HW 1: Advanced Topics in Machine Learning

Instructor: René Vidal, E-mail: rvidal@cis.jhu.edu

Due 2/17/10 in class

1. If $S \in \mathbb{R}^{n \times n}$ is a real symmetric matrix then prove that:
   (a) All eigenvalues of S Are real, i.e., $\sigma(S) \subset \mathbb{R}$.
   (b) Let (λ, v) be an eigenvalue-eigenvector pair. If $\lambda_i \neq \lambda_j$, then $v_i \perp v_j$; i.e., eigenvectors corresponding to distinct eigenvalues are orthogonal.
   (c) There always exist n orthonormal eigenvectors of S, which form a basis of $\mathbb{R}^n$.
   (d) $S > 0$ ($S \geq 0$) if $\lambda_i > 0$ ($\lambda_i \geq 0$) $\forall i = 1, 2, \ldots, n$; i.e., S is positive (semi-)definite if all eigenvalues are positive (non-negative).
   (e) If $S \geq 0$ and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, then $\max_{\|x\|_2=1} < x, Sx >= \lambda_1$ and $\min_{\|x\|_2=1} < x, Sx >= \lambda_n$.
2. Prove the following:
   (a) Consider the problem $Ax = b$ with $A \in \mathbb{R}^{m \times n}$ of rank $p \leq r \triangleq \min\{m, n\}$. Prove that the solution x^* that minimizes $\|Ax - b\|_2$ is given by $x^* = A^\dagger b$, where $A^\dagger$ is the generalized (Moore Penrose) inverse.
   (b) Let $A = U \Sigma V^T$ be the SVD of A. Let $B = U \Sigma_p V^T$, where Σ_p denotes the matrix obtained from Σ by setting to zero its elements on the diagonal after the p^{th} entry. Show that $\|A - B\|_F^2 = \sigma_{p+1}^2 + \ldots + \sigma_r^2$, where $\|\cdot\|_F$ indicates the Frobenius norm. Furthermore, show that such a norm is the minimum achievable over all matrices $B \in \mathbb{R}^{m \times n}$ of rank p, i.e.,

$$\min_{B:\text{rank}(B)=p} \|A - B\|_F^2 = \sigma_{p+1}^2 + \ldots + \sigma_r^2.$$

3. Let $\boldsymbol{x}$ be a random vector with covariance matrix $\Sigma_{\boldsymbol{x}}$. Consider a linear transformation of $\boldsymbol{x}$:

$$\boldsymbol{y} = W^\top \boldsymbol{x}, \tag{1}$$

   where $\boldsymbol{y} \in \mathbb{R}^d$ and W is a $D \times d$ orthogonal matrix, i.e., $W^\top W = I_d$. 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 (normalized) 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 (normalized) eigenvectors of $\Sigma_{\boldsymbol{x}}$.
4. 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 $u_1, u_2 \in (S_1 \cap S_2)^\perp$ with $u_1 \in S_1$ and $u_2 \in S_2$ 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. Then 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 $u_1 = U_1 c_1$ and $u_2 = U_2 c_2$. Show that you need to optimize $\cos(\theta) = c_1^\top U_1^\top U_2 c_2$ subject to $\|c_1\| = \|c_2\| = 1$. Show (using Lagrange multipliers) that a necessary condition for optimality is

$$\begin{bmatrix} 0 & U_1^\top U_2 \\ U_2^\top U_1 & 0 \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \lambda \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \tag{3}$$

Deduce from here that $\sigma = \lambda^2$ is a singular value of $U_1^\top U_2$ with c_2 as singular vector.

5. (a) PowerFactorization can be used to compute the principal components of a zero-mean dataset with missing entries. How would you modify the method to deal with the case where the data is not zero mean?

   (b) Implement PowerFactorization for nonzero mean data with missing entries. Use the following format

   Function `[x,U,Y] = pf(X,d,W)`

   Parameters

   `X` $D \times N$ matrix whose columns are the data points
   `d` dimension of low-dimensional representation
   `W` $D \times N$ matrix whose ij entry is equal to 1 if Xij is given and to 0 if Xij is missing

   Returned values

   `x` $D \times 1$ vector containing the mean of the data
   `U` $D \times d$ matrix containing the basis for the subspace
   `Y` $d \times N$ matrix containing the principal components

   (c) Generate and plot a dataset `X` with $D = 3$; $d = 2$; $N = 100$; `x=ones(D,1);` `[Q,R]=qr(randn(D,d));` `U=Q(:,1:d);` `Y=randn(d,N);` `Y=Y-mean(Y,2)*ones(1,N);` `X=x*ones(1,N)+U*diag(3*(0:d-1)+1)*Y;` `plot3(X(1,:),X(2,:),X(3,:),'.')`
   Compute the mean and principal components of the dataset `x=mean(X,2);` `[U,S,V]=svd(X-x*ones(1,N));` `U=U(:,1:d);` `Y=S(1:d,1:d)*V(:,1:d)';`
   Apply `pf` with `W1=ones(D,N);` `[x1,U1,Y1]=pf(X,d,W1)`. Do you get the same results as PCA?
   Apply `pf` with `W2=ones(D,N)-eye(D,N);` `[x2,U2,Y2]=pf(X,d,W2)`. Do you get the same results as PCA? Can you reconstruct the missing entries in X perfectly or approximately?
   Plot the low-dimensional representations `Y, Y1, Y2` and comment on their similarities and differences
   `figure(1), plot(Y(1,:),Y(2,:),'or');figure(2), plot(Y1(1,:),Y1(2,:),'xg');`
   `figure(3), plot(Y2(1,:),Y2(2,:),'sb');axis('equal');`