LECTURE 10: THE EM ALGORITHM (CONTD)

STAT 545: Intro. to Computational Statistics

Vinayak Rao

Purdue University

September 24, 2018

Consider a space $\mathbb{X}$. E.g. $\mathbb{R}, \mathbb{R}^d$ or $\mathbb{N}$.

$\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_D(x)]$: (feature) vector of sufficient statistics

$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D]$: vector of natural parameters

Exponential family models

Consider a space $\mathbb{X}$. E.g. $\mathbb{R}, \mathbb{R}^d$ or $\mathbb{N}$.

$\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_D(x)]$: (feature) vector of sufficient statistics

$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D]$: vector of natural parameters

Exponential family distribution:

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x))$$

Exponential family models

Consider a space $\mathbb{X}$. E.g. $\mathbb{R}, \mathbb{R}^d$ or $\mathbb{N}$.

$\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_D(x)]$: (feature) vector of sufficient statistics

$\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D]$: vector of natural parameters

Exponential family distribution:

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\theta)} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x))$$

$h(x)$ is the base-measure or base distribution.

$Z(\theta) = \int h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x)) \mathrm{d}x$ is the normalization constant.

Examples

The normal distribution:

$$
\begin{aligned}
p(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\
&= \frac{\exp\left(-\frac{\mu^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right)
\end{aligned}
$$

Examples

The normal distribution:

$$
\begin{aligned}
p(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \\
&= \frac{\exp\left(-\frac{\mu^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right)
\end{aligned}
$$

The Poisson distribution:

$$
\begin{aligned}
p(x|\lambda) &= \frac{\lambda^x \exp(-\lambda)}{x!} \\
&= \exp(-\lambda)\frac{1}{x!}\exp(\log(\lambda)x)
\end{aligned}
$$

Minimal exponential family

Sufficient statistics are linearly independent

Consider a K-component discrete distribution $\pi = (\pi_1, \ldots, \pi_K)$

$$p(X) = \prod_{c=1}^{K} \pi_i^{\delta(X=c)} = \exp(\sum_{c=1}^{K} \delta(X = c) \log \pi_c)$$

Minimal exponential family

Sufficient statistics are linearly independent

Consider a K-component discrete distribution $\pi = (\pi_1, \ldots, \pi_K)$

$$p(X) = \prod_{c=1}^{K} \pi_i^{\delta(X=c)} = \exp(\sum_{c=1}^{K} \delta(X=c) \log \pi_c)$$

Is it minimal?

$$\begin{aligned} p(X) &= \pi_K \exp(\sum_{c=1}^{K-1} \delta(X=c) \log \pi_c/\pi_K) \\ &= \frac{1}{Z} \exp(\sum_{c=1}^{K-1} \delta(X=c)\theta_c) \end{aligned}$$

Maximum-likelihood estimation

Given N i.i.d. observations $X \equiv \{x_1, \ldots, x_N\}$, the likelihood is

$$\begin{aligned}
\mathcal{L}(X|\boldsymbol{\theta}) &= \prod_{i=1}^{N} \frac{1}{Z(\theta)} h(x_i) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x_i)) \\
&= \left(\frac{1}{Z(\theta)}\right)^N \left(\prod_{i=1}^{N} h(x_i)\right) \exp(\boldsymbol{\theta}^\top \sum_{i=1}^{N} \boldsymbol{\phi}(x_i))
\end{aligned}$$

Given N i.i.d. observations $X \equiv \{x_1, \ldots, x_N\}$, the likelihood is

$$\begin{aligned}\mathcal{L}(X|\boldsymbol{\theta}) &= \prod_{i=1}^{N} \frac{1}{Z(\theta)} h(x_i) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x_i)) \\ &= \left(\frac{1}{Z(\theta)}\right)^N \left(\prod_{i=1}^{N} h(x_i)\right) \exp(\boldsymbol{\theta}^\top \sum_{i=1}^{N} \boldsymbol{\phi}(x_i))\end{aligned}$$

The log-likelihood $\ell(X|\boldsymbol{\theta}) = \log \mathcal{L}(X|\boldsymbol{\theta})$ is

$$\ell(X|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \left(\sum_{i=1}^{N} \boldsymbol{\phi}(x_i)\right) - N \log Z(\theta) + \sum_{i=1}^{N} \log h(x_i)$$

Maximum-likelihood estimation

Given N i.i.d. observations $X \equiv \{x_1, \ldots, x_N\}$, the likelihood is

$$\begin{aligned}\mathcal{L}(X|\boldsymbol{\theta}) &= \prod_{i=1}^{N} \frac{1}{Z(\theta)} h(x_i) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x_i)) \\ &= \left(\frac{1}{Z(\theta)}\right)^N \left(\prod_{i=1}^{N} h(x_i)\right) \exp(\boldsymbol{\theta}^\top \sum_{i=1}^{N} \boldsymbol{\phi}(x_i))\end{aligned}$$

The log-likelihood $\ell(X|\boldsymbol{\theta}) = \log \mathcal{L}(X|\boldsymbol{\theta})$ is

$$\ell(X|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \left(\sum_{i=1}^{N} \boldsymbol{\phi}(x_i)\right) - N \log Z(\theta) + \sum_{i=1}^{N} \log h(x_i)$$

To calculate a maximum likelihood estimate, we only need the sum of the suff. statistics.

