DATA 100: Vitamin 5 Solutions

March 2, 2019

1 Linear Algebra

1.1 Orthogonal and orthonormal vectors

Are the following vectors orthogonal, orthonormal or neither?
[1/2, 1/5], [-1/2, 5/4]

- ☑ Orthogonal
- ☐ Orthonormal
- ☐ Neither

Explanation: The dot product of these two vectors is 0, but their lengths are not equal to 1. Therefore, they are orthogonal.

1.2 Matrix Algebra

If Q is an m x m matrix and $Q^TQ = I$, which of the following are possible?

- ☑ Q is the identity matrix.
- ☑ Q is orthonormal.
- ☐ Q is not orthonormal.

Explanation: Since the transpose of the identity matrix and the multiplication of the identy matrix by itself are equal to the identity matrix, the first option is correct. The second option is also correct by the definition of the orthonormal matrix: all columns have length 1, and all columns are orthogonal.

2 Principal Component Analysis

2.1 True or False?

The singular value for the first principal component (PC) is always (select all that apply):

- ☑ Greater than or equal to zero
- ☐ Greater than one
- ☐ Larger than the other singular values
- ☑ At least as large as the other singular values
- ☐ None of these

Explanation: PCA identifies the PC of a data set in decreasing order of variance, which is defined as the PC's eigenvalue. The first PC has the largest or is equal to the largest variance of all PCs. Option two is incorrect since the lower bound on the variance explained by any PC is 0. Option three is incorrect since the first PC has the largest or equal to the largest variance of all PCs. Therefore, option four is correct. Note that the singular value of the first PC is its standard deviation.

2.2 Scree plots

Given that that PCA was applied to some data set, which of the following statement(s) describe the use of a scree plot?

- ☑ Compare the amount of total variance explained by each PC.
- ☐ Identify clusters in the data.
- ☑ Assess whether a 2-d scatter plot is an appropriate visualization of the data.
- ☐ Visualize the data using a lower-dimensional projection.
- ☐ Identify the most informative variables in the data.

Explanation: A scree plot illustrates the variance explained by the largest PCs. This visualization is used to compare the eigenvalues of the PCs, to assess the dimensionality reducing ability of PCA and to identify the correct number of PCs to create the most informative lower dimensional representation of the original data set. A scatter plot should be used to visualize a lower dimensional representation of the data. The loadings vectors of the PCs should be used to identify the most informative variables.

2.3 Centering and Scaling

For which of the following reason(s) are the variables (i.e. the columns) of a data set centered and scaled prior to using PCA?

- ☐ Singular value decomposition only applies to centered matrices
- ☐ Centering and scaling removes outliers from the data
- ☑ The singular values only express the total variance of the data after centering and scaling

Explanation: Centering and scaling the variables of a data set prior to using PCA ensures that the variables are comparable. This process does not remove outliers and is not required for the using singular value decomposition.