The EM Algorithm

Pramook Khungurn

August 26, 2015

This is the "untangled" version of "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models" by Jeff A. Blimes. Even the tutorial is very gentle to begin with, I had a hard time reading and understanding it without writing this document up as I read.

1 The Problem

- We have a density function $p(\mathbf{x}|\Theta)$ governed by a set of parameters Θ.
  We also have a data set $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ drawn i.i.d. from the distribution.
- The density of the sampled data $\mathcal{X}$ is given by:

$$p(\mathcal{X}|\Theta) = \prod_{i=1}^{N} p(\mathbf{x}_i|\Theta).$$

- We define $\mathcal{L}(\Theta|\mathcal{X}) = p(\mathcal{X}|\Theta)$ and view it as a function of Θ.
  We call it the *likelihood* of the parameter Θ.
- In the *maximum likelihood problem*, we wish to find Θ that maximizes $\mathcal{L}(\Theta|\mathcal{X})$. That is, we want to find

$$\Theta^* = \arg\max_{\Theta} \mathcal{L}(\Theta|\mathcal{X}).$$

- However, the EM algorithm tries to maximize $\log(\mathcal{L}(\Theta|\mathcal{X}))$ instead because it's much easier.
- Depending on the form of $p(\mathcal{X}|\Theta)$, the problem can be easy or hard. The EM algorithm at least works well when p is a mixture of Gaussians or Hidden Markov Model.

2 General EM

- The EM algorithm is a general method of finding the maximum-likelihood estimate of parameters of an underlying distribution from a given data set when the data is incomplete or has missing values.
- The algorithm has two common application scenarios:
  - The data is actually missing due to the observation process.
  - The likelihood function is analytically intractable but can be simplified by assuming additional but *missing* hidden parameters.
- $\mathcal{X}$ denotes the observed data, which we also call the *incomplete data*.

- We assume a complete data set $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ exists. That is, each data item $\mathbf{z}_i$ has two components: $\mathbf{x}_i$ and $\mathbf{y}_i$, where $\mathbf{x}_i$ is the observed data and $\mathbf{y}_i$ is the hidden or missing data.
- We also assume the complete data set is generated by a joint distribution function:

  $$p(\mathbf{z}|\Theta) = p(\mathbf{x}, \mathbf{y}|\Theta) = p(\mathbf{y}|\mathbf{x}, \Theta)p(\mathbf{x}|\Theta).$$

  The joint density often "arise" from the marginal density $p(\mathbf{x}|\Theta)$ and the assumption of hidden variables and parameter value guesses.
- We can now define a likelihood function:

  $$\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) = p(\mathcal{X}, \mathcal{Y}|\Theta).$$

  This function is actually a random variable because the hidden data $\mathcal{Y}$ is unknown. Thus, we can think of the new likelihood function as

  $$\mathcal{L}(\Theta|\mathcal{Z}) = h_{\mathcal{X},\Theta}(\mathcal{Y})$$

  for some some function $h_{\mathcal{X},\Theta}(\cdot)$ where $\mathcal{X}$ and Θ are constants and $\mathcal{Y}$ is a random variable.
- The orignal likelihood function $\mathcal{L}(\Theta|\mathcal{X})$ is called the *incomplete-data likelihood function*.
- The EM algorithm is an iterative refinement algorithm. Suppose that we already have a parameter estimate $\Theta^{(i-1)}$. It will find a better estimate $\Theta^{(i)}$ and iterate this process until convergence.
- To find a better estimate, it computes the expected value of the complete-data log-likelihood function with respect to the unknown variable $\mathcal{Y}$ using the observed data $\mathcal{X}$ and the current parameter estmate $\Theta^{(i-1)}$:

  $$Q(\Theta, \Theta^{(i-1)}) = E\big[\log(\mathcal{X}, \mathcal{Y}|\Theta) \;\big|\; \mathcal{X}, \Theta^{(i-1)}\big].$$

  Here,
  - Θ is a parameter we would like to adjust.
  - $\Theta^{(i-1)}$ and $\mathcal{X}$ are constants.
  - $\mathcal{Y}$ is a random variable which is distributed by some marginal distribution $f(\mathcal{Y}|\mathcal{X}, \Theta^{(i-1)})$.