MLE for exponential families

$$\ell(X|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \left(\sum_{i=1}^{N} \phi(x_i) \right) - N \log Z(\theta) + \sum_{i=1}^{N} \log h(x_i)$$

MLE for exponential families

$$\ell(X|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \left(\sum_{i=1}^{N} \phi(x_i) \right) - N \log Z(\theta) + \sum_{i=1}^{N} \log h(x_i)$$

At MLE of θ_d, the dth component of $\boldsymbol{\theta}$: $\frac{\partial \ell(X|\boldsymbol{\theta})}{\partial \theta_d} = 0$.

MLE for exponential families

$$\ell(X|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \left(\sum_{i=1}^{N} \boldsymbol{\phi}(x_i) \right) - N \log Z(\theta) + \sum_{i=1}^{N} \log h(x_i)$$

At MLE of θ_d, the dth component of $\boldsymbol{\theta}$: $\frac{\partial \ell(X|\boldsymbol{\theta})}{\partial \theta_d} = 0$.

$$\sum_{i=1}^{N} \phi_d(x_i) = N \frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_d}$$

MLE for exponential families

$$\ell(X|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \left(\sum_{i=1}^{N} \boldsymbol{\phi}(x_i)\right) - N \log Z(\theta) + \sum_{i=1}^{N} \log h(x_i)$$

At MLE of θ_d, the dth component of $\boldsymbol{\theta}$: $\frac{\partial \ell(X|\boldsymbol{\theta})}{\partial \theta_d} = 0$.

$$\sum_{i=1}^{N} \boldsymbol{\phi}_d(x_i) = N \frac{\partial \log Z(\boldsymbol{\theta})}{\partial \theta_d}$$

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\phi}_d(x_i) &= \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial Z(\boldsymbol{\theta})}{\partial \theta_d} = \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial \int h(x)\exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x))\mathrm{d}x}{\partial \theta_d} \\
&= \frac{1}{Z(\boldsymbol{\theta})} \int h(x) \frac{\partial \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x))}{\partial \theta_d} \mathrm{d}x \\
&= \int \frac{1}{Z(\boldsymbol{\theta})} h(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x)) \phi_d(x) \mathrm{d}x
\end{aligned}$$

MLE for exponential families

Match empirical and population averages of $\phi(x)$:

$$\frac{1}{N}\sum_{i=1}^{N}\phi_d(x_i) = \mathbb{E}_{\theta_{MLE}}[\phi_d(x)]$$

MLE for exponential families

Match empirical and population averages of $\phi(x)$:

$$\frac{1}{N}\sum_{i=1}^{N}\phi_d(x_i) = \mathbb{E}_{\theta_{MLE}}[\phi_d(x)]$$

RHS: ‘moment parameters’ of the exponential distribution.

Thus: θ_{MLE} are natural parameters corresponding to empirical moment parameters (‘moment matching’).

MLE for exponential families

Match empirical and population averages of $\boldsymbol{\phi}(x)$:

$$\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\phi}_d(x_i) = \mathbb{E}_{\theta_{MLE}}[\phi_d(x)]$$

RHS: 'moment parameters' of the exponential distribution.

Thus: θ_{MLE} are natural parameters corresponding to empirical moment parameters ('moment matching').

Is this a maximum?

- is second derivative (Hessian) negative (negative definite)?

We can show $\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log Z(X|\boldsymbol{\theta}) = \mathrm{Cov}(\phi_i, \phi_j)$, and the Hessian of $\ell(X|\theta)$ is $-N$ times the feature covariance matrix

Example

The 1-d Gaussian: $\phi = [x \quad x^2]$
Moment parameters are mean and mean squared
Easy to find corresponding natural parameters

Quite often, it is not the case.
However, we will restrict ourselves to cases where it is.

Example

The 1-d Gaussian: $\phi = [x \quad x^2]$
Moment parameters are mean and mean squared
Easy to find corresponding natural parameters

Quite often, it is not the case.
However, we will restrict ourselves to cases where it is.

Can you do it for the Poisson?

$$p(x|\lambda) = \frac{1}{x!}\lambda^x \exp(-\lambda)$$

Missing data in exponential family distributions

Let samples from the exponential family have two parts: $[x\ y]$.

Feature vector $\boldsymbol{\phi}([x\ y]) := \boldsymbol{\phi}(x, y)$.

$$P(x, y|\theta) := P([x\ y]|\boldsymbol{\theta}) = \frac{h(x, y)}{Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

Let samples from the exponential family have two parts: $[x\ y]$.

