Chapter 7: EM algorithm in exponential families: JAW 4.30-32

7.1 (i) The EM Algorithm finds MLE's in problems with latent variables (sometimes called "missing data"): things you wish you could observe, but cannot.

(ii) Recall homework example (with added parameter σ^2): Y_i i.i.d. from mixture $\frac{1}{2}N(-\theta,\sigma^2)+\frac{1}{2}N(\theta,\sigma^2)$

$$\begin{aligned}
f_Y(y;\theta,\sigma^2) &= \frac{1}{2}(2\pi\sigma^2)^{-\frac{1}{2}}\exp(-y^2/(2\sigma^2))\exp(-\theta^2/(2\sigma^2)) \\
&\quad (\exp(-y\theta/\sigma^2)+\exp(y\theta/\sigma^2)) \\
\ell_n(\theta,\sigma^2) &= \text{const} - (n/2)\log(\sigma^2) - (\sum_i Y_i^2)/(2\sigma^2) - n\theta^2/(2\sigma^2) \\
&\quad + \sum_i \log((\exp(-Y_i\theta/\sigma^2)+\exp(Y_i\theta/\sigma^2))
\end{aligned}$$

(iv) What is the sufficient statistic?
How would we estimate (θ,σ^2) ?

(v) Now suppose each Y_i caries a "flag" $Z_i = -1$ or 1 as obsn i comes from $N(-\theta,\sigma^2)$ or $N(\theta,\sigma^2)$. Let $X_i=(Y_i,Z_i)$.

$$\begin{aligned}
f_X(y,z;\theta,\sigma^2) &= \frac{1}{2}(2\pi\sigma^2)^{-\frac{1}{2}}\exp(-(y-\theta z)^2/2\sigma^2) \\
\ell_{c,n}(\theta,\sigma^2) &= \sum_i \log f(Y_i,Z_i;\theta,\sigma^2) \\
&= \text{const} - (n/2)\log(\sigma^2) - \sum_i (Y_i-\theta Z_i)^2/2\sigma^2 \\
&= \text{const} \ -(n/2)\log(\sigma^2) - (2\sigma^2)^{-1}(\sum Y_i^2 - 2\theta\sum Y_i Z_i + \theta^2 n)
\end{aligned}$$

(vi) What now is sufficient statistic, if X_i were observed?
How now could you estimate (θ,σ^2) ?
The Z_i are "latent variables"; X_i are "complete data"

(vii) $\mathrm{E}(\ell_{c,n}|Y)$ requires only

$$\mathrm{E}(Z_i|Y_i) = \frac{\phi((y_i-\theta)/\sigma)-\phi((y_i+\theta)/\sigma)}{\phi((y_i-\theta)/\sigma)+\phi((y_i+\theta)/\sigma)}$$

7.2 Defining the EM algorithm

(i) Data random variables Y, with probability measure Q_θ, with $\theta \in \Theta \subset \Re^k$. Suppose Y has density f_θ w.r.t. some σ-finite measure μ (usually Lebegue measure (pdf) or counting measure (pmf)).

(ii) $\ell(\theta) = \log L(\theta) = \log f_\theta(y)$, for observed data $Y = y$. Suppose maximization of $\ell(\theta)$ is messy/impossible; $k > 1$ but not huge.

(iii) Suppose we augment the random variables to "complete data" X: $Y = Y(X)$. Suppose X has density $g_\theta(x)$. Then

$$\begin{aligned} L(\theta) &= f_\theta(y) = \int_{x:y(x)=y} g_\theta(x) d\mu(x) \\ \textbf{and } \ell_c(\theta) &= \log g_\theta(x) \end{aligned}$$

is known as the "complete-data likelihood".

(iv) Let $Q(\theta;\theta^*) = \mathrm{E}_{\theta^*}(\ell_c(\theta) \mid Y(X) = y)$. Then

$$\begin{aligned} g_\theta(X) &= h_\theta(X|Y=y) f_\theta(y) \\ \ell_c(\theta; X) &= \log h_\theta(X|Y=y) + \ell(\theta; y) \\ Q(\theta;\theta^*) &= H(\theta;\theta^*) + \ell(\theta) \quad \forall\ y \\ \textbf{where } H(\theta;\theta^*) &= \mathrm{E}_{\theta^*}(\log h_\theta(X|Y=y) \mid Y(X) = y) \end{aligned}$$

(v) EM algorithm is:
E-step: at current estimate θ^* compute $Q(\theta;\theta^*)$.
M-step: Maximize $Q(\theta;\theta^*)$ w.r.t. θ to obtain new estimate $\tilde{\theta}$.
Set $\theta^* = \tilde{\theta}$ and repeat ad nauseam.

