Expectation Maximization

- Introduction to EM algorithm

TLT-5906 Advanced Course in Digital Transmission

Jukka Talvitie, M.Sc. (eng)
jukka.talvitie@tut.fi
Department of Communication Engineering
Tampere University of Technology

5.12.2013
Outline

- Expectation Maximization (EM) algorithm
  - Motivation, background
  - Where the EM can be used?

- EM principle
  - Formal definition
  - How the algorithm really works?
  - Coin toss example
  - About some practical issues

- More advanced examples
  - Line fitting with EM algorithm
  - Parameter estimation of multivariate Gaussian mixture

- Conclusions
Consider classical line fitting problem:

- Assume below measurements of a linear model $y = ax + b + n$ (here the line parameters are $a$ and $b$ and $n$ is zero mean noise)
Motivation

- We use LS (Least Squares) to find the best fit:
- Is this the best solution?
Motivation

- LS would be the Best Linear Unbiased Estimator, if the noise would be uncorrelated with fixed variance.
- Here, actually the noise term is correlated and the actual linear model of this realization can be seen below as the black line.
  - Here the LS gives too much weight for a group of samples in the middle.
Motivation

- Taking the correlation of the noise term into account, we can use Generalized LS method and the result can be improved considerably.
- However, in many cases we do not know the correlation model.
  - It is hidden in the observations and we cannot access it directly.
  - Therefore, e.g., here we would need to estimate simultaneously the covariance and the line parameters.
- This sort of problems might quite quickly become very complicated.
  - How to estimate the covariance without knowing the line parameters and vice versa?
- Intuitive (heuristic) solution:
  - Iteratively estimate the other parameter, and then the other, and continue...
  - No guarantee for the performance in this case (e.g., compared to maximum likelihood (ML) solution).
- The EM algorithm provides the ML solution for these sort of problems.
Expectation Maximization Algorithm

- Presented by Dempster, Laird and Rubin in [1] in 1977
  - Basically the same principle was already proposed earlier by some other authors in specific circumstances
- EM algorithm is an iterative estimation algorithm that can derive the maximum likelihood (ML) estimates in the presence of missing/hidden data ("incomplete data")
  - e.g. the classical case is the Gaussian mixture, where we have a set of unknown Gaussian distributions (see example later on)

Many-to-one mapping [2]

\[ X(y) \] 

\[ X: \text{underlying space} \]
\[ x: \text{complete data (required for ML)} \]
\[ Y: \text{observation space} \]
\[ y: \text{observation} \]

\[ x \] is observed only by means of \( y(x) \). \( X(y) \) is a subset of \( X \) determined by \( y \).
Expectation Maximization Algorithm

- The basic functioning of the EM algorithm can be divided into two steps (the parameter to be estimated is $\theta$):
  - Expectation step (E-step)
    - Take the expected value of the complete data given the observation and the current parameter estimate $\hat{\theta}_k$
    $$Q(\theta, \hat{\theta}_k) = E \left\{ \log f(x \mid \theta) \mid y, \hat{\theta}_k \right\}$$
  - Maximization step (M-step)
    - Maximize the $Q$-function in the E-step (basically, the data of the E-step is used as it were measured observations)
    $$\hat{\theta}_{k+1} = \arg \max_{\theta} Q(\theta \mid \hat{\theta}_k)$$
- The likelihood of the parameter is increased at every iteration
  - EM converges towards some local maximum of the likelihood function
An example: ML estimation vs. EM algorithm [3]

- We wish to estimate the variance of $S$:
  - observation $Y = S + N$
    - $S$ and $N$ are normally distributed with zero means and variances $\theta$ and 1, respectively
    - Now, $Y$ is also normally distributed (zero mean with variance $\theta + 1$)
- ML estimate can be easily derived:

\[
\hat{\theta}_{ML} = \arg \max_\theta (p(y | \theta))
\]

\[
= \max \{0, y^2 - 1\}
\]

- The zero in above result becomes from the fact that we know that the variance is always non-negative
An example: ML estimation vs. EM algorithm

- The same with the EM algorithm
  - complete data is now included in $S$ and $N$
  - E-step is then:

$$Q(\theta, \hat{\theta}_k) = E\left[ \ln p(s, n | \theta) | y, \hat{\theta}_k \right]$$

- the logarithmic probability distribution for the complete data is then

$$\ln p(s, n | \theta) = \ln p(n) + \ln(p(s | \theta))$$

$$= C - \frac{1}{2} \ln \theta - \frac{S^2}{2\theta}$$

$$\rightarrow$$

$$Q(\theta, \hat{\theta}_k) = C - \frac{1}{2} \ln \theta - \frac{E[S^2 | Y, \hat{\theta}_k]}{2\theta}$$

(C contains all the terms independent of $\theta$)
M-step:
- maximize the E-step
- We set the derivative to zero and get (use results from math tables: conditional means and variances, “Law of total variance”)

\[
\hat{\theta}_{k+1} = E\left[ S^2 \mid Y, \hat{\theta}_k \right] = E^2 \left[ S \mid Y, \hat{\theta}_k \right] + \text{var} \left[ S \mid Y, \hat{\theta}_k \right]
\]

\[
= \left( \frac{\hat{\theta}_k}{\hat{\theta}_k + 1} Y \right)^2 + \frac{\hat{\theta}_k}{\hat{\theta}_k + 1}
\]

