CSCI 1950-F Homework 9:
EM for Factor Analysis & Regression

Brown University, Spring 2012

Homework due at 12:00pm on May 3, 2012

Question 1:

The MovieLens dataset (http://movielens.org) contains ratings for M movies, recorded as integers between 1 and 5, for a community of U users. Most users have only rated a small subset of the total movie library, so the data is stored as a sparse $M \times U$ matrix X, where M is the number of movies and U is the number of users. For this assignment we have extracted a small subset of the overall database, containing $M = 500$ movie titles and $U = 943$ users. The rows of the training and test data represent the same users, but with the ratings split between the two matrices.

Your job is to predict user ratings for the entries in the test rating matrix given the training rating matrix. To accomplish this you'll first find a low-dimensional representation of the ratings by learning a factor analysis model. Letting $x_i \in \mathbb{R}^M$ denote the ratings for user i, and x_{ij} the rating that user i gives to movie j, we have:

$$\begin{aligned} z_i &\sim N(0, I_K), \\ x_i &\sim N(Wz_i + \mu, \Psi). \end{aligned}$$

Here, $z_i \in \mathbb{R}^K$ is the latent vector specifying the low-dimensional representation of movie rating vector x_i, where $K \ll M$. The movie ratings are related to this low-dimensional space by the factor loading matrix $W \in \mathbb{R}^{M \times K}$. The mean rating of each movie is encoded in $\mu \in \mathbb{R}^M$, and $\Psi \in \mathbb{R}^{M \times M}$ is an unknown *diagonal* noise covariance matrix. To simplify the later derivations, we can rewrite the model in terms of the following modified variables:

$$\begin{aligned} \widetilde{z}_i &= \begin{bmatrix} z_i \\ 1 \end{bmatrix} \in \mathbb{R}^{K+1} \\ \widetilde{W} &= \begin{bmatrix} W & \mu \end{bmatrix} \in \mathbb{R}^{M \times K+1} \end{aligned}$$

The movie rating likelihood can then be equivalently expressed as

$$x_i \sim N(\widetilde{W}\widetilde{z}_i, \Psi)$$

We have used an extra constant dimension in the latent space to encode the mean μ.

To learn the factor analysis model we will use the Expectation Maximization (EM) algorithm. Let x_i^o denote the sparse subset of movie ratings for user i that are available for training. We can encode the rating pattern via a set of binary variables:

$$r_{ij} = \begin{cases} 1 & \text{if } x_{ij} \in x_i^o \\ 0 & \text{otherwise} \end{cases}$$

Because the covariance matrix Ψ is diagonal in the factor analysis model, the likelihood distribution factorizes. Letting $\widetilde{W}_j^T$ denote row j of $\widetilde{W}$, we have

$$p(x_i \mid z_i, W, \mu, \Psi) = N(x_i \mid \widetilde{W}\widetilde{z}_i, \Psi) = \prod_{j=1}^{M} N(x_{ij} \mid \widetilde{W}_j^T \widetilde{z}_i, \Psi_{jj})$$

The log-likelihood of the observed ratings for user i, given all the parameters and their low-dimensional coordinates z_i, is then

$$\log p(x_i^o \mid z_i, \widetilde{W}, \Psi) = C' - \frac{1}{2} \sum_{j=1}^{M} r_{ij} \left(\log \Psi_{jj} + \Psi_{jj}^{-1} (x_{ij} - \widetilde{W}_j^T \widetilde{z}_i)^2 \right)$$

where C' is some log-normalization constant, independent of the parameters. The complete data log-likelihood, including the observed ratings x_i^o and low-dimensional coordinates z_i for all U users, and using C to collect log-normalization constants, is then

$$\log p(x^o, z \mid \widetilde{W}, \Psi) = C - \frac{1}{2} \sum_{i=1}^{U} \left[\widetilde{z}_i^T \widetilde{z}_i + \sum_{j=1}^{M} r_{ij} \left(\log \Psi_{jj} + \Psi_{jj}^{-1} (x_{ij} - \widetilde{W}_j^T \widetilde{z}_i)^2 \right) \right]$$

The following questions walk you through the steps of deriving the EM algorithm. You might find it helpful to refer to the textbook sections on EM for dense factor analysis.

a) *Derive the expected value of the complete data log likelihood with respect to some distribution $q(\widetilde{z})$. Simplify the form of your answer as much as possible. The final result should depend on only two sufficient statistics of $\widetilde{z}_i$, $\mathbb{E}[\widetilde{z}_i]$ and $\mathbb{E}[\widetilde{z}_i \widetilde{z}_i^T]$.*

b) *Determine the M-step update of the error variance Ψ_{jj} for movie j. Calculate the partial derivative of the expected complete data log-likelihood with respect to Ψ_{jj}, set to zero, and solve for the optimal estimate.*

c) *Determine the M-step update of the factor loading weights $\widetilde{W}_j$ for movie j. Calculate the gradient of the expected complete data log-likelihood with respect to $\widetilde{W}_j$, set to zero, and solve for the optimal estimate.* Hint: *The form of your solution should be similar to the least squares estimates in a linear regression model.*

d) *Determine explicit formulas for $\mathbb{E}[\widetilde{z}_i]$ and $\mathbb{E}[\widetilde{z}_i \widetilde{z}_i^T]$, the E-step expectations needed to compute the M-step parameter updates. To perform these calculations, it can be helpful to revert to the original factor analysis formulation using z_i, μ, and W. First, note that*

$$\mathbb{E}[\widetilde{z}_i] = \begin{bmatrix} \mathbb{E}[z_i] \\ 1 \end{bmatrix}, \qquad \mathbb{E}[\widetilde{z}_i \widetilde{z}_i^T] = \begin{bmatrix} \mathbb{E}[z_i z_i^T] & \mathbb{E}[z_i] \\ \mathbb{E}[z_i]^T & 1 \end{bmatrix},$$

and the joint distribution of z_i and x_i^o is a multivariate normal:

$$\begin{bmatrix} z_i \\ x_i^o \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \mu_i^o \end{bmatrix}, \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \right)$$