Feature vector $\boldsymbol{\phi}([x\ y]) := \boldsymbol{\phi}(x, y)$.

$$P(x, y|\theta) := P([x\ y]|\boldsymbol{\theta}) = \frac{h(x, y)}{Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

We observe only x. What is the posterior over y?

$$P(y|x, \theta) = \frac{P(x, y|\boldsymbol{\theta})}{P(x|\boldsymbol{\theta})} = \frac{h(x, y)}{P(x|\theta) Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

Missing data in exponential family distributions

Let samples from the exponential family have two parts: $[x\ y]$.

Feature vector $\boldsymbol{\phi}([x\ y]) := \boldsymbol{\phi}(x, y)$.

$$P(x, y|\theta) := P([x\ y]|\boldsymbol{\theta}) = \frac{h(x, y)}{Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

We observe only x. What is the posterior over y?

$$P(y|x, \theta) = \frac{P(x, y|\boldsymbol{\theta})}{P(x|\boldsymbol{\theta})} = \frac{h(x, y)}{P(x|\theta)Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

An exponential family distrib. over y (remember x is fixed) with:

- feature vector $\boldsymbol{\phi}_x(y) = \boldsymbol{\phi}(x, y)$
- base distribution $h_x(y) = h(x, y)$

Missing data in exponential family distributions

Let samples from the exponential family have two parts: $[x\ y]$.

Feature vector $\boldsymbol{\phi}([x\ y]) := \boldsymbol{\phi}(x, y)$.

$$P(x, y|\theta) := P([x\ y]|\boldsymbol{\theta}) = \frac{h(x, y)}{Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

We observe only x. What is the posterior over y?

$$P(y|x, \theta) = \frac{P(x, y|\boldsymbol{\theta})}{P(x|\boldsymbol{\theta})} = \frac{h(x, y)}{P(x|\theta) Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(x, y))$$

An exponential family distrib. over y (remember x is fixed) with:

- feature vector $\boldsymbol{\phi}_x(y) = \boldsymbol{\phi}(x, y)$
- base distribution $h_x(y) = h(x, y)$

Not necessarily easy to work with, but will restrict to this case.

y_i with different x_i belong to different exp. fam. distrbs.

What about the marginal likelihood $P(x|\boldsymbol{\theta}) = \int P(x, y|\boldsymbol{\theta})\mathrm{d}y$?

MLE with missing data

What about the marginal likelihood $P(x|\boldsymbol{\theta}) = \int P(x, y|\boldsymbol{\theta})\mathrm{d}y$?

- Not an exponential family distribution!
- Calculating derivatives is messy
- No nice closed form expression for MLE

MLE with missing data

What about the marginal likelihood $P(x|\boldsymbol{\theta}) = \int P(x, y|\boldsymbol{\theta})\mathrm{d}y$?

- Not an exponential family distribution!
- Calculating derivatives is messy
- No nice closed form expression for MLE

One approach: the EM algorithm.

An algorithm for MLE in exp. fam. distribs. with missing data

MLE with missing data

What about the marginal likelihood $P(x|\boldsymbol{\theta}) = \int P(x, y|\boldsymbol{\theta})\mathrm{d}y$?

- Not an exponential family distribution!
- Calculating derivatives is messy
- No nice closed form expression for MLE

One approach: the EM algorithm.

An algorithm for MLE in exp. fam. distribs. with missing data

Problem: Given observations i.i.d. $X = \{x_1, \ldots, x_N\}$ from $P(x|\boldsymbol{\theta})$ where $P(X, Y|\boldsymbol{\theta})$ is exponential family, maximize w.r.t. $\boldsymbol{\theta}$

$$\ell(X|\boldsymbol{\theta}) = \log P(X|\theta) = \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta})$$

MLE in latent variable models

$$\ell(X|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i$$

MLE in latent variable models

$$\ell(X|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i$$

$$= \sum_{i=1}^{N} \log \int q_i(y_i) \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(for arbitrary densities } q_i(y_i))$$

MLE in latent variable models

$$
\begin{aligned}
\ell(X|\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i \\
&= \sum_{i=1}^{N} \log \int q_i(y_i) \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(for arbitrary densities } q_i(y_i)\text{)} \\
&\geq \sum_{i=1}^{N} \int q_i(y_i) \log \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(Jensen's inequality)}
\end{aligned}
$$

MLE in latent variable models

$$
\begin{aligned}
\ell(X|\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i \\
&= \sum_{i=1}^{N} \log \int q_i(y_i) \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(for arbitrary densities } q_i(y_i)\text{)} \\
&\geq \sum_{i=1}^{N} \int q_i(y_i) \log \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(Jensen's inequality)} \\
&= \sum_{i=1}^{N} \int q_i(y_i) \log P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i - \sum_{i=1}^{N} \int q_i(y_i) \log q_i(y_i) \mathrm{d}y_i
\end{aligned}
$$

