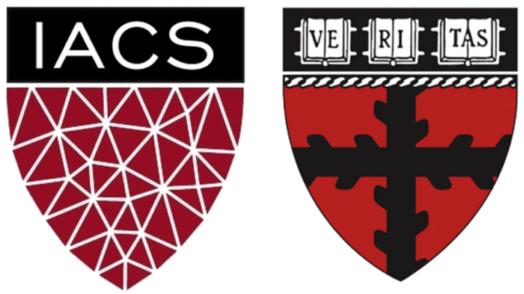
Lecture #9: Latent Variable Models and MLE

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020


IACS
VE RI TAS

Outline

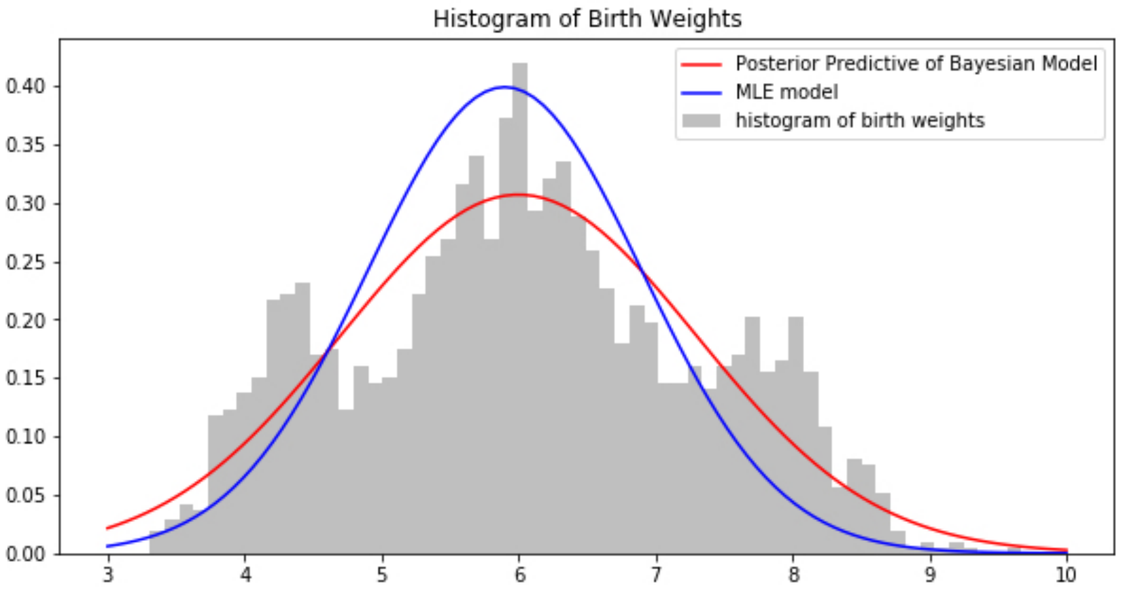
1. Motivation for Latent Variable Models
2. Common Latent Variable Models
3. Maximum Likelihood Estimation for Latent Variable Models: Expectation Maximization
4. Mixture of Gaussians

Motivation for Latent Variable Models

A Model for Birth Weights

Recall our model for birth weigths, $Y_1, \ldots, Y_N$. We *posited* that the birth weights are iid normally distributed with known σ^2, $Y_n \sim \mathcal{N}(\mu, 1)$.

Compare the maximum likelihood model and the Bayesian model for bith weight. Which model would you use to make clinical decisions? What's hard about this comparison?



A Similarity Measure for Distributions: Kullback–Leibler Divergence

Visually comparing models to the ***empirical distribution*** of the data is impractical. Fortunately, there are a large number of quantitative measures for comparing two distributions, these are called ***divergence measures***. For example, the ***Kullback–Leibler (KL) Divergence*** is defined for two distributions $p(\theta)$ and $q(\theta)$ supported on Θ as:

$$D_{\mathrm{KL}}[q \parallel p] = \int_{\Theta} \log\left[\frac{q(\theta)}{p(\theta)}\right] q(\theta) d\theta$$

The KL-divergence $D_{\mathrm{KL}}[q \parallel p]$ is bounded below by 0, which happens if and only if $q = p$. The KL-divergence has information theoretic interpretations that we will explore later in the course.