  The RHS can be rewritten as:

  $$E\big[\log(\mathcal{X}, \mathcal{Y}|\Theta) \;\big|\; \mathcal{X}, \Theta^{(i-1)}\big] = \int_{\mathcal{Y}\in\mathbb{Y}} \log\big(p(\mathcal{X}, \mathcal{Y}|\Theta)\big) f(\mathcal{Y}|\mathcal{X}, \Theta^{(i-1)})\, \mathrm{d}\mathcal{Y}$$

  where $\mathbb{Y}$ is the set of all values $\mathcal{Y}$ can take on.
- In the best case, the marginal distribution f is a simple expression.

  In the worst case, the density is hard to obtain.

  Often, the density used is actually is $f(\mathcal{Y}, \mathcal{X}|\Theta^{(i-1)}) = f(\mathcal{Y}|\mathcal{X}, \Theta^{(i-1)})f(\mathcal{X}|\Theta^{(i-1)})$. This distribution does not affect subsequent step because $f(\mathcal{X}|\Theta^{(i-1)})$ is a constant.
- The determination of the expected value $Q(\Theta, \Theta^{(i-1)})$ is called the *E-step* of the algorithm.
- The second step is to maximize the expectation we just computed. That is, the new parameter estimate is given by:

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

  This step is called the *M-step*.
- Each iteration of the algorithm consists of an E-step followed by an M-step.
- An iteration is guaranteed to increase the log-likelihood, and the algorithm is guaranteed to converge.

3 Learning Mixtures Models by EM

- We assume that the distribution p is given by:

$$p(\mathbf{x}|\Theta) = \sum_{i=1}^{M} \alpha_i p_i(\mathbf{x}|\theta_i)$$

  where the parameters are $\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_M, \theta_1, \theta_2, \ldots, \theta_M)$ with the constraint that $\sum_{i=1}^{M} \alpha_i = 1$. In other words, we have M probability distributions that are mixed together with weights α_i.

- The incomplete-data log-likelihood function for this density is given by:

$$\log(\mathcal{L}(\Theta|\mathcal{X})) = \log\left(\prod_{i=1}^{N} p(\mathbf{x}_i|\Theta)\right) = \sum_{i=1}^{N} \log(p(\mathbf{x}_i|\Theta)) = \sum_{i=1}^{N} \log\left(\sum_{j=1}^{M} \alpha_j p_j(\mathbf{x}_i|\theta_j)\right)$$

  which is difficult to optimize because it contains the log of the sum.

- We now think of $\mathcal{X}$ as incomplete. We define the hidden data $\mathcal{Y} = (y_1, y_2, \ldots, y_N)$ so that y_i is the "component" which generates $\mathbf{x}_i$. In other words, $y_i \in \{1, 2, \ldots, M\}$, and $y_i = k$ if the ith observed data was generated by the kth mixture component. Note that the set $\mathbb{Y}$ of values $\mathcal{Y}$ can take on is the set $\{1, 2, \ldots, M\}^N$.

- If we know the values of $\mathcal{Y}$, the likelihood becomes:

$$\log(\mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y})) = \sum_{i=1}^{N} \log(\alpha_{y_i} p_{y_i}(\mathbf{x}_i|\theta_{y_i})).$$

  Depending on the form of the component densities, the above log-likelihood might be maximized easily.

- The problem is that we don't know $\mathcal{Y}$, so we assume that it is a random variable.

- Let $\Theta^g = (\alpha_1^g, \alpha_2^g, \ldots, \alpha_M^g, \theta_1^g, \theta_2^g, \ldots, \theta_M^g)$ be the current (guessed) parameter estimation.

