Covariance and Correlation

Parthiban Rajendran

November 3, 2018

```
The tikzmagic extension is already loaded. To reload it, use:
  %reload_ext tikzmagic
```

1 Visualization

Now that we have seen TIA is already doing a good job on giving us a measure of the linearity, we shall come to the core of this section. We have not yet visualized the totality of the rectangles. We could have done this earlier, but I wanted to instill a strong sense of what rectangles are we dealing with and why they are whole representative of the dataset though we have taken only half of all possible rectangles. We initially decided how do we color the rectangles, based on positive or negative relationship as a convention, and then looked in detail, what are the rectangles to be plotted. Let us consider the sample sets as below. Recall these were the same sample sets we saw in the beginning of this section. Note the TIA is already calculated indicating us the kind of relationship.

The Sample Sets

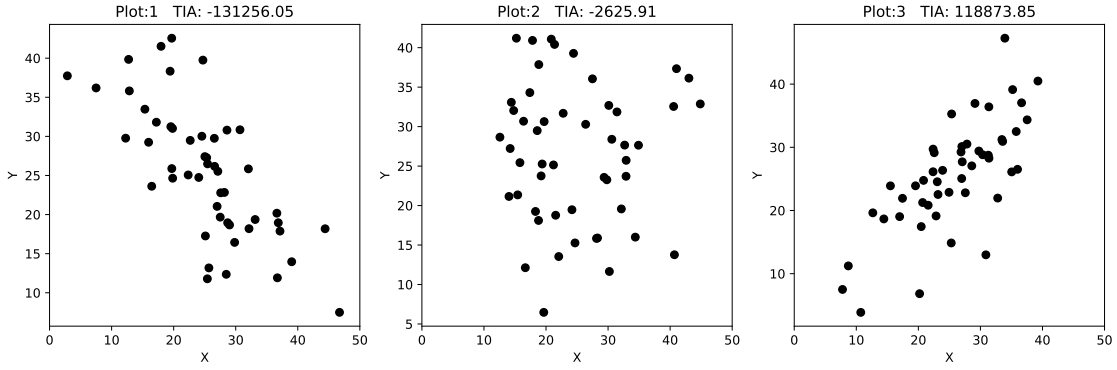

Figure 1: The Sample Sets

As per TIA from Figure 2, Plot 1 is highly negatively correlated, Plot 2 is somewhat negative, and Plot 3 is positively correlated. Though visually Plots 1 and 3 look like not having much difference in their *slope* or *rate*, our TIA gives a wide difference in value. This is because, TIA between sample sets are not comparable (we will solve that soon in correlation, but remember this problem). That is, given a sample set, say Plot 1, having -148859 is one of infinite no of possibilities among that sample set, with perfectly linear positive, negative and 0 TIA as one of those. Simiarly for sample set in Plot 2 and so on. Below are the sample sample plots with colored rectangles laid over them. Remember, if N is the size of sample set, or no of (x, y) pairs, then the number of rectangles we have drawn is $N(N-1)/2$. And as we already saw, only because of this limited rectangles, we get the output as below without neutralization issues.

