Quizzes due May 4, 2021 15:01 +03

Assignment Prep

Load the **GSE5859Subset** data:

```
library(GSE5859Subset)
data(GSE5859Subset)
```

Question 1

1/1 point (graded)

Which of the following are true about the singular value decomposition matrices $\mathbf{YV} = \mathbf{UD}$?

Check ALL correct answers.

- ☑ $\mathbf{Y}$ is the original data matrix
- ☑ $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices
- ☑ $\mathbf{D}$ is a diagonal matrix with decreasing values
- □ This equation can also be written as $\mathbf{Y} = \mathbf{U}^\top \mathbf{DV}$
- ☑ $\mathbf{UD}$ are the new coordinates for the projection $\mathbf{YV}$

✓

Explanation

Option D is false: $\mathbf{Y} = \mathbf{UDV}^\top$. All other choices are correct.

Submit

You have used 1 of 5 attempts

ℹ Answers are displayed within the problem

Question 2

3/3 points (graded)

Compute the SVD of `geneExpression`. Save it as the variable `s`.

A: What is the first entry of `s$d`?

2794.669

✓ **Answer:** 2794.669

2794.669

B: The proportion of variabiilty in the data explained by the `x` th column of $\mathbf{U}$ is equal to `s$d[x]^2` divided by the sum of all `s$d` values. What proportion of variability is explained by the first column of $\mathbf{U}$?

Report your answer as a value between 0 and 1.

0.9980186

✓ **Answer:** 0.9980186

0.9980186

C: Compute the mean of each row of `geneExpression` as a vector `m`. What is the correlation between `m` and the first column of `s$u`?

-1

✓ **Answer:** -1

-1

Explanation

Part A:

```
s = svd(geneExpression)
s$d[1]
```

```
## [1] 2794.669
```

Part B:

```
s$d[1]^2/(sum(s$d^2))
```

```
## [1] 0.9980186
```

Part C:

```
m = rowMeans(geneExpression)
cor(m, s$u[,1])
```

```
## [1] -1
```

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

Question 3

1/1 point (graded)

Which of the following are true about the singular value decomposition matrices $\mathbf{YV} = \mathbf{UD}$?

Check ALL correct answers.

- □ Over 99.9% of the variability in $\mathbf{U}$ is explained by the row means of `geneExpression`.
- ☑ The row means are almost perfectly correlated with the first column of $\mathbf{U}$ aside from a sign change.
- □ The row means are completely uncorrelated with the first column of $\mathbf{U}$.
- ☑ Most of the variability in the gene expression matrix is driven by average expression levels of each gene rather than biological differences between samples.
- ☑ Removing the row means before computing the SVD would help reveal the underlying biological signal.

✓

Explanation

Option A is false, but it's very close. 99.8% of the variability is explained.

Option C is false. A correlation of 0 is no correlation, but a correlation of -1 or 1 is almost perfect correlation.

All other options are true.

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

Question 4

4/4 points (graded)

Define `y` as `geneExpression - rowMeans(geneExpression)`, then compute the SVD of `y` and save the result as `s`.

A: What is the first entry of `s$d`?

58.42845

✓ **Answer:** 58.42845

58.42845

B: What proportion of variability is explained by the first column of $\mathbf{U}$?

Report your answer as a value between 0 and 1.

0.2196728

✓ **Answer:** 0.2196728

0.2196728

C: Calculate the proportion of variability explained by each column of $\mathbf{U}$. How many individual columns explain more than 5% of the variability?

6

✓ **Answer:** 6

6

D: What percent of variability is explained by the first 10 columns of $\mathbf{U}$?

0.772049

✓ **Answer:** 0.772049

0.772049

Explanation

Part A:

```
y = geneExpression - rowMeans(geneExpression)
s = svd(y)
s$d[1]
```

```
## [1] 58.42845
```

Part B:

```
s$d[1]^2/sum(s$d^2)
```

```
## [1] 0.2196728
```

Part C:

```
sum(s$d^2/sum(s$d^2) > .05)
```

```
## [1] 6
```

Part D:

```
sum(s$d[1:10]^2/sum(s$d^2))
```

```
## [1] 0.772049
```

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

Question 5

1/1 point (graded)

Confirm that $\mathbf{Y} = \mathbf{UDV}^\top$. First, multiply the matrices within `s` to regenerate the data matrix. Then, subtract the regenerated matrix from the original matrix to find residuals. Find and report the residual with the maximum absolute value to show the matrices are identical to within numerical error.

What is the value of the residual with the maximum absolute value?

Copy and paste your answer exactly from R.

9.967374e-14

✓ **Answer:** 9.967374e-14

9.967374×10^{-14}

Explanation

```
y2 = s$u %*% diag(s$d) %*% t(s$v)
resid = y - y2
max(abs(resid))
```

```
## [1] 9.967374e-14
```

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

Question 6

4/4 points (graded)

Let `z = s$d * t(s$v)`. Compare the distance between columns 1 and 2 in `geneExpression` (the original matrix), `y` (the de-trended matrix with row means subtracted), and `z`.

A: What is the distance between columns 1 and 2 in `geneExpression`?

30.15294

✓ **Answer:** 30.15294

30.15294

B: What is the distance between columns 1 and 2 in `y`?

30.15294

✓ **Answer:** 30.15294

30.15294

C: What is the distance between columns 1 and 2 in `z`?

30.15294

✓ **Answer:** 30.15294

30.15294

D: What is the distance between columns 1 and 2 in `z` using only the first 10 columns as an approximation?

22.66855

✓ **Answer:** 22.66855

22.66855

Explanation

Part A:

```
sqrt(crossprod(geneExpression[,1]-geneExpression[,2]))
```

```
##          [,1]
## [1,] 30.15294
```

Part B:

```
sqrt(crossprod(y[,1]-y[,2]))
```

```
##          [,1]
## [1,] 30.15294
```

Part C:

```
z = s$d * t(s$v)
sqrt(crossprod(z[,1]-z[,2]))
```

```
##          [,1]
## [1,] 30.15294
```

Part D:

```
sqrt(crossprod(z[1:10,1]-z[1:10,2]))
```

```
##          [,1]
## [1,] 22.66855
```

Submit

You have used 1 of 5 attempts

Answers are displayed within the problem

Question 7

1/1 point (graded)

Perform MDS on the original `geneExpression` data:

```
d = dist(t(geneExpression))
mds = cmdscale(d)
```

Make an MDS plot using `mds` and color the points by `fdate = factor(sampleInfo$date)`.

Which of the following is true about the relationship of date to the MDS plot?

○ The second dimension of the MDS plot appears to correlate with date, with earlier dates at the bottom and later dates at the top.

○ The second dimension of the MDS plot appears to correlate with date, with earlier dates at the top and later dates at the bottom.

● The first dimension of the MDS plot appears to correlate with date, with earlier dates to the left and later dates to the right

○ The first dimension of the MDS plot appears to correlate with date, with earlier dates to the right and later dates to the left.

○ There is no apparent relationship between date and the two dimensions of the MDS plot.

✓

Submit

You have used 1 of 2 attempts

✓ Correct (1/1 point)