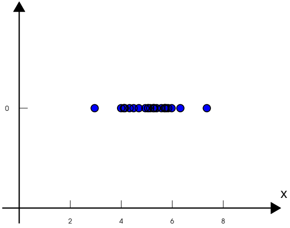
Expectation Maximization without tears!

Josephine Sullivan + the web

Some stuff you probably already know

Parameter estimation

Have n independent draws $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $p(\mathbf{x} \mid \Theta)$.

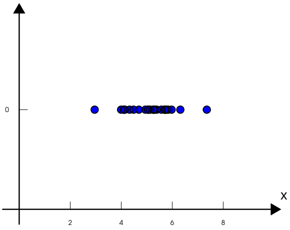


$\longleftarrow$ **1D example**

Each $\mathbf{x}_i \sim N(\mathbf{x} \mid \boldsymbol{\mu}, \Sigma)$ where $\Theta = (\boldsymbol{\mu}, \Sigma)$

Parameter estimation

Have n independent draws $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $p(\mathbf{x} \,|\, \Theta)$.



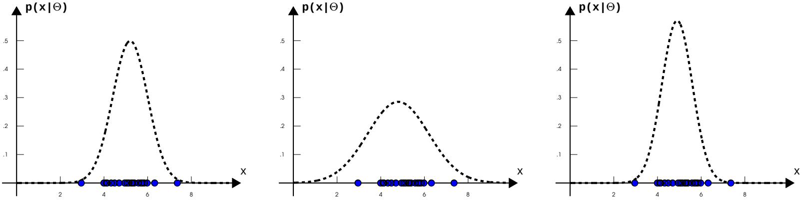
$\longleftarrow$ **1D example**

$$\text{Each} \quad \mathbf{x}_i \sim N(\mathbf{x} \,|\, \boldsymbol{\mu}, \Sigma) \quad \text{where } \Theta = (\boldsymbol{\mu}, \Sigma)$$

Want to estimate the parameters Θ from the $\mathbf{x}_i$'s

Parameter estimation

Have n independent draws $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $p(\mathbf{x} \mid \Theta)$.



$\Theta = (5.2, .8)$ $\Theta = (4.8, 1.4)$ $\Theta = (4.9, .7)$

Want to estimate the parameters Θ from the $\mathbf{x}_i$'s.

HOW??

Maximum Likelihood Estimation (MLE)

Choose the Θ which maximizes the likelihood of your data:

$$\Theta^* = \arg\max_{\Theta} p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \Theta)$$

Maximum Likelihood Estimation (MLE)

Choose the Θ which maximizes the likelihood of your data:

$$\begin{aligned} l(\Theta; \mathbf{X}) &\equiv p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \Theta) \\ &= \prod_{i=1}^{n} p(\mathbf{x}_i \mid \Theta) \quad \leftarrow \text{assuming independent samples} \end{aligned}$$

Maximum Likelihood Estimation (MLE)

Choose the Θ which maximizes the likelihood of your data:

$$l(\Theta; \mathbf{X}) \equiv p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \mid \Theta)$$
$$= \prod_{i=1}^{n} p(\mathbf{x}_i \mid \Theta) \quad \leftarrow \text{assuming independent samples}$$

Easier to work with the log-likelihood

$$L(\Theta; \mathbf{X}) = \log\left(l(\Theta; \mathbf{X})\right) = \sum_{i=1}^{n} \log\left(p(\mathbf{x}_i \mid \Theta)\right)$$

Maximum Likelihood Estimation (MLE)

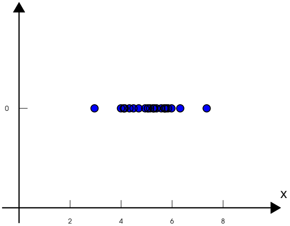
Choose the Θ which maximizes the likelihood of your data:

Note

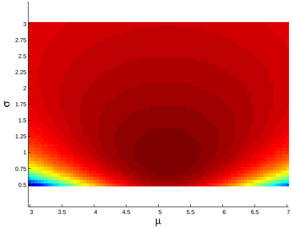
$$\Theta^* = \arg\max_{\Theta} l(\Theta; \mathbf{X}) = \arg\max_{\Theta} L(\Theta; \mathbf{X})$$

An example Log-likelihood function

Our 1D example of points drawn from $N(\mu, \Sigma)$



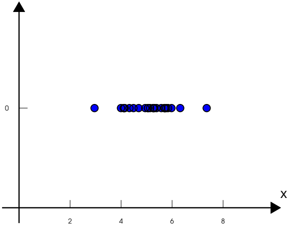
$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$



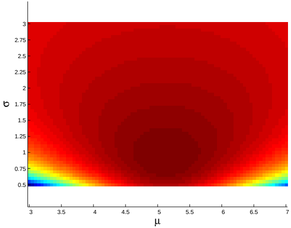
Log-likelihood: $L(\Theta; \mathbf{X})$

An example Log-likelihood function

Our 1D example of points drawn from $N(\mu, \Sigma)$



$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$



Log-likelihood: $L(\Theta; \mathbf{X})$

Want to find the maximum of this function $L(\Theta; \mathbf{X})$.

MLE for a Normal distribution

The formula for a normal distribution for $\mathbf{x} \in \mathcal{R}^d$:

$$p(\mathbf{x} \mid \Theta) = (2\pi)^{-\frac{d}{2}} \, |\Sigma|^{-\frac{1}{2}} \; \exp\left(-.5(\mathbf{x} - \boldsymbol{\mu})^t \, \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

MLE for a Normal distribution

The formula for a normal distribution for $\mathbf{x} \in \mathcal{R}^d$:

$$p(\mathbf{x} \mid \Theta) = (2\pi)^{-\frac{d}{2}} \, |\Sigma|^{-\frac{1}{2}} \, \exp\left(-.5(\mathbf{x} - \boldsymbol{\mu})^t \, \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

The log-likelihood of our n data-points is

$$\begin{aligned} L(\Theta; \mathbf{X}) &= \sum_{i=1}^{n} \log\left(p(\mathbf{x}_i \mid \Theta)\right) \\ &= \sum_{i=1}^{n} \left[-\frac{d}{2} \log(2\pi) - \frac{1}{2} \log\left(|\Sigma|\right) - .5(\mathbf{x}_i - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \\ &= -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log\left(|\Sigma|\right) - .5 \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\ &= -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log\left(|\Sigma|\right) - .5 \, \mathrm{tr}\left[\sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})^t \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right] \end{aligned}$$

MLE for a Normal distribution

$$
\begin{aligned}
L(\Theta; \mathbf{X}) &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\left(|\Sigma|\right) - .5\,\mathrm{tr}\left[\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right] \\
&= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\left(|\Sigma|\right) - .5\,\mathrm{tr}\left[\sum_{i=1}^{n}\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right] \\
&= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\left(|\Sigma|\right) - .5\,\mathrm{tr}\left[\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right]
\end{aligned}
$$

Note Σ is a symmetric positive definite matrix. Thus $\Sigma = T^t T$ therefore

$$
\begin{aligned}
L(\Theta; \mathbf{X}) &= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\left(\left|T^t T\right|\right) - .5\,\mathrm{tr}\left[(T^t T)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right] \\
&= -\frac{nd}{2}\log(2\pi) - n\log\left(|T|\right) - .5\,\mathrm{tr}\left[(T^t T)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right]
\end{aligned}
$$

Remember

How do we analytically solve for an optimum?

