AI – Machine Learning

Artificial Intelligence
Research Group

Maximum likelihood

- The fitting (learning) of models has been achieved by minimizing a ***sum of squares*** for ***regression***, or by minimizing ***cross-entropy*** for ***classification***.
- In fact, both of these minimizations are instances of the ***maximum likelihood*** approach to fitting.

Maximum likelihood is based on the likelihood function:

$$L(\theta;\mathbf{Z}) = \prod_{i=1}^{N} g_{\theta}(z_i)$$

Log-likelihood:

$$l(\theta;\mathbf{Z}) = \sum_{i=1}^{N} \log g_{\theta}(z_i)$$

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T y \qquad \widehat{\sigma} = \frac{1}{N}\sum(y_i - \hat{\mu}(x_i))^2$$

Bayesian methods

In the Bayesian approach to inference, we specify a sampling model $\Pr(\mathbf{Z}|\theta)$ (density or probability mass function) for our data given the parameters, and a ***prior distribution*** for the parameters $\Pr(\theta)$ reflecting our knowledge about θ before we see the data. We then compute the posterior distribution

$$\Pr(\theta \mid \mathbf{Z}) = \frac{\Pr(\mathbf{Z} \mid \theta) \cdot \Pr(\theta)}{\int \Pr(\mathbf{Z} \mid \theta) \cdot \Pr(\theta)\, d\theta}$$

The Bayesian approach differs from the standard ("***frequentist***") method for inference in its use of a prior distribution to ***express the uncertainty*** present before seeing the data, and to allow the uncertainty remaining after seeing the data to be expressed in the form of a posterior distribution.

Bayesian methods

The posterior distribution also provides the basis for predicting the values of a future observation z^{new}, via the ***predictive*** distribution:

$$\Pr(z^{\text{new}} \mid \mathbf{Z}) = \int \Pr(z^{\text{new}} \mid \theta) \cdot \Pr(\theta \mid \mathbf{Z}) d\theta$$

The maximum likelihood approach would use $\Pr(z^{\text{new}} \mid \hat{\theta})$, the data density evaluated at the maximum likelihood estimate, to predict future data. Unlike the ***predictive distribution***, this does not account for the uncertainty in estimating θ.

Bayesian methods

We assume that is σ^2 know, and a prior for the coefficients β as a Gaussian prior centered at zero