MLE in latent variable models

$$
\begin{aligned}
\ell(X|\boldsymbol{\theta}) &= \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i \\
&= \sum_{i=1}^{N} \log \int q_i(y_i) \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(for arbitrary densities } q_i(y_i)\text{)} \\
&\geq \sum_{i=1}^{N} \int q_i(y_i) \log \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(Jensen's inequality)} \\
&= \sum_{i=1}^{N} \int q_i(y_i) \log P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i - \sum_{i=1}^{N} \int q_i(y_i) \log q_i(y_i) \mathrm{d}y_i \\
&= \sum_{i=1}^{N} \int q_i(y_i) \log P(x_i|\boldsymbol{\theta}) \mathrm{d}y_i + \sum_{i=1}^{N} \int q_i(y_i) \log \frac{P(y_i|x_i, \boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i
\end{aligned}
$$

$$\ell(X|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log P(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \int P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i$$

$$= \sum_{i=1}^{N} \log \int q_i(y_i) \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(for arbitrary densities } q_i(y_i)\text{)}$$

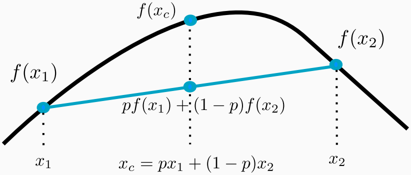
$$\geq \sum_{i=1}^{N} \int q_i(y_i) \log \frac{P(x_i, y_i|\boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i \quad \text{(Jensen's inequality)}$$

$$= \sum_{i=1}^{N} \int q_i(y_i) \log P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i - \sum_{i=1}^{N} \int q_i(y_i) \log q_i(y_i) \mathrm{d}y_i$$

$$= \sum_{i=1}^{N} \int q_i(y_i) \log P(x_i|\boldsymbol{\theta}) \mathrm{d}y_i + \sum_{i=1}^{N} \int q_i(y_i) \log \frac{P(y_i|x_i, \boldsymbol{\theta})}{q_i(y_i)} \mathrm{d}y_i$$

$$= \ell(X|\boldsymbol{\theta}) \quad - \quad \sum_{i=1}^{N} \mathrm{KL}(q_i(y_i) \| P(y_i|X, \boldsymbol{\theta}))$$

Let $f(x)$ be a concave real-valued function defined on $\mathbb{X}$.

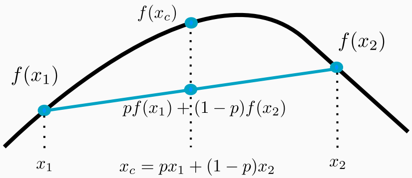


Concave: Non-positive 2nd-derivative (non-increasing deriv.)

A chord always lies below the function.

E.g. logarithm (defined on $\mathbb{R}^+$).

Let $f(x)$ be a concave real-valued function defined on $\mathbb{X}$.



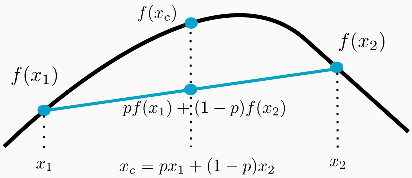
Concave: Non-positive 2nd-derivative (non-increasing deriv.)

A chord always lies below the function.

E.g. logarithm (defined on $\mathbb{R}^+$).

Jensen: for any prob. vector $p = (p_1, \ldots, p_K)$ and any set of points $(x_1, \ldots, x_K)$, $\quad f(\sum_{i=1}^K p_i x_i) \geq \sum_{i=1}^K p_i f(x_i)$

Let $f(x)$ be a concave real-valued function defined on $\mathbb{X}$.



Concave: Non-positive 2nd-derivative (non-increasing deriv.)
A chord always lies below the function.

E.g. logarithm (defined on $\mathbb{R}^+$).

Jensen: for any prob. vector $p = (p_1, \ldots, p_K)$ and any set of points $(x_1, \ldots, x_K)$, $\quad f(\sum_{i=1}^K p_i x_i) \geq \sum_{i=1}^K p_i f(x_i)$

In fact, for a prob. density $p(x)$, $f(\int_{\mathbb{X}} xp(x)\mathrm{d}x) \geq \int_{\mathbb{X}} f(x)p(x)\mathrm{d}x$

Defining $Q(Y) = \prod_{i=1}^{N} q_i(y_i)$,

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}) &\geq \sum_{i=1}^{N} \int q_i(y_i) \log P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i + \sum_{i=1}^{N} H(q_i) \\
&= \sum_{i=1}^{N} \mathbb{E}_{q_i}[\log P(x_i, y_i|\boldsymbol{\theta})] + \sum_{i=1}^{N} H(q_i) \\
&= \ell(X|\boldsymbol{\theta}) - \sum_{i=1}^{N} \mathrm{KL}(q_i(y_i) \| P(y_i|X, \boldsymbol{\theta})) \\
&:= \mathcal{F}_X(\boldsymbol{\theta}, Q(\cdot))
\end{aligned}$$

