# ECE586/AI586 Applied Matrix Analysis - Homework 3

Winter 2024
Instructor: Xiao Fu

School of Electrical Engineering and Computer Science
Oregon State University

March 7, 2025

**Woonki Kim**
**kimwoon@oregonstate.edu**

**Q1.** (The Eckart-Young Theorem) Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and the full SVD $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Define $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Assume that $k \leq \text{rank}(\mathbf{A}) = r$. Show that

$$\mathbf{A}_k = \arg \min_{\mathbf{B}:\, \text{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F.$$

(Hint: Do not use Weyl's inequality directly; if you want to use it, prove it first.) (25%)

**A1.**
Any matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$ can be written as:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V^T}$$

where $\mathbf{U}$ and $\mathbf{V}$ is orthogonal.
Let $r$ be the number of nonzero singular values:

$$\sigma_1 \geq \ldots \sigma_r > 0, \quad \sigma_{r+1} = \ldots = 0$$

Then:

$$\mathbf{M} = \begin{bmatrix} \mathbf{U_1} & \mathbf{U_2} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{\Sigma}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{V_1}^T \\ \mathbf{V_2}^T \end{bmatrix}$$

where

$$\tilde{\mathbf{\Sigma}} = \text{Diag}(\sigma_1, \cdots, \sigma_r)$$

$$\mathbf{U}_1 = \begin{bmatrix} \mathbf{u}_1, \ldots, \mathbf{u}_r \end{bmatrix} \in \mathbb{R}^{m \times r}, \quad \mathbf{U}_2 = \begin{bmatrix} \mathbf{u}_{r+1}, \ldots, \mathbf{u}_m \end{bmatrix} \in \mathbb{R}^{m \times (m-r)}$$

Thus:

$$\mathbf{M} = \mathbf{U}_1 \tilde{\mathbf{\Sigma}} \mathbf{V_1}^T$$

Writing in outerproduct form:

$$\mathbf{M} = \sum_{i}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i}^T$$

$$\mathbf{V}_1 = \begin{bmatrix} \mathbf{v}_1, \ldots, \mathbf{v}_r \end{bmatrix} \in \mathbb{R}^{n \times r}, \quad \mathbf{V}_2 = \begin{bmatrix} \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n \end{bmatrix} \in \mathbb{R}^{n \times (n-r)}$$

Since rank($\mathbf{B}$) = (number of nonzero singular values), any matrix $\mathbf{B}$ with rank at most $k$ can be written as:

$$\mathbf{B} = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

To discriminate from the formula of $\mathbf{A}$, let $\mathbf{B}$:

$$\mathbf{B} = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T$$

where, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 0$ are singular values of $\mathbf{B}$,
$\mathbf{x}_i$ and $\mathbf{y}_i$ are orthonormal left and right singular vectors.

Using thin SVD form, the optimization problem can be reformulated as:

$$\arg \min_{\mathbf{B}:\,\mathrm{rank}(\mathbf{B})\leq k} \|\mathbf{A} - \mathbf{B}\|_F.$$

$$= \arg \min_{\sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T} \| \left( \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \right) - \left( \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T \right) \|_F$$

Optimization problem's solution is invariant to square, thus squaring:

$$\arg \min_{\sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T} \| \left( \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \right) - \left( \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T \right) \|_F^2$$

Looking at the objective function:

$$\| \left( \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \right) - \left( \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T \right) \|_F^2$$

$$= \mathrm{Tr} \left( \left( \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T - \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T \right)^T \left( \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T - \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T \right) \right).$$

Expanding further:

$$= \mathrm{Tr} \left( \left( \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \right)^T \left( \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^T \right) \right)$$

$$-2\,\mathrm{Tr} \left( \left( \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \right)^T \left( \sum_{j=1}^{k} \lambda_j \mathbf{x}_j \mathbf{y}_j^T \right) \right)$$

$$+\,\mathrm{Tr} \left( \left( \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mathbf{y}_i^T \right)^T \left( \sum_{j=1}^{k} \lambda_j \mathbf{x}_j \mathbf{y}_j^T \right) \right).$$

- First term:

$$\mathrm{Tr}\left(\sum_{i=1}^{r}\sum_{j=1}^{r}\sigma_i\sigma_j(\mathbf{u}_i\mathbf{v}_i^T)^T(\mathbf{u}_j\mathbf{v}_j^T)\right).$$

Since $\mathbf{u}_i$ and $\mathbf{v}_i$ are orthonormal vectors:

$$\sum_{i=1}^{r}\sigma_i^2.$$

- Second term:

$$-2\,\mathrm{Tr}\left(\sum_{i=1}^{r}\sum_{j=1}^{k}\sigma_i\lambda_j(\mathbf{u}_i\mathbf{v}_i^T)^T(\mathbf{x}_j\mathbf{y}_j^T)\right).$$