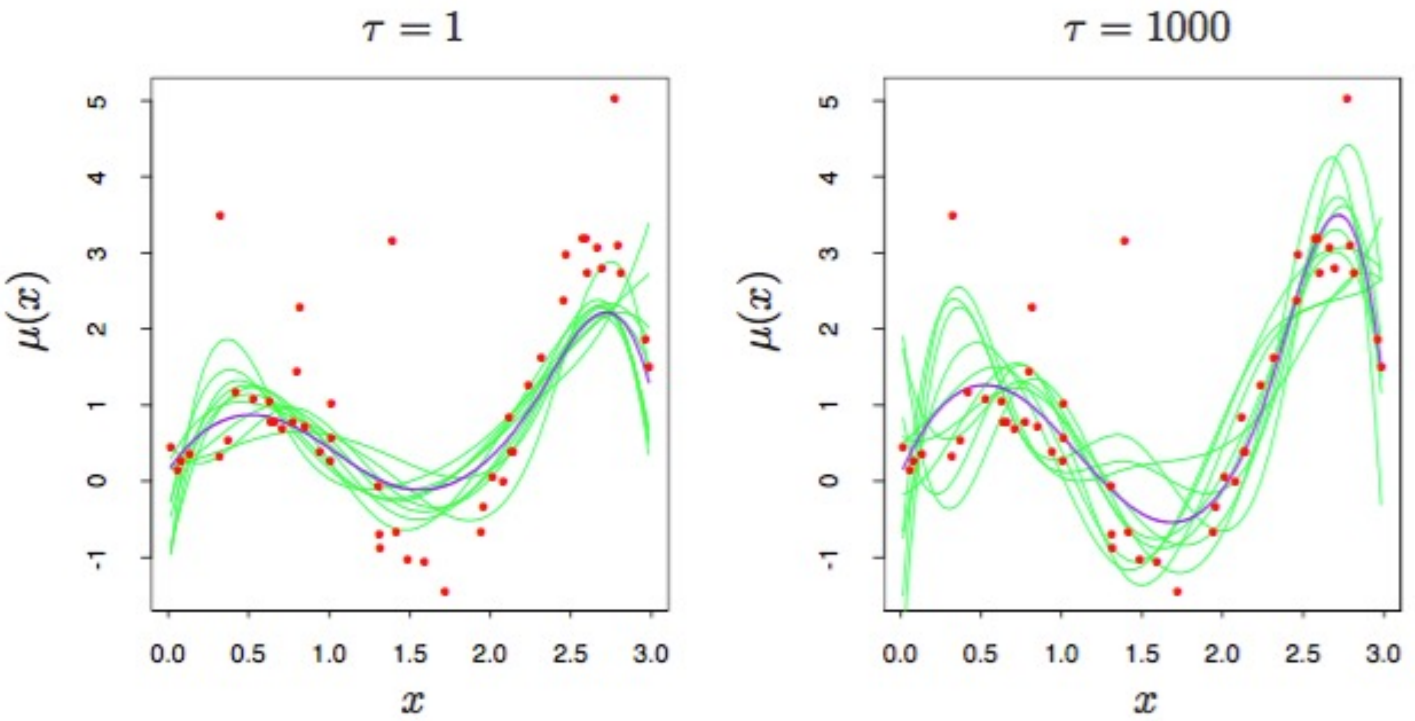
$$\beta \sim N(0, \tau\Sigma)$$

The posterior distribution for β is also Gaussian, with mean and covariance

$$\mathrm{E}(\beta \mid \mathbf{Z}) = \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma^{-1} \right)^{-1} \mathbf{H}^T y$$

$$\mathrm{cov}(\beta \mid \mathbf{Z}) = \left(\mathbf{H}^T\mathbf{H} + \frac{\sigma^2}{\tau}\Sigma^{-1} \right)^{-1} \sigma^2$$

