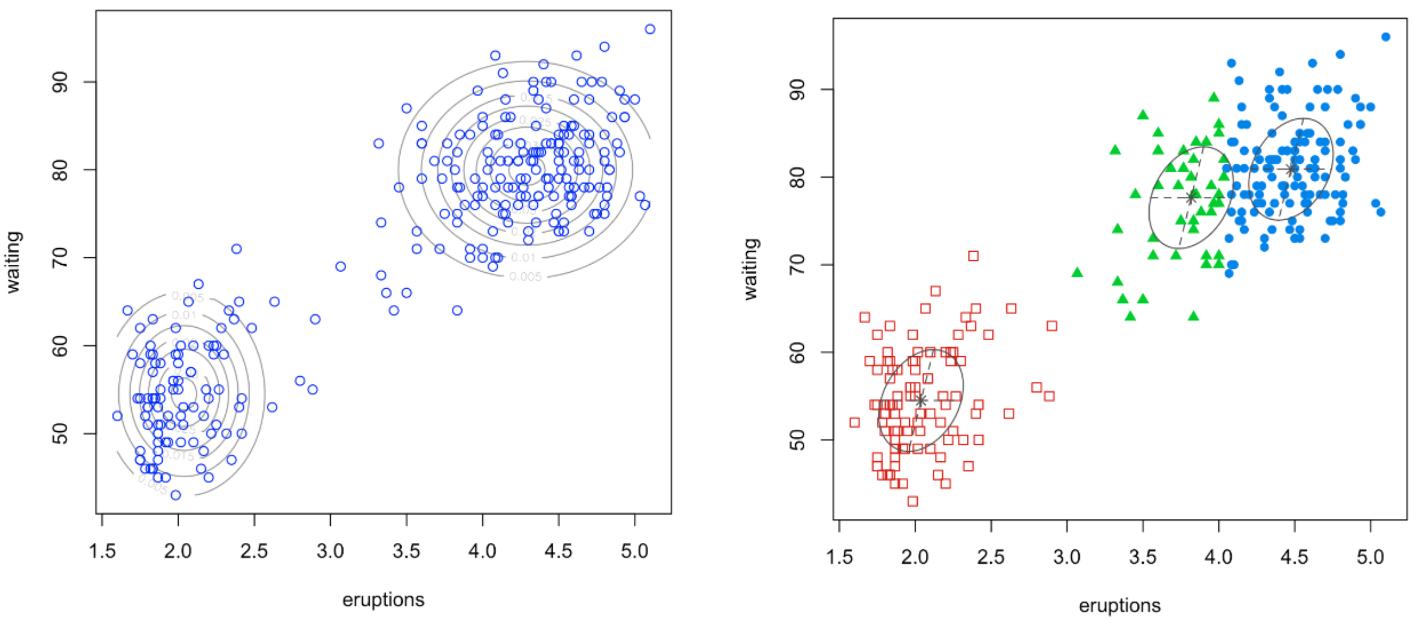
Model-based Clustering

Model-based clustering refers to clustering a set of data points $(x_1, \ldots, x_n)$ by fitting a mixture model on this data set, where each cluster corresponds to a component of the mixture model.



Mixture Models

- Consider a mixture model with K components, whose pdf is given by

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x \mid \theta_k),$$

where the mixing weight π_k is between 0 and 1 and $\sum_k \pi_k = 1$, and $f_k(\cdot \mid \theta_k)$ is a pdf with parameter θ_k.

Scenario 2: the two-dimensional data $X \in \mathbf{R}^2$ in each class is generated from a mixture of 10 different bivariate Gaussian distributions with uncorrelated components and different means, i.e.,

$$X|Y = k, Z = l \sim \mathcal{N}(\mathbf{m}_{kl}, s^2\mathbf{I}_2),$$

where $k = 0, 1$, $l = 1 : 10$, $P(Y = k) = 1/2$, and $P(Z = 1) = 1/10$. In other words, given $Y = k$, X follows a mixture distribution with density function

$$\frac{1}{10}\sum_{l=1}^{10}\left(\frac{1}{\sqrt{2\pi s^2}}\right)^2 e^{-\|\mathbf{x}-\mathbf{m}_{kl}\|^2/(2s^2)}.$$

- A random sample from the mixture model above can be generated by the following two steps:

  1. Generate Z from a multinomial distribution with $P(Z = k) = \pi_k$ and $k = 1, 2, \ldots, K$.
  2. Conditioning on $Z = k$, generate X from f_k, the k-th component.

