Eugene Shvarts
STA 250

Homework 3 – Optimization & EM will be shortly emailed.

1. For this Q, code and calculations ~~are attached~~.

2. For $i = 1:n$, consider the heirarchal model $\begin{cases} Y_i \mid \tau_i \overset{iid}{\sim} N_p(\mu, \frac{1}{\tau_i}\Psi) \\ \tau_i \sim \frac{1}{\nu}\chi^2_\nu \end{cases}$

Then we have an AA model for EM with
$Y_0 := Y = \langle Y_1 \cdots Y_n \rangle$, $Y_m := \tau = \langle \tau_1 \cdots \tau_n \rangle$, $\theta := \langle \mu, \Psi \rangle$,
assuming ν is provided. Note $\mu \in \mathbb{R}^p$, $\Psi \in \mathbb{R}^{p \times p}$.

We are asked to give the updates $\theta^{(t)} \rightsquigarrow \theta^{(t+1)}$.

First, we give $\log p(Y_0, Y_m \mid \theta) = \log p(Y \mid \tau, \mu, \Psi) + \log p(\tau)$,
suppressing those terms independent of μ, Ψ:

$$p(Y \mid \tau, \mu, \Psi) = \prod_{i=1}^{n} \left(\frac{\tau_i}{2\pi}\right)^{p/2} |\Psi|^{-\frac{1}{2}} \exp\left\{-\frac{\tau_i}{2}(Y_i - \mu)^T \Psi^{-1}(Y_i - \mu)\right\}$$

$$p(\tau) = \prod_{i=1}^{n} \nu^{-1} 2^{-\frac{\nu}{2}} \left(\Gamma\left(\tfrac{\nu}{2}\right)\right)^{-1} \tau_i^{\frac{\nu}{2}-1} \exp\left\{-\frac{\tau_i}{2}\right\}$$

Note in particular that $p(\tau)$ will not affect the maximization of $Q(\theta \mid \theta')$.
Taking the log, the relevant terms are:

$$\sum_{i=1}^{n} \left\{-\tfrac{1}{2}\log|\Psi| - \frac{\tau_i}{2}(Y_i - \mu)^T \Psi^{-1}(Y_i - \mu)\right\}$$

$$= -\frac{1}{2}\left\{n \log|\Psi| + \sum_{i=1}^{n} \tau_i (Y_i - \mu)^T \Psi^{-1}(Y_i - \mu)\right\}.$$

Now, we examine the second term. A quadratic form can be

* $\operatorname{tr}(cA) = \sum_i (cA)_{ii} = \sum_i cA_{ii} = c\operatorname{tr}A$

seen as the trace of a certain matrix product:

$$\operatorname{tr}(\underbrace{xx^T}_{\text{data matrix}}A) = \sum_i (xx^T)_{ii} A_{ii} = \sum_i x_i A_{ii} x_i = \sum_i (x^TA)_i x_i = x^TAx$$

Hence, we may rewrite this term as

$$\sum_{i=1}^n \tau_i \operatorname{tr}\left((y_i-u)(y_i-u)^T\Psi^{-1}\right) \overset{*}{=} \sum_{i=1}^n \operatorname{tr}\left(\tau_i (y_i-u)(y_i-u)^T\Psi^{-1}\right)$$

by additivity of trace

$$= \operatorname{tr}\left\{\underbrace{\left(\sum_{i=1}^n \tau_i (y_i-u)(y_i-u)^T\right)}_{\text{call this matrix } B.}(\Psi^{-1})\right\}$$

Then our log-likelihood looks like

$-\frac{1}{2}\left\{n \log|\Psi| + \operatorname{tr}(B\Psi^{-1})\right\}$. Now we consider

$Q(\theta|\theta') = \mathbb{E}_{\tau|Y,\theta'}\left[\log p(Y,\tau|\theta) \mid Y,\theta'\right]$, where θ' is a particular value of u, Ψ. Considering only the same terms, τ appears only linearly, in the expression for B:

$$\mathbb{E}[B \mid Y,\theta'] = \sum_{i=1}^n \mathbb{E}[\tau_i \mid Y,\theta'](y_i-u)(y_i-u)^T.$$

For convenience we denote $\tau_i' := \mathbb{E}[\tau_i \mid Y,\theta']$, and we are given the fact that

$$\tau_i' = \frac{\nu+p}{\nu + (y_i-u')^T(\Psi')^{-1}(y_i-u')}$$

Denote the corresponding matrix B'. Then (up to relevant terms), $Q(\theta|\theta') = -\frac{1}{2}\left\{n\log|\Psi| + \operatorname{tr}(B'\Psi^{-1})\right\}$.