Note: The KL-divergence is defined in terms of the pdf's of p and q. If p is a distribution from which we only have samples and not the pdf (like the empirical distribution), we can nontheless estimate $D_{\mathrm{KL}}[q \parallel p]$. Techniques that estimate the KL-divergence from samples are called ***non-parametric***. We will use them later in the course.

Class Membership as a Latent Variable

We observe that there are three ***clusters*** in the data. We posit that there are three ***classes*** of infants in the study: infants with low birth weights, infants with normal birth weights and those with high birth weights. The numbers of infants in the classes are not equal.

For each observation Y_n, we model its class membership Z_n as a categorical variable,

$$Z_n \sim Cat(\pi),$$

where π_i in $\pi = [\pi_1, \pi_2, \pi_3]$ is the class proportion. Note that we don't have the class membership Z_n in the data! So Z_n is called a ***latent variable***.

Depending on the class, the n-th birth weight Y_n will have a different normal distribution,

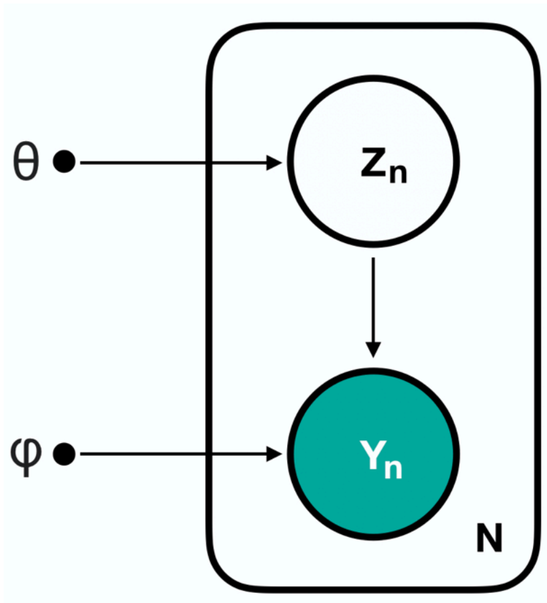
$$Y_n | Z_n \sim \mathcal{N}\left(\mu_{Z_n}, \sigma^2_{Z_n}\right)$$

where μ_{Z_n} is one of the three class means $[\mu_1, \mu_2, \mu_3]$ and $\sigma^2_{Z_n}$ is one of the three class variances $[\sigma^2_1, \sigma^2_2, \sigma^2_3]$.

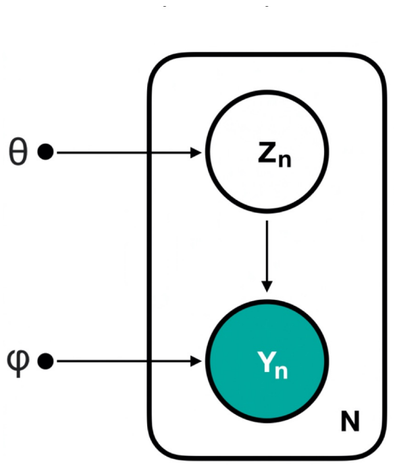
Common Latent Variable Models

Latent Variable Models

Models that include an observed variable Y and at least one unobserved variable Z are called ***latent variable models***. In general, our model can allow Y and Z to interact in many different ways. Today, we will study models with one type of interaction:



$$
\begin{aligned}
Z_n &\sim p(Z|\theta) \\
Y_n|Z_n &\sim p(Y|Z,\phi) \\
n &= 1,\ldots,N
\end{aligned}
$$



Item-Response Models

In ***item-response models***, we measure an real-valued unobserved trait Z of a subject by performing a series of experiments with binary observable outcomes, Y:

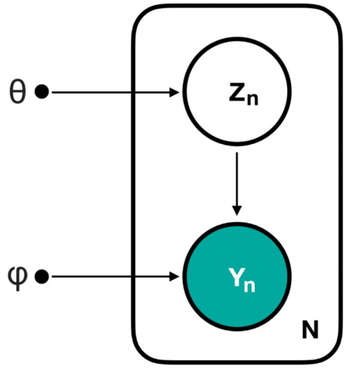
$$\begin{aligned} Z_n &\sim \mathcal{N}(\mu, \sigma^2), \\ \theta_n &= g(Z_n) \\ Y_n|Z_n &\sim Ber(\theta_n), \end{aligned}$$

where $n = 1, \ldots, N$ and g is some fixed function of Z_n.

Applications

Item response models are used to model the way "underlying intelligence" Z relates to scores Y on IQ tests.

