Discussion 10B

1. Orthonormality and Least Squares

Recall that, if $U \in \mathbb{R}^{m \times n}$ is a tall matrix (i.e. $m \geq n$) with orthonormal columns, then

$$U^\top U = I_{n \times n} \tag{1}$$

However, it is not necessarily true that $UU^\top = I_{m \times m}$. In this discussion, we will deal with "orthonormal" matrices, where the term "orthonormal" refers to a matrix that is square with orthonormal columns and rows. Furthermore, for an orthonormal matrix U,

$$U^\top U = UU^\top = I_{n \times n} \implies U^{-1} = U^\top \tag{2}$$

This discussion will cover some useful properties that make orthonormal matrices favorable, and we will see a "nice" matrix factorization that leverages orthonormal matrices and helps us speed up least squares.

(a) Suppose you have a real, square, $n \times n$ orthonormal matrix U. You also have real vectors $\vec{x}_1$, $\vec{x}_2$, $\vec{y}_1$, $\vec{y}_2$ such that

$$\vec{y}_1 = U\vec{x}_1 \tag{3}$$

$$\vec{y}_2 = U\vec{x}_2 \tag{4}$$

This is analogous to a change of basis. Show that, in this new basis, the inner products are preserved. **Calculate $\langle \vec{y}_1, \vec{y}_2 \rangle = \vec{y}_2^\top \vec{y}_1 = \vec{y}_1^\top \vec{y}_2$ in terms of $\langle \vec{x}_1, \vec{x}_2 \rangle = \vec{x}_2^\top \vec{x}_1 = \vec{x}_1^\top \vec{x}_2$.**

Solution: Since we have defined the y vectors, we can substitute their expressions into $\vec{y}_2^\top \vec{y}_1$:

$$\begin{aligned}
\langle \vec{y}_1, \vec{y}_2 \rangle &= \vec{y}_2^\top \vec{y}_1 && (5)\\
&= (U\vec{x}_2)^\top U\vec{x}_2 && (6)\\
&= \vec{x}_2^\top \underbrace{U^\top U}_{I_{n\times n}} \vec{x}_1 && (7)\\
&= \vec{x}_2^\top \vec{x}_1 && (8)\\
&= \langle \vec{x}_1, \vec{x}_2 \rangle && (9)
\end{aligned}$$

Note that in going from eq. (7) to eq. (8), we used eq. (2).

(b) Using the change of basis defined in part **1.**a, show that, in the new basis, the norms are preserved. **Express $\|\vec{y}_1\|^2$ and $\|\vec{y}_2\|^2$ in terms of $\|\vec{x}_1\|^2$ and $\|\vec{x}_2\|^2$.**

Solution: Recall that we can write the norm squared as

$$\|\vec{v}\|^2 = \vec{v}^\top \vec{v} = \langle \vec{v}, \vec{v} \rangle \tag{10}$$

We can directly use the method from part **1.**a to show that

$$\|\vec{y}_i\|^2 = \langle \vec{y}_i, \vec{y}_i \rangle \tag{11}$$

$$= \vec{y}_i^\top \vec{y}_i \tag{12}$$
$$= \vec{x}_i^\top U^\top U \vec{x}_i \tag{13}$$
$$= \vec{x}_i^\top \vec{x}_i \tag{14}$$
$$= \|\vec{x}_i\|^2 \tag{15}$$

for $i \in \{1, 2\}$.

(c) Suppose you observe data coming from the model $y_i = \vec{a}^\top \vec{x}_i$, and you want to find the linear scale-parameters (each a_i). We are trying to learn the model $\vec{a}$. You have m data points $(\vec{x}_i, y_i)$, with each $\vec{x}_i \in \mathbb{R}^n$. Each $\vec{x}_i$ is a different input vector that you take the inner product of with $\vec{a}$, giving a scalar y_i.

Set up a matrix-vector equation of the form $X\vec{a} = \vec{y}$ for some X and $\vec{y}$, and propose a way to estimate $\vec{a}$.

