Exam 1: Unsupervised Learning (600.692)

Instructor: René Vidal

April 3, 2017

1. **Model Selection for PCA.** Assume you are given a matrix $X \in \mathbb{R}^{D \times N}$ whose columns lie approximately in a low-dimensional subspace of *unknown dimension* $d < D$. Let $X = U_X \Sigma_X V_X^\top$ be the SVD of X. You would like to approximate X by a low-rank matrix $A \in \mathbb{R}^{D \times N}$ and you consider the following optimization problem

$$\min_A \quad \|X - A\|_F^2 + \tau \|A\|_*^2, \tag{1}$$

where $\tau > 0$ is a fixed parameter.

(a) **(15 points)** Let $A = U_A \Sigma_A V_A^\top$ be the SVD of A. Show that an optimal solution for U_A and V_A with Σ_A held constant is given by $U_A = U_X$ and $V_A = V_X$.

Solution: If Σ_A is constant, so is $\|A\|_* = \sigma(A)^\top \mathbf{1}$, where $\sigma(A) \in \mathbb{R}^D$ is the vector of singular values of A. Thus the problem reduces to minimizing the first term of the objective with respect to U_A and V_A only, i.e.,

$$\begin{aligned}
\min_{U_A, V_A} \|X - A\|_F^2 &\equiv \min_{U_A, V_A} \|U_X \Sigma_X V_X^\top - U_A \Sigma_A V_A^\top\|_F^2 \\
&\equiv \min_{U_A, V_A} \|\Sigma_X - U_X^\top U_A \Sigma_A V_A^\top V_X\|_F^2 \quad \text{(Frobenius norm is invariant under rotation)} \\
&\equiv \min_{U,V} \|\Sigma_X - U \Sigma_A V^\top\|_F^2 \quad (U_X^\top U_A = U, V_X^\top V_A = V) \\
&\equiv \min_{U,V} \|\Sigma_X\|_F^2 - 2\langle \Sigma_X, U \Sigma_A V^\top \rangle + \|U \Sigma_A V^\top\|_F^2 \\
&\equiv \min_{U,V} \|\Sigma_X\|_F^2 - 2\langle \Sigma_X, U \Sigma_A V^\top \rangle + \|\Sigma_A\|_F^2 \quad (U, V \text{ are orthogonal matrices}) \\
&\equiv \max_{U,V} \langle \Sigma_X, U \Sigma_A V^\top \rangle \quad \text{(omitting constant terms w.r.t. } U \text{ and } V)
\end{aligned}$$

Now, using Von Neumann's theorem we have that $\langle \Sigma_X, U \Sigma_A V^\top \rangle \le \sum_{i=1}^{\min(D,N)} \sigma_i(X) \sigma_i(A)$, where $\sigma_i(X)$ and $\sigma_i(A)$, $i = 1, 2, \ldots$, are the singular values of X and A, respectively, and that equality occurs (i.e., the maximum with respect to U and V is achieved) when $U = I$ and $V = I$, i.e.,

$$U = I, V = I \implies U_X^\top U_A = I, V_X^\top V_A = I \implies U_A = U_X, V_A = V_X.$$

(b) **(35 points)** Let $\bar{\sigma}_d(X)$ be the average of the top d singular values of X, where d is the largest integer such that $\sigma_d(X) > \frac{\tau d}{1 + \tau d} \bar{\sigma}_d(X)$. Show that an optimal solution for A is:

$$A = U_X \mathcal{S}_\mu(\Sigma_X) V_X^\top \qquad \text{where} \qquad \mu = \frac{\tau d}{1 + \tau d} \bar{\sigma}_d(X) \tag{2}$$

and $\mathcal{S}_\mu(Y) = \operatorname{argmin}_A \frac{1}{2}\|Y - A\|_F^2 + \mu \|A\|_1$ is the shrinkage thresholding operator.

Hint: Show that an optimal solution for Σ_A satisfies $(I_d + \tau \mathbf{1}\mathbf{1}^\top)\sigma(A) = \sigma_{1:d}(X)$, where $\sigma(A) \in \mathbb{R}^d$ is the vector of singular values of A (similarly for X). Show also that $(I_d + \tau \mathbf{1}\mathbf{1}^\top)^{-1} = (I_d - \frac{\tau}{1 + \tau d} \mathbf{1}\mathbf{1}^\top)$.

