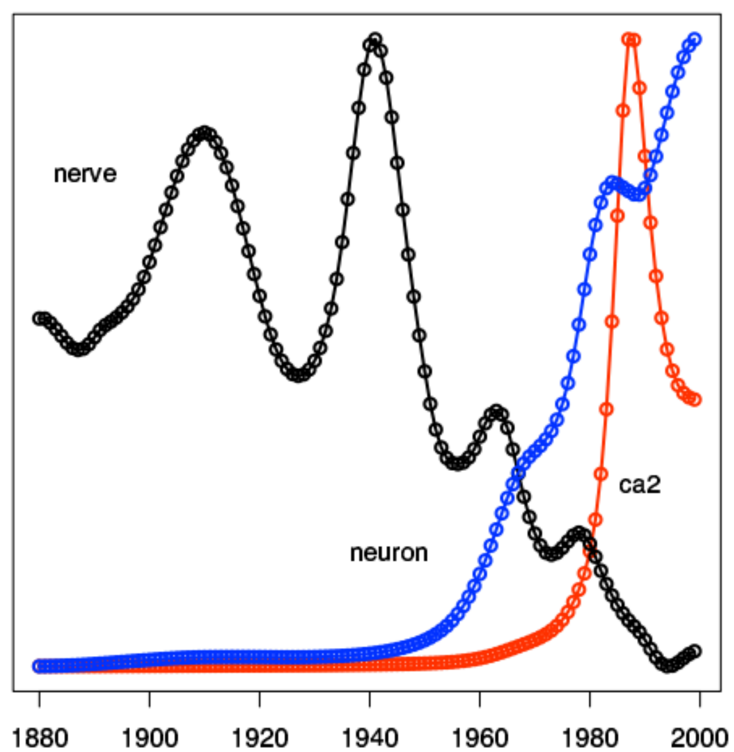
1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball''
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

From "Dynamic Topic Models" Blei & Lafferty 2006

Dynamic topic modeling (over time)

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
brain	movement	brain	movement	movement	stimulate	record	respons	response	respons	cell	cell	neuron
movement	eye	eye	brain	sound	muscle	nerve	record	stimulate	cell	neuron	channel	active
action	right	movement	sound	muscle	sound	stimulate	stimulate	record	potential	response	neuron	brain
right	hand	right	nerve	active	movement	response	nerve	condition	stimul	active	ca2	cell
eye	brain	left	active	nerve	response	muscle	muscle	active	neuron	brain	active	fig
hand	left	hand	muscle	stimulate	nerve	electrode	active	potential	active	stimul	brain	response
left	action	nerve	left	fiber	frequency	active	frequency	stimulus	nerve	muscle	receptor	channel
muscle	muscle	vision	eye	reaction	fiber	brain	electrode	nerve	eye	system	muscle	receptor
nerve	sound	sound	right	brain	active	fiber	potential	subject	record	nerve	respons	synapse
sound	experiment	muscle	nervous	response	brain	potential	study	eye	abstract	receptor	current	signal

"Neuroscience"



1887 Mental Science
1900 Hemianopsia in Migraine
1912 A Defence of the ``New Phrenology"
1921 The Synchronal Flashing of Fireflies
1932 Myoesthesis and Imageless Thought
1943 Acetylcholine and the Physiology of the Nervous System
1952 Brain Waves and Unit Discharge in Cerebral Cortex
1963 Errorless Discrimination Learning in the Pigeon
1974 Temporal Summation of Light by a Vertebrate Visual Receptor
1983 Hysteresis in the Force-Calcium Relation in Muscle
1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

From “Dynamic Topic Models” Blei & Lafferty 2006

General Expectation Maximization

- consider fitting a (general) mixture distribution
  - training data: $\{x_1, \ldots, x_n\}$ (or it could be $\{(x_1, y_1), \ldots, (x_n, y_n)\}$)
  - suppose each sample is drawn i.i.d. from a distribution that a cluster z_i for the sample x_i is first drawn with probability $\pi = \{\pi_1, \ldots, \pi_k\}$ and then the sample x_i is drawn according to its cluster membership with
    $$p(x_i, z_i = k; w = \{w_1, \ldots, w_K\}, \pi = \{\pi_1, \ldots, \pi_K\})$$
    and we only observe x_i's and not z_i's
  - to maximize the log-likelihood given by
    $$\ell(w, \pi) \;=\; \sum_{i=1}^{n} \log\Big(\underbrace{\sum_{k=1}^{K} p(x_i, z_i = k; w, \pi)}_{p(x;w,\pi)} \Big)$$

General Expectation Maximization

- Randomly initialize $w^{(0)} = \{w_1^{(0)}, \ldots, w_K^{(0)}\}, \pi^{(0)} = \{\pi_1^{(0)}, \ldots, \pi_K^{(0)}\}$
- Repeat for t=1,…,T
  - E-step: given w, π, find r_{ik}'s

$$r_{ik} = \mathbb{P}(z_i = k \,|\, x_i; w^{(t-1)}, \pi^{(t-1)})$$

$$= \frac{\mathbb{P}(z_i = k, x_i; w^{(t-1)}, \pi^{(t-1)})}{\mathbb{P}(x_i; w^{(t-1)}, \pi^{(t-1)})}$$

$$= \frac{\mathbb{P}(z_i = k, x_i; w^{(t-1)}, \pi^{(t-1)})}{\sum_{k'=1}^{K} \mathbb{P}(z_i = k', x_i; w^{(t-1)}, \pi^{(t-1)})}$$

  - M-step: given r_{ik}'s find $w^{(t)}, \pi^{(t)}$

$$\pi_k^{(t)} = \frac{1}{n} \sum_{i=1}^{n} r_{ik} \quad \text{for } k \in \{1, \ldots, K\}$$

$$w_k^{(t)} = \arg\max_{w_k} \sum_{i=1}^{n} r_{ik} \log \mathbb{P}(x_i \,|\, z_i = k; w_k) \quad \text{for } k \in \{1, \ldots, K\}$$