Numerical Analysis II - Assignment I

Francis Fregeau

January 2023

A.1

Let M be a $n \times m$ matrix where $n = \tau \cdot a$ and $m = \lambda \cdot b$, with $\{\tau, \lambda, a, b\}$ being natural numbers. Let $A_{i,j}$ be the i^{th} row-wise and j^{th} column-wise sub-matrix contained within M, with $\dim(A_{i,j}) = (a, b)$.

Let us swap the sub-matrices $A_{i,j}$ with $A_{j,i}$, i.e.:

$$A_{i,j} \to A_{j,i}$$

$$A_{j,i} \to A_{i,j}$$

Let $k \leq \min(n, m)$. The k^{th} row of M used to be $\cup_{j=1}^{\lambda} \text{row}_k(A_{i,j})$ for some $(i-1) \cdot a \leq k \leq i \cdot a$. However, it is now $\cup_{i=1}^{\tau} \text{col}_k(A_{i,j})$ for some $(j-1) \cdot b \leq k \leq j \cdot b$.

As an example, consider the first pre-swap row of M which was the union of the first rows of $\cup_{j=1}^{\lambda} U_{1,j}$. It is obvious from the picture that the first post-swap row is now the first column of $\cup_{i=1}^{\tau} U_{i,1}$. Since choosing the first row was made without loss of generality, we can conclude that the k^{th} row of M is now its k^{th} column, so long as $k \leq \min(m, n)$. In other words, M is now M^{T} as a result of the sub-matrices swaps.

Lastly, the order of operation doesn't matter for conjugate transpose since

$$U^{\text{T}} + V^{\text{T}} = (U + V)^{\text{T}}$$

and as such, one can simply applies the conjugate operator to M prior to making the swaps, which is akin to applying said operator on every $A_{i,j}$.

A.2

Let $v \in \mathbb{C}^n = \lambda \cdot a$ and $w \in \mathbb{C}^m = \tau \cdot b$, where $\|a\|_2 = \|b\|_2 = 1$. The SVD of $M = vw^* \in \mathbb{C}^{n \times m}$ is

$$M = \lambda\tau \cdot U\Sigma V^*$$

where the first column of U is a, the first row of V^* is b and all entries of Σ are null save for $\Sigma_{1,1}$. (This is simply a convoluted way of writing down $M = vw^* = \lambda\tau \cdot ab^*$.) Since both of these vectors have unit norm, then surely $\Sigma_{1,1} = 1$. As such, we can simply let $\Sigma_{1,1} = \lambda\tau$ and drop the $\lambda\tau$ scalar. In other words, the sole singular value of M is the product of the norms of v and w. The aforementioned quantity is precisely $\|M\|_2$, which indicates that $\|M\|_2 = \|v\|_2 \cdot \|w\|_2$ as desired.

As for the second part of the exercise, $\|M\|_\infty$ is the supremum of the sums of individual rows of $|M|$, whereas $\|v\|_\infty = \max_i(|v_i|)$ and $\|w\|_1 = \sum_{i=1}^m |w_j|$. It suffices to observe that

$$\begin{aligned} |r_i| &= |\sum_{j=1}^m v_i \cdot w_j| \\ &\leq \sum_{j=1}^m |v_i| \cdot |w_j| \\ &\leq \max_i(|v_i|) \cdot \sum_{j=1}^m |w_j| \\ &= \|v\|_\infty \|w\|_1 \end{aligned}$$

so to deduce that

$$\|M\|_\infty = \max_i(|r_i|) \leq \|v\|_\infty \|w\|_1$$

A.3

By the SVD theorem, let

$$A = U\Sigma_A U^*$$
$$B = V\Sigma_B V^*$$

Since

$$A = QBQ^*$$
$$= (QV)\,\Sigma_B\,(V^*Q^*)$$

and

$$(QV)\,(QV)^*$$
$$= (QV)\,(V^*Q^*)$$
$$= \mathrm{I}$$

It follows that

$$A = (QV)\,\Sigma_B\,(V^*Q^*)$$

is in fact the SVD of A, and by uniqueness, it must be the case that

$$(QV)\,\Sigma_B\,(V^*Q^*) = U\Sigma_A U^*$$
$$\Rightarrow QV = U$$
$$\Rightarrow \Sigma_A = \Sigma_B$$

Therefore the statement is true.