A Two Components Gaussian Mixture

Consider a simple case where $K = 2$, $x_i \in \mathbb{R}$, and each component is a Gaussian distribution with mean μ_k and variance σ_k^2, i.e., a one-dimensional two-component Gaussian mixture model. The pdf is given by

$$p(x|\theta) = \pi\phi_{\mu_1,\sigma_1^2}(x) + (1-\pi)\phi_{\mu_2,\sigma_2^2}(x). \tag{1}$$

where

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

and $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \pi)$ denotes all parameters of this mixture model.

Given n training samples $\mathbf{x} = (x_1, \ldots, x_n)$, the log-likelihood is

$$\log p(\mathbf{x}|\theta) = \sum_{i=1}^{n} \log \Big[\pi \phi_{\mu_1, \sigma_1^2}(x_i) + (1-\pi)\phi_{\mu_2, \sigma_2^2}(x_i)\Big]. \tag{2}$$

The MLE of the parameter $\theta = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ is defined to be

$$\hat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta} \log p(\mathbf{x}|\theta),$$

which is not easy to compute. Why? Log-likelihood of a single normal pdf takes a derivative friendly form,

$$\log \phi_{\mu, \sigma^2}(x) = -\frac{1}{2}\log \sigma^2 - \frac{(x-\mu)^2}{2\sigma^2} + \text{const.},$$

but log-likelihood of a weighted summation of normal pdfs does not.

The calculation is much easier if we *knew* which component x_i belongs to.

Introduce the latent variable $Z_i = 1$ or 2.

$$\begin{aligned} Z_i &\sim \mathsf{Bern}(\pi) \\ X_i \mid Z_i = k &\sim \mathsf{N}(\mu_k, \sigma_k^2). \end{aligned}$$

The likelihood of the *full data* $(\mathbf{x}, \mathbf{z})$ is given by

$$\prod_{i=1}^{n} \left[\pi \phi_{\mu_1, \sigma_1^2}(x_i)\right]^{\{z_i=1\}} \left[(1-\pi)\phi_{\mu_2, \sigma_2^2}(x_i)\right]^{\{z_i=2\}}.$$

The log-likelihood is given by

$$\begin{aligned} &\sum_i \mathbf{1}_{\{z_i=1\}} \left[\log \phi_{\mu_1, \sigma_1^2}(x_i) + \log \pi\right] + \mathbf{1}_{\{z_i=2\}} \left[\log \phi_{\mu_2, \sigma_2^2}(x_i) + \log(1-\pi)\right] \\ = &\sum_{i:z_i=1} \left[\log \phi_{\mu_1, \sigma_1^2}(x_i) + \log \pi\right] + \sum_{i:z_i=2} \left[\log \phi_{\mu_2, \sigma_2^2}(x_i) + \log(1-\pi)\right] \end{aligned}$$

The MLE for $\theta = (\mu_{1:2}, \sigma^2_{1:2}, \pi)$ is given by

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i:z_i=1} x_i, \quad \hat{\sigma}^2_1 = \frac{1}{n_1} \sum_{i:z_i=1} (x_i - \hat{\mu}_1)^2,$$

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_{i:z_i=2} x_i, \quad \hat{\sigma}^2_2 = \frac{1}{n_2} \sum_{i:z_i=2} (x_i - \hat{\mu}_2)^2,$$

and $\hat{\pi} = n_1/n$. Why the MLE of π is n_1/n?

$$\begin{aligned} n_1 \log \pi + (n - n_1) \log(1-\pi) &\propto \frac{n_1}{n} \log \pi + \left(1 - \frac{n_1}{n}\right) \log(1-\pi) \\ &= \frac{n_1}{n} \log \frac{\pi}{n_1/n} + (1 - \frac{n_1}{n}) \log \frac{(1-\pi)}{1 - n_1/n} + C \end{aligned}$$

where C is a constant not depending on π and the sum is the negative KL distance between two distributions, which is non-positive and is zero only if $\pi = n_1/n$.