  Given Θ^g, we can compute $p_j(\mathbf{x}_i|\theta_j^g)$ for each i and j. The coefficient α_j can be thought of as a probability for component j. By Bayes's rule, we have that

$$p(y_i|\mathbf{x}_i, \Theta^g) = \frac{\alpha_{y_i}^g p_{y_i}(\mathbf{x}_i|\theta_{y_i}^g)}{p(\mathbf{x}_i|\theta^G)} = \frac{\alpha_{y_i}^g p_{y_i}(\mathbf{x}_i|\theta_{y_i}^g)}{\sum_{k=1}^{M} \alpha_k^g p_k(\mathbf{x}_i|\theta_k^g)}.$$

  Moreover,

$$p(\mathcal{Y}|\mathcal{X}, \Theta^g) = \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \Theta^g).$$

- We now have the expression for $Q(\Theta, \Theta^g)$:

$$\begin{aligned}
Q(\Theta, \Theta^g) &= E[\log(\mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \Theta^g] \\
&= \sum_{\mathcal{Y}\in\mathbb{Y}} \log(\mathcal{L}(\Theta|\mathcal{X}, \mathcal{Y}))p(\mathcal{Y}|\mathcal{X}, \Theta^g) = \sum_{\mathcal{Y}\in\mathbb{Y}} \log(p(\mathcal{X}, \mathcal{Y}|\Theta))p(\mathcal{Y}|\mathcal{X}, \Theta^g) \\
&= \sum_{\mathcal{Y}\in\mathbb{Y}} \Big(\sum_{i=1}^{N} \log(\alpha_{y_i} p_{y_i}(\mathbf{x}_i|\theta_{y_i})) \Big) \Big(\prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \Big) \\
&= \sum_{\mathcal{Y}\in\mathbb{Y}} \Big(\sum_{i=1}^{N} \log(\alpha_{y_i} p_{y_i}(\mathbf{x}_i|\theta_{y_i})) \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \Big) \\
&= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \Big(\sum_{i=1}^{N} \log(\alpha_{y_i} p_{y_i}(\mathbf{x}_i|\theta_{y_i})) \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \Big) \\
&= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \Big(\sum_{i=1}^{N} \sum_{\ell=1}^{M} \delta_{\ell,y_i} \log(\alpha_\ell p_\ell(\mathbf{x}_i|\theta_\ell)) \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \Big) \\
&= \sum_{i=1}^{N} \sum_{\ell=1}^{M} \log(\alpha_\ell p_\ell(\mathbf{x}_i|\theta_\ell)) \Big(\sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{\ell,y_i} \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \Big)
\end{aligned}$$

  Now, note that

$$\begin{aligned}
&\sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{\ell,y_i} \prod_{j=1}^{N} p(y_j|x_j, \Theta^g) \\
&= \Big(\sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_{i-1}=1}^{M} \sum_{y_{i+1}=1}^{M} \cdots \sum_{y_N=1}^{M} \prod_{j\neq i} p(y_j|x_j, \Theta^g) \Big) p(\ell|\mathbf{x}_i, \Theta^g) \\
&= p(\ell|\mathbf{x}_i, \Theta^g) \prod_{j\neq i} \Big(\sum_{y_j=1}^{M} p(y_j|x_j, \Theta^g) \Big) \\
&= p(\ell|\mathbf{x}_i, \Theta^g)
\end{aligned}$$

  because $\sum_{y_j=1}^{M} p(y_j|x_j, \Theta^g) = 1$. Thus, we have

$$\begin{aligned}
Q(\Theta, \Theta^g) &= \sum_{i=1}^{N} \sum_{\ell=1}^{M} \log(\alpha_\ell p_{y_i}(\mathbf{x}_i|\theta_\ell)) p(\ell|\mathbf{x}_i, \Theta^g) \\
&= \sum_{i=1}^{N} \sum_{\ell=1}^{M} \log(\alpha_\ell) p(\ell|\mathbf{x}_i, \Theta^g) + \sum_{i=1}^{N} \sum_{\ell=1}^{M} \log(p_\ell(\mathbf{x}_i|\theta_\ell)) p(\ell|\mathbf{x}_i, \Theta^g) \qquad (1)
\end{aligned}$$

- Now, we have to optimize the above expression. We can optimize the expression involving α_ℓ and the one involving θ_ℓ separately.