- Take derivative of function wrt each variable.

Remember

How do we analytically solve for an optimum?

- Take derivative of function wrt each variable.
- Set each derivative to zero.

Remember

How do we analytically solve for an optimum?

- Take derivative of function wrt each variable.
- Set each derivative to zero.
- Solve the set of simultaneous equations if possible.

MLE for a Normal distribution

For our Normal distribution

$$L(\Theta; \mathbf{X}) = -\frac{nd}{2}\log(2\pi) - n\log\left(|T|\right) - .5\,\mathrm{tr}\left[(T^t T)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right]$$

Take derivative of function wrt each variable:

$$\frac{\partial\, L(\Theta; \mathbf{X})}{\partial\, \boldsymbol{\mu}} = \sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu})$$

$$\frac{\partial\, L(\Theta; \mathbf{X})}{\partial\, T} = -nT^{-t} + T(T^t T)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t(T^t T)^{-1}$$

Remember: *The Matrix Cookbook* is your friend.

MLE for a Normal distribution

For our Normal distribution

$$L(\Theta; \mathbf{X}) = -\frac{nd}{2}\log(2\pi) - n\log(|T|) - .5\,\mathrm{tr}\left[(T^t T)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right]$$

Set each derivative to zero:

$$\mathbf{0} = \Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})$$

$$\mathbf{0} = -nT^{-t} + T(T^t T)^{-1}\left[\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right](T^t T)^{-1}$$

Remember: *The Matrix Cookbook* is your friend.

MLE for a Normal distribution

For our Normal distribution

$$L(\Theta; \mathbf{X}) = -\frac{nd}{2}\log(2\pi) - n\log(|T|) - .5\,\mathrm{tr}\left[(T^t T)^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^t\right]$$

Solve the set of simultaneous equations if possible:

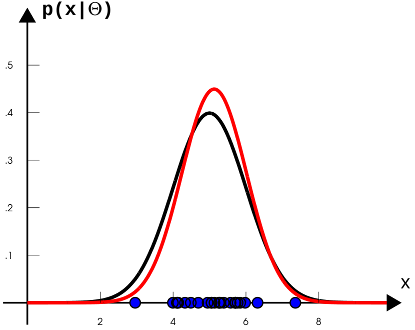
$$\boldsymbol{\mu}^* = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

$$T^{*t}T^* = \Sigma^* = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}^*)(\mathbf{x}_i - \boldsymbol{\mu}^*)^t$$

Remember: *The Matrix Cookbook* is your friend.

MLE for a Normal distribution

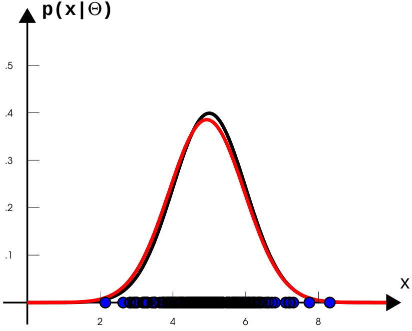
Back to our 1D example:



Red curve is the MLE pdf ($n = 25$)
Black curve is the ground truth

MLE for a Normal distribution

Estimate becomes better as n increases



Red curve is the MLE pdf ($n = 200$)
Black curve is the ground truth

Some more stuff you probably already know

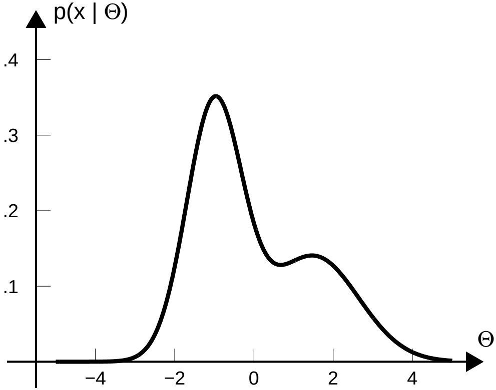
Limitations of Normal distributions

Unfortunately Normal distributions are not very expressive.

They can only accurately represent distributions with one mode.

Limitations of Normal distributions

Unfortunately Normal distributions are not very expressive.



What do we do in this situation ??

Gaussian Mixture Models (GMM)

They can accurately represent any distribution.

Mathematical definition

$$p(\mathbf{x} \mid \Theta) = \sum_{k=1}^{K} \pi_k \, N(\mathbf{x}_k; \boldsymbol{\mu}_k, \Sigma_k)$$

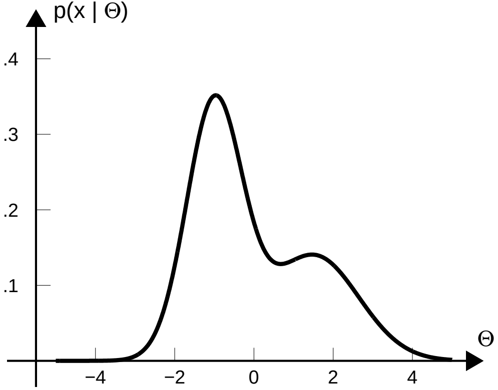
where

$$\sum_{k=1}^{K} \pi_k = 1 \quad \text{and} \quad \pi_k \geq 0 \text{ for } k = 1, \ldots, K$$

and $\Theta = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K, \pi_1, \ldots, \pi_K)$

Gaussian Mixture Models (GMM)

They can accurately represent any distribution.

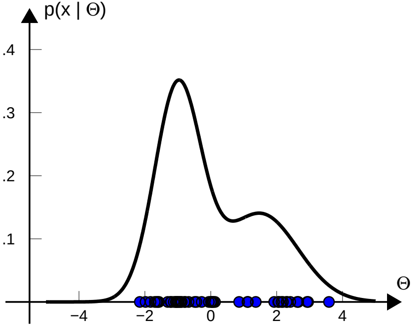


$$p(x \mid \Theta) = \alpha \mathcal{N}(x \mid \mu_1, \sigma_1^2) + (1 - \alpha) \mathcal{N}(x \mid \mu_2, \sigma_2^2)$$

$$\Theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2) = (.6, -1, .5, 1.5, 1.3)$$

Parameter estimation for a GMM

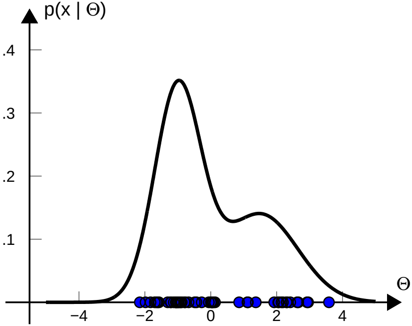
Given n independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a GMM.



