A Tour on EM Algorithm

Tianxing He

1 Introduction

Since its formal introduction in 1977 by Dempster et al. [4], the EM algorithm has become a standard methodology for ML estimation. It has steadily become more and more popular and is being used in an increasing number of applications. The first publications in IEEE journals making reference to the EM algorithm appeared in 1988 and dealt with the problem of tomographic reconstruction of photon limited images [18], [12]. Since then, the EM algorithm has become a popular tool for statistical signal processing used in a wide range of applications, such as recovery and segmentation of images and video, image modelling, carrier frequency synchronization, and channel estimation in communications and speech recognition[14].
Assume that x are the observations and θ the unknown parameters of a model that generated x. The ML estimate is obtained as

$$\hat{\theta}_{ML} = \arg\max_{\theta} L(\theta) = ln(p(x;\theta)) \tag{1}$$

However, in many cases the likelihood function $p(x;\theta)$ is complex and is either difficult or immpossible to directly optimize. In such cases the computation of this likelihood is greatly facilitated by the introduction of hidden variables z. These random variables act as links that connect the observations to the unknown parameters via Bayes law. The choice of hidden variables is problem dependent. However, as their name suggests, these variables are not observed and they provide enough information about the observations so that the conditional probability $p(x|z)$ is easy to compute. Apart from this role, hidden variables play another role in statistical modeling. They are an important part of the probabilistic mechanism that is assumed to have generated the observations and can be described very succinctly by a graph that is termed graphical model. Once hidden variables and a prior probability for them $p(z;\theta)$ have been introduced, one can obtain the likelihood or the marginal likelihood as it is called at times by integrating out (marginalizing) the hidden variables according to

$$L(\theta) = lnp(x;\theta) = ln \int p(x,z;\theta) = ln \int P(x|z;\theta)p(z;\theta) \tag{2}$$

Despite the simplicity of the above formulation, in most cases of interest the integral in (2) is either impossible or very difficult to compute in closed form. Thus, the main effort in Bayesian Inference is concentrated on techniques that allow us to bypass or approximately evaluate this integral.
Such methods can be classified into two broad categories[5]. The first is numerical sampling methods also known as Monte Carlo techniques and the second is deterministic approximations. Here we won't address Monte Carlo methods[1], [17]. On the other hand, the EM algorithm is a Bayesian inference methodology that assumes knowledge of the posterior $p(z|x;\theta)$ and iteratively maximizes the likelihood function without explicitly computing it.
One of the most insightful explanations of EM, that provides a deeper understanding of its operation than the intuition of alternating between variables, is in terms of lower-bound maximization

[16].
Assuming now we somehow have a parameter $\theta^{(p)}$, using JenSen's inequaility and following equation(2), we have

$$L(\theta) = ln \int P(x, z; \theta) = ln \int \frac{P(x, z; \theta)q(z)}{q(z)} \geq \int q(z) ln \frac{P(x, z; \theta)}{q(z)} \tag{3}$$

Where $q(z)$ could be any distribution. For EM Algorithm, we take $q(z) = p(z|x; \theta^{(p)})$, so

$$\begin{aligned} \int q(z) ln \frac{P(x, z; \theta)}{q(z)} &= \int p(z|x; \theta^{(p)}) ln \frac{P(x, z; \theta)}{p(z|x; \theta^{(p)})} \\ = \int p(z|x; \theta^{(p)}) ln P(x, z; \theta) &- \int p(z|x; \theta^{(p)}) ln p(z|x; \theta^{(p)}) \end{aligned} \tag{4}$$

Notice the second term is constant. We now define the **auxiliary function**

$$Q(\theta, \theta^{(p)}) = \int p(z|x; \theta^{(p)}) ln P(x, z; \theta) \tag{5}$$

Assume we have $\theta^{(p)}$, in the E-step, we deride the anxiliary function, then in the M-step, we maximize the anxiliary function to get $\theta^{(p+1)}$. One get the inituition of EM in equation (3) : we are maximizing the lower bound of $L(\theta)$.

$$\begin{aligned} &EM\ Alorithm \\ &E - Step : compute\ p(z|x; \theta^{old}) \\ &M - Step : Evaluate\ \theta^{new} = \arg\max_{\theta} Q(\theta, \theta^{old}) \end{aligned} \tag{6}$$

