# Comprehension Check: Matrix Factorization

In this exercise set, we will be covering a topic useful for understanding matrix factorization: the singular value decomposition (SVD). SVD is a mathematical result that is widely used in machine learning, both in practice and to understand the mathematical properties of some algorithms. This is a rather advanced topic and to complete this exercise set you will have to be familiar with linear algebra concepts such as matrix multiplication, orthogonal matrices, and diagonal matrices.

The SVD tells us that we can **decompose** an $N \times p$ matrix $Y$ with $p < N$ as

$$Y = UDV^\top$$

with $U$ and $V$ **orthogonal** of dimensions $N \times p$ and $p \times p$ respectively and $D$ a $p \times p$ **diagonal** matrix with the values of the diagonal decreasing:

$$d_{1,1} \geq d_{2,2} \geq \ldots d_{p,p}$$

In this exercise, we will see one of the ways that this decomposition can be useful. To do this, we will construct a dataset that represents grade scores for 100 students in 24 different subjects. The overall average has been removed so this data represents the percentage point each student received above or below the average test score. So a 0 represents an average grade (C), a 25 is a high grade (A+), and a -25 represents a low grade (F). You can simulate the data like this:

```
set.seed(1987)
#if using R 3.6 or later, use `set.seed(1987, sample.kind="Rounding")` instead
n <- 100
k <- 8
Sigma <- 64  * matrix(c(1, .75, .5, .75, 1, .5, .5, .5, 1), 3, 3)
m <- MASS::mvrnorm(n, rep(0, 3), Sigma)
m <- m[order(rowMeans(m), decreasing = TRUE),]
y <- m %x% matrix(rep(1, k), nrow = 1) + matrix(rnorm(matrix(n*k*3)), n, k*3)
colnames(y) <- c(paste(rep("Math",k), 1:k, sep="_"),
                 paste(rep("Science",k), 1:k, sep="_"),
                 paste(rep("Arts",k), 1:k, sep="_"))
```

Our goal is to describe the student performances as succinctly as possible. For example, we want to know if these test results are all just a random independent numbers. Are all students just about as good? Does being good in one subject  imply you will be good in another? How does the SVD help with all

this? We will go step by step to show that with just three relatively small pairs of vectors we can explain much of the variability in this $100 \times 24$ dataset.

## Q1

1.0/1.0 point (graded)

You can visualize the 24 test scores for the 100 students by plotting an image:

```
my_image <- function(x, zlim = range(x), ...){
        colors = rev(RColorBrewer::brewer.pal(9, "RdBu"))
        cols <- 1:ncol(x)
        rows <- 1:nrow(x)
        image(cols, rows, t(x[rev(rows),,drop=FALSE]), xaxt = "n", yaxt = "n",
                        xlab="", ylab="",  col = colors, zlim = zlim, ...)
        abline(h=rows + 0.5, v = cols + 0.5)
        axis(side = 1, cols, colnames(x), las = 2)
}

my_image(y)
```

How would you describe the data based on this figure?

- The test scores are all independent of each other.
- The students that are good at math are not good at science.
- The students that are good at math are not good at arts.
- **The students that test well are at the top of the image and there seem to be three groupings by subject.** ✔
- The students that test well are at the bottom of the image and there seem to be three groupings by subject.

Submit

You have used 1 of 2 attempts

Answers are displayed within the problem

## Q2

1.0/1.0 point (graded)
You can examine the correlation between the test scores directly like this:

```
my_image(cor(y), zlim = c(-1,1))
range(cor(y))
axis(side = 2, 1:ncol(y), rev(colnames(y)), las = 2)
```

Which of the following best describes what you see?

