The Expectation-Maximization algorithm

David Bolin
Chalmers University of Technology

October 14, 2014

The EM-algorithm

- We have earlier seen how the introduction of auxiliary variables can simplify estimation in several cases:
  - Censored and truncated data
  - Binary regression models
  - Latent variabel models such as normal-variance mixtures
- We used data augmentation to derive MCMC estimators for these problem.
- If we are only interested in finding MAP/ML parameter estimates, another alternative is the EM-algorithm
- The main reference is Dempster, Laird & Rubin (1977).
- "proposed many times in special circumstances".

The EM-algorithm

Basic setup:

- We have observed some data y.
- Additional data z is "missing".
- The estimation problem would be "easy" if z was known.

In principle we could write out the posterior given only the observed data as

$$\pi(\theta|y) = \int \pi(\theta|y,z)\pi(z|y)\,\mathrm{d}z.$$

However the integral over the unknown data is often hard to compute.

The EM-algorithm provides a method for finding the MAP estimate of the parameters

The EM-algorithm

1. Choose some initial guess of the parameters, θ_{guess}.
2. Write down the log-posterior assuming that all the data is known, $\log \pi(\theta|y,z)$.
3. Compute the expected value of the log-posterior over the auxiliary variables, $Q(\theta,\theta_{\text{guess}}) = \mathsf{E}(\log \pi(\theta|y,z)|y,\theta_{\text{guess}})$.
4. Q can now be seen as the average possible value of the log-posterior given known observations and guessed parameters.
5. Update our guess of the parameters by maximasing $Q(\theta,\theta_{\text{guess}})$.
6. Repeat from 3.

The result is the Expectation-Maximization algorithm.

The EM algorithm

Choose a starting value $\theta^{(0)}$ and repeat for $i = 1, 2, \dots$ until convergence.

E-step Compute the expectation of the log-posterior with respect to the unknown data

$$Q(\theta, \theta^{(i-1)}) = \mathsf{E}\left(\log \pi(\theta | y, z) | y, \theta^{(i-1)}\right).$$

M-step Compute $\theta^{(i)} = \arg\max_{\theta} Q(\theta, \theta^{(i-1)})$.

Remarks:

- Under weak smoothness conditions, the algorithm will converge to a local maxima of the posterior.
- We have presented the algorithm in the Bayesian setting, in the original frequentist setting, the log-posterior is replaced with the log-likelihood.

Example — How to optimize your dart strategy

- Darts is enjoyed both as a pub game and as a professional competitive activity.
- Most players aim for the highest scoring region of the board, regardless of their level of skill.
- Recently Tibshirani, Price, and Taylor (2010) investigated whether this is the optimal strategy

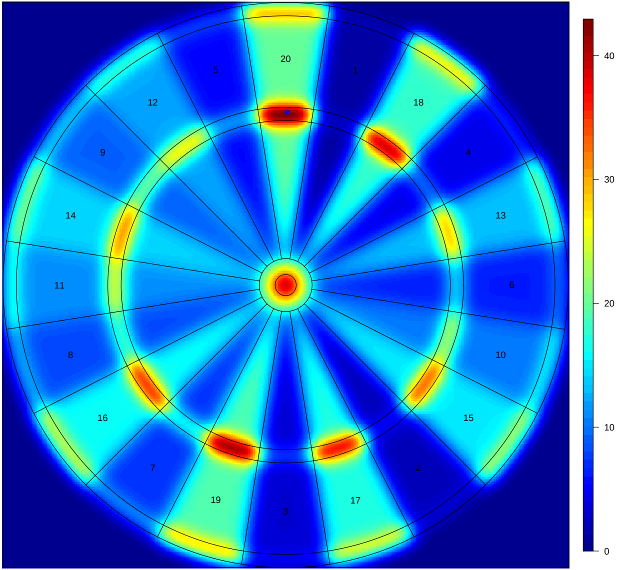
Darts: Setup

- Let the center of the board correspond to the origin
- Let μ be the location where we aim and let Z denote the location where the dart actually hits the board
- A simple model is that $Z \sim \mathsf{N}(\mu, \sigma^2 I)$ where σ^2 represents our accuracy.
- Let $s(Z)$ denote the score we get from Z.
- The goal is now to choose where we aim (μ) in order to maximize

