The Introduction to Expectation Maximization Algorithm

R94922022 郭煜楓

R94944004 劉智杰

R94922013 楊恕先

R94922006 莊上墀

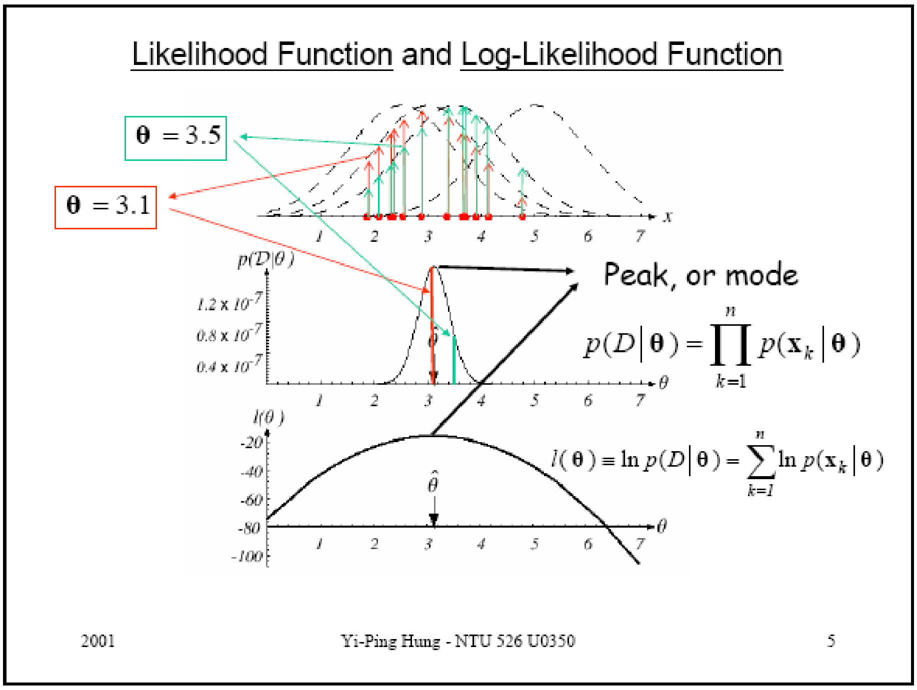
Outline

I. INTRODUCTION

A. What's maximum likelihood?

簡單地說，maximum likelihood 是用來推論如何從觀察到的樣本(samples)中，推測出整個群組最合理的分佈狀況。Ex.在資訊系中，隨機抽樣六個學生的身高，如何推出整個資訊系學生身高分佈的真正情況。

如下圖的紅點：



假設那些 samples 是 Gaussian distribution，那我們要如何找出合適的 mean 跟 variance？

首先我們先定義 likelihood function：

$$p(D\,|\,\theta)=\prod p(x_k\,|\,\theta)$$
$$x_k\ : sample\ data$$
$$\theta : the\ distribution\ \ parameter\ \ ex.\ mean\ \mu,\ variance\ \sigma$$

那麼 log likelihood function, i.e.,

$$\log p(D\mid\theta)=\sum\log p(x_k\mid\theta).$$

由上面可以明顯的體會出，當 $\log p(D\mid\theta)$ 為最大時，此時 distribution parameter θ 最合理。

B. The relation between EM and maximum likelihood

In a word , EM is a general method to finding maximum likelihood estimate of the parameter of a distribution from a given data when the data is incomplete or has missing values.

上面那句言簡意賅的解釋出 EM 最基本的觀念，EM 其實就是找 maximum likelihood，為一不一樣是 EM 可能會有未知的 sample 點。

解釋一下何謂 missing values。舉例來說：

◆ 給予四個人的身高跟體重，但其中一個人的體重未知(miss)，推測整體的 distribution?

◆ 任意給予十個人，但尚未做 classify，此時他屬於哪一個 class 是未知的 (miss, hidden)?

C. The EM algorithm

這裡主要是提出 EM 演算法一個基本的觀念，EM 就如 maximum likelihood 一樣是要找最適合的 distribution：

$$\theta^* = \arg\max_\theta \ln P(X \mid \theta)$$
$$= \arg\max_\theta \ln \sum_z P(X, z \mid \theta)$$
$$z : hidden\ data$$

The EM algorithm:

It defines a lower bound on log likelihood, then iteratively increases the low bound by alternating between

E-step: maximize it with respect to the distribution of hidden variables

M-step: maximize it with respect to the parameter θ

簡單地說，EM 就是先由目前的 distribution θ_n ，推測出 missing data z 最適合的 distribution，再由已知的 data + missing data z 去推測下一步整筆資料的 distribution θ_{n+1} 。

II. DERIVATION OF EM ALGORITHM

A. The derivation of EM

1. Total

從 EM 的 log likelihood function 開始推起

$$
\begin{aligned}
L(\theta) &= \ln P(X \mid \theta) \\
&= \ln \sum_z P(X, z \mid \theta) \\
&= \ln \sum_z P(X, z \mid \theta) \frac{Q(z)}{Q(z)} \\
&\geq \sum_z Q(z) \ln \frac{P(X, z \mid \theta)}{Q(z)}, \; by\, Jensen's\, inequality
\end{aligned}
$$

因為 ln Σp() 並不好計算，因此幸運地，可以給予一個 Q(z) 去得

E: Maximize it with respect to Q

M: Maximize it with respect to θ

接下來會對 E-step 跟 M- step 分別做說明。

2. E-step

E-step: Maximize it with respect to Q (z) when $\theta=\theta_n$

當θ=θn，先推測下一步 Q(z) 最適合的 distribution 為什麼樣子。

$$Q_n(z) = \arg\max_{Q(z)} \sum_z Q(z) \ln \frac{P(X, z \mid \theta_n)}{Q(z)}$$

$$use\ lagrange\ with\ constrant\ \sum_z Q(z) = 1$$

$$=> Q_n(z) = P(z \mid X, \theta_n)$$

我們可以用 Lagrange method 來找到極值，下面為使用 Lagrange 的推導過程

$$
\begin{aligned}
&G(Q(z)) = \lambda(1-\sum_z Q(z)) + \sum_z Q(z)\ln P(z,X\mid\theta_n) - \sum_z Q(z)\ln Q(z) \\
&\frac{\partial G}{\partial Q(z)} = -\lambda + \ln P(z,X\mid\theta_n) - \ln Q(z) - 1 \\
&=> \quad \ln Q(z) = \ln P(z,X\mid\theta_n) - (\lambda+1) \\
&=> \quad Q(z) = \frac{P(z,X\mid\theta_n)}{e^{\lambda+1}} \\
&and \quad \sum_z Q(z) = \sum_z \frac{P(z,X\mid\theta_n)}{e^{\lambda+1}} = 1 \\
&=> \sum_z P(z,X\mid\theta_n) = e^{\lambda+1} \\
&so, \ Q_n(z) = \frac{P(z,X\mid\theta_n)}{\sum_z P(z,X\mid\theta_n)} = \frac{P(z,X\mid\theta_n)}{P(X\mid\theta_n)} \\
&\qquad\qquad = P(z\mid X,\theta_n)
\end{aligned}
$$