Solution #1: Let us first show that for any d, $(I_d + \tau \mathbf{1}\mathbf{1}^\top)^{-1} = (I_d - \frac{\tau}{1+\tau d}\mathbf{1}\mathbf{1}^\top)$, which is equivalent to showing that $(I_d + \tau \mathbf{1}\mathbf{1}^\top)(I_d - \frac{\tau}{1+\tau D}\mathbf{1}\mathbf{1}^\top) = I$. We have

$$\begin{aligned}
(I + \tau \mathbf{1}\mathbf{1}^\top)(I - \frac{\tau}{1+\tau d}\mathbf{1}\mathbf{1}^\top) &= I - \frac{\tau}{1+\tau d}\mathbf{1}\mathbf{1}^\top + \tau \mathbf{1}\mathbf{1}^\top - \frac{\tau^2}{1+\tau d}\mathbf{1}\mathbf{1}^\top\mathbf{1}\mathbf{1}^\top \\
&= I - \frac{\tau}{1+\tau d}\mathbf{1}\mathbf{1}^\top + \tau \mathbf{1}\mathbf{1}^\top - \frac{\tau^2 d}{1+\tau d}\mathbf{1}\mathbf{1}^\top \\
&= I - (\frac{\tau}{1+\tau d} - \tau + \frac{d\tau^2}{1+\tau d})\mathbf{1}\mathbf{1}^\top \\
&= I \qquad \text{(This shows one of the hints).}
\end{aligned}$$

Now, following the results from part (a) we have that

$$\begin{aligned}
\min_{\Sigma_A \geq 0} \|X - A\|_F^2 + \tau \|A\|_*^2 \equiv \min_{\Sigma_A \geq 0} \|\Sigma_X - \Sigma_A\|_F^2 + \tau \|\Sigma_A\|_*^2 &\equiv \min_{\sigma(A)\geq 0} \|\sigma(X) - \sigma(A)\|_2^2 + \tau \|\sigma(A)\|_1^2 \\
&\equiv \min_{\sigma(A)\geq 0} \sum_i (\sigma_i(X) - \sigma_i(A))^2 + \tau \Big(\sum_i \sigma_i(A)\Big)^2.
\end{aligned}$$

For $k = 1, \ldots, K \doteq \min(D, N)$, let $\lambda_k \geq 0$ be the KKT multiplier for the constraint $\sigma_k(A) \geq 0$. The KKT conditions are given by

$$\forall k = 1, \ldots K, \;\; \lambda_k \geq 0, \sigma_k(A) \geq 0, \lambda_k \sigma_k(A) = 0, 2(\sigma_k(A) - \sigma_k(X)) + 2\tau \sum_{i=1}^K \sigma_i(A) - \lambda_k = 0.$$

Let $d = |\{k : \sigma_k(A) > 0\}|$ be the number of nonzero $\sigma_k(A)$. For $k = 1, \ldots, d$ we have $\lambda_k = 0$, hence

$$\begin{aligned}
\sigma_k(A) + \tau \sum_{i=1}^d \sigma_i(A) = \sigma_k(X) &\implies (I_d + \tau \mathbf{1}\mathbf{1}^\top)\sigma(A) = \sigma_{1:d}(X) \qquad \text{(This shows one of the hints)} \\
&\implies \sigma(A) = (I_d - \frac{\tau}{1+\tau d}\mathbf{1}\mathbf{1}^\top)\sigma_{1:d}(X) \\
&\implies \sigma_k(A) = \sigma_k(X) - \frac{\tau}{1+\tau d}\sum_{i=1}^d \sigma_i(X) \implies \Sigma_A = \mathcal{S}_\mu(\Sigma_X).
\end{aligned}$$

