The Expectation-Maximization Algorithm

April 25, 2006

Contents

1 Introduction

Problem: estimation in the presence of nuisance parameters

Consider the following problem: we transmit a sequence $\mathbf{a}$ of N data symbols, belonging to a constellation Ω (for instance $\Omega = \{-1, +1\}$) over a channel. The symbols are drawn iid and uniformly from Ω. At the receiver we have the following observation

$$\mathbf{r} = h\mathbf{a} + \mathbf{n}$$

where h is a channel gain, $h \in \mathbb{R}$, and $\mathbf{n}$ is a vector of N iid normal RVs, zero-mean, variance σ^2. Our goal is to detect the data sequence $\mathbf{a}$.

Solution - attempt 1

The optimal way to proceed (as we know from detection theory) is to determine the MAP estimate of $\mathbf{a}$

$$\begin{aligned}
\hat{\mathbf{a}}_{MAP} &= \arg\max_{\mathbf{a}} p\left(\mathbf{a}|\,\mathbf{r}\right) \\
&= \arg\max_{\mathbf{a}} \int p\left(\mathbf{a}, h|\,\mathbf{r}\right) dh \\
&= \arg\max_{\mathbf{a}} \int p\left(\mathbf{r}|\,\mathbf{a}, h\right) p\left(\mathbf{a}\right) p\left(h\right) dh \\
&= \arg\max_{\mathbf{a}} \int \prod_{k=1}^{N} p\left(r_k|\,a_k, h\right) p\left(h\right) dh
\end{aligned}$$

and at this point we are stuck, since this integral is generally very hard, and $p\left(h\right)$ is perhaps not known.

Solution - attempt 2

A common way receivers are designed is as follows: we first estimate h (leading to $\hat{h}$), and then determine $\mathbf{a}$, assuming h to be known:

$$\begin{aligned}
\hat{\mathbf{a}} &= \arg\max_{\mathbf{a}} p\left(\mathbf{a}|\, \mathbf{r}, \hat{h}\right) \\
&= \arg\max_{\mathbf{a}} p\left(\mathbf{r}|\, \mathbf{a}, \hat{h}\right) \\
&= \arg\max_{\mathbf{a}} \prod_{k=1}^{N} p\left(r_k|\, a_k, \hat{h}\right) \\
&= \arg\max_{\mathbf{a}} \sum_{k=1}^{N} \log p\left(r_k|\, a_k, \hat{h}\right)
\end{aligned}$$

so that

$$\hat{a}_k = \arg\max_{a_k \in \Omega} \log p\left(r_k|\, a_k, \hat{h}\right).$$

So how do we find $\hat{h}$? Suppose h has a zero-mean Gaussian prior (which may or may not be known). Then

$$\hat{h}_{MAP} = \arg\max_{h} \sum_{\mathbf{a}} p\left(\mathbf{r}|\, \mathbf{a}, h\right) p\left(h\right)$$

and

$$\hat{h}_{ML} = \arg\max_{h} \sum_{\mathbf{a}} p\left(\mathbf{r}|\, \mathbf{a}, h\right)$$

As this summation is over $|\Omega|^N$ sequences, it is again very hard to compute.

2 The EM algorithm

The Expectation-Maximization (EM) algorithm is a technique that solves ML and MAP problems iteratively. To obtain an estimate of a parameter θ, the EM algorithm generates a sequence of estimate $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots$, staring from a well-chose initial estimate $\hat{\theta}^{(0)}$. Many variations exist. We will focus on the most common ones.

2.1 MAP version

The MAP estimate of a parameter θ from an observation x is obtained by maximizing the a posteriori distribution:

$$\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max_{\theta} p\left(\theta|\, x\right) \\
&= \arg\max_{\theta} p\left(x|\, \theta\right) p\left(\theta\right)
\end{aligned}$$

When the solution is hard to obtain, we can introduce so-called missing or unobserved data (say z), in such a way that if z were known, the estimation problem would be easier to solve ($p\left(\theta|\, x, z\right)$ is easy to maximize w.r.t. θ). Once we have introduced the missing data, we can execute the EM algorithm. Starting from an initial estimate of θ, $\hat{\theta}^{(0)}$, the EM algorithm iterates between the E-step and the M-step:

E-step:

$$Q_{MAP}\left(\theta|\, \hat{\theta}^{(n)}\right) = \int p\left(z|\, x, \hat{\theta}^{(n)}\right) \log p\left(z, x, \theta\right) dz. \tag{1}$$

M-step:

$$\hat{\theta}^{(n)} = \arg\max_{\theta} Q_{MAP}\left(\theta | \, \hat{\theta}^{(n)}\right)$$

Hence, the EM algorithm generates a sequence of estimates, $\hat{\theta}^{(0)}, \hat{\theta}^{(1)}, \ldots$. When this sequence converges, we name $\hat{\theta}^{(+\infty)}$ a *solution* of the EM algorithm. The missing data z should be chosen such that the E-step is easy to compute.