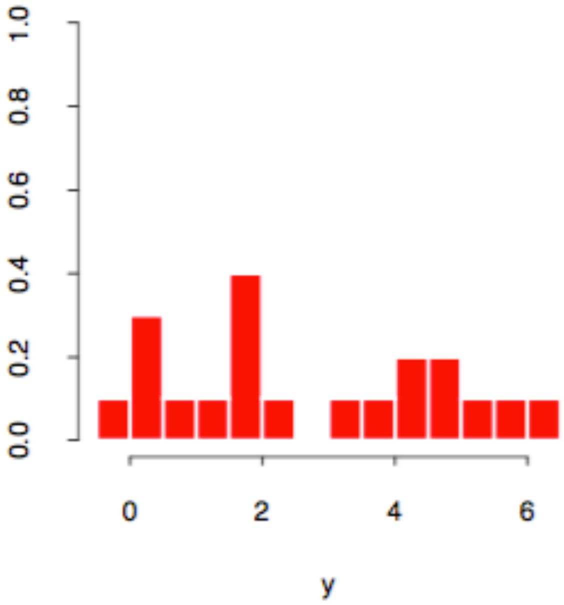
Prior information

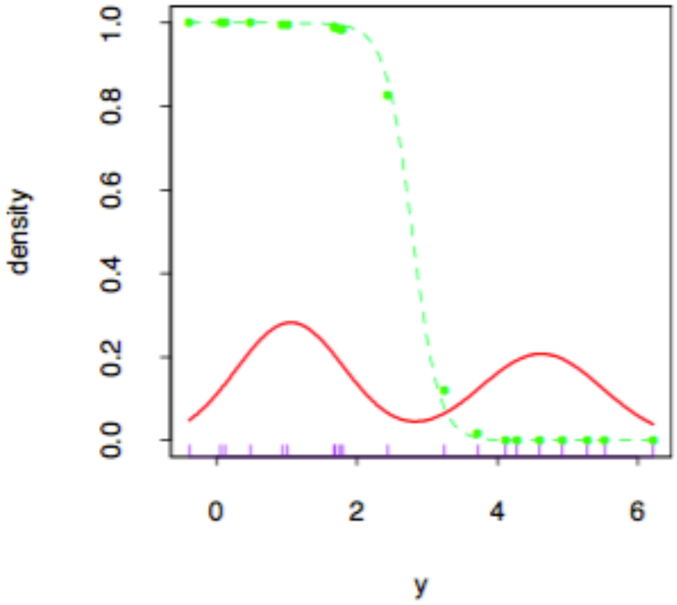


The distribution with $\tau \to \infty$ is called a ***noninformative prior*** for θ. In Gaussian models, maximum likelihood and parametric ***bootstrap analyses*** tend to agree with Bayesian analyses that use a noninformative prior for the free parameters.

The EM algorithm

The EM algorithm is a popular tool for simplifying ***difficult maximum likelihood*** problems.





We model Y as a mixture of two normal distributions:

$$Y_1 \sim N(\mu_1, \sigma_1^2) \qquad Y_2 \sim N(\mu_2, \sigma_2^2)$$

$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2 \qquad \Delta \in \{0,1\}, \Pr(\Delta = 1) = \pi$$

The EM algorithm

Let $\phi_\theta(x)$ denote the normal density with parameters $\theta = (\mu, \sigma^2)$. Then the density of Y is

$$g_Y(y) = (1-\pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y)$$

Now suppose we wish to fit this model to the data by maximum likelihood. The parameters are

$$\theta = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

The log-likelihood based on the N training cases is

$$l(\theta; \mathbf{Z}) = \sum_{i=1}^{N} \log\left[(1-\pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)\right]$$

The EM algorithm

We consider unobserved latent variables Δ_i taking values 0 or 1: if $\Delta_i = 1$ then Y_i comes from model 2, otherwise it comes from model 1. Suppose we knew the values of the Δ_i's. Then the log-likelihood would be

$$l(\theta;\mathbf{Z},\Delta) = \sum_{i=1}^{N} \log\left[(1-\Delta_i)\log\phi_{\theta_1}(y_i) + \Delta_i \log\phi_{\theta_2}(y_i)\right]$$

