CS281: HOMEWORK #5

MIKE SCHACHTER

Problem 1. Given k fixed distributions $p_1(x)$,...,$p_k(x)$, consider the problem of fitting the mixture weights $\theta = (\pi_1, ..., \pi_k)$ to the mixture distribution:

$$p(x|\theta) = \sum_{i=1}^{k} \pi_i p_i(x)$$

Prove that given n i.i.d. samples of X, the loglikelihood $l(\theta|\mathcal{D})$ is concave in θ.
The log likelihood function for a set of data $x = (x^1...x^n)$ is given as:

$$l(x|\theta) = \sum_{i=1}^{n} log \left(\sum_{j=1}^{k} \pi_j p_j(x^i) \right)$$

The first derivative of this function with respect to π_m is given as:

$$\frac{\partial l}{\partial \pi_m} = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} \pi_j p_j(x^i) \right)^{-1} p_m(x^i)$$

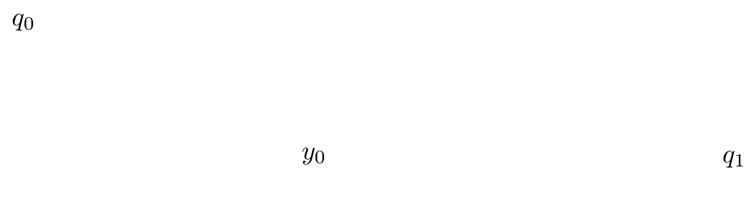
and the second is:

$$\frac{\partial^2 l}{\partial \pi_m^2} = - \sum_{i=1}^{n} \left(p_m(x^i) \right)^{-1} (\pi_m)^{-2}$$

Because $p_m(x^i), \pi_m \geq 0$, the second derivative is always negative, so the log likelihood is concave.

Problem 2. A common modification of HMMs involve using mixture models for the emission probabilities $p(y_t|q_t)$. Assume that $y_t \in \mathbb{R}^d$.

(a) Draw the graphical model for the modified HMM.



An MxM transition matrix A governs the transitions between states, where $a_{ij} = p(q^j_{t+1} = 1 | q^i_t = 1)$. The emission probability of producing an observed state y_t given a hidden state q_t is a mixture of Gaussians:

$$p(y_t | q^j_t = 1) = \sum_{i=1}^{M} \pi_i \mathcal{N}(y_t | \mu_i, \Sigma_i)^{q^i_t} = \pi_j \mathcal{N}(y_t | \mu_j, \Sigma_j) = f_j(y_t)$$

The initial state q_0 has a distribution $\omega = (\omega_1, ..., \omega_M)$, where $\sum_{i=1}^{M} \omega_i = 1$ and $\omega_i = p(q^i_0 = 1)$. As in the course book, we let ω_{q_0} be indexed by whatever the value of q_0 actually is; the same goes for something like $a_{q_t q_{t+1}}$. The parameters of the model at iteration k are $\theta^{(k)} = (\omega^{(k)}, \pi^{(k)}, A^{(k)}, \mu^{(k)}, \Sigma^{(k)})$.

(b) Write the expected complete log likelihood for the model and identify the expecations that are needed in the E step.

The complete likelihood is given as:

$$p(q, y) = \omega_{q_0} \prod_{t=0}^{T-1} a_{q_t q_{t+1}} \prod_{t=0}^{T} f_{q_t}(y_t)$$

Taking the log gives:

$$l(q, y) = log\left(\omega_{q_0}\right) + \sum_{t=0}^{T-1} log\left(a_{q_t q_{t+1}}\right) + \sum_{t=0}^{T} log\left(f_{q_t}(y_t)\right)$$

We need to take the expectation of the complete log likelihood with respect a distribution. For this step, we will have a set of fixed parameters $\theta^{(k)}$, and take the expectation with respect to the density $r = p(q|y, \theta^{(k)})$:

$$\langle l(q, y) \rangle_r = \langle log\left(\omega_{q_0}\right) \rangle_r + \left\langle \sum_{t=0}^{T-1} log\left(a_{q_t q_{t+1}}\right) \right\rangle_r + \left\langle \sum_{t=0}^{T} log\left(f_{q_t}(y_t)\right) \right\rangle_r$$

(c) Outline an algorithm for computing the E step, relating it to the standard alpha and beta recursions.