3. M-step

M-step: Maximize it with respect to θ

$$
\begin{aligned}
\theta^{n+1} &= \arg\max_\theta l(\theta) \\
&= \arg\max_\theta \sum_z P(z\mid X,\theta^n)\ln\frac{P(X,z\mid\theta)}{P(z\mid X,\theta^n)} \\
&= \arg\max_\theta \sum_z P(z\mid X,\theta^n)\ln P(X,z\mid\theta) \\
&= \arg\max_\theta E_{z\mid X,\theta^n}(\ln P(X,z\mid\theta))
\end{aligned}
$$

當求到這裡，照著上面將已知的式子都帶進去之後，做一次微分，即可得到下一步的θn+1。

B. The no decreasing feature of EM

這邊是要證明一下，為什麼 EM 的 iteration 會越來越好。

$$
\begin{aligned}
\max_{\theta} l(\theta) &= \max_{\theta} \sum_{z} P(z \mid X, \theta^{n}) \ln \frac{P(X, z \mid \theta)}{P(z \mid X, \theta^{n})}, (M-step) \\
&\geq \sum_{z} P(z \mid X, \theta^{n}) \ln \frac{P(X, z \mid \theta^{n})}{P(z \mid X, \theta^{n})}, (E-step) \\
&= l(\theta^{n})
\end{aligned}
$$

附上最後兩步比較詳細的推導過程：

$$
\begin{aligned}
&\sum_{z} P(z \mid X, \theta^{n}) \ln \frac{P(X, z \mid \theta^{n})}{P(z \mid X, \theta^{n})} - l(\theta^{n}) \\
&= \sum_{z} P(z \mid X, \theta^{n}) \ln \frac{P(X, z \mid \theta^{n})}{P(z \mid X, \theta^{n})} - \ln P(X \mid \theta^{n}) \\
&= \sum_{z} P(z \mid X, \theta^{n}) \ln \frac{P(X, z \mid \theta^{n})}{P(z \mid X, \theta^{n})} - \sum_{z} P(z \mid X, \theta^{n}) \ln P(X \mid \theta^{n}) \\
&= \sum_{z} P(z \mid X, \theta^{n}) \ln \frac{P(X, z \mid \theta^{n})}{P(z \mid X, \theta^{n}) P(X \mid \theta^{n})} \\
&= \sum_{z} P(z \mid X, \theta^{n}) \ln \frac{P(X, z \mid \theta^{n})}{P(z, X \mid \theta^{n})} \\
&= \sum_{z} P(z \mid X, \theta^{n}) \ln 1 = 0
\end{aligned}
$$

III. SUMMARY OF EM

EM 演算法的整個流程如下：

Do n = n+1

E-step: compute

$$Q_n(z) = P(z \mid X, \theta), \quad E_{z|X,\theta^n}(\ln P(X, z \mid \theta))$$

M-step:

$$\theta^{n+1} = \arg\max_{\theta} E_{z|X,\theta^n}(\ln P(X, z \mid \theta))$$

Until

$$E_{z|X,\theta^n}(\ln P(X, z \mid \theta^{n+1})) - E_{z|X,\theta^{n-1}}(\ln P(X, z \mid \theta^n)) < Threshold$$

IV. AN EXAMPLE OF EM ALGORITHM

Expectation-Maximization for a 2D Normal Model

來源：

《Pattern Classification》

Richard O. Duda, Peter E. Hart, David G. Stork

Chapter 3, Example 2

假設一個集合中有四組資料，每組資料包含兩個變數：

$$D = \{\vec{x}_1, \vec{x}_2, \vec{x}_3, \vec{x}_4\} = \left\{ \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} * \\ 4 \end{pmatrix} \right\}$$

假設此二變數符合 Gaussian distribution，且兩變數間無交互關係，即 diagonal covariance 為零。可設定 θ 為：

$$\theta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^{\ 2} \\ \sigma_2^{\ 2} \end{pmatrix}$$

設定 θ^0 的值。假設兩變數的 Gaussian distribution 以原點為中心、$\Sigma = 1$，即：

$$\theta^0 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$$

現在要找出第一個修正的估測 θ^1，也就是說，須計算

$$Q(\theta;\theta^0)$$

接續前一節的 summery，E-step 可表示為下式：

$$Q(\theta;\theta^t)=\mathrm{E}_{Z|X,\theta^t}\left\{\ln p\left(X,z\mid\theta\right)\right\}$$

將資料套用至 E-step，可得:

$$\begin{aligned}Q(\theta;\theta^0)&=\mathrm{E}_{x_{41}|x_{42}=4,\theta^0}\left\{\ln p\left(\vec{x}_1,\vec{x}_2,\vec{x}_3,\begin{pmatrix}x_{41}\\4\end{pmatrix}\mid\theta\right)\right\}\\&=\int_{-\infty}^{\infty}\left[\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)+\ln p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}\mid\theta\right)\right]p\left(x_{41}\mid x_{42}=4,\theta^0\right)dx_{41}\end{aligned}$$