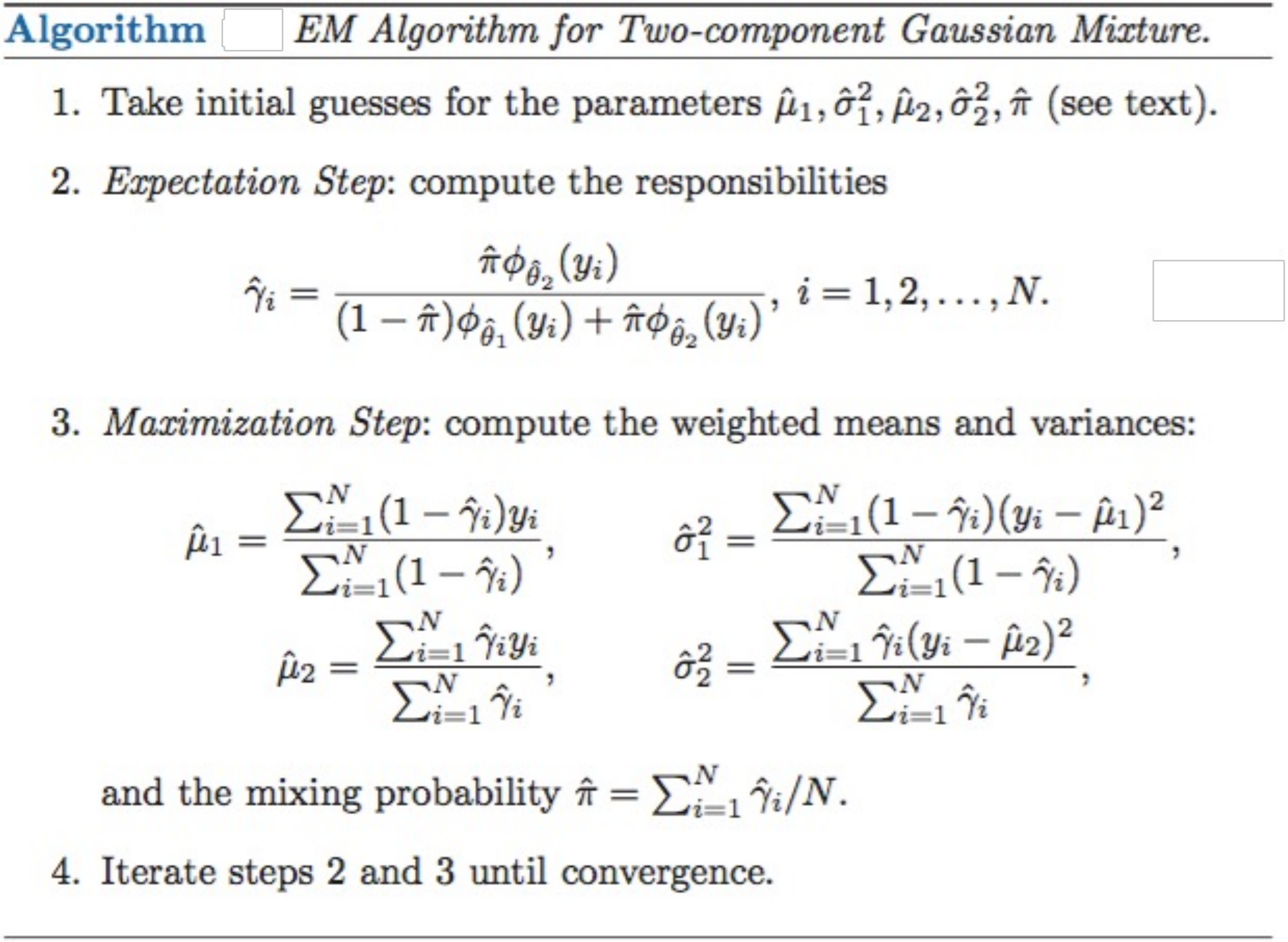
$$+\sum_{i=1}^{N} \log\left[(1-\Delta_i)\log(1-\pi) + \Delta_i \log\pi\right]$$

Since the values of the Δ_i's are actually unknown, we proceed in an iterative fashion, substituting for each Δ_i its expected value

$$\gamma_i(\theta) = \mathrm{E}(\Delta_i \mid \theta, \mathbf{Z}) = \Pr(\Delta_i = 1 \mid \theta, \mathbf{Z})$$

also called the ***responsibility*** of model 2 for observation i.

The EM (or Baum-Welch) algorithm



Algorithm *EM Algorithm for Two-component Gaussian Mixture.*

1. Take initial guesses for the parameters $\hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2, \hat{\pi}$ (see text).

2. *Expectation Step*: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \dots, N.$$

3. *Maximization Step*: compute the weighted means and variances:

$$\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i) y_i}{\sum_{i=1}^N (1-\hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^N (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^N (1-\hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^N \hat{\gamma}_i y_i}{\sum_{i=1}^N \hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^N \hat{\gamma}_i (y_i - \hat{\mu}_2)^2}{\sum_{i=1}^N \hat{\gamma}_i},$$

   and the mixing probability $\hat{\pi} = \sum_{i=1}^N \hat{\gamma}_i / N$.

4. Iterate steps 2 and 3 until convergence.

Why does the EM algorithm work?