A.4

The Eckart–Young–Mirsky theorem states that the optimal low-rank approximation with respect to the Frobenius norm of A can be obtained by using the SVD.

$$AA^* = \begin{bmatrix} 1+|z|^2 & z \\ z^* & 1 \end{bmatrix}$$

$$A^*A = \begin{bmatrix} 1+|z|^2 & z \\ z^* & 1 \end{bmatrix}$$

$$\mathbf{det}\left(AA^* - \lambda \mathrm{I}\right) = \lambda^2 - \left(2+|z|^2\right)\lambda + 1 = 0$$

$$\lambda_1 = \frac{2+|z|^2+\sqrt{\left(2+|z|^2\right)^2-4}}{2}$$

$$\lambda_2 = \frac{2+|z|^2-\sqrt{\left(2+|z|^2\right)^2-4}}{2}$$

I'll settle for whatever partial marks I can get by simply stating that the outer product of the unit eigenvectors of AA^* and A^*A tied to the eigenvalue $\lambda = \max\left(\lambda_1,\ \lambda_2\right)$ multiplied by λ will give the optimal rank 1 approximation.

A.5

$$
\begin{aligned}
&\mathbf{det}\,(B)\\
&= \mathbf{det}\,(U\Sigma_B V^*)\\
&= \pm\prod_{i=1}^{n}(\Sigma_B)_{i,i}\\
&|\mathbf{det}\,(A)\,|\\
&= |\mathbf{det}\,(K\Sigma_A W^*)\,|\\
&= |\prod_{i=1}^{n}(\Sigma_A)_{i,i}\,|
\end{aligned}
$$

Thus either $\mathbf{det}\,(U) = \mathbf{det}\,(V) = 1$ or $\mathbf{det}\,(U) = \mathbf{det}\,(V) = -1$.

$$
\begin{aligned}
&\|A - B\|_F\\
&= \mathbf{Tr}\,(AA^* - AB^* - BA^* + BB^*)
\end{aligned}
$$

?

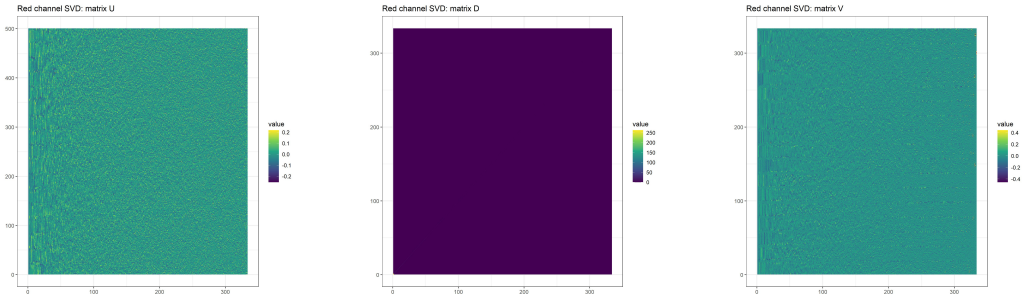
A.6

?

B.1.1

See section B.1.1 of the R_code_B.1.R file. Codes can be found here: `https://github.com/c-Stats/Numerical_Analysis_II_A1`.

