Inference and Representation, Fall 2017

Problem Set 4: PCA & Factor Analysis

Due: Wednesday, October 25, 2017 at 3pm (as a PDF and .zip files uploaded in Gradescope.)

Important: *See problem set policy on the course web site.* Your submission should consist of 2 files: a PDF containing your solutions, and a .zip file containing your source code.

---

1. *Non-negative Matrix Factorization (NMF).* [2, 3] 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 $\mathcal{L}(\mathbf{W}, \mathbf{H})$:

$$w_{ik} \leftarrow w_{ik} \frac{\sum_j h_{kj} x_{ij}/(WH)_{ij}}{\sum_j h_{kj}}\ , \tag{3}$$

$$h_{kj} \leftarrow h_{kj} \frac{\sum_i w_{ik} x_{ij}/(WH)_{ij}}{\sum_j w_{kj}}\ ,. \tag{4}$$

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

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

where $c_{kij} = \frac{w^t_{ik}h^t_{kj}}{\sum_{k'\le r} w^t_{ik'}h^t_{k'j}}$ 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. *NMF vs PCA on images.* The following questions refer to the code and data `hw4.zip`. The MNIST dataset (in `mnist_all.mat`) contains images of handwritten digits labeled with their associated numeric value. The file called `nmf.ipynb` has code performing the following tasks: (a) plot the singular vectors corresponding to the top 10 singular values of the data, and (b) project the training data and the test data (obtained using `load_test_data` in `mnist_tools.py`) onto the first $k = 8$ principal components and run nearest neighbors for each test image in this lower dimensional space. (The PCA code for this problem was adapted from Homework 1 in [1] .)

   (a) Now apply the NMF algorithm with $r \in \{3, 6, 10\}$ and plot the rows of H (using `plot_image_grid` in `plot_tools.py`)

   (b) Project the training data and the test data (obtained using `load_test_data` in `mnist_tools.py`) onto the r rows, and run nearest neighbors for each test image in this lower dimensional space. Include your choice for r, and the plots of your nearest neighbor results in your submitted homework document.

   (c) Comment on the differences between the PCA and NMF (To understand this better, plot the coefficients of a random image for each digit in terms of the first 10 principal components, on one plot and in terms of NMF for $r = 10$ on another plot).

   Include a .zip file with your online homework submission containing all of your source code.

   Keep in mind that the data points in the training and test data are given as rows.

3. *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 $-1 \le \tilde{\sigma}_{ij} \le 1$.

   (b) Show how Factor Analysis applied to the covariance matrix Σ of X and the corresponding Factor Analysis applied to the correlation matrix $\tilde{\Sigma}$ of X are related. Interpret this result. Does the same phenomena hold if you apply PCA to Σ and $\tilde{\Sigma}$?

   (c) Construct an example 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 recover it. (*Hint:* Construct data such that one variable is scaled differently from the rest, and use the previous result.)

(d) Construct an example where the factor analysis method fails to reveal the true correlation structure of the data, but PCA is robust. (*Hint:* Think what happens when there is weak or absent correlation).

(e) Consider the centered Factor Analysis model $X = AY + \epsilon$, with $X = (X_1, \dots, X_L)$ and $J < L$ uncorrelated factors $Y = (Y_1, \dots, Y_J)$, where $\mathbb{E}(\epsilon_i) = 0$, $\mathrm{var}(\epsilon_i) = \beta_i$. Write down the joint data likelihood of the model when both Y and ϵ are jointly Gaussian, and specify the loss that dictates how to obtain the parameters of the model ($A \in \mathbb{R}^{N \times L}$ and $\beta \in \mathbb{R}^L$) via MLE.

(f) Are the parameters of the model uniquely specified? Justify your answer.

(g) If one supposes that $\beta_i = \beta_0$ for $i = 1 \dots L$, give an algorithm to estimate the MLE parameters in that case.

4. *Factor Analysis vs PCA on personality data.* The following questions also refer to the code and data `hw4.zip`. The complete the function `FactorAnalysis` in the file `fa.ipynb` to perform Factor Analysis of the psychological data. Include the plots generated by `FactorAnalysis.ipynb` in your submission. Briefly describe the differences between the PCA and Factor Analysis results.

References

[1] NYU Center for Data Science. Ds-ga 1013 course, optimization-based data analysis. *Available online at* *http://www.cims.nyu.edu/~cfgranda/pages/OBDA_fall17/index.html*, Fall 2017.

[2] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[3] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT Press, 2000.