- Third term:

$$\mathrm{Tr}\left(\sum_{i=1}^{k}\sum_{j=1}^{k}\lambda_i\lambda_j(\mathbf{x}_i\mathbf{y}_i^T)^T(\mathbf{x}_j\mathbf{y}_j^T)\right).$$

Since $\mathbf{x}_i$ and $\mathbf{y}_i$ are orthonormal vectors:

$$\sum_{i=1}^{k}\lambda_i^2.$$

Thus:

$$\|\mathbf{A}-\mathbf{B}\|_F^2=\sum_{i=1}^{r}\sigma_i^2-2\sum_{i=1}^{k}\sigma_i\lambda_i\langle\mathbf{x}_i,\mathbf{u}_i\rangle\langle\mathbf{y}_i,\mathbf{v}_i\rangle+\sum_{i=1}^{k}\lambda_i^2.$$

To minimize this expression, second term should be as large as possible.
The second term in the equation,

$$-2\sum_{i=1}^{k}\sigma_i\lambda_i\langle\mathbf{x}_i,\mathbf{u}_i\rangle\langle\mathbf{y}_i,\mathbf{v}_i\rangle,$$

Since $\mathbf{u}_i$ and $\mathbf{v}_i$,$\mathbf{x}_i$ and $\mathbf{y}_i$ are orthonormal, thus unit vectors. Thus by Cauchy-Schwarz inequality:

$$|\langle\mathbf{x}_i,\mathbf{u}_i\rangle|\leq 1.$$

$$|\langle\mathbf{y}_i,\mathbf{v}_i\rangle|\leq 1,$$

where equality holds when:

$$\mathbf{x}_i=\mathbf{u}_i,\quad \mathbf{y}_i=\mathbf{v}_i.$$

Substituting these into the equation:

$$\|\mathbf{A}-\mathbf{B}\|_F^2=\sum_{i=1}^{r}\sigma_i^2-2\sum_{i=1}^{k}\sigma_i\lambda_i+\sum_{i=1}^{k}\lambda_i^2.$$

Dividing sigma term:

$$\sum_{i=k+1}^{r} \sigma_i^2 + \sum_{i=1}^{k} \sigma_i^2 - 2\sum_{i=1}^{k} \sigma_i \lambda_i + \sum_{i=1}^{k} \lambda_i^2.$$

Integrating summation over $i$ to $k$:

$$\sum_{i=k+1}^{r} \sigma_i^2 + \sum_{i=1}^{k} (\sigma_i^2 - 2\sigma\lambda + \lambda_i^2)$$

Expressing in square form:

$$\sum_{i=k+1}^{r} \sigma_i^2 + \sum_{i=1}^{k} (\sigma_i - \lambda_i)^2$$

Square term's minimum is obviously 0, thus it is minimum when:

$$\lambda_i = \sigma_i, \quad \text{for } i = 1, \ldots, k$$

Substituting this choice,

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{i=1}^{r} \sigma_i^2 - 2\sum_{i=1}^{k} \sigma_i \lambda_i + \sum_{i=1}^{k} \lambda_i^2.$$

$$= \sum_{i=1}^{r} \sigma_i^2 - 2\sum_{i=1}^{k} \sigma_i^2 + \sum_{i=1}^{k} \sigma_i^2.$$

Simplifying,

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{i=1}^{r} \sigma_i^2 - \sum_{i=1}^{k} \sigma_i^2 = \sum_{i=k+1}^{r} \sigma_i^2$$

Thus the optimal value for the optimization problem is:

$$\|\mathbf{A} - \mathbf{B}\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2.$$

Recall that frobenius norm could be expressed in summation form of $\sigma$:

$$||\mathbf{X}||_F^2 = ||\mathbf{U\Sigma V}^T||_F^2$$

Since frobenius norm is invariant in orthogonal transformation(It just rotates):

$$||\mathbf{X}||_F^2 = ||\mathbf{U\Sigma V}^T||_F^2 = ||\mathbf{\Sigma}||_F^2 = \sum_i \sigma_i^2$$

And since:

$$\mathbf{A} - \mathbf{A}_k = \sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T.$$

Thus:

$$\|\mathbf{A} - \mathbf{A}_k\|_F^2 = \sum_{i=k+1}^{r} \sigma_i^2$$

Thus optimal value holds when :

$$\mathbf{B} = \mathbf{A}_k$$

Thus:

$$\mathbf{A}_k = \arg \min_{\mathbf{B}:\, \mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_F.$$