Here, μ_i^o contains the entries of μ corresponding to the observed ratings x_i^o. If there are M_i observations for user i, the submatrices $A \in \mathbb{R}^{K \times K}$, $B \in \mathbb{R}^{K \times M_i}$, and $C \in \mathbb{R}^{M_i \times M_i}$. Determine formulas for A, B, and C using the basic definition of covariance in terms of expected values. Then, solve for $\mathbb{E}[z_i]$ and $\mathbb{E}[z_i z_i^T]$ using standard identities for conditional distributions of multivariate normals (see textbook).

Question 2:

Because it can be time-consuming to correctly implement the EM algorithm from Question 1, we have provided an implementation for you. You are encouraged to review the code and make connections between the update equations and Matlab code.

In this question, we evaluate the empirical performance of several methods for predicting movie ratings. Note that the training and test sets correspond to the same set of users, but in the training set only some ratings are observed. When computing test root mean square error (RMSE) in the questions below, be sure to include only the ratings that were not observed in the training set. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{1}{N_h} \sum_{i=1}^{U} \sum_{x_{ij} \in x_i^h} (x_{ij} - \hat{x}_{ij})^2}$$

where x_i^h is the set of test ratings for user i, N_h is the total number of ratings in the test dataset, and $\hat{x}_{ij}$ is the rating predicted by the model under evaluation.

a) *We first consider a very simple baseline. For each movie in the corpus, compute the average of the observed, training ratings. Then for each test item, we simply predict the mean rating of the corresponding movie. Calculate the test RMSE for this method.*

b) *Next, we consider a simple heuristic method for dimensionality reduction with sparse data. First, fill in the missing entries of the training movie rating matrix using the mean predictions from part (a). Apply principal component analysis (PCA) to this matrix using Matlab's* `princomp` *function, and consider the top $K = \{1, 2, \ldots, 15\}$ principal components. Use these low-dimensional representations to reconstruct the missing ratings, and plot RMSE versus K.*

c) *For the heuristic dimensionality reduction method of part (b), what should the performance approach as $K \to M$, the number of movies?*

d) *Run the provided EM factor analysis code to estimate W, μ, and Ψ for $K = \{1, 2, \ldots, 15\}$. For each K, run EM for 100 training iterations, use the equations from Question 1(d) to calculate $\mathbb{E}[\widetilde{z}_i \mid x_i^o]$ for each user, and then use these low-dimensional coordinates to reconstruct the missing ratings. Plot RMSE versus K, and compare to the PCA method from part (b). What choice of K leads to the best performance?*

Question 3:

We now revisit the Bayesian linear regression model from homework 4. As before, given M basis functions $\phi_j(x)$, $j = 1, \ldots, M$, we model the dependence of response variables y_i on input covariates x_i as follows:

$$p(y_i \mid x_i, w, \beta) = \mathcal{N}(y_i \mid w^T \phi(x_i), \beta^{-1}) \qquad p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I_M)$$

Rather than searching over a discrete grid of potential values for the hyperparameters α and β, we will instead estimate them via the EM algorithm. In the E-step, we compute the expected value of certain statistics of the M-dimensional vector of regression coefficients w. In the M-step, we use these to produce new estimates of α and β. The previously distributed solutions for homework 4 may be useful.

a) *For the normal distributions assumed above, derive the form of the expected complete-data log likelihood,* $\mathbb{E}[\log p(y, w \mid x, \alpha, \beta)]$, *given* N *observations* $y = (y_1, y_2, \ldots, y_N)$ *of inputs* $x = (x_1, x_2, \ldots, x_N)$. *It may be helpful to examine the EM derivation for the factor analysis model. What particular statistics do we need to determine the expectations of in the E-step, in order to concretely evaluate this expression?*

b) *Take the derivative of the expression in part (a) with respect to* α, *set it to zero, and determine the M-step update of* α.

c) *Take the derivative of the expression in part (a) with respect to* β, *set it to zero, and determine the M-step update of* β.

d) *Using the equations from the preceding parts, implement the EM algorithm for this model. Consider two different families of basis functions, the polynomial and radial basis functions from homework 4, both with order* $L = 50$ *(*$M = L + 1$, *including the bias feature* $\phi_0(x_i) = 1$*). For each family, initialize* $\alpha^{(0)} = 0.01$, $\beta^{(0)} = 0.0025$, *and run EM until changes in the likelihood fall below* 10^{-6}. *Plot the resulting sequences of parameters* $\alpha^{(t)}$ *and* $\beta^{(t)}$, *as well as the corresponding likelihoods* $p(y \mid \alpha^{(t)}, \beta^{(t)})$ *(see the formula from homework 4, problem 3(a)). What is the test accuracy of the resulting models, using the squared-error loss from homework 4?*