HW3 solutions

Problem 1 (a) The likelihood function for θ is

$$L(\theta) = \prod_{j=1}^{n} \frac{(x_j\theta + r_j)^{y_j}}{y_j!} e^{-(x_j\theta + r_j)}$$

(b) The conditional p.d.f for $z_1 = (z_{j1})'_{1 \leq j \leq n}$ is

$$p(z_1|x, r, y, \theta^{(t)}) = \prod_{j=1}^{n} \frac{(x_j\theta^{(t)})^{z_{j1}} r_j^{y_j - z_{j1}}}{(x_j\theta^{(t)} + r_j)^{y_j}} C_{y_j}^{z_j}$$

then the conditional expectation of under $p(z_1|x, r, y, \theta^{(t)})$

$$\mathbb{E}_{p(z_1|x,r,y,\theta^{(t)})} \log p(z_1, y|x, r, \theta) = \log\theta \sum_{j=1}^{n} \frac{y_j x_j \theta^{(t)}}{x_j\theta^{(t)} + r_j} - \theta \sum_{i=1}^{n} x_i + C$$

where C does not depend on θ. Then the update rule for $\theta^{(t)}$ is

$$\theta^{(t+1)} = \arg\min_{\theta} \mathbb{E}_{p(z_1|x,r,y,\theta^{(t)})} \log p(z_1, y|x, r, \theta) = \frac{\sum\limits_{j=1}^{n} \frac{y_j x_j \theta^{(t)}}{x_j\theta^{(t)} + r_j}}{\sum\limits_{i=1}^{n} x_i}$$

(c) The MLE of θ is $\hat{\theta} = 5.606063396561341$.
(d) The observed Fisher information is

$$I_{observed} = -\frac{\partial^2 l(\theta)}{\partial\theta^2} = \sum_{j=1}^{n} \frac{y_j x_j^2}{(x_j\theta + r_j)^2}$$

and the complete information is

$$I_{complete} = \mathbb{E}_{p(z_1|y,\theta)}(-\nabla^2 \log p(y, z_1|\theta)) = \sum_{j=1}^{n} \frac{x_j y_j}{(x_j\theta + r_j)\theta}$$

Then the fraction of missing information is

$$\frac{I_{missing}}{I_{complete}} = \frac{I_{complete} - I_{observed}}{I_{complete}} \approx 0.0638$$

Problem 2 (a) We have the log-likelihood

$$\ell(\mu, \sigma; Y, X) = -\frac{n}{2} \ln(2\pi\sigma^2) - \sum_{j=1}^{n} \frac{(x_j - \mu)^2}{2\sigma^2}.$$

We derive the EM algorithm as follows:

- **E-step**: We first derive the posterior of x_j. By the Bayes formula

$$p_{x_j|y_j}(x_j; y_j) \propto p_{y_j|x_j}(y_j; x_j) p_{x_j}(x_j) \propto \frac{\delta(|x_j| - y_j)}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}},$$

where $\delta(\cdot)$ is the delta function. From which we yield

$$x_j|y_j = \begin{cases} y_j, & \text{with probability } p_j, \\ -y_j, & \text{with probability } 1 - p_j. \end{cases}$$

Where $p_j = \frac{e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}}{e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}+e^{-\frac{(y_j+\mu)^2}{2\sigma^2}}}$. Thus the Q-function is

$$\begin{aligned}Q^{(t)}(\mu,\sigma) =&\mathbb{E}\left[\ell(\mu,\sigma;Y,X)\middle|Y,\mu^{(t)},\sigma^{(t)}\right]\\ =&-\frac{n}{2}\ln(2\pi\sigma^2)-\sum_{j=1}^n\frac{p_j(y_j-\mu)^2+(1-p_j)(y_j+\mu)^2}{2\sigma^2}.\end{aligned}$$

- **M-step**: Note that

$$\begin{aligned}\frac{\partial Q^{(t)}(\mu,\sigma)}{\partial\mu} =&\sum_{j=1}^n\frac{p_j(y_j-\mu)-(1-p_j)(y_j+\mu)}{\sigma^2},\\ \frac{\partial Q^{(t)}(\mu,\sigma)}{\partial\sigma^2} =&-\frac{n}{2\sigma^2}+\sum_{j=1}^n\frac{p_j(y_j-\mu)^2+(1-p_j)(y_j+\mu)^2}{2\sigma^4}.\end{aligned}$$