7.3 Why does EM work?

(i) Recall (see Kullback-Leibler info) that for any densities p **and** q **of r.v.** Z**,** $\mathrm{E}_q(\log p(Z)) = \int \log(p(z))q(z)d\mu(z)$ **is maximized w.r.t** p **by** $p = q$**.** **(**$K(q;p) = \mathrm{E}_q \log(q/p) \geq 0$**.)**

(ii) Hence $H(\theta;\theta^*) \leq H(\theta^*,\theta^*)$ **for all** $\theta,\ \theta^*$**.**

(iii) Now with new/old estimates $\tilde{\theta}$**,** θ^*

$$
\begin{aligned}
\ell(\tilde{\theta}) - \ell(\theta^*) &= Q(\tilde{\theta};\theta^*) - Q(\theta*;\theta^*) \\
&\qquad - (H(\tilde{\theta};\theta^*) - H(\theta^*;\theta^*)) \text{ by } \mathbf{7.2(iv)} \\
&\geq H(\theta^*;\theta^*) - H(\tilde{\theta};\theta^*) \text{ by M-step} \\
&\geq 0 \text{ by (ii). Also} \\
\ell(\tilde{\theta}) - \ell(\theta^*) &> 0, \text{ unless } h_{\tilde{\theta}}(X|Y) = h_{\theta^*}(X|Y)
\end{aligned}
$$

(iv) Thus each step of EM cannot decrease $\ell(\theta)$ **and usually increases** $\ell(\theta)$**.**

(v) If the MLE $\hat{\theta}$ **is the unique stationary point of** $\ell(\theta)$ **in the interior of the space, then** $\tilde{\theta} \to \hat{\theta}$

(vi) In practice, EM is very robust, but can be very slow, especially in final stages: cgce is first-order.

(vii) Caution: we do NOT "use expectations to impute the missing data"
We compute the expected complete-data log-likelihood. This normally involves using conditional expectations to impute the complete-data sufficient statistics. This is NOT the same thing – see hwk. And it could be more complicated than this – although not if we have chosen sensible "complete-data".

7.4 A multinomial example

(i) Bernstein (1928) used population data to validate the hypothesis that human ABO blood tyoes are determined by 3 alleles, A, B and O at a single genetic locus, rather than being 2 independent factors A/not-A, B/not-B.

(ii) Suppose that the population frequencies of the A, B and O are p, q and r $(p+q+r=1)$;
we want to estimate (p, q, r).

(iii) We assume that the types of the two alleles carried by an individual are independent (Hardy-Weinberg Equil: 1908), and that individuals are independent ("unrelated").

(iv) ABO blood types are determined as follows:

blood type	genotype	freq.	type	geno.	freq.
A	AA	p^2	**A**	AO	$2pr$
B	BB	q^2	**B**	BO	$2qr$
AB	AB	$2pq$	**O**	OO	r^2

(v) $Y \sim M_4(n, (p^2+2pr, q^2+2qr, 2pq, r^2))$
$X \sim M_6(n, (p^2, 2pr, q^2, 2qr, 2pq, r^2))$

(vi) $\ell(p, q, r)$ **easy to evaluate but hard to max.**

$$\begin{aligned} \ell(p,q,r) &= \textbf{const} + y_A \log(p^2+2pr) + y_B \log(q^2+2qr) \\ &\quad + y_{AB}\log(2pq) + y_O \log(r^2) \end{aligned}$$

(v) $\ell_c(p, q, r)$ **is easy to maximize:**

$$\begin{aligned} \ell_c(p,q,r) &= \textbf{const} + x_{AA}\log(p^2) + x_{AO}\log(2pr) + x_{BB}\log(q^2) \\ &\quad + x_{BO}\log(2qr) + x_{AB}\log(2pq) + x_{OO}\log(r^2) \\ &= \textbf{const} + (2x_{AA} + x_{AO} + x_{AB})\log p \;\; + \\ &\quad (2x_{BB} + x_{BO} + x_{AB})\log q \;\; + (2x_{OO} + x_{AO} + x_{BO})\log r \end{aligned}$$

(vi) E-step: $x^*_{AA} = \mathrm{E}_{p,q,r}(X_{AA}|Y = y) = \frac{p^2}{p^2+2pr} y_A = \frac{p}{p+2r} y_A$ **etc.**