$\longleftarrow$ **training data**

Parameter estimation for a GMM

Given n independent samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from a GMM.



$\longleftarrow$ **training data**

Can still use MLE to estimate Θ from the $\mathbf{x}_i$'s, but...

Attempt 1: Analytic Solution

Attempt 1: Parameter estimation for a GMM

The log-likelihood of the data is

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, \Sigma_k) \right)$$

(Note: We'll assume K is known and fixed.)

Attempt 1: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Let's try to maximize $L(\Theta; \mathbf{X})$ analytically subject to the constraint $\sum_k \pi_k = 1$ and each $\Sigma_k = T_k^t T_k$. Construct the Lagrangian $\mathcal{L}(\Theta, \lambda; \mathbf{X})$.

$$\mathcal{L}(\Theta, \lambda; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right) + \lambda \left(1 - \sum_{k=1}^{K} \pi_k \right)$$

Attempt 1: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Let's try to maximize $L(\Theta; \mathbf{X})$ analytically subject to the constraint $\sum_k \pi_k = 1$ and each $\Sigma_k = T_k^t T_k$. Construct the Lagrangian $\mathcal{L}(\Theta, \lambda; \mathbf{X})$.

Take derivatives for $k = 1, \ldots, K$:

$$\frac{\partial\, \mathcal{L}(\Theta, \lambda; \mathbf{X})}{\partial\, \boldsymbol{\mu}_k} = \sum_{i=1}^{n} \frac{\pi_k N(\mathbf{x}_i; \boldsymbol{\mu}_k, T_k^t T_k)}{GMM(\mathbf{x}_i; \Theta)} \; (T_k^t T_k)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)$$

$$\frac{\partial\, \mathcal{L}(\Theta, \lambda; \mathbf{X})}{\partial\, T_k} = \text{something complicated.....}$$

etc

Attempt 1: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Let's try to maximize $L(\Theta; \mathbf{X})$ analytically subject to the constraint $\sum_k \pi_k = 1$ and each $\Sigma_k = T_k^t T_k$. Construct the Lagrangian $\mathcal{L}(\Theta, \lambda; \mathbf{X})$.

Set derivatives to zero:

$$\sum_{i=1}^{n} \frac{\pi_k N(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)}{GMM(\mathbf{x}_i; \Theta)} \Sigma_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) = \mathbf{0}$$

etc

Attempt 1: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Let's try to maximize $L(\Theta; \mathbf{X})$ analytically subject to the constraint $\sum_k \pi_k = 1$ and each $\Sigma_k = T_k^t T_k$. Construct the Lagrangian $\mathcal{L}(\Theta, \lambda; \mathbf{X})$.

Solve the set of simultaneous equations

NO ANALYTIC SOLUTION

Attempt 2: Newton based iterative optimzation

Attempt 2: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Could try to maximize $L(\Theta; \mathbf{X})$ iteratively using Newton's Method.

After all $L(\Theta; \mathbf{X})$ is a **scalar valued function** of a **vector** Θ of **variables**.

Attempt 2: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Could try to maximize $L(\Theta; \mathbf{X})$ iteratively using Newton's Method.

After all $L(\Theta; \mathbf{X})$ is a **scalar valued function** of a **vector** Θ of **variables**.

One iteration

- Have a current estimate $\Theta^{(t)}$.

Attempt 2: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Could try to maximize $L(\Theta; \mathbf{X})$ iteratively using Newton's Method.

After all $L(\Theta; \mathbf{X})$ is a **scalar valued function** of a **vector** Θ of **variables**.

One iteration

- Have a current estimate $\Theta^{(t)}$.
- Approximate $L(\Theta; \mathbf{X})$ in neighbourhood of $\Theta^{(t)}$ with a paraboloid.

Attempt 2: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Could try to maximize $L(\Theta; \mathbf{X})$ iteratively using Newton's Method.

After all $L(\Theta; \mathbf{X})$ is a **scalar valued function** of a **vector** Θ of **variables**.

One iteration

- Have a current estimate $\Theta^{(t)}$.
- Approximate $L(\Theta; \mathbf{X})$ in neighbourhood of $\Theta^{(t)}$ with a paraboloid.
- $\Theta^{(t+1)}$ is set to maximum of the paraboloid.

Attempt 2: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Could try to maximize $L(\Theta; \mathbf{X})$ iteratively using Newton's Method.

After all $L(\Theta; \mathbf{X})$ is a **scalar valued function** of a **vector** Θ of **variables**.

Comments

- Should find a local maximum. ✓

Attempt 2: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k) \right)$$

Could try to maximize $L(\Theta; \mathbf{X})$ iteratively using Newton's Method.

After all $L(\Theta; \mathbf{X})$ is a **scalar valued function** of a **vector** Θ of **variables**.

Comments

- Should find a local maximum. ✓
- Convergence fast if $\Theta^{(t)}$ close to an optimum. ✓

Attempt 2: Parameter estimation for a GMM

$$L(\Theta; \mathbf{X}) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k N(x_i; \boldsymbol{\mu}_k, T_k^t T_k)\right)$$

Could try to maximize $L(\Theta; \mathbf{X})$ iteratively using Newton's Method.

After all $L(\Theta; \mathbf{X})$ is a **scalar valued function** of a **vector** Θ of **variables**.

Comments

- Should find a local maximum. ✓
- Convergence fast if $\Theta^{(t)}$ close to an optimum. ✓
- If $\Theta^{(0)}$ far away from a local maximum method can fail. Paraboloid approximation process can hit problems. ✗

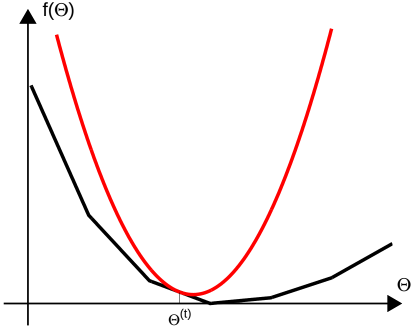
What other options are there??

Now for, what may seem like, a slight diversion

Defintion of Majorization

A function $g(\Theta\,;\,\Theta^{(t)})$ majorizes a function $f(\Theta)$ at $\Theta^{(t)}$ if

$f(\Theta^{(t)}) = g(\Theta^{(t)}\,;\,\Theta^{(t)})$ **and** $f(\Theta) \leq g(\Theta\,;\,\Theta^{(t)})$ for all Θ



$\longleftarrow$ $g(\Theta;\,\Theta^{(t)})$ majorizes $f(\Theta)$

The MM Algorithm

To **minimize** an objective function $f(\Theta)$:

- The MM algorithm is a prescription for constructing optimization algorithms.