2 Examples

2.1 Guassian Mixture Model

The Guassian Mixture distribution is a linear superposition of Guassians:

$$p(x) = \Sigma_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{7}$$

Subject to:

$$\Sigma_{k=1}^{K} \pi_k = 1 \tag{8}$$

Given a Guassian Mixture model, we introduce K-dimensional binary random variable z which only one element z_k is euqal to 1 and the others are all 0.

$$z = (0, 0, .., 1, 0, ..0) \tag{9}$$

So there are K possible states for z.And we let

$$p(z_k = 1) = \pi_k \tag{10}$$

That is to say,

$$p(z) = \prod_{k=1}^{K} {\pi_k}^{z_k} \tag{11}$$

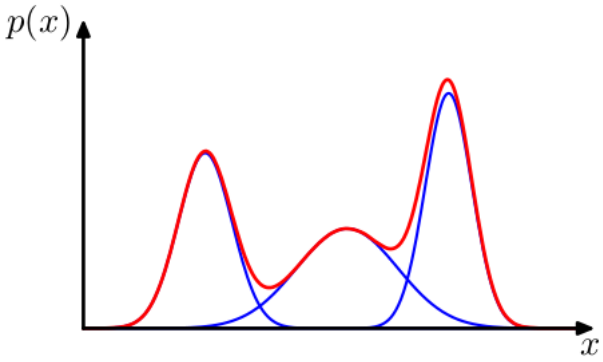

Figure 1: The Guassian Mixture Model

Then, we define the conditional distribution of x given a particular z:

$$p(x|z_k = 1) = \mathcal{N}(x|\mu_k, \Sigma_k) \tag{12}$$

which can also be written as:

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \tag{13}$$

Now we can easily compute the marginal distribution of x

$$\begin{aligned} p(x) = \Sigma_z p(x|z)p(z) &= \Sigma_z \prod_{k=1}^{K} \mathcal{N}(x|\mu_k, \Sigma_k)^{z_k} \prod_{k=1}^{K} {\pi_k}^{z_k} \\ &= \Sigma_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \end{aligned}$$

Now, instead of working with $p(x)$ we can work with $p(x, z) = p(x|z)p(z)$, which will lead to

Figure 2: $p(x) = \Sigma_z p(x, z)$

significant simplification when we are introducing the EM algorithm.
If we have a sequence of observances X, now we apply the EM algorithm.

$$lnp(X|\pi, \mu, \Sigma) = \Sigma_{n=1}^{N} ln\{\Sigma_{k=1}^{K} \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)\} \tag{14}$$

Remembering the anxiliary function is

$$Q(\theta, \theta^{old}) = \int p(z|x; \theta^{old}) lnP(x, z; \theta) \tag{15}$$

For GMM

$$Q(\theta, \theta^{old}) = \Sigma_Z p(Z|X, \theta^{old}) lnp(X, Z|\theta) \tag{16}$$

Z is a vector of length N, indicating which Guassian component each observance is from.
For GMM, we have the constraint that $\Sigma\pi_i = 1$, using the lagrange multiplier, we should maximize

$$\Lambda = \Sigma_Z p(Z|X, \theta^{old}) lnp(X, Z|\theta) + \lambda(\Sigma_k \pi_k - 1) \tag{17}$$

Now we set the derivative of each variable to zero:

$$\frac{\partial \Lambda}{\partial \lambda} = \Sigma_k \pi_k - 1 = 0 \tag{18}$$

For simplicity, we consider each x_i, which means, we consider $lnp(X, Z|\theta)$ as $\Sigma_i ln(p(x_i, z_i))$, unfold, and look at each i. Because the sequence is generated independently, the term$p(Z|X, \theta^{old})$ can be regarded as $Const * \gamma(z_{ik})$.

