Stat501_Homework1

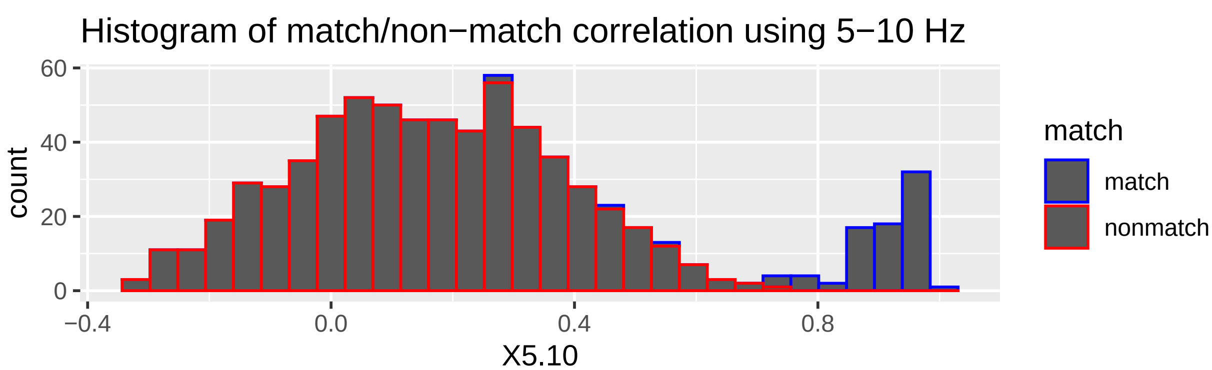
Kelby Kies

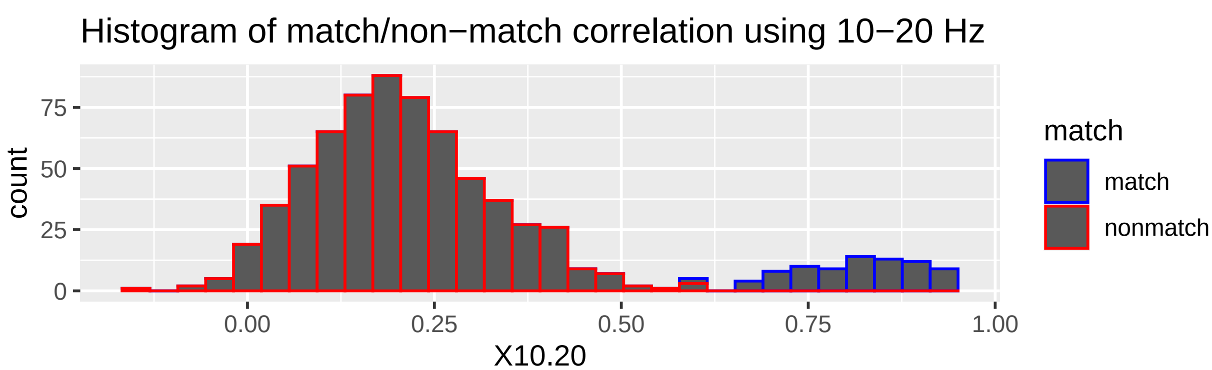
2/17/2021

Question #1:

Part A:

I:: For each frequency band,make histograms of the correlations of the match and non-match pairs in one figure.

