Discussion 10A

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

$$U^\top U = I_{n\times n} \quad (1)$$

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,

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

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

$$\vec{y}_1 = U\vec{x}_1 \quad (3)$$
$$\vec{y}_2 = U\vec{x}_2 \quad (4)$$

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 $\langle \vec{x}_1, \vec{x}_2\rangle = \vec{x}_2^\top \vec{x}_1 = \vec{x}_1^\top \vec{x}_2$.**

want to show $\langle y_1, y_2\rangle = \langle x_1, x_2\rangle$

$\langle y_1, y_2\rangle = \langle Ux_1, Ux_2\rangle$
$= (Ux_1)^T (Ux_2)$
$= (x_1^T U^T)(Ux_2)$
$= x_1^T U^T U x_2$
$= x_1^T I x_2$
$= x_1^T x_2$
$= \langle x_1, x_2\rangle$

$(Ux_1)^T = x_1^T U^T$

(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 $\|\vec{x}_2\|^2$.**

$\langle x, x\rangle = \|x\|^2$

$\|y_1\|^2 = \|x_1\|^2$
$\|y_2\|^2 = \|x_2\|^2$

$\|y_1\|^2 = \|Ux_1\|^2$
$= \langle Ux_1, Ux_1\rangle$
$= (x_1^T U^T)(Ux_1)$ $\quad U^T U = I$
$= x_1^T x_1$
$= \langle x_1, x_1\rangle$
$= \|x_1\|^2$

$\|y_1\|^2 = \langle y_1, y_1\rangle$
$= \langle x_1, x_1\rangle$
$= \|x_1\|^2$

rule
algorithm

(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 $\vec{a}$.

$$x_i \longrightarrow y_i$$

$$a^T x_1 \overset{?}{=} x_1^T a$$

$$\underbrace{\begin{bmatrix} a_1 & a_2 & \cdots \end{bmatrix}}_{a} \underbrace{\begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \end{bmatrix}}_{x_1} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots \end{bmatrix}}_{x_1} \underbrace{\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}}_{a}$$

$$Xa = y$$

(x_1, y_1)

$$\left.\begin{aligned} y_1 &= a^T x_1 \\ y_2 &= a^T x_2 \\ &\vdots \\ y_m &= a^T x_m \end{aligned}\right\} \longrightarrow \begin{bmatrix} - & x_1^T & - \\ - & x_2^T & - \\ & \vdots & \\ - & x_m^T & - \end{bmatrix} \begin{bmatrix} | \\ a \\ | \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

m scalar equations

1 vector equation

$$Xa = y$$

$$\hat{a} = (X^TX)^{-1}X^Ty$$

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

$$X = MV^\top \tag{5}$$

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.

$$\hat{a} = (X^TX)^{-1}X^Ty$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

$$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$$

$$= \left((MV^T)^T (MV^T) \right)^{-1} (MV^T)^T y$$

$$= \left(V \underbrace{M^T M} V^T \right)^{-1} (MV^T)^T y$$

$$= (V^T)^{-1} (M^TM)^{-1} V^{-1} \left[(V^T)^T M^T \right] y$$

$$(V^T)^T = V$$

$$(V^{-1})^T = V$$

$$V^{-1} = V^T$$

$$= V(M^TM)^{-1} V^{-1} \left[VM^T \right] y$$

$$\hat{a} = V(M^TM)^{-1}M^T y$$

(e) Now suppose that we have the matrix

$$\begin{bmatrix} \vec{x}_1^\top \\ \vec{x}_2^\top \\ \vdots \\ \vec{x}_m^\top \end{bmatrix} := X = U\Sigma V^\top. \tag{6}$$

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.

Contributors:

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

$$\hat{a} = V(M^T M)^{-1} M^T y \qquad X = M V^T$$

$$X = U \Sigma V^T$$

$$M = U \Sigma$$

$$\hat{a} = V(M^T M)^{-1} M^T y \qquad M^T = \Sigma^T u^T$$

$$= V(\Sigma^T u^T u \Sigma)^{-1} \Sigma^T u^T y$$

$$= V(\Sigma^T \Sigma)^{-1} \Sigma^T u^T y$$

$$\Sigma^T = \begin{bmatrix} \sigma_1 & 0 & \cdots & & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ & \ddots & & & \\ 0 & & \sigma_n & & 0 \end{bmatrix} \qquad \Sigma = \underset{(m \times n)}{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ & 0 & \end{bmatrix}}$$

$$\Sigma^T \Sigma = \underset{(n \times m)}{\begin{bmatrix} \sigma_1 & 0 & \cdots & & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ & \ddots & & & \\ 0 & & \sigma_n & & 0 \end{bmatrix}} \underset{(m \times n)}{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ & 0 & \end{bmatrix}} = \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & 0 \\ 0 & & \ddots \\ & & & \sigma_n^2 \end{bmatrix}$$

$$\Sigma^T\Sigma = \underbrace{\begin{bmatrix} \sigma_1 & 0 & \cdots & & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ & \mathbf{0} & \ddots & & \mathbf{0} \\ & & & \sigma_n & \end{bmatrix}}_{(n\times m)} \underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ & \mathbf{0} & \end{bmatrix}}_{(m\times n)} = \begin{pmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \mathbf{0} \\ \mathbf{0} & & \ddots \\ & & & \sigma_n^2 \end{pmatrix}$$

$$(\Sigma^T\Sigma)^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_n^2} \end{bmatrix}$$

$$\begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$$

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}\begin{pmatrix} 1/2 & 0 \\ 0 & 1/3 \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\hat{a} = V(\Sigma^T\Sigma)^{-1}\Sigma^T U^T y$$

$$\begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_n^2} \end{bmatrix}\begin{bmatrix} \sigma_1 & & 0 & \\ & \ddots & & \mathbf{0} \\ 0 & & \sigma_n & \end{bmatrix}$$

$$= V[\quad]U^T y$$

SUMMER

links.eecs16b.org/nima-dis