An Improved EM Algorithm and Semi-Blind Channel Identification for Affinely Precoded Communication Systems

Chi Li
chi.li@anu.edu.au

Department of Information Engineering,
Research School of Information Sciences and Engineering,
The Australian National University

September 28, 2006
Outline

Channel Identification
  Channel and precoders
  Channel model
  Geometric property of channel model

EM Algorithm
  Why EM algorithm
  What is EM algorithm
  Disadvantages

EM on Real projective space
  Formulation of the channel problem
  Optimization on real projective space
  The EM algorithm on real projective space

Conclusion and future work
What is the communication system?
What is the channel?
What is the channel identification problem?
What is the precoder?
What is semi-blind channel identification?
Affine precoders

Affine precoder is one of the most popular precoders, the mathematical model is

\[ x = As + b, \quad A \in \mathbb{R}^{(n+L-1) \times p}, \quad b \in \mathbb{R}^{(n+L-1)} \]  \hspace{1cm} (1)

where the affinely precoded symbols \( s \in \mathbb{R}^p \) is introduced to form the encoded symbols \( x \). \( A \) and \( b \) are known to the receiver. Different choices of matrix \( A \) allow many popular precoder schemes including the well-known OFDM (orthogonal frequency division multiplexing).
The channel model is given by

$$z = H_c(h)x + n$$

(2)

where $L$ dimensional vector $h = [h_0, h_1, \ldots, h_{L-1}]^T$ is known as the channel parameter which is to be estimated. $H_c$ here is an $n$-by-$(n + L - 1)$ Toeplitz matrix having $[h_{L-1}, h_{L-2}, \ldots, h_0, 0, \ldots, 0]$ as the first row. $n$ is referred to as additive white Gaussian noise (AWGN) which follows the zero mean normal distribution $\mathcal{N}(0, \Sigma)$, where the covariance matrix $\Sigma = \sigma^2 I_n$. This is the common setting in engineering. For simplicity, we set everything to be real number, however, it is easy to extend to complex numbers.
With the affine precoder, the channel model is updated to

$$z = H_c(As + b) + n \quad (3)$$

In the scheme of semi-blind channel identification, the symbols $s$ are normally distributed with $\mathcal{N}(0, V)$. 
The scale ambiguity problem is associated with all linear and some affine precoders. Scale ambiguity comes in because for any non-zero \( \lambda \in \mathbb{R} \), \( \mathbf{y} = \mathbf{h} * \mathbf{A} \mathbf{s} = (\lambda \mathbf{h}) * \mathbf{A}(\lambda^{-1} \mathbf{s}) \), which means it is not possible to determine the magnitude of the channel.

The ambiguity can be removed if and only if \( \mathbf{b} \) is orthogonal to the range space of \( \mathbf{A} \), that is, \((I - \mathbf{A} \mathbf{A}^+) \mathbf{b} \neq 0\) where \( \mathbf{A}^+ \) is the Moore-Penrose pseudo-inverse.

We only consider the case with scale ambiguity, i.e. \((I - \mathbf{A} \mathbf{A}^+) \mathbf{b} = 0\), even though the other case can also be solved up to an unknown scaling factor.
When scale ambiguity is present, the channel $h$ is treated the same as $h' = \lambda h$.  

*This is the real projective space!*  

Real projective space of dimension $L - 1$, denoted by $\mathbb{RP}^{L-1}$, is the collection of equivalent classes of $\mathbb{R}^L - \{0\}$, where $h, h' \in \mathbb{R}^L - \{0\}$ are equivalent if there exists a $\lambda \in \mathbb{R}$ such that $h' = \lambda h$. An element in $\mathbb{RP}^{L-1}$ is written as $[h]$.

- It is easy to see, the $\mathbb{RP}^1$ is the upper semi-circle.
- Real projective space is the very fundamental framework to study channel identification problem.
- Real projective space is important to algebraic geometry and differential geometry.
Why EM?

- For an incomplete data set, it is impossible to evaluate the complete data log-likelihood. EM algorithm provides an iterative procedure to compute a series of easier log-likelihood;
- Not only the incomplete data, EM algorithm can be used for missing information, truncated distributions, censored or grouped or clustered observations.
- Even there is no missing information, some log-likelihood functions are hard to evaluate. EM algorithm can also be applied.

After the paper by Dempster, Laird, and Rubin (1977), EM algorithm is one of the top 3 fields being studied in the history of statistics.
What is EM?

Let $X$ be the complete data, $Y$ be the incomplete data, $\theta \in \Theta$ is the parameter to be estimated. The parameter space $\Theta$ is assumed to be the Euclidean space.

Let

$$k(x|y; \theta) = \frac{f_c(x; \theta)}{f(y; \theta)}$$

be the conditional density of $X$ given $Y = y$ where $f$ and $f_c$ be the complete and incomplete data pdf respectively.
The incomplete data log-likelihood is given by

\[
\log L(\theta) = \log f(y; \theta) = \log f_c(x; \theta) - \log k(x|y; \theta) = \log L_c(\theta) - \log k(x|y; \theta)
\]

Taking the expectation of both sides w.r.t. the conditional distribution of \(X\) given \(Y = y\), using the fit \(\theta^{(k)}\) for \(\theta\), we have

\[
\log L(\theta) = E_{\theta^{(k)}}(\log L_c(\theta)|y) - E_{\theta^{(k)}}(\log k(x|y; \theta)) = Q(\theta; \theta^{(k)}) - H(\theta; \theta^{(k)})
\]

(5)
Description of EM

It is not hard to show that the function $Q$ defined dominates the log-likelihood function, thus the EM algorithm can be sketched below.