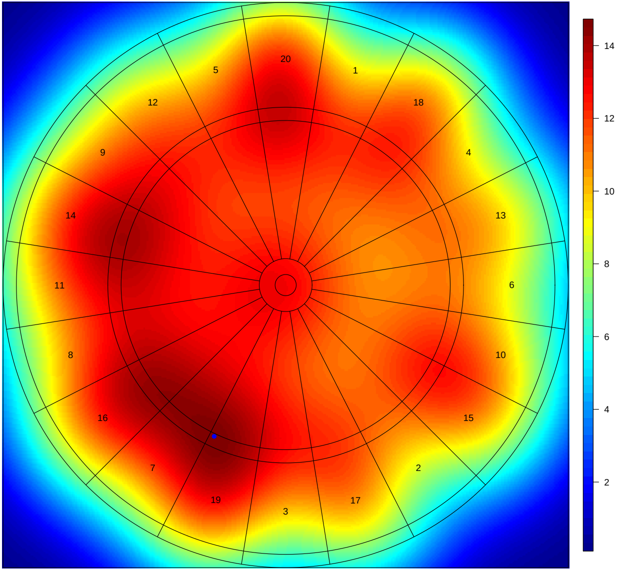
$$\mathsf{E}(s(Z)) = \int s(Z) \frac{1}{2\pi\sigma^2} \exp\left(-\frac{1}{2}\|Z-\mu\|^2\right) dZ$$

- If we know σ^2, we can calculate the expected score as a function of μ (e.g. using Fourier transforms)

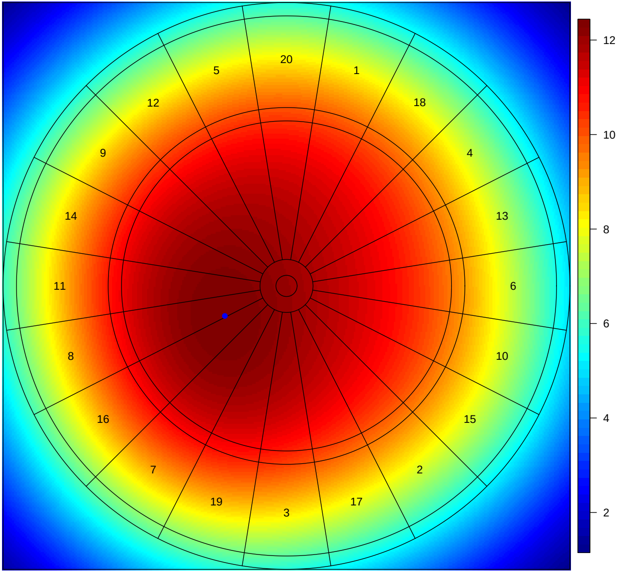
Expected score for $\sigma = 5$ mm



Expected score for $\sigma = 30$ mm



Expected score for $\sigma = 60$ mm



Darts: Estimation

- Where we should aim depends on how accurate we are!
- We need to estimate our own accuracy σ^2 in order to find the optimal strategy
- Throw n darts, aiming at bullseye ($\mu = 0$)
- Estimating σ^2 is trivial if we record the positions of the darts:

$$\sigma^2_{\mathsf{MLE}} = \frac{1}{2n}\sum_{i=1}^{n}(Z_{i,x}^2 + Z_{i,y}^2)$$

- But this is not realistic to do at the pub!
- Instead, we just record the score and use the EM algorithm to estimate σ^2.

Darts: The algorithm

Let $X = s(Z)$ denote the score.

We have

$$Q(\sigma, \sigma^{(i)}) = -n \log(\sigma^2) + \frac{1}{2\sigma^2} \sum_{j=1}^{n} \mathsf{E}(Z_{j,x}^2 + Z_{j,y}^2 | X_j, \sigma^{(i)})$$

Calculating $\frac{\partial Q}{\partial \sigma^2} = 0$ gives

$$(\sigma^2)^{(i+1)} = \frac{1}{2n} \sum_{j=1}^{n} \mathsf{E}(Z_{j,x}^2 + Z_{j,y}^2 | X_j, \sigma^{(i)})$$

Thus, in order to estimate σ^2 we iterate:

1. Calculate $\mathsf{E}(Z_{j,x}^2 + Z_{j,y}^2 | X_j, \sigma^{(i)})$ for $j = 1, \ldots, n$.
2. Set $(\sigma^2)^{(i+1)} = \frac{1}{2n} \sum_{j=1}^{n} \mathsf{E}(Z_{j,x}^2 + Z_{j,y}^2 | X_j, \sigma^{(i)})$