Kullback-Leibler Distance

The KL distance between two distributions, $p(\cdot)$ and $q(\cdot)$, is defined to be

$$\int p(x) \log \frac{p(x)}{q(x)} dx, \qquad \text{or} \qquad \sum_{j=1}^{m} p_j \log \frac{p_j}{q_j}$$

for continuous and discrete cases, respectively. Note that KL distance is not symmetric.

Using Jensen's inequality, we can show that

$$KL(p\|q) = \mathbb{E}_{p(X)} \log \frac{p(X)}{q(X)} = \mathbb{E}_{p(X)} \left[-\log \frac{q(X)}{p(X)} \right] \geq -\log \left(\mathbb{E}_{p(X)} \frac{q(X)}{p(X)} \right) = 0.$$

So $KL(p\|q) \geq 0$ and $= 0$ iff p and q are the same distribution (up to a measure zero set).

However, we do not observe z_i's. Consider the following iterative scheme: start with some initial guess of θ, then

a) calculate the corresponding distribution of Z_i:

$$
\begin{aligned}
P(Z_i = 1 \mid x_i, \theta) &= \gamma_i = \frac{\pi \phi_{\mu_1, \sigma_1^2}(x_i)}{\pi \phi_{\mu_1, \sigma_1^2}(x_i) + (1-\pi)\phi_{\mu_2, \sigma_2^2}(x_i)}, \\
P(Z_i = 2 \mid x_i, \theta) &= 1 - \gamma_i.
\end{aligned}
$$

b) Now, for each point x_i, instead of allocating it to component 1 or 2, we count its γ_i fraction to component 1 and $(1-\gamma_i)$ fraction to component 2, and update $\theta = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ as follows

$$\hat{\mu}_1 = \frac{1}{\gamma_+}\sum_i \gamma_i x_i, \quad \hat{\sigma}_1^2 = \frac{1}{\gamma_+}\sum_i \gamma_i (x_i - \hat{\mu}_1)^2,$$

$$\hat{\mu}_2 = \frac{1}{n-\gamma_+}\sum_i (1-\gamma_i) x_i, \quad \hat{\sigma}_2^2 = \frac{1}{n-\gamma_+}\sum_i (1-\gamma_i)(x_i - \hat{\mu}_2)^2,$$

$$\hat{\pi} = \gamma_+/n$$

We can iterative the two steps until the value of θ gets stabilized. Is the returned value of θ the MLE that maximizes the marginal likelihood $p(\mathbf{x}|\theta)$?

The EM Algorithm

The Expectation-Maximization (EM) algorithm is an iterative method that finds the MLE by enlarging the sample with unobserved latent data.

Suppose our observed data is $\mathbf{x}$ with log-likelihood $\log p(\mathbf{x}|\theta)$ that depends on unknown parameter θ. Using latent variable $\mathbf{z}$, the log-likelihood can be written as

$$\log p(\mathbf{x}|\theta) = \log \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}|\theta) = \log \sum_{\mathbf{z}} p(\mathbf{z}|\theta) p(\mathbf{x}|\mathbf{z}, \theta). \tag{3}$$

Direct maximization of (3) is quite difficult due to the sum inside the logarithm.

In the EM algorithm, we pretend we knew $\mathbf{Z}$, then we can maximize log of the joint likelihood

$$\log p(\mathbf{x}, \mathbf{Z}|\theta) = \log p(\mathbf{Z}|\theta) + \log p(\mathbf{x}|\mathbf{Z}, \theta).$$

Each iteration of the EM algorithm involves two steps, the E-step and the M-step.

- E-step: Let θ_0 denote the current value of θ. Find $p(\mathbf{Z}|\mathbf{x}, \theta_0)$, the distribution of the latent variable $\mathbf{Z}$ given the data $\mathbf{x}$ and θ_0, and then calculate the following expectation

$$g(\theta) = \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log p(\mathbf{x}, \mathbf{Z}|\theta)$$

which is

$$\sum_{\mathbf{z}} p(\mathbf{Z} = \mathbf{z}|\mathbf{x}, \theta_0) \log p(\mathbf{x}, \mathbf{z}|\theta), \quad \text{or} \int p(\mathbf{z}|\mathbf{x}, \theta_0) \log p(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z}.$$

- M-step: Find θ_1 that maximizes $g(\theta)$.

- Replace θ_0 by θ_1 and repeat the above E and M steps until convergence.

