Lecture Notes of Multivariate Statistics
Lecture 01

Weizhong Zhang
School of Data Science, Fudan University

February 22, 2023

1 Review of Linear Algebra

Theorem 1.1 (QR Factorization). *Prove the following results for Gram-Schmidt orthogonalization*

1. $r_{jj} \neq 0$ *for all* $i = 1, \ldots, n$
2. $\|\mathbf{q}_i\|_2 = 1$ *for all* $i = 1, \ldots, n$
3. $\mathbf{q}_i^\top \mathbf{q}_j = 0$ *for all* $i = 1, \ldots, n$ *and* $j < i$.

Proof. **Part 1:** Since each $\mathbf{q}_i$ is a linear combination of $\{\mathbf{a}_1, \cdots, \mathbf{a}_i\}$, the entry r_{jj} is zero means

$$r_{jj} = \left\| \mathbf{a}_j - \sum_{i=1}^{j-1} r_{ij} \mathbf{q}_i \right\|_2 = 0,$$

then $\mathbf{a}_j$ must be a linear combination of $\{\mathbf{a}_1, \cdots, \mathbf{a}_{j-1}\}$, which validates the full rank assumption on $\mathbf{A}$.

Part 2: Just use the expression of r_{jj}.

Part 3: Recall that $r_{ij} = \mathbf{q}_i^\top \mathbf{a}_j$ for any $i \neq j$. We can verify

$$\mathbf{q}_1^\top \mathbf{q}_2 = \frac{\mathbf{q}_1^\top (\mathbf{a}_2 - r_{12}\mathbf{q}_1)}{r_{22}} = \frac{\mathbf{q}_1^\top (\mathbf{a}_2 - (\mathbf{q}_1^\top \mathbf{a}_2)\mathbf{q}_1)}{r_{22}} = \frac{\mathbf{q}_1^\top \mathbf{a}_2 - (\mathbf{q}_1^\top \mathbf{a}_2)\mathbf{q}_1^\top \mathbf{q}_1}{r_{22}} = 0$$

Suppose for $\mathbf{q}_i^\top \mathbf{q}_j = 0$ for all $\mathbf{q}_i^\top \mathbf{q}_j = 0$ for all $i = 1, \ldots, n' - 1$ and $j < i$. Then for all $k = 1, 2, \ldots, n' - 1$, we have

$$\mathbf{q}_k^\top \mathbf{q}_{n'} = \frac{\mathbf{q}_k^\top \mathbf{a}_{n'} - \sum_{i=1}^{n'-1} r_{in'} \mathbf{q}_k^\top \mathbf{q}_i}{r_{n'n'}} = \frac{\mathbf{q}_k^\top \mathbf{a}_{n'} - r_{kn'} \mathbf{q}_k^\top \mathbf{q}_k}{r_{n'n'}} = \frac{\mathbf{q}_k^\top \mathbf{a}_{n'} - r_{kn'}}{r_{n'n'}} = 0$$

Then we prove the result by induction. □

Theorem 1.2. *Prove* $\|\mathbf{A}\|_2 = \sigma_1$.

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be full SVD of $\mathbf{A}$. Then

$$\|\mathbf{A}\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \left\|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \mathbf{x}\right\|_2 = \sup_{\|\mathbf{x}\|_2 = 1} \left\|\mathbf{\Sigma}\mathbf{V}^\top \mathbf{x}\right\|_2$$

Then let $\mathbf{y} = \mathbf{V}^\top \mathbf{x}$. Since $\mathbf{V}$ is orthogonal matrix, we have $\|\mathbf{y}\|_2 = \left\|\mathbf{V}^\top \mathbf{x}\right\|_2 = \|\mathbf{x}\|_2 = 1$. Hence,

$$\sup_{\|\mathbf{x}\|_2 = 1} \left\|\mathbf{\Sigma}\mathbf{V}^\top \mathbf{x}\right\|_2 = \sup_{\|\mathbf{y}\|_2 = 1} \|\mathbf{\Sigma}\mathbf{y}\|_2 = \sup_{\|\mathbf{y}\|_2 = 1} \sqrt{\sum_{i=1}^{r} (\sigma_i y_i)^2} \leq \sigma_1.$$

We attain the maximum by taking $\mathbf{y} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ and the corresponding $\mathbf{x}$ is $\mathbf{V} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ $\square$

Theorem 1.3 (Cholesky Factorization)**.** *The symmetric positive-definite matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ has the decomposition of the form*

$$\mathbf{A} = \mathbf{L}\mathbf{L}^\top$$

where $\mathbf{L} \in \mathbb{R}^{\times n}$ is a lower triangular matrix with real and positive diagonal entries.