Convergence

- Under some regularity conditions, one can show that the EM algorithm converges to an extremum or a saddle point of the a posteriori distribution $p\left(\theta | \, x\right)$
- The a posteriori probability of successive estimates is non-decreasing:

$$\log p\left(\hat{\theta}^{(n+1)} \middle| \, x\right) \geq \log p\left(\hat{\theta}^{(n)} \middle| \, x\right). \tag{2}$$

2.2 ML version

The ML estimate of a parameter θ from an observation x is obtained by maximizing the likelihood function:

$$\hat{\theta}_{ML} \quad = \quad \arg\max_{\theta} p\left(x | \, \theta\right)$$

The ML version is very similar, but uses a slightly different Q-function.

E-step:

$$Q_{ML}\left(\theta | \, \hat{\theta}^{(n)}\right) = \int p\left(z | \, x, \hat{\theta}^{(n)}\right) \log p\left(z, x | \, \theta\right) dz. \tag{3}$$

M-step:

$$\hat{\theta}^{(n)} = \arg\max_{\theta} Q_{ML}\left(\theta | \, \hat{\theta}^{(n)}\right)$$

Convergence

Similar convergence properties: extremum of likelihood, non-decreasing likelihood of successive estimates.

2.3 Relation between MAP and ML version

Since $p\left(z, x, \theta\right) = p\left(z, x | \, \theta\right) p\left(\theta\right)$, we can re-write (1) as

$$\begin{aligned} Q_{MAP}\left(\theta | \, \hat{\theta}^{(n)}\right) &= \int p\left(z | \, x, \hat{\theta}^{(n)}\right) \log\left(p\left(z, x | \, \theta\right) p\left(\theta\right)\right) dz \\ &= \int p\left(z | \, x, \hat{\theta}^{(n)}\right) \log\left(p\left(z, x | \, \theta\right)\right) dz + \int p\left(z | \, x, \hat{\theta}^{(n)}\right) \log\left(p\left(\theta\right)\right) dz \\ &= Q_{ML}\left(\theta | \, \hat{\theta}^{(n)}\right) + \log\left(p\left(\theta\right)\right) \end{aligned}$$

3 Example

As we will see, it is important to simplify the E-step as much as possible. This includes removing terms and factors that do not depend on θ. For this reason we will often (ab)use the notation '$\propto$'. We will assume $h \sim \mathcal{N}\left(0, \sigma_h^2\right)$. This distribution may or may not be known to the receiver.

3.1 Method 1

Let us return to our example. We set $\mathbf{a} \leftrightarrow \theta$ and $h \leftrightarrow z$. Since $\mathbf{a}$ is uniformly distributed over Ω^N, we can use the ML-version of the EM algorithm:

$$Q_{ML}\left(\mathbf{a}|\,\hat{\mathbf{a}}^{(n)}\right) = \int p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right) \log p\left(h, \mathbf{r}|\,\mathbf{a}\right) dh$$

E-step

Since $\mathbf{a}$ and h are independent, $p\left(h, \mathbf{r}|\,\mathbf{a}\right) = p\left(\mathbf{r}|\,h, \mathbf{a}\right) p\left(h\right)$m so that

$$\begin{aligned} Q_{ML}\left(\mathbf{a}|\,\hat{\mathbf{a}}^{(n)}\right) &= \int p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right) \log p\left(\mathbf{r}|\,h, \mathbf{a}\right) dh \\ &+ \int p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right) \log p\left(h\right) dh \end{aligned}$$

As the latter term is independent of $\mathbf{a}$, it will not affect the M-step, so it can be dropped. This gives us

$$Q_{ML}\left(\mathbf{a}|\,\hat{\mathbf{a}}^{(n)}\right) = \int p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right) \log p\left(\mathbf{r}|\,h, \mathbf{a}\right) dh. \tag{4}$$