MLE in latent variable models

Defining $Q(Y) = \prod_{i=1}^{N} q_i(y_i)$,

$$
\begin{aligned}
\ell(X|\boldsymbol{\theta}) &\geq \sum_{i=1}^{N} \int q_i(y_i) \log P(x_i, y_i|\boldsymbol{\theta}) \mathrm{d}y_i + \sum_{i=1}^{N} H(q_i) \\
&= \sum_{i=1}^{N} \mathbb{E}_{q_i}[\log P(x_i, y_i|\boldsymbol{\theta})] + \sum_{i=1}^{N} H(q_i) \\
&= \ell(X|\boldsymbol{\theta}) - \sum_{i=1}^{N} \mathrm{KL}(q_i(y_i) \| P(y_i|X, \boldsymbol{\theta})) \\
&:= \mathcal{F}_X(\boldsymbol{\theta}, Q(\cdot))
\end{aligned}
$$

$\mathcal{F}_X(\boldsymbol{\theta}, Q(\cdot))$ is a lower bound to the log-likelihood $\ell(X|\boldsymbol{\theta})$. Sometimes called 'variational free energy' and is function of $\boldsymbol{\theta}$ and the 'variational distribution' $Q(Y)$ (X is fixed).

Our original goal was to maximize the log-likelihood:

$$\boldsymbol{\theta}_{MLE} = \text{argmax } \ell(X|\theta)$$

EM algorithm: maximize the lower-bound instead

$$(\boldsymbol{\theta}^*, Q^*) = \text{argmax } \mathcal{F}_X(\theta, Q(\cdot))$$

Hopefully easier, since all summations are outside logarithms.

Strategy: Coordinate ascent.

Alternately maximize w.r.t Q and θ

First find best lower-bound given the current $\boldsymbol{\theta}_s$.

Optimize this lower-bound to find $\boldsymbol{\theta}_{s+1}$.

Maximizing $\mathcal{F}_X(\boldsymbol{\theta}, Q)$ with $\boldsymbol{\theta}$ fixed:

- Recall $\mathcal{F}_X(\boldsymbol{\theta}, Q) = \ell(X|\boldsymbol{\theta}) - \sum_{i=1}^{N} \mathrm{KL}(q_i(y_i) \| P(y_i|x_i, \boldsymbol{\theta}))$

Maximizing $\mathcal{F}_X(\boldsymbol{\theta}, Q)$ with $\boldsymbol{\theta}$ fixed:

- Recall $\mathcal{F}_X(\boldsymbol{\theta}, Q) = \ell(X|\boldsymbol{\theta}) - \sum_{i=1}^{N} \mathrm{KL}(q_i(y_i) \| P(y_i|x_i, \boldsymbol{\theta}))$
- Solution: set $q_i(y_i) = P(y_i|x_i, \boldsymbol{\theta}) \qquad \text{for} \;\; i = 1, \ldots, N$
- Recall: $P(\cdot|x_i, \boldsymbol{\theta})$ is an exponential family distribution with natural parameters $\boldsymbol{\theta}$ and feature vector $\boldsymbol{\phi}(x_i, \cdot)$

Maximizing $\mathcal{F}_X(\boldsymbol{\theta}, Q)$ with Q fixed:

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \sum_{i=1}^{N} \mathbb{E}_{q_i}[\log P(x_i, y_i|\boldsymbol{\theta})] + \sum_{i=1}^{N} H(q_i)$$

The entropy terms $H(q_i)$ don't depend on $\boldsymbol{\theta}$. Ignore.

Maximizing $\mathcal{F}_X(\boldsymbol{\theta}, Q)$ with Q fixed:

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \sum_{i=1}^{N} \mathbb{E}_{q_i}[\log P(x_i, y_i|\boldsymbol{\theta})] + \sum_{i=1}^{N} H(q_i)$$

The entropy terms $H(q_i)$ don't depend on $\boldsymbol{\theta}$. Ignore.

$$\log P(x_i, y_i|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \phi(x_i, y_i) + \log h(x_i) - \log Z(\boldsymbol{\theta})$$

Maximizing $\mathcal{F}_X(\boldsymbol{\theta}, Q)$ with Q fixed:

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \sum_{i=1}^{N} \mathbb{E}_{q_i}[\log P(x_i, y_i|\boldsymbol{\theta})] + \sum_{i=1}^{N} H(q_i)$$

The entropy terms $H(q_i)$ don't depend on $\boldsymbol{\theta}$. Ignore.

$$\log P(x_i, y_i|\boldsymbol{\theta}) = \boldsymbol{\theta}^\top \phi(x_i, y_i) + \log h(x_i) - \log Z(\boldsymbol{\theta})$$

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \sum_{i=1}^{N} \boldsymbol{\theta}^\top \mathbb{E}_{q_i}[\phi(x_i, y_i)] - N \log Z(\boldsymbol{\theta}) + const$$

The Expectation-Maximization algorithm

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \boldsymbol{\theta}^\top \sum_{i=1}^{N} \mathbb{E}_{q_i}[\phi(x_i, y_i)] - N \log Z(\boldsymbol{\theta}) + \mathit{const}$$