Item response models can also be used to model the way "suicidality" Z relates to answers on mental health surveys. Building a good model may help to infer when a patient is at psychiatric risk based on in-take surveys at points of care through out the health-care system.



Factor Analysis Models

In ***factor analysis models***, we posit that the observed data Y with many measurements is generated by a small set of unobserved factors Z:

$$Z_n \sim \mathcal{N}(0, I),$$
$$Y_n | Z_n \sim \mathcal{N}(\mu + \Lambda Z_n, \Phi),$$

where $n = 1, \ldots, N, Z_n \in \mathbb{R}^{D'}$ and $Y_n \in \mathbb{R}^D$. We typically assume that D' is much smaller than D.

Applications

Factor analysis models are useful for biomedical data, where we typically measure a large number of characteristics of a patient (e.g. blood pressure, heart rate, etc), but these characteristics are all generated by a small list of health factors (e.g. diabetes, cancer, hypertension etc). Building a good model means we may be able to infer the list of health factors of a patient from their observed measurements.

Maximum Likelihood Estimation for Latent Variable Models: Expectation Maximization

MODEL

$x_n \sim p(x_n|z_n, \theta)$
$(x_n = f(z_n, \varepsilon_n, \theta))$
parameters: θ, β

Likelihood over q's:

$$\log \prod_{n=1}^{N} p(y_n|x_n, \beta)\, p(x_n|\theta) = \sum_{n=1}^{N} \log \underbrace{p(y_n|x_n, \beta) + \log p(x_n|\theta)}_{\text{can we evaluate this?}}$$

Likelihood over observed data:

$$\log \prod_{n=1}^{N} p(y_n|\theta, \beta) = \sum_{n=1}^{N} \log \int p(y_n|x_n, \beta)\, p(x_n|\theta)\, dx_n$$

$$= \sum_{n=1}^{N} \log \int p(y_n|x_n, \beta)\, \frac{p(x_n|\theta)}{q(x_n)}\, q(x_n)\, dx_n \quad \text{does the log hide?}$$

$$= \sum_{n=1}^{N} \log \underbrace{\mathbb{E}_{x_n \sim q(x_n)}\left[p(y_n|x_n, \beta)\, \frac{p(x_n|\theta)}{q(x_n)} \right]}_{\text{is this hard?}}$$

The maximum likelihood objective:

$$\theta_{ML}, \theta_{ML} = \arg\max_{\theta, \beta} \ell(\theta, \beta) = \arg\max_{\theta, \beta} \sum_{n=1}^{N} \log \underbrace{\mathbb{E}_{x_n \sim p(x_n|\theta)}\left[p(y_n|x_n, \beta) \right]}_{\text{is this hard?}}$$

Trying out the optimization:

$$\nabla_{\theta,\beta}\, \ell(\theta, \beta) = \sum_{n=1}^{N} \nabla_{\theta,\beta} \log \mathbb{E}_{x_n \sim p(x_n|\theta)}\left[p(y_n|x_n, \beta) \right]$$

$$= \sum_{n=1}^{N} \nabla_{\theta,\beta} \log \mathbb{E}_{x_n \sim p(x_n|\theta)}\left[p(y_n|x_n, \beta) \right]$$

$$\because \nabla_{\theta,\beta} \log f = \frac{\nabla_{\theta,\beta} f}{f}$$

$$= \sum_{n=1}^{N} \frac{\nabla_{\theta,\beta} \mathbb{E}_{x_n \sim p(x_n|\theta)}\left[p(y_n|x_n, \beta) \right]}{\mathbb{E}_{x_n \sim p(x_n|\theta)}\left[p(y_n|x_n, \beta) \right]}$$

can we MC estimate this?

Calculus Facts we need:

1. $\mathbb{E}_{x \sim p(x)}[f(x)] \rightarrow$ Footnote
2. Properties of $\mathbb{E}$:
   - A $\mathbb{E}_{x \sim p(x)}[a f(x)] = a \frac{1}{C} \sum_{x \sim p(x)} f(x) = a \sum_{x \sim p(x)} \mathbb{E}[f(x)]$
   - (Jensen's inequality) $\rightarrow$ B. $\log \mathbb{E}_{x \sim p(x)}[f(x)] \geq \mathbb{E}_{x \sim p(x)}[\log f(x)]$
   - C $\nabla_\theta \mathbb{E}_{x \sim p(x|\theta)}[f(x,w)] = \mathbb{E}_{x \sim p(x|\theta)}[\nabla_\theta f(x,w)]$
3. $D_{KL}[q(x) \| p(x)] = \left[\log \frac{q(x)}{p(x)}\right]$ Aside

$$= \mathbb{E}_{x \sim q(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

An objective we can actually work with:

$$\max_{\theta,\beta} \ell(\theta, \beta) = \max_{\theta,\beta} \sum_{n=1}^{N} \log \mathbb{E}_{x_n \sim p(x_n|\theta)}\left[p(y_n|x_n, \beta) \right]$$

$$= \max_{\theta,\beta} \sum_{n=1}^{N} \log \int p(y_n|x_n, \beta)\, p(x_n|\theta)\, dx_n$$

$$= \max_{\theta,\beta} \sum_{n=1}^{N} \log \int p(y_n|x_n, \beta)\, \frac{p(x_n|\theta)}{q(x_n)}\, q(x_n)\, dx_n$$

introduce auxiliary variables $p(x_n)$! q is some distr. of your choice.

$$= \max_{\theta,\beta} \sum_{n=1}^{N} \log \underbrace{\mathbb{E}_{x_n \sim q(x_n)}\left[p(y_n|x_n, \beta)\, \frac{p(x_n|\theta)}{q(x_n)} \right]}_{\text{is this log helping?}}$$

does $\nabla_{\theta,\beta}$ commute with $\mathbb{E}_{x \sim q(x)}$?

$$\geq \max_{\theta,\beta,q} \sum_{n=1}^{N} \underbrace{\mathbb{E}_{x_n \sim q(x_n)}\left[\log \left(p(y_n|x_n, \beta)\, \frac{p(x_n|\theta)}{q(x_n)} \right) \right]}_{\text{The Evidence Lower Bound ELBO}(\theta, \beta, q)}$$

Idea Instead of maximizing the log-likelihood, we maximize the lower bound ELBO.

Why you Should

When ELBO is maximized $\ell(\theta, \beta)$ is guaranteed to be as big.

Why you Shouldn't

If ELBO is maximized then $\ell(\theta, \beta)$ can be far from optimized

How to maximize ELBO: coordinate ascent

I. maximize θ, β, fixing q^t

$$\theta^*, \beta^* = \arg\max_{\theta,\beta} \text{ELBO}(\theta, \beta, q^t)$$

(M-step)

$$= \arg\max_{\theta,\beta} \sum_{n=1}^{N} \mathbb{E}_{x_n \sim q^t(x_n)}\left[\log \left(\frac{p(y_n|x_n, \beta)\, p(x_n|\theta)}{q^t(x_n)} \right) \right]$$

$$= \arg\max_{\theta,\beta} \sum_{n=1}^{N} \left(\mathbb{E}_{x_n \sim q^t(x_n)}\left[\log \left(p(y_n|x_n, \beta)\, p(x_n|\theta) \right) \right] - \mathbb{E}_{x_n \sim q^t(x_n)}\left[\log q^t(x_n) \right] \right)$$

irrelevant for $\max_{\theta,\beta}$

$$\equiv \arg\max_{\theta,\beta} \sum_{n=1}^{N} \mathbb{E}_{x_n \sim q^t(x_n)}\left[\log \left(p(y_n|x_n, \beta)\, p(x_n|\theta) \right) \right]$$

is this problem easier?

II. maximize q, fixing θ^t, β^t

$$q^t = \arg\max_{q} \text{ELBO}(\theta^t, \beta^t, q)$$

Note $\ell(\theta^t, \beta^t) - \text{ELBO}(\theta^t, \beta^t, q) = \sum_{n=1}^{N} \left[\log p(y_n|\theta^t, \beta^t) - \mathbb{E}_{x_n \sim q(x_n)}\left[\log \frac{p(y_n, x_n|\theta^t, \beta^t)}{q(x_n)} \right] \right]$

$$= \sum_{n=1}^{N} \left[\mathbb{E}_{x_n \sim q(x_n)}\left[\log p(y_n|\theta^t, \beta^t) \right] - \mathbb{E}_{x_n \sim q(x_n)}\left[\log \frac{p(y_n, x_n|\theta^t, \beta^t)}{q(x_n)} \right] \right]$$