以 general Gaussian Distribution 取代，可得：

$$\begin{aligned}Q(\theta;\theta^0)&=\int_{-\infty}^{\infty}\left[\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)+\ln p\left(\begin{pmatrix}x_{41}\\4\end{pmatrix}\mid\theta\right)\right]p\left(x_{41}\mid x_{42}=4,\theta^0\right)dx_{41}\\
&=\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)+\int_{-\infty}^{\infty}\ln\left\{\frac{1}{2\pi\left|\begin{pmatrix}\sigma_1&0\\0&\sigma_2\end{pmatrix}\right|}e^{-\frac{1}{2}\left[\frac{(x_{41}-\mu_1)^2}{\sigma_1^2}+\frac{(4-\mu_2)^2}{\sigma_2^2}\right]}\right\}\frac{1}{2\pi\left|\begin{pmatrix}1&0\\0&1\end{pmatrix}\right|}e^{-\frac{1}{2}\left[\frac{(x_{41}-0)^2}{1^2}+\frac{(4-0)^2}{1^2}\right]}dx_{41}\\
&=\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)+\int_{-\infty}^{\infty}\left\{-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{1}{2}\left[\frac{(x_{41}-\mu_1)^2}{\sigma_1^2}+\frac{(4-\mu_2)^2}{\sigma_2^2}\right]\right\}\frac{1}{2\pi\left|\begin{pmatrix}1&0\\0&1\end{pmatrix}\right|}e^{-\frac{1}{2}\left[\frac{(x_{41}-0)^2}{1^2}+\frac{(4-0)^2}{1^2}\right]}dx_{41}\\
&=\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{(4-\mu_2)^2}{2\sigma_2^2}+\int_{-\infty}^{\infty}\left\{-\frac{1}{2}\left[\frac{(x_{41}-\mu_1)^2}{\sigma_1^2}\right]\right\}\frac{1}{2\pi\left|\begin{pmatrix}1&0\\0&1\end{pmatrix}\right|}e^{-\frac{1}{2}\left[\frac{(x_{41}-0)^2}{1^2}+\frac{(4-0)^2}{1^2}\right]}dx_{41}\\
&=\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{(4-\mu_2)^2}{2\sigma_2^2}+\int_{-\infty}^{\infty}\left\{-\frac{1}{2}\left[\frac{x_{41}^2-2x_{41}\mu_1+\mu_1^2}{\sigma_1^2}\right]\right\}\frac{1}{2\pi\left|\begin{pmatrix}1&0\\0&1\end{pmatrix}\right|}e^{-\frac{1}{2}\left[\frac{(x_{41}-0)^2}{1^2}+\frac{(4-0)^2}{1^2}\right]}dx_{41}\\
&=\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{(4-\mu_2)^2}{2\sigma_2^2}-\frac{\mu_1^2}{2\sigma_1^2}+\int_{-\infty}^{\infty}\left\{-\frac{1}{2}\left[\frac{x_{41}^2}{\sigma_1^2}\right]\right\}\frac{1}{2\pi\left|\begin{pmatrix}1&0\\0&1\end{pmatrix}\right|}e^{-\frac{1}{2}\left[\frac{(x_{41}-0)^2}{1^2}+\frac{(4-0)^2}{1^2}\right]}dx_{41}\\
&=\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{(4-\mu_2)^2}{2\sigma_2^2}-\frac{\mu_1^2+1}{2\sigma_1^2}\end{aligned}$$

完成第一次的 E-step 後，可得：

$$Q\left(\theta;\theta^{0}\right)=\sum_{i=1}^{3}\ln p\left(\vec{x}_i\mid\theta\right)-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{\left(4-\mu_2\right)^2}{2{\sigma_2}^2}-\frac{{\mu_1}^2+1}{2{\sigma_1}^2}$$

將之展開化簡：

$$\begin{aligned}
Q\left(\theta;\theta^{0}\right)=&\ln\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)\exp\left(-\frac{1}{2}\frac{\left(0-\mu_1\right)^2}{{\sigma_1}^2}-\frac{1}{2}\frac{\left(2-\mu_2\right)^2}{{\sigma_2}^2}\right)\\
&+\ln\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)\exp\left(-\frac{1}{2}\frac{\left(1-\mu_1\right)^2}{{\sigma_1}^2}-\frac{1}{2}\frac{\left(0-\mu_2\right)^2}{{\sigma_2}^2}\right)\\
&+\ln\left(\frac{1}{2\pi\sigma_1\sigma_2}\right)\exp\left(-\frac{1}{2}\frac{\left(2-\mu_1\right)^2}{{\sigma_1}^2}-\frac{1}{2}\frac{\left(2-\mu_2\right)^2}{{\sigma_2}^2}\right)\\
&-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{\left(4-\mu_2\right)^2}{2{\sigma_2}^2}-\frac{{\mu_1}^2+1}{2{\sigma_1}^2}\\
=&-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{1}{2}\frac{\left(0-\mu_1\right)^2}{{\sigma_1}^2}-\frac{1}{2}\frac{\left(2-\mu_2\right)^2}{{\sigma_2}^2}\\
&-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{1}{2}\frac{\left(1-\mu_1\right)^2}{{\sigma_1}^2}-\frac{1}{2}\frac{\left(0-\mu_2\right)^2}{{\sigma_2}^2}\\
&-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{1}{2}\frac{\left(2-\mu_1\right)^2}{{\sigma_1}^2}-\frac{1}{2}\frac{\left(2-\mu_2\right)^2}{{\sigma_2}^2}\\
&-\ln\left(2\pi\sigma_1\sigma_2\right)-\frac{\left(4-\mu_2\right)^2}{2{\sigma_2}^2}-\frac{{\mu_1}^2+1}{2{\sigma_1}^2}
\end{aligned}$$

完成 E-step 後，最大化上式，即進行 M-step。

使用微分取得極值。

設

$$\begin{aligned}
&\frac{\partial\,Q\left(\theta;\theta^{0}\right)}{\partial\,\mu_1}=0\\
&\Rightarrow\frac{2\left(0-\mu_1\right)}{-2{\sigma_1}^2}+\frac{2\left(1-\mu_1\right)}{-2{\sigma_1}^2}+\frac{2\left(2-\mu_1\right)}{-2{\sigma_1}^2}+\frac{2\mu_1}{-2{\sigma_1}^2}=0\\
&\Rightarrow\mu_1=0.75
\end{aligned}$$

設

$$\frac{\partial Q(\theta;\theta^0)}{\partial \mu_2}=0$$

$$\Rightarrow \frac{2(2-\mu_2)}{-2\sigma_2{}^2}+\frac{2(0-\mu_2)}{-2\sigma_2{}^2}+\frac{2(2-\mu_2)}{-2\sigma_2{}^2}+\frac{2(4-\mu_2)}{-2\sigma_2{}^2}=0$$

$$\Rightarrow \mu_1=2$$

設

$$\frac{\partial Q(\theta;\theta^0)}{\partial \sigma_1}=0$$

$$\Rightarrow \frac{1}{-\sigma_1}+\frac{(0-\mu_1)^2}{\sigma_1{}^3}+\frac{1}{-\sigma_1}+\frac{(1-\mu_1)^2}{\sigma_1{}^3}+\frac{1}{-\sigma_1}+\frac{(2-\mu_1)^2}{\sigma_1{}^3}+\frac{1}{-\sigma_1}+\frac{1+\mu_1{}^2}{\sigma_1{}^3}=0$$

$$replace\ \mu_1\ with\ 0.75$$

$$\Rightarrow \frac{4}{-\sigma_1}+\frac{60}{16\sigma_1{}^3}=0$$

$$\Rightarrow \sigma_1{}^2=\frac{60}{64}=0.9375$$

設

$$\frac{\partial Q(\theta;\theta^0)}{\partial \sigma_2}=0$$

$$\Rightarrow \frac{1}{-\sigma_2}+\frac{(2-\mu_2)^2}{\sigma_2{}^3}+\frac{1}{-\sigma_2}+\frac{(0-\mu_2)^2}{\sigma_2{}^3}+\frac{1}{-\sigma_2}+\frac{(2-\mu_2)^2}{\sigma_2{}^3}+\frac{1}{-\sigma_2}+\frac{(4-\mu_2)^2}{\sigma_2{}^3}=0$$

