Assignment 2

Special Course on Unsuprevised Machine Learning: Probabilistic Factor Analysis

May 22, 2017

Practicalities

The assignment contains three problems, and will be marked for a total of 7 points. Submit your assignment in the form of a pdf document by 8:00 am on Monday 29th May 2017 to suleiman.khan@helsinki.fi and joseph.sakaya@helsinki.fi.

Problem 1 (3 points)

Read the following papers, provide a brief summary of each and contrast BPCA and BCCA in the context of the two articles.

- Bayesian PCA `https://papers.nips.cc/paper/1549-bayesian-pca.pdf`
- Bayesian CCA `http://www.jmlr.org/papers/volume14/klami13a/klami13a.pdf`

Problem 2 (2 points)

Data set

You are given a data set containing two paired matrices $\mathbf{X}1 \in \mathcal{R}^{N \times D_1}$ and $\mathbf{X}2 \in \mathcal{R}^{N \times D_2}$, where $N = 78$, $D_1 = 1106$ and $D_2 = 1106$.

Description: The data represents drugs responses of two cancer patients. One of them is blood cancer (also denoted as GeneExpression.HL60) and the other is breast cancer (denoted as GeneExpression.MCF7). For each patient you are provided gene expression measurements of a set of genes ($D_1, D_2 = 1106$) over a set of drugs ($N = 78$). Therefore, each cell in the matrix indicates if the gene (column) is up-regulated (positive) or down-regulated (negative) as a result of the drug treatment (row). It is hypothesised that such data can be used to understand the mechanisms (genes) through which drugs affect the patients.

Access: R users can access the data set as:

```
> load(url("https://www.cs.helsinki.fi/u/sakaya/tutorial/data/UML.RData"))
> X1 <- GeneExpression.HL60 #Blood Cancer
> X2 <- GeneExpression.MCF7 #Breast Cancer
> dim(X1) #verify that the dimensions of the data are 78 x 1106
> dim(X2) #verify that the dimensions of the data are 78 x 1106
> X1[1:3,1:5] #examine a few values
```

Python and other users can access it at: `https://www.cs.helsinki.fi/u/sakaya/tutorial/data/UML.tar.gz`.

Task

Use Bayesian Canonical Correlation Analysis (BCCA) to identify responses of drugs that are shared between the two cancers as well as those specfic to any one. In BCCA, this corresponds to investigating the dependencies between the two matrices.

Your solution should consist of the following:

- To interpret the results, plot the group-sparsity paramters (Hint: $\boldsymbol{\alpha}$) and comment on the components (shared and specific). Recall, a component is active when the $\boldsymbol{\alpha}$ goes low.
- Choose and plot a component k that captures dependencies between the two matrices.
  - Collect the row indices of the 10 highest and lowest scores of the k^{th} component of $\mathbf{Z}$, and
  - Collect the column indices of the 20 highest scores of the k^{th} component of $\mathbf{W}_{1:D_1}$.
  - Subset the original data matrix **X**1 , using the row and column indices, and plot the heat map of the resulting sub-matrix (in R you can use heatmap.2).
  - Collect the column indices of the 20 highest scores of the k^{th} component of $\mathbf{W}_{(1:D_2)+D_1}$.
  - Subset the original data matrix **X**2 , using the row and column indices, and plot the heat map of the resulting sub-matrix (in R you can use heatmap.2).
- Comment on the dependencies identified by BCCA; submit the plots as well.
- You can use the BCCA and starter codes from exercise 2 and 3 for running the model and plotting functions.

Problem 3 (2 points)

CP and Tucker-3 are two widely studied tensor factorization approaches. The distributional assumptions of Bayesian CP factorization of a tensor $\mathcal{X} \in \mathcal{R}^{N \times D \times L}$, were formulated in the course and can be represented as:

$$\begin{aligned}
x_{n,l,d} &\sim \mathcal{N}(\sum_{k=1}^{K} \mathbf{z}_k \circ \mathbf{w}_k \circ \mathbf{u}_k, \tau^{-1}) \\
z_{n,k} &\sim \mathcal{N}(0,1) \\
u_{l,k} &\sim \mathcal{N}(0,1) \\
w_{d,k} &\sim h_k \ \mathcal{N}(0, (\alpha_{d,k})^{-1}) + (1-h_k)\delta_0 \\
h_k &\sim Bernoulli(\pi_k) \\
\pi_k &\sim Beta(a^{\pi}, b^{\pi}) \\
\alpha_{d,k} &\sim Gamma(a^{\alpha}, b^{\alpha}) \\
\tau &\sim Gamma(a^{\tau}, b^{\tau})
\end{aligned}$$

where $\mathbf{Z} \in \mathcal{R}^{N \times K}$, $\mathbf{W} \in \mathcal{R}^{D \times K}$, $\mathbf{U} \in \mathcal{R}^{L \times K}$, and $\circ$ indicates the outer products.

The Tucker-3 factorization of a mode-3 tensor $\mathcal{X} \in \mathcal{R}^{N \times D \times L}$ models the complex interactions of factors in each mode through a core tensor $\mathcal{G} \in \mathcal{R}^{K1 \times K2 \times K3}$ and is defined as

$$\mathcal{X} \sim \sum_{p=1}^{K_1} \sum_{q=1}^{K_2} \sum_{r=1}^{K_3} g_{p,q,r} \mathbf{z}_p \circ \mathbf{w}_q \circ \mathbf{u}_r$$

where $\mathbf{Z} \in \mathcal{R}^{N \times K1}$, $\mathbf{W} \in \mathcal{R}^{D \times K2}$, $\mathbf{U} \in \mathcal{R}^{L \times K3}$, and $\circ$ indicates outer products. While the Tucker-3 is a more flexible model, its solutions suffer from the rotational ambiguity problem due to the full core tensor $\mathbf{G}$.

Task: Use your skills from the course to write the distributional assumptions of a Bayesian formulation for Tucker-3 factorization. Your model should be able to identify rotationally unique solutions as well automatically identify the number of components. Note: the number of components can be different in each mode.

Submit the distributional assumptions of your model along with justifications of your choices. You are *not* required to implement the model in STAN.

Additional Support: To learn more about CP and Tucker-3 factorizations see the lecture material and recommended additional reading in the material section of course webpage.