Name coined by *David R. Hunter* and *Kenneth Lange*

The MM Algorithm

To **minimize** an objective function $f(\Theta)$:

- The MM algorithm is a prescription for constructing optimization algorithms.
- An MM algorithm creates a surrogate function that majorizes the **objective function**. When the surrogate function is minimized the **objective** function is decreased.

Name coined by *David R. Hunter* and *Kenneth Lange*

The MM Algorithm

To **minimize** an objective function $f(\Theta)$:

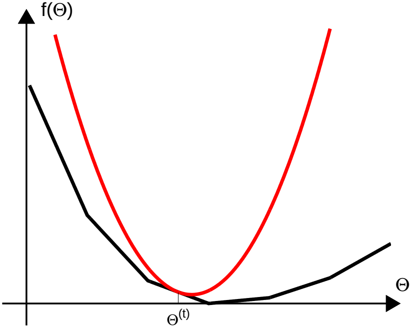
- The MM algorithm is a prescription for constructing optimization algorithms.
- An MM algorithm creates a surrogate function that majorizes the **objective function**. When the surrogate function is minimized the **objective** function is decreased.
- When minimizing MM $\equiv$ majorize/minimize.

Name coined by *David R. Hunter* and *Kenneth Lange*

Some definitions

A function $g(\Theta;\, \Theta^{(t)})$ majorizes the function $f(\Theta)$ at $\Theta^{(t)}$ if

$f(\Theta^{(t)}) = g(\Theta^{(t)};\, \Theta^{(t)})$ **and** $f(\Theta) \leq g(\Theta;\, \Theta^{(t)})$ for all Θ

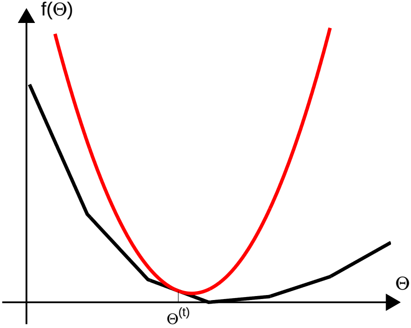


$\longleftarrow$ $g(\Theta; \Theta^{(t)})$ majorizes $f(\Theta)$

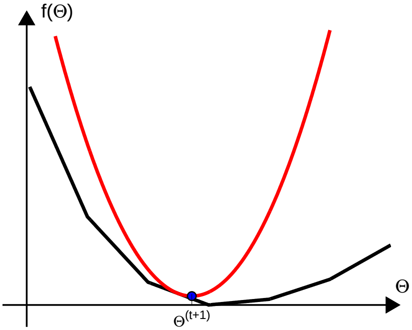
Some definitions

Let

$$\Theta^{(t+1)} = \arg\min_{\Theta} g(\Theta; \Theta^{(t)})$$



Majorize function



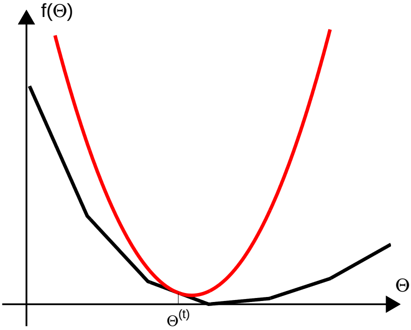
Find minimum of majorizing function

Some definitions

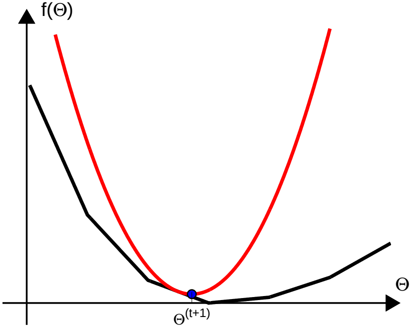
Let

$$\Theta^{(t+1)} = \arg\min_{\Theta} g(\Theta; \Theta^{(t)})$$

(so should choose a $g(\Theta; \Theta^{(t)})$ which is easy to minimize)



Majorize function



Find minimum of majorizing function

Descent Properties

MM minimization algorithm satisfies the descent property as

$$
\begin{aligned}
f(\Theta^{(t+1)}) &\leq g(\Theta^{(t+1)};\, \Theta^{(t)}), \quad \text{as } f(\Theta) \leq g(\Theta;\, \Theta^{(t)})\ \forall \Theta \\
&\leq g(\Theta^{(t)};\, \Theta^{(t)}), \quad \text{as } \Theta^{(t+1)} \text{ minimizes } g(\Theta;\, \Theta^{(t)}) \\
&= f(\Theta^{(t)})
\end{aligned}
$$

In summary

$$
f(\Theta^{(t+1)}) \leq f(\Theta^{(t)})
$$

Descent Properties

MM minimization algorithm satisfies the descent property as

$$\begin{aligned} f(\Theta^{(t+1)}) &\leq g(\Theta^{(t+1)}; \Theta^{(t)}), \quad \text{as } f(\Theta) \leq g(\Theta; \Theta^{(t)}) \ \forall \Theta \\ &\leq g(\Theta^{(t)}; \Theta^{(t)}), \quad \text{as } \Theta^{(t+1)} \text{ minimizes } g(\Theta; \Theta^{(t)}) \\ &= f(\Theta^{(t)}) \end{aligned}$$

In summary

$$f(\Theta^{(t+1)}) \leq f(\Theta^{(t)})$$

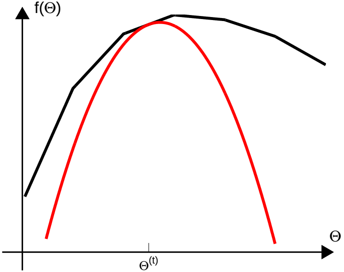
The descent property makes the MM algorithm very stable.

Algorithm converges to local minima or saddle point.

Maximizing a function

To **maximize** an objective function $f(\Theta)$:

- MM algorithm creates a surrogate function that minorize the **objective function**. When the surrogate function is maximized the **objective** function is **increased**.

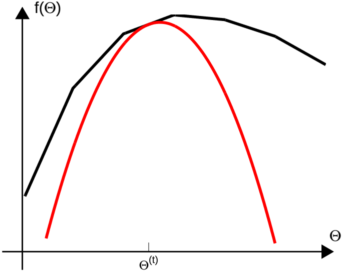


Red curve minorize the black curve

Maximizing a function

To **maximize** an objective function $f(\Theta)$:

- MM algorithm creates a surrogate function that minorize the **objective function**. When the surrogate function is maximized the **objective** function is **increased**.