↪

HW 3 cont.

2. cont. Now, given $\theta^{(t)}$, the E-step is computing $\tau'(\theta^{(t)}) = (\tau_1' \cdots \tau_n')$, using $\theta^{(t)}$ as the value of θ'.
The M-step is finding the argmax wrt each parameter.

To find $\mu^{(t+1)}$, we need to find the maximum of the (negative-semidefinite) quadratic form $-\frac{1}{2}\operatorname{tr}(B'\Psi^{-1}) = -\frac{1}{2}\sum_{i=1}^{n} \tau_i'(y_i-\mu)^T\Psi^{-1}(y_i-\mu)$.
Fortunately such a form always has a unique maximizer, found by observing the stationary point. In general, applying $\frac{\partial}{\partial x_\kappa} := \partial_\kappa$ to x^TAx yields

$$\partial_\kappa(x^TAx) = \partial_\kappa \sum_{ij} A_{ij}x_ix_j = \sum_{ij} A_{ij}[x_i\delta_{j\kappa} + x_j\delta_{i\kappa}]$$

$$= \sum_j [A_{j\kappa} + A_{\kappa j}]x_j = (A+A^T)_{\kappa\cdot}x$$

where $\kappa\cdot$ subscript denotes the κ^{th} row. Applying this in our case, concatenating the p row-equations, and setting to 0 yields

$\sum_{i=1}^{n} \tau_i'(\Psi^{-1}+\Psi^{-1^T})(y_i-\mu) = 0$. As a covariance matrix is symmetric, pos-semi-definite, this simplifies to

$\sum_{i=1}^{n} \tau_i'\Psi^{-1}y_i = \sum_{i=1}^{n} \tau_i'\Psi^{-1}\mu$, and so

$$\mu^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_i' y_i}{\sum_{i=1}^{n} \tau_i'}$$

For Ψ, we are supplied that $\operatorname{argmax}_A\{n\log|A^{-1}| - \operatorname{tr}(A^{-1}B)\} = \frac{1}{n}B$.

Clearly our $Q(\theta|\theta')$ can easily be rearranged into this form.

(4). $|\Psi^{-1}| = \frac{1}{|\Psi|}$ since $|\Psi||\Psi^{-1}| = |\Psi\Psi^{-1}| = 1$

Doing so yields $\frac{1}{2}\{n \log|\Psi^{-1}| - tr(\Psi^{-1}B')\}$,

so $\Psi^{(t+1)} = \frac{1}{2n}B'$ (where B' has τ' evaluated at $\theta^{(t)}$.) □

3. For $i = 1:n$, consider the heirarchal model $\begin{cases} Y_i|\lambda_i \sim Pois(\lambda_i) \\ \lambda_i|\beta \sim Gamma(1,\beta) \end{cases}$

This is an SA model $Y_m = \lambda$, $Y_o = Y$, $\theta = \beta$.

For reference, $Y \sim Pois(\lambda)$, $p(y) = \frac{1}{y!} e^{-\lambda}\lambda^y$

$X \sim Gamma(\alpha,\beta)$, $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp\{-\beta x\}$

(a) Derive the EM model for β in the SA framework.

To compute $Q(\beta|\beta')$, we need the β-having terms of $\log p(Y,\lambda|\beta)$.
These are: $\sum_i \{\log\beta - \lambda_i\beta\} = n\log\beta - \beta\sum_i \lambda_i$
Taking the conditional expectation, we have
$n\log\beta - \beta\sum_i \mathbb{E}[\lambda_i|Y,\beta']$, so we need Bayes rule.