The observed data is $\mathbf{Z}$, having log-likelihood $l(\theta;\, \mathbf{Z})$ depending on parameters θ. The latent or ***missing data*** is $\mathbf{Z}^m$, so that the complete data is $\mathbf{T} = (\mathbf{Z}, \mathbf{Z}^m)$ with log-likelihood $l_0(\theta;\, \mathbf{T})$, l_0 based on the ***complete density***. In the mixture problem $(\mathbf{Z}, \mathbf{Z}^m) = (\mathbf{y}, \Delta)$, and $l_0(\theta;\, \mathbf{T})$ is given in

$$l(\theta;\mathbf{Z},\Delta) = \sum_{i=1}^{N} \log\left[(1-\Delta_i)\log\phi_{\theta_1}(y_i) + \Delta_i \log\phi_{\theta_2}(y_i)\right]$$
$$+\sum_{i=1}^{N} \log\left[(1-\Delta_i)\log(1-\pi) + \Delta_i \log\pi\right]$$

Why does the EM algorithm work?

Since $\Pr(\mathbf{Z}^m \mid \mathbf{Z}, \theta') = \dfrac{\Pr(\mathbf{Z}^m, \mathbf{Z} \mid \theta')}{\Pr(\mathbf{Z} \mid \theta')}$

We can write $\Pr(\mathbf{Z} \mid \theta') = \dfrac{\Pr(\mathbf{T} \mid \theta')}{\Pr(\mathbf{Z}^m \mid \mathbf{Z}, \theta')}$

In terms of log-likelihoods, we have

$$l(\theta'; \mathbf{Z}) = l_0(\theta'; \mathbf{T}) - l_1(\theta'; \mathbf{Z}^m \mid \mathbf{Z})$$

Where l_1 is based on the conditional density $\Pr(\mathbf{Z}^m \mid \mathbf{Z}, \theta')$

Taking conditional expectations with respect to the distribution of $\mathbf{T}|\mathbf{Z}$ governed by parameter θ gives

$$\begin{aligned} l(\theta'; \mathbf{Z}) &= \mathrm{E}[l_0(\theta'; \mathbf{T}) \mid \mathbf{Z}, \theta] - \mathrm{E}[l_1(\theta'; \mathbf{Z}^m \mid \mathbf{Z}) \mid \mathbf{Z}, \theta] \\ &\equiv Q(\theta'; \theta) - R(\theta'; \theta) \end{aligned}$$

Why does the EM algorithm work?

In the *M* step, the EM algorithm maximizes $Q(\theta', \theta)$ over θ', rather than the actual objective function $l(\theta'; \mathbf{Z})$. Why does it succeed in maximizing $l(\theta'; \mathbf{Z})$?

If θ' maximizes $Q(\theta', \theta)$, we see that

$$l(\theta';\mathbf{Z}) - l(\theta;\mathbf{Z}) = [Q(\theta';\theta) - Q(\theta;\theta)] - [R(\theta';\theta) - R(\theta;\theta)]$$

$$\geq 0$$

Hence the EM iteration never decrease the log-likelihood.

This argument also makes it clear that a full maximization in the *M* step is not necessary: we need only to find a value $\hat{\theta}^{(j+1)}$ so that $Q(\theta' \mid \hat{\theta}^{(j)})$ increases as a function of the first argument

The EM algorithm in general

Algorithm *The EM Algorithm.*

1. Start with initial guesses for the parameters $\hat{\theta}^{(0)}$.

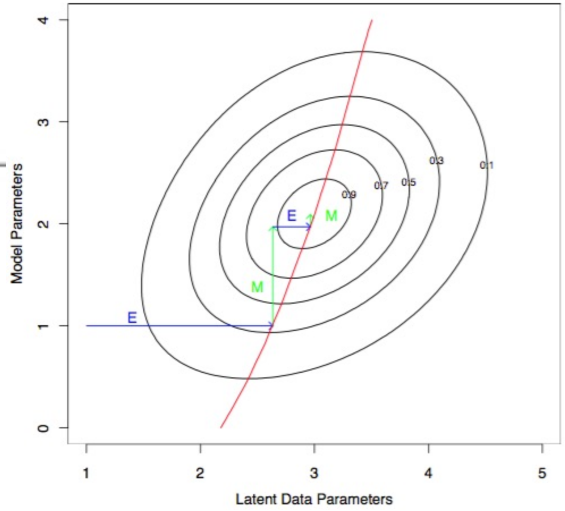
2. *Expectation Step*: at the jth step, compute