- To optimize α_ℓ, we introduce Lagrange multiplier λ with the constraint that $\sum_\ell \alpha_\ell = 1$, and solve the following equation:

$$\frac{\partial}{\partial \alpha_\ell} \Big[\sum_{\ell=1}^{M} \sum_{i=1}^{N} \log(\alpha_\ell) p(\ell|\mathbf{x}_i, \Theta^g) + \lambda \Big(\sum_{\ell} \alpha_\ell - 1 \Big) \Big] = 0$$

or

$$\sum_{i=1}^{N} \frac{1}{\alpha_\ell} p(\ell|\mathbf{x}_i, \Theta^g) + \lambda = 0$$
$$\sum_{i=1}^{N} p(\ell|\mathbf{x}_i, \Theta^g) + \alpha_\ell \lambda = 0$$
$$\sum_{\ell=1}^{M}\sum_{i=1}^{N} p(\ell|\mathbf{x}_i, \Theta^g) + \sum_{\ell=1}^{M} \alpha_\ell \lambda = 0$$
$$\sum_{i=1}^{N}\sum_{\ell=1}^{M} p(\ell|\mathbf{x}_i, \Theta^g) + \lambda \sum_{\ell=1}^{M} \alpha_\ell = 0$$
$$\sum_{i=1}^{N} 1 + \lambda = 0$$
$$\lambda = -N.$$

This means that

$$\alpha_\ell = \frac{1}{N} \sum_{i=1}^{N} p(\ell|\mathbf{x}_i, \Theta^g).$$

- Optimizing for θ_ℓ depends on the form of p_ℓ. An important special case where we can find θ_ℓ analytically is given in the next section.

4 Learning Gaussian Mixture Models with EM

- In the Gaussian mixture model, the component density assumes the form:

$$p_\ell(\mathbf{x}|\mu_\ell, \Sigma_\ell) = \frac{1}{(2\pi)^{d/2}(\det \Sigma_\ell)^{1/2}} \exp\bigg(-\frac{1}{2}(\mathbf{x}-\mu_\ell)^T \Sigma_\ell^{-1}(\mathbf{x}-\mu_\ell)\bigg).$$

  Here, the parameters θ_ℓ is (μ_ℓ, Σ_ℓ).

- We now wish to find the update expression for μ_ℓ and Σ_ℓ.

- We have that

$$\sum_{\ell=1}^{M}\sum_{i=1}^{N} \log(p_\ell(\mathbf{x}_i|\theta_\ell)) p(\ell|\mathbf{x}_i, \Theta^g)$$
$$= \sum_{\ell=1}^{M}\sum_{i=1}^{N} \bigg(-\frac{1}{2}\log(\det \Sigma_\ell) - \frac{1}{2}(\mathbf{x}_i-\mu_\ell)^T \Sigma_\ell^{-1}(\mathbf{x}_i-\mu_\ell) - \frac{d}{2}\log(2\pi)\bigg) p(\ell|\mathbf{x}_i, \Theta^g). \tag{2}$$

- For μ_ℓ, we compute the following partial derivative:

$$\frac{\partial}{\partial \mu_\ell}\bigg(\sum_{\ell=1}^{M}\sum_{i=1}^{N} \bigg(-\frac{1}{2}\log(\det \Sigma_\ell) - \frac{1}{2}(\mathbf{x}_i-\mu_\ell)^T \Sigma_\ell^{-1}(\mathbf{x}_i-\mu_\ell) - \frac{d}{2}\log(2\pi)\bigg) p(\ell|\mathbf{x}_i, \Theta^g)\bigg)$$
$$= \frac{\partial}{\partial \mu_\ell}\bigg(\sum_{i=1}^{N} \bigg(-\frac{1}{2}(\mathbf{x}_i-\mu_\ell)^T \Sigma_\ell^{-1}(\mathbf{x}_i-\mu_\ell)\bigg) p(\ell|\mathbf{x}_i, \Theta^g)\bigg)$$
$$= -\frac{1}{2}\sum_{i=1}^{N} \bigg(\frac{\partial}{\partial \mu_\ell}(\mathbf{x}_i-\mu_\ell)^T \Sigma_\ell^{-1}(\mathbf{x}_i-\mu_\ell)\bigg) p(\ell|\mathbf{x}_i, \Theta^g).$$

Using Lemma 5.2, we have that

$$\frac{\partial}{\partial \mu_\ell}(\mathbf{x}_i - \mu_\ell)^T \Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell) = (\Sigma^{-1} + \Sigma^{-T})(\mathbf{x}_i - \mu_\ell)\frac{\partial(\mathbf{x}_i - \mu_\ell)}{\partial \mu_\ell} = -2\Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell).$$

