1 On centering of the data

In Chapter 10 there is a problem with centering of the data. It starts already in Definition 10.1.0.1 which states that the variance of a set of vectors equals the sum of the squares of their norms. This is usually not the case. The variance of a set of vectors is supposed to measure how much the vectors vary from each other. Thus a proper definition of the variance is something along the following lines (remember that mathematical definitions are not set in stone, there are conventions and preferences).

Definition 1.1. *The variance of a set of vectors* $\{v_1, v_2, ..., v_N\}$ *is*

$$\frac{1}{2N} \sum_{i,j=1}^{N} |v_i - v_j|^2.$$

Here $|v|$ *is the euclidean length of vector* $v = (v^1, v^2, ..., v^d) \in \mathbb{R}^d$ *given by*

$$\sqrt{\sum_{k=1}^{d} (v^k)^2}.$$

Let's compare it to the one proposed in the book which is $\sum_{i=1}^{N} |v_i|^2$. We calculate the Newton's binomial

$$\begin{aligned} \frac{1}{2N} \sum_{i,j=1}^{N} |v_i - v_j|^2 &= \frac{1}{2N} \sum_{i,j=1}^{N} \left(|v_i|^2 + |v_j|^2 - 2v_i \cdot v_j \right) \\ &= \frac{1}{2N} \sum_{i,j=1}^{N} |v_i|^2 + \frac{1}{2N} \sum_{i,j=1}^{N} |v_j|^2 - \frac{1}{N} \sum_{i,j=1}^{N} v_i \cdot v_j, \end{aligned} \tag{1.1}$$

where $\cdot$ denotes the inner product. We have

$$\frac{1}{2N} \sum_{i,j=1}^{N} |v_i|^2 = \frac{1}{2N} \sum_{j=1}^{N} \left(\sum_{i=1}^{N} |v_i|^2 \right) = \frac{1}{2N} N \sum_{i=1}^{N} |v_i|^2 = \frac{1}{2} \sum_{i=1}^{N} |v_i|^2$$

Similarly

$$\frac{1}{2N} \sum_{i,j=1}^{N} |v_j|^2 = \frac{1}{2} \sum_{j=1}^{N} |v_j|^2 = \frac{1}{2} \sum_{i=1}^{N} |v_i|^2.$$

Therefore (1.1) leads to

$$\frac{1}{2N} \sum_{i,j=1}^{N} |v_i - v_j|^2 = \sum_{i=1}^{N} |v_i|^2 - \frac{1}{N} \sum_{i,j=1}^{N} v_i \cdot v_j,$$

which means that the total variance defined here and the total variance defined in the book are different unless

$$\frac{1}{N} \sum_{i,j=1}^{N} v_i \cdot v_j = 0.$$

However

$$\frac{1}{N}\sum_{i,j=1}^{N} v_i \cdot v_j = \frac{1}{N}\left(\sum_{i=1}^{N} v_i\right)\cdot\left(\sum_{i=j}^{N} v_j\right) \tag{1.2}$$

and we see that if the data are centered, that is if their mean is equal to 0, that is if $\frac{1}{N}\sum_{i=1}^{N} v_i = 0$, then we have 0 in (1.2), and thus

$$\frac{1}{2N}\sum_{i,j=1}^{N} |v_i - v_j|^2 = \sum_{i=1}^{N} |v_i|^2.$$

Otherwise **it is not true**.

Conclusions

- Throughout Ch. 10 we should assume that the data are always centered, that is the sums of entries in each row equals to 0. If it is not the case, we have to center the data. We do it by calculating the average and subtracting it.

  Example.

  $$A = \begin{bmatrix} 4 & 1 & -2 & -3 \\ 1 & 1 & 0 & -2 \end{bmatrix} \qquad \text{is centered.}$$

  On the other hand

  $$B = \begin{bmatrix} -1 & 1 & 2 \\ -1 & 2 & 2 \end{bmatrix} \qquad \text{is not centered.}$$

  The mean is

  $$M = \frac{1}{3}\begin{bmatrix} -1+1+2 \\ -1+2+2 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2 \\ 3 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ 1 \end{bmatrix}.$$

  Thus to center B we simply subtract M from each column of B obtaining

  $$\hat{B} = \begin{bmatrix} -\frac{5}{3} & \frac{1}{3} & \frac{4}{3} \\ -2 & 1 & 1 \end{bmatrix}.$$

- The centering has to be done everywhere in Ch 10. For example Ex. 10.1 and Ex. 10.2 work just fine because the matrix is already centered. On the other hand Ex 10.3 requires centering of the matrix. Otherwise what we call total variance will not be the true total variance.

- To compare the viability of Definition 1.1 and Definition 10.1.0.1 for the total variance just consider 100 vectors $[1, 0, 0]^T$. This is literally the same vector taken 100 times so the variance should be 0. And according to Definition 1.1 it is 0, while according to Definition 10.1.0.1 it is 100.

- As a side note, in Ch 10.5 there is a comment on the centering but Justin decided that for simplicity we should "completely ignore" this issue. So the problem is recognized but ignored for simplicity. I admit that there is some merit to this simplification and I respect it. I also completely disagree. In my opinion it doesn't look good from the mathematical perspective (since we end up with a bad definition of total variance) and from the perspective of applications (since in practice this may lead to carelessness, which nobody outside the university will be happy to overlook).

2 SVD: freedom in the choice of eigenvectors

There is also a problem in Chapter 9, which appears in Ex 9.1 and 9.2 and generally in the SVD "algorithm". Let us recall it:

1. Given A first we calculate AA^T and its eigenvalue-eigenvector pairs with unit eigenvectors.

2. At this point you should rearrange the eigenvalues (and corresponding eigenvectors) in the nonincreasing order.

3. Then we do the same for $A^T A$ but there is a catch. If v_i is a unit eigenvector corresponding to λ_i then so is $-v_i$. We need to make sure that the eigenvecor v_i corresponds to the proper choice of eigenvecor u_i for AA^T! The answer is (hidden) in Theorem 9.3.1.1: having u_i calculated in the previous step we take

$$v_i = \frac{A^T u_i}{\|A^T u_i\|}. \tag{2.1}$$

   In particular it means that if we take $-u_i$ instead of u_i, we will also have to take $-v_i$ instead of v_i. This is very important, since otherwise it doesn't work! The eigenvectors corresponding to eigenvalues 0 can be chosen in the standard way. This has to be done also if there are multiple eigenvectors corresponding to the same eigenvalue. We need to identify which eigenvecor of AA^T corresponds to which eigenvector of $A^T A$. We do it by taking v_i as in (2.1). Fortunately, formula (2.1) **ensures** that v_i is an unit eigenvector of $A^T A$ corresponding to λ_i, so in practice we only need to find the eigenvectors corresponding to the 0 eigenvalue and we actually have less computations to do.

4. Then we proceed as in examples Ex 9.1 and 9.2.