---

**Q2.** (The Eckart-Young Theorem) Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and the full SVD $\mathbf{A} = \mathbf{U\Sigma V}^T$. Define $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Assume that $k \leq \mathrm{rank}(\mathbf{A}) = r$. Show that

$$\mathbf{A}_k = \arg \min_{\mathbf{B}:\, \mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_2.$$

Note: This is different from Q1, since the distance is measured using the matrix 2-norm here. (25%)

**A2.**
Since:

$$\|\mathbf{A} - \mathbf{A}_k\|_2 \leq \|\mathbf{A} - \mathbf{B}\|_2, \quad \forall \text{ rank-}k \text{ matrices } \mathbf{B}.$$

Let $\mathbf{B} \in \mathbb{R}^{m \times n}$ and $\mathrm{rank}(\mathbf{B}) = k$.
And we have proven in Homework1 that the null space $\mathcal{N}(\mathbf{B}) \subset \mathbb{R}^n$ dimension:

$$\dim \mathcal{N}(\mathbf{B}) = n - k$$

Now, considering the $n \times (k+1)$ matrix $\mathbf{V}_{k+1}$:

$$\mathbf{V}_{k+1} = [\mathbf{v}_1 \, \mathbf{v}_2 \, \ldots \, \mathbf{v}_{k+1}],$$

since it has (k+1) orthonormal vectors:

$$\mathrm{rank}(\mathbf{V_{k+1}}) = k + 1$$

While $\mathcal{R}(\mathbf{V}_{k+1})$, is the subspace spanned by these $k+1$ column vectors:

$$\mathcal{R}(\mathbf{V}_{k+1}) = \mathrm{span}\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k+1}\}.$$

Since each $\mathbf{v}_i$ is a vector in $\mathbb{R}^n$, their span is a subspace of $\mathbb{R}^n$, meaning:

$$\mathcal{R}(\mathbf{V}_{k+1}) \subset \mathbb{R}^n.$$

Since by the property of dimension:

$$\dim S + \dim T = \dim(S + T) + \dim(S \cap T).$$

Applying this to this case:

$$\dim(\mathcal{N}(\mathbf{B}) + \mathcal{R}(\mathbf{V}_{k+1})) = \dim \mathcal{N}(\mathbf{B}) + \dim \mathcal{R}(\mathbf{V}_{k+1}) - \dim(\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1}))$$

Substituting in actual dimension:

$$\dim(\mathcal{N}(\mathbf{B}) + \mathcal{R}(\mathbf{V}_{k+1})) = (n-k) + (k+1) - \dim(\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1}))$$

$$= (n+1) - \dim(\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1}))$$

Moving all to left side:

$$\dim(\mathcal{N}(\mathbf{B}) + \mathcal{R}(\mathbf{V}_{k+1})) - (n+1) + \dim(\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1})) = 0$$

Since $\mathcal{N}(\mathbf{B}) + \mathcal{R}(\mathbf{V}_{k+1})$ is at most $n$-dimensional (both are a subspace of $\mathbb{R}^n$):

$$\dim(\mathcal{N}(\mathbf{B}) + \mathcal{R}(\mathbf{V}_{k+1})) \leq n$$

Substitute in the inequality to equality:

$$n - (n+1) + \dim(\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1})) \geq 0$$

Thus:

$$\dim(\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1})) \geq 1$$

Which means that $\mathcal{N}(\mathbf{B})$ and $\mathcal{R}(\mathbf{V}_{k+1})$ must have a nontrivial intersection:

$$\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1}) \neq \{0\}.$$

Since the intersection $\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1})$ is nontrivial, there exists at least one nonzero vector $\mathbf{w}$ in this intersection:

$$\mathbf{w} \in \mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V}_{k+1}),$$

which means:

- $\mathbf{w}$ is in the null space of $\mathbf{B}$ thus, $\mathbf{Bw} = \mathbf{0}$
- $\mathbf{w}$ is also in the range space of $\mathbf{V}_{k+1}$, so it can be written as a linear combination of the columns of $\mathbf{V}_{k+1}$

Since this subspace contains nonzero vectors, we can always normalize $\mathbf{w}$ so that $\|\mathbf{w}\|_2 = 1$, meaning there exists a unit vector in this intersection.
Since $\mathbf{w} \in \mathcal{R}(\mathbf{V}_{k+1})$:

$$\mathbf{w} = \sum_{i=1}^{k+1} w_i \mathbf{v}_i,$$

where the $\mathbf{w}$:

$$\sum_{i=1}^{k+1} w_i^2 = 1.$$

By the definition of the matrix norm:

$$\|\mathbf{M}\|_p = \max_{\|\mathbf{x}\|_p \leq 1} \|\mathbf{M}\mathbf{x}\|_p$$

Thus 2-norm satisfies:

$$\|\mathbf{M}\|_2 = \max_{\|\mathbf{x}\|_2 \leq 1} \|\mathbf{M}\mathbf{x}\|_2$$

expressing with inequality:

$$\|\mathbf{M}\|_2 \geq \|\mathbf{M}\mathbf{x}\|_2$$

Since squaring does not change inequality:

$$\|\mathbf{M}\|_2^2 \geq \|\mathbf{M}\mathbf{x}\|_2^2$$

Substituting in $\mathbf{A} - \mathbf{B}$ and $\mathbf{w}$(since $\mathbf{w}$ is unit vector):