Next we show that

$$g(\theta_1) \geq g(\theta_0) \implies p(\mathbf{x}|\theta_1) \geq p(\mathbf{x}|\theta_0),$$

that is, each iteration of the EM algorithm increases (or at least doesn't decrease) the marginal likelihood $p(\mathbf{x}|\theta)$. Recall $g(\theta) = \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log p(\mathbf{x}, \mathbf{Z}|\theta)$.

$$\begin{aligned} g(\theta_1) - g(\theta_0) &= \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log \frac{p(\mathbf{x}, \mathbf{Z}|\theta_1)}{p(\mathbf{x}, \mathbf{Z}|\theta_0)} = \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log \frac{p(\mathbf{x}|\theta_1)p(\mathbf{Z}|\mathbf{x}, \theta_1)}{p(\mathbf{x}|\theta_0)p(\mathbf{Z}|\mathbf{x}, \theta_0)} \\ &= \log \frac{p(\mathbf{x}|\theta_1)}{p(\mathbf{x}|\theta_0)} - \mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log \frac{p(\mathbf{Z}|\mathbf{x}, \theta_0)}{p(\mathbf{Z}|\mathbf{x}, \theta_1)} \end{aligned}$$

where the 2nd term is the Kullback-Leibler distance between two distributions which is always non-negative. So

$$\log \frac{p(\mathbf{x}|\theta_1)}{p(\mathbf{x}|\theta_0)} = \underbrace{g(\theta_1) - g(\theta_0)}_{\geq 0} + \underbrace{\mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta_0} \log \frac{p(\mathbf{Z}|\mathbf{x}, \theta_0)}{p(\mathbf{Z}|\mathbf{x}, \theta_1)}}_{\geq 0}.$$

An Alternative View of EM

The EM algorithm is essentially an MM algorithm (Neal and Hinton, 1998).

Consider the following objective function

$$F(q,\theta) = \mathbb{E}_{q(\mathbf{Z})} \log \frac{p(\mathbf{x},\mathbf{Z}|\theta)}{q(\mathbf{Z})}, \tag{4}$$

where q denotes any pdf/pmf of $\mathbf{Z}$ and $\mathbb{E}_{q(\mathbf{Z})}$ denotes an expectation of $\mathbf{Z}$ taken with respect of q. The objective function (4) can be re-expressed as

$$F(q,\theta) = \mathbb{E}_{q(\mathbf{Z})} \log \frac{p(\mathbf{x}|\theta)p(\mathbf{Z}|\mathbf{x},\theta)}{q(\mathbf{Z})} = \log p(\mathbf{x}|\theta) - \mathbb{E}_{q(\mathbf{Z})} \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{x},\theta)}.$$

So $\max_{q,\theta} F(q,\theta)$ is achieved by setting $\theta = \hat{\theta}_{\mathsf{mle}}$ and q to be $p(\mathbf{z}|\mathbf{x},\hat{\theta}_{\mathsf{mle}})$. In other words, we can obtain $\hat{\theta}_{\mathsf{mle}}$ as a byproduct of maximizing $F(q,\theta)$.

Next we show that EM can be viewed as a coordinate descent algorithm on $F(q, \theta)$: at the t-th iteration,

- E-step: $q^{t+1} = \arg\max_q F(q, \theta^t) = p(\mathbf{z}|\mathbf{x}, \theta^t)$;
- M-step: $\theta^{t+1} = \arg\max_\theta F(q^{t+1}, \theta) = \arg\max_\theta \Big[\mathbb{E}_{\mathbf{Z}|\mathbf{x},\theta^t} \log p(\mathbf{x}, \mathbf{Z}|\theta)\Big]$.