- The test scores are independent.
- Test scores in math and science are highly correlated but scores in arts are not.
- There is high correlation between tests in the same subject but no correlation across subjects.
- **There is correlation among all tests, but higher if the tests are in science and math and even higher within each subject.** ✔

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

## Q3

1.0/1.0 point (graded)
Remember that orthogonality means that $U^\top U$ and $V^\top V$ are equal to the identity matrix. This implies that we can also rewrite the decomposition as

$$YV = UD \text{ or } U^\top Y = DV^\top$$

We can think of $YV$ and $U^\top V$ as two transformations of $Y$ that preserve the total variability of $Y$ since $U$ and $V$ are orthogonal.

Use the function `svd` to compute the SVD of `y`. This function will return $U$, $V$, and the diagonal entries of $D$.

```
s <- svd(y)
names(s)
```

You can check that the SVD works by typing:

```
y_svd <- s$u %*% diag(s$d) %*% t(s$v)
max(abs(y - y_svd))
```

Compute the sum of squares of the columns of $Y$ and store them in `ss_y`. Then compute the sum of squares of columns of the transformed $YV$ and store them in `ss_yv`. Confirm that `sum(ss_y)` is equal to `sum(ss_yv)`.

What is the value of `sum(ss_y)` (and also the value of `sum(ss_yv)`)?

175435

✔ **Answer:** 175435

175435

**Explanation**

```
ss_y <- apply(y^2, 2, sum)
ss_yv <- apply((y%*%s$v)^2, 2, sum)
sum(ss_y)
sum(ss_yv)
```

Submit

You have used 1 of 10 attempts

Answers are displayed within the problem

## Q4

1.0/1.0 point (graded)

We see that the total sum of squares is preserved. This is because $V$ is orthogonal. Now to start understanding how $YV$ is useful, plot `ss_y` against the column number and then do the same for `ss_yv`.

What do you observe?

- `ss_y` and `ss_yv` are decreasing and close to 0 for the 4th column and beyond.
- ✔ `ss_yv` is decreasing and close to 0 for the 4th column and beyond.
- `ss_y` is decreasing and close to 0 for the 4th column and beyond.
- There is no discernible pattern to either `ss_y` or `ss_yv`.

✔

**Explanation**

The plots can be made using `plot(ss_y)` and `plot(ss_yv)`. We see that the variability of the columns of $YV$ is decreasing. Furthermore, we see that, relative to the first three, the variability of the columns beyond the third is almost 0.

Submit

You have used 1 of 2 attempts

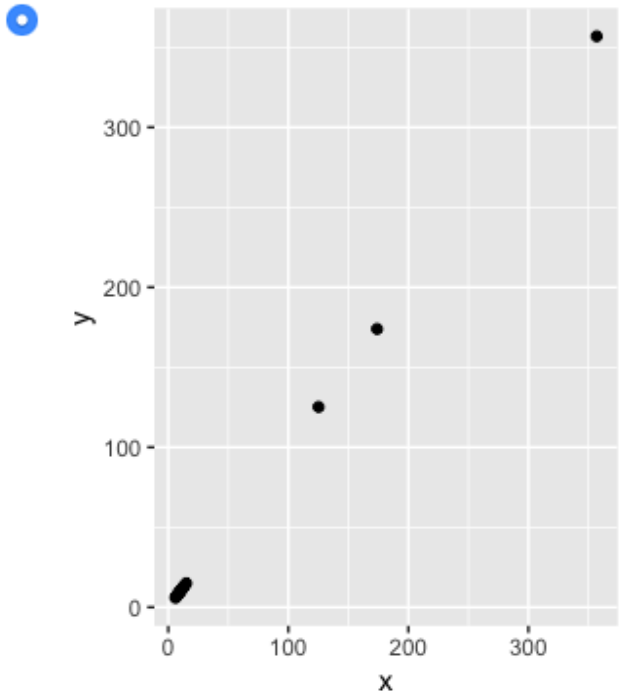
Answers are displayed within the problem