Let us first focus on $\log p\left(\mathbf{r}|\,h, \mathbf{a}\right)$. Since the noise is iid Gaussian,

$$\begin{aligned} \log p\left(\mathbf{r}|\,h, \mathbf{a}\right) &\propto -\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left(r_k - h a_k\right)^2 \\ &= -\frac{1}{2\sigma^2} \sum_{k=1}^{N} \left(r_k^2 + h^2 a_k^2 - 2 h r_k a_k\right) \\ &\propto -\sum_{k=1}^{N} \left(h^2 a_k^2 - 2 h r_k a_k\right). \end{aligned} \tag{5}$$

Now, we look into $p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right)$. Using Bayes' rule, we know that

$$\begin{aligned} p\left(h, \mathbf{r}|\,\hat{\mathbf{a}}^{(n)}\right) &= p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right) p\left(\mathbf{r}|\,\hat{\mathbf{a}}^{(n)}\right) \\ &\text{and} \\ &= p\left(\mathbf{r}|\,h, \hat{\mathbf{a}}^{(n)}\right) p\left(h|\,\hat{\mathbf{a}}^{(n)}\right) \\ &= p\left(\mathbf{r}|\,h, \hat{\mathbf{a}}^{(n)}\right) p\left(h\right) \end{aligned}$$

so that

$$p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right) = C p\left(\mathbf{r}|\,h, \hat{\mathbf{a}}^{(n)}\right) p\left(h\right)$$

where C is a constant (independent of h). Taking into account (5), it is clear that we only require the first and second order moments of $p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right)$ to evaluate $Q_{ML}\left(\mathbf{a}|\,\hat{\mathbf{a}}^{(n)}\right)$. We find that $p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right)$ is Gaussian with mean M and variance Σ^2. Let us find M and Σ^2:

$$\begin{aligned} p\left(h|\,\mathbf{r}, \hat{\mathbf{a}}^{(n)}\right) &= C p\left(\mathbf{r}|\,h, \hat{\mathbf{a}}^{(n)}\right) p\left(h\right) \\ &= C' \exp\left(-\frac{1}{2\sigma^2} \left\|\mathbf{r} - h\hat{\mathbf{a}}^{(n)}\right\|^2 - \frac{1}{2\sigma_h^2} h^2\right) \\ &= C'' \exp\left(-\frac{1}{2\sigma^2} \left(h^2 \left\|\hat{\mathbf{a}}^{(n)}\right\|^2 - 2h\mathbf{r}^T \hat{\mathbf{a}}^{(n)}\right) - \frac{1}{2\sigma_h^2} h^2\right) \\ &= C'' \exp\left(-\frac{1}{2}\left(h^2\right) \left(\frac{\left\|\hat{\mathbf{a}}^{(n)}\right\|^2}{\sigma^2} + \frac{1}{\sigma_h^2}\right) + \frac{1}{\sigma^2} h\mathbf{r}^T \hat{\mathbf{a}}^{(n)}\right) \end{aligned}$$

which implies that $p\left(h|\,\mathbf{r},\hat{\mathbf{a}}^{(n)}\right)$ is a Gaussian distribution with variance $\Sigma^2 = 1/\left(\left(\frac{\left\|\hat{\mathbf{a}}^{(n)}\right\|^2}{\sigma^2} + \frac{1}{\sigma_h^2}\right)\right)$ so that

$$p\left(h|\,\mathbf{r},\hat{\mathbf{a}}^{(n)}\right) = C'' \exp\left(-\frac{h^2}{2\Sigma^2} + \frac{h}{\Sigma^2}\left(\frac{\Sigma^2}{\sigma^2}\mathbf{r}^T\hat{\mathbf{a}}^{(n)}\right)\right)$$

from which we see that the mean of $p\left(h|\,\mathbf{r},\hat{\mathbf{a}}^{(n)}\right)$ is given by

$$\begin{aligned} M &= \frac{\Sigma^2}{\sigma^2}\mathbf{r}^T\hat{\mathbf{a}}^{(n)} \\ &= \frac{\sigma_h^2}{\left\|\hat{\mathbf{a}}^{(n)}\right\|^2\sigma_h^2 + \sigma^2}\mathbf{r}^T\hat{\mathbf{a}}^{(n)} \end{aligned}$$