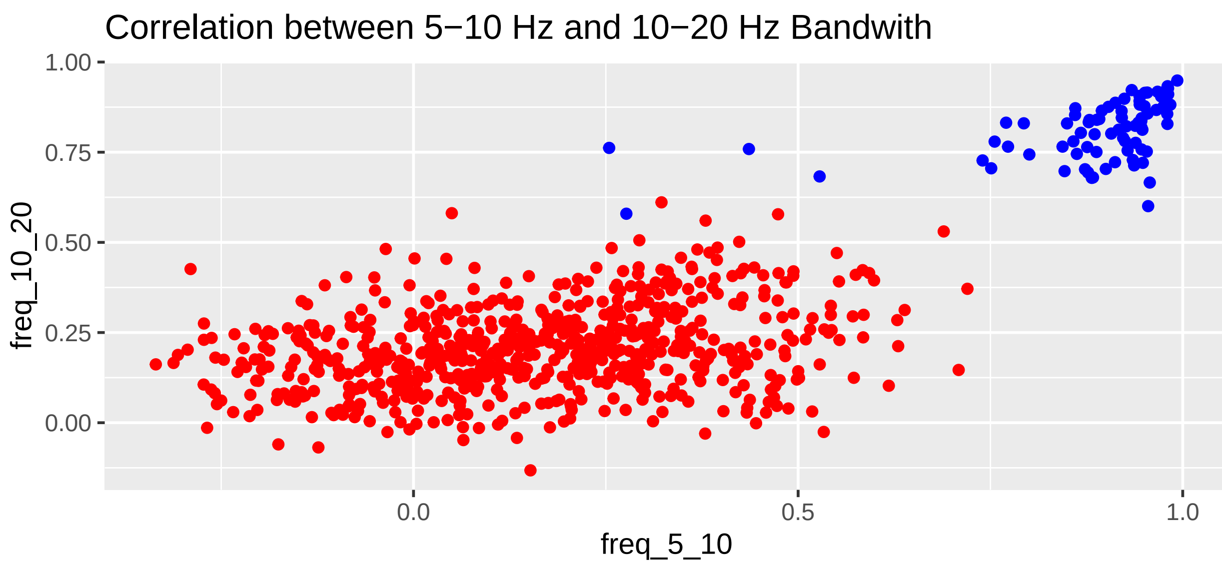




Comment on the separation of the matches and non-matches.:

The majority of images have a low correlation which makes sense because the majority of the image pairs (knife tip/blade) will not match. However, you can see that image pairs that have a high correlation and those values are coordinate to the matched knife tip to knife base. We see that matches/nonmatches are not completely separated, but for the most part there is clear separation between matches/nonmatches.

II::For each pair,and using color to denote match/non-match, plot the correlation between the 5-10 and 10-20 band pairs.

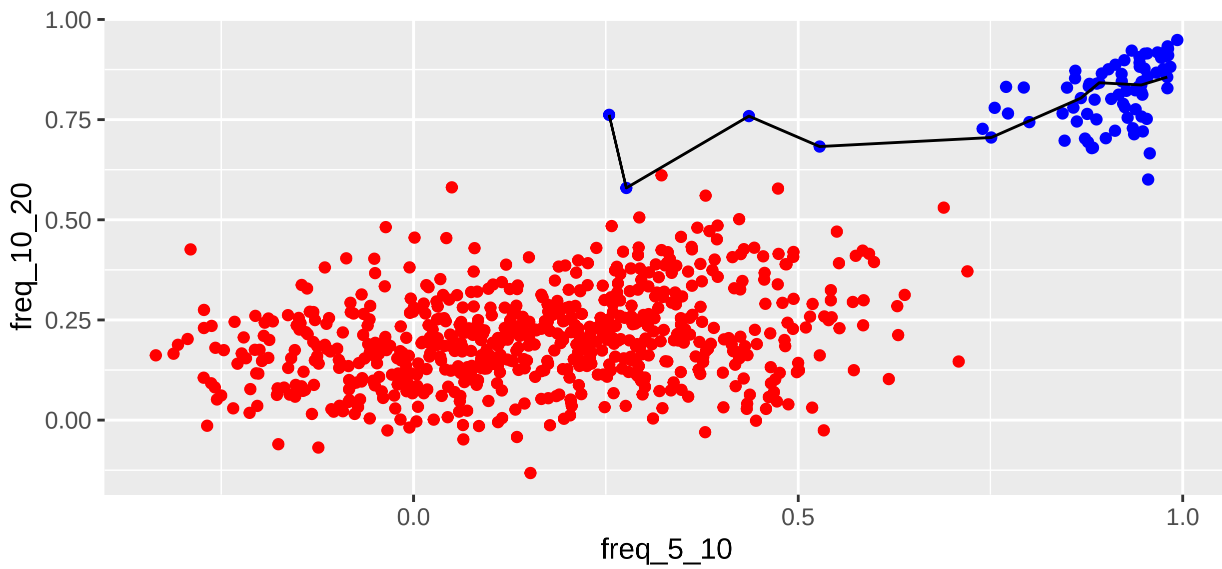


Comment on the separation of the matches/non-matches in the bivariate plot vis-a-vis the univariate plots:

Both plots show the separation of the matches from the non-matches. The univariate plots from part A.i. show a bimodal curve indicating matching image pairs have a high correlation, while non-matches have a low correlation. These plots also might be good for comparing the differences between the frequencies.