Red curve minorize the black curve

- When maximizing MM $\equiv$ minorize/maximize.

Big Question?

How do you **majorize** or **minorize** a function??

Big Question?

How do you **majorize** or **minorize** a function??

Here are some generic tricks and tools

- Jensen's inequality
- Chord above the graph property of a convex function
- Supporting hyperplane property of a convex function
- Quadratic upper bound principle
- Arithmetic-geometric mean inequality
- The Cauchy-Schwartz inequality

Big Question?

How do you **majorize** or **minorize** a function??

Here are some generic tricks and tools

- Jensen's inequality
- Chord above the graph property of a convex function
- Supporting hyperplane property of a convex function
- Quadratic upper bound principle
- Arithmetic-geometric mean inequality
- The Cauchy-Schwartz inequality

Presume it would take some practice to use these tricks.

Big Question?

How do you **majorize** or **minorize** a function??

Here are some generic tricks and tools

- Jensen's inequality
- Chord above the graph property of a convex function
- Supporting hyperplane property of a convex function
- Quadratic upper bound principle
- Arithmetic-geometric mean inequality
- The Cauchy-Schwartz inequality

Presume it would take some practice to use these tricks.

But....

But wait...

You probably have minorized via **Jensen's Inequality**!

Remember **Jensen's Inequality:**

- $h(\cdot)$ be a concave function,

But wait...

You probably have minorized via **Jensen's Inequality**!

Remember **Jensen's Inequality:**

- $h(\cdot)$ be a concave function,
- have K non-negative numbers $\pi_1, \dots, \pi_K$ with $\sum_k \pi_i = 1$,

But wait...

You probably have minorized via **Jensen's Inequality**!

Remember **Jensen's Inequality:**

- $h(\cdot)$ be a concave function,
- have K non-negative numbers $\pi_1, \ldots, \pi_K$ with $\sum_k \pi_i = 1$,
- K arbitrary numbers $a_1, \ldots, a_K$

But wait...

You probably have minorized via **Jensen's Inequality**!

Remember **Jensen's Inequality:**

- $h(\cdot)$ be a concave function,
- have K non-negative numbers $\pi_1, \ldots, \pi_K$ with $\sum_k \pi_i = 1$,
- K arbitrary numbers $a_1, \ldots, a_K$

then

$$h\left(\sum_{k=1}^{K} \pi_k \, a_k\right) \geq \sum_{k=1}^{K} \pi_k \, h(a_k)$$

Finally we're getting to **E**xpectation**M**aximization

- The EM algorithm is a MM algorithm.

Finally we're getting to **E**xpectation**M**aximization

- The EM algorithm is a MM algorithm.
- Use Jensen's inequality to minorize the log-likelihood.

Finally we're getting to **E**xpectation**M**aximization

- The EM algorithm is a MM algorithm.
- Use Jensen's inequality to minorize the log-likelihood.

Here's how we minorize. Step 1:

$$L(\Theta; \mathbf{X}) = \log\left(p(\mathbf{X} \mid \Theta\right) = \log\left(\sum_{j=1}^{n_z} p\left(\mathbf{X}, \mathbf{Z} = \mathbf{z}_j \mid \Theta\right)\right) \quad \leftarrow \text{introduce discrete variable } \mathbf{Z}$$

$$f^{(t)}(\mathbf{Z}) \text{ a pdf} \rightarrow = \log\left(\sum_{j=1}^{n_z} f^{(t)}(\mathbf{Z} = \mathbf{z}_j) \frac{p\left(\mathbf{X}, \mathbf{Z} = \mathbf{z}_j \mid \Theta\right)}{f^{(t)}(\mathbf{Z} = \mathbf{z}_j)}\right)$$

$$\text{Jensen's inequality} \rightarrow \geq \sum_{j=1}^{n_z} f^{(t)}(\mathbf{Z} = \mathbf{z}_j) \log\left(\frac{p\left(\mathbf{X}, \mathbf{Z} = \mathbf{z}_j \mid \Theta\right)}{f^{(t)}(\mathbf{Z} = \mathbf{z}_j)}\right)$$

Finally we're getting to **E**xpectation**M**aximization

- The EM algorithm is a MM algorithm.
- Use Jensen's inequality to minorize the log-likelihood.

Here's how we minorize. Step 1:

$$L(\Theta;\mathbf{X}) = \log\left(p(\mathbf{X}\,|\,\Theta)\right) = \log\left(\sum_{j=1}^{n_z} p\left(\mathbf{X},\mathbf{Z}=\mathbf{z}_j\,|\,\Theta\right)\right) \quad \leftarrow \text{introduce discrete variable } \mathbf{Z}$$

$$f^{(t)}(\mathbf{Z}) \text{ a pdf} \rightarrow = \log\left(\sum_{j=1}^{n_z} f^{(t)}(\mathbf{Z}=\mathbf{z}_j)\frac{p\left(\mathbf{X},\mathbf{Z}=\mathbf{z}_j\,|\,\Theta\right)}{f^{(t)}(\mathbf{Z}=\mathbf{z}_j)}\right)$$

$$\text{Jensen's inequality} \rightarrow \geq \sum_{j=1}^{n_z} f^{(t)}(\mathbf{Z}=\mathbf{z}_j)\log\left(\frac{p\left(\mathbf{X},\mathbf{Z}=\mathbf{z}_j\,|\,\Theta\right)}{f^{(t)}(\mathbf{Z}=\mathbf{z}_j)}\right)$$

$$L(\Theta;\mathbf{X}) \geq \sum_{j=1}^{n_z} f^{(t)}(\mathbf{Z}=\mathbf{z}_j)\log\left(\frac{p(\mathbf{X},\mathbf{Z}=\mathbf{z}_j\,|\,\Theta)}{f^{(t)}(\mathbf{Z}=\mathbf{z}_j)}\right)$$

Find $f^{(t)}(\mathbf{Z})$

Here's how we minorize. Step 2:

The lower bound must touch the log-likelihood at $\Theta^{(t)}$

$$L(\Theta^{(t)}; \mathbf{X}) = \sum_{j=1}^{n_z} f^{(t)}(\mathbf{Z} = \mathbf{z}_j) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}=\mathbf{z}_j \mid \Theta^{(t)})}{f^{(t)}(\mathbf{Z}=\mathbf{z}_j)} \right)$$

Find $f^{(t)}(\mathbf{Z})$

Here's how we minorize. Step 2:

The lower bound must touch the log-likelihood at $\Theta^{(t)}$

$$L(\Theta^{(t)}; \mathbf{X}) = \sum_{j=1}^{n_z} f^{(t)}(\mathbf{Z} = \mathbf{z}_j) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}{=}\mathbf{z}_j \,|\, \Theta^{(t)})}{f^{(t)}(\mathbf{Z}{=}\mathbf{z}_j)} \right)$$