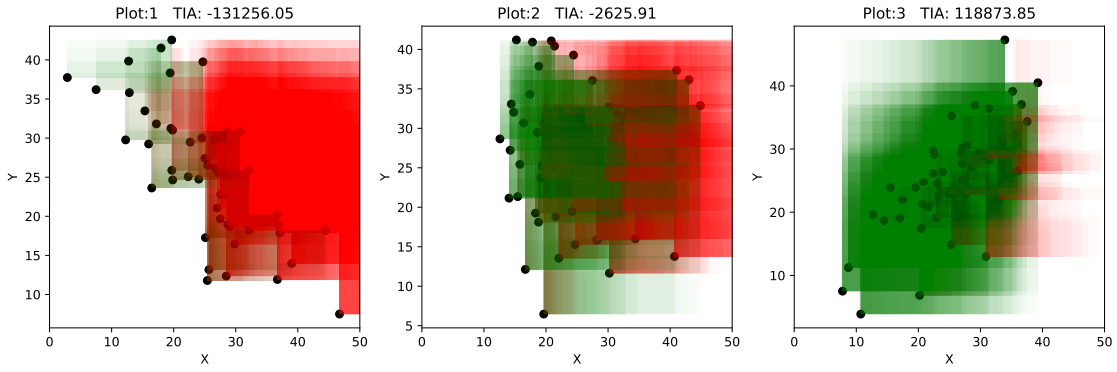

Figure 2: The Visualization of Covariance

I think, Figure 2 speaks for itself :) Plot 1,which has highly negative linear relationship among its sample sets, has more red rectangles than green. Plot 2, which is very less linearity in any direction, shows an almost equal mix of red and green, of course the accurate measure is reflected in its TIA though. Plot 3, which has a positive linear relationship, obviously has lot more green. Figure 3 gives total area of red and green separately, giving us better glimpse of the *net* relationship underneath. The TIA is just the difference between the total green area and red area.

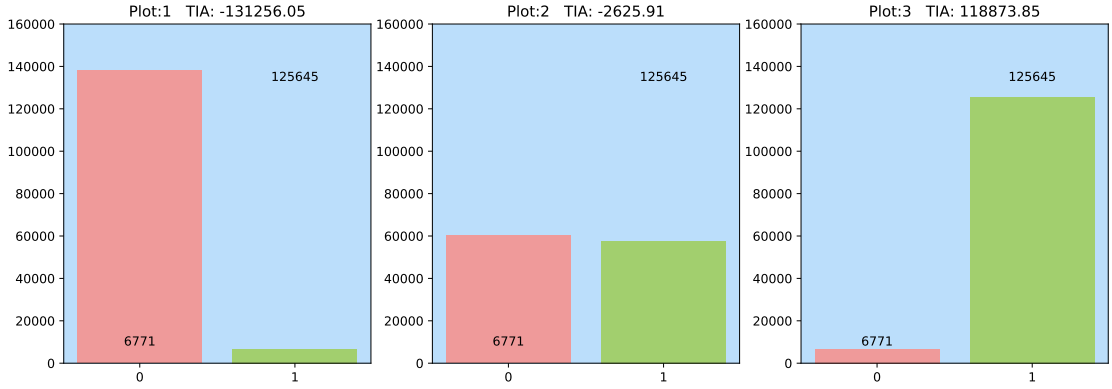

Figure 3: The separated total area
Green indicates Positive

2 Expected value of TIA

For any given sample set, we are typically interested not in the total of the sample set, but most probable or best representative candidate of that sample set. In our case, our sample set of TIA, is not individual pairs (x_i, y_i), but a function of them, a product $(x_i - x_j)(y_i - y_j)$. That is, using 2 if,

$$h(X,Y) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)$$

then, we are interested in $E[h(X,Y)]$
As per expectation formula,

$$E[h(X,Y)] = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)p(x_i, y_i) \tag{1}$$

Note, we are not interested in expected value of *number of rectangles* or *red colored rectangles* etc. The area of rectangles carry the measure and each rectangle might have different area. We are thus interested in the *expected value* of the area, given the *total* interested area.

Expectation needs a *joint probability mass function* $p(X, Y)$ associated with $h(X, Y)$. Recall the rectangle graph for $N = 6$ and replace with area A_{ij} (could also call as product, P_{ij} but just to avoid notational confusion with probability let us stick with area).

y_1	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}
y_2	A_{21}	A_{22}	A_{23}	A_{24}	A_{25}	A_{26}
y_3	A_{31}	A_{32}	A_{33}	A_{34}	A_{35}	A_{36}
y_4	A_{41}	A_{42}	A_{43}	A_{44}	A_{45}	A_{46}
y_5	A_{51}	A_{52}	A_{53}	A_{54}	A_{55}	A_{56}
y_6	A_{61}	A_{62}	A_{63}	A_{64}	A_{65}	A_{66}
y/x	x_1	x_2	x_3	x_4	x_5	x_6

$N = 6$,
Total rectangles $= N^2$
TIA $= \sum_{i=1}^{N} \sum_{j=i+1}^{N} A_{ij}$

Figure 4: The Number of Area Components

Assuming each *area* has equal probability, given the number of area, each A_{ij} will have a probability of $\frac{1}{N^2}$ as there are N^2 area components possible. Thus, 1 becomes,

$$E[h(X,Y)] = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)p(x_i, y_i) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j) \tag{2}$$

Ladies and Gentlemen. That $E[h(X, Y)]$ is called **Covariance** of X and Y , shortly called $\text{Cov}(\mathbf{X}, \mathbf{Y})$. Also note, the alternative form we saw earlier in equation 3, could also be used to derive covariance as below.