The simplification $\Sigma_\ell^{-1} + \Sigma_\ell^{-T} = 2\Sigma_\ell^{-1}$ comes from the fact that Σ_ℓ is a symmetric matrix, so is its inverse. So, we have that the partial derivative is given by:

$$-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{\partial}{\partial \mu_\ell}(\mathbf{x}_i - \mu_\ell)^T \Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell)\right)p(\ell|\mathbf{x}_i, \Theta^g) = \sum_{i=1}^{N}\Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell)p(\ell|\mathbf{x}_i, \Theta^g).$$

Setting the above expression equal to zero, we have:

$$\begin{aligned}
\sum_{i=1}^{N}\Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell)p(\ell|\mathbf{x}_i, \Theta^g) &= 0 \\
\Sigma_\ell^{-1}\sum_{i=1}^{N}(\mathbf{x}_i - \mu_\ell)p(\ell|\mathbf{x}_i, \Theta^g) &= 0 \\
\sum_{i=1}^{N}(\mathbf{x}_i - \mu_\ell)p(\ell|\mathbf{x}_i, \Theta^g) &= 0 \\
\sum_{i=1}^{N}\mathbf{x}_i p(\ell|\mathbf{x}_i, \Theta^g) - \sum_{i=1}^{N}\mu_\ell p(\ell|\mathbf{x}_i, \Theta^g) &= 0 \\
\mu_\ell \sum_{i=1}^{N} p(\ell|\mathbf{x}_i, \Theta^g) &= \sum_{i=1}^{N}\mathbf{x}_i p(\ell|\mathbf{x}_i, \Theta^g) \\
\mu_\ell &= \frac{\sum_{i=1}^{N}\mathbf{x}_i p(\ell|\mathbf{x}_i, \Theta^g)}{\sum_{i=1}^{N} p(\ell|\mathbf{x}_i, \Theta^g)}.
\end{aligned}$$

- For Σ_ℓ, we rewrite (2), dropping the term involing $d/2$ because it doesn't show up in the derivative.

$$\begin{aligned}
&\sum_{\ell=1}^{M}\sum_{i=1}^{N}\left(-\frac{1}{2}\log(\det \Sigma_\ell) - \frac{1}{2}(\mathbf{x}_i - \mu_\ell)^T\Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell)\right)p(\ell|\mathbf{x}_i, \Theta^g) \\
&= \frac{1}{2}\sum_{\ell=1}^{M}\left(\log(\det(\Sigma_\ell^{-1}))\sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g) - \sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g)(\mathbf{x}_i - \mu_\ell)^T\Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell)\right).
\end{aligned}$$

Now, using Lemma 5.3, we rewrite the above expression further:

$$\begin{aligned}
&\frac{1}{2}\sum_{\ell=1}^{M}\left(\log(\det(\Sigma_\ell^{-1}))\sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g) - \sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g)(\mathbf{x}_i - \mu_\ell)^T\Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell)\right) \\
&= \frac{1}{2}\sum_{\ell=1}^{M}\left(\log(\det(\Sigma_\ell^{-1}))\sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g) - \sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g)\mathrm{tr}\left(\Sigma_\ell^{-1}(\mathbf{x}_i - \mu_\ell)(\mathbf{x}_i - \mu_\ell)^T\right)\right) \\
&= \frac{1}{2}\sum_{\ell=1}^{M}\left(\log(\det(\Sigma_\ell^{-1}))\sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g) - \sum_{i=1}^{N}p(\ell|\mathbf{x}_i, \Theta^g)\mathrm{tr}\left(\Sigma_\ell^{-1}N_{\ell,i}\right)\right)
\end{aligned}$$

where $N_{\ell,i} = (\mathbf{x}_i - \mu_\ell)(\mathbf{x}_i - \mu_\ell)^T$.