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \boldsymbol{\theta}^\top \sum_{i=1}^{N} \mathbb{E}_{q_i}[\boldsymbol{\phi}(x_i, y_i)] - N \log Z(\boldsymbol{\theta}) + const$$

To maximize $\mathcal{F}_X$ w.r.t. θ_d, solve $\frac{\partial}{\partial \theta_d} \mathcal{F}_X(\boldsymbol{\theta}, Q) = 0$:

- Solution: set θ^* to match moments (compare w. fully observed case)

  $\mathbb{E}_{\theta^*}[\phi_d(x, y)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q_i}[\boldsymbol{\phi}_d(x_i, y_i)]$

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \boldsymbol{\theta}^\top \sum_{i=1}^{N} \mathbb{E}_{q_i}[\boldsymbol{\phi}(x_i, y_i)] - N \log Z(\boldsymbol{\theta}) + \mathit{const}$$

To maximize $\mathcal{F}_X$ w.r.t. θ_d, solve $\frac{\partial}{\partial \theta_d}\mathcal{F}_X(\boldsymbol{\theta}, Q) = 0$:

- Solution: set θ^* to match moments (compare w. fully observed case)

$$\mathbb{E}_{\theta^*}[\phi_d(x, y)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q_i}[\boldsymbol{\phi}_d(x_i, y_i)]$$

The Expectation-Maximization algorithm

$$\mathcal{F}_X(\boldsymbol{\theta}, Q) = \boldsymbol{\theta}^\top \sum_{i=1}^{N} \mathbb{E}_{q_i}[\boldsymbol{\phi}(x_i, y_i)] - N \log Z(\boldsymbol{\theta}) + const$$

To maximize $\mathcal{F}_X$ w.r.t. θ_d, solve $\frac{\partial}{\partial \theta_d} \mathcal{F}_X(\boldsymbol{\theta}, Q) = 0$:

- Solution: set θ^* to match moments (compare w. fully observed case)

$$\mathbb{E}_{\theta^*}[\phi_d(x, y)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{q_i}[\boldsymbol{\phi}_d(x_i, y_i)]$$

$q_i(y_i) = P(y_i | X, \boldsymbol{\theta}^{old})$, an exponential family distribution whose moment parameters can be calculated (by assumption).

Step s of the EM algorithm

Current parameters: $\boldsymbol{\theta}^s, Q^s(Y) = \prod_{i=1}^N q_i^s(y_i)$

Step s of the EM algorithm

Current parameters: $\boldsymbol{\theta}^s, Q^s(Y) = \prod_{i=1}^N q_i^s(y_i)$
E-step:
For $i = 1, \ldots, N$:

- Set $q_i^{s+1}(y_i) = P(y_i|X, \boldsymbol{\theta}^s)$.
  Exp. fam. distrib. with suff. stats $\boldsymbol{\phi}(x_i, \cdot)$, natural params $\boldsymbol{\theta}^s$
- Calculate $\mathbb{E}_{q_i^{s+1}}[\boldsymbol{\phi}(x_i, y_i)]$ (Expectation)

Current parameters: $\boldsymbol{\theta}^s, Q^s(Y) = \prod_{i=1}^N q_i^s(y_i)$

E-step:

For $i = 1, \ldots, N$:

- Set $q_i^{s+1}(y_i) = P(y_i|X, \boldsymbol{\theta}^s)$.
  Exp. fam. distrib. with suff. stats $\boldsymbol{\phi}(x_i, \cdot)$, natural params $\boldsymbol{\theta}^s$
- Calculate $\mathbb{E}_{q_i^{s+1}}[\boldsymbol{\phi}(x_i, y_i)]$ (Expectation)

M-step (Maximization):

- Set $\boldsymbol{\theta}^{s+1}$ so that $\mathbb{E}_{\boldsymbol{\theta}^{s+1}}[\phi(x, y)] = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{q_i^s}[\phi(x_i, y_i)]$.

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y) \| P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\ell(X|\boldsymbol{\theta}^{s-1}) = \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s)$$

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}^{s-1}) &= \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^s)
\end{aligned}$$

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}^{s-1}) &= \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^{s+1})
\end{aligned}$$

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}^{s-1}) &= \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^{s+1}) \\
&= \ell(X|\boldsymbol{\theta}^{s})
\end{aligned}$$

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}^{s-1}) &= \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^{s+1}) \\
&= \ell(X|\boldsymbol{\theta}^{s})
\end{aligned}$$

Can also show that local maxima of $\mathcal{F}_X$ are local maxima of ℓ.

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}^{s-1}) &= \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^{s+1}) \\
&= \ell(X|\boldsymbol{\theta}^{s})
\end{aligned}$$

Can also show that local maxima of $\mathcal{F}_X$ are local maxima of ℓ.

Variants of E and M-steps.

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}^{s-1}) &= \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^{s+1}) \\
&= \ell(X|\boldsymbol{\theta}^{s})
\end{aligned}$$

Can also show that local maxima of $\mathcal{F}_X$ are local maxima of ℓ.