$$\begin{aligned} \frac{\partial \Lambda}{\partial \pi_k} &= \frac{Const * \partial ln(\pi_k) \Sigma_i \gamma(z_{ik}) + \lambda \pi_k}{\partial \pi_k} \\ &= \frac{1}{\pi_k} \Sigma_i \gamma(z_{ik}) + \lambda = 0 \end{aligned} \tag{19}$$

$$\begin{aligned} \frac{\partial \Lambda}{\partial \mu_k} &= \frac{Const * \Sigma_i \gamma(z_{ik}) \partial(-1/2 * (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k))}{\partial \mu_k} \\ &= Const * \Sigma_i \gamma(z_{ik}) \frac{-1}{2} ((x_i - u_k)^T \Sigma_k^{-1} + \Sigma_k^{-1} (x_i - \mu_k)) \\ &= Const * \Sigma_i \gamma(z_{ik}) ((x_i - \mu_k)^T \Sigma_k^{-1}) = 0 \end{aligned} \tag{20}$$

Note that in the second step we use a constraint that Σ is symmetric, so is Σ^{-1}.
The derivation for $\partial\Sigma$ is more complex, we need to use some conclusion from Appendix A. Now we can get the update formula for each variable by solving these equations:
E-step Evaluate the responsibilities using the current parameter values.

$$\gamma(z_{nk}) = \frac{\mathcal{N}(x_n|\mu_k, \Sigma_k)}{\Sigma_{k=1}^K \pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)} \tag{21}$$

M-step Re-estimate the parameters using the current responsibilities.

$$\begin{aligned} \mu_k^{new} &= \frac{1}{N_k} \Sigma_{n=1}^N \gamma(z_{nk}) x_n \\ \Sigma_k^{new} &= \frac{1}{N_k} \gamma_k(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T \\ \pi_k^{new} &= \frac{N_k}{N} \end{aligned} \tag{22}$$

Likelihood Recalculate the log likelihood function to see if it converges, if not, go to step 1 again. Where

$$N_k = \Sigma_i \gamma(z_{ik}) \tag{23}$$

2.2 HMM-GMM in ASR

2.2.1 Introduce the HMM-GMM model

New machine learning algorithms can lead to significant advances in automatic speech recognition(ASR).The biggest advance occurred nearly four decades ago with the introduction of the EM

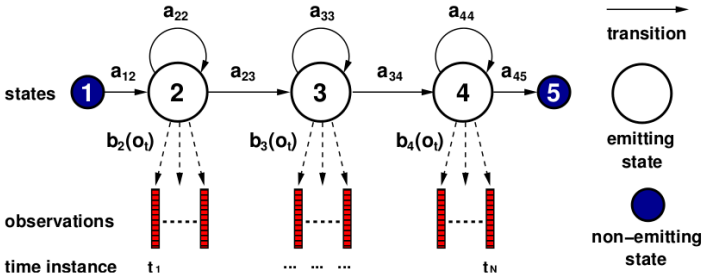

Figure 3: The HMM-GMM model

algorithm for training HMMs[10]. Most current speech recognition systems[8] use hidden Markov models to deal with the temporal variability of speech and Gaussian mixture models(GMM) to determine how well each state of each HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. An alternative way to evaluate the fit is to use a feed-forward neural network that takes several frames of coefficients as input and produces posterior porbabilities over HMM states as output. GMMs have a number of advantages that make them suitable for modeling the probability distributions over vectors of input features that are associated with each state of an HMM. With enough components, they can model probability distributions to any required level of accuracy, and they are fairly easy to fit to data using the EM algorithm.
Despite all their advantages, GMMs have a serious shortcomings. They are statistically[6] inefficient for modeling data that lie on or near a nonlinear manifold in the data space. For example, modeling the set of points that lie very close to the surface of a sphere only requires a few parameters using an appropriate model class, but it requires a very large number of diagonal Gaussians or a fairly large number of full-covariance Gaussians. Speech is produced by modulating a relatively small number of parameters of a dynamical system, and this implies that its true underlying structure is much lower-dimensional than is immediately apparent in a window that contains hundreds of coefficients. It is believed, therefore, that other types of model may work better than GMMs for acoustic modeling if they can more effectively exploit information embedded in a large window of frames.

We use a HMM to model a *context dependent phone*. a_{ij} is the state transition probability. $b(o)$ is a single Guassian distribution that models the observance that this state.
We call $O = (o_1, o_2, ..., o_T)$ the observance of a HMM, and $w = (w_1, w_2, ..., w_T)$ the hidden state sequence.