(vii) M-step: $\tilde{p} \; = \; (2n)^{-1}(2x^*_{AA} + x^*_{AO} + y_{AB})$**,**
$\tilde{q} \; = \; (2n)^{-1}(2x^*_{BB} + x^*_{BO} + y_{AB})$**,** $\tilde{r} = 1 - \tilde{p} - \tilde{q}$**.**

(viii) This method know to geneticists in 1950s: "gene-counting". (EM algorithm dates to 1977: Dempster,Laird & Rubin)

7.5 A mixture example (see prev hwk)

(i) Y_i **i.i.d, with pdf** $f(y; \theta, \psi) = \theta f_1(y; \psi) + (1 - \theta) f_2(y; \psi)$

(ii) $\ell(\theta, \psi) \; = \; \Sigma_i \log(\theta f_1(y_i; \psi) + (1 - \theta) f_2(y_i; \psi))$

(iii) Let $Z_i = I(Y_i \sim f_1)$**.** $P(Z_i = 1) = \theta$**,**
$\ell_c(\theta, \psi) = \Sigma_i(Z_i \log \theta + (1 - Z_i) \log(1 - \theta) + Z_i \log f_1(y_i; \psi) + (1 - Z_i) \log f_2(y_i; \psi))$

(iv) ℓ_c **is linear in** Z_i**, so E-step requires only**

$$\delta_i \; = \; \mathrm{E}(Z_i|Y) = \frac{\theta f_1(y_i; \psi)}{\theta f_1(y_i; \psi) + (1 - \theta) f_2(y_i; \psi)}$$

(v) M-step: $\tilde{\theta} \; = \; \Sigma_i \, \delta_i / n$ **and** $\tilde{\psi}$ **maximizes**
$\Sigma_i \, \delta_i \log f_1(y_i; \psi) + (1 - \delta_i) \log f_2(y_i; \psi))$

(vi) Example: $f_j(y_i; \psi) \; = \; \psi_j^{-1} \exp(-y_i / \psi_j)$
$\Sigma_i \, \delta_i \log f_1(y_i; \psi) \; = \; \log \psi_1 \, \Sigma_i \, \delta_i - \Sigma_i \, \delta_i y_i$
$\tilde{\psi}_1 \; = \; \Sigma_i \, \delta_i y_i / (\Sigma_i \, \delta_i)$**,** $\tilde{\psi}_2 \; = \; \Sigma_i (1 - \delta_i) y_i / \Sigma_i (1 - \delta_i)$**.**

(vii) Be careful about identifiability – exchanging the probs and labels on components gives same mixture: e.g. fix $\psi_1 < \psi_2$**.**

7.6 Other types of example

(i) Missing data – actual

Caution: we do NOT "use expectations to impute the missing data".

(ii) Variance component models (see hwk 9)

(a) $Y = AZ + e$**,** $e \sim N_n(0, \tau^2)$**,** $Z \sim N_r(0, \sigma^2 G)$**,** A **is** $n \times r$ **matrix.**

$Y \sim N_n(0, \sigma^2 AGA' + \tau^2 I)$

(b) $X = (Y, Z)$**:**

$$\begin{aligned} \ell_c(\sigma^2, \tau^2) &= -(n/2)\log(\tau^2) - (r/2)\log(\sigma^2) \\ &\quad -(2\tau^2)^{-1}(y - Az)'(y - Az) - (2\sigma^2)^{-1} z' G^{-1} z \end{aligned}$$

(c) $\tilde{\sigma^2} = r^{-1}\mathrm{E}(z'G^{-1}z|y)$**,** $\tilde{\tau^2} = n^{-1}\mathrm{E}((Y - AZ)'(Y - AZ)|Y)$**, but** $\mathrm{E}(z'G^{-1}z|y) \neq \mathrm{E}(z|y)'G^{-1}\mathrm{E}(z|y)$**.**

(d) Note $X \sim N_{n+r}(0, V)$**, so we have usual formulae**
$\mathrm{E}(Z|Y) = V_{zy}V_{yy}^{-1}Y$**,** $\mathrm{var}(Z|Y) = V_{zz} - V_{zy}V_{yy}^{-1}V_{yz}$**.**

(e) Also, if $\mathrm{E}(W) = 0$**,**
$\mathrm{E}(W'BW) = \mathrm{E}(\Sigma_{i,j} W_i W_j B_{ij}) = \Sigma_{i,j} \mathrm{var}(W)_{ij} B_{ij} = \mathbf{tr}(\mathrm{var}(W)B)$**.**

(iii) Censored data, age-of-onset-data, competing risks models, etc.