$$replace\ \mu_2\ with\ 2$$

$$\Rightarrow \frac{4}{-\sigma_2}+\frac{8}{\sigma_2{}^3}=0$$

$$\Rightarrow \sigma_2{}^2=2$$

即可完成第一次的 EM，得到：

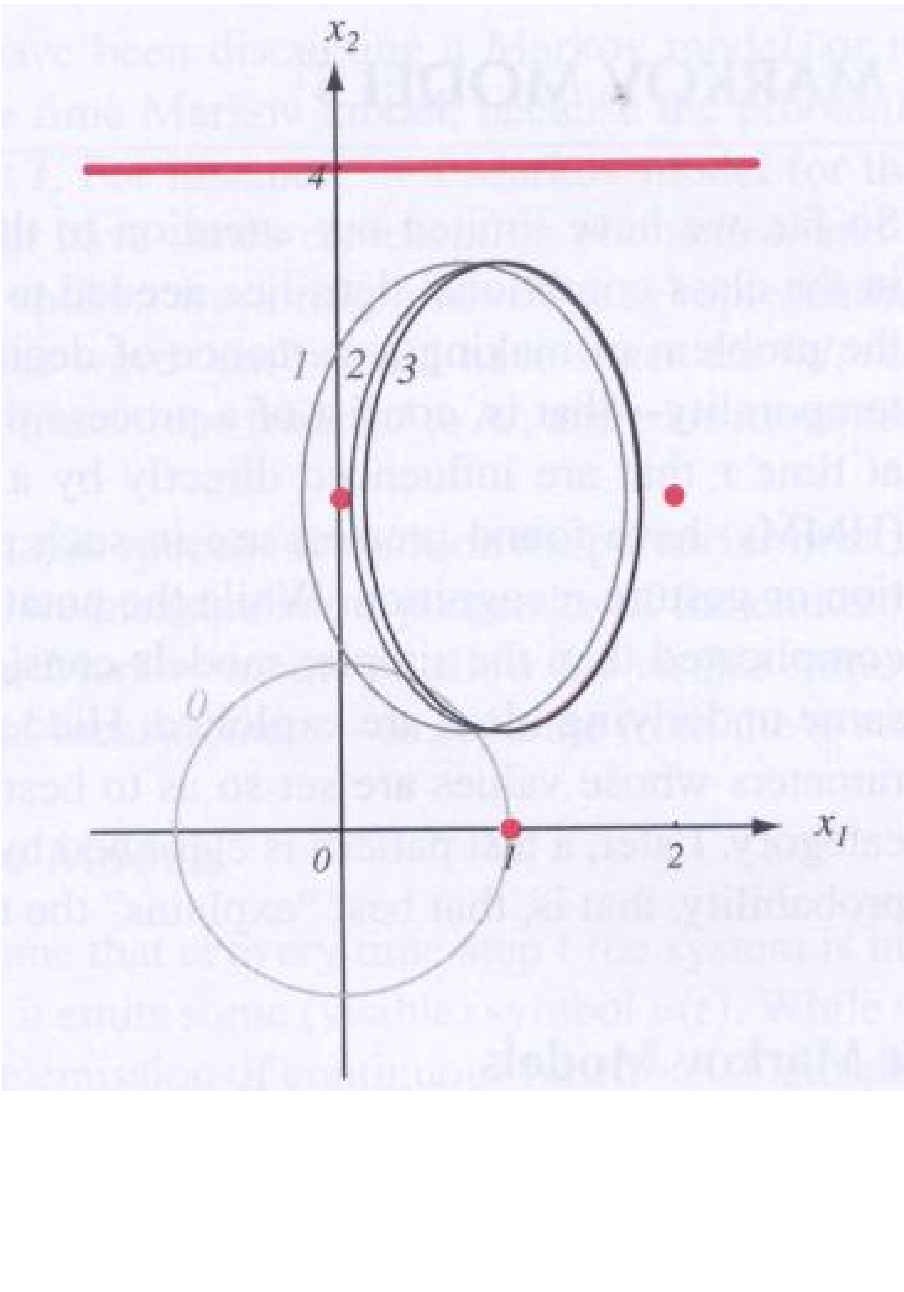
$$\theta^1=\begin{pmatrix}0.75\\2\\0.9375\\2\end{pmatrix}$$

接下來的動作都使用相同的概念與方法，但將會需要多餘的計算（這是因為 θ^0 設定成最易於計算的值）。但無論做幾次，因為第一個變數與第二個變數互相獨立，因此 μ_2 永遠為 2。

第三次 iteration 後，EM algorithm 將逐漸收斂至

$$\theta^3 = \begin{pmatrix} 1 \\ 2 \\ 0.6667 \\ 2 \end{pmatrix}$$

以下為三次 iterations 後，θ 的變化。



V. A PRACTICAL EXAMPLE OF EM ALGORITHM

論文： Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying

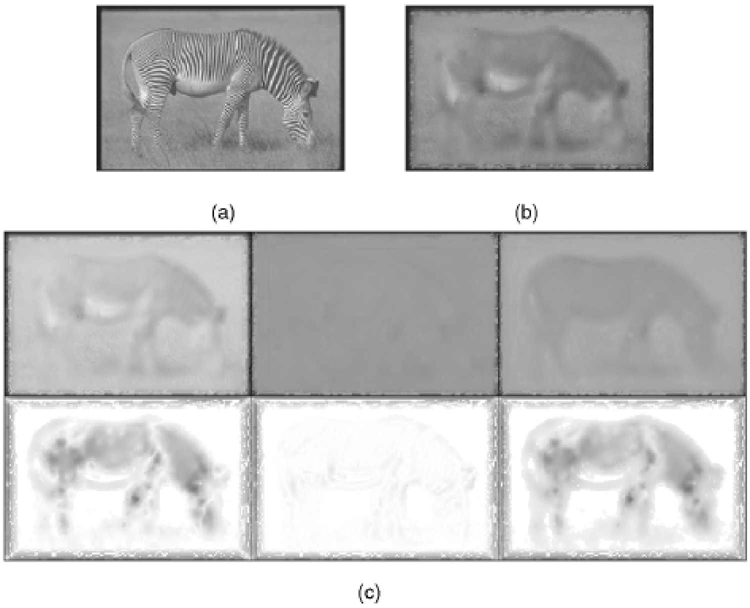
作者： Chad Carson, Member, IEEE, Serge Belongie, Member, IEEE, Hayit Greenspan, Member, IEEE, and Jitendra Malik, Member, IEEE