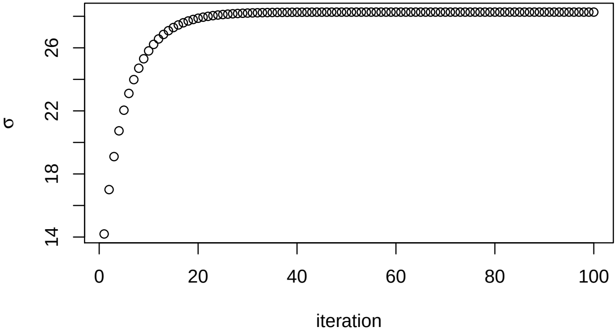
Calculating $\mathsf{E}(Z_x^2 + Z_y^2 | X, \sigma^2)$

We can describe X as being achieved by landing in $\cup_j A_j$, where each region A_j can be expressed as $[r_{j,1} r_{j,2}] \times [\theta_{j,1}, \theta_{j,2}]$ in polar co-ordinates.

Thus,

$$
\begin{aligned}
\mathsf{E}(Z_x^2 + Z_y^2 | X, \sigma^2) &= \mathsf{E}(Z_x^2 + Z_y^2 | Z \in \cup_j A_j, \sigma^2) \\
&= \frac{\sum_j \int\int_{A_j} (x^2 + y^2) e^{-(x^2+y^2)/2\sigma^2} dxdy}{\sum_j \int\int_{A_j} e^{-(x^2+y^2)/2\sigma^2} dxdy} \\
&= \frac{\sum_j \int_{r_{j,1}}^{r_{j,2}} \int_{\theta_{j,1}}^{\theta_{j,2}} r^3 e^{-r^2/2\sigma^2} d\theta dr}{\sum_j \int_{r_{j,1}}^{r_{j,2}} \int_{\theta_{j,1}}^{\theta_{j,2}} r e^{-r^2/2\sigma^2} d\theta dr} \\
&= \frac{\sum_j (r_{j,1}^2 + 2\sigma^2) e^{-r_{j,1}^2/2\sigma^2} - (r_{j,2}^2 + 2\sigma^2) e^{-r_{j,2}^2/2\sigma^2}}{\sum_j e^{-r_{j,1}^2/2\sigma^2} - e^{-r_{j,2}^2/2\sigma^2}}
\end{aligned}
$$

Darts: Results

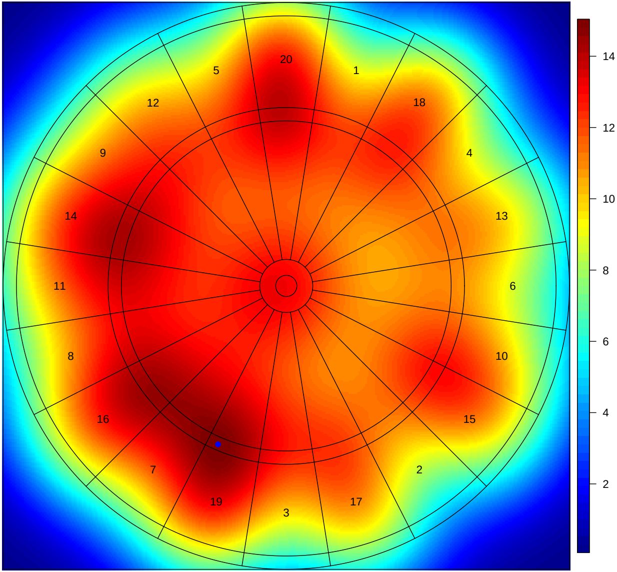


- Results based on 100 measurements

$$12, 16, 19, 3, 17, 1, 25, 19, 17, 50, 18, \ldots$$

- Implementation available in the R package `darts`

Resulting heat map



Gaussian mixture models

A classical application of the EM algorithm.

- Assume that we have observations from one of several Gaussian distributions, called classes.
- The prior probability of data coming from class k is w_k.
- The distribution of each class is $[y|\text{from class k}] \sim \mathsf{N}(\mu_k, \Sigma_k)$.
- This generates a Gaussian mixture model with density

$$\pi(y|w, \mu, \Sigma) = \sum_{k=1}^{K} w_k \pi(y|\text{from class k}, \mu_k, \Sigma_k).$$

- Possible usages
  - Modeling heavy tailed distributions.
  - Classification/clustering of data.

