EECS 127/227AT **Fall 2020**

Solutions for Discussion 3

1 Weighted least squares

You need funding for your startup, for which you do some research to understand how startup funding works. You have past funding data from K different startup incubators. Each of the incubators bases its response to any startup proposal that comes to it on the same n features. The data you have from startup i is thus a matrix $A_i \in \mathbb{R}^{m_i \times n}$ and a vector $b_i \in \mathbb{R}^{m_i}$, where m_i is the number of proposals made to incubator i for which you have data and, for each $1 \leq j \leq m_i$, the j-th row of A_i gives the features for the j-th proposal made to that incubator and the j-th coordinate of b_i gives the corresponding response from that incubator.

You aim to come up with an assessment of the best possible weight to assign to each feature to explain all this data, assuming that at each incubator the response to each proposal it gets is a weighted linear combination of the values of the features of that proposal. With this in mind, you come up with a least-squares methodology for penalizing the residuals of the fit as $\|A_i x - b_i\|_2^2$ at incubator i, but since you have more faith in the correctness of the data available to you from some of the incubators than others, you incorporate this prior knowledge by forming an overall weighted least squares objective given by

$$f(x) = \lambda_1 \|A_1 x - b_1\|_2^2 + \lambda_2 \|A_2 x - b_2\|_2^2 + \cdots + \lambda_K \|A_K x - b_K\|_2^2,$$

where $\lambda_i > 0$ for $i \in \{1, 2, \cdots, K\}$.

(a) Show that the overall objective function $f(x)$ can be expressed as a new least square objective $\|Ax - b\|_2^2$, where you will express A and b in terms of λ_i, A_i, b_i, $i \in \{1, 2, \cdots, K\}$.

Solution:

The weighted least squares objective can be written as,

$$f(x) = \|\sqrt{\lambda_1} A_1 x - \sqrt{\lambda_1} b_1\|_2^2 + \cdots + \|\sqrt{\lambda_K} A_K x - \sqrt{\lambda_K} b_K\|_2^2.$$

The individual norms can be stacked together to have,

$$f(x) = \left\| \begin{bmatrix} \sqrt{\lambda_1} A_1 x - \sqrt{\lambda_1} b_1 \\ \sqrt{\lambda_2} A_2 x - \sqrt{\lambda_2} b_2 \\ . \\ . \\ \sqrt{\lambda_K} A_K x - \sqrt{\lambda_K} b_K \end{bmatrix} \right\|_2^2 = \left\| \begin{bmatrix} \sqrt{\lambda_1} A_1 \\ \sqrt{\lambda_2} A_2 \\ . \\ . \\ \sqrt{\lambda_K} A_K \end{bmatrix} x - \begin{bmatrix} \sqrt{\lambda_1} b_1 \\ \sqrt{\lambda_2} b_2 \\ . \\ . \\ \sqrt{\lambda_K} b_K \end{bmatrix} \right\|_2^2 .$$

Comparing this with $\|Ax - b\|_2^2$ gives

$$A = \begin{bmatrix} \sqrt{\lambda_1} A_1 \\ \sqrt{\lambda_2} A_2 \\ . \\ . \\ \sqrt{\lambda_K} A_K \end{bmatrix},$$

$$b = \begin{bmatrix} \sqrt{\lambda_1} b_1 \\ \sqrt{\lambda_2} b_2 \\ . \\ . \\ \sqrt{\lambda_K} b_K \end{bmatrix}.$$

(b) What is a criterion for existence of an unique solution for $\min_x f(x)$, assuming a solution exists, i.e. assuming that $b \in \mathcal{R}(A)$? Express your answer in terms of rows/columns of A as applicable.
Solution:
The criterion is that A should have linearly independent columns.