$p(\lambda_i|Y,\beta) \propto p(Y_i|\lambda_i)\,p(\lambda_i|\beta)$ as a function of λ_i; that is
$p(\lambda_i|Y,\beta) \propto e^{-\lambda_i}\lambda_i^{Y_i} e^{-\beta\lambda_i} = \lambda_i^{Y_i} e^{-(\beta+1)\lambda_i} \sim Gamma(Y_i+1, \beta+1)$

If $X \sim Gamma(\alpha,\beta)$, $\mathbb{E}[X] = \frac{\alpha}{\beta}$. Hence,
$Q(\beta|\beta') = n\log\beta - \beta\sum_i \frac{Y_i+1}{\beta'+1}$ for fixed β'.

Taking $\beta' = \beta^{(t)}$, our iteration is found by the argmax:

↪

(5)

let $\bar{y} := \frac{1}{n}\sum_i y_i$

HW 3 cont.

3(a) cont. $\partial_\beta Q(\beta|\beta^{(t)}) = \frac{n}{\beta} - \sum_i \frac{y_i+1}{\beta^{(t)}+1} \Rightarrow 0$, thus $\underline{\beta^{(t+1)} = \frac{\beta^{(t)}+1}{\bar{y}+1}}$.

(Here the E-step was performed analytically.

(b) Construct an AA model which preserves $p(Y|\beta)$ in the above model.

Here we need to trivialize the distribution of $\lambda|\beta$; because it is part of an exponential family, consider $\tilde{\lambda}_i = \beta\lambda_i$ as a transformation.

Then $p(\tilde{\lambda}_i \le x) = p(\beta\lambda_i \le x) = p(\lambda_i \le \frac{x}{\beta}) = \int_0^{\frac{x}{\beta}} \beta e^{-\beta\lambda}\, d\lambda$

$= 1 - e^{-x}$.

Hence $p(\tilde{\lambda}_i) = e^{-\tilde{\lambda}_i}$, or $\tilde{\lambda}_i \sim \text{Gamma}(1,1)$.

Consequently, our model becomes $\begin{cases} y_i|\tilde{\lambda}_i, \beta \sim \text{Pois}(\frac{1}{\beta}\tilde{\lambda}_i) \\ \tilde{\lambda}_i \sim \text{Gamma}(1,1) \end{cases}$

Now, our log-likelihood (with only β-relevant terms) is

$$\sum_i \left\{ -\frac{1}{\beta}\tilde{\lambda}_i + y_i \log\frac{1}{\beta} \right\} = -\frac{1}{\beta}\sum_i \tilde{\lambda}_i - n\bar{y}\log\beta$$

Taking the conditional expectation, we have

$-\frac{1}{\beta}\sum_i \mathbb{E}[\tilde{\lambda}_i | Y, \beta'] - n\bar{y}\log\beta$, so we need Bayes rule.

$p(\tilde{\lambda}_i|Y,\beta) \propto p(y_i|\tilde{\lambda}_i,\beta)\,p(\tilde{\lambda}_i)$ as a function of $\tilde{\lambda}_i$; that is,

$p(\tilde{\lambda}_i|Y,\beta) \propto e^{-\frac{1}{\beta}\tilde{\lambda}_i}(\frac{1}{\beta}\tilde{\lambda}_i)^{y_i} \cdot e^{-\tilde{\lambda}_i} \sim \text{Gamma}(y_i+1, \frac{1}{\beta}+1)$.

So, $\mathbb{E}[\tilde{\lambda}_i|Y,\beta] = \frac{y_i+1}{\frac{1}{\beta}+1} = \underline{\frac{\beta}{\beta+1}(y_i+1)}$, Putting this together,

$$Q(\beta|\beta') = -\frac{1}{\beta}\left(\frac{\beta'}{\beta'+1}\right)n(\bar{y}+1) - n\bar{y}\log\beta.$$

So to find the argmax,

(6)

$$\partial_\beta Q(\beta|\beta^{(t)}) = \frac{1}{\beta^2}\left(\frac{\beta^{(t)}}{\beta^{(t)}+1}\right) n(\bar{Y}+1) - \frac{1}{\beta} n\bar{Y} \Rightarrow 0.$$

Thus $\underline{\beta^{(t+1)} = \frac{\beta^{(t)}}{\beta^{(t)}+1}\left(1+\frac{1}{\bar{Y}}\right)}$

(c) Using the SA and AA from (a),(b), derive the corresponding Interwoven EM algorithm.