Much of the notation and ideas shown here come from Jeff Bilm's "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models" (1998). We can evaluate each expectation individually. First the initial state:

$$\langle log\left(\omega_{q_0}\right) \rangle_r = \sum_q p(q|y, \theta^{(k)}) log(\omega_{q_0}) = \sum_{i=1}^{M} \omega^{(k)}_i log(\omega_i)$$

The summation over all possible q contains all possible q_0, so we exchange it for a sum over all each of M states. Next we can deal with the second expectation:

$$\begin{aligned}
\left\langle \sum_{t=0}^{T-1} log\left(a_{q_t q_{t+1}}\right) \right\rangle_r &= \sum_q p(q|y, \theta^{(k)}) \sum_{t=0}^{T-1} log\left(a_{q_t q_{t+1}}\right) \\
&= \sum_{i,j=1}^{m} \sum_{t=0}^{T-1} log\left(a_{ij}\right) p(q^i_t = 1, q^j_{t+1} = 1 | y, \theta^{(k)}) \\
&= \sum_{i,j=1}^{m} \sum_{t=0}^{T-1} log\left(a_{ij}\right) \xi^{ij(k)}_t
\end{aligned}$$

Lastly, we can deal with the emission term:

$$\begin{aligned}\left\langle \sum_{t=0}^{T} log\left(f_{q_t}(y_t)\right)\right\rangle_r &= \sum_q p(q|y,\theta^{(k)}) \sum_{t=0}^{T} log\left(f_{q_t}(y_t)\right) \\ &= \sum_{i=1}^{M}\sum_{t=0}^{T} log\left(f_{q_t}(y_t)\right) p(q_t^i = 1|y,\theta^{(k)})\end{aligned}$$

Now we'll relate things to the alpha and beta step. It's well known that the probability $p(q_t|y,\theta)$ can be written in terms of recursive functions $\alpha(q_t)$ and $\beta(q_t)$, where:

$$\alpha(q_{t+1}) = p(y_{0:t}, q_t) = \sum_{q_t'} \alpha(q_t) a_{q_t q_{t+1}} p(y_{t+1}|q_{t+1})$$

$$\beta(q_t) = p(y_{t+1:T}|q_t) = \sum_{q_{t+1}} \beta(q_{t+1}) a_{q_t q_{t+1}} p(y_{t+1}|q_{t+1})$$

The alpha recursion and beta recursion are initialized with:

$$\alpha(q_0) = p(y_0|q_0)\omega_{q_0}$$

$$\beta(q_T) = 1$$

Using these definitions gives:

$$p(q_t|y,\theta) = \frac{\alpha(q_t)\beta(q_t)}{\sum_{q_t}\alpha(q_t)\beta(q_t)}$$

The term $p(q_t|y,\theta)$ is used in the expecation of the third term in the expected complete log likelihood. Another probability that can be written in terms of alpha and beta functions is:

$$p(q_t, q_{t+1}) = \frac{\alpha(q_t)\beta(q_{t+1}) a_{q_t q_{t+1}} p(y_{t+1}|q_{t+1})}{\sum_{q_t}\alpha(q_t)\beta(q_t)}$$

This is used in the second term of the expected complete log likelihood.

(d) Write down the equations that implement the M step.