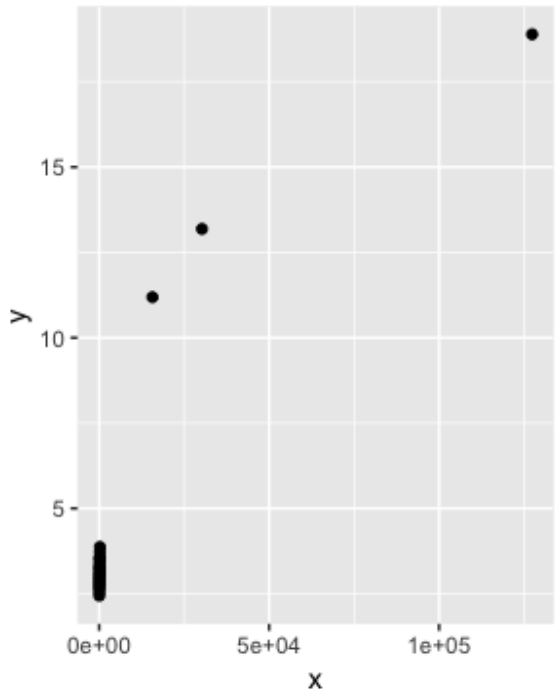
## Q5

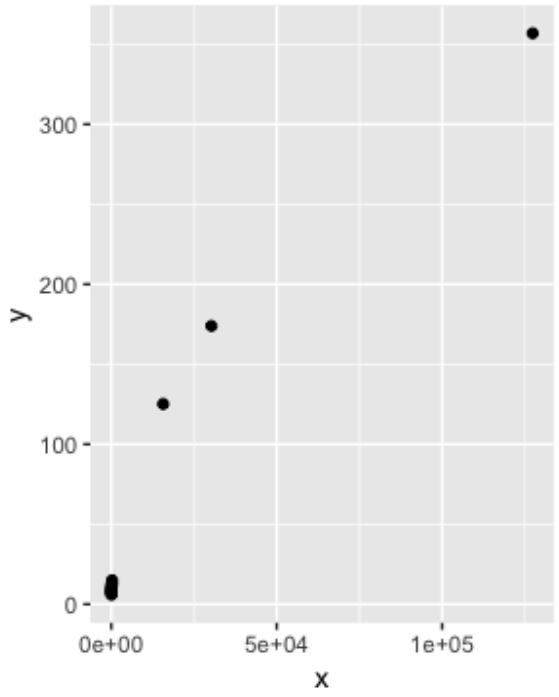
1/1 point (graded)

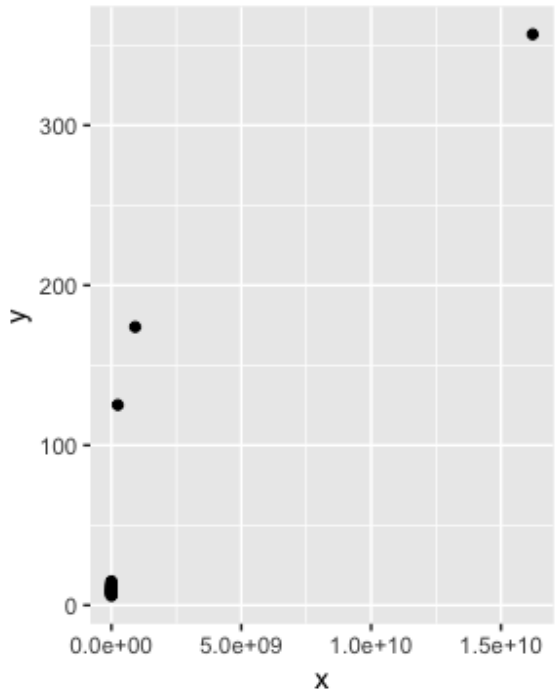
Now notice that we didn't have to compute `ss_yv` because we already have the answer. How? Remember that $YV = UD$ and because $U$ is orthogonal, we know that the sum of squares of the columns of $UD$ are the diagonal entries of $D$ squared. Confirm this by plotting the square root of `ss_yv` versus the diagonal entries of $D$.