出自： IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 24, NO. 8, AUGUST 2002

這篇主要的目的是要用 EM 的方式做 Image Segmentation，之後可應用在 Image Querying 上面。而此處主要講如何用 EM 來做 segmentation。

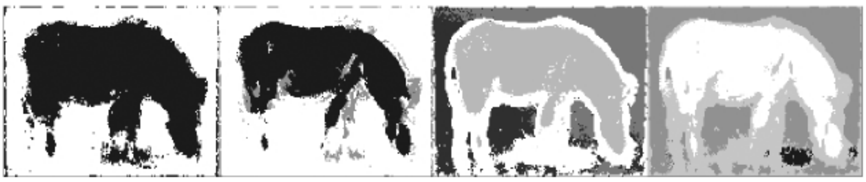
首先，他對圖的每個 pixel，取出八個特徵(feature)，分別是 color(L, a, b 三個)、texture(anisotropy, polarity, and contrast 三個)、加上 position(x, y 兩個)，總共八個 feature。

假設 pixel feature 的分布是 mixture Gaussians，利用 EM 來找出那些 Gaussian distribution 的 parameter。實際例子如下：



(a)原圖 (b)影像經過適當的 smooth (c)取出六個 feature，上面三個是分別是 L、a、b，下面三個分別是 anisotropy、polarity 和 contrast，範圍從 0(白)到 1(黑)。

每個 pixel 除了六個 feature 外，還有它的座標，所以總共八個 feature。



上圖為假設有 2、3、4、5 個 Gaussian 分別做出來的情況。

以下來說數學式子：

假設有 k 個 Gaussian

$$f(x|\Theta) = \sum_{i=1}^{k} \alpha_i f_i(x|\Theta_i)$$

x is a feature vector

α's represent the mixing weights

Θ represents the collection of $(\alpha_1,\dots,\alpha_k, \Theta_1,\dots, \Theta_k)$

f_i is a multivariate Gaussian density parameterized by Θ_i (μ_i and Σ_i)

d(維度)=8

$$f_i(x|\Theta_i) = \frac{1}{(2\pi)^{d/2} \det \Sigma_i^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$$

The update equations :

$$\alpha_i^{new} = \frac{1}{N}\sum_{j=1}^{N} p(i|x_j, \Theta^{old})$$

$$\mu_i^{new} = \frac{\sum_{j=1}^{N} x_j p(i|x_j, \Theta^{old})}{\sum_{j=1}^{N} p(i|x_j, \Theta^{old})}$$

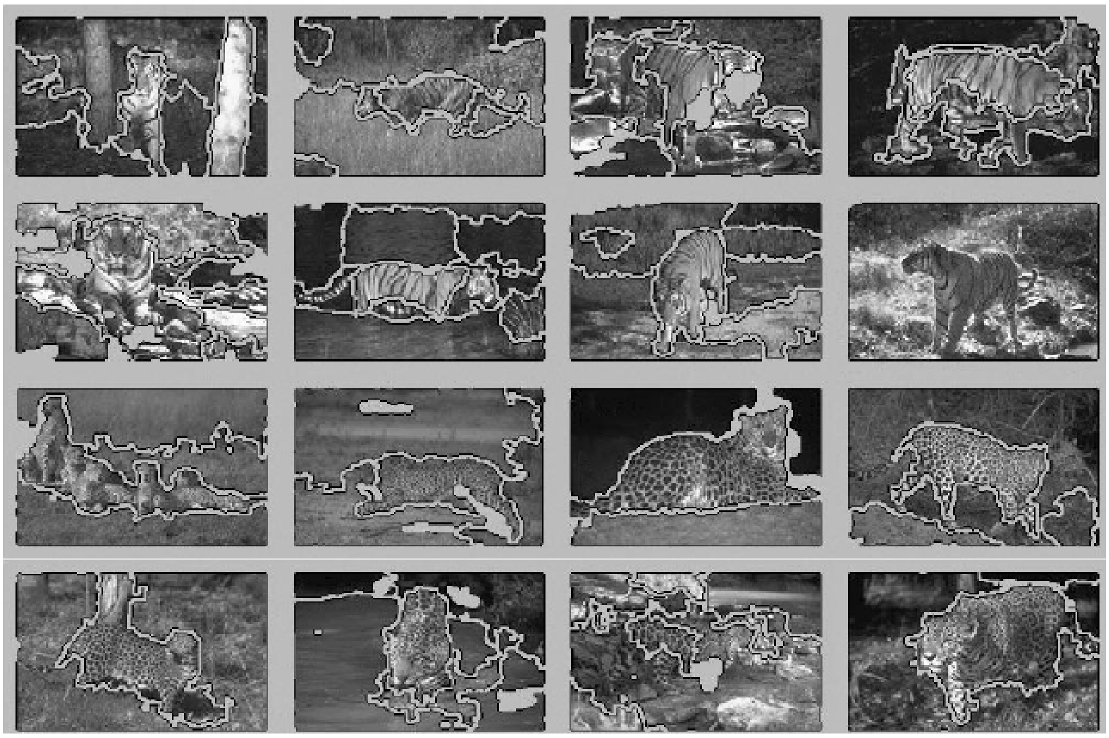
$$\Sigma_i^{new} = \frac{\sum_{j=1}^{N} p(i|x_j, \Theta^{old})(x_j - \mu_i^{new})(x_j - \mu_i^{new})^T}{\sum_{j=1}^{N} p(i|x_j, \Theta^{old})}$$