  Hence we update μ,σ^2 by

$$\begin{aligned}\mu^{(t+1)} =&\frac{1}{n}\sum_{j=1}^n(2p_j-1)y_j,\\ (\sigma^2)^{(t+1)} =&\frac{1}{n}\sum_{j=1}^n y_j^2-(\mu^{(t+1)})^2.\end{aligned}$$

(b) We observe that

$$(\mu^{(t)},(\sigma^2)^{(t)})\to\begin{cases}(-2.123,4.267), & \mu^{(0)}<0,\\ (0,8.777), & \mu^{(0)}=0,\\ (2.123,4.267), & \mu^{(0)}>0.\end{cases}$$

This is because we always have $y_j>0$ while

$$p_j\begin{cases}>\frac{1}{2}, & \mu>0,\\ =\frac{1}{2}, & \mu=0,\\ <\frac{1}{2}, & \mu<0,\end{cases}$$

together with $p_j(y_j,\mu)=1-p_j(y_j,-\mu)$. Thus starting with $\mu^{(0)}=0$ will always get $\mu^{(t+1)}=0$ while different sign of $\mu^{(0)}$ will get same esstimation of $\hat{\sigma}^2$ but different sign in $\hat{\mu}$.

(c) We have the log-likelihood

$$\ell(\mu,\sigma;Y)=-\frac{n}{2}\ln(2\pi\sigma^2)+\sum_{j=1}^n\ln\left(e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}+e^{-\frac{(y_j+\mu)^2}{2\sigma^2}}\right)$$

with gradient

$$\begin{aligned}\frac{\partial\ell(\mu,\sigma;\boldsymbol{Y})}{\partial\mu} =&\frac{1}{\sigma^2}\sum_{j=1}^n\frac{(y_j-\mu)e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}-(y_j+\mu)e^{-\frac{(y_j+\mu)^2}{2\sigma^2}}}{e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}+e^{-\frac{(y_j+\mu)^2}{2\sigma^2}}},\\ \frac{\partial\ell(\mu,\sigma;\boldsymbol{Y})}{\partial\sigma^2} =&-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{j=1}^n\frac{(y_j-\mu)^2e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}+(y_j+\mu)^2e^{-\frac{(y_j+\mu)^2}{2\sigma^2}}}{e^{-\frac{(y_j-\mu)^2}{2\sigma^2}}+e^{-\frac{(y_j+\mu)^2}{2\sigma^2}}}.\end{aligned}$$

Here's the figure of $\ell^*-\ell$ as functions of the number of iterations of EM algorithm and gradient descent From which we can see that the EM algorithm converges much faster than the gradient descent.

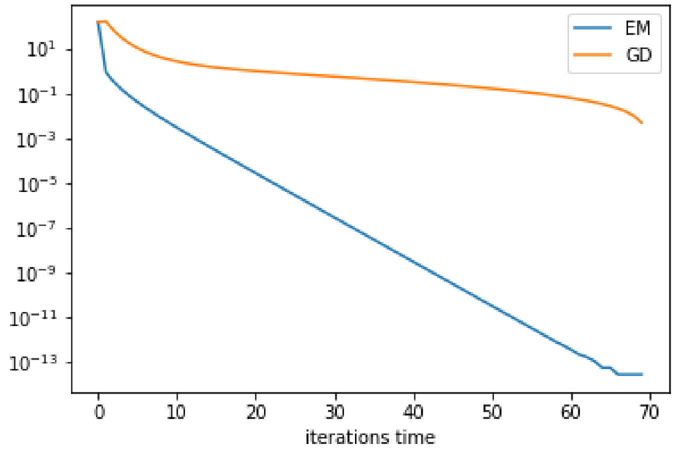


Figure 1: $\ell^* - \ell$ as functions of the number of iterations with initial point $\mu^{(0)} = 1, (\sigma^2)^{(0)} = 1$, the step size of gradient descent is 0.05).