(Gaussian) Mixture Models (cont.)

- If we knew the class belonging, z_j, of each observation y_i the problem would be trivial.
- Thus the problem consists of two parts:
  1. Determine the class belongings z.
  2. Estimating the parameters $\theta = \{w, \mu, \Sigma\}$.
- Assuming flat priors for the parameters, we get

$$\begin{aligned}\log \pi(\theta|y,z) &= \log \prod_{j=1}^{n} w_{z_i} \pi(y_j|z_j, \mu_k, \Sigma_k) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{K} \mathbf{1}(z_j = k) \log\left(w_k \pi(y_j|z_j = k, \mu_k, \Sigma_k)\right).\end{aligned}$$

(Gaussian) Mixture Models — E-step

Taking the conditional expectation, we get

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= \mathsf{E}_z\left(\log \pi(\theta|y,z)|y,\theta^{(i)}\right) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{K} w_{jk} \log\left(w_k \pi(y_j|z_j = k, \mu_k, \Sigma_k)\right) \end{aligned}$$

where

$$w_{jk} = \mathsf{E}(\mathbf{1}(z_j = k)|y, \theta^{(i)}) = \mathsf{P}(z_j = k|y, \theta^{(i)})$$

is the posterior probability for observation j belonging to class k:

$$\begin{aligned} w_{jk} &= \frac{\pi(z_j = k, y|\theta^{(i)})}{\pi(y|\theta^{(i)})} = \frac{\pi(y|z_j = k, \theta^{(i)})\, \pi(z_j = k|\theta^{(i)})}{\sum_k \pi(z_j = k, y|\theta^{(i)})} \\ &= \frac{\pi(y|z_j = k, \theta^{(i)})\, w_k}{\sum_k \pi(y|z_j = k, \theta^{(i)})\, w_k} \end{aligned}$$

Gaussian Mixture Models — M-step

Having calculated the expectation of the log-likelihood we have that

$$
\begin{aligned}
Q(\theta, \theta^{(i)}) &= \mathsf{E}\Big(\log \pi(\theta|y, z)|y, \theta^{(i)}\Big) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{K} w_{ik}\bigg(\log(w_k) + \log\Big(\pi(y_j|z_j = k, \mu_k, \Sigma_k)\Big)\bigg) \\
&= \sum_{j=1}^{n}\sum_{k=1}^{K} w_{jk}\bigg(\log(w_k) - \frac{1}{2}\log\det\Sigma_k - \frac{d}{2}\log(2\pi) - \\
&\qquad \frac{(y_j - \mu_k)^\top \Sigma_k^{-1}(y_j - \mu_k)}{2}\bigg).
\end{aligned}
$$

Gaussian Mixture Models — M-step

The new estimates of $\{\pi, \mu, \Sigma\}$ are obtained by maximizing the Q-function.

Differentiating the function and setting the derivatives equal to zero yields:

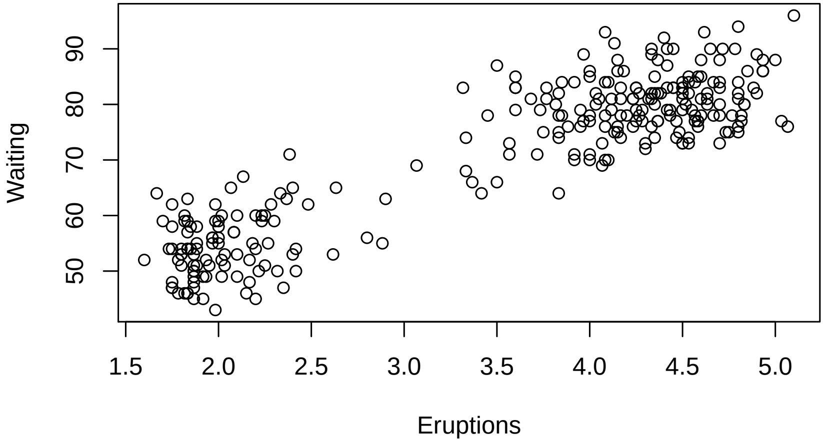
$$w_k^{(i+1)} = \frac{1}{n}\sum_{j=1}^{n} \mathsf{P}(z_j = k | y_j, \theta^{(i)}) = \frac{1}{n}\sum_{j=1}^{n} w_{jk}$$

$$\mu_k^{(i+1)} = \frac{1}{n w_k}\sum_{j=1}^{n} w_{jk} y_i$$

$$\Sigma_k^{(i+1)} = \frac{1}{n w_k}\sum_{j=1}^{n} w_{jk} (y_j - \mu_k)^\top (y_j - \mu_k).$$