Setting to zero the derivative of the first term with respect to ω_j, with Lagrange constraint $\sum_{i=1}^{M}\omega_i = 1$:

$$\begin{aligned}\frac{\partial}{\partial\omega_j}\left[\sum_{i=1}^{M}\omega_i^{(k)} log(\omega_i) - \lambda(1 - \sum_{i=1}^{M}\omega_i)\right] &= 0 \\ \omega_j^{(k+1)} &= \frac{\omega_j^{(k)}}{\lambda} \\ \omega_j^{(k+1)} &= \omega_j^{(k)}\end{aligned}$$

(Taking into account all ω_i gives a Lagrange multiplier of $\lambda = 1$). Doing the same thing to the second term gives:

$$\begin{aligned}
\frac{\partial}{\partial a_{xy}}\left[\sum_{i,j=1}^{m}\sum_{t=0}^{T-1} log\,(a_{ij})\,\xi_t^{ij(k)} - \lambda(1-\sum_{j=1}^{M} a_{xj})\right] &= 0\\
\frac{1}{a_{xy}}\sum_{t=0}^{T-1}\xi_t^{xy(k)} &= \lambda\\
a_{xy}^{(k+1)} &= \frac{\sum_{t=0}^{T-1}\xi_t^{xy(k)}}{\sum_{j=1}^{M}\sum_{t=0}^{T-1}\xi_t^{xj(k)}}
\end{aligned}$$

The third term needs to be differentiated with respect to μ_x , Σ_x, and π_x. First, μ_x:

$$\begin{aligned}
\frac{\partial}{\partial\mu_x}\left[\sum_{i=1}^{M}\sum_{t=0}^{T} log\,(f_{q_t}(y_t))\,p(q_t^i=1|y,\theta^{(k)})\right] &= 0\\
\frac{\partial}{\partial\mu_x}\left[\sum_{i=1}^{M}\sum_{t=0}^{T} -log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right)\sum_{j=1}^{M}\left(\frac{1}{2}(y_t-\mu_j)^T\Sigma_j^{-1}(y_t-\mu_j)^T + log(\pi_j)\right)p(q_t^i=1|y,\theta^{(k)})\right] &= 0\\
\sum_{i=1}^{M}\sum_{t=0}^{T}\left(\frac{\partial}{\partial\mu_x}\left[(y_t-\mu_j)^T\Sigma_j^{-1}(y_t-\mu_j)^T\right]\right)p(q_t^i=1|y,\theta^{(k)}) &= 0\\
\sum_{i=1}^{M}\sum_{t=0}^{T}\left((y_t-\mu_x)^T\Sigma_x^{-1}\right)p(q_t^i=1|y,\theta^{(k)}) &= 0
\end{aligned}$$

Well... that's probably not right. Gonna stop here... It's been a fun semester though!

Problem 3. Two 2-dimensional data sets supplied in pca1.dat and pca2.dat are generated by choosing a line through the origin and then choosing random samples from a univariate Gaussian distribution along that line. These were corrupted by (1) using an additive two-dimensional Gaussian with equal covariances in the y_1 and y_2 directions (pca1.dat) and (2) by using additive two-dimensional Gaussian with greater covariance in the y_2 than in the y_1 direction (pca2.dat).

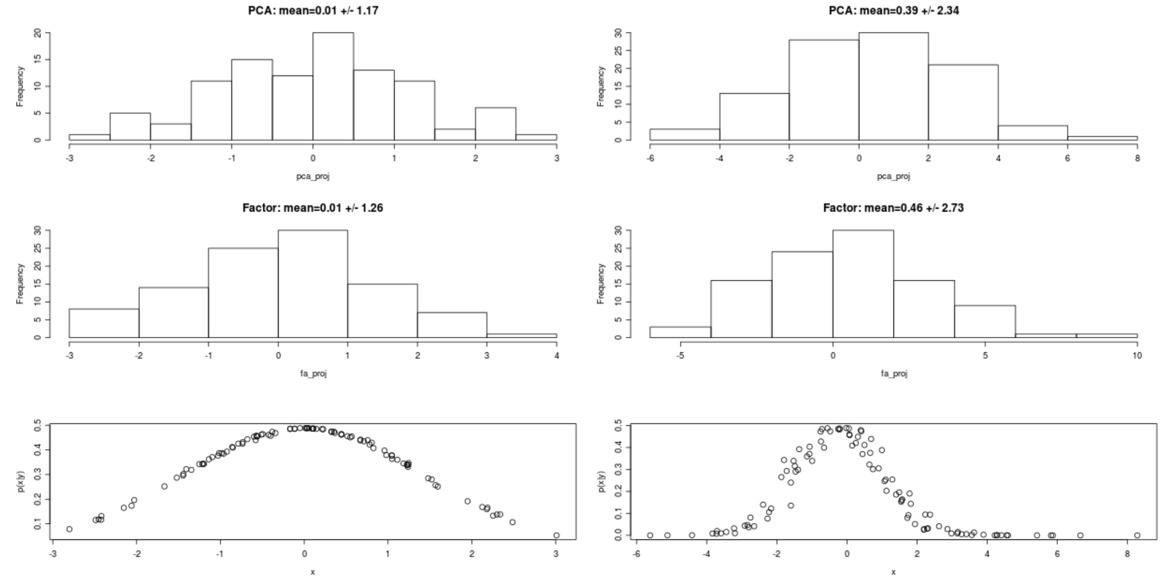
(a) Write an R implementation of PCA. For each data set compute the sample covariance matrix, determine the principle eigenvector, project that data onto the corresponding subspace.