(iv) Hidden states, latent variables:

Models in Genetics, Biology, Climate modelling, Environmental modelling.

(v) General auxiliary variables: the latent variables do not have to mean anything – they are simply a tool, s.t. that the complete-data log-likelihood is easy.

7.7 Likelihood in Exponential families

(i) See Chapter 2, for definitions, the natural sufficient statistics T_j,
the natural parameter space and parametrization π_j.

(iii) Moment formulae. see 2.2 (vii)

$$\begin{aligned} \mathrm{E}(t_j(X)) &= -\frac{\partial \log c(\pi)}{\partial \pi_j} \\ \mathrm{Cov}(t_j(X), t_j(X)) &= -\frac{\partial^2 \log c(\pi)}{\partial \pi_j \partial \pi_l} \end{aligned}$$

(iv) Likelihood equation for exponential family – see 4.6

$$\frac{\partial \ell}{\partial \pi_j} = n(n^{-1} \sum_1^n t_j(X_i) - \mathrm{E}(t_j(X)))$$

The natural sufficient statistics T_j are set equal to their expectations $n\tau_j$.

(v) The information results (see 4.6)

$$\begin{aligned} I(\pi) = J(\pi) &= \mathrm{var}(T_1, ..., T_k) \\ I(\tau) &= (\mathrm{var}(T_1, ..., T_k))^{-1} \quad \text{where } \tau_j = \mathrm{E}(t_j(X)) \end{aligned}$$

$(T_1, ..., T_k)$ **achieves (multiparameter) CRLB for** $(\tau_1, ..., \tau_k)$**.**

(v) Suppose complete-data X has exp.fam. form: for n-sample $T_j(X) = \Sigma_{i=1}^n t_j(X_i)$

$$\begin{aligned} \log g_\theta(X) &= \log c(\theta) + \sum_{j=1}^k \pi_j(\theta) T_j(X) + \log w(X) \\ Q(\theta; \theta^*) &= \log c(\theta) + \sum_{j=1}^k \pi_j(\theta) \mathrm{E}_{\theta^*}(T_j(X)|Y) + \mathrm{E}_{\theta^*}(\log w(X)|Y). \end{aligned}$$

7.8 EM for exponential families

(i) In natural parametrization π_j:

$$\begin{aligned} Q(\pi;\pi^*) &= \log c(\pi) + \sum_{j=1}^{k} \pi_j \mathrm{E}_{\pi^*}(T_j(X)|Y) \\ \frac{\partial Q}{\partial \pi_j} &= \mathrm{E}_{\pi^*}(T_j(X)|Y) + \frac{\partial}{\partial \pi_j} \log c(\pi) \\ &= \mathrm{E}_{\pi^*}(T_j(X)|Y) - \mathrm{E}_{\pi}(T_j(X)) \end{aligned}$$

Thus EM iteratively fits unconditioned to conditioned expectations of T_j. At MLE $\mathrm{E}_{\pi^*}(T_j(X)|Y) = \mathrm{E}_{\pi^*}(T_j(X))$.

(ii) Recall

$$\begin{aligned} \ell(\pi) &= \log g_\pi(X) - \log h_\pi(X|Y) \\ \text{but} \quad h_\pi(X|Y) &= \frac{w(X)\exp(\sum_j \pi_j t_j(X))}{\int_{y(X)=y} w(X)\exp(\sum_j \pi_j t_j(X)) dX} \\ &= c^*(\pi;Y) w(X) \exp(\sum_j \pi_j t_j(X)) \\ \text{so} \quad \ell(\pi) &= \log c(\pi) - \log c^*(\pi;Y) \end{aligned}$$

(iii) Hence, differentiating this:

$$\frac{\partial \ell}{\partial \pi_j} = -\mathrm{E}_\pi(T_j) + \mathrm{E}_\pi(T_j|Y) \qquad \text{At MLE}: \ \mathrm{E}_\pi(T_j) = \mathrm{E}_\pi(T_j|Y)$$

(iv) Differentiating again:

$$-\frac{\partial^2 \ell}{\partial \pi_j \partial \pi_l} = \mathrm{Cov}(T_j, T_l) - \mathrm{Cov}((T_j, T_l)|Y)$$

If Y determines X, $\mathrm{var}(T(X)|Y) = 0$, and then observed information is $\mathrm{var}(T)$ as for any exp fam.
If Y tells nothing about X, $\mathrm{var}(T(X)|Y) = \mathrm{var}(T(X))$, and observed information is 0.
"Information lost" due to observing Y not X is $\mathrm{var}(T(X)|Y)$.