Which of these plots is correct?




15
10
5
y
0e+00
5e+04
1e+05
x


300
200
100
0
y
0e+00
5e+04
1e+05
x




✔

**Explanation**
This plot can be generated using the following code:

```
data.frame(x = sqrt(ss_yv), y = s$d) %>%
ggplot(aes(x,y)) +
geom_point()
```

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

## Q6

1.0/1.0 point (graded)
So from the above we know that the sum of squares of the columns of $Y$ (the total sum of squares) adds up to the sum of `s$d^2` and that the transformation $YV$ gives us columns with sums of squares equal to `s$d^2`. Now compute the percent of the total variability that is explained by just the first three columns of $YV$.

What proportion of the total variability is explained by the first three columns of $YV$?
Enter a decimal, **not** the percentage.

0.988 ✔ **Answer:** 0.988

0.988

**Explanation**
The total variability explained can be calculated using the following code: `sum(s$d[1:3]^2) / sum(s$d^2)`. We see that almost 99% of the variability is explained by the first three columns of $YV = UD$. So we get the sense that we should be able to explain much of the variability and structure we found while exploring the data with a few columns.

Submit You have used 1 of 10 attempts

Answers are displayed within the problem

## Q7

1/1 point (graded)
Before we continue, let's show a useful computational trick to avoid creating the matrix `diag(s$d)`. To motivate this, we note that if we write $U$ out in its columns $[U_1, U_2, \ldots, U_p]$ then $UD$ is equal to

$$UD = [U_1 d_{1,1}, U_2 d_{2,2}, \ldots, U_p d_{p,p}]$$

Use the `sweep` function to compute $UD$ without constructing `diag(s$d)` or using matrix multiplication.

Which code is correct?

- `identical(t(s$u %*% diag(s$d)), sweep(s$u, 2, s$d, FUN = "*"))`
- **`identical(s$u %*% diag(s$d), sweep(s$u, 2, s$d, FUN = "*"))`** ✔
- `identical(s$u %*% t(diag(s$d)), sweep(s$u, 2, s$d, FUN = "*"))`
- `identical(s$u %*% diag(s$d), sweep(s$u, 2, s, FUN = "*"))`

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

## Q8

1.0/1.0 point (graded)

We know that $U_1 d_{1,1}$, the first column of $UD$, has the most variability of all the columns of $UD$. Earlier we looked at an image of $Y$ using `my_image(y)`, in which we saw that the student to student variability is quite large and that students that are good in one subject tend to be good in all. This implies that the average (across all subjects) for each student should explain a lot of the variability. Compute the average score for each student, plot it against $U_1 d_{1,1}$, and describe what you find.

What do you observe?

- There is no relationship between the average score for each student and $U_1 d_{1,1}$.
- There is a linearly decreasing relationship between the average score for each student and $U_1 d_{1,1}$.
- **There is a linearly increasing relationship between the average score for each student and $U_1 d_{1,1}$.** ✔
- There is an exponentially increasing relationship between the average score for each student and $U_1 d_{1,1}$.
- There is an exponentially decreasing relationship between the average score for each student and $U_1 d_{1,1}$.

**Explanation**
You can generate the plot using `plot(-s$u[,1]*s$d[1], rowMeans(y))`.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

## Q9

1/1 point (graded)
We note that the signs in SVD are arbitrary because:

$$UDV^\top = (-U)\, D(-V)^\top$$

With this in mind we see that the first column of $UD$ is almost identical to the average score for each student except for the sign.

This implies that multiplying $Y$ by the first column of $V$ must be performing a similar operation to taking the average. Make an image plot of $V$ and describe the first column relative to others and how this relates to taking an average.

How does the first column relate to the others, and how does this relate to taking an average?