2.2.2 Apply the EM algorithm

We will now assume that $b(o)$ is Guassian (instead of Guassian Mixture) for simplicity. As before, we unfold the anxiliary function.

$$\begin{aligned} Q_{ML}(M_{k+1}; \hat{M}_k) &= \sum_w log p(O, w|M_{k+1}) p(w|O, \hat{M}_k) \\ &= \sum_w log(a_{w_0 w_1} \prod_t a_{w_{t-1} w_t} b_{w_t}(o_t)) P(w|O, \hat{M}_k) \\ Q_{ML}(M_{k+1}; \hat{M}_k) &= \sum_{t,j} \gamma_j(t) log b_j(o_t) + \sum_{t,i,j} \varepsilon_{ij}(t) log a_{ij} \end{aligned} \tag{24}$$

Where $\gamma_j(t) = P(w_t = j|O, \hat{M}_k)$ and $\varepsilon_{ij}(t) = P(w_{t-1} = i, w_t = j|O, \hat{M}_k)$. And the transistion is simply via take each a and b out and add all out their cooresponding w.
To calculate γ_j and ε_{ij}, we need to calculate $\alpha_j(t) = p(o_{1-t}, w_t = j|\hat{M}_k)$ and $\beta_j(t) = p(o_{t+1-T}|w_t = j, \hat{M}_k)$ They can be calculated efficiently:

$$\begin{aligned} \alpha_j(t) &= (\sum_{i=2}^{N-1} \alpha_i(t-1)a_{ij})b_j(o_t) \\ \beta_j(t) &= \sum_{i=2}^{N-1} a_{ji}b_i(o_{t+1})\beta_i(t+1) \end{aligned} \tag{25}$$

Then we have

$$\begin{aligned} \gamma_j(t) &= \frac{\alpha_j(t)\beta_j(t)}{p(O|\hat{M}_k)} \\ \varepsilon_{ij}(t) &= \frac{\alpha_i(t-1)a_{ij}b_j(o_t)\beta_j(t)}{p(O|\hat{M}_k)} \end{aligned} \tag{26}$$

Having got the γ and ε, we finally could try to maximize

$$Q_{ML}(M_{k+1}; \hat{M}_k) = \sum_{t,j} \gamma_j(t) log b_j(o_t) + \sum_{t,i,j} \varepsilon_{ij}(t) log a_{ij} \tag{27}$$

First let's consider a_{i*}, now we should maxmize $\sum_{t,j} \varepsilon_{ij}(t) log a_{ij}$ subject to $\sum_j a_{ij} = 1$, we could use the lagrange multiplier, I omit the derivation here.

$$\hat{a}_{ij} = \frac{\sum_{t=2}^{T} \varepsilon_{ij}(t)}{\sum_{t=1}^{T} \gamma_i(t)} \tag{28}$$

Now we maximize the other component $\sum_{t,j} \gamma_j(t) log b_j(o_t)$.We set $G(\mu_j, \Sigma_j) = \sum_t \gamma_j(t) log b_j(o_t)$. And we maximize it by solving $\frac{\partial G}{\partial \mu_j} = 0$. Then we get,

$$\hat{\mu_j} = \frac{\sum_t (\gamma_j(t) o_t)}{\sum_t \gamma_j(t)} \tag{29}$$

3 Properities of the EM algorithm

3.1 Monotonous Increase of the Likelihood

We follow the exposition of the EM in [2] and [15]. It is straightforward to show that the log-likelihood can be written as

$$ln p(x; \theta) = F(q, \theta) + KL(q||p) \tag{30}$$

where

$$F(q, \theta) = \int q(z) ln \frac{p(x, z; \theta)}{q(z)} \tag{31}$$

and

$$KL(q||p) = \int -q(z) ln \frac{p(z|x; \theta)}{q(z)} \tag{32}$$

Notice that we always have $KL(q||p) \geq 0$ as long as q is a distribution.
Now, assuming we have a $\theta^{(p)}$, and let $q(z) = p(z|x;\theta^{(p)})$, we can get $\theta^{(p+1)}$ through an iteration:

$$\theta^{(p+1)} = \arg\max_{\theta} F(q,\theta) \tag{33}$$

Notice this is equal to maximizing the anxiliary function.
Now, one can easily find out that if we just take $\theta^{(p+1)} = \theta^{(p)}$, we have$F(q,\theta^{(p+1)}) = L(\theta^{(p)})$, because $KL(q||p) = 0$. Since we are maximizing F, we have

$$F(q,\theta^{(p+1)}) \geq L(\theta^{(p)}) \tag{34}$$

Remembering that we always have $KL(q||p) \geq 0$, we get

$$L(\theta^{(p+1)}) \geq L(\theta^{(p)}) \tag{35}$$

3.2 Convergence to a Local Maxima

With the exception of a few specific cases, the EM algorithm is not guaranteed to converge to a global maximizer of the likehood[3, 20]. Under some regularity conditions on the likelihood $L(\theta)$ and on the parameters set Θ, it is possible, however, to show the parameter sequence obtained by EM converges to a local maximizer of $L(\theta)$ or at least, to a stationary point of $L(\theta)$. Necessary conditions for the convergence of the EM algorithm and related theorems can be found in [19].

4 Variants of the EM Algorithm

4.1 Penalized Likelihood Estimation

The EM algorithm can be straightforwardly modified to compute penalized likelihood estimates, that is, estimates of the form

$$\theta = \arg\max_{\theta}[L(\theta) + G(\theta)] \tag{36}$$

The penalty term $G(\theta)$ could represent, for example, the logarithm of a prior on if a Bayesian approach is used and the maximum a posteriori(MAP) estimate of is desired instead of the ML estimate.(Someone put it "poor mans bayesian inference[5]") The EM algorithm for penalized-likelihood estimation can be obtained by replacing the M-step with

$$\theta^{(p+1)} = \arg\max_{\theta}[Q(\theta,\theta^{(p)}) + G(\theta)] \tag{37}$$

It is straightforward to see that the monotonicity property is preserved. Examples of penalized likelihood estimation to be found in [7].

4.2 Variational EM

One can bypass the requirement of exactly knowing $p(z|x;\theta)$ by assuming an appropriate q(z) in the decomposition (30)[5]. In the E-step q(z) is found such that it maximizes $F(q,\theta)$ keeping θ fixed. To perform this maximization, a particular form of q(z) must be assumed. In certain cases it is possible to assume knowledge of the form of $q(z;\omega)$, where is a set of parameters. Thus, the

lower bound $F(\omega, \theta)$ becomes a function of these parameters and is maximized with respect to in the E-step and with respect to in the M-step. The process can be summarized as below[11]:

$$\begin{gathered} Variational\ EM \\ Variational\ E-Step: compute\ \hat{q} = \arg\max_{q} F(q, \theta^{old}) \\ Variational\ M-Step: Evaluate\ \theta^{new} = \arg\max_{\theta} F(\hat{q}, \theta) \end{gathered} \tag{38}$$

One can easily deduce the monotonicity property is still preserved, since we can first prove

$$F(\hat{q}, \theta^{old}) \geq L(\theta^{old}) \tag{39}$$

and then

$$F(\hat{q}, \theta^{new}) \geq F(\hat{q}, \theta^{old}) \tag{40}$$

because of the argmax.
Although there are no approximations in the variational theory, variational methods can be used to find approximate solutions in Bayesian inference problems. This is done by assuming that the functions over which optimization is performed have specific forms. For example, we can assume only quadratic functions or functions that are linear combinations of fixed basis functions. For Bayesian inference, a particular form that has been used with great success is the factorized one, see [9] and [13]. In which case we assume

$$q(z) = \prod_{i=1}^{M} q_i(z_i) \tag{41}$$

From that assumption we could deride [5]

$$q_j^*(z_j) = \frac{exp(< lnp(x, z; \theta) >_{i \neq j})}{\int exp(< lnp(x, z; \theta) >_{i \neq j}) dz_j} \tag{42}$$

