Advanced Topics in Machine Learning (600.692)
Homework 2: Principal Component Analysis

Instructor: René Vidal

Due Date: 02/28/2014, 11.59PM Eastern

READING MATERIAL: Chapter 2 and Appendix B.4 of GPCA book.

1. **Statistical PCA for Non-Zero Mean Random Variables.** Let $\boldsymbol{x} \in \mathbb{R}^D$ be a random vector. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x}) \in \mathbb{R}^D$ and $\Sigma_{\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \in \mathbb{R}^{D \times D}$ be, respectively, the mean and the covariance of $\boldsymbol{x}$. Define the principal components of $\boldsymbol{x}$ as the random variables $y_i = \boldsymbol{u}_i^\top \boldsymbol{x} + a_i \in \mathbb{R}$, $i = 1, \ldots, d \leq D$, where $\boldsymbol{u}_i \in \mathbb{R}^D$ is a unit norm vector, $a_i \in \mathbb{R}$, and $\{y_i\}_{i=1}^n$ are zero mean, uncorrelated random variables whose variances are such that $\text{Var}(y_1) \geq \text{Var}(y_2) \geq \cdots \geq \text{Var}(y_d)$. Assuming that the eigenvalues of $\Sigma_{\boldsymbol{x}}$ are different from each other, show that
   (a) $a_i = -\boldsymbol{u}_i^\top \boldsymbol{\mu}_{\boldsymbol{x}}$, $i = 1, \ldots, d$.
   (b) $\boldsymbol{u}_1$ is the eigenvector of $\Sigma_{\boldsymbol{x}}$ corresponding to its largest eigenvalue.
   (c) $\boldsymbol{u}_2^\top \boldsymbol{u}_1 = 0$ and $\boldsymbol{u}_2$ is the eigenvector of Σ corresponding to its second largest eigenvalue.
   (d) $\boldsymbol{u}_i^\top \boldsymbol{u}_j = 0$ for all $i \neq j$ and $\boldsymbol{u}_i$ is the eigenvector of $\Sigma_{\boldsymbol{x}}$ corresponding to its i-th largest eigenvalue.

2. **Properties of PCA.** Let $\boldsymbol{x} \in \mathbb{R}^D$ be a random vector with covariance matrix $\Sigma_{\boldsymbol{x}} \in \mathbb{R}^{D \times D}$. Consider a linear transformation of $\boldsymbol{x}$:
$$\boldsymbol{y} = W^\top \boldsymbol{x}, \tag{1}$$
   where $\boldsymbol{y} \in \mathbb{R}^d$ and $W \in \mathbb{R}^{D \times d}$ has orthonormal columns. Let $\Sigma_{\boldsymbol{y}} = W^\top \Sigma_{\boldsymbol{x}} W$ be the covariance matrix for $\boldsymbol{y}$. Show that
   (a) The trace of $\Sigma_{\boldsymbol{y}}$ is maximized by $W = U_d$, where U_d consists of the first d unit eigenvectors of $\Sigma_{\boldsymbol{x}}$.
   (b) The trace of $\Sigma_{\boldsymbol{y}}$ is minimized by $W = \tilde{U}_d$, where $\tilde{U}_d$ consists of the last d unit eigenvectors of $\Sigma_{\boldsymbol{x}}$.

3. **Subspace Angles.** Given two d-dimensional subspaces S_1 and S_2 in $\mathbb{R}^D$, define the largest subspace angle θ_1 between S_1 and S_2 to be the largest possible sharp angle ($< 90°$) formed by any two vectors $\boldsymbol{u}_1, \boldsymbol{u}_2 \in (S_1 \cap S_2)^\perp$ with $\boldsymbol{u}_1 \in S_1$ and $\boldsymbol{u}_2 \in S_2$ respectively. Let $U_1 \in \mathbb{R}^{D \times d}$ be an orthogonal matrix whose columns form a basis for S_1 and similarly U_2 for S_2. Show that if σ_1 is the smallest non-zero singular value of the matrix $W = U_1^\top U_2$, then we have
$$\cos(\theta_1) = \sigma_1. \tag{2}$$
   Similarly, one can define the rest of the subspace angles as $\cos(\theta_i) = \sigma_i$, $i = 2, \ldots, d$ from the rest of the singular values of W.