**E-Step** Calculate $Q(\theta; \theta^{(k)})$, where

$$Q(\theta; \theta^{(k)}) = E_{\theta^{(k)}}(\log L_c(\theta) | y) \quad (6)$$

**M-Step** Choose $\theta^{(k+1)}$ to be any value of $\theta \in \Omega$ that maximize $Q(\theta; \theta^{(k)})$; that is

$$\theta^{(k+1)} = \arg \max_{\theta \in \Theta} Q(\theta; \theta^{(k)}). \quad (7)$$
Disadvantages

- Sometimes it fails to converge;
- Sometimes converge very slow. The EM algorithm can have linear convergence and the speed is based on how many information is lost.
- The convergence rate of EM algorithm is the convergence rate of the E-step. That is to say, if the maximum value in M-step can not be easily obtained, the convergence rate will be very slow or fail to converge.

Many speed up methods have been proposed for special problems, but no single algorithm can be used extensively.
Revisit channel model

The affine precoded channel model without scale ambiguity is

$$z = H_c A s + n$$  \hspace{1cm} (8)

Note that $n$ and $s$ are normally distributed. Thus, $z$ is also a Gaussian random vector. Furthermore, $y = [s^T, z^T]^T$ also follows normal distribution with mean 0 and covariance matrix $R$. Let $H = H_c A$,

$$R = \begin{bmatrix} V & VH^T \\ HV & \Sigma + HVH^T \end{bmatrix}$$  \hspace{1cm} (9)
The Q function

The two steps of EM algorithm are to compute Q function and maximize it. In this channel identification problem, Q function can be computed as

\[ Q(h; h^{(k)}) = E[\log f(s, z|h)|z, h^{(k)}] \]  \hspace{1cm} (10)

Notice that \( s|z \) also follows normal distribution \( \mathcal{N}(u, B) \), where \( u \) and \( B \) are functions of \( h \) and is known as constant after each iteration. Define \( c = B + uu^T \) for the later use.

Remember, our goal is to maximize the Q function at each round of EM iteration.
 Optimization on Manifolds

Optimization on manifolds is sometimes called geometric optimization. It has long been studied and recently brought to engineering area by Manton, Hueper et.al.. The key difference between the methods by Manton et.al. and previous researchers is the former does not need the Riemannian structure of manifold, thus reduced the computational complexity.

Manton used parametrization to solve the problem. Local parametrization is the smooth map $\mu_p : \mathbb{R}^n \rightarrow M$ associated with each point $p$ on the manifold $M$.

The idea is: when choosing a search step in an optimization algorithm, use parametrization to project it to Euclidean space where everything is well defined. After that, project it back to the manifold.
After some calculation, to maximize the $Q$ function is equivalent to minimize another cost function $\varphi : \mathbb{RP}^{L-1} \to \mathbb{R}$ defined as

$$\varphi(h) = tr\{(V^{-1} + H^T\Sigma^{-1}H)c^{(k)}\} - 2z^T\Sigma^{-1}Hu^{(k)}$$

(11)

Steepest descent method along with the Armijo line search is used in finding the local minimum.
Tangent space as the local parametrization

One of the natural local parametrization is the projection to the tangent space. The tangent space \( T_{[h]} \) at \([h] \in \mathbb{RP}^{L-1}\) is the \( L-1 \) dimensional subspace \( T_{[h]} = \{v: h^T v = 0\} \) of \( \mathbb{R}^L \).

Given the standard Euclidean inner product, the projection of the gradient onto the tangent space \( T_{[h]} \) is given by

\[
\tilde{\nabla} \varphi(h) = (I - hh^T) \nabla \varphi(h)
\]  

(12)
1. Choose $h^{(0)} \in \mathbb{R}^L$ such that $\|h\| = 1$

2. Compute the E-step which becomes the cost function $\varphi(h)$
   
   2.1 Compute the steepest descent direction
   
   $t = -\widehat{\nabla \varphi(h)} = -(I - hh^T) \nabla \varphi(h)$. If $\|t\|$ is sufficiently small then stop.

   2.2 Set the step size $\gamma := 1$.
   
   If $\varphi(h) - \varphi(h + 2\gamma t) \geq \gamma \|t\|^2$ then set $\gamma := 2\gamma$ and repeat step (2.2).

   2.3 If $\varphi(h) - \varphi(h + \gamma t) < \frac{1}{2} \gamma \|t\|^2$ then set $\gamma := \frac{1}{2} \gamma$ and repeat step (2.3).

   2.4 Set $h := \frac{h + \gamma t}{\|h + \gamma t\|}$. Go to step (2.1).

3. Set $h^{(k+1)} = h$, go to step 2. Stop when $\|h^{(k+1)} - h^{(k)}\|$ is sufficiently small.
Conclusion and future work

- Simulation result shows the algorithm works very well;
- The geometric properties are exploited in the optimization algorithm which result in fast convergence. We are going to extend the algorithm to arbitrary manifold;
- Parameter space $\Theta$ is assumed to be a manifold. We are going to consider the situation when the sample space is also a manifold;
- The steepest descent method is used. We are going to consider Newton-like methods as well;
- We are going to develop the theory of stochastic processes on manifolds to be applied in filtering;
- We are looking for the applications to finance.
Thank You!