$$= \sum_{n=1}^{N} \left[\mathbb{E}_{x_n \sim q(x_n)}\left[\log \frac{p(y_n|\theta^t, \beta^t)\, q(x_n)}{p(y_n, x_n|\theta^t, \beta^t)} \right] \right]$$

$$= \sum_{n=1}^{N} \left[\mathbb{E}_{x_n \sim q(x_n)}\left[\log \frac{q(x_n)}{p(x_n|y_n, \theta^t, \beta^t)} \right] \right]$$

$$= \sum_{n=1}^{N} D_{KL}\left[q(x_n) \,\|\, p(x_n|y_n, \theta^t, \beta^t) \right]$$

$$q^t = \arg\max_{q} \text{ELBO}(\theta^t, \beta^t, q) = \arg\min_{q} D_{KL}[q(x_n) \,\|\, p(x_n|y_n, \theta^t, \beta^t)]$$

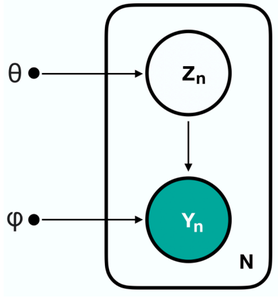
$$q^t = p(x_n|y_n, \theta^t, \beta^t)$$

Are we done?

Maximizing each coordinate once is not sufficient, we need to iterate!

The Expectation Maximization Algorithm

The ***exepectation maximization (EM) algorithm*** maximize the ELBO of the model,



$$Z_n \sim p(Z|\theta)$$
$$Y_n|Z_n \sim p(Y|Z,\phi)$$
$$n = 1, \dots, N$$

1. **Initialization:** Pick θ_0, ϕ_0.
2. Repeat $i = 1, \dots, I$ times:

   E-Step:
   $$q_{\text{new}}(Z_n) = \underset{q}{\text{argmax}}\ ELBO(\theta_{\text{old}}, \phi_{\text{old}}, q) = p(Z_n|Y_n, \theta_{\text{old}}, \phi_{\text{old}})$$

   M-Step:
   $$\theta_{\text{new}}, \phi_{\text{new}} = \underset{\theta,\phi}{\text{argmax}}\ ELBO(\theta, \phi, q_{\text{new}})$$
   $$= \underset{\theta,\phi}{\text{argmax}} \sum_{n=1}^{N} \mathbb{E}_{Z_n \sim p(Z_n|Y_n,\theta_{\text{old}},\phi_{\text{old}})} \left[\log(p(y_n, Z_n|\phi, \theta)\right].$$

Example: EM for the Gaussian Mixture Model of Birth Weight

Solving the M-Step

We see that the optimization problem in the M-step:
$\mu_{\text{new}}, \sigma^2_{\text{new}}, \pi_{\text{new}} = \underset{\mu,\sigma^2,\pi}{\text{argmax}}\ ELBO(\mu, \sigma^2, \pi, q_{\text{new}})$ is equivalent to two problems

$$1. \quad \underset{\mu,\sigma^2}{\text{argmax}} \sum_{n=1}^{N}\sum_{k=1}^{K} r_{n,k} \log \mathcal{N}(y_n; \mu_k, \sigma_k^2)$$

$$2. \quad \underset{\pi}{\text{argmax}} \sum_{n=1}^{N}\sum_{k=1}^{K} r_{n,k}\pi_k$$

We can solve each optimization problem analytically by finding stationary points of the gradient (or the Lagrangian):

- $\mu_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k} y_n$
- $\sigma^2_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k}(y_n - \mu_{\text{new}})^2$
- $\pi_{\text{new}} = \frac{\sum_{n=1}^{N} r_{n,k}}{N}$