   Hint: Following the derivation of statistical PCA, find first the smallest angle (largest cosine = largest variance) and then find the second smallest angle all the way to the largest angle (smallest variance). As your proceed, the vectors that achieve the second smallest angle need to be chosen to be perpendicular to the vectors that achieve the smallest angle and so forth, as we did in statistical PCA. Also, let $\boldsymbol{u}_1 = U_1 \boldsymbol{c}_1$ and $\boldsymbol{u}_2 = U_2 \boldsymbol{c}_2$. Show that you need to optimize $\cos(\theta) = \boldsymbol{c}_1^\top U_1^\top U_2 \boldsymbol{c}_2$ subject to $\|\boldsymbol{c}_1\| = \|\boldsymbol{c}_2\| = 1$. Show (using Lagrange multipliers) that a necessary condition for optimality is
$$\begin{bmatrix} 0 & U_1^\top U_2 \\ U_2^\top U_1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix} = \lambda \begin{bmatrix} \boldsymbol{c}_1 \\ \boldsymbol{c}_2 \end{bmatrix}. \tag{3}$$
   Deduce from here that $\sigma = \lambda^2$ is a singular value of $U_1^\top U_2$ with $\boldsymbol{c}_2$ as singular vector.

4. **Ranking of Webpages.** PCA is actually used to rank webpages on the Internet by many popular search engines. One way to see this is to view the Internet as a directed graph $G = (V, E)$, where every webpage, denoted as p_i, is a node in V, and every hyperlink from p_i to p_j, denoted as e_{ij}, is a directed edge in E. We can assign each webpage p_i an "authority" score x_i and a "hub" score y_i. The "authority" score x_i is a scaled sum of the "hub" scores of other webpages pointing to webpage p_i. The "hub" score is the scaled sum of the "authority" scores of other webpages that webpage p_i is pointing out to. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be the vector of authority scores and hub scores, respectively. Also, let A be the adjacent matrix of the graph G, i.e., $A_{ij} = 1$ if $e_{ij} \in E$ and $A_{ij} = 0$ otherwise and consider the following algorithm:

Algorithm 1 (Ranking webpages)

Choose a random vector $\boldsymbol{x}$, and repeat the following two steps

(a) $\boldsymbol{y}' \leftarrow A\boldsymbol{x}$, $\boldsymbol{y} \leftarrow \frac{\boldsymbol{y}'}{\|\boldsymbol{y}'\|}$

(b) $\boldsymbol{x}' \leftarrow A^\top \boldsymbol{y}$, $\boldsymbol{x} \leftarrow \frac{\boldsymbol{x}'}{\|\boldsymbol{x}'\|}$

Answer the following questions.

(a) Given the definitions of hubs and authorities, justify the algorithm.

(b) Show that unit-norm eigenvectors of $AA^\top$ (for $\boldsymbol{y}$) and $A^\top A$ (for $\boldsymbol{x}$) give fixed points of the algorithm.

(c) Show that, in general, $\boldsymbol{y}$ and $\boldsymbol{x}$ converge to the unit-norm eigenvectors associated with the maximum eigenvalue of $AA^\top$ and $A^\top A$, respectively. Explain why not any other eigenvector and why the normalization steps in the algorithm are necessary.

(d) Explain how $\boldsymbol{y}$ and $\boldsymbol{x}$ can be computed from the singular value decomposition of A. Under what circumstances would the given algorithm be preferable to using the SVD?

In the literature, this is known as the *Hypertext Induced Topic Selection* (HITS) algorithm. The same algorithm can also be used to rank any competitive sports such as football teams and chess players.

5. **PPCA by Maximum Likelihood.** Study the proof of Theorem 2.8 in great detail and show the missing piece that is left as an exercise to the reader. More specifically, let $\lambda_1, \ldots, \lambda_D$ be the eigenvalues of a covariance matrix $\Sigma \in \mathbb{R}^{D \times D}$. Let $\pi : \{1, \ldots, D\} \to \{1, \ldots, D\}$ be a permutation of the first D integers. We would like to choose d eigenvalues $\lambda_{\pi[1]}, \ldots, \lambda_{\pi[d]}$ such that the discarded ones $\lambda_{\pi[d+1]}, \ldots, \lambda_{\pi[D]}$ minimize

$$\mathcal{M}(\pi) = \log\left(\frac{\sum_{i=d+1}^{D} \lambda_{\pi[i]}}{D - d}\right) - \frac{\sum_{i=d+1}^{D} \log \lambda_{\pi[i]}}{D - d}. \tag{4}$$

Use Jensen's inequality to show that $\mathcal{M}$ is nonnegative and the concavity of the $\log$ function to prove that $\mathcal{M}$ is minimized by choosing $\lambda_{\pi[i]}$, $i = d + 1, \ldots, D$ to be contiguous in magnitude.

Submission instructions. Please follow the same instructions as in HW1.