At the steady state \((\hat{\theta}_{k+1} = \hat{\theta}_k)\) we get the same value for the estimate as in ML estimation \((\max\{0, y^2 - 1\})\)

What about the convergence? What if we choose the initial value \(\hat{\theta}_0 = 0\)
An example: ML estimation vs. EM algorithm

- In the previous example, the ML estimate could be solved in a closed form expression
  - In this case there was no need for EM algorithm, since the ML estimate is given in a straightforward manner (we just showed that the EM algorithm converges to the peak of the likelihood function)

- Next we consider a coin toss example:
  - The target is to figure out the probability of heads for two coins
  - ML estimate can be directly calculated from the results

- We will raise the bets a little bit higher and assume that we don’t even know which one of the coins is used for the sample set?
  - i.e. we are estimating the coin probabilities without knowing which one of the coins is being tossed
An example: Coin toss) [4]

- We have two coins: A and B
- The probabilities for heads are $\theta_A$ and $\theta_B$
- We have 5 measurement sets including 10 coin tosses in each set
- If we know which of the coins are tossed in each set, we can calculate the ML probabilities for $\theta_A$ and $\theta_B$
- If we don’t know which of the coins are tossed in each set, ML estimates cannot be calculated directly
  → EM algorithm

Binomial distribution used to calculate probabilities:

$\binom{n}{k} p^k (1-p)^{n-k}$

ML method (if we know the coins):

$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$

$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$

Example calculations for the first set ($\hat{\theta}_A^{(0)} = 0.6$, $\hat{\theta}_B^{(0)} = 0.5$)

$0.6^5 \cdot 0.5^5 = 0.201$

$0.5^5 \cdot 0.5^5 = 0.246$

Expectation Maximization

1. Initialization
2. E-step
3. M-step

$\hat{\theta}_A^{(1)} = \frac{21.3}{21.3 + 8.6} = 0.71$

$\hat{\theta}_B^{(1)} = \frac{11.7}{11.7 + 8.4} = 0.58$

$\hat{\theta}_A^{(2)} = \frac{21.3}{21.3 + 8.6} = 0.71$

$\hat{\theta}_B^{(2)} = \frac{11.7}{11.7 + 8.4} = 0.58$

$\hat{\theta}_A^{(3)} = \frac{21.3}{21.3 + 8.6} = 0.71$

$\hat{\theta}_B^{(3)} = \frac{11.7}{11.7 + 8.4} = 0.58$

$\hat{\theta}_A^{(4)} = \frac{21.3}{21.3 + 8.6} = 0.71$

$\hat{\theta}_B^{(4)} = \frac{11.7}{11.7 + 8.4} = 0.58$

$\hat{\theta}_A^{(5)} = \frac{21.3}{21.3 + 8.6} = 0.71$

$\hat{\theta}_B^{(5)} = \frac{11.7}{11.7 + 8.4} = 0.58$
About some practical issues

- Although many examples in the literature are showing excellent results using the EM algorithm, the reality is often less glamorous
  - As the number of uncertain parameters increase in the modeled system, even the best available guess (in ML sense) might not be adequate
  - NB! This is not the algorithm’s fault. It still provides the best possible solution in ML sense
- Depending on the form of the likelihood function (provided in the E-step) the convergence rate of the EM might vary considerably
- Notice, that the algorithm converges towards a local maximum
  - To locate the global peak one must use different initial guesses for the estimated parameters or use some other more advanced methods to find out the global peak
  - With multiple unknown (hidden/latent) parameters the number of local peaks usually increases
Further examples

- Line Fitting (showed only in the lecture)
- Parameter estimation of multivariate Gaussian mixture
  - See additional pdf-file for the
    - Problem definition
    - Equations
      - Definition of the log-likelihood function
      - E-step
      - M-step
  - See additional Matlab m-file for the illustration of
    - The example in numerical form
      - Dimensions and value spaces for each parameter
    - The iterative nature of the EM algorithm
      - Study how parameters change at each iteration
    - How initial guesses for the estimated parameters affect the final result
Conclusions

- EM finds iteratively ML estimates in estimation problems with hidden (incomplete) data
  - likelihood increases at every step of the iteration process
- Algorithm consists of two iteratively taken steps:
  - Expectation step (E-step)
    - Take the expected value of the complete data given the observation and the current parameter estimate
  - Maximization step (M-step)
    - Maximize the \( Q \)-function in the E-step (basically, the data of the E-step is used as it were measured observations)
- Algorithm converges to the local maximum
  - Global maximum can be elsewhere
- See reference list for literature regarding use cases of EM algorithm in the Communications
  - These are the references [5]-[16] (not mentioned in the previous slides)
References


Some communications related papers using the EM algorithm (continues in the next slide):

References