Problem 3 (a) Our notations: $1 \le k \le K = 4$ is index of ancestor population; $1 \le i \le M = 100$ is the index of individual; N_i is the number of genes in individual i; $1 \le j \le N = 200$ is the index of genotype locus. Our variational distribution is

$$q(\theta, z|\gamma, \phi) = \prod_{i=1}^{M} q(\theta_i|\gamma_i) \prod_{n=1}^{N_i} q(z_{in}|\phi_{in})$$

The updules are as follows:

- Update γ:

$$q(\theta_i|\gamma_i) \propto \exp(\mathbb{E}_{q(\theta_{-i},z)} \log p(w, z, \theta)) \propto \exp(\sum_{k=1}^{K}(\alpha_k - 1 + \sum_{n=1}^{N_i} \phi_{ink}) \log \theta_{ik})$$
$$\implies \gamma_{ik}^{\star} = \alpha_k + \sum_{n=1}^{N_i} \phi_{ink}$$

  Since $w_{dn_1} = w_{dn_2}$ implies $\phi_{in_1k} = \phi_{in_2k}$, we can compute the last term using the data matrix D and a compressed version of ϕ in $M_{N\times K}$.

- Update ϕ:

$$q(z_{in}|\phi_{in}) \propto \exp(\mathbb{E}_{q(\theta,-z_{in})} \log p(w, z, \theta)) \propto \exp\left(\sum_{k=1}^{K} 1_{z_{in}=k} \left[\mathbb{E}_{\theta_i} \log \theta_{ik} + \sum_{j=1}^{N} w_{in}^{j} \log \beta_{kj}\right]\right)$$
$$\implies \phi_{ink}^{\star} \propto \exp(\mathbb{E}_{\theta_i} \log \theta_{ik} + \log \beta_{kw_{in}}) = \beta_{kw_{in}} \exp(\psi(\gamma_{ik}) - \psi(\sum_{k=1}^{K} \gamma_{ik}))$$

  where ψ is the digamma function. We should normalize *phi* so that $\sum_{k=1}^{K} \phi_{ink} = 1$.

(b) For individual 1, $n_1 = 71$. We run LDA inference to find ϕ for each genotype locus occuring in individual. The result is shown in Figure 2.

(c) The matrix Θ constructed by LDA inference is shown in Figure 3.

(d) The number of iterations needed to get convergence for each individual is plotted in Figure 4. The total number is 6326.

(e) We compare the behaviors of LDA inference when $\alpha = 0.01, 1, 10$. For each α, the Θ matrix is plotted, and it seems that larger α diversifies the ancestor assignments.

The mean numbers of iteration needed for convergence are $64.55, 36.85, 15.3$ respectively. Together with Figure 5, we can conclude that LDA inference converges faster as α gets larger.