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \geq \|(\mathbf{A} - \mathbf{B})\mathbf{w}\|_2^2.$$

Since $\mathbf{w} \in N(\mathbf{B})$, we have $\mathbf{B}\mathbf{w} = \mathbf{0}$, thus:

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \geq \|(\mathbf{A} - \mathbf{B})\mathbf{w}\|_2^2 = \|\mathbf{A}\mathbf{w}\|_2^2.$$

Using the singular value decomposition $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$:

$$\|\mathbf{A}\mathbf{w}\|_2^2 = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w})^T(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{w})$$

$$= \mathbf{w}^T\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T\mathbf{w} = \sum_{i=1}^{k+1} w_i v_i \sigma_i^2 v_i w_i$$

Substituting $\mathbf{w} = \sum_{i=1}^{k+1} w_i \mathbf{v}_i$:

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \geq \|\mathbf{A}\mathbf{w}\|_2^2 = \sum_{i=1}^{k+1} \sigma_i^2 w_i^2$$

Since $\sigma_1 \geq \sigma_2 \geq \ldots$:

$$\sum_{i=1}^{k+1} \sigma_i^2 w_i^2 \geq \sigma_{k+1}^2 \sum_{i=1}^{k+1} w_i^2.$$

Since $\sum_{i=1}^{k+1} w_i^2 = 1$:

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \geq \|\mathbf{A}\mathbf{w}\|_2^2 \geq \sigma_{k+1}^2.$$

Thus:

$$\|\mathbf{A} - \mathbf{B}\|_2^2 \geq \sigma_{k+1}^2.$$

Finally, equality holds when $\mathbf{B} = \mathbf{A_k}$

$$\|\mathbf{A} - \mathbf{A}_k\|_2^2 = \sigma_{k+1}^2.$$

Thus:

$$\mathbf{A}_k = \arg \min_{\mathbf{B}:\, \mathrm{rank}(\mathbf{B}) \leq k} \|\mathbf{A} - \mathbf{B}\|_2$$

---

**Q3.** Consider the nonnegative matrix factorization problem:

$$\mathbf{X} = \mathbf{W}\mathbf{H}^T,$$

where $\mathbf{W} \in \mathbb{R}^{M \times R}$ and $\mathbf{H} \in \mathbb{R}^{N \times R}$, with $\mathrm{rank}(\mathbf{W}) = \mathrm{rank}(\mathbf{H}) = R$. Assume that $\mathbf{H} \geq 0$ and the separability condition holds for $\mathbf{H}$. Denote $\boldsymbol{\Lambda} = \{\ell_1, \ldots, \ell_R\}$ such that

$$\mathbf{H}(\boldsymbol{\Lambda}, :) = \mathbf{I}_R.$$

(a) Assume $\mathbf{H}\mathbf{1} = \mathbf{1}$ and $\mathbf{H} \geq 0$. Consider the following problem:

$$\min_{\mathbf{C} \in \mathbb{R}^{N \times N}} \|\mathbf{C}\|_{\mathrm{row}-0}$$

$$\text{s.t. } \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \mathbf{C} \geq 0, \quad \mathbf{1}^T\mathbf{C} = \mathbf{1}^T,$$

where $\|\mathbf{C}\|_{\mathrm{row}-0}$ counts the number of nonzero rows of $\mathbf{C}$. Assume the separability condition holds for $\mathbf{H}$. Denote $\boldsymbol{\Lambda} = \{\ell_1, \ldots, \ell_R\}$ such that

$$\mathbf{H}(\boldsymbol{\Lambda}, :) = \mathbf{I}_R$$

Denote $\mathbf{C}^\star$ as the solution of (1) and

$$\hat{\boldsymbol{\Lambda}} = \mathrm{supp}(\mathbf{C}^\star),$$

where $\mathrm{supp}(\mathbf{C}^\star)$ extracts the indices of the nonzero rows of $\mathbf{C}^\star$. Show that:

$$\hat{\mathbf{W}} = \mathbf{X}(:, \hat{\boldsymbol{\Lambda}}) = \mathbf{W}\boldsymbol{\Pi},$$

where $\boldsymbol{\Pi}$ is a permutation matrix. (25%)