We take the derivative of the above expression with respect to Σ_ℓ^{-1}, making use of Lemma 5.7 and Lemma 5.8:

$$\begin{aligned}
&\frac{1}{2}\Big(\sum_{i=1}^{N} p(\ell|\mathbf{x}_i,\Theta^g)(2\Sigma_\ell - \mathrm{diag}(\Sigma_\ell)) - \sum_{i=1}^{N} p(\ell|\mathbf{x}_i,\Theta^g)(2N_{\ell,i} - \mathrm{diag}(N_{\ell,i}))\Big)\\
&= \frac{1}{2}\Big(2\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)(\Sigma_\ell - N_{\ell,i}) - \mathrm{diag}\Big(\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)(\Sigma_\ell - N_{\ell,i})\Big)\Big)\\
&= \frac{1}{2}(2S - \mathrm{diag}(S))
\end{aligned}$$

where $S = \sum_{i=1}^{N} p(\ell|x_i,\Theta^g)(\Sigma_\ell - N_{\ell,i})$. Setting the derivative equal to zero, we have the equation $2S - \mathrm{diag}(S) = 0$ and consequently $S = 0$. Hence,

$$\begin{aligned}
&\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)(\Sigma_\ell - N_{\ell,i}) = 0\\
&\Sigma_\ell \sum_{i=1}^{N} p(\ell|x_i,\Theta^g) = \sum_{i=1}^{N} p(\ell|x_i,\Theta^g) N_{\ell,i}\\
&\Sigma_\ell = \frac{\sum_{i=1}^{N} p(\ell|x_i,\Theta^g) N_{\ell,i}}{\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)} = \frac{\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)(\mathbf{x}_i - \mu_\ell)(\mathbf{x}_i - \mu_\ell)^T}{\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)}.
\end{aligned}$$

- Hence, the update rules for learning mixture of Gaussians is:

$$\begin{aligned}
\alpha_\ell^{\text{new}} &= \frac{1}{N}\sum_{i=1^N} p(\ell|\mathbf{x}_i,\Theta^g)\\
\mu_\ell^{\text{new}} &= \frac{\sum_{i=1}^{N} \mathbf{x}_i p(\ell|\mathbf{x}_i,\Theta^g)}{\sum_{i=1}^{N} p(\ell|\mathbf{x}_i,\Theta^g)}\\
\Sigma_\ell^{\text{new}} &= \frac{\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)(\mathbf{x}_i - \mu_\ell^{\text{new}})(\mathbf{x}_i - \mu_\ell^{\text{new}})^T}{\sum_{i=1}^{N} p(\ell|x_i,\Theta^g)}.
\end{aligned}$$

5 Some Matrix Identities

- The last section used some matrix identities which might not be obvious. We prove them in this section.
- **Lemma 5.1.** $\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j$

Proof.

$$\begin{aligned}
\mathbf{x}^T AX &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\
&= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \sum_{j=1}^n a_{1j}x_j \\ \sum_{j=1}^n a_{2j}x_j \\ \vdots \\ \sum_{j=1}^n a_{nj}x_j \end{bmatrix} \\
&= x_1 \sum_{j=1}^n a_{1j}x_j + x_2 \sum_{j=1}^n a_{2j}x_j + \cdots + x_n \sum_{j=1}^n a_{nj}x_j \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j.
\end{aligned}$$

- **Lemma 5.2.** $\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} = (A + A^T)\mathbf{x}$

  Proof. Using Lemma 5.1, we have that, for a fixed k,

$$\mathbf{x}^T Z\mathbf{x} = \sum_{i=1}^N \sum_{j=1}^N a_{ij}x_i x_j = a_{kk}x_k^2 + \sum_{j\neq k} a_{kj}x_k x_j + \sum_{i\neq k} \sum_{j=1}^n a_{ij}x_i x_j$$

  Thus,

$$\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial x_k} = 2a_{kk}x_k + \sum_{j\neq k} a_{kj}x_j + \sum_{i\neq k} a_{ik}x_i = \sum_{j=1}^n a_{kj}x_j + \sum_{i=1}^n a_{ik}x_i = \sum_{i=1}^n (a_{ki} + a_{ik})x_i.$$