The alternative view of EM

- provides a justification for some variants of EM algorithms such as generalized EM (GEM) where only partial implementation of the E or M steps is performed
- can handle cases where we have some special constraints on the latent variable (Graca et al, 2007)
- motivates variational EM algorithms

Variational EM

- Given θ, the optimal choice for q is $p(\mathbf{z}|\mathbf{x},\theta)$, which maximizes

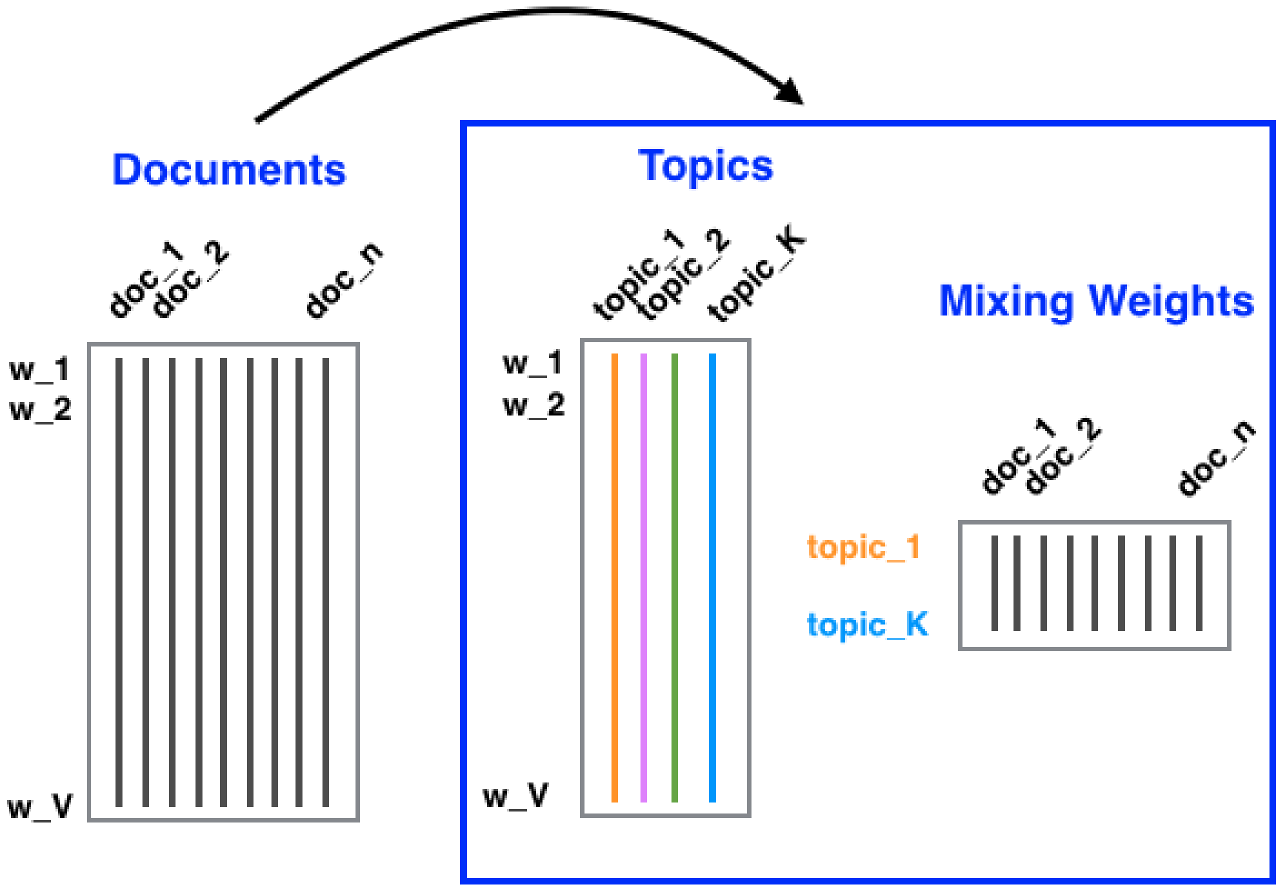
$$F(q,\theta) = \mathbb{E}_{q(\mathbf{Z})} \log \frac{p(\mathbf{x}|\theta)p(\mathbf{Z}|\mathbf{x},\theta)}{q(\mathbf{Z})}.$$

But $p(\mathbf{z}|\mathbf{x},\theta)$ may not be easy to obtain and approximation is needed for tractable computation.