$$Q(\theta', \hat{\theta}^{(j)}) = \mathrm{E}(\ell_0(\theta'; \mathbf{T})|\mathbf{Z}, \hat{\theta}^{(j)})$$

   as a function of the dummy argument θ'.

3. *Maximization Step*: determine the new estimate $\hat{\theta}^{(j+1)}$ as the maximizer of $Q(\theta', \hat{\theta}^{(j)})$ over θ'.

4. Iterate steps 2 and 3 until convergence.



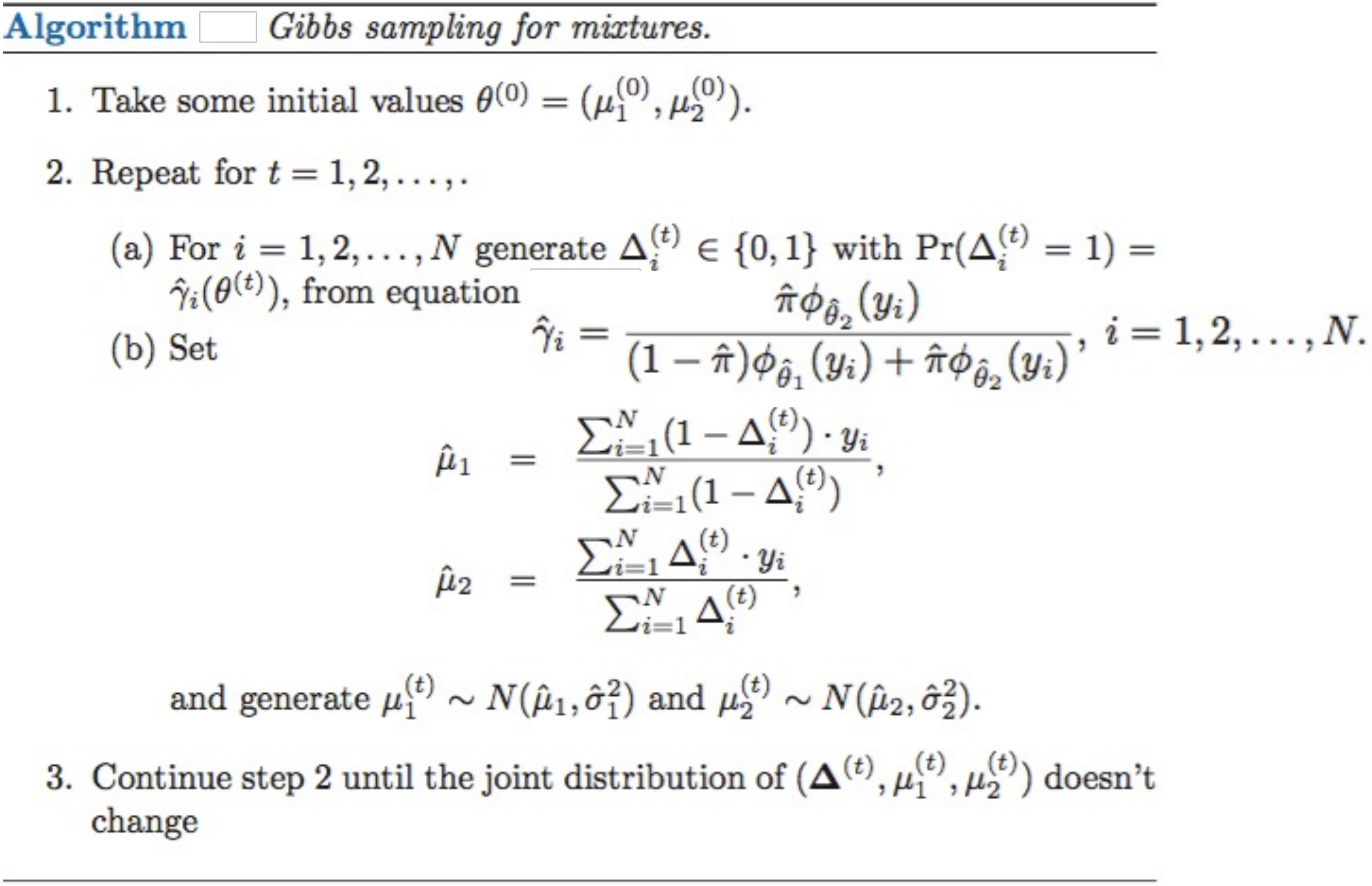
Markov chain Monte Carlo approach

- Having defined a Bayesian model, one would like to draw samples from the resulting ***posterior distribution***, in order to make ***inferences*** about the ***parameters***.
- This is often a difficult computational problem. We discuss the ***Markov chain Monte Carlo*** (MCMC) approach to posterior sampling. We will see that ***Gibbs sampling***, an MCMC procedure, is closely related to the EM algorithm.
- We have random variables $U_1, U_2, \ldots, U_K$, and we wish to draw a sample from their ***joint distribution***. Suppose this is difficult to do, but it is easy to simulate from ***the conditional distributions*** $\Pr(U_j | U_1, U_2, \ldots, U_{j-1}, U_{j+1}, \ldots, U_K), j = 1, 2, \ldots, K$.

Markov chain Monte Carlo approach

- The ***Gibbs sampling*** procedure alternatively simulates from each of these distributions and when the process stabilizes, provides a sample from the desired ***joint distribution***.
- Note that we don't need to know the explicit form of the conditional densities, but just need to be able to ***sample*** from them.
- Gibbs sampling will be helpful if it is easy to sample from the ***conditional distribution*** of each parameter given the other parameters and $\mathbf{Z}$.
- There is a close connection between ***Gibbs sampling*** from a ***posterior*** and the ***EM algorithm*** in exponential family models. The key is to consider the latent data $\mathbf{Z}^m$ from the EM procedure to be another parameter for the Gibbs sampler.