(b) Assume $\mathbf{H}\mathbf{1} = \mathbf{1}$ and $\mathbf{H} \geq 0$. Consider the following problem:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_{\infty,1}$$

subject to:

$$\mathbf{X} = \mathbf{X}\mathbf{C}, \quad \mathbf{C} \geq 0, \quad \mathbf{1}^T\mathbf{C} = \mathbf{1}^T,$$

where,$\|\mathbf{C}\|_{\infty,1} = \sum_{\ell=1}^{N} \|\mathbf{C}(\ell, :)\|_\infty$ Denote $\mathbf{C}^\star$ as the solution of (2). Also assume that all the rows of $\mathbf{H}$ are different, i.e., $\mathbf{H}(\ell, :) \neq \mathbf{H}(k, :)$ for all $k \neq \ell$. Show that:

$$\mathrm{supp}(\mathbf{C}^\star) = \boldsymbol{\Lambda},$$

where $\mathrm{supp}(\mathbf{C}^\star)$ extracts the indices of the nonzero rows of $\mathbf{C}^\star$. (25%)

**Remark:** The formulations in (1) and (2) offer alternative approaches for handling the separable NMF problem. Particularly, (2) is a convex optimization-based formulation, which automatically implies that there exists a polynomial-time algorithm for solving the separable NMF problem.

**A3.**

**(a)** Before getting into the problem check $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{H}$ is all non-negative:

We have

$$\mathbf{C} \geq 0, \quad \mathbf{H} \geq 0$$

By constraint:

$$\mathbf{X} = \mathbf{X}\mathbf{C}$$

we can say that $\mathbf{X}$ is non-negative, since $\mathbf{C}$ is non-negative and if $\mathbf{X}$ holds negative elements the equation cannot hold.

With the equation:

$$\mathbf{X} = \mathbf{W}\mathbf{H}^T$$

$\mathbf{W}$ is also non-negative, since both $\mathbf{X}$ and $\mathbf{H}$ is non-negative.

1. First consider separability condition:

$$\mathbf{H}(\mathbf{\Lambda}, :) = \mathbf{I}_R.$$

where $\mathbf{H}(\mathbf{\Lambda}, :) \in \mathbb{R}^{R \times R}$

Since we assume $\mathbf{H} \geq 0$ and $\text{rank}(\mathbf{H}) = R$ implying $N > R$, we can express $\mathbf{H}$ as:

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & \cdots & * & * & * \\ 0 & 1 & 0 & \cdots & * & * & * \\ 0 & 0 & 1 & \cdots & * & * & * \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & \cdots & * & * & * \end{bmatrix}$$

where $\mathbf{H} \in \mathbb{R}^{N \times R}$ with the first $R$ rows form the identity matrix, and the remaining rows contain nonnegative values expressed as *.

Now consider:

$$\mathbf{X} = \mathbf{W}\mathbf{H}^T$$

if we select a single column at index $\ell_r$ from $\mathbf{X}$, we get:

$$\mathbf{X}(:, \ell_r)$$

Which can be expressed as:

$$\mathbf{X}(:, \ell_r) = \mathbf{W}\mathbf{H}(\ell_r, :)^T.$$

where $\mathbf{H}(\ell_r, :)^T$ represents the $\ell_r$-th row of $\mathbf{H}^T$.

Since we know that $\mathbf{H}(\mathbf{\Lambda}, :) = \mathbf{I}_R$, the row at index $\ell_r$ must be the $r$-th standard basis vector:

$$\mathbf{H}(\ell_r, :) = \mathbf{e}_r^T = [0, 0, \ldots, 1, \ldots, 0].$$

where the 1 is in the $r$-th position.

Thus:

$$\mathbf{H}(\ell_r, :)^T = \mathbf{e}_r \in \mathbb{R}^R.$$

And this standard basis vector $\mathbf{e}_r$ selects the $r$-th column of $\mathbf{W}$:

$$\mathbf{W}\mathbf{e}_r = \mathbf{w}_r.$$

where $\mathbf{w}_r$ is the $r$-th column of $\mathbf{W}$.