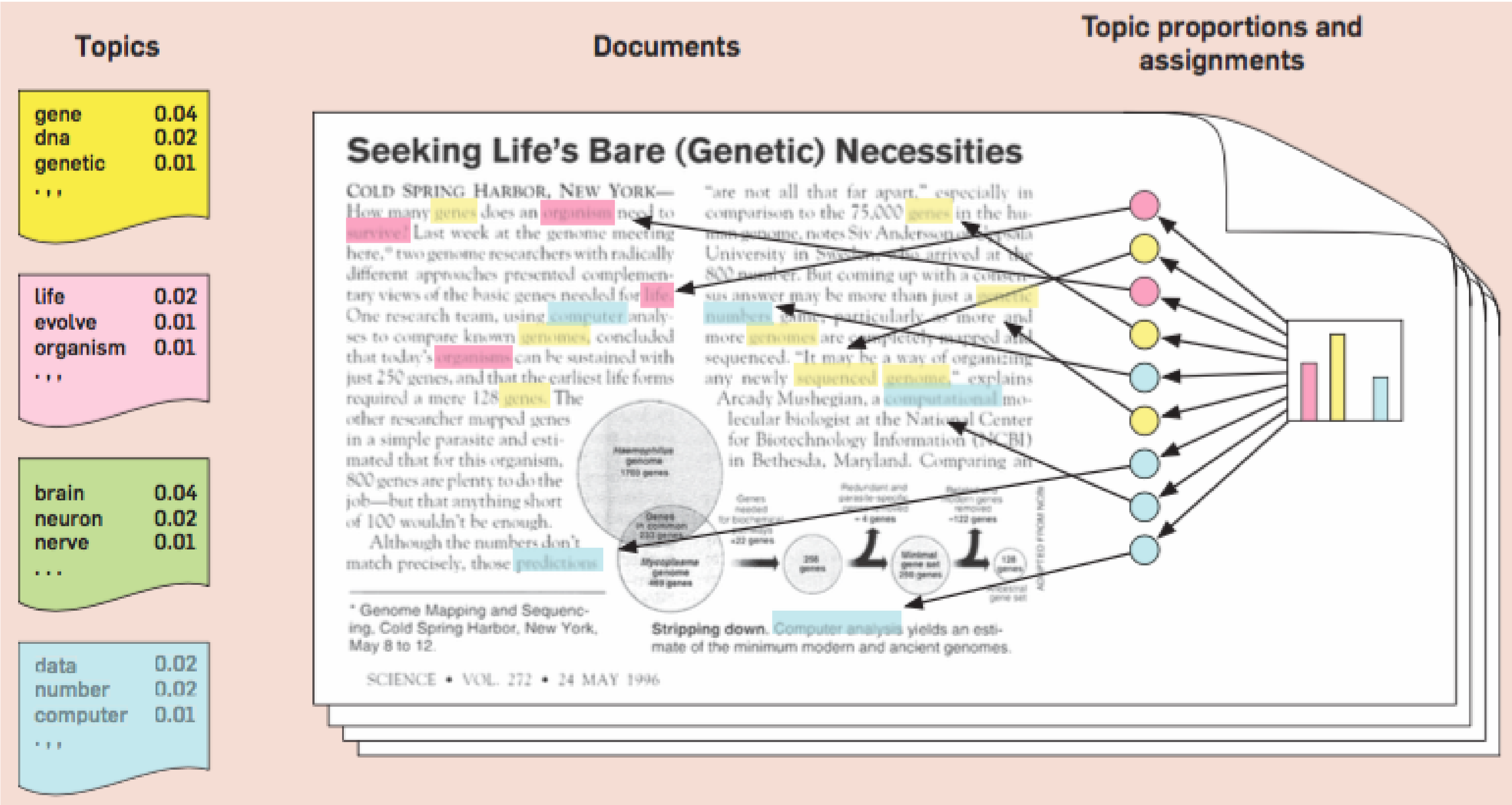
- For example, we can optimize $F(q,\theta)$ subject to constraint that $q(\mathbf{z})$ can be factorized as $\prod_{i=1}^n q_i(z_i)$. Then we can apply coordinate descent over $(\theta, q_1, \ldots, q_n)$ to maximize

$$F(\theta, q_1, \cdots, q_n) = \mathbb{E}_{q_1,\ldots,q_n} \log \frac{p(\mathbf{x}|\theta)p(\mathbf{Z}|\mathbf{x},\theta)}{q_1(Z_1)\cdots q_n(Z_n)}.$$

Latent Dirichlet Allocation (LDA)



Each topic is a distribution over words
Each word is a draw from a topic
Each document is a mixture of topics



Source: Blei (2012) "Probabilistic topic models"