(c) Show that any optimal solution, call it x^*, to $\min_x f(x)$ satisfies the equation

$$(\lambda_1 A_1^\top A_1 + \lambda_2 A_2^\top A_2 + \cdots + \lambda_K A_K^\top A_K)x^* = \lambda_1 A_1^\top b_1 + \lambda_2 A_2^\top b_2 + \cdots + \lambda_K A_K^\top b_K.$$

Solution:
We have the optimization problem in the form

$$\min_x \quad f(x) = \min_x \|Ax - b\|_2^2.$$

Taking the first derivative and setting it to 0 gives the requirement

$$\begin{aligned} \frac{d}{dx}(x^\top A^\top - b^\top)(Ax - b)|_{x=x^*} &= 0 \\ \Rightarrow 2A^\top Ax^* - 2A^\top b &= 0 \\ \Rightarrow A^\top Ax^* &= A^\top b. \end{aligned}$$

Plugging in the values of A and b as found from part (a), we have the requirement

$$(\lambda_1 A_1^\top A_1 + \lambda_2 A_2^\top A_2 + \cdots + \lambda_K A_K^\top A_K)x^* = \lambda_1 A_1^\top b_1 + \lambda_2 A_2^\top b_2 + \cdots + \lambda_K A_K^\top b_K.$$

2 Moore-Penrose pseudoinverse, least squares and the SVD

Let $A \in \mathbb{R}^{m\times n}$ with rank r. We will assume that $r \geq 1$, i.e. A is not the zero matrix. We know that $r \leq \min(m, n)$. As you have seen in the lectures, A has a singular value decomposition (SVD) of the form

$$A = U\tilde{\Sigma}V^T = \begin{bmatrix} U_{\mathcal{R}} & U_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \Sigma & 0^{r\times(n-r)} \\ 0^{(m-r)\times r} & 0^{(m-r)\times(n-r)} \end{bmatrix} \begin{bmatrix} V_{\mathcal{R}}^T \\ V_{\mathcal{N}}^T \end{bmatrix},$$

where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, i.e. $U^TU = UU^T = I_m$ and $V^TV = VV^T = I_n$, and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix with the r singular values of A (which are all strictly positive) along the main diagonal in decreasing order. Further we have $U_{\mathcal{R}} \in \mathbb{R}^{m\times r}$, with its columns giving an orthonormal basis for $\mathcal{R}(A)$; $U_{\mathcal{N}} \in \mathbb{R}^{m\times(m-r)}$, with its columns giving an orthonormal basis for $\mathcal{N}(A^T)$; $V_{\mathcal{R}} \in \mathbb{R}^{n\times r}$, with its columns giving an orthonormal basis for $\mathcal{R}(A^T)$; and $V_{\mathcal{N}} \in \mathbb{R}^{n\times(n-r)}$, with its columns giving an orthonormal basis for $\mathcal{N}(A)$. (If either $r = m$ or $r = n$ or both, the corresponding matrices $U_{\mathcal{N}}$ and/or $V_{\mathcal{N}}$ will not need to be defined.)

We also know that the representation

$$A = U_{\mathcal{R}}\Sigma V_{\mathcal{R}}^T,$$

which follows from the above, is called a compact form SVD of A.

In the lectures we defined

$$A^\dagger := V_{\mathcal{R}}\Sigma^{-1}U_{\mathcal{R}}^T,$$

which is called the *Moore-Penrose pseudoinverse* of A.

In this problem we will explore some connections between the Moore-Penrose pseudoinverse, the least squares problem and the SVD.

(a) Suppose $A = a \in \mathbb{R}^{m\times 1}$ is a nonzero column vector. What is $A^\dagger$ (i.e. $a^\dagger$)?

Solution:
Here the rank of a, viewed as a matrix, is $r = 1$. The compact form SVD of the column vector $a \in \mathbb{R}^{m\times 1}$ has $U_{\mathcal{R}} = \frac{1}{\|a\|_2}a$, $\Sigma = \|a\|_2$ and $V_{\mathcal{R}} = 1$. Note that Σ is a number (i.e. a 1×1 matrix), and since $r = n = 1$, the matrix $V_{\mathcal{N}}$ is not defined. If we were interested in the full form SVD of a, we could choose $U_{\mathcal{N}}$ to be any matrix in $\mathcal{R}^{(m-1)\times m}$ whose columns form an orthonormal basis for the orthogonal complement of the 1-dimensional subspace of $\mathbb{R}^m$ spanned by a (which the same as the null space of a^T).

The Moore-Penrose inverse of a is then

$$a^\dagger = V_{\mathcal{R}}\Sigma^{-1}U_{\mathcal{R}}^T = \frac{1}{\|a\|_2^2}a^T.$$

(b) Suppose $A = a^T \in \mathbb{R}^{1\times n}$ is a nonzero row vector. What is $A^\dagger$ (i.e. $a^{T\dagger}$)?