Proof. For $n = 1$, it is trivial. Suppose it holds for $n-1$, then any $\widetilde{\mathbf{A}} \in \mathbb{R}^{(n-1)\times(n-1)}$ can be written as

$$\widetilde{\mathbf{A}} = \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top$$

where $\widetilde{\mathbf{L}} \in \mathbb{R}^{(n-1)\times(n-1)}$ is a lower triangular matrix with real and positive diagonal entries. Consider the case of n such that

$$\mathbf{A} = \begin{bmatrix} \widetilde{\mathbf{A}} & \mathbf{a} \\ \mathbf{a}^\top & \alpha \end{bmatrix} = \begin{bmatrix} \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top & \mathbf{a} \\ \mathbf{a}^\top & \alpha \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad \text{where } \mathbf{a} \in \mathbb{R}^{n-1}, \quad \alpha \in \mathbb{R}.$$

Let

$$\mathbf{L}_1 = \begin{bmatrix} \widetilde{\mathbf{L}}^{-1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

We have

$$\mathbf{L}_1^{-1}\mathbf{A}\mathbf{L}_1^{-\top} = \begin{bmatrix} \widetilde{\mathbf{L}}^{-1} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{L}}\widetilde{\mathbf{L}}^\top & \mathbf{a} \\ \mathbf{a}^\top & \alpha \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{L}}^{-\top} & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{b} \\ \mathbf{b}^\top & \alpha \end{bmatrix} \triangleq \mathbf{B} \in \mathbb{R}^{n \times n} \quad \text{where } \mathbf{b} \in \widetilde{\mathbf{L}}^{-1}\mathbf{a} \in \mathbb{R}^{n-1}.$$

Let

$$\mathbf{L}_2 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{b}^\top & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Then

$$\mathbf{L}_2^{-1}\mathbf{B}\mathbf{L}_2^{-\top} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{b}^\top & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{b} \\ \mathbf{b}^\top & \alpha \end{bmatrix} \begin{bmatrix} \mathbf{I} & -\mathbf{b} \\ \mathbf{0} & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \alpha - \mathbf{b}^\top\mathbf{b} \end{bmatrix}.$$

Since $\mathbf{A}$ is positive-definite, we have

$$\alpha - \mathbf{b}^\top\mathbf{b} = \alpha - \mathbf{a}^\top\widetilde{\mathbf{L}}^{-\top}\widetilde{\mathbf{L}}^{-1}\mathbf{a} = \alpha - \mathbf{a}^\top\widetilde{\mathbf{L}}^{-\top}\widetilde{\mathbf{L}}^{-1}\mathbf{a} = \alpha - \mathbf{a}^\top\widetilde{\mathbf{A}}^{-1}\mathbf{a} > 0.$$

Let $\alpha - \mathbf{b}^\top\mathbf{b} = \lambda^2$, where $\lambda > 0$. Hence, we have

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \alpha - \mathbf{b}^\top\mathbf{b} \end{bmatrix} = \mathbf{L}_3\mathbf{L}_3^\top, \quad \text{where } \mathbf{L}_3 = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \lambda \end{bmatrix}$$

which means $\mathbf{A} = \mathbf{L}\mathbf{L}^\top \in \mathbb{R}^{n \times n}$ where $\mathbf{L} = \mathbf{L}_1\mathbf{L}_2\mathbf{L}_3 \in \mathbb{R}^{n \times n}$ is a lower triangular matrix with real and positive diagonal entries. $\square$

Theorem 1.4. *Suppose $\nabla^2 f(\mathbf{x})$ is continuous in an open neighborhood of $\mathbf{x}^*$ and that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \succ \mathbf{0}$. Then $\mathbf{x}^*$ is a strict local minimizer of f.*