Solution: Since $y = \vec{a}^\top \vec{x}$ means that $y = \vec{x}^\top \vec{a}$, we can stack the equations with the following definitions:

$$X := \begin{bmatrix} \vec{x}_1^\top \\ \vec{x}_2^\top \\ \vdots \\ \vec{x}_m^\top \end{bmatrix} \qquad \vec{y} := \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{16}$$

Then, we have $\vec{y} = X\vec{a}$. Note that $X \in \mathbb{R}^{m \times n}$, and $\vec{y} \in \mathbb{R}^m$. We can estimate $\vec{a}$ using least squares. Applying the standard least squares formula, we can find our estimate $\widehat{\vec{a}}$ by computing

$$\widehat{\vec{a}} = \left(X^\top X\right)^{-1} X^\top \vec{y}. \tag{17}$$

(d) Let's suppose that we can write our X matrix from part **1.c** as

$$X = MV^\top \tag{18}$$

for some matrix $M \in \mathbb{R}^{m \times n}$ and some orthonormal matrix $V \in \mathbb{R}^{n \times n}$. **Find an expression for $\widehat{\vec{a}}$ from the previous part, in terms of M and $V^\top$.**

Note: take this form as a given. We will go over how to find such a V and M later.

Solution: From the previous part, we have

$$\widehat{\vec{a}} = \left(X^\top X\right)^{-1} X^\top \vec{y}. \tag{19}$$

Plugging in $X = MV^\top$, we have

$$\widehat{\vec{a}} = \left(\left(MV^\top\right)^\top \left(MV^\top\right)\right)^{-1} \left(MV^\top\right)^\top \vec{y} \tag{20}$$
$$= \left(VM^\top MV^\top\right)^{-1} VM^\top \vec{y} \tag{21}$$
$$= \left(V^\top\right)^{-1} \left(M^\top M\right)^{-1} (V)^{-1} VM^\top \vec{y} \tag{22}$$
$$= V\left(M^\top M\right)^{-1} M^\top \vec{y} \tag{23}$$

(e) Now suppose that we have the matrix

$$\begin{bmatrix} \vec{x}_1^\top \\ \vec{x}_2^\top \\ \vdots \\ \vec{x}_m^\top \end{bmatrix} := X = U\Sigma V^\top. \tag{24}$$

where $U \in \mathbb{R}^{m\times m}$ is an orthonormal matrix, and $V \in \mathbb{R}^{n\times n}$ is an orthonormal matrix. Here, $\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}$. Here we assume that we have more data points than the dimension of our space (that is, $m > n$). Also, the transformation V in part e) is the same V in this factorized representation.

Set up a least squares formulation for estimating $\vec{a}$ and find the solution to the least squares. Why might this factorization help us compute $\widehat{\vec{a}}$ faster?

Note: again, take this factorization as a given. We will go over how to find U, Σ, and V later.

Solution: From the previous part, we know

$$\widehat{\vec{a}} = V\left(M^\top M\right)^{-1} M^\top \vec{y} \tag{25}$$

Here, $M = U\Sigma$ by pattern matching terms. Plugging this in,

$$\begin{aligned}
\widehat{\vec{a}} &= V\left((U\Sigma)^\top (U\Sigma)\right)^{-1} (U\Sigma)^\top \vec{y} && (26)\\
&= V\left(\Sigma^\top U^\top U\Sigma\right)^{-1} \Sigma^\top U^\top \vec{y} && (27)\\
&= V\left(\Sigma^\top \Sigma\right)^{-1} \Sigma^\top U^\top \vec{y} && (28)\\
&= V\left(\begin{bmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}\right)^{-1} \Sigma^\top U^\top \vec{y} && (29)\\
&= V\left(\begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}\right)^{-1} \Sigma^\top U^\top \vec{y} && (30)
\end{aligned}$$

$$= V \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n^2} \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \dots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n & 0 & \cdots & 0 \end{bmatrix} U^\top \vec{y} \tag{31}$$

$$= V \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sigma_n} & 0 & \cdots & 0 \end{bmatrix} U^\top \vec{y} \tag{32}$$

The nice part about this matrix factorization is that we can compute our least squares estimate really quickly (owing to the diagonal nature of $\Sigma^\top \Sigma$), since inverting an arbitrarily large matrix is computationally expensive. In particular, we only need to take the reciprocal of the diagonal elements of $\Sigma^\top \Sigma$ when computing the matrix inverse. Multiplying this with $\Sigma^\top$ adds the extra $\vec{0}$ columns.

Contributors:

- Neelesh Ramachandran.
- Kuan-Yun Lee.
- Anant Sahai.
- Kumar Krishna Agrawal.