Theoretical Background: Negative Correlations and Equality of Means

Chris Bentz

04 February 2021

Session Info

Give the session info (reduced).

```
## [1] "R version 3.6.3 (2020-02-29)"
```

```
## [1] "x86_64-pc-linux-gnu"
```

Load Libraries

If the libraries are not installed yet, you need to install them using, for example, the command: install.packages("ggplot2").

```
library(MASS)
library(ggplot2)
library(plyr)
library(GGally)
library(rstatix)
```

Give the package versions.

```
##  rstatix   GGally     plyr  ggplot2     MASS
##  "0.6.0"  "2.0.0"  "1.8.6"  "3.3.3" "7.3-53"
```

Introduction

We here give a proof by counterexample that negative correlations between measurements of complexity in different domains (i.e. complexity trade-offs) do not strictly entail equi-complexity (in the sense of equality of mean complexity) of languages on which these measurements are taken.

Theoretical Background

Let us assume we have n languages for which we measure complexities in two domains, e.g. syntax and morphology, such that we get two samples of measurments $m = (m_1, m_2, \ldots, m_n)$ and $s = (s_1, s_2, \ldots, s_n)$. This situation is illustrated in the table below.

language	m	s
L_1	m_1	s_1
L_2	m_2	s_2
L_3	m_3	s_3
L_4	m_4	s_4
L_5	m_5	s_5
...	...	...
L_n	m_n	s_n

Trade-offs as Negative Correlations

Trade-offs are here conceptualized as negative correlations. We here choose the Pearson (r) and Spearman (ρ) correlation coefficients as examples. While the former measures linear dependence, the latter is a non-parametric rank correlation.

Across the languages L_1 to L_n the Pearson correlation coefficient between the complexity measurements in the two domains (e.g. morphology and syntax) is defined as

$$r_{ms} = \frac{\sum_{i=1}^{n}(m_i - \overline{m})(s_i - \overline{s})}{\sqrt{\sum_{i=1}^{n}(m_i - \overline{m})^2}\sqrt{\sum_{i=1}^{n}(s_i - \overline{s})^2}}, \tag{1}$$

where n is the number of data points in the paired samples, m_i and s_i are individual measurements in the respective domain, and $\overline{m}$ and $\overline{s}$ are the arithmetic means of the samples (columns in the table above), i.e. complexity measurements in a certain domain, with

$$\overline{m} = \frac{1}{n}\sum_{i=1}^{n} m_i, \tag{2}$$

and

$$\overline{s} = \frac{1}{n}\sum_{i=1}^{n} s_i. \tag{3}$$

We have a negative correlation $r_{ms} < 0$ iff the numerator is negative, i.e.

$$\sum_{i=1}^{n}(m_i - \overline{m})(s_i - \overline{s}) < 0. \tag{4}$$

Note that the denominator cannot be negative.

The Spearman correlation coefficient, on the other hand, is then defined as

$$\rho_{ms} = 1 - \frac{6\sum_{i=1}^{n}(\mathrm{rank}(m_i) - \mathrm{rank}(s_i))^2}{n(n^2-1)}, \tag{5}$$

where rank() is a function that gives the rank of the respective value when the values are ranked in ascending order (i.e. the smallest receives rank 1, the second smallest rank 2 etc.). Note that this definition only holds for distinct integers being ranked, which we will assume here for simplicity. We get a negative correlation iff

$$\frac{6\sum_{i=1}^{n}(\mathrm{rank}(m_i) - \mathrm{rank}(s_i))^2}{n(n^2-1)} > 1. \tag{6}$$

Equi-Complexity as Equality of Means

Furthermore, we conceptualize equi-complexity of languages L_1 to L_n here as equality of arithmetic means. For instance, if language L_1 has a morphological complexity of $m_1 = 5$ and a syntactic complexity of $s_1 = 1$, and language L_2 has $m_2 = 1$ and $s_2 = 5$, then the arithmetic mean complexity is 3 for both. Hence, they are considered overall equally complex. Assume more generally that m_j and s_j as well as m_k and s_k represent complexity measurements for two languages L_j and L_k. We would then consider the languages equi-complex iff

$$\overline{L_j} = \overline{L_k}, \tag{7}$$

where

$$\overline{L_j} = \frac{1}{d}(m_j + s_j), \tag{8}$$

and

$$\overline{L_k} = \frac{1}{d}(m_k + s_k). \tag{9}$$

Here d is the number of different domains for which we measure complexity, i.e. d = 2 in our example of morphology and syntax.

Proof

We here proof by counterexample that neither a negative Pearson nor a negative Spearman correlation between two samples of complexity measurements in different domains strictly entail equality of means for the respective languages from which these measurments were taken. In other words, we will disproof by counterexample the claim that

$$(r_{ms} < 0), (\rho_{ms} < 0) \vdash (\overline{L_j} = \overline{L_k}). \tag{10}$$