Solution:
The rank of a^T, viewed as a matrix, is 1. The compact form SVD of a^T has $U_{\mathcal{R}} = 1$, $\Sigma = \|a\|_2$ and $V_{\mathcal{R}} = \frac{1}{\|a\|_2}a$. Hence

$$a^{T\dagger} = V_{\mathcal{R}}\Sigma^{-1}U_{\mathcal{R}}^T = \frac{1}{\|a\|_2^2}a.$$

(c) Now consider the linear system $Ax = b$, to be solved for $x \in \mathbb{R}^n$ given $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. In general this equation may not have a solution and even if it does, it may not have a unique solution. In the least squares methodology we attempt to find the "best possible approximate solution", using as a metric to measure how "good" a solution is the sum of the squares of the residual errors, i.e. $\|Ax - b\|_2^2$. (Here $b - Ax$ is called the *residual* and x is called the vector of *regression coefficients*.) If there are many solutions even with this approximation criterion in place (as would happen only if $\mathcal{N}(A)$ is non-trivial) we may further try to focus attention on the one for which $\|x\|_2$ is the smallest.

Based on the SVD representation of A, which we assume to have rank $r \geq 1$ as in the earlier parts of this problem, simplify $\|Ax - b\|_2^2$ so that it is in terms of Σ, $V_\mathcal{R}$, $U_\mathcal{R}$, and $U_\mathcal{N}$.

Solution:
Since any orthornormal matrix preserves the norm of any vector it acts on, we have

$$\begin{aligned}
\|Ax - b\|_2^2 &= \|U^T(Ax - b)\|_2^2 \\
&= \|U^T(U\tilde{\Sigma}V^T x - b)\|_2^2 \\
&= \|\tilde{\Sigma}V^T x - U^T b\|_2^2 \\
&= \left\| \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_\mathcal{R}^T \\ V_\mathcal{N}^T \end{bmatrix} x - \begin{bmatrix} U_\mathcal{R}^T \\ U_\mathcal{N}^T \end{bmatrix} b \right\|_2^2 \\
&= \left\| \begin{bmatrix} \Sigma V_\mathcal{R}^T x \\ 0 \end{bmatrix} - \begin{bmatrix} U_\mathcal{R}^T b \\ U_\mathcal{N}^T b \end{bmatrix} \right\|_2^2 \\
&= \left\|\Sigma V_\mathcal{R}^T x - U_\mathcal{R}^T b\right\|_2^2 + \left\|U_\mathcal{N}^T b\right\|_2^2
\end{aligned}$$

Notice that we have separated the contributions to the overall squared residual error $\|Ax - b\|_2^2$ between two components. One is the contribution in the range space of A and the other is the contribution from the null space of A. We have no control over $\|U_\mathcal{N}^T b\|_2^2$, so let's focus on the other term, which we have control over since it includes the variable x.

(d) Solve for an optimal x^* that minimizes $\|Ax - b\|_2^2$.
Solution:
As mentioned, to minimize the sum $\left\|\Sigma V_\mathcal{R}^T x - U_\mathcal{R}^T b\right\|_2^2 + \left\|U_\mathcal{N}^T b\right\|_2^2$ we can only manipulate $\left\|\Sigma V_\mathcal{R}^T x - U_\mathcal{R}^T b\right\|_2^2$ by the choice of x. What we would ideally like is for this term to be 0. So we ask ourselves if we can find x^* solving the equation

$$\Sigma V_\mathcal{R}^T x^* = U_\mathcal{R}^T b.$$

We see that there is a solution, given by

$$x^* := V_\mathcal{R} \Sigma^{-1} U_\mathcal{R}^T b,$$

because

$$\Sigma V_\mathcal{R}^T V_\mathcal{R} \Sigma^{-1} U_\mathcal{R}^T b = \Sigma \Sigma^{-1} U_\mathcal{R}^T b = U_\mathcal{R}^T b.$$

Notice that

$$x^* = A^\dagger b.$$

Of course, in case $\mathcal{N}(A)$ is nonempty any choice of x of the form $A^\dagger b + z$, where $z \in \mathcal{N}(A)$, will also be optimal.