Thus:

$$\mathbf{X}(:, \ell_r) = \mathbf{w}_r.$$

Thus, if we select exactly the columns indexed by $\mathbf{\Lambda}$:

$$\mathbf{X}(:, \mathbf{\Lambda}) = \mathbf{W}.$$

<u>2.Now consider the constraint $\mathbf{X} = \mathbf{X}\mathbf{C}$</u>

Substituting $\mathbf{X} = \mathbf{W}\mathbf{H}^T$:

$$\mathbf{W}\mathbf{H}^T = (\mathbf{W}\mathbf{H}^T)\mathbf{C}.$$

Since $\mathbf{W}$ is a full column rank:

$$\mathbf{W}^\dagger\mathbf{W}\mathbf{H}^T = \mathbf{W}^\dagger(\mathbf{W}\mathbf{H}^T)\mathbf{C}.$$

Thus:

$$\mathbf{H}^T = \mathbf{H}^T\mathbf{C}.$$

Here by assumption:

$$\mathbf{H}(\mathbf{\Lambda}, :) = \mathbf{I}_R.$$

Thus, we can write $\mathbf{H}^T$ in block form:

$$\mathbf{H}^T = \begin{bmatrix} \mathbf{I}_R \\ \mathbf{H}_{\text{other}}^T \end{bmatrix}.$$

Substitute in block form into the equation:

$$\begin{bmatrix} \mathbf{I}_R \\ \mathbf{H}_{\text{other}}^T \end{bmatrix} (\mathbf{I} - \mathbf{C}) = 0.$$

- For the First $R$ Rows (Corresponding to $\mathbf{\Lambda}$):

  $$\mathbf{I}_R(\mathbf{I} - \mathbf{C}) = 0.$$

  Since $\mathbf{I}_R$ is an identity matrix, this forces:

  $$(\mathbf{I} - \mathbf{C})(\ell, :) = 0, \quad \forall \ell \in \mathbf{\Lambda}.$$

  This means that for rows indexed by $\mathbf{\Lambda}$, we must have:

  $$\mathbf{I} - \mathbf{C} = 0 \quad (\mathbf{C}(\ell, :) = \mathbf{I}(\ell, :)).$$

  Thus, the rows indexed by $\mathbf{\Lambda}$ in $\mathbf{C}$ form an identity matrix.

- For the Other Rows:

  $$\mathbf{H}_{\text{other}}^T(\mathbf{I} - \mathbf{C}) = 0.$$

  which implies each row of $\mathbf{I} - \mathbf{C} \in \mathcal{N}(\mathbf{H}_{\text{other}}^T)$.
  Since $\mathbf{H}^T$ is full column rank $R$, the rows indexed by $\mathbf{\Lambda}$ (which form $\mathbf{I}_R$) already span the row space. This means that the remaining rows in $\mathbf{H}_{\text{other}}^T$ do not contribute to a larger row space Thus, the null space of $\mathbf{H}_{\text{other}}^T$ is exactly the subspace spanned by the identity rows indexed by $\mathbf{\Lambda}$.
  Thus:

  $$\mathcal{N}(\mathbf{H}_{\text{other}}^T) = \text{span}\{\text{rows of } \mathbf{I}_R\}.$$

  Since the rows of $\mathbf{I}_R$ are already a basis, this means the basis of $\mathcal{N}(\mathbf{H}_{\text{other}}^T)$ is given by the rows of $\mathbf{I}_R$
  Since each row of $\mathbf{I} - \mathbf{C}$ must lie in $\mathcal{N}(\mathbf{H}_{\text{other}}^T)$, and we just showed that $\mathcal{N}(\mathbf{H}_{\text{other}}^T)$ is spanned by the rows of $\mathbf{I}_R$:

  $$\mathbf{C}(k, :) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{C}(\ell, :).$$

  Since $\mathbf{C}(\ell, :) = \mathbf{I}_R$ for $\ell \in \mathbf{\Lambda}$:

  $$\mathbf{C}(k, :) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{I}_R.$$

  Thus, every remaining row of $\mathbf{C}$ is a linear combination of the identity rows.
  Since $\mathbf{C}(\ell, :) = \mathbf{I}_R$ for $\ell \in \mathbf{\Lambda}$:

  $$\mathbf{C}(k, :) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{I}_R.$$

3.Now showing the relationship between $\mathbf{\Lambda}$ and $\hat{\mathbf{\Lambda}}$

Recall $\mathbf{C}^*$ is the solution for optimization problem and the set $\hat{\mathbf{\Lambda}}$ is the set of indices corresponding to the nonzero rows of $\mathbf{C}^\star$:

$$\hat{\mathbf{\Lambda}} = \text{supp}(\mathbf{C}^\star),$$

- $\|\mathbf{C}\|_{\text{row}-0}$ counts the number of non-zero rows in $\mathbf{C}$. Thus minimizing this ensures that only a minimal number of rows in $\mathbf{C}$ remain nonzero.
- And by constraint, for rows indexed by $\mathbf{\Lambda}$, we must have:

  $$\mathbf{I} - \mathbf{C} = 0 \quad (\mathbf{C}(\ell, :) = \mathbf{I}(\ell, :)).$$

  which means the rows indexed by $\mathbf{\Lambda}$ in $\mathbf{C}$ form an identity matrix.
- And also by constraint, each remaining row of $\mathbf{C}$ (rows not in $\mathbf{\Lambda}$):

  $$\mathbf{C}(k, :) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{C}(\ell, :),$$

  where $\alpha_{k\ell}$ are some coefficients.