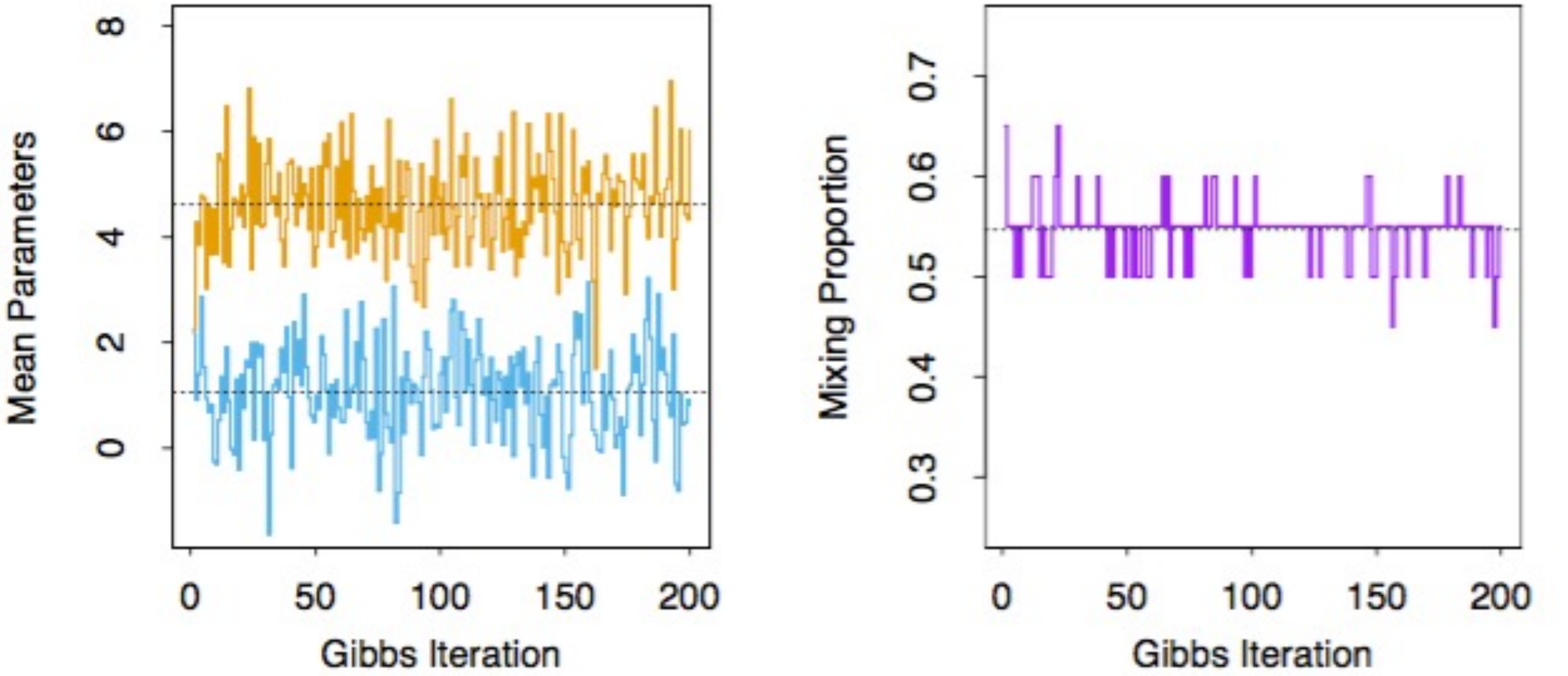
Gibbs sampling for mixtures



Algorithm *Gibbs sampling for mixtures.*

```
1. Take some initial values $\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)})$.
2. Repeat for $t = 1, 2, \ldots,$.
    (a) For $i = 1, 2, \ldots, N$ generate $\Delta_i^{(t)} \in \{0, 1\}$ with $\Pr(\Delta_i^{(t)} = 1) = \hat{\gamma}_i(\theta^{(t)})$, from equation
        $\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1-\hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\hat{\theta}_2}(y_i)}, \quad i = 1, 2, \ldots, N.$
    (b) Set
        $\hat{\mu}_1 = \frac{\sum_{i=1}^N (1-\Delta_i^{(t)}) \cdot y_i}{\sum_{i=1}^N (1-\Delta_i^{(t)})},$
        $\hat{\mu}_2 = \frac{\sum_{i=1}^N \Delta_i^{(t)} \cdot y_i}{\sum_{i=1}^N \Delta_i^{(t)}},$
        and generate $\mu_1^{(t)} \sim N(\hat{\mu}_1, \hat{\sigma}_1^2)$ and $\mu_2^{(t)} \sim N(\hat{\mu}_2, \hat{\sigma}_2^2)$.
3. Continue step 2 until the joint distribution of $(\mathbf{\Delta}^{(t)}, \mu_1^{(t)}, \mu_2^{(t)})$ doesn't change
```

Gibbs sampling for mixtures



Mixture example. (Left panel:) 200 values of the two mean parameters from Gibbs sampling; horizontal lines are drawn at the maximum likelihood estimates $\hat{\mu}_1$, $\hat{\mu}_2$. (Right panel:) Proportion of values with $\Delta_i = 1$, for each of the 200 Gibbs sampling iterations; a horizontal line is drawn at $\sum_i \hat{\gamma}_i/N$.

Any questions?

AI Research Group
Fudan University