Since $A^\dagger b$ lies in $\mathcal{R}(A^T)$ because the columns of $V_\mathcal{R}$ lie in $\mathcal{R}(A^T)$ (in fact they form an orthonormal basis for $\mathcal{R}(A^T)$), we see that if we impose the additional requirement that our solution must have minimum ℓ_2 norm, then $A^\dagger b$ is the unique optimal solution. This is because we have

$$\|A^\dagger b + z\|_2^2 = \|A^\dagger b\|_2^2 + \|z\|_2^2$$

for any $z \in \mathcal{N}(A)$, by virtue of the fact that $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ are orthogonal subspaces (in fact they are orthogonal complements of each other).

(e) What is the error at the optimal x^*?

Solution:

The error at the optimal x^* is given by $\|U_\mathcal{N}^T b\|_2^2$

(f) Assume now that $r = n$. Show that the residual error $b - Ax$ at any optimal choice of regression coefficients x (which we know must be of the form $x^* + z$ for some $z \in \mathcal{N}(A)$) satisfies the equation

$$\begin{bmatrix} I_m & A \\ A^T & 0^{n\times n} \end{bmatrix} \begin{bmatrix} b - Ax \\ x \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

Solution:

We know that $x = x^* + z$ for some $z \in \mathcal{N}(A)$, because we are given that x is an optimal vector of regression coefficients. We can write

$$\begin{bmatrix} I_m & A \\ A^T & 0^{n\times n} \end{bmatrix} \begin{bmatrix} b - A(x^* + z) \\ x^* + z \end{bmatrix} = \begin{bmatrix} b - A(x^* + z) + A(x^* + z) \\ A^T b - A^T A(x^* + z) \end{bmatrix} \overset{(a)}{=} \begin{bmatrix} b \\ A^T b - A^T A A^\dagger b \end{bmatrix} \overset{(b)}{=} \begin{bmatrix} b \\ 0 \end{bmatrix}.$$

Here, in step (a), we have used $Az = 0$ (because $z \in \mathcal{N}(A)$) and substituted $x^* = A^\dagger b$. In step (b) we have used $AA^\dagger = U_\mathcal{R} U_\mathcal{R}^T$, which holds because $r = n$, and we have observed that

$$b - U_\mathcal{R} U_\mathcal{R}^T b \in \mathcal{N}(A^T),$$

which can be proved by showing that it is in the orthogonal complement of $\mathcal{R}(A)$, which, in turn, follows from

$$U_\mathcal{R}^T (b - U_\mathcal{R} U_\mathcal{R}^T b) = 0,$$

because the columns of $U_\mathcal{R}$ form a basis for $\mathcal{R}(A)$ (in fact they form an orthonormal basis for $\mathcal{R}(A)$).

3 PCA and low-rank compression

We are given an $m \times n$ matrix $X = [x_1, \ldots, x_n]$, with $x_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$ being the data points. (Thus, each feature corresponds to a row of X.) We assume that the data matrix is centered, in the sense that $x_1 + \ldots + x_n = 0$.

$C := \frac{1}{n} X X^\top$ is called the covariance matrix of the data, since the empirical variance of the data points along a direction $u \in \mathbb{R}^m$ is $u^\top C u$.

Consider the following three problems:

(P_1) Find a line going through the origin that maximizes the empirical variance of the collection of data points projected on the line. More explicitly, P_1 is the problem

$$\max_{u \in \mathbb{R}^m : u^\top u = 1} \quad u^\top C u.$$

(P_2) Find a line going through the origin that minimizes the sum of squares of the distances from the points to their projections on this line. More explicitly, P_2 is the problem

$$\min_{u \in \mathbb{R}^m : u^\top u = 1} \sum_{i=1}^{n} \min_{\alpha_i \in \mathbb{R}} \|x_i - \alpha_i u\|_2.$$

(P_3) Find the best approximation of rank at most 1 to the data matrix in Frobenius norm. More explicitly, P_3 is the problem

$$\min_{X' : \text{rank}(X') \leq 1} \|X - X'\|_F.$$

In this exercise, you are asked to show the equivalence between these three problems.