From this constraint can calculate $f^{(t)}(\mathbf{Z})$. It is:

$$f^{(t)}(\mathbf{Z}) = p(\mathbf{Z} \,|\, \mathbf{X}, \Theta^{(t)})$$

(Derivation is straight-forward)

EM as MM summary

The log-likelihood function $L(\Theta; \mathbf{X})$ at $\Theta^{(t)}$ is minorized by

$$g(\Theta; \Theta^{(t)}) = \sum_{j=1}^{n_z} p(\mathbf{Z} = \mathbf{z}_j \mid \mathbf{X}, \Theta^{(t)}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z} = \mathbf{z}_j \mid \Theta)}{p(\mathbf{Z} = \mathbf{z}_j \mid \mathbf{X}, \Theta^{(t)})} \right)$$

EM as MM summary

The log-likelihood function $L(\Theta; \mathbf{X})$ at $\Theta^{(t)}$ is minorized by

$$g(\Theta; \Theta^{(t)}) = \sum_{j=1}^{n_z} p(\mathbf{Z} = \mathbf{z}_j \mid \mathbf{X}, \Theta^{(t)}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}=\mathbf{z}_j \mid \Theta)}{p(\mathbf{Z}=\mathbf{z}_j \mid \mathbf{X}, \Theta^{(t)})} \right)$$

Maximizing the surrogate function, $g(\Theta; \Theta^{(t)})$, involves:

$$\begin{aligned}
\Theta^{(t+1)} &= \arg\max_{\Theta} \; g(\Theta; \Theta^{(t)}) \\
&= \arg\max_{\Theta} \; \sum_{j=1}^{n_z} p(\mathbf{Z} = \mathbf{z}_j \mid \mathbf{X}, \Theta^{(t)}) \log\left(p(\mathbf{X}, \mathbf{Z} = \mathbf{z}_j \mid \Theta)\right) \\
&= \overbrace{\arg\max_{\Theta} \; \underbrace{E_{p(\mathbf{Z} \mid \mathbf{X}, \Theta^{(t)})}\left[\log\left(p(\mathbf{X}, \mathbf{Z} \mid \Theta)\right)\right]}_{\text{Expectation Step}}}^{\text{Maximization Step}}
\end{aligned}$$

The latent/hidden variables $\mathbf{Z}$

There seemed to be some magic in this derivation!

What are the **Z**'s and where did they come from??

Answer:

The latent/hidden variables $\mathbf{Z}$

There seemed to be some magic in this derivation!

What are the $\mathbf{Z}$'s and where did they come from??

Answer:

- $\mathbf{Z}$ is a random variable whose pdf conditioned on $\mathbf{X}$ is completely determined by Θ.

The latent/hidden variables $\mathbf{Z}$

There seemed to be some magic in this derivation!

What are the $\mathbf{Z}$'s and where did they come from??

Answer:

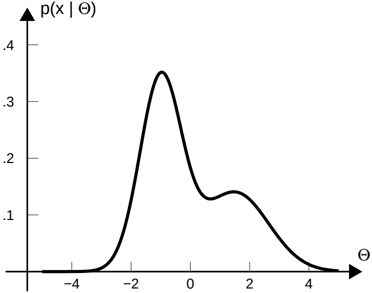
- $\mathbf{Z}$ is a random variable whose pdf conditioned on $\mathbf{X}$ is completely determined by Θ.
- Choice of $\mathbf{Z}$ should make the maximization step **easy**.

Back to our GMM parameter estimation and EM

Attempt 3: Parameter estimation for a GMM

Let's look at a tutorial example using EM:

$$p(x \mid \Theta) = \alpha \, \mathcal{N}(x \mid \mu_1, \sigma_1^2) + (1 - \alpha) \, \mathcal{N}(x \mid \mu_2, \sigma_2^2)$$

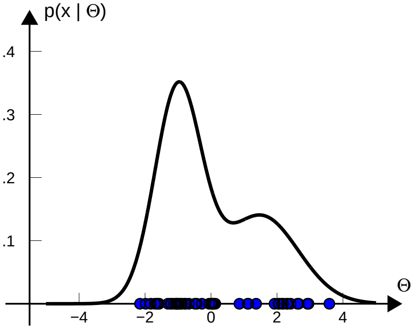


$\longleftarrow$ **Ground truth distribution**

where $\Theta = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2) = (.6, -1, .5, 1.5, 1.3)$

Attempt 3: Parameter estimation for a GMM

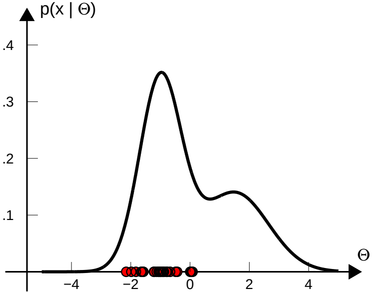
Say all the parameters of Θ are known except α. Then we are given n samples $\mathbf{X} = (x_1, x_2, \ldots, x_n)$ independently drawn from $p(x \mid \Theta)$. Using these samples and EM we can estimate α.



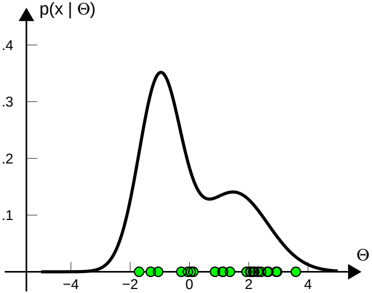
$\longleftarrow$ **training data**

Attempt 3: Parameter estimation for a GMM

If we knew which samples were generated by which component, life would be so much simpler!



Component 1 samples



Component 2 samples

Attempt 3: EM Solution

Introduce hidden/latent variables:

$\mathbf{Z} = (z_1, \ldots, z_n)$ is a vector of hidden variables.

Each $z_i \in \{0, 1\}$ indicates component generating x_i.

Attempt 3: EM Solution

Introduce hidden/latent variables:

$\mathbf{Z} = (z_1, \ldots, z_n)$ is a vector of hidden variables.

Each $z_i \in \{0, 1\}$ indicates component generating x_i.

E-step:

- Update posteriors for the hidden variables:

$$p(z_i = 0 \mid x_i, \alpha^{(t)}) = \frac{p(x_i \mid \mu_1, \sigma_1)\, \alpha^{(t)}}{p(x_i \mid \mu_1, \sigma_1)\, \alpha^{(t)} + p(x_i \mid \mu_2, \sigma_2)\, (1 - \alpha^{(t)})}$$

Attempt 3: EM Solution

Introduce hidden/latent variables:

$\mathbf{Z} = (z_1, \ldots, z_n)$ is a vector of hidden variables.

Each $z_i \in \{0, 1\}$ indicates component generating x_i.