Example: EM for the Gaussian Mixture Model of Birth Weight

All Together

Initialization: Pick any π, μ, σ^2

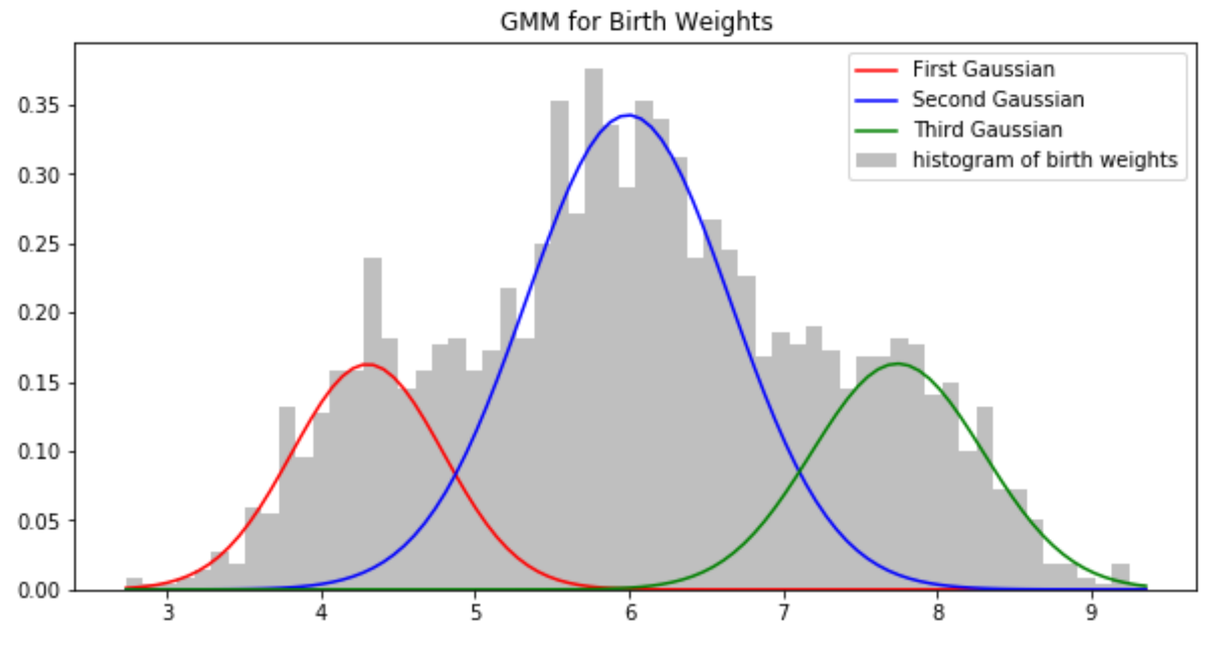
E-Step: Compute $r_{n,k} = \dfrac{\pi_{k,\text{old}}\ \mathcal{N}(y_n; \mu_{k,\text{old}}, \sigma^2_{k,\text{old}})}{\mathcal{Z}}$, where $\mathcal{Z} = \sum_{k=1}^{K} \pi_{k,\text{old}}\ \mathcal{N}(y_n; \mu_{k,\text{old}}, \sigma^2_{k,\text{old}})$.

M-Step: Compute model parameters:

- $\mu_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k}\, y_n$
- $\sigma^2_{\text{new}} = \frac{1}{\sum_{n=1}^{N} r_{n,k}} \sum_{n=1}^{N} r_{n,k} (y_n - \mu_{\text{new}})^2$
- $\pi_{\text{new}} = \frac{\sum_{n=1}^{N} r_{n,k}}{N}$

Implementing EM for the Gaussian Mixture Model of Birth Weight

In [3]:
```
fig, ax = plt.subplots(1, 1, figsize=(10, 5))
ax.hist(y, bins=60, density=True, color='gray', alpha=0.5, label='histogram of b
irth weights')
ax.plot(x, pi_current[0] * sp.stats.norm(mu_current[0], sigma_current[0]**0.5).p
df(x), color='red', label='First Gaussian')
ax.plot(x, pi_current[1] * sp.stats.norm(mu_current[1], sigma_current[1]**0.5).p
df(x), color='blue', label='Second Gaussian')
ax.plot(x, pi_current[2] * sp.stats.norm(mu_current[2], sigma_current[2]**0.5).p
df(x), color='green', label='Third Gaussian')
ax.set_title('GMM for Birth Weights')
ax.legend(loc='best')
plt.show()
```



Sanity Check: Log-Likelihood During Training

Remember that ploting the MLE model against actual data is not always an option (e.g. high-dimensional data).

A sanity check for that your EM algorithm has been implemented correctly is to plot the observed data log-likelihood over the iterations of the algorithm:

$$\ell_y(\mu, \sigma^2, \pi) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \mathcal{N}(y_n; \mu_k, \sigma_k^2)\pi_k$$

In [4]:
```
fig, ax = plt.subplots(1, 1, figsize=(10, 3))
ax.plot(range(len(log_lkhd)), log_lkhd, color='red', alpha=0.5)
ax.set_title('observed data log-likelihood over iterations of EM')
plt.show()
```