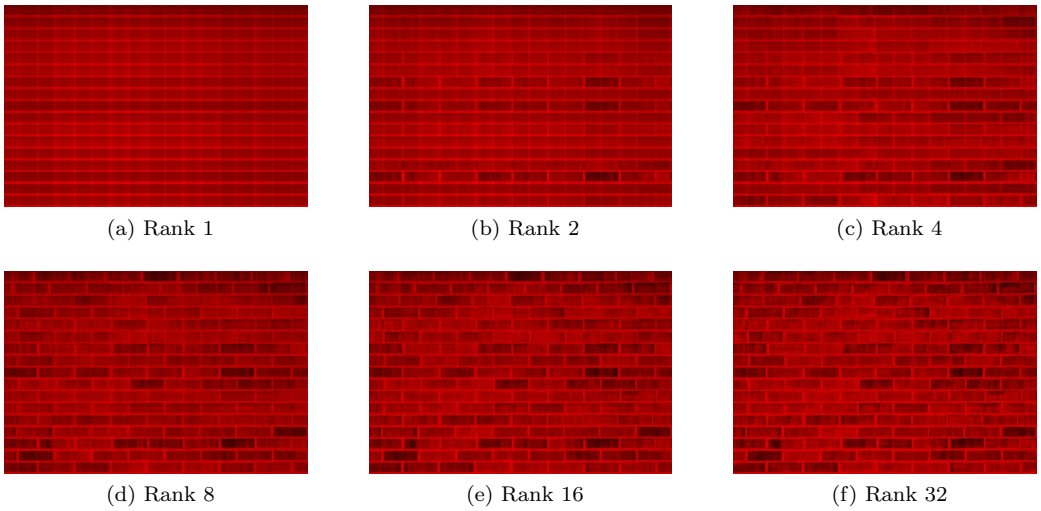
B.1.2



(a) U matrix (b) D matrix (c) V matrix

Figure 1: Red channel's SVD decomposition

B.1.3



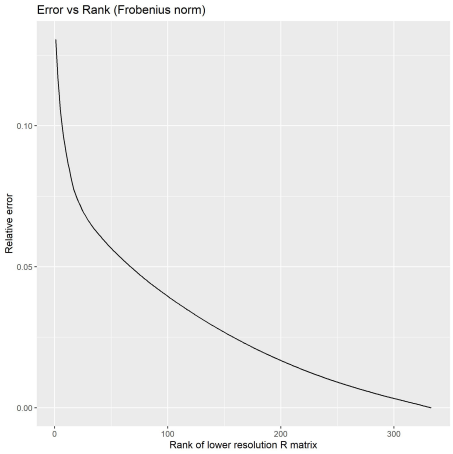
(a) Rank 1 (b) Rank 2 (c) Rank 4

(d) Rank 8 (e) Rank 16 (f) Rank 32

Figure 2: Red channel's Rank-n approximations

The picture becomes clearer as more outer products are added to form the approximation. The rank 32 approximation looks very similar to the full red channel picture.

B.1.4

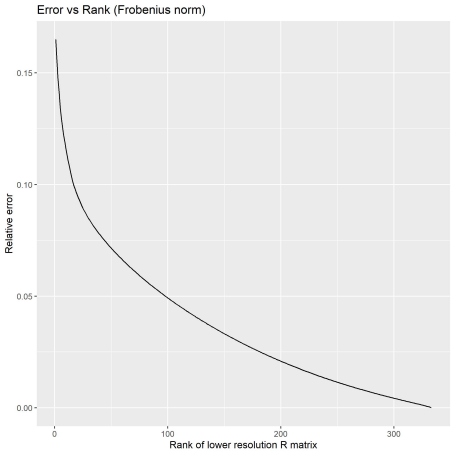


(a) Errors

Figure 3: Scaled absolute errors

The errors go down as expected, albeit at a decreasing rate with respect to the approximation rank.

B.1.5



(a) Errors

Figure 4: Scaled squared errors

The results are very similar to those of B.1.4.