Proof. Because the Hessian is continuous and positive definite at x^*, we can choose a radius $r > 0$ so that $\nabla^2 f(\mathbf{x})$ remains positive definite for all $\mathbf{x}$ in the open ball $\mathcal{D} = \{\mathbf{z} : \|\mathbf{z} - \mathbf{x}^*\|_2 < r\}$. Taking any nonzero vector $\mathbf{p}$ with $\|\mathbf{p}\|_2 < r$, we have $\mathbf{x}^* + \mathbf{p} \in \mathcal{D}$ and so

$$f(\mathbf{x}^* + \mathbf{p}) = f(\mathbf{x}^*) + \mathbf{p}^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p} = f(\mathbf{x}^*) + \frac{1}{2}\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p},$$

where $\mathbf{z} = \mathbf{x}^* + t\mathbf{p}$ for some $t \in (0, 1)$. Since $\mathbf{z} \in \mathcal{D}$, we have $\mathbf{p}^\top \nabla^2 f(\mathbf{z})\mathbf{p} > 0$, and therefore $f(\mathbf{x}^* + \mathbf{p}) > f(\mathbf{x}^*)$, giving the result. □

Theorem 1.5. *Suppose $\mathbf{x}^*$ is a local minimizer of twice differentiable $f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$ is continuous in an open neighborhood of $\mathbf{x}^*$, then $\nabla^2 f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*) \succeq \mathbf{0}$.*

Proof. Suppose for contradiction that $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Define the vector $p = -\nabla f(\mathbf{x}^*)$, which leads to that $\mathbf{p}^\top \nabla f(\mathbf{x}^*) < 0$. Because ∇f is continuous near $\mathbf{x}^*$, there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla f(\mathbf{x}^* + t\mathbf{p}) < 0,$$

for all for any $t \in [0, T]$. We have by Taylor's theorem that

$$f(\mathbf{x}^* + \bar{t}\mathbf{p}) = f(\mathbf{x}^*) + \bar{t}\mathbf{p}^\top \nabla f(x^* + t\mathbf{p}),$$

for some $t \in (0, \bar{t})$. Therefore, $f(x^* + \bar{t}\mathbf{p}) < f(x^*)$ for all $\bar{t} \in (0, T]$. We have found a direction leading away from x^* along which f decreases, so x^* is not a local minimizer, and we have $\nabla^2 f(\mathbf{x}) = \mathbf{0}$.

For contradiction, assume that $\nabla^2 f(\mathbf{x}^*)$ is not positive semidefinite. Then we can choose a vector $\mathbf{p}$ such that $\mathbf{p}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{p} < 0$. Because $\nabla^2 f(\mathbf{x})$ is continuous near $\mathbf{x}^*$, there is a scalar $T > 0$ such that

$$\mathbf{p}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{p})\mathbf{p} < 0$$

for all $t \in [0, T]$. By doing a Taylor series expansion around x^*, we have for all $\bar{t} \in (0, T]$ and some $t \in (0, \bar{t})$ that

$$f(\mathbf{x}^* + \bar{t}\mathbf{p}) = f(\mathbf{x}^*) + \bar{t}\mathbf{p}^\top \nabla f(\mathbf{x}^*) + \frac{1}{2}\bar{t}^2\mathbf{p}^\top \nabla^2(\mathbf{x}^* + t\mathbf{p})\bar{t}^2\mathbf{p} < f(\mathbf{x}^*).$$

We have found a direction from $\mathbf{x}^*$ along which f is decreasing, and so again, $\mathbf{x}^*$ is not a local minimizer. □

Theorem 1.6. *Given $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$, the solution of minimization problem*

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2.$$

is $\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^n$

Proof. The Hessian of $f(\mathbf{x})$ is $\mathbf{A}^\top\mathbf{A} \succeq \mathbf{0}$, which means $f(\mathbf{x})$ is convex. Let $\mathbf{A} = \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top$ be the condense SVD, where r is the rank of $\mathbf{A}$. Since $\nabla f(\mathbf{x}) = \mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b}$, we only needs to solve the linear system

$$\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}.$$

We denote the solution of $\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}$ be

$$\mathcal{X} = \left\{\mathbf{x} : \mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b} = \mathbf{0}\right\}.$$