  Moreover, we have that

$$\begin{aligned}
\frac{\partial(\mathbf{x}^T A\mathbf{x})}{\partial \mathbf{x}} &= \begin{bmatrix} \partial(\mathbf{x}^T A\mathbf{x})/\partial x_1 \\ \partial(\mathbf{x}^T A\mathbf{x})/\partial x_2 \\ \vdots \\ \partial(\mathbf{x}^T A\mathbf{x})/\partial x_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n (a_{1i} + a_{i1})x_i \\ \sum_{i=1}^n (a_{2i} + a_{i2})x_i \\ \vdots \\ \sum_{i=1}^n (a_{ni} + a_{in})x_i \end{bmatrix} \\
&= \begin{bmatrix} a_{11} + a_{11} & a_{12} + a_{21} & \cdots & a_{1n} + a_{n1} \\ a_{21} + a_{12} & a_{22} + a_{22} & \cdots & a_{2n} + a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} + a_{1n} & a_{n2} + a_{2n} & \cdots & a_{nn} + a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\
&= (A + A^T)\mathbf{x}.
\end{aligned}$$

- **Lemma 5.3.** $\mathbf{x}^T A\mathbf{x} = \mathrm{tr}(A\mathbf{x}\mathbf{x}^T)$

Proof. We just have to show that $tr(A\mathbf{x}\mathbf{x}^T) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j$. Note that

$$A\mathbf{x}\mathbf{x}^T = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$$
$$= \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1x_1 & x_1x_2 & \cdots & x_1x_n \\ x_2x_1 & x_2x_2 & \cdots & x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_nx_1 & x_nx_2 & \cdots & x_nx_n \end{bmatrix}.$$

The diagonal members of the above matrix, from top to bottom, are:

$$\begin{gathered} a_{11}x_1x_1 + a_{12}x_2x_1 + \cdots + a_{1n}x_nx_1 \\ a_{21}x_1x_2 + a_{22}x_2x_2 + \cdots + a_{2n}x_nx_2 \\ \vdots \\ a_{n1}x_1x_n + a_{n2}x_2x_n + \cdots + a_{nn}x_nx_n. \end{gathered}$$

If we add all of these up, we get $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_ix_j$ as required. □

- Let $f(A)$ be a function sending matrix $A \in \mathbb{R}^{n\times n}$ to a real number.
  We define the partial derivative $\partial f(A)/\partial A$ to be a matrix whose (i,j)-entry is equal to $\partial f(A)/\partial a_{ij}$.

- **Lemma 5.4.** *If A is a general matrix (no relationship between the entries), then $\partial \det(A)/\partial a_{ij} = C_{ij}$ where C_{ij} is the (i,j)-cofactor of A.*

  Proof. Recall that $\det(A) = \sum_{i=1}^{n} a_{ij}C_{ij}$. Notice that no terms except $a_{ij}C_{ij}$ can contain a_{ij}. Moreover, C_{ij} does not contain a_{ij} itself. Hence, $\partial \det(A)/\partial a_{ij} = C_{ij}$. □

- **Lemma 5.5.** *If A is a symmetric matrix, then*

$$\frac{\partial \det(A)}{\partial a_{ij}} = \begin{cases} C_{ij}, & \text{if } i = j \\ 2C_{ij}, & \text{if } i \neq j \end{cases}$$

  where C_{ij} is the (i,j)-cofactor of A.

  Proof. If $i = j$, recall that $\det(A) = \sum_{i=1}^{n} a_{ij}C_{ij}$. Notice that no terms $a_{ij}C_{ij}$ in the sum can contain a_{ii} except for $a_{ii}C_{ii}$. Moreover, C_{ii} does not contain a_{ii} in itself. Therefore, $\partial \det(A)/\partial a_{ii} = C_{ii} = C_{ij}$.
  If $i \neq j$, we have that

$$\begin{aligned} \frac{\partial \det(A)}{\partial a_{ij}} &= \frac{\partial}{\partial a_{ij}}\Bigg(\sum_{\pi\in\mathfrak{S}_n} \operatorname{sgn}(\pi) \prod_{k=1}^{n} a_{k\pi(k)} \Bigg) \\ &= \frac{\partial}{\partial a_{ij}}\Bigg(\sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) a_{ij}^2 \prod_{k\neq i,j} a_{k\pi(k)} + \sum_{\substack{\pi(i)=j \\ \pi(j)\neq i}} \operatorname{sgn}(\pi) a_{ij} \prod_{k\neq i} a_{k\pi(k)} \\ &\quad + \sum_{\substack{\pi(i)\neq j \\ \pi(j)=i}} \operatorname{sgn}(\pi) a_{ij} \prod_{k\neq j} a_{k\pi(k)} + \sum_{\substack{\pi(i)\neq j \\ \pi(j)\neq i}} \operatorname{sgn}(\pi) \prod_{k=1}^{n} a_{k\pi(k)} \Bigg) \\ &= 2 \sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) a_{ij} \prod_{k\neq i,j} a_{k\pi(k)} + \sum_{\substack{\pi(i)=j \\ \pi(j)\neq i}} \operatorname{sgn}(\pi) \prod_{k\neq i} a_{k\pi(k)} + \sum_{\substack{\pi(i)\neq j \\ \pi(j)=i}} \operatorname{sgn}(\pi) \prod_{k\neq j} a_{k\pi(k)}. \end{aligned}$$