- The first column is very variable, which implies that the first column of YV is the sum of the rows of Y multiplied by some non-constant function, and is thus not proportional to an average.
- The first column is very variable, which implies that the first column of YV is the sum of the rows of Y multiplied by some non-constant function, and is thus proportional to an average.
- **The first column is very close to being a constant, which implies that the first column of YV is the sum of the rows of Y multiplied by some constant, and is thus proportional to an average.**
- The first three columns are all very close to being a constant, which implies that these columns are the sum of the rows of Y multiplied by some constant, and are thus proportional to an average.

✔

**Explanation**
The image plot can be made using `my_image(s$v)`.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

The following four exercises are all **ungraded** and are provided to give you an additional opportunity to practice working with matrices in a continuation of the exercises with this dataset.

We recommend that you attempt to write the code on your own **before** hitting "submit" and viewing the answers.

## Q10 - UNGRADED

0 points possible (ungraded)
We already saw that we can rewrite $UD$ as

$$U_1 d_{1,1} + U_2 d_{2,2} + \cdots + U_p d_{p,p}$$

with $U_j$ the j-th column of $U$. This implies that we can rewrite the entire SVD as:

$$Y = U_1 d_{1,1} V_1^\top + U_2 d_{2,2} V_2^\top + \cdots + U_p d_{p,p} V_p^\top$$

with $V_j$ the jth column of $V$. Plot $U_1$, then plot $V_1^\top$ using the same range for the y-axis limits, then make an image of $U_1 d_{1,1} V_1^\top$ and compare it to the image of $Y$. Hint: use the `my_image` function defined above. Use the `drop=FALSE` argument to assure the subsets of matrices are matrices.

**Explanation**
The plot can be made using the following code:

```
plot(s$u[,1], ylim = c(-0.25, 0.25))
plot(s$v[,1], ylim = c(-0.25, 0.25))
with(s, my_image((u[, 1, drop=FALSE]*d[1]) %*% t(v[, 1, drop=FALSE])))
my_image(y)
```

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

## Q11 - UNGRADED

0 points possible (ungraded)
We see that with just a vector of length 100, a scalar, and a vector of length 24, we can actually come close to reconstructing the a $100 \times 24$ matrix. This is our first matrix factorization:

$$Y \approx d_{1,1} U_1 V_1^\top$$

In the exercise in Q6, we saw how to calculate the percent of total variability explained. However, our approximation only explains the observation that good students tend to be good in all subjects. Another aspect of the original data that our approximation does not explain was the higher similarity we observed within subjects. We can see this by computing the difference between our approximation and original data and then computing the correlations. You can see this by running this code:

```
resid <- y - with(s,(u[, 1, drop=FALSE]*d[1]) %*% t(v[, 1, drop=FALSE]))
my_image(cor(resid), zlim = c(-1,1))
axis(side = 2, 1:ncol(y), rev(colnames(y)), las = 2)
```

Now that we have removed the overall student effect, the correlation plot reveals that we have not yet explained the within subject correlation nor the fact that math and science are closer to each other than to the arts. So let's explore the second column of the SVD.

Repeat the previous exercise (Q10) but for the second column: Plot $U_2$, then plot $V_2^\top$ using the same range for the y-axis limits, then make an image of $U_2 d_{2,2} V_2^\top$ and compare it to the image of `resid`.

**Explanation**
The plot can be made using the following code:

```
plot(s$u[,2], ylim = c(-0.5, 0.5))
plot(s$v[,2], ylim = c(-0.5, 0.5))
with(s, my_image((u[, 2, drop=FALSE]*d[2]) %*% t(v[, 2, drop=FALSE])))
my_image(resid)
```

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

## Q12 - UNGRADED

0 points possible (ungraded)

The second column clearly relates to a student's difference in ability in math/science versus the arts. We can see this most clearly from the plot of `s$v[,2]`. Adding the matrix we obtain with these two columns will help with our approximation:

$$Y \approx d_{1,1} U_1 V_1^\top + d_{2,2} U_2 V_2^\top$$

We know it will explain `sum(s$d[1:2]^2)/sum(s$d^2) * 100` percent of the total variability. We can compute new residuals like this:

```
resid <- y - with(s,sweep(u[, 1:2], 2, d[1:2], FUN="*") %*% t(v[, 1:2]))
my_image(cor(resid), zlim = c(-1,1))
axis(side = 2, 1:ncol(y), rev(colnames(y)), las = 2)
```

and see that the structure that is left is driven by the differences between math and science. Confirm this by first plotting $U_3$, then plotting $V_3^\top$ using the same range for the y-axis limits, then making an image of $U_3 d_{3,3} V_3^\top$ and comparing it to the image of `resid`.