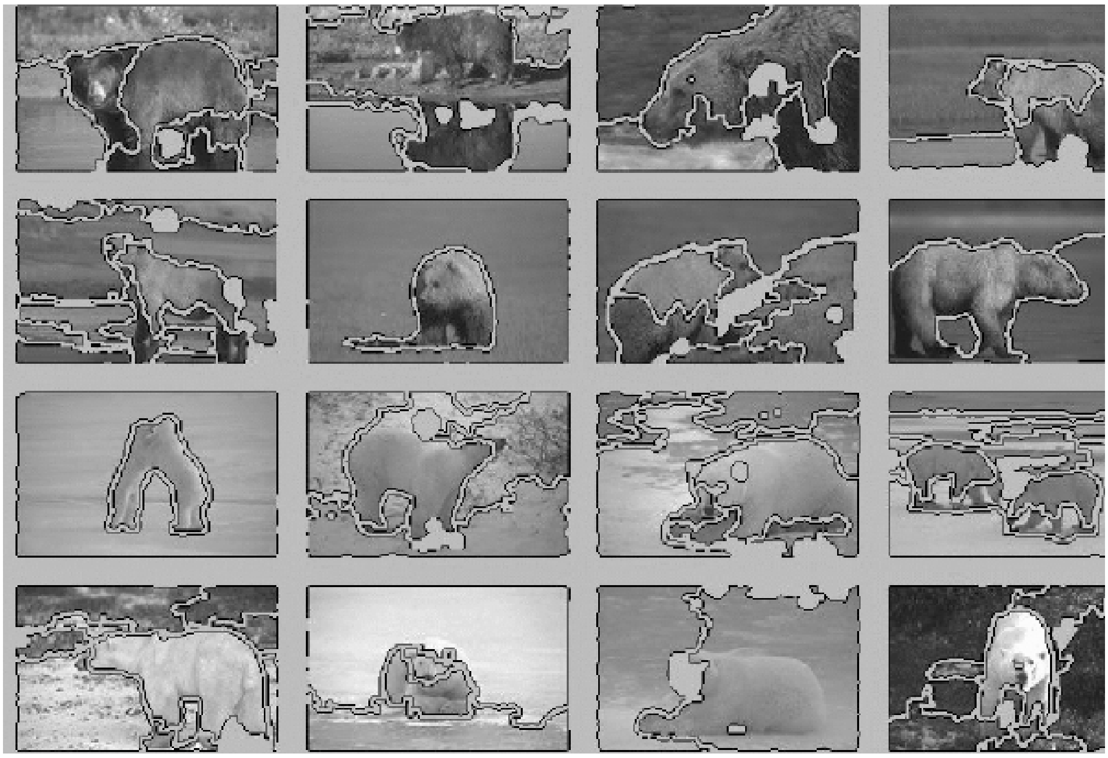
where $p(i|x_j, \Theta)$ is the probability that Gaussian i fits the pixel x_j , given the data Θ

$$p(i|x_j, \Theta) = \frac{\alpha_i f_i(x_j|\Theta_i)}{\sum_{k=1}^{N} \alpha_k f_k(x_j|\Theta_k)}$$

重複做到 $\log L(\Theta|X) = \log \prod_{k=1}^{N} f(x_k|\Theta)$ 增加少於 1%。

以下為實際做出來的樣子(在此設 k=4)





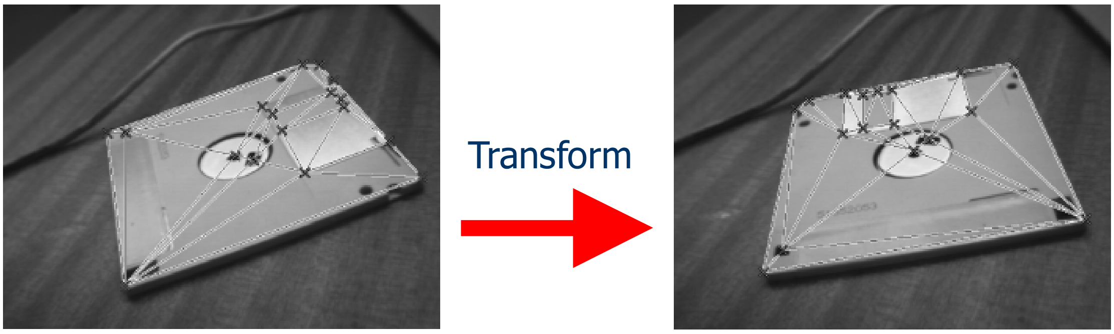
VI. ANOTHER EXAMPLE OF EM ALGORITHM

論文： Graph Matching With a Dual-Step EM Algorithm

作者： Andrew D.J. Cross and Edwin R. Hancock, University of York

出自： PAMI, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL 20, NO. 11, NOVEMBER 1998

這篇 paper 敘述如何利用 EM 演算法解決 2D graph matching 問題，如下圖例：



左右兩張圖分別以兩個角度拍攝同一張磁片，所以可將右圖視為是左圖乘上某一個 transform matrix 的結果，Graph matching 問題即是試圖找出這兩張圖相對應的特徵點，並藉由這些相對應的特徵點推出 transform matrix。這個問題的難度在於，我們試圖在兩張圖的特徵點集合間找出相對應關係(feature points correspondence matches)的方法，其實是利用了 transform matrix，而這個未知的 transform matrix 卻正是我們之所以要找出對應特徵點的原因，形成一個雞生蛋蛋生雞的問題。

前人的作法大致為先預估一組特徵點對應關係，接著以一些技巧消除明顯預估錯誤的組合，再拿剩下的對應關係去推導 transform matrix 中的各項係數。在這類作法中，雖然明顯的對應錯誤已被消除，不過剩下來的對應關係仍然無法保證

其正確性，所以推得的 transform matrix 係數自然有一定誤差。

由於上述作法無法保證結果的正確性，這篇 paper 放棄這種先預估再消去明顯錯誤的作法，他們以更新取代消去，一邊預估特徵點的對應關係，一邊以目前預估的關係計算 transform matrix，再拿剛求得的 transform matrix 衡量之前的預估是否正確，並更正之前的錯誤得到新的一組特徵點對應，如此不斷循環下去，最後 transform matrix 和特徵點對應將會收斂。顯而易見的，這個過程即是由 EM 完成，題目中所謂的 dual-step 即是指更新 transform matrix 和特徵點間的對應關係兩者。

和一般 EM 演算法不同的是，一般 EM 演算法是藉由已觀測到的 sample 推量 likelihood 最大的那組參數，但由於兩張圖的特徵點要如何對應是我們自己預估的，並沒有受到限制，也就是說這個問題中並沒有所謂已經觀測到的 sample，所以並不能直覺的套用 EM 演算法來計算 transform matrix 這組參數。當然，我們還是需要一些限制才能去衡量目前的參數的 likelihood，作者利用的是特徵點間的結構關係，例如左圖的某特徵點 A 以三條 edge 和另外三個特徵點相連，則一個好的對應關係應該會讓這個特徵點 A 對應到同樣和三個特徵點相鄰的特徵點 A'。