The bivariate plot shows that the matches(blue) and non-matches(red) are not as separated as we may think. It displays a positive correlation between the two frequency bands. For the most part the two frequency bands are separated, but there are a few matches that seem almost like potential outliers.

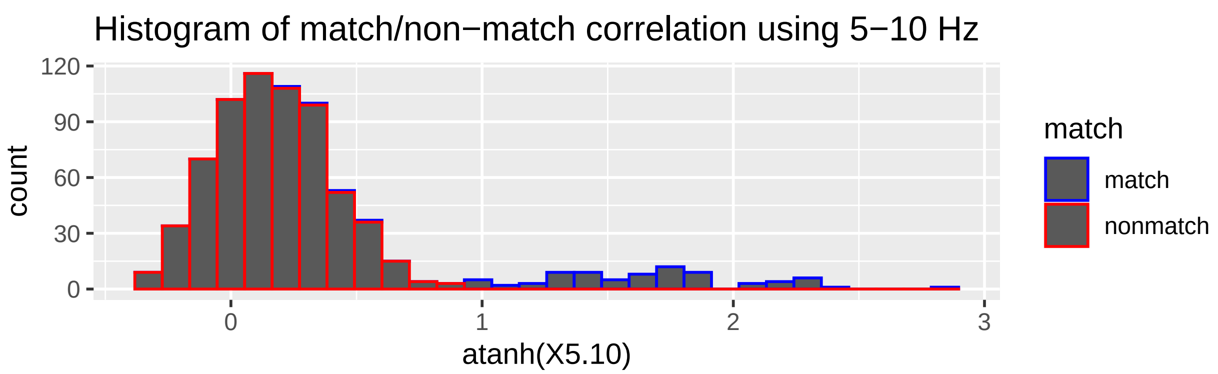
III:: Connect all the 9 correlation pairs from this set by means of a line

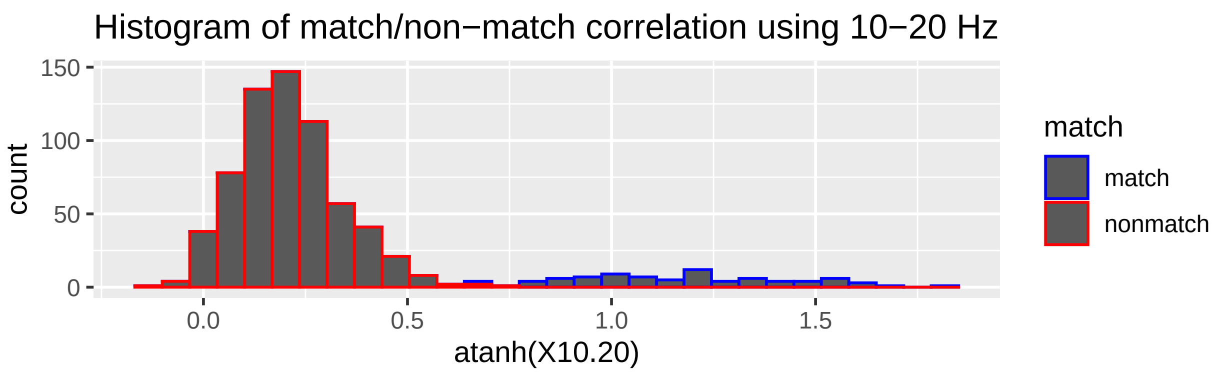


Comment on separation when all 9 correlation pairs are considered together: I think that when all 9 pairs are considered it shows the separation between matches/nonmatches better than before. Although the line is not as clean cut as we would like to see.