Therefore, d is the largest integer such that $\sigma_k(X) > \frac{\tau}{1+\tau d}\sum_{i=1}^d \sigma_i(X)$ for all $k = 1, \ldots, d$. Since the sequence $\sigma_k(X)$ is non increasing, so is the sequence $\sigma_k(A)$, and hence d is the largest integer such that $\sigma_d(X) > \frac{\tau}{1+\tau d}\sum_{i=1}^d \sigma_i(X)$. Note that such an integer exists since for $d = 1$ we have $\sigma_1(X) > \frac{\tau}{1+\tau}\sigma_1(X)$. On the other hand, for $k = d+1, \ldots, K$ we have that $\lambda_k \geq 0$, because

$$\begin{aligned}
\frac{\lambda_k}{2} = \tau \sum_{i=1}^d \sigma_i(A) - \sigma_k(X) &= \tau \Big(\sum_{i=1}^d \sigma_i(X) - \frac{\tau d}{1+\tau d}\sum_{i=1}^d \sigma_i(X)\Big) - \sigma_k(X) \\
&= \frac{\tau}{1+\tau d}\sum_{i=1}^d \sigma_i(X) - \sigma_k(X) \geq 0.
\end{aligned}$$

Therefore, the optimal solution for the optimization problem is given by

$$A = U_A \Sigma_A V_A^\top = U_X \mathcal{S}_\mu(\Sigma_X) V_X^\top. \tag{3}$$

Solution #2: Recall that if $A = U\Sigma V^\top$ is the rank r compact SVD of A, then $\partial\|A\|_* = UV^\top + W$, where $U \in \mathbb{R}^{D\times r}$, $V \in \mathbb{R}^{N\times r}$ and $W \in \mathbb{R}^{D\times N}$ are such that $U^\top W = 0$, $WV = 0$ and $\|W\|_2 \leq 1$. Let us decompose the SVD of X as $U_X = [U_1 \; U_2]$, $V_X = [V_1 \; V_2]$ and $\Sigma_X = \mathrm{diag}(\Sigma_1, \Sigma_2)$, where $U_1 \in \mathbb{R}^{D\times r}$, $U_2 \in \mathbb{R}^{D\times(D-r)}$, $\Sigma_1 \in \mathbb{R}^{r\times r}$, $\Sigma_2 \in \mathbb{R}^{(D-r)\times(N-r)}$, $V_1 \in \mathbb{R}^{N\times r}$ and $V_2 \in \mathbb{R}^{N\times(N-r)}$, so that $X = U_1\Sigma_1 V_1^\top + U_2\Sigma_2 V_2^\top$. Since the objective function is strictly convex, the global minimizer is unique and must satisfy

$$A - X + \tau\|A\|_*\partial\|A\|_* \ni 0 \iff A + \tau\|A\|_*\partial\|A\|_* \ni X. \tag{4}$$

To find the global minimizer, we must find (U, Σ, V, W) such that $U^\top W = 0$, $WV = 0$, $\|W\|_2 \le 1$ and

$$\begin{aligned} U\Sigma V^\top + \tau\|\Sigma\|_*(U\Sigma V^\top + W) &= X \\ U(\Sigma + \tau\|\Sigma\|_* I_r)V^\top + \tau\|\Sigma\|_* W &= U_1\Sigma_1 V_1^\top + U_2\Sigma_2 V_2^\top. \end{aligned}$$

A candidate solution is $U = U_1$, $\Sigma + \tau\|\Sigma\|_* I_r = \Sigma_1$, $V = V_1$, and $\tau\|\Sigma\|_* W = U_2\Sigma_2 V_2^\top$, from which we have

$$\|\Sigma\|_* + \tau\|\Sigma\|_* r = \|\Sigma_1\|_* \implies \|\Sigma\|_* = \frac{\|\Sigma_1\|_*}{1+\tau r} \implies \Sigma = \Sigma_1 - \frac{\tau}{1+\tau r}\|\Sigma_1\|_* I_r$$

$$W = \frac{U_2\Sigma_2 V_2^\top}{\tau\|\Sigma\|_*} = \frac{U_2\Sigma_2 V_2^\top}{\tau\|\Sigma_1\|_*}(1+\tau r) \implies \|W\|_2 = \frac{\sigma_{r+1}(X)}{\tau\|\Sigma_1\|_*}(1+\tau r) \le 1 \implies \sigma_{r+1}(X) \le \frac{\tau}{1+\tau r}\|\Sigma_1\|_*.$$