Note that when $\sigma_h^2 \to +\infty$, then $M \to \mathbf{r}^T\hat{\mathbf{a}}^{(n)}/\left\|\hat{\mathbf{a}}^{(n)}\right\|^2$, as we would expect. We find that

$$\begin{aligned} \int p\left(h|\,\mathbf{r},\hat{\mathbf{a}}^{(n)}\right) h\,dh &= M \\ &= \tilde{h} \end{aligned}$$

and

$$\begin{aligned} \int p\left(h|\,\mathbf{r},\hat{\mathbf{a}}^{(n)}\right) h^2 dh &= M^2 + \Sigma^2 \\ &= \widetilde{h^2} \end{aligned}$$

Using M and Σ^2 to compute (4):

$$Q_{ML}\left(\mathbf{a}|\,\hat{\mathbf{a}}^{(n)}\right) = -\sum_{k=1}^{N}\left(\widetilde{h^2}a_k^2 - 2\tilde{h}r_k a_k\right).$$

M-step

To maximize this function w.r.t. $\mathbf{a}$, we find that the k-th component is given by

$$\hat{a}_k^{(n+1)} = \arg\max_{a\in\Omega}\left(2\tilde{h}r_k a_k - \widetilde{h^2}a_k^2\right)$$

Comments

- We see that this approach requires knowledge of $p\left(h|\,\mathbf{r},\hat{\mathbf{a}}^{(n)}\right)$. When $p(h)$ is known, this distribution can be found. When $p(h)$ is Gaussian, the EM algorithm leads to an elegant solution.
- When $\Omega = \{-1,+1\}$, we see that $Q_{ML}\left(\mathbf{a}|\,\hat{\mathbf{a}}^{(n)}\right) = \sum_{k=1}^{N}\left(\tilde{h}r_k a_k\right)$, in which case the M-step becomes ML detection of a_k, given a channel estimate $\hat{h} = \tilde{h}$.

3.2 Method 2

As we mentioned earlier, it is sometimes more convenient to estimate h, considering $\mathbf{a}$ as a nuisance parameter. We set $\mathbf{a} \leftrightarrow z$ and $h \leftrightarrow \theta$. We now have a choice to perform MAP estimation or ML estimation, depending on whether or not we know $p(h)$. We focus on MAP. By replacing the prior with a constant (or letting $\sigma_h^2 \to \infty$), we find ML.

E-step

$$Q_{MAP}\left(h|\,\hat{h}^{(n)}\right)=\int p\left(\mathbf{a}|\,\mathbf{r},\hat{h}^{(n)}\right)\log p\left(\mathbf{a},\mathbf{r},h\right)dz \tag{6}$$

Now, $p\left(\mathbf{a},\mathbf{r},h\right)=p\left(\mathbf{r}|\,\mathbf{a},h\right)p\left(\mathbf{a}\right)p\left(h\right)$, so that

$$\begin{aligned}\log p\left(\mathbf{a},\mathbf{r},h\right) &\propto -\frac{1}{2\sigma_h^2}h^2-\frac{1}{2\sigma^2}\left\|\mathbf{r}-h\mathbf{a}\right\|^2\\ &= -\frac{1}{2\sigma_h^2}h^2-\frac{1}{2\sigma^2}h^2\sum_k a_k^2+\frac{1}{\sigma^2}h\sum_k r_k a_k.\end{aligned}$$

We see that to evaluate (6), we require the first and second order moments of the marginals $p\left(a_k|\,\mathbf{r},\hat{h}^{(n)}\right)$:

$$\begin{aligned}Q_{MAP}\left(h|\,\hat{h}^{(n)}\right) &= -\frac{1}{2\sigma_h^2}h^2-\frac{1}{2\sigma^2}\left\{h^2\sum_k\int a_k^2 p\left(\mathbf{a}|\,\mathbf{r},\hat{h}^{(n)}\right)da_k-2h\sum_k r_k\int a_k p\left(\mathbf{a}|\,\mathbf{r},\hat{h}^{(n)}\right)da_k\right\}\\ &= -\frac{1}{2\sigma_h^2}h^2-\frac{1}{2\sigma^2}\left\{h^2\sum_k\int a_k^2 p\left(a_k|\,\mathbf{r},\hat{h}^{(n)}\right)da_k-2h\sum_k r_k\int a_k p\left(a_k|\,\mathbf{r},\hat{h}^{(n)}\right)da_k\right\}\\ &= -\frac{1}{2\sigma_h^2}h^2-\frac{1}{2\sigma^2}\left\{h^2\sum_{k=1}^N\sum_{a\in\Omega}a^2\times p\left(a_k=a|\,\mathbf{r},\hat{h}^{(n)}\right)-2h\sum_{k=1}^N r_k\sum_{a\in\Omega}a\times p\left(a_k=a|\,\mathbf{r},\hat{h}^{(n)}\right)\right\}\\ &= -\frac{1}{2\sigma_h^2}h^2-\frac{1}{2\sigma^2}\left\{h^2\sum_k\widetilde{a_k^2}-2h\sum_k r_k\tilde{a}_k\right\}\end{aligned}$$