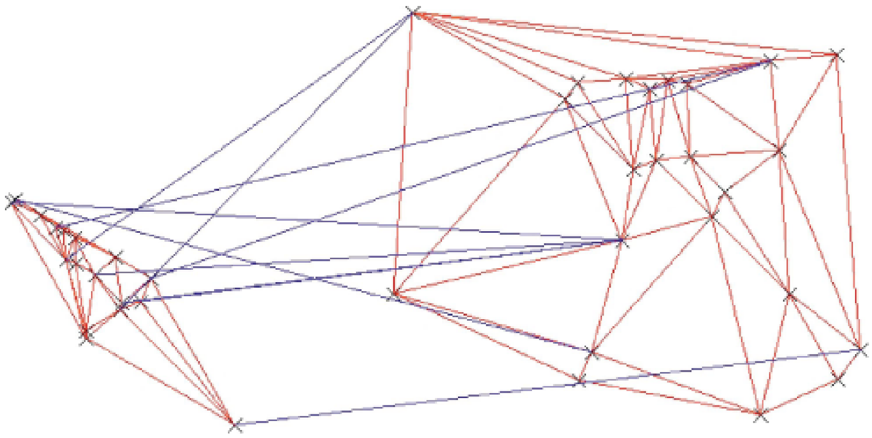
接下來將詳細介紹這篇 paper 的作法。

首先，我們假設這裡出現的 transform matrix 均為 perspective matrix：

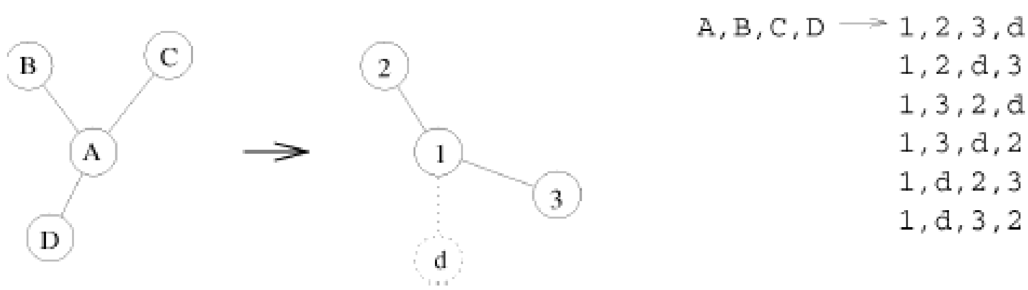
$$\Phi^{(n)} = \begin{pmatrix} \phi_{1,1}^{(n)} & \phi_{1,2}^{(n)} & \phi_{1,3}^{(n)} \\ \phi_{2,1}^{(n)} & \phi_{2,2}^{(n)} & \phi_{2,3}^{(n)} \\ \phi_{3,1}^{(n)} & \phi_{3,2}^{(n)} & \phi_{3,3}^{(n)} \end{pmatrix}$$

所以我們的目標是求出上列 matrix 中的九個參數。

我們將兩張圖的特徵點以 bipartite graph 表示，並分別對這兩個特徵點集合進行三角化，結果由紅線表示；而藍線代表目前預測的對應關係。由於拍攝時可能受到雜訊干擾，兩張圖找出來的特徵點個數不一定完全相同，所以在實際進行配對時會為少的那方補上 dummy node d 以方便對應，當然，出現 dummy node 的對應組合其 likelihood 自然較一般組合為低。



由於兩張圖可能是從完全不同的角度拍攝的，以特徵點間實際的連結狀況進行表示(如某 edge 和某 edge 的夾角度數) 是不可行的，所以他們定義了一個抽象化的表示法稱為 dictionary，Dictionary 並不受圖片 translation、scaling、rotation 的影響，只紀錄哪些點是以某個中心點相連的，如下圖：



在找出特徵點、將特徵點三角化、補上 dummy node 並排列出 dictionary 中的各種對應可能後，我們要為 dictionary 中的每種對應可能算出他們各自的機率：

$$P(\Gamma_{i,j}) = \sum_{S\in\Theta_j} P(\Gamma_{i,j}|S).P(S)$$

$$P(\Gamma_{i,j}|S) = \prod_{(k,l)\in S} P(f(k)\,|\,l). \qquad P(S) = \frac{1}{|\Theta_j|}$$

$$P(f(k)\,|\,l) = \begin{cases} (1-P_\phi)(1-P_e) & \text{if } f(k)=l \\ (1-P_\phi)P_e & \text{if } f(k)\neq l \text{ and } l\neq \text{dummy} \\ P_\phi & \text{if } k=\text{dummy or } l=\text{dummy} \end{cases}$$

$$P_e = P_\phi = \frac{2\big||\mathcal{M}|-|\mathcal{D}|\big|}{\big||\mathcal{M}|+|\mathcal{D}|\big|}$$

$$P(\Gamma_{i,j}|S) = \left[(1-P_\phi)(1-P_e)\right]^{R_{i,j}-H(\Gamma_{i,j},S)-\Psi(\Gamma_{i,j})} \times \left[(1-P_\phi)P_e\right]^{H(\Gamma_{i,j},S)} \times \left[P_\phi\right]^{\Psi(\Gamma_{i,j})}.$$

$$H(\Gamma_{i,j},S) = \sum_{(k,l)\in S}\left(1-s_{k,l}^{(n)}\right)$$

$$\Psi(\Gamma_{i,j}) = \left||C_i^D|-|C_j^M|\right|$$

$$R_{i,j} = \max\left[|C_i^D|,|C_j^M|\right]$$

其中 i, j 分別為兩張圖經三角化後的特徵點，$P(\Gamma_{i,j})$ 代表 i 對應到 j 的機率，S 為與 i, j 相鄰的點的集合。經貝氏定理展開後，P(S)即為 dictionary 中 j 出現的機率，而前面那項又可分為三種 case 討論，分別是 k 正確對應到 l、k 沒有對應到 l、或是 k 雖對應到 l 但 k, l 其中一個是 dummy node。P_ϕ和P_e分別代表結構上出錯的機率(如三個 edge 的點對應到只有兩個 edge 的點)和初始時的錯誤。

得到評量特徵點對應關係是否恰當的 $P(\Gamma_{i,j})$ 後，接下來將介紹如何在 EM 架構下使用之前推導出的數學工具，首先，我們將要 maximize 的 likelihood 定義為 $p(\mathbf{w}|f,\Phi) = \prod_{l\in D} p(\bar{w}_l|f,\Phi)$，f 為特徵點的對應關係，Φ為 transform matrix 的

係數，w 為 input data graph。上式可拆開為各 subgraph likelihood 的乘積，進一步推導可得下式：

$$p(\mathrm{w}|f,\Phi)=\prod_{i\in\mathcal{D}}p(\bar{w}_i|f,\Phi)$$

$$p(\bar{w}_i|f,\Phi)=\sum_{j\in\mathcal{M}}p(\bar{w}_i,\bar{z}_j|f,\Phi)$$

$$p(\bar{w}_i,\bar{z}_j|f,\Phi)=p(\bar{w}_i,\bar{z}_j|\Phi)^{s_{i,j}}\rho^{1-s_{i,j}}$$

$$\boxed{p(\mathrm{w}|f,\Phi)=\prod_{i\in\mathcal{D}}\sum_{j\in\mathcal{M}}p(\bar{w}_i,\bar{z}_j|\Phi)^{s_{i,j}}\rho^{1-s_{i,j}}}$$