For this solution, the objective function reduces to the following function of r

$$\begin{aligned} \|A - X\|_F^2 + \tau\|A\|_*^2 = \tau^2\|\Sigma\|_*^2(\|\Sigma\|_*^2 + \frac{\|\Sigma_2\|_*^2}{\tau^2\|\Sigma\|_*^2}) + \tau\|\Sigma\|_*^2 &= \tau^2\|\Sigma\|_*^4 + \|\Sigma_2\|_*^2 + \tau\|\Sigma\|_*^2 \\ &= \frac{\tau^2}{(1+\tau r)^2}\|\Sigma_1\|_*^4 + \frac{\tau}{1+\tau r}\|\Sigma_1\|_*^2 + \|\Sigma_2\|_*^2, \end{aligned}$$

which is minimized by the largest possible integer d such that $\sigma_d(X) > \frac{\tau}{1+\tau d}\|\Sigma_1\|_* = \frac{\tau}{1+\tau d}\sum_{i=1}^d \sigma_i(X) \doteq \mu$. Such an integer exists, since for $d = 1$ we have $\sigma_1(X) > \frac{\tau}{1+\tau}\sigma_1(X)$.
Therefore, the optimal solution for the optimization problem is given by

$$A = U\Sigma V^\top = U_1(\Sigma_1 - \mu I_d)V_1^\top = U_X \mathcal{S}_\mu(\Sigma_X)V_X^\top. \tag{5}$$

(c) **(5 points)** What is the estimate of the subspace dimension given by the above model selection approach? What is the estimate when $\tau \to \infty$?

Solution: The estimate of the subspace dimension is the number of singular values of X that are bigger than or equal to $\mu = \frac{\tau d}{1+\tau d}\bar{\sigma}_d(X)$. When $\tau \to \infty$, $d = 0$ because the condition $\sigma_1(X) > \frac{\tau}{1+\tau}\sigma_1(X)$ is no longer valid.

(d) **(5 points)** Discuss the advantages of this model selection approach versus the one discussed in class

$$\min_A \quad \frac{1}{2}\|X - A\|_F^2 + \tau\|A\|_*, \tag{6}$$

where an optimal solution is given by $A = U_X \mathcal{S}_\tau(\Sigma_X)V_X^\top$. **Hint:** consider the case $\tau \to \infty$.

Solution: The approach discussed in class uses a fixed threshold to estimate the dimension, which is independent on scalings of the data. The approach proposed here has the advantage that it uses a data-dependent threshold that is adapted to scalings of the data, which may facilitate tuning the τ parameter.

2. **PCA with missing entries and outliers.** Let $L_0 \in \mathbb{R}^{D\times N}$ be a low-rank matrix, i.e., $\operatorname{rank}(L_0) \ll \min\{D, N\}$. Let $E_0 \in \mathbb{R}^{D\times N}$ be a sparse matrix, i.e., its number of nonzero entries is $\|E_0\|_0 \ll ND$. Suppose you are given a subset of the entries of $X = L_0 + E_0 \in \mathbb{R}^{D\times N}$ indexed by a set $\Omega \subseteq \{1, \ldots, D\} \times \{1, \ldots, N\}$. To recover L_0 and E_0 you consider the optimization problem, where $\tau > 0$ and $\lambda > 0$ are fixed parameters:

$$\min_{L,E} \quad \frac{1}{2}\|L\|_F^2 + \tau\|L\|_* + \frac{\lambda}{2}\|E\|_F^2 + \lambda\tau\|E\|_1 \quad \text{such that} \quad P_\Omega(X) = P_\Omega(L + E). \tag{7}$$

(a) **(30 points)** Write down the Lagrangian for this problem and use it to give a detailed derivation of the following (dual ascent) algorithm for solving the above problem

$$L^{k+1} = \mathcal{D}_\tau(Z^k) \tag{8}$$
$$E^{k+1} = \mathcal{S}_\tau(Z^k/\lambda) \tag{9}$$
$$Z^{k+1} = Z^k + \delta_k P_\Omega(X - L^{k+1} - E^{k+1}), \tag{10}$$

where $\mathcal{S}_\tau(Y) = \operatorname{argmin}_A \frac{1}{2}\|Y - A\|_F^2 + \tau\|A\|_1$ is the shrinkage thresholding operator, $\mathcal{D}_\tau(Y) = \operatorname{argmin}_A \frac{1}{2}\|Y - A\|_F^2 + \tau\|A\|_*$ is the singular value thresholding operator, $Z \in \mathbb{R}^{D\times N}$ is the matrix of Lagrange multipliers initialized as $Z^0 = \mathbf{0}$, and $\delta_k > 0$ is a sequence of real numbers.