$$E[h(X,Y)] = \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j)p(x_i, y_i) = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)(y_i - y_j) \tag{3}$$

Covariance of discrete X and Y with p(X Y) uniform

Given X and Y are discrete variables of sample size N, and $p(X, Y) = \frac{1}{N^2}$,

$$\text{Cov}(X,Y) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=i+1}^{N} (x_i - x_j)(y_i - y_j) \tag{4}$$

$$\text{Cov}(X,Y) = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_j)(y_i - y_j) \tag{5}$$

3 Standard Formula

What we have seen so far, is a deformed form of covariance which numerically gave us the same results as a standard formula. It is mathematically possible to show that,

$$\mathrm{Cov}(X,Y) = \sum_{i=1}^{N}\sum_{j=i+1}^{N}(x_i - x_j)(y_i - y_j)p(x_i,y_i) = \sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})p(x_i,y_i) \tag{6}$$

The derivation is proven by yuli2012. At the time of this writing, the doubts in the derivation is not yet cleared, if and once it is done, this section should be enriched with a proper derivation. Till then, this is a discontinuity in our understanding. The visualization of standard formula is slightly different because it involves mean, so all rectangles have one corner at mean position $(\overline{x},\overline{y})$. The visualization is shown in figure 5. The top 3 rows from our deformed formula and bottom 3 using standard formula. One could observe, the rectangles in plots 3,4,and 5 are centered around the mean (shown in dotted lines), thus giving a better viusal perception of the measure (no of red or green rectangles, which is more). We did not start with this visualization only because, there was no intuition to introduce mean in the equation out of no where.

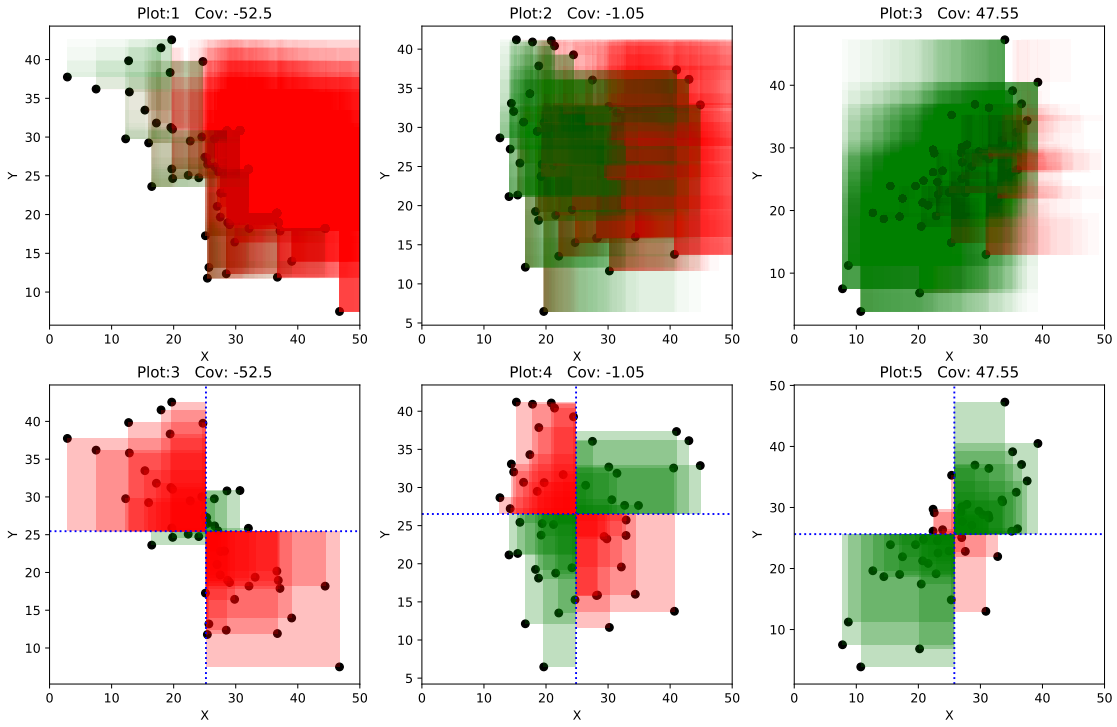

Figure 5: The Visualization of deformed and standard formula for Covariance