B.1.6

Lower rank approximations for the red channel are of mediocre quality when there are a lot of small details in the picture, as demonstrated below using a glorious rendition of Donald Trump riding a giant eagle whose wings are equipped with state of the art air to ground missiles. The 32 rank rendering fails to adequately capture Mount Rushmore's details, as well as the majestic forest sitting beneath the 45^{th} president of the United States.

(a) Original (b) Red channel, rank 32

Figure 5: Trump's red channel SVD approximation

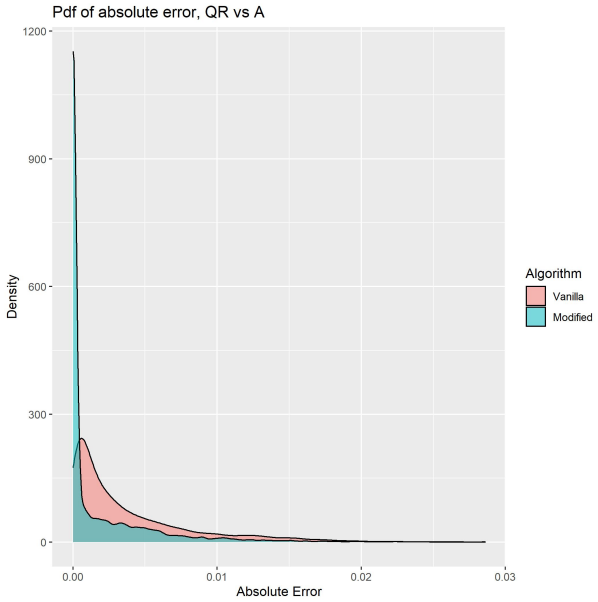
In contrast, the rank 32 picture reasonably approximates the famous painting of Napoleon as Emperor of the French by François Gérard, save for the small motifs of the imperial attire on display and his majesty's face.

(a) Original (b) Red channel, rank 32

Figure 6: Emperor Napoleon's red channel SVD approximation

B.2

Functions used to produce the plot shown below are located in the __main__.R file located within the B.2 folder.



(a) Vanilla vs modified algorithm

Figure 7: Pdf of absolute QR-decomposition errors

The modified algorithm produces smaller errors as evidenced by the sharp peak in density around 0, and a much slimmer right tail.

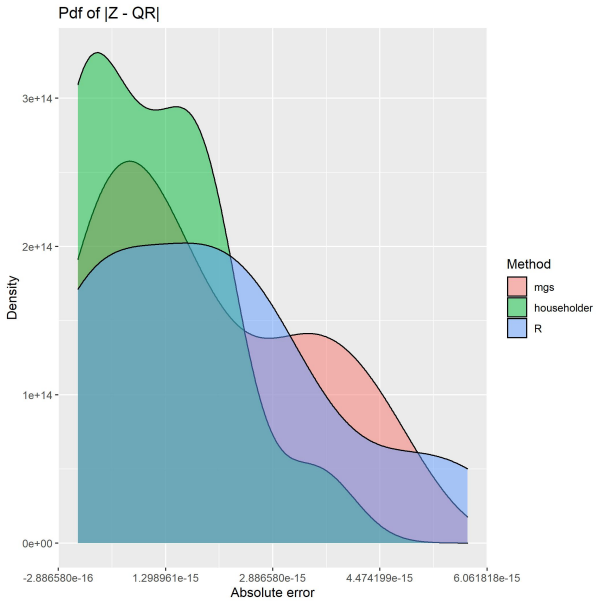
B.3.1

See the __main__.R file located within the B.3 folder.

B.3.2

See the __main__.R file located within the B.3 folder.

B.3.3



(a) Mgs, HH and base R

Figure 8: Pdf of absolute QR-decomposition errors

The householder method produces errors of lower magnitudes as evidenced by the higher mass located on the left half of the horizontal axis coupled with a slimmer right tail. The mgs method and R's base function seem to have relatively similar performance, save for the former's pdf having two sharp bumps rising above the later's.