# Inference and Representation, Fall 2016

## Problem Set 5: EM & Factor Analysis

**Due: Monday, October 24, 2016 at 3pm (as a PDF document uploaded in Gradescope.)**

**Important:** *See problem set policy on the course web site.*

---

1. *Non-negative Matrix Factorization (NMF)* [Lee and Seung'99] is an alternative to PCA when data and factors can be cast as non-negative. We seek to factorize the $N \times p$ data matrix $\mathbf{X}$ as

$$\mathbf{X} \approx \mathbf{W}\ \mathbf{H}\ , \tag{1}$$

where $\mathbf{W}$ is $N \times r$ and $\mathbf{H}$ is $r \times p$, with $r \leq \max(N, p)$, and we assume that $x_{ij}, w_{ik}, h_{kj} \geq 0$.

(a) Suppose that $x_{ij} \in \mathbb{N}$. If we model each random variable $x_{ij}$ as a Poisson random variable with mean $(WH)_{ij}$, show that the log-likelihood of the model is (up to a constant)

$$\mathcal{L}(\mathbf{W}, \mathbf{H}) = \sum_{i,j} [x_{ij} \log((WH)_{ij}) - (WH)_{ij}]\ . \tag{2}$$

The following alternating algorithm (Lee, Seung, '01) converges to a local maximum of $\mathcal{L}(\mathbf{W}, \mathbf{H})$:

$$\begin{aligned} w_{ik} &\leftarrow w_{ik} \frac{\sum_j h_{kj} x_{ij}/(WH)_{ij}}{\sum_j h_{kj}}\ , & (3) \\ h_{kj} &\leftarrow h_{kj} \frac{\sum_i w_{ik} x_{ij}/(WH)_{ij}}{\sum_j w_{kj}}\ ,. & (4) \end{aligned}$$

We shall study this algorithm and prove its correctness.

A function $g(x, y)$ is said to minorize a function $f(x)$ if

$$\forall\ (x, y)\ ,\ g(x, y) \leq f(x)\ ,\ g(x, x) = f(x)\ .$$

(a) Show that under the update

$$x^{t+1} = \arg\max_x g(x, x^t)$$

the sequence $f_t = f(x^t)$ is non-decreasing.

(b) Using concavity of the logarithm, show that for any set of $r$ values $y_k \geq 0$ and $0 \leq c_k \leq 1$ with $\sum_{k \leq r} c_k = 1$,

$$\log\left(\sum_{k \leq r} y_k\right) \geq \sum_{k \leq r} c_k \log(y_k / c_k)\ .$$

(c) Deduce that

$$\log\left(\sum_{k\leq r} w_{ik}h_{kj}\right) \geq \sum_{k\leq r} c_{kij}\log(w_{ik}h_{kj}/c_{kij}) \ ,$$

where $c_{kij} = \frac{w_{ik}^t h_{kj}^t}{\sum_{k'\leq r} w_{ik'}^t h_{k'j}^t}$ and $t$ is the current iteration.

(d) Ignoring constants, show that

$$g(\mathbf{W},\mathbf{H};\mathbf{W}^t,\mathbf{H}^t) = \sum_{i,j,k}[x_{ij}c_{kij}(\log w_{ik} + \log h_{kj}) - w_{ik}h_{kj}]$$

minorizes $\mathcal{L}(\mathbf{W},\mathbf{H})$.

(e) Finally, derive the update steps (3) by setting to zero the partial derivatives of $g$.

2. *Factor Analysis, Covariance and Correlation.* Recall that the covariance and correlation of two random variables $X_i$, $X_j$ defined respectively as

$$\sigma_{i,j} = \mathbb{E}((X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)) \ , \ \tilde{\sigma}_{i,j} = \frac{\sigma_{i,j}}{\sqrt{\sigma_{i,i}\sigma_{j,j}}} \ .$$

(a) Show that $-1 \leq \tilde{\sigma}_{ij} \leq 1$.

(b) Show the relationship between Factor Analysis applied to the covariance matrix $\Sigma$ of $X$ and the corresponding Factor Analysis applied to the correlation matrix $\tilde{\Sigma}$ of $X$. Do you obtain the same relationship than Principal Component Analysis?

(c) Construct an example (A) with three random variables exhibiting some correlation, such that the leading principal component fails to detect that correlation but the leading factor analysis direction does, and an example (B) where the detected principal component aligns better with the underlying factor than the leading factor analysis direction.

3. A common modification of the hidden Markov model involves using mixture models for the emission probabilities $p(\mathbf{y}_t|q_t)$, where $q_t$ refers to the state for time $t$ and $\mathbf{y}_t$ to the observation for time $t$.

Suppose that $\mathbf{y}_t \in \mathbb{R}^n$ and that the emission distribution is given by a mixture of Gaussians for each value of the state. To be concrete, suppose that the $q_t$ can take $K$ discrete states and each mixture has $M$ components. Then,

$$p(\mathbf{y}_t \mid q_t) = \sum_{j=1}^{M} b_{q_t j}\left(\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_{q_t j}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mu_{q_t j})^T\Sigma_{q_t j}^{-1}(\mathbf{y}_t - \mu_{q_t j})\right\}\right)$$

where $\mathbf{b}_i \in [0,1]^M$ denotes the mixing weights for state $i$ ($\sum_{j=1}^{M} b_{ij} = 1$ for $i = 1,\ldots K$), $\mu_{ij} \in \mathbb{R}^n$ and $\Sigma_{ij} \in \mathbb{R}^{n\times n}$.

Let $\pi \in \mathbb{R}^K$ be the probability distribution for the initial state $q_0$, and $A \in \mathbb{R}^{K\times K}$ be the transition matrix of the $q_t$'s. In this problem you will derive an EM algorithm for learning the parameters $\{b_{ij}, \mu_{ij}, \Sigma_{ij}\}$ and $A, \pi$.

(a) The EM algorithm is substantially simpler if you introduce auxiliary variables $z_t \in \{1,\ldots,M\}$ denoting which mixture component the $t$'th observation is drawn from. Draw the graphical model for this modified HMM, identifying clearly the additional latent variables that are needed.

(b) Write the expected complete log likelihood for the model and identify the expectations that you need to compute in the E step. *Show all steps of your derivation.*

(c) Give an algorithm for computing the E step.

*Hint: Reduce the inference problem to something you know how to do, such as sum-product belief propagation in tree-structured pairwise MRFs.*

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