Variants of E and M-steps.

Partial M-step: Update $\boldsymbol{\theta}$ to increase (rather than maximize) $\mathcal{F}_X$

EM algorithm never decreases log-likelihood

Recall $\mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^{s-1}) = \ell(X|\boldsymbol{\theta}^{s-1}) - \mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta}^{s-1}))$.
After the E-step, $Q^s(Y) = P(Y|X, \boldsymbol{\theta}^{s-1}))$

$$\begin{aligned}
\ell(X|\boldsymbol{\theta}^{s-1}) &= \mathcal{F}_X(\boldsymbol{\theta}^{s-1}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^s) \\
&\leq \mathcal{F}_X(\boldsymbol{\theta}^{s}, Q^{s+1}) \\
&= \ell(X|\boldsymbol{\theta}^{s})
\end{aligned}$$

Can also show that local maxima of $\mathcal{F}_X$ are local maxima of ℓ.

Variants of E and M-steps.

Partial M-step: Update $\boldsymbol{\theta}$ to increase (rather than maximize) $\mathcal{F}_X$
Partial E-step: Update Q to decrease $\mathrm{KL}(Q(Y)\|P(Y|X, \boldsymbol{\theta})$ (rather reduce to 0). E.g. update just one or a few q_i.

A toy example:

A mixture of two Gaussians, $\mathcal{N}(x|m, 1)$ and $\mathcal{N}(x|5 - m, 2)$.
First has probability 0.6, the second 0.4.

A toy example:

A mixture of two Gaussians, $\mathcal{N}(x|m, 1)$ and $\mathcal{N}(x|5 - m, 2)$.
First has probability 0.6, the second 0.4.

We observe a single data point $x = 1$.
What is the ML estimate of m?

A toy example:

A mixture of two Gaussians, $\mathcal{N}(x|m, 1)$ and $\mathcal{N}(x|5 - m, 2)$.
First has probability 0.6, the second 0.4.

We observe a single data point $x = 1$.
What is the ML estimate of m?

What is the hidden variable?

A toy example:

A mixture of two Gaussians, $\mathcal{N}(x|m, 1)$ and $\mathcal{N}(x|5-m, 2)$.
First has probability 0.6, the second 0.4.

We observe a single data point $x = 1$.
What is the ML estimate of m?

What is the hidden variable?
Is the overall model exponential family?

A toy example:

A mixture of two Gaussians, $\mathcal{N}(x|m,1)$ and $\mathcal{N}(x|5-m,2)$.
First has probability 0.6, the second 0.4.

We observe a single data point $x = 1$.
What is the ML estimate of m?

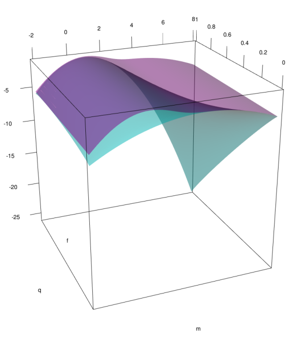
What is the hidden variable?
Is the overall model exponential family?
What is the posterior distribution over the hidden variable?

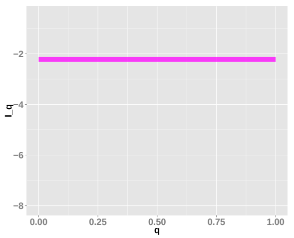
A mixture of two Gaussians, $\mathcal{N}(x|m, 1)$ and $\mathcal{N}(x|5-m, 2)$.
First has probability 0.6, the second 0.4.

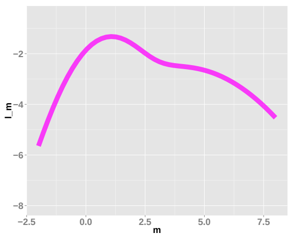
We observe a single data point $x = 1$.
What is the ML estimate of m?

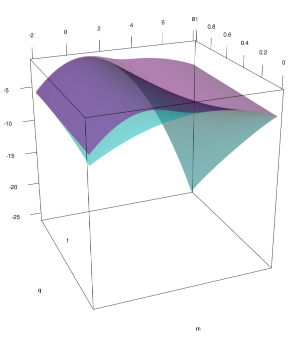
What is the hidden variable?
Is the overall model exponential family?
What is the posterior distribution over the hidden variable?
If we knew the hidden variable, what is the MLE?