Solution: The Lagrangian is formulated as

$$\mathcal{L}(L, E, Z) = \quad \frac{1}{2}\|L\|_F^2 + \tau\|L\|_* + \frac{\lambda}{2}\|E\|_F^2 + \lambda\tau\|E\|_1 + \langle Z, P_\Omega(X) - P_\Omega(L + E)\rangle.$$

Solving for L with E and Z fixed we obtain

$$\begin{aligned}
\min_L \mathcal{L}(L, E, Z) &\equiv \min_L(\frac{1}{2}\|L\|_F^2 + \tau\|L\|_* - \langle Z, P_\Omega(L)\rangle) \\
&\equiv \min_L(\frac{1}{2}\|L\|_F^2 + \tau\|L\|_* - \langle P_\Omega(Z), L\rangle) \\
&\equiv \min_L(\frac{1}{2}\|L\|_F^2 + \tau\|L\|_* - \langle P_\Omega(Z), L\rangle) \\
&\equiv \min_L(\frac{1}{2}\|P_\Omega(Z) - L\|_F^2 + \tau\|L\|_* - \|P_\Omega(Z)\|_F^2) \\
&\equiv \min_L(\frac{1}{2}\|P_\Omega(Z) - L\|_F^2 + \tau\|L\|_*)
\end{aligned}$$

The solution to the above problem is given by $L^{k+1} = \min_L \mathcal{L}(L, E^k, Z^k) = \mathcal{D}_\tau(P_\Omega(Z^k)) = \mathcal{D}_\tau(Z^k)$ (update only observed values)

Solving for E with L and Z fixed we obtain

$$\begin{aligned}
\min_E \mathcal{L}(L, E, Z) &\implies \min_E(\frac{\lambda}{2}\|E\|_F^2 + \tau\lambda\|E\|_1 - \langle Z, P_\Omega(E)\rangle) \\
&\implies \min_E(\frac{\lambda}{2}\|E\|_F^2 + \tau\lambda\|E\|_1 - \langle P_\Omega(Z), E\rangle) \\
&\implies \min_E(\frac{1}{2}\|E\|_F^2 + \tau\|E\|_1 - \langle \frac{1}{\lambda}P_\Omega(Z), E\rangle) \\
&\implies \min_E(\frac{1}{2}\|\frac{1}{\lambda}P_\Omega(Z) - E\|_F^2 + \tau\|E\|_1 - \|P_\Omega(Z)\|_F^2) \\
&\implies \min_E(\frac{1}{2}\|\frac{1}{\lambda}P_\Omega(Z) - E\|_F^2 + \tau\|E\|_1)
\end{aligned}$$

The solution to the above problem is given by $E^{k+1} = \min_E \mathcal{L}(L^k, E, Z^k) = \mathcal{S}_\tau(\frac{1}{\lambda}P_\Omega(Z^k)) = \mathcal{S}_\tau(\frac{1}{\lambda}Z^k)$ (update only observed values)

The update of Z is given by gradient descent

$$\begin{aligned}
Z^{k+1} &= Z^k + \delta_k \frac{\partial \mathcal{L}(L^{k+1}, E^{k+1}, Z)}{\partial Z}|_{Z=Z^k} \\
&= Z^k + \delta_k P_\Omega(X - L^{k+1} - E^{k+1})
\end{aligned}$$

(b) **(10 points)** What parameter should be increased to make L low rank and make E sparse? Can you guess sufficient conditions on L_0 and E_0 under which $L^* = L_0$ and $E^* = E_0$ with overwhelming probability.

Solution: To make L low rank we should increase τ. To make E sparse we should increase $\lambda\tau$. The conditions for correct recovery should require L_0 to be τ incoherent w.r.t the set of sparse matrices and E_o should be sufficiently sparse.