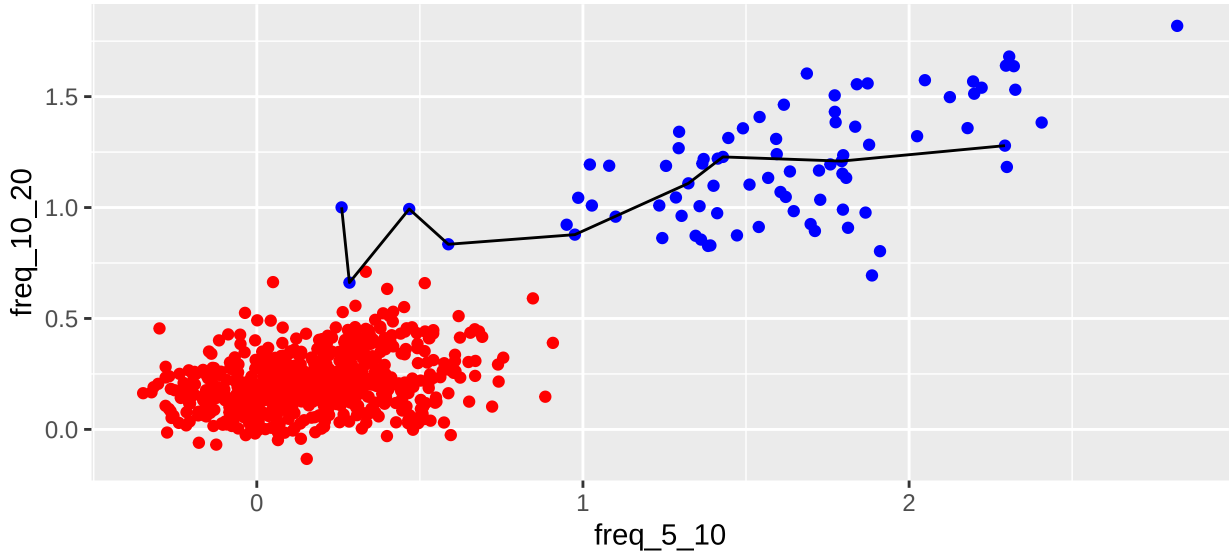
Part B:

The inverse hyperbolic tangent transformation is often used on correlations to stabilize its variance. Repeat the exercise in the previous part, but use the inverse hyperbolic tangent transformed correlations.Comment.





Comment: By using the inverse hyperbolic tangent transformed correlations it really helped show the difference between the matches/nonmatches. The inverse hyperbolic tangent transformation really stretched out the variance of the match knife tip/blade pairs, and condensed the variance of nonmatches.



Comment:The matched image pairs seem to be more spread out than before. The nonmatches are clustered tighter than before which helps exaggerate the separation between matches/nonmatches. The line here looks similar to what we had before except it is stretched along with the variance of the match pairs.

Question #2:

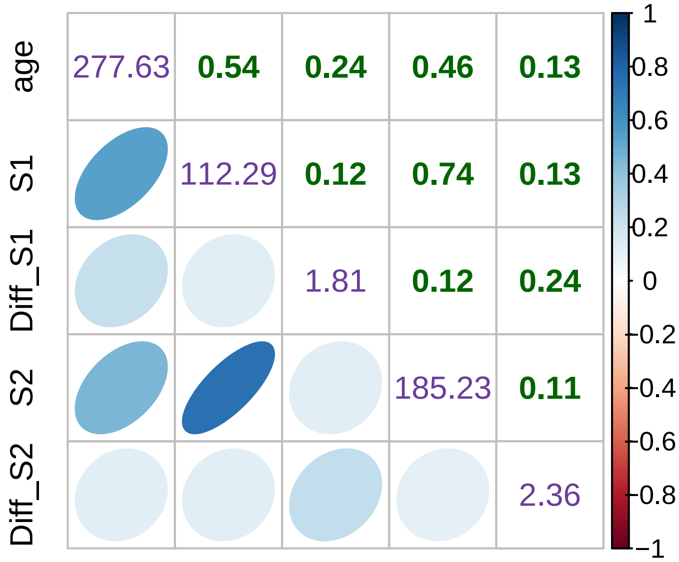
Part A:

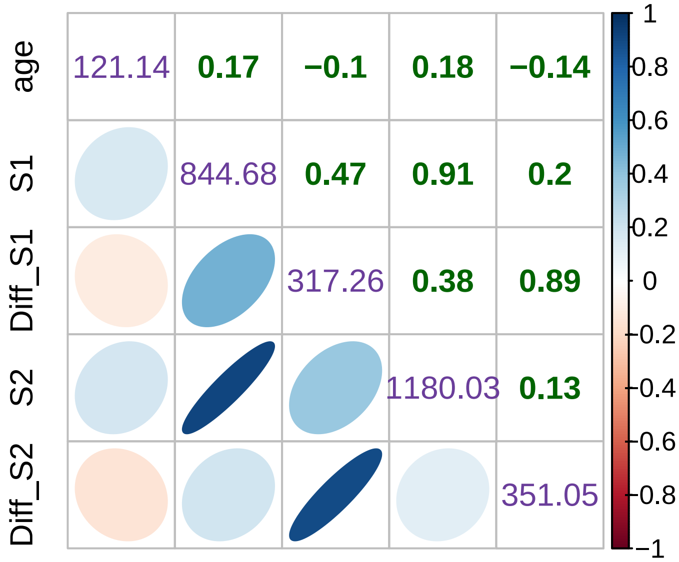
```
## [1] "Calculate the means for each group:"
```

```
##         sclerosis_mean normal_mean
## age          37.985507    42.06897
## S1          147.289855   178.26897
## Diff_S1       1.562319    12.27586
## S2          195.602899   236.93103
## Diff_S2       1.620290    13.08276
```

Part B:

Use the plotcorr function to display the correlation matrix for each group. Comment.





Comment: We can see that all variables have a positive correlation while Diff_S1/age and Diff_S2/age are negatively correlated. S2/S1 & Diff_S2/Diff_S1 have a very strong correlation while Diff_S1/age, Diff_S2/age, S1/age, S2/age, Diff_S2/S1 and Diff_S2/S2 have a very weak correlation.

Question 3:

Part A:

Read in the TIFF image and convert to a dataset of 375000 3-dimensional observations.

```
## [1] "Top 10 rows of RGB Values matrix"
```

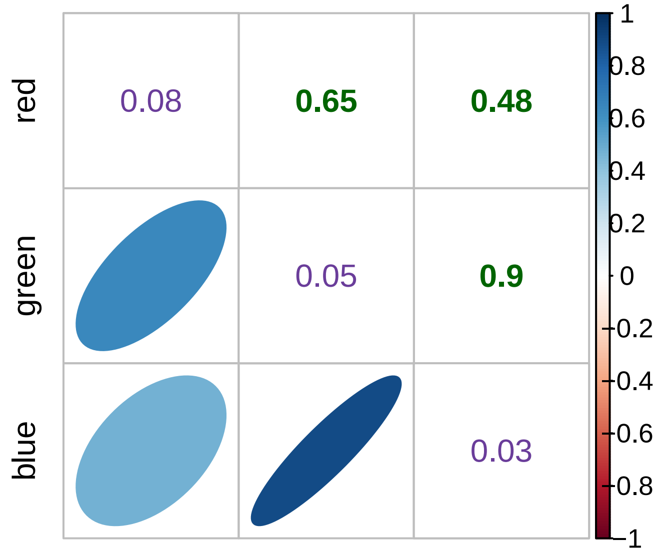
```
##            red     green       blue
## [1,] 0.7960784 0.1294118 0.03137255
## [2,] 0.7764706 0.1176471 0.01568627
## [3,] 0.7803922 0.1333333 0.02352941
## [4,] 0.8117647 0.1647059 0.03921569
## [5,] 0.8313725 0.1764706 0.04313725
## [6,] 0.8274510 0.1490196 0.01568627
```

```
## [1] "Dimensions of RGB Values matrix"
```

```
## [1] 375000      3
```

Part B:

Use the plotcorr function to display the correlation matrix between the three colors. Comment.



Comment: We can see that all colors have a positive correlation with each other. Blue and Green have a very strong correlation, while the lowest correlation was between blue and red.