where $\mathfrak{S}_n$ is the set of permutations over $\{1, 2, \ldots, n\}$. Note that

$$2 \sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) a_{ij} \prod_{k \neq i,j} a_{k\pi(k)} = \sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) \prod_{k \neq i} a_{k\pi(k)} + \sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) \prod_{k \neq j} a_{k\pi(k)}.$$

Hence,

$$\begin{aligned}
&2 \sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) a_{ij} \prod_{k \neq i,j} a_{k\pi(k)} + \sum_{\substack{\pi(i)=j \\ \pi(j)\neq i}} \operatorname{sgn}(\pi) \prod_{k \neq i} a_{k\pi(k)} + \sum_{\substack{\pi(i)\neq j \\ \pi(j)=i}} \operatorname{sgn}(\pi) \prod_{k \neq j} a_{k\pi(k)} \\
&= \left(\sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) \prod_{k \neq i} a_{k\pi(k)} + \sum_{\substack{\pi(i)=j \\ \pi(j)\neq i}} \operatorname{sgn}(\pi) \prod_{k \neq i} a_{k\pi(k)} \right) \\
&\quad + \left(\sum_{\substack{\pi(i)=j \\ \pi(j)=i}} \operatorname{sgn}(\pi) \prod_{k \neq j} a_{k\pi(k)} + \sum_{\substack{\pi(i)\neq j \\ \pi(j)=i}} \operatorname{sgn}(\pi) \prod_{k \neq j} a_{k\pi(k)} \right) \\
&= \sum_{\pi(i)=j} \operatorname{sgn}(\pi) \prod_{k \neq j} a_{k\pi(k)} + \sum_{\pi(j)=i} \operatorname{sgn}(\pi) \prod_{k \neq i} a_{k\pi(k)} \\
&= C_{ij} + C_{ji} = 2C_{ij}.
\end{aligned}$$

The last line follows from the fact that A is symmetric, so the cofactors are symmetric as well. □

- **Corollary 5.6.** *If A is a symmetric matrix, then*

$$\frac{\partial \log(\det(A))}{\partial a_{ij}} = \begin{cases} C_{ij}/\det(A), & \text{if } i = j \\ 2C_{ij}/\det(A), & \text{if } i \neq j \end{cases}$$

- **Corollary 5.7.** *If A is a symmetric matrix, then*

$$\frac{\partial \log(\det(A))}{\partial A} = 2A^{-1} - \operatorname{diag}(A^{-1}).$$

- **Lemma 5.8.** *If A is symmetric, then*

$$\frac{\partial \operatorname{tr}(AB)}{\partial A} = B + B^T - \operatorname{diag}(B).$$

Proof. Let the rows of A be $a_1^T, a_2^T, \ldots a_n^T$. The the columns of B be $b_1, b_2, \ldots, b_n$. We have that $\operatorname{tr}(AB) = \sum_{i=1}^n a_i^T b_i = \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ji}$.

Consider the expression $\frac{\partial \operatorname{tr}(AB)}{\partial a_{ij}}$. There are two cases:

- If $i = j$, then the coefficient of a_{ij} in $\operatorname{tr}(AB)$ is $b_{ji} = b_{ii}$. Thus, the derivative with respect to a_{ij} is b_{ii}.
- If $i \neq j$, then the coefficient of $a_{ij} = a_{ji}$ is $b_{ji} + b_{ij}$. Thus, the derivative with respect to a_{ij} is $b_{ij} + b_{ji}$.

All in all, we have that the derivative with respect to A is equal to $B + B^T - \operatorname{diag}(B)$. □

References