5 Appendix

5.1 A:The derivative of the Guassian likelihood with respect to covariance matrix

First we list some math tools we will use:

$$\frac{\partial}{\partial A} ln|A| = (A^{-1})^T \tag{43}$$

$$\frac{\partial [X^{-1}A]}{\partial X} = -X^{-1}A^T X^{-1} \tag{44}$$

We have the Guassian likelihood:

$$ln(\mathcal{N}(x|\mu, \Sigma)) = -\frac{D}{2} ln(2\pi) + (-\frac{1}{2}) ln(|\Sigma|) + (-\frac{1}{2})(x-\mu)^T \Sigma^{-1} (x-\mu) \tag{45}$$

Now we differentiate it with respect to Σ, we get

$$\begin{aligned} \frac{\partial ln(\mathcal{N}(x|\mu, \Sigma))}{\partial \Sigma} &= \Sigma^{-1} + \frac{\partial Tr[\Sigma^{-1}(x-\mu)(x-\mu)^T]}{\partial \Sigma} \\ &= \Sigma^{-1} - \Sigma^{-1}(x-\mu)(x-\mu)^T \Sigma^{-1} \end{aligned} \tag{46}$$

By setting the derivative to zero, we get

$$\Sigma = (x-\mu)(x-\mu)^T \tag{47}$$

References

[1] A. Doucet C. Andrieu, N. de Freitas and M. Jordan. An introduction to mcmc for machine learning. *Mach. Learn.*, pages vol. 50, no. 1, pp. 543, 2003.

[2] C.Bishop. *Pattern Recongnition and Machine Learning*. Springer-Verlag, 2006.

[3] Christophe Couvreur. The em algorithm: A guided tour. In *2d IEEE Europen Workshop on Computationaly Intensive Methods in Control and Signal Processing*, 1996.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Roy. Statis. Soc. A*, pages vol. 39, no. 1, pp. 138, 1977.

[5] Aristidis C. Likas Dimitris G. Tzikas and Nikolaos P. Galatsanos. The variational approximation for bayesian inference. *IEEE SIGNAL PROCESSING MAGAZINE*, page vol. 131, 2008.

[6] Yu Dong Andrew Senior Vincent Vanhoucke Abdel-rahman Mohamed Navdeep Patrick Nguyen Tara Sainath and Brian Kingsbury GeoffreyJaitly, Li Deng. Deep neural networks for acoustic modeling in speech recognition. *IEEE SIGNAL PROCESSING MAGAZINE*, 2012.

[7] Peter J. Green. On use of the em for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B*, 1990.

[8] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*, 2012.

[9] T.S. Jaakkola. Variational methods for inference and learning in graphical models. *Ph.D. dissertation, Dept. Elect. Eng. Comp. Sci., MIT*, 1997.

[10] James Glass Sanjeev Khudanpur Chin-Hui Lee Nelson Morgan Janet M. Baker, Li Deng and Douglas OShaughnessy. Developments and directions in speech recognition and understanding. *IEEE SIGNAL PROCESSING MAGAZINE*, 2009.

[11] J.Sun. A gentle introduction to latent variable model and variational em algorithm. 2009.

[12] Z. Liang and H. Hart. Bayesian reconstruction in emission computerized tomography. *IEEE Trans. Nuclear Sci*, pages vol. 35, no. 1, pp. 788792, 1988.

[13] T. Jaakkola M. Jordan, Z. Gharamani and L. Saul. An introduction to variational methods for graphical models. *Learning in Graphical Models*, pages 105–162, 1998.

[14] G. McLachlan and T. Krishnan. The em algorithm and extensions. *Wiley*, 1997.

[15] R.M. Neal and G.E. Hinton. A view of the em algorithm that justifies incremental, sparse and other variants. *Learning in Graphical Models*, 1998.

[16] Neal.R and Hinton.G. A view of the em algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, 1998.

[17] C. Robert and G. Cassela. *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.

[18] L. Byars W. Jones and M. Casey. Positron emission tomographic images and expectation maximization: A vlsi architecture for multiple iterations per second. *IEEE Trans. Nuclear Sci*, pages vol. 35, no. 1, pp. 620624, 1977.

[19] C.F.Jeff Wu. On the convergence properties of the em algorithm. *The Annals of Statics*, pages vol.11 No.1 pp.95–103, 1983.

[20] Lei Xu and Michael I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural Computation*, 8:129–151, 1995.