E-step:

- Update posteriors for the hidden variables:

$$p(z_i = 0 \,|\, x_i, \alpha^{(t)}) = \frac{p(x_i \,|\, \mu_1, \sigma_1)\, \alpha^{(t)}}{p(x_i \,|\, \mu_1, \sigma_1)\, \alpha^{(t)} + p(x_i \,|\, \mu_2, \sigma_2)\,(1 - \alpha^{(t)})}$$

- Calculate the conditional expectation

$$g(\alpha; \alpha^{(t)}) = \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z} \,|\, \mathbf{X}, \alpha^{(t)})\, \log\left(\frac{p(\mathbf{X}, \mathbf{Z} \,|\, \alpha)}{p(\mathbf{Z} \,|\, \mathbf{X}, \alpha^{(t)})}\right)$$

Attempt 3: EM Solution

Introduce hidden/latent variables:

$\mathbf{Z} = (z_1, \ldots, z_n)$ is a vector of hidden variables.

Each $z_i \in \{0, 1\}$ indicates component generating x_i.

E-step:

- Update posteriors for the hidden variables:

$$p(z_i = 0 \mid x_i, \alpha^{(t)}) = \frac{p(x_i \mid \mu_1, \sigma_1)\, \alpha^{(t)}}{p(x_i \mid \mu_1, \sigma_1)\, \alpha^{(t)} + p(x_i \mid \mu_2, \sigma_2)\, (1 - \alpha^{(t)})}$$

- Calculate the conditional expectation

$$g(\alpha; \alpha^{(t)}) = \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \alpha^{(t)}) \ \log \left(\frac{p(\mathbf{X}, \mathbf{Z} \mid \alpha)}{p(\mathbf{Z} \mid \mathbf{X}, \alpha^{(t)})} \right)$$

M-step: Find $\arg\max_{\alpha} g(\alpha; \alpha^{(t)})$ which gives:

$$\alpha^{(t+1)} = \frac{\sum_i p(z_i = 0 \mid \mathbf{x}_i, \alpha^{(t)})}{n}$$

Attempt 3: EM expectation calculation

$$\sum_{\text{all } \mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \alpha^{(t)}) \; \log\left(p(\mathbf{X}, \mathbf{Z} \mid \alpha)\right)$$

$$= \sum_{\text{all } \mathbf{Z}} \left[\prod_{s=1}^{n} p(z_s \mid x_s, \alpha^{(t)}) \sum_{i=1}^{n} \log\left(p(x_i \mid z_i, \alpha)\, p(z_i | \alpha)\right) \right]$$

$$= \sum_{j_1=0}^{1} \cdots \sum_{j_n=0}^{1} \left[\prod_{s=1}^{n} p(z_s = j_s \mid x_s, \alpha^{(t)}) \sum_{i=1}^{n} \log\left(p(x_i \mid z_i = j_i, \alpha)\, p(z_i = j_i | \alpha)\right) \right]$$

$$= \sum_{i=1}^{n} \left[\left(\prod_{s=1, s\neq i}^{n} \underbrace{\sum_{j_s=0}^{1} p(z_s = j_s \mid x_s, \alpha^{(t)})}_{=1} \right) p(z_i = j_i \mid x_i, \alpha^{(t)}) \log\left(p(x_i \mid z_i = j_i, \alpha)\, p(z_i = j_i | \alpha)\right) \right]$$

$$= \sum_{i=1}^{n} \sum_{j_i=0}^{1} p(z_i = j_i \mid x_i, \alpha^{(t)}) \log\left(p(x_i \mid z_i = j_i, \alpha)\, p(z_i = j_i | \alpha)\right)$$

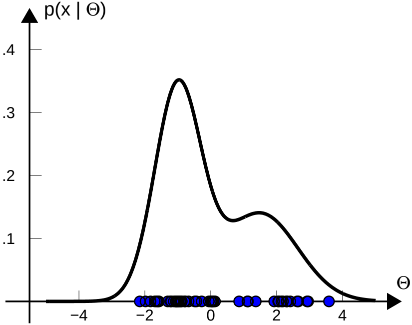
$$= \sum_{i=1}^{n} \sum_{j_i=0}^{1} p(z_i = j_i \mid x_i, \alpha^{(t)}) \log\left(N(x_i \mid \mu_{j_i}, \sigma_{j_i})\, \alpha^{1-j_i} (1-\alpha)^{j_i} \right)$$

Attempt 3: EM maximization process

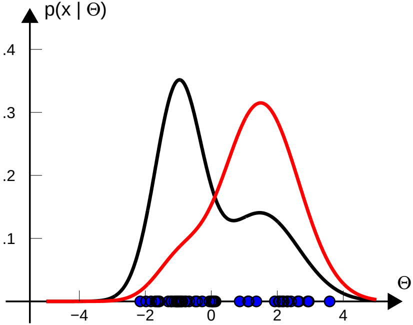
$$
\begin{aligned}
\frac{\partial \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \alpha^{(t)}) \; \log\left(p(\mathbf{X}, \mathbf{Z} \mid \alpha)\right)}{\partial \alpha} &= \sum_{i=1}^{n} \sum_{j_i=0}^{1} p(z_i = j_i \mid x_i, \alpha^{(t)}) \, \frac{\partial \, \log\left(\alpha^{1-j_i}(1-\alpha)^{j_i}\right)}{\partial \, \alpha} \\
&= \sum_{i=1}^{n} \sum_{j_i=0}^{1} p(z_i = j_i \mid x_i, \alpha^{(t)}) \, \left(\frac{1-j_i}{\alpha} - \frac{j_i}{1-\alpha}\right) \\
&= \sum_{i=1}^{n} \sum_{j_i=0}^{1} p(z_i = j_i \mid x_i, \alpha^{(t)}) \, (1 - j_i - \alpha) \\
&= (1-\alpha) \sum_{i=1}^{n} \sum_{j_i=0}^{1} p(z_i = j_i \mid x_i, \alpha^{(t)}) - \sum_{i=1}^{n} \sum_{j_i=0}^{1} p(z_i = j_i \mid x_i, \alpha^{(t)}) \, j_i \\
&= n(1-\alpha) - \sum_{i=1}^{n} p(z_i = 1 \mid x_i, \alpha^{(t)}) \\
&= -n\,\alpha + n - \sum_{i=1}^{n} (1 - p(z_i = 0 \mid x_i, \alpha^{(t)})) \\
&= \sum_{i=1}^{n} p(z_i = 0 \mid x_i, \alpha^{(t)}) - n\,\alpha = 0
\end{aligned}
$$

Therefore $\alpha^{(t+1)} = \frac{\sum_{i=1}^{n} p(z_i = 0 \mid x_i, \alpha^{(t)})}{n}$

Attempt 3: EM Solution starting point



Ground truth distribution



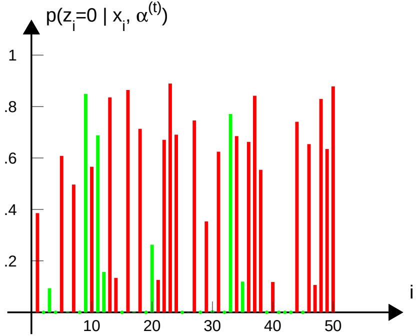
Initial guess of distribution with $\alpha^{(0)} = .1$

Remember $g(\alpha; \alpha^{(t)})$ minorizes $\log\left(p(\mathbf{X} \mid \alpha)\right)$ at $\alpha^{(t)}$.