All this conditions can be achieved by letting $\mathbf{C}$ to have the rows indexed by $\mathbf{\Lambda}$ in $\mathbf{C}$ to form an identity matrix, and setting all coefficient $\alpha_{k\ell} = 0$, making other remaining row of $\mathbf{C}$'s element to zero.

Thus the optimal $\mathbf{C}$ looks like:

$$\mathbf{C}^* = \begin{bmatrix} \mathbf{H}^T \\ 0 \end{bmatrix}$$

Thus the nonzero rows of $\mathbf{C}^\star$ are exactly $\mathbf{\Lambda}$:

$$\text{supp}(\mathbf{C}^\star) = \hat{\mathbf{\Lambda}} = \mathbf{\Lambda}.$$

4.Now consider $\hat{\mathbf{W}}$

By definition, $\hat{\mathbf{W}}$ is the matrix obtained by selecting the columns of $\mathbf{X}$ indexed by $\hat{\mathbf{\Lambda}}$:

$$\hat{\mathbf{W}} = \mathbf{X}(:, \hat{\mathbf{\Lambda}}).$$

From 3. we proved that:

$$\hat{\mathbf{\Lambda}} = \mathbf{\Lambda}.$$

Thus:

$$\hat{\mathbf{W}} = \mathbf{X}(:, \mathbf{\Lambda}).$$

From 1. we proved that:

$$\mathbf{X}(:, \boldsymbol{\Lambda}) = \mathbf{W}$$

Finally:

$$\hat{\mathbf{W}} = \mathbf{X}(:, \boldsymbol{\Lambda}) = \mathbf{W}$$

But here:

- $\boldsymbol{\Lambda}$ is a fixed set of $R$ indices.
- $\hat{\boldsymbol{\Lambda}}$ is also a set of $R$ indices, but they may be selected in a different order by the optimization process.

Thus even though we have identified the correct indices in $\boldsymbol{\Lambda}$, we cannot guarantee that they appear in the same order in $\hat{\boldsymbol{\Lambda}}$.

So we use permutation:

$$\hat{\mathbf{W}} = \mathbf{W}\boldsymbol{\Pi}.$$

where $\boldsymbol{\Pi}$ is an $R \times R$ permutation matrix, meaning it only reorders the columns of $\mathbf{W}$.

Thus:

$$\hat{\mathbf{W}} = \mathbf{W}\boldsymbol{\Pi}.$$

**(b)** 1. First consider the objective function

We want to minimize:

$$\|\mathbf{C}\|_{\infty,1} = \sum_{\ell=1}^{N} \|\mathbf{C}(\ell, :)\|_{\infty}.$$

Mixed norm $\|\mathbf{C}\|_{\infty,1}$:

$$\|\mathbf{C}\|_{\infty,1} = \sum_{\ell=1}^{N} \|\mathbf{C}(\ell, :)\|_{\infty}.$$

where:

$$\|\mathbf{C}(\ell, :)\|_{\infty} = \max_j \sum_{j=1}^{n} |C(\ell, j)|$$

Here:

- $\|\mathbf{C}(\ell, :)\|_{\infty}$ takes the maximum absolute value in row $\ell$.

- And summing over all rows means the objective function gathers row-wise large values in $\mathbf{C}$.
- And by minimizing this objective function, it ensures row sparsity in $\mathbf{C}$, meaning that $\mathbf{C}$ should have as few nonzero rows as possible.

2.Now consider the constraint $\mathbf{X} = \mathbf{X}\mathbf{C}$ (Same as problem (a))

Substituting $\mathbf{X} = \mathbf{W}\mathbf{H}^T$:

$$\mathbf{W}\mathbf{H}^T = (\mathbf{W}\mathbf{H}^T)\mathbf{C}.$$

Since $\mathbf{W}$ is a full column rank:

$$\mathbf{W}^\dagger\mathbf{W}\mathbf{H}^T = \mathbf{W}^\dagger(\mathbf{W}\mathbf{H}^T)\mathbf{C}.$$

Thus:

$$\mathbf{H}^T = \mathbf{H}^T\mathbf{C}.$$

Here by assumption, $\mathbf{H}$ satisfies the separability condition, meaning there exists an index set $\mathbf{\Lambda} = \{\ell_1, \ldots, \ell_R\}$ such that:

$$\mathbf{H}(\mathbf{\Lambda}, :) = \mathbf{I}_R.$$

Thus, we can write $\mathbf{H}^T$ in block form as:

$$\mathbf{H}^T = \begin{bmatrix} \mathbf{I}_R \\ \mathbf{H}_{\text{other}}^T \end{bmatrix}.$$