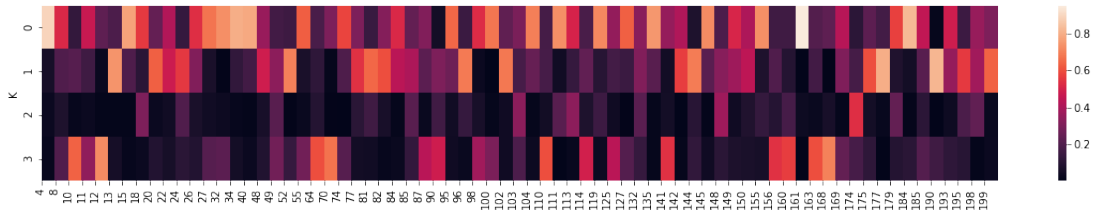


Figure 2: ϕ matrix for individual 1. $\alpha = 0.01$

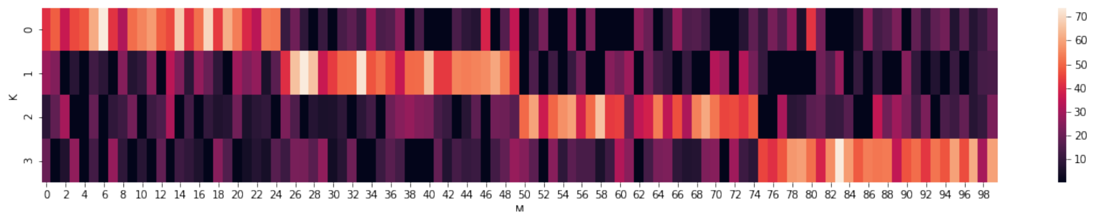


Figure 3: Θ matrix for the dataset. $\alpha = 0.01$

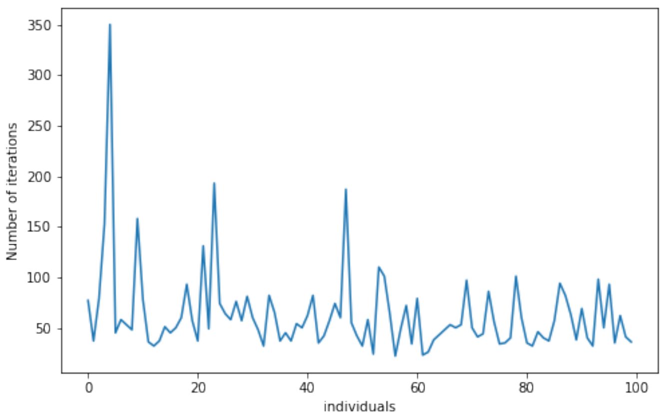


Figure 4: The number of iterations needed to get convergence. $\alpha = 0.01$

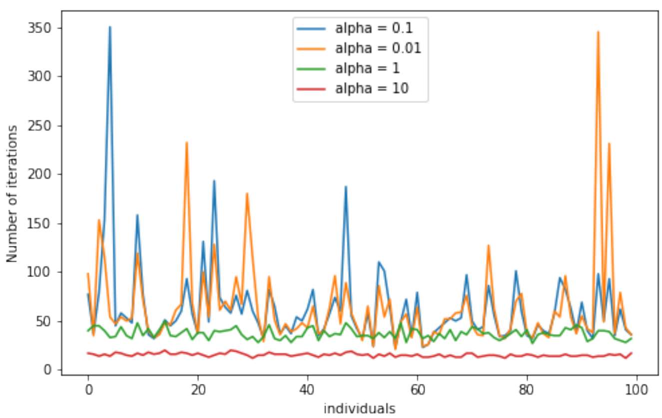
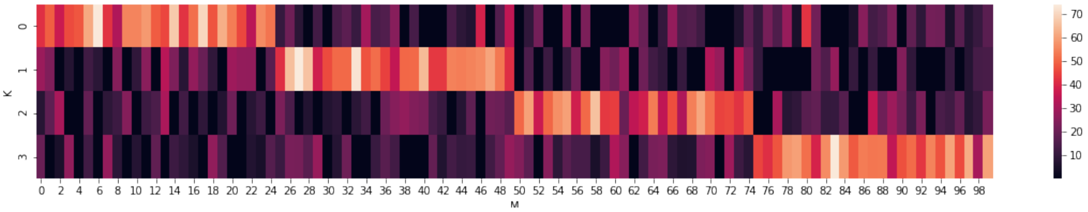
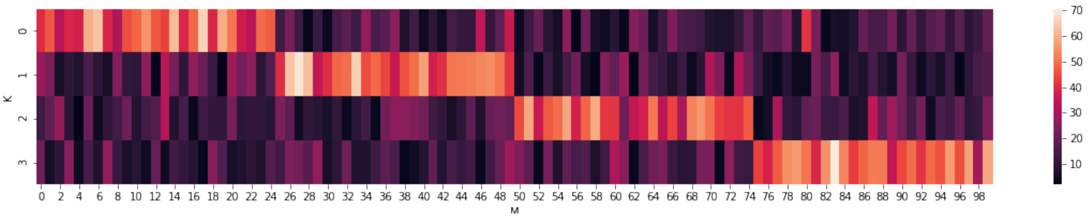


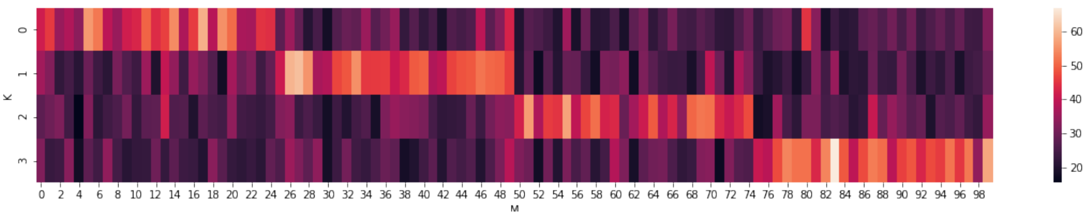
Figure 5: The number of iterations needed to get convergence with different α.



(a) $\alpha = 0.01$



(b) $\alpha = 1$



(c) $\alpha = 10$