Question 4:

Show that the $d^2(\mathbf{X}^\circ, \mathbf{Y}^\circ) = 2 - 2 * Corr(\mathbf{X}, \mathbf{Y})$

Definitions:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ X_p \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_p \end{bmatrix}$$

$$\mathbf{X}^\circ = \frac{(\mathbf{X} - \mathbf{X}' \cdot \mathbf{1} \cdot \mathbf{1}/p)}{||(\mathbf{X} - \mathbf{X}' \cdot \mathbf{1} \cdot \mathbf{1}/p)||} = \frac{(\mathbf{X} - \overline{\mathbf{X}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2}}$$

$$\mathbf{Y}^\circ = \frac{(\mathbf{Y} - \mathbf{Y}' \cdot \mathbf{1} \cdot \mathbf{1}/p)}{||(\mathbf{Y} - \mathbf{Y}' \cdot \mathbf{1} \cdot \mathbf{1}/p)||} = \frac{(\mathbf{Y} - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(Y_i - \overline{\mathbf{Y}})^2}}$$

Show that the $d^2(\mathbf{X}^\circ, \mathbf{Y}^\circ) = 2 - 2 * Corr(\mathbf{X}, \mathbf{Y})$

$$d^2(\mathbf{X}^\circ, \mathbf{Y}^\circ) = \left(\sqrt{\sum_{j=1}^{p}(X_j^\circ - Y_j^\circ)^2}\right)^2 = \sum_{j=1}^{p}(X_j^\circ - Y_j^\circ)^2$$

Plug in $\mathbf{X}^\circ$ & $\mathbf{Y}^\circ$.

$$= \sum_{i=1}^{p} (\frac{(X_i - \overline{\mathbf{X}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2}} - \frac{(Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(Y_i - \overline{\mathbf{Y}})^2}})^2$$

$$= \sum_{i=1}^{p} (\frac{(X_i - \overline{\mathbf{X}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2}})^2 - 2 \cdot \frac{(X_i - \overline{\mathbf{X}}) \cdot (Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2} \cdot \sqrt{\sum_{i=1}^{p}(Y_i - \overline{\mathbf{Y}})^2}} + (\frac{(Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(Y_i - \overline{\mathbf{Y}})^2}})^2$$

$$= \sum_{i=1}^{p} (\frac{(X_i - \overline{\mathbf{X}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2}})^2 - 2 \cdot \frac{(X_i - \overline{\mathbf{X}}) \cdot (Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2 \cdot (Y_i - \overline{\mathbf{Y}})^2}} + (\frac{(Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(Y_i - \overline{\mathbf{Y}})^2}})^2$$

$$= \sum_{i=1}^{p} \frac{(X_i - \overline{\mathbf{X}})^2}{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2} - 2 \cdot \sum_{i=1}^{p} \frac{(X_i - \overline{\mathbf{X}}) \cdot (Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2 \cdot (Y_i - \overline{\mathbf{Y}})^2}} + \sum_{i=1}^{p} \frac{(Y_i - \overline{\mathbf{Y}})^2}{\sum_{i=1}^{p}(Y_i - \overline{\mathbf{Y}})^2}$$

$$= 1 - 2 \cdot \sum_{i=1}^{p} \frac{(X_i - \overline{\mathbf{X}}) \cdot (Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2 \cdot (Y_i - \overline{\mathbf{Y}})^2}} + 1$$

$$= 2 - 2 \cdot \sum_{i=1}^{p} \frac{(X_i - \overline{\mathbf{X}}) \cdot (Y_i - \overline{\mathbf{Y}})}{\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2 \cdot (Y_i - \overline{\mathbf{Y}})^2}}$$

$$= 2 - 2 \cdot \frac{(\frac{1}{p-1}) \sum_{i=1}^{p}(X_i - \overline{\mathbf{X}}) \cdot (Y_i - \overline{\mathbf{Y}})}{(\frac{1}{p-1})\sqrt{\sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2 \cdot (Y_i - \overline{\mathbf{Y}})^2}}$$

$$= 2 - 2 \cdot \frac{(\frac{1}{p-1}) \sum_{i=1}^{p}(X_i - \overline{\mathbf{X}}) \cdot (Y_i - \overline{\mathbf{Y}})}{\sqrt{(\frac{1}{p-1})^2 \sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2 \cdot (Y_i - \overline{\mathbf{Y}})^2}}$$

$$= 2 - 2 \cdot \frac{s_{xy}}{\sqrt{(\frac{1}{p-1}) \sum_{i=1}^{p}(X_i - \overline{\mathbf{X}})^2}\sqrt{(\frac{1}{p-1}) \sum_{i=1}^{p}(Y_i - \overline{\mathbf{Y}})^2}}$$

$$= 2 - 2 \cdot \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

$$= 2 - 2 \cdot Corr(\mathbf{X}, \mathbf{Y})$$