Substitute in block form into the equation:

$$\begin{bmatrix} \mathbf{I}_R \\ \mathbf{H}_{\text{other}}^T \end{bmatrix} (\mathbf{I} - \mathbf{C}) = 0.$$

- For the First $R$ Rows (Corresponding to $\mathbf{\Lambda}$):

  $$\mathbf{I}_R(\mathbf{I} - \mathbf{C}) = 0.$$

  Since $\mathbf{I}_R$ is an identity matrix thus:

  $$(\mathbf{I} - \mathbf{C})(\ell, :) = 0, \quad \forall \ell \in \mathbf{\Lambda}.$$

  This means that for rows indexed by $\mathbf{\Lambda}$:

  $$\mathbf{I} - \mathbf{C} = 0 \quad (\mathbf{C}(\ell, :) = \mathbf{I}(\ell, :)).$$

  Thus, the rows indexed by $\mathbf{\Lambda}$ in $\mathbf{C}$ form an identity matrix.

- For the Other Rows:

$$\mathbf{H}_{\text{other}}^T(\mathbf{I} - \mathbf{C}) = 0.$$

which implies each row of $\mathbf{I} - \mathbf{C} \in \mathcal{N}(\mathbf{H}_{\text{other}}^T)$.
Since $\mathbf{H}^T$ is full column rank $R$, the rows indexed by $\mathbf{\Lambda}$ (which form $\mathbf{I}_R$) already span the row space. This means that the remaining rows in $\mathbf{H}_{\text{other}}^T$ do not contribute to a larger row space Thus, the null space of $\mathbf{H}_{\text{other}}^T$ is exactly the subspace spanned by the identity rows indexed by $\mathbf{\Lambda}$.
Thus:

$$\mathcal{N}(\mathbf{H}_{\text{other}}^T) = \text{span}\{\text{rows of } \mathbf{I}_R\}.$$

Since the rows of $\mathbf{I}_R$ are already a basis, this means the basis of $\mathcal{N}(\mathbf{H}_{\text{other}}^T)$ is given by the rows of $\mathbf{I}_R$
Since each row of $\mathbf{I} - \mathbf{C}$ must lie in $\mathcal{N}(\mathbf{H}_{\text{other}}^T)$, and we just showed that $\mathcal{N}(\mathbf{H}_{\text{other}}^T)$ is spanned by the rows of $\mathbf{I}_R$:

$$\mathbf{C}(k,:) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{C}(\ell,:).$$

Since $\mathbf{C}(\ell,:) = \mathbf{I}_R$ for $\ell \in \mathbf{\Lambda}$:

$$\mathbf{C}(k,:) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{I}_R.$$

Thus, every remaining row of $\mathbf{C}$ is a linear combination of the identity rows.
Since $\mathbf{C}(\ell,:) = \mathbf{I}_R$ for $\ell \in \mathbf{\Lambda}$:

$$\mathbf{C}(k,:) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{I}_R.$$

<u>3.Now the summing up objective function and constraints</u>

- By minimizing $\|\mathbf{C}\|_{\infty,1}$ ensures that only a minimal number of rows in $\mathbf{C}$ remain nonzero.
- And by constraint, for rows indexed by $\mathbf{\Lambda}$:

$$\mathbf{I} - \mathbf{C} = 0 \quad (\mathbf{C}(\ell,:) = \mathbf{I}(\ell,:)).$$

which means the rows indexed by $\mathbf{\Lambda}$ in $\mathbf{C}$ form an identity matrix.
- And also by constraint, each remaining row of $\mathbf{C}$ ( rows not in $\mathbf{\Lambda}$):

$$\mathbf{C}(k,:) = \sum_{\ell \in \mathbf{\Lambda}} \alpha_{k\ell} \mathbf{C}(\ell,:),$$

where $\alpha_{k\ell}$ are coefficients.

This can be achieved by letting $\mathbf{C}$ to have the rows indexed by $\mathbf{\Lambda}$ in $\mathbf{C}$ to form an identity matrix, and setting all coefficient $\alpha_{k\ell} = 0$, making other remaining row of $\mathbf{C}$'s element to zero.

Thus the optimal $\mathbf{C}$ looks like:

$$\mathbf{C}^* = \begin{bmatrix} \mathbf{H}^T \\ 0 \end{bmatrix}$$

So the nonzero rows of $\mathbf{C}^\star$ are exactly $\mathbf{\Lambda}$, thus:

$$\mathrm{supp}(\mathbf{C}^\star) = \mathbf{\Lambda}.$$