See the figures for part (c) of this problem and the script "homework5_mschachter.R".

(b) Write an R implementation of factor analysis using the EM algorithm discussed in chapter 14 and in class. Once the parameters are determined, for each data point compute the posterior probability $p(x|y)$; the factor analysis equivalent of projecting onto the principle subspace.

The figure in part (c) has this.

(c) Compute the fits for both data sets and plot the resulting projections. What changes and what stays the same?



The left column is for pca1.dat, the right for pca2.dat. The top histograms are 1D projections from the top eigenvector of PCA. The second row are histograms for 1D projections of factor analysis. The third column is the posterior probability $p(x|y)$ from factor analysis. Note in the left column that the means and standard deviations of the projections are very similar between PCA and factor analysis. The right column, the data for pca2.dat, shows that PCA projects differently than factor analysis. The fits are as follows:

pca1.dat	PCA	Factor Analysis
Λ	$\begin{bmatrix} 0.69 \\ 0.72 \end{bmatrix}$	$\begin{bmatrix} 0.68 \\ 0.84 \end{bmatrix}$
ψ	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.29 & 0 \\ 0 & 0.1 \end{bmatrix}$

pca2.dat	PCA	Factor Analysis
Λ	$\begin{bmatrix} 0.16 \\ 0.99 \end{bmatrix}$	$\begin{bmatrix} 0.54 \\ 1.09 \end{bmatrix}$
ψ	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	$\begin{bmatrix} 0.44 & 0 \\ 0 & 4.21 \end{bmatrix}$

Problem 4. Consider an Ising model with binary variables $X_s \in \{-1, 1\}$ and factorization of the form $p(x; \theta) \propto exp\{\sum_{s\in V} \theta_s x_s + \sum_{(s,t)\in E} \theta_{st} x_s x_t\}$. Also assume a 2D grid with toroidal boundary conditions.

(a) Derive the Gibb's sampling updates for the model, implement with $\theta_{st} = 0.2$ and $\theta_s = 0.2 + (-1)^s$. Run a burn in period of 1000 iterations, then sample for 1000 iterations, forming Monte Carlo estimates $\hat{\mu}_s$ of the moments $\mathbb{E}[X_s]$ at each node. Output a 7x7 matrix of the estimated moments, repeating a few times to provide an idea of variability in the estimate

First let's derive the Gibb's update. In order to do so, we need do determine $p(x_k = 1|N(x_k))$, where x_k is the node being updated, and $N(x_k)$ is the set of neighbors. We can apply Bayes rule to give:

$$p(x_k|N(x_k)) = \frac{p(N(x_k), x_k)}{p(N(x_k)|x_k = 1) + p(N(x_k)|x_k = 0)}$$

The probabilities in the numerator and denominator can be expanded using marginalization:

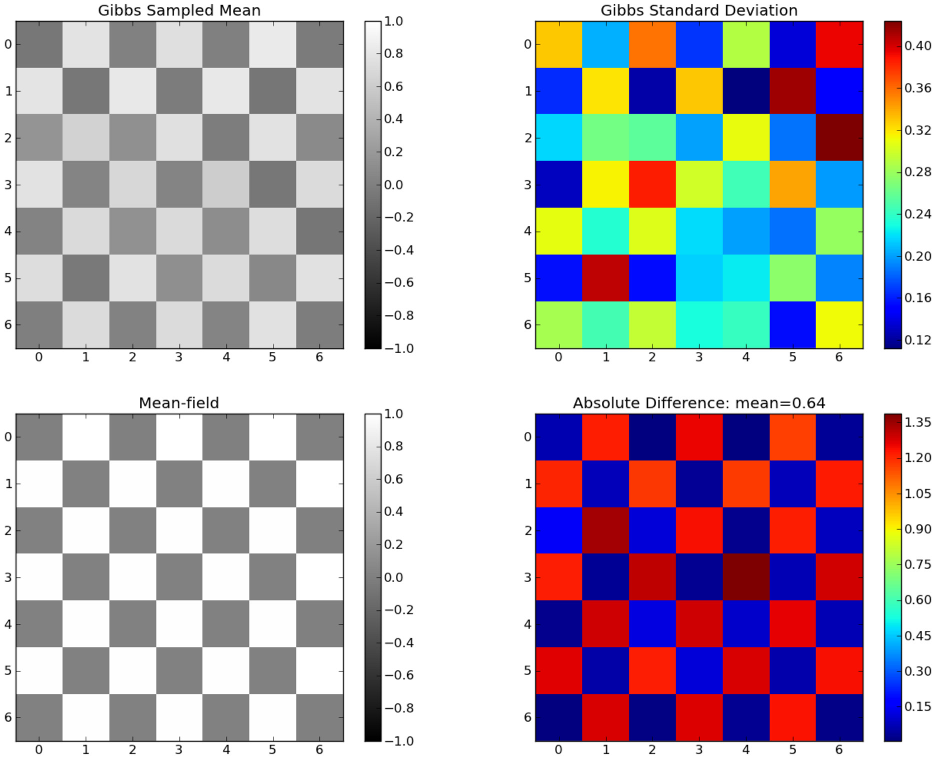
$$\begin{aligned} p(N(x_k)|x_k) &\propto \sum_x exp\{\sum_{s\in V} \theta_s x_s + \sum_{(s,t)\in E} \theta_{st} x_s x_t\} \\ &\propto exp\{\sum_{s\in N(x_k)} \theta_s x_s + \sum_{(s,t)\in N(x_k)} \theta_{st} x_s x_t\} \left(\sum_x exp\{\sum_{s\notin N(x_k)} \theta_s x_s + \sum_{(s,t)\notin N(x_k)} \theta_{st} x_s x_t\} \right) \end{aligned}$$

We separate out the terms that belong to $N(x_k)$ and ones that don't. We can use this approach for $p(x_k|N(x_k))$ to get:

$$\begin{aligned} p(x_k = 1|N(x_k)) &= \frac{p(N(x_k), x_k = 1)}{p(N(x_k)|x_k = 1) + p(N(x_k)|x_k = 0)} \\ &= \left(1 + \frac{p(N(x_k)|x_k = 0)}{p(N(x_k), x_k = 1)}\right)^{-1} \\ &= \left(1 + \frac{1}{exp\{\theta_k x_k + \sum_{t\in N(x_k)} \theta_{kt} x_k x_t\}}\right)^{-1} \\ &= \left(1 + exp\{\theta_k x_k + \sum_{t\in N(x_k)} \theta_{kt} x_k x_t\}\right)^{-1} \end{aligned}$$

See part (b) for plots of the mean estimations and comparisons with the mean-field approach.

(b) Derive the naive mean-field updates and implement them. Compute the average difference in absolute value between the mean field moments and the Gibb's estimates.



Left column: estimates using Gibbs sampling (top) and mean-field estimates (bottom). Right column: standard deviation of Gibb's estimates for 10 samples (top); absolute difference between Gibbs and mean-field samples (bottom). Average absolute difference is reported as 0.64.