其中ρ為 outlier 的 desity，D 為 data graph，M 為 model graph。紅線處的推導過程是假設了 outlier 的的機率和 coordinates 無關，為一 uniform density。由上式可繼續推得 EM 演算法中必須的 conditional log-likelihood $Q\left(\Phi^{(n+1)}|\Phi^{(n)}\right)$：

$$\boxed{Q\left(\Phi^{(n+1)}|\Phi^{(n)}\right)=\sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{M}}P\left(\bar{z}_j|\bar{w}_i,\Phi^{(n)}\right)\left[\zeta_{i,j}^{(n)}\left(\ln p\left(\bar{w}_i,\bar{z}_j|\Phi^{(n+1)}\right)-\ln\rho\right)+\ln\rho\right]}$$

$$\boxed{p\left(\bar{w}_i,\bar{z}_j|\Phi^{(n)}\right)=\frac{1}{(2\pi)^{\frac{3}{2}}\sqrt{|\Sigma|}}\exp\left[-\frac{1}{2}\epsilon_{i,j}\left(\Phi^{(n)}\right)^{T}\Sigma^{-1}\epsilon_{i,j}\left(\Phi^{(n)}\right)\right]}$$

最後，定義更新 hidden/missing data 的事後機率 E-step 為：

$$P\left(\vec{z}_j \middle| \vec{w}_i, \Phi^{(n+1)}\right) = \frac{\alpha_{i,j}^{(n)} p\left(\vec{w}_i, \vec{z}_j \middle| \Phi^{(n)}\right)}{\sum_{j' \in \mathcal{M}} \alpha_{j'}^{(n)} p\left(\vec{w}_i, \vec{z}_{j'} \middle| \Phi^{(n)}\right)} \qquad \alpha_{i,j}^{(n+1)} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} P\left(\vec{z}_j \middle| \vec{w}_i, \Phi^{(n)}\right)$$

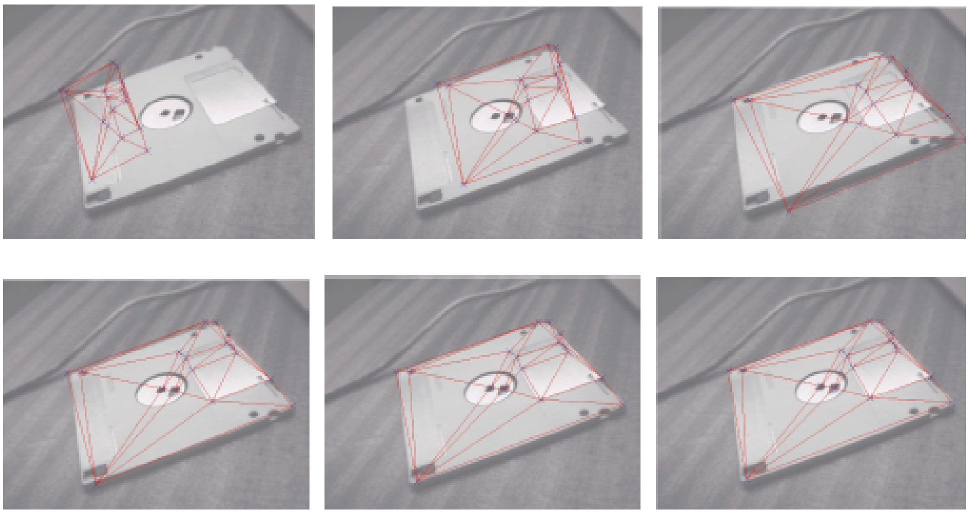
而 M-step 因為需要同時更新 transform matrix 和 correspondence matches 所以有兩個步驟：

$$f^{(n+1)}(i) = \arg\max_{j \in \mathcal{M}} P\left(\vec{z}_j \middle| \vec{w}_i, \Phi^{(n)}\right) \zeta_{i,j}^{(n+1)}$$

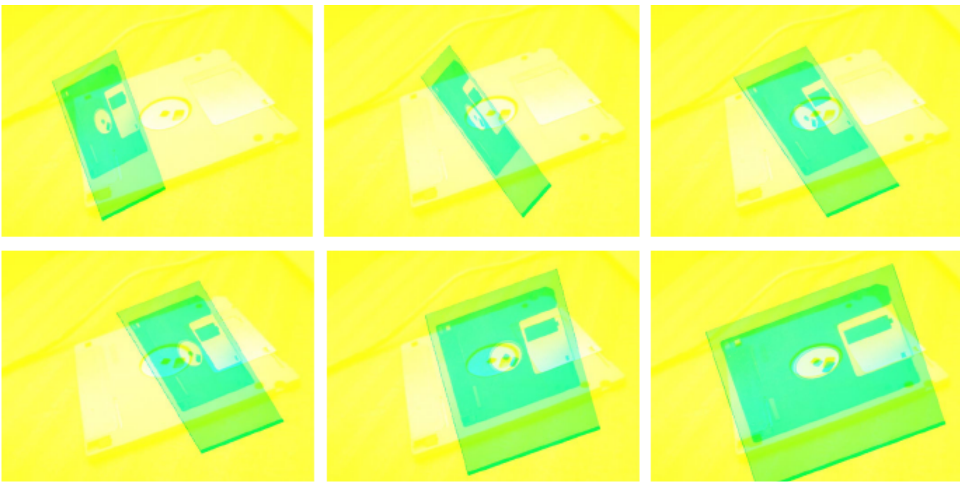
$$\Phi^{(n+1)} = \left[\sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{M}} P\left(\vec{z}_j \middle| \vec{w}_i, \Phi^{(n)}\right) \zeta_{i,j}^{(n)} \vec{z}_j U^T \vec{z}_j^T \Sigma^{-1} \right]^{-1} \times \left[\sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{M}} P\left(\vec{z}_j \middle| \vec{w}_i, \Phi^{(n)}\right) \zeta_{i,j}^{(n)} \vec{w}_i U^T \vec{z}_j^T \Sigma^{-1} \right].$$

其中 $\zeta_{i,j}^{(n+1)}$ 為前述的 structural probability、$f^{n+1}(i)$為第 n+1 個 iteration 時 feature point i 所對應到的點 j。而$\Phi^{(n+1)}$ 的推導過程中加入了 perspective matrix 特性，若選用 affine matrix 則會得到不同結果。

結果：



經過 6 個 iteration 後的結果，紅線代表 model graph 經不斷更正後，和 data graph 的磁片越來越接近，如下圖所示：



(a) (b) (c)

Fig. 13. Aerial image registration. (a) The digital map. (b) The registration with the high altitude image. (c) The registration with the low altitude image.

另一實驗，電子地圖和兩張空照圖的比對。