Let's take SA as A1, AA as A2. The IEM Q-function is

$$Q_I(\theta|\theta') = \mathbb{E}_{A2}\left[\mathbb{E}_{A1}\left[\log P_{A1}(Y_o, Y_m|\theta) \,\middle|\, Y_o, \tilde{Y}_m, G_{A2}(\theta')\right] \,\middle|\, Y_o, \theta'\right],$$

where $\mathbb{E}_{A2}$ integrates $\tilde{Y}_m | Y_o, \theta'$, $\mathbb{E}_{A1}$ integrates $Y_m | Y_o, \tilde{Y}_m, G_{A2}(\theta')$, and G_{A2} maps θ to its EM update under A2.

$\log P_{A1}$ up to β-terms is $n\log\beta - \beta\sum_i \lambda_i$, so taking $\mathbb{E}_{A1}$ yields $n\log\beta - \beta\sum_i \mathbb{E}_{A1}[\lambda_i | Y, \tilde{\lambda}, G_{A2}(\beta')]$

The relationship between $\lambda_i, \tilde{\lambda}$ is deterministic, so we have the same A1 expectation as before: $\frac{Y_i+1}{\beta+1}$. Conditioning on β, we get $\mathbb{E}_{A1}[\lambda_i | Y, \tilde{\lambda}, G_{A2}(\beta')] = (Y_i+1) \Big/ \left[\left(\frac{\beta'}{\beta'+1}\right)\left(1+\frac{1}{\bar{Y}}\right)+1\right]$

Because $\tilde{\lambda}$ doesn't appear, $\mathbb{E}_{A2}$ has no effect, so:

$$Q_I(\beta|\beta') = n\left[\log\beta - \beta\frac{1}{\left(\frac{\beta'}{\beta'+1}\right)\left(1+\frac{1}{\bar{Y}}\right)+1}\left(\bar{Y}+1\right)\right]$$

$$= n\left[\log\beta - \beta\frac{1}{\left(\frac{\beta'}{\beta'+1}\right)\frac{1}{\bar{Y}} + \frac{1}{1+\bar{Y}}}\right].$$

up to β-terms.

↪

HW 3 cont.

3 (c) cont. To find the argmax, $\partial_\beta Q(\beta|\beta^{(t)}) = n\left[\frac{1}{\beta} - \frac{1}{\left(\frac{\beta^{(t)}}{\beta^{(t)}+1}\right)\frac{1}{\bar{y}} + \frac{1}{\bar{y}+1}}\right] \Rightarrow 0$,

so $\underline{\beta^{(t+1)} = \frac{\beta^{(t)}}{\beta^{(t)}+1}\frac{1}{\bar{y}} + \frac{1}{1+\bar{y}}}$.

(d) Derive the observed-data log-likelihood and compute the MLE. (we use the SA model)

We know $p(y|\beta) = \int p(y|\lambda,\beta)\, p(\lambda|\beta)\, d\lambda$

$= \prod_i \int p(y_i|\lambda_i) p(\lambda_i|\beta) d\lambda_i$,

so that $\log p(y|\beta) = \sum_i \log \int p(y_i|\lambda_i) p(\lambda_i|\beta) d\lambda_i$.

$\int p(y_i|\lambda_i) p(\lambda_i|\beta) d\lambda_i = \frac{\beta}{y_i!} \underbrace{\int e^{-\lambda_i} \lambda_i^{y_i} e^{-\beta\lambda_i} d\lambda_i}$

$\sim \text{Gamma}(y_i+1, \beta+1)$

$= \frac{\beta}{y_i!}\frac{y_i!}{(\beta+1)^{y_i+1}} = \underline{\frac{\beta}{\beta+1}(\beta+1)^{-y_i}}$

Hence $\log p(y|\beta) = \sum_i \log\beta - (y_i+1)\log(\beta+1)$

$= n\left[\log\beta - (\bar{y}+1)\log(\beta+1)\right]$

To find the MLE, $\partial_\beta \log p(y|\beta) = n\left[\frac{1}{\beta} - (\bar{y}+1)\frac{1}{\beta+1}\right] \Rightarrow 0$,

then $\underline{\hat{\beta} = \frac{1}{\bar{y}}}$

(e) Compare the linear rate of convergence of each EM algorithm, and discuss when each will perform well.