So we need to determine $p\left(a_k|\,\mathbf{r},\hat{h}^{(n)}\right)$.

$$\begin{aligned}p\left(a_k|\,\mathbf{r},\hat{h}^{(n)}\right) &= \frac{1}{p\left(\mathbf{r}|\,\hat{h}^{(n)}\right)}p\left(\mathbf{r}|\,a_k,\hat{h}^{(n)}\right)p\left(a_k\right)\\ &= C\exp\left(-\frac{1}{2\sigma^2}\left(r_k-ha_k\right)^2\right)\end{aligned}$$

for some constant C. This constant can be found by determining $f\left(a_k=a\right)=\exp\left(-\frac{1}{2\sigma^2}\left(r_k-ha\right)^2\right)$, $\forall a\in\Omega$, so that $C=1/\sum_{a\in\Omega}f\left(a_k=a\right)$.

M-step

We now need to maximize

$$\begin{aligned}Q_{MAP}\left(h|\,\hat{h}^{(n)}\right) &= -\frac{1}{2\sigma_h^2}h^2-\frac{1}{2\sigma^2}\left\{h^2\sum_k\widetilde{a_k^2}-2h\sum_k r_k\tilde{a}_k\right\}\\ &= h^2\underbrace{\left(-\frac{1}{2\sigma^2}\left(\sum_k\widetilde{a_k^2}+\frac{\sigma^2}{\sigma_h^2}\right)\right)}_{\alpha}-2h\underbrace{\left(-\frac{1}{2\sigma^2}\sum_k r_k\tilde{a}_k\right)}_{\beta}\\ &= h^2\alpha-2h\beta\end{aligned}$$

w.r.t. h, giving

$$\begin{aligned}\hat{h}^{(n+1)} &= \frac{\beta}{\alpha}\\ &= \frac{\sum_k r_k\tilde{a}_k}{\sum_k\widetilde{a_k^2}+\frac{\sigma^2}{\sigma_h^2}}\end{aligned}$$

Comments

- When $p(h)$ is unknown, we find the ML-version of the EM algorithm by simply letting $\sigma_h^2 \to +\infty$:

$$Q_{ML}\left(h|\,\hat{h}^{(n)}\right) = -h^2 \sum_k \widetilde{a_k^2} + 2h \sum_k r_k \tilde{a}_k$$

  and

$$\hat{h}^{(n+1)} = \frac{\sum_k r_k \tilde{a}_k}{\sum_k \widetilde{a_k^2}}.$$

- When $\Omega = \{-1, +1\}$, $a_k^2 = 1$ so that

$$\begin{aligned} Q_{MAP}\left(h|\,\hat{h}^{(n)}\right) &= -\frac{1}{2\sigma_h^2}h^2 - \frac{1}{2\sigma^2}\left\{h^2 N - 2h\sum_k r_k \tilde{a}_k\right\} \\ \hat{h}^{(n+1)} &= \frac{\sum_k r_k \tilde{a}_k}{N + \frac{\sigma^2}{\sigma_h^2}} \end{aligned}$$

  and

$$\begin{aligned} Q_{ML}\left(h|\,\hat{h}^{(n)}\right) &= -h^2 N + 2h\sum_k r_k \tilde{a}_k \\ \hat{h}^{(n+1)} &= \frac{\sum_k r_k \tilde{a}_k}{N} \end{aligned}$$

- After a few iterations (say K), we can make final decisions w.r.t. the data symbols

$$\hat{a}_k = \arg\max_{a \in \Omega} p\left(a_k = a|\,\mathbf{r}, \hat{h}^{(K)}\right)$$