Let's plot what happens as EM update $\alpha^{(t)}$...

EM one iteration

Compute posterior probabilities of the hidden variables



Graph shows $p(z_i = 0 \,|\, x_i, \alpha^{(0)})$ of each hidden variable.

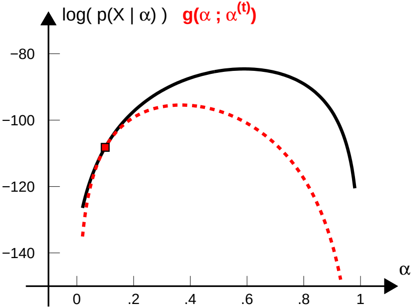
Red $\Longrightarrow$ sample really generated by component 1

Green $\Longrightarrow$ sample really generated by component 2

EM one iteration

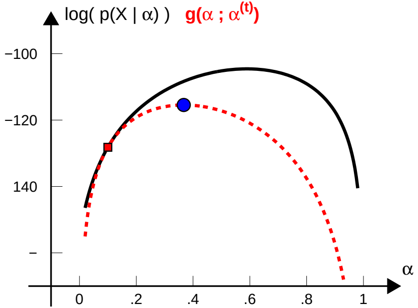
Compute the expectation minorizing the log-likelihood at $\alpha^{(0)} = .1$

$$\color{red}{g(\alpha; \alpha^{(t)})} = \sum_{\text{all } \mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \alpha^{(t)}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z} \mid \alpha)}{p(\mathbf{Z} \mid \mathbf{X}, \alpha^{(t)})} \right)$$



EM one iteration

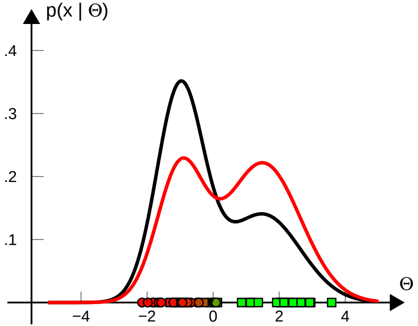
Calculate maximum of $g(\alpha; \alpha^{(0)})$



Maximum of $g(\alpha; \alpha^{(0)})$ gives $\alpha^{(1)} = .3672$

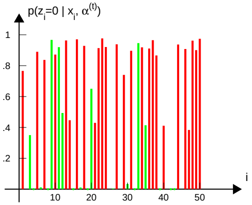
EM one iteration

The estimate of the GMM with $\alpha^{(1)} = .3672$

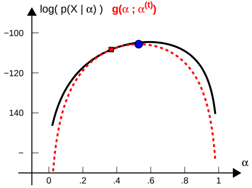


EM Iterations

Iteration 2

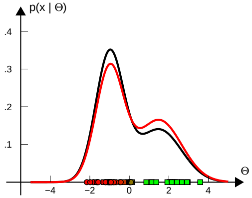


Posterior probabilities



$g(\alpha; \alpha^{(1)})$

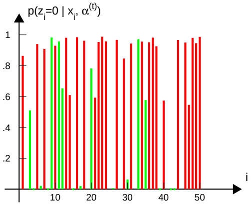
$\alpha^{(2)} = .5287$



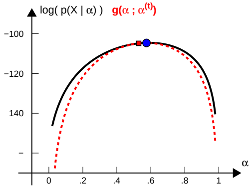
Current GMM estimate

EM Iterations

Iteration 3

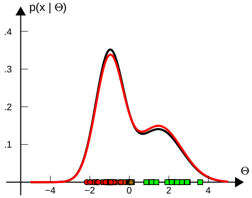


Posterior probabilities



$g(\alpha; \alpha^{(2)})$

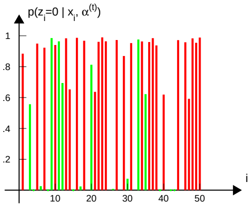
$\alpha^{(3)} = .5748$



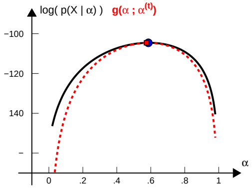
Current GMM estimate

EM Iterations

Iteration 4

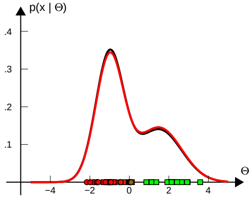


Posterior probabilities



$g(\alpha; \alpha^{(3)})$

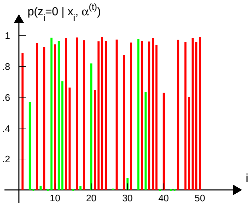
$\alpha^{(4)} = .5859$



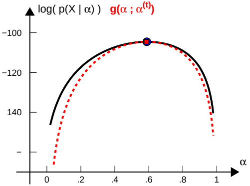
Current GMM estimate

EM Iterations

Iteration 5

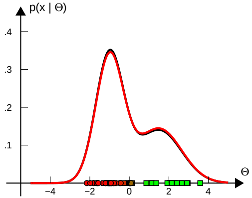


Posterior probabilities



$g(\alpha; \alpha^{(4)})$

$\alpha^{(5)} = .5885$



Current GMM estimate

Comments on EM

Design Issues

- The choice of hidden/latent variable **Z** is the most important issue for EM.

Implementation Issues

Comments on EM

Design Issues

- The choice of hidden/latent variable $\mathbf{Z}$ is the most important issue for EM.
- Choice must be done so the maximization step is easy.

Implementation Issues

Comments on EM

Design Issues

- The choice of hidden/latent variable $\mathbf{Z}$ is the most important issue for EM.
- Choice must be done so the maximization step is easy.
- Or at least easier than the maximization of the log-likelihood function.

Implementation Issues

Comments on EM

Design Issues

- The choice of hidden/latent variable $\mathbf{Z}$ is the most important issue for EM.
- Choice must be done so the maximization step is easy.
- Or at least easier than the maximization of the log-likelihood function.

Implementation Issues

- Calculation of the conditional expectation may be taxing.

Comments on EM

Design Issues

- The choice of hidden/latent variable $\mathbf{Z}$ is the most important issue for EM.
- Choice must be done so the maximization step is easy.
- Or at least easier than the maximization of the log-likelihood function.

Implementation Issues

- Calculation of the conditional expectation may be taxing.
- Convergence of EM can be slow near the local optimum.