The SA update is $\beta \mapsto \frac{\beta+1}{\bar{\gamma}+1}$. Its fixed points satisfy $\beta^* = \frac{\beta^*+1}{\bar{\gamma}+1}$, so that $\beta^* = \frac{1}{\bar{\gamma}}$ as desired. The linear convergence rate is found by

$$\lim_{\beta\to\beta^*} \frac{f(\beta)-\beta^*}{\beta-\beta^*} = \lim_{\beta\to\frac{1}{\bar{\gamma}}} \frac{\frac{\beta+1}{\bar{\gamma}+1} - \frac{1}{\bar{\gamma}}}{\beta - \frac{1}{\bar{\gamma}}} = \frac{\bar{\gamma}\beta+\bar{\gamma}-\bar{\gamma}-1}{(\bar{\gamma}\beta-1)(1+\bar{\gamma})} \to \frac{1}{1+\bar{\gamma}}$$

The AA update is $\beta \mapsto \frac{\beta}{\beta+1}\left(1+\frac{1}{\bar{\gamma}}\right)$. Its fp. satisfy $\beta^* = \frac{\beta^*}{\beta^*+1}\left(1+\frac{1}{\bar{\gamma}}\right)$, so that $\beta^* = 0$ or $\frac{1}{\bar{\gamma}}$. The derivative of the map is positive, and $1+\frac{1}{\bar{\gamma}} > 1$, so that 0 is unstable and $\frac{1}{\bar{\gamma}}$ is stable. The linear rate is

$\left(f'(0) = 1+\frac{1}{\bar{\gamma}} > 1\right)$ $\left(f'\left(\frac{1}{\bar{\gamma}}\right) = \frac{1}{1+\frac{1}{\bar{\gamma}}} < 1\right)$

$$\lim_{\beta\to\frac{1}{\bar{\gamma}}} \frac{\frac{\beta}{\beta+1}\left(1+\frac{1}{\bar{\gamma}}\right) - \frac{1}{\bar{\gamma}}}{\beta - 1/\bar{\gamma}} = \frac{\beta(\bar{\gamma}+1)-(\beta+1)}{(\beta+1)(\bar{\gamma}\beta-1)} = \frac{1}{\beta+1} \to \frac{\bar{\gamma}}{1+\bar{\gamma}}$$

The IEM update is $\beta \mapsto \frac{\beta}{\beta+1}\frac{1}{\bar{\gamma}} + \frac{1}{1+\bar{\gamma}}$. Its fp. satisfy $\beta^* = \frac{\beta^*}{\beta^*+1}\frac{1}{\bar{\gamma}} + \frac{1}{1+\bar{\gamma}}$, so that β^* satisfies a quadratic eqn. with solns. $\beta^* = \frac{1}{\bar{\gamma}}, -\frac{\bar{\gamma}}{1+\bar{\gamma}}$. Because negative β are inadmissible, and the f.p. at $1/\bar{\gamma}$ is stable, we have

$\left(f'(1/\bar{\gamma}) = \frac{\bar{\gamma}}{(1+\bar{\gamma})^2} < 1\right)$

the desired result. The linear rate is

$$\lim_{\beta\to\frac{1}{\bar{\gamma}}} \frac{\frac{\beta}{\beta+1}\frac{1}{\bar{\gamma}} + \frac{1}{1+\bar{\gamma}} - \frac{1}{\bar{\gamma}}}{\beta - 1/\bar{\gamma}} = \frac{\beta\bar{\gamma}+\bar{\gamma}-\bar{\gamma}-1}{(\beta+1)(\bar{\gamma}+1)(\beta\bar{\gamma}-1)} = \frac{1}{(\beta+1)(\bar{\gamma}+1)} \to \frac{\bar{\gamma}}{(1+\bar{\gamma})^2}$$

Notice that this rate is the product of the individual rates, and so guaranteed to outperform either. When $\bar{\gamma} \ll 1$, the AA method beats the SA, and is on par with IEM. When $\bar{\gamma} \gg 1$, the SA method beats the AA, and IEM $\sim \frac{1}{2(1+\bar{\gamma})}$ is about twice as fast as SA. When $\bar{\gamma} \approx 1$, SA $\sim \frac{1}{2}$, AA $\sim \frac{1}{2}$, and IEM $\sim \frac{1}{4}$; again about twice as fast.