We can verify that $\hat{\mathbf{x}} = \mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}$ is the solution of the linear system because

$$\begin{aligned}
&\mathbf{A}^\top\mathbf{A}\hat{\mathbf{x}} - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top\mathbf{A}\left(\mathbf{A}^\dagger\mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger\mathbf{A})\mathbf{y}\right) - \mathbf{A}^\top\mathbf{b} \\
=&\mathbf{A}^\top(\mathbf{A}\mathbf{A}^\dagger - \mathbf{I})\mathbf{b} + \mathbf{A}^\top\mathbf{A}\left(\mathbf{I} - \mathbf{A}^\dagger\mathbf{A}\right)\mathbf{y} \\
=&\mathbf{V}_r\mathbf{\Sigma}_r\mathbf{U}_r^\top(\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top\mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^\top - \mathbf{I})\mathbf{b} + \mathbf{V}_r\mathbf{\Sigma}_r\mathbf{U}_r^\top\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top\left(\mathbf{I} - \mathbf{V}_r\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^\top\mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^\top\right)\mathbf{y}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{V}_r \boldsymbol{\Sigma}_r \mathbf{U}_r^\top (\mathbf{U}_r \mathbf{U}_r^\top - \mathbf{I})\mathbf{b} + \mathbf{V}_r \boldsymbol{\Sigma}_r^2 \mathbf{V}_r^\top \left(\mathbf{I} - \mathbf{V}_r \mathbf{V}_r^\top\right) \mathbf{y} \\
&= \mathbf{V}_r \boldsymbol{\Sigma}_r (\mathbf{U}_r^\top - \mathbf{U}_r^\top)\mathbf{b} + \mathbf{V}_r \boldsymbol{\Sigma}_r^2 \left(\mathbf{V}_r^\top - \mathbf{V}_r^\top\right) \mathbf{y} \\
&= \mathbf{0}.
\end{aligned}
$$

Hence, we have $\mathcal{X}_1 \subseteq \mathcal{X}$, where $\mathcal{X}_1 = \left\{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger \mathbf{b} + (\mathbf{I} - \mathbf{A}^\dagger \mathbf{A})\mathbf{y}, \quad \mathbf{y} \in \mathbb{R}^n\right\}$.

We also have

$$
\begin{aligned}
& \mathbf{A}^\top \mathbf{A} \mathbf{x} - \mathbf{A}^\top \mathbf{b} = \mathbf{0} \\
\Longleftrightarrow & \mathbf{V}_r \boldsymbol{\Sigma}_r^2 \mathbf{V}_r^\top \mathbf{x} - \mathbf{V}_r \boldsymbol{\Sigma}_r \mathbf{U}_r^\top \mathbf{b} = \mathbf{0} \\
\Longleftrightarrow & \boldsymbol{\Sigma}_r^2 \mathbf{V}_r^\top \mathbf{x} - \boldsymbol{\Sigma}_r \mathbf{U}_r^\top \mathbf{b} = \mathbf{0} \\
\Longleftrightarrow & \mathbf{V}_r^\top \mathbf{x} = \boldsymbol{\Sigma}_r^{-1} \mathbf{U}_r^\top \mathbf{b} \\
\Longleftrightarrow & \mathbf{V}_r \mathbf{V}_r^\top \mathbf{x} = \mathbf{V}_r \boldsymbol{\Sigma}_r^{-1} \mathbf{U}_r^\top \mathbf{b} \\
\Longleftrightarrow & \mathbf{x} - (\mathbf{I} - \mathbf{V}_r \mathbf{V}_r^\top)\mathbf{x} = \mathbf{A}^\dagger \mathbf{b} \\
\Longleftrightarrow & \mathbf{x} = \mathbf{A}^\dagger \mathbf{b} + (\mathbf{I} - \mathbf{V}_r \mathbf{V}_r^\top)\mathbf{x}
\end{aligned}
$$

Hence, we have $\mathcal{X} = \left\{\mathbf{x} : \mathbf{x} = \mathbf{A}^\dagger \mathbf{b} + (\mathbf{I} - \mathbf{V}_r \mathbf{V}_r^\top)\mathbf{x}\right\} \subseteq \mathcal{X}_1$. In conclusion, we have $\mathcal{X} = \mathcal{X}_1$. $\square$