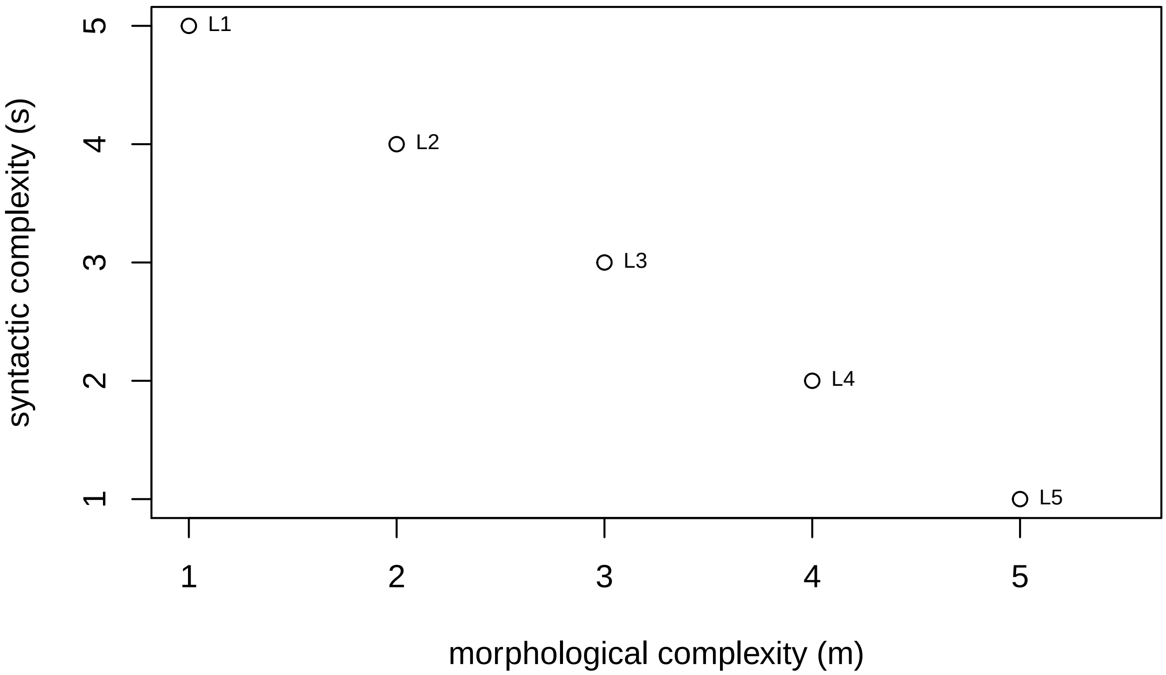
Firstly, assume that we have a perfect negative Pearson and Spearman correlation between two samples, i.e. $r_{ms} = \rho_{ms} = -1$. For example, assume that m = (1, 2, 3, 4, 5), while the syntactic complexity can be perfectly linearly predicted by $s = 6 - 1m$ across five different languages. In other words, the syntactic complexity values are a linear transformation of the morphological complexity values. We thus have the following table of measurements per language and domain:

```
language = c("L1", "L2", "L3", "L4", "L5")
m <- c(1, 2, 3, 4, 5)
s <- 6 + -1*m
example.df <- data.frame(language, m, s)
print(example.df)
```

```
##   language m s
## 1       L1 1 5
## 2       L2 2 4
## 3       L3 3 3
## 4       L4 4 2
## 5       L5 5 1
```

This data set is visualized in the plot below.

```
plot(example.df$m, example.df$s, xlab = "morphological complexity (m)",
     ylab = "syntactic complexity (s)", xlim = c(1, 5.5))
text(example.df$m, example.df$s, labels = example.df$language, cex = 0.7, pos = 4)
```



In this case, we indeed have equality of means across the languages, namely

$$\overline{L}_1 = \overline{L}_2 = \overline{L}_3 = \overline{L}_4 = \overline{L}_5 = 3. \tag{11}$$

Note that in this particular case, we also have $\overline{m} = \overline{s} = 3$. However, it is of course not necessarily the case that the means for languages and the means of measurements per domain are equal.

For the numerator of the Pearson correlation we have

$$\begin{aligned}\sum_{i=1}^{n}(m_i - \overline{m})(s_i - \overline{s}) = \\ (1-3)(5-3) + (2-3)(4-3) + (3-3)(3-3) + (4-3)(2-3) + (5-3)(1-3) = \\ -4 - 1 + 0 - 1 - 4 = -10.\end{aligned} \tag{12}$$

The condition for a negative Pearson correlation in equation (4) is hence fullfilled. In fact, in this particular case it is a perfect negative correlation ($r_{ms} = -1$) since the denominator of the formular for the Pearson coefficient in equation (1) evaluates to 10.

This result can be double-checked with the R function cor().

```
cor(example.df$m, example.df$s, method = "pearson")
```

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

In order to evaluate the Spearman condition (equation (6)) for a negative correlation we might first create a table with the rankings of complexity measurements in ascending order. Note, however, that in our particular case, the rankings coincide with the sample values themselves, i.e. the ranking for m is rank(m_1) = 1, rank(m_2) = 2, etc. Hence, we can simply take the original values as the output of rank().

$$\begin{aligned} \frac{6\sum_{i=1}^{n}(\text{rank}(m_i) - \text{rank}(s_i))^2}{n(n^2-1)} &= \\ \frac{6((1-5)^2 + (2-4)^2 + (3-3)^2 + (4-2)^2 + (5-1)^2)}{5(5^2-1)} &= \\ \frac{6((-4)^2 + (-2)^2 + (0)^2 + (2)^2 + (4)^2)}{5(24)} &= \\ \frac{240}{120} &= 2 \end{aligned} \tag{13}$$

We thus also get a perfect negative Spearman correlation of $\rho_{ms} = -1$.

This result can be double-checked with the R function cor().

```
cor(example.df$m, example.df$s, method = "spearman")
```

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

To summarize, in our example data set we have perfect negative correlations between the measurements in the two domains, i.e. a perfect trade-off between morphological and syntactic complexity, and we have equality of means across the languages, i.e. overall equi-complexity.