Example — Old Faithful

Data



- Time before eruption and duration of eruption.
- A mixture model with two classes seems reasonable

Results

```
library(mixtools)
data(faithful)
res = mvnormalmixEM(faithful)
```

- Estimated probabilities: $w_1 = 0.356$, $w_2 = 0.644$.
- Estimated parameters for first class:

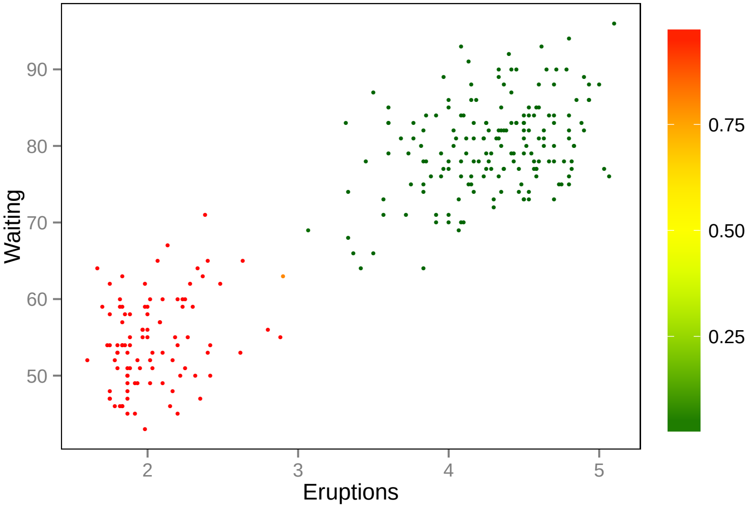
$$\mu_1 = \begin{pmatrix} 2.04 \\ 54.48 \end{pmatrix}, \quad \Sigma_1 = \begin{pmatrix} 0.0692 & 0.4352 \\ 0.4352 & 33.6973 \end{pmatrix}$$

- Estimated parameters for second class:

$$\mu_2 = \begin{pmatrix} 4.29 \\ 79.97 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 0.170 & 0.941 \\ 0.941 & 36.046 \end{pmatrix}$$

Gaussian Mixture Models — Old Faithful

Posterior probability for first class:



The Monte Carlo EM algorithm

In some applications, the E-step is complex and does not admit a closed form solution. It is then natural to approximate it using Monte Carlo methods.

The MCEM algorithm

Choose a starting value $\theta^{(0)}$ and repeat for $i = 1, 2, \ldots$ until convergence.

MC E-step Draw $z^{(1)}, \ldots, z^{(M)}$ from $\pi(z|y, \theta^{(i)})$, and let

$$Q(\theta, \theta^{(i-1)}) = \frac{1}{M} \sum_{m=1}^{M} \log \pi(\theta|y, z^{(m)})$$

M-step Update the parameter estimate

$$\theta^{(i)} = \arg\max_{\theta} Q(\theta, \theta^{(i-1)}).$$

The Expectation Conditional Maximization algorithm

When we have several unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, the M-step may not admit a closed form solution. In this case, one can replace it with p conditional maximization steps

The ECM algorithm

Choose a starting value $\boldsymbol{\theta}^{(0)}$ and repeat for $i = 1, 2, \ldots$ until convergence.

E-step $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(i-1)}) = \mathsf{E}_z \left(\log \pi(\boldsymbol{\theta}|y, z)|y, \boldsymbol{\theta}^{(i-1)} \right).$

CM-step For $j = 1, \ldots, p$, compute

$$\theta_j^{(i)} = \arg\max_{\theta} Q(\boldsymbol{\theta}_{-j}^{(i-1)}, \boldsymbol{\theta}^{(i-1)}).$$

where $\boldsymbol{\theta}_{-j}^{(i)} = (\theta_1^{(i)}, \ldots, \theta_{j-1}^{(i)}, \theta, \theta_{j+1}^{(i)}, \ldots, \theta_p^{(i)})$