**Explanation**

This plot can be made using the following code:

```
plot(s$u[,3], ylim = c(-0.5, 0.5))
plot(s$v[,3], ylim = c(-0.5, 0.5))
with(s, my_image((u[, 3, drop=FALSE]*d[3]) %*% t(v[, 3, drop=FALSE])))
my_image(resid)
```

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

## Q13 - UNGRADED

0 points possible (ungraded)

The third column clearly relates to a student's difference in ability in math and science. We can see this most clearly from the plot of `s$v[,3]`. Adding the matrix we obtain with these two columns will help with our approximation:

$$Y \approx d_{1,1}U_1V_1^\top + d_{2,2}U_2V_2^\top + d_{3,3}U_3V_3^\top$$

We know it will explain: `sum(s$d[1:3]^2)/sum(s$d^2) * 100` percent of the total variability. We can compute new residuals like this:

```
resid <- y - with(s,sweep(u[, 1:3], 2, d[1:3], FUN="*") %*% t(v[, 1:3]))
my_image(cor(resid), zlim = c(-1,1))
axis(side = 2, 1:ncol(y), rev(colnames(y)), las = 2)
```

We no longer see structure in the residuals: they seem to be independent of each other. This implies that we can describe the data with the following model:

$$Y = d_{1,1}U_1V_1^\top + d_{2,2}U_2V_2^\top + d_{3,3}U_3V_3^\top + \varepsilon$$

with $\varepsilon$ a matrix of independent identically distributed errors. This model is useful because we summarize of $100 \times 24$ observations with $3 \times (100 + 24 + 1) = 375$ numbers.

Furthermore, the three components of the model have useful interpretations:

1 - the overall ability of a student

2 - the difference in ability between the math/sciences and arts

3 - the remaining differences between the three subjects.

The sizes $d_{1,1}$, $d_{2,2}$ and $d_{3,3}$ tell us the variability explained by each component. Finally, note that the components $d_{j,j}U_jV_j^\top$ are equivalent to the jth principal component.

Finish the exercise by plotting an image of $Y$, an image of $d_{1,1}U_1V_1^\top + d_{2,2}U_2V_2^\top + d_{3,3}U_3V_3^\top$ and an image of the residuals, all with the same `zlim`.

**Explanation**

These plots can be made using the following code:

```
y_hat <- with(s,sweep(u[, 1:3], 2, d[1:3], FUN="*") %*% t(v[, 1:3]))
my_image(y, zlim = range(y))
my_image(y_hat, zlim = range(y))
my_image(y - y_hat, zlim = range(y))
```

Submit

You have used 1 of 1 attempt

Answers are displayed within the problem

Ask your questions or make your comments about Matrix Factorization here! **Remember, one of the best ways to reinforce your own learning is by explaining something to someone else, so we encourage you to answer each other's questions (without giving away the answers, of course).**

Some reminders:

- Search the discussion board before posting to see if someone else has asked the same thing before asking a new question.
- Please be specific in the title and body of your post regarding which question you're asking about to facilitate answering your question.
- Posting snippets of code is okay, but posting full code solutions is not.
- If you do post snippets of code, please format it as code for readability. If you're not sure how to do this, there are instructions in a pinned post in the "general" discussion forum.

## Discussion: Matrix Factorization

Show Discussion

**Topic:** Section 6: Model fitting and recommendation systems / 6.3.2: Matrix Factorization