(a) Given $x_0, u \in \mathbb{R}^m$ such that $u^\top u = 1$, consider the problem of projecting a point $x \in \mathbb{R}^m$ on the line $\mathcal{L} := \{x_0 + \alpha u \ : \ \alpha \in \mathbb{R}\}$.

Show that the projected point z^* is given by

$$z^* = x_0 + \alpha^* u,$$

where we define

$$\alpha^* := (x - x_0)^\top u.$$

Further, show that the squared distance $\|z^* - x\|_2^2$ is equal to $\|x - x_0\|_2^2 - \alpha^{*2}$. (By the definition of the projection, this is the minimal squared distance from x to any point on the line $\mathcal{L}$.)

Solution:

Finding the projection of point x on line $\mathcal{L}$ corresponds to solving the following problem:

$$\alpha^* = \min_{\alpha \in \mathbb{R}} \|x_0 + \alpha u - x\|_2.$$

Since $u^T u = 1$, the squared objective can be written as

$$\|x_0 + \alpha u - x\|_2^2 = \alpha^2 - 2\alpha(x - x_0)^\top u + \|x - x_0\|_2^2.$$

We see that we need to find the minimum in a single-variable quadratic equation in α, and we find the optimizer to be

$$\alpha^* = (x - x_0)^\top u.$$

At the optimum value α^*, the squared objective function, which equals the minimum squared distance $\|z - x\|_2^2$ over all $z \in \mathcal{L}$, takes the claimed value:

$$\|x_0 + \alpha^* u - x\|_2^2 = \|x - x_0\|_2^2 - ((x - x_0)^\top u)^2.$$

(b) Show that problems P_1, P_2 are equivalent.

Solution:

P_2 minimizes the sum of squared distances from the points to their projections on a line passing through the origin $\mathcal{L} = \{\alpha u : \alpha \in \mathbb{R}\}$. This problem can be written as:

$$\min_{u \,:\, u^\top u = 1} \sum_{i=1}^n \min_{\alpha_i} \|x_i - \alpha_i u\|_2^2, \tag{1}$$

where $\alpha_i^* = x_i^\top u$, as we found in the previous part (note that $x_0 = 0$ because $\mathcal{L}$ passes through the origin). From the previous part, we obtain the equivalent form:

$$\min_{u \,:\, u^\top u = 1} \sum_{i=1}^n \|x_i\|_2^2 - (x_i^\top u)^2.$$

Since the x_i are fixed, solving this problem is equivalent to solving the problem

$$\min_{u \,:\, u^\top u = 1} \sum_{i=1}^n -(u^\top x_i)(x_i^\top u),$$

which is the same as the problem

$$\max_{u \,:\, u^\top u = 1} \sum_{i=1}^n u^\top x_i x_i^\top u.$$

This problem can be written as a variance maximization problem:

$$\max_{u \,:\, u^\top u = 1} \; u^\top C u,$$

where $C := XX^\top = (1/n) \sum_{i=1}^n x_i x_i^\top$ is the covariance matrix associated with the centered data. The above is exactly P_1, which maximizes the variance of the set of projections of the data points on a line.

(c) Show that P_3 is equivalent to P_1.

Hint: The data matrix is rank-one if and only if it can be expressed as the outer product of two nonzero vectors.

Solution:

When a data matrix X' has rank at most 1, all the points are on a line going through the origin; this means that there exists $u \in \mathbb{R}^m$, with $u^\top u = 1$, such that, for every i, there exists a scalar α_i such that

$$x'_i = \alpha_i u.$$

Thus,

$$X' = [x'_1, \ldots, x'_n] = [\alpha_1 u, \ldots, \alpha_n u] = u\alpha^\top,$$

with $\alpha^\top = [\alpha_1, \dots, \alpha_n]$. (Note that the case $\alpha = 0$ corresponds to when X' has rank 0, i.e. is the zero matrix.)

Now P_3 is the problem of minimizing the Frobenius norm of the matrix $X - u\alpha^\top$, over u, α, where, by a simple rescaling, we can always impose the condition $u^\top u = 1$. Thus, P_3 can be written as

$$\min_{\alpha,u} \|X - u\alpha^\top\|_F = \min_{\alpha,u:u^\top u=1} \sum_{i=1}^{n} \|x_i - \alpha_i u\|_2^2$$

which is exactly P_2.