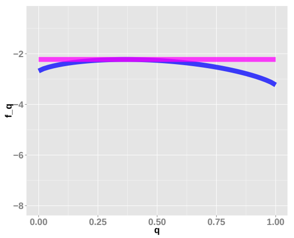
The EM algorithm

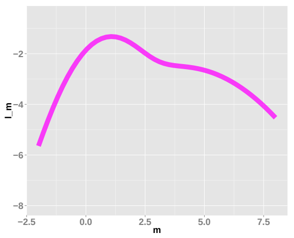




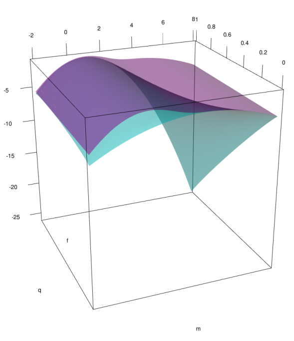


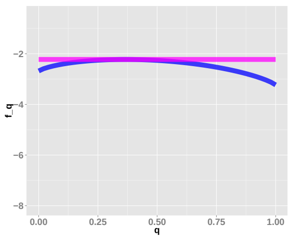


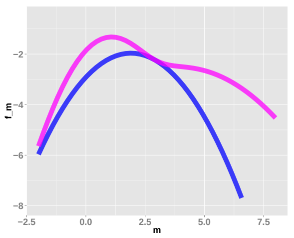




Initialize $m = 2.9$.

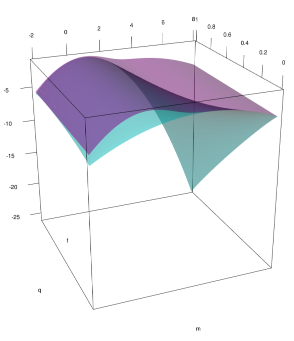


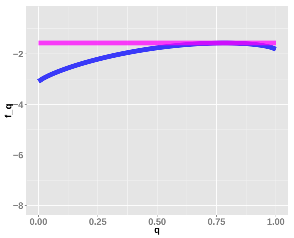


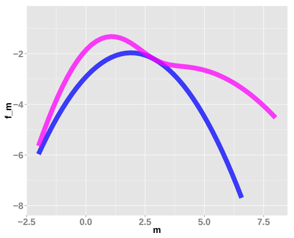


Initialize $m = 2.9$.

Set $q = 0.37$.





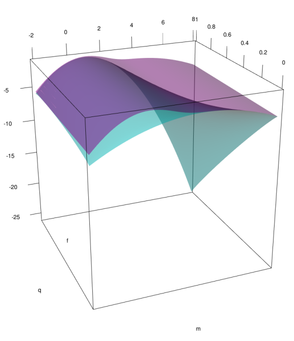


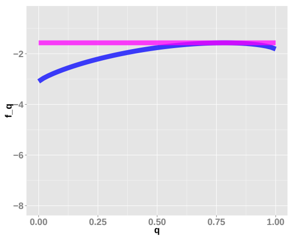
Initialize $m = 2.9$.

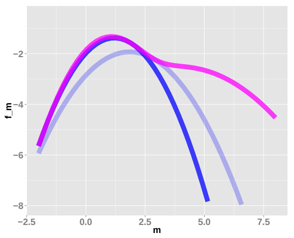
Set $q = 0.37$.

Set $m = 1.88$.

The EM algorithm







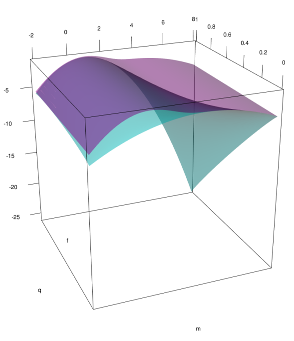
Initialize $m = 2.9$.

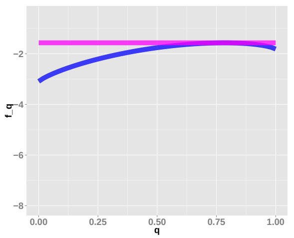
Set $q = 0.37$.

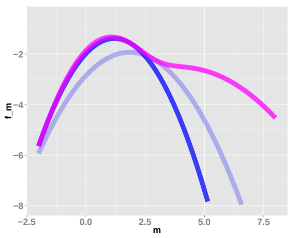
Set $m = 1.88$.

Set $q = 0.775$.

The EM algorithm







Initialize $m = 2.9$.

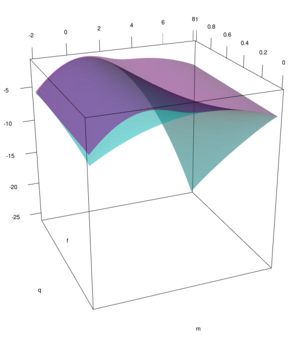
Set $q = 0.37$.

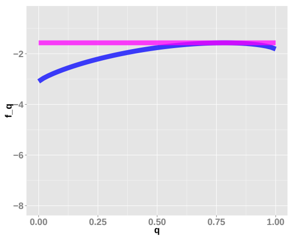
Set $m = 1.88$.

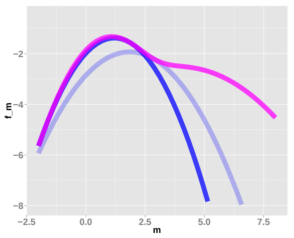
Set $q = 0.775$.

Set $m = 1.265$.

The EM algorithm







Initialize $m = 2.9$.

Set $q = 0.37$.

Set $m = 1.88$.

Set $q = 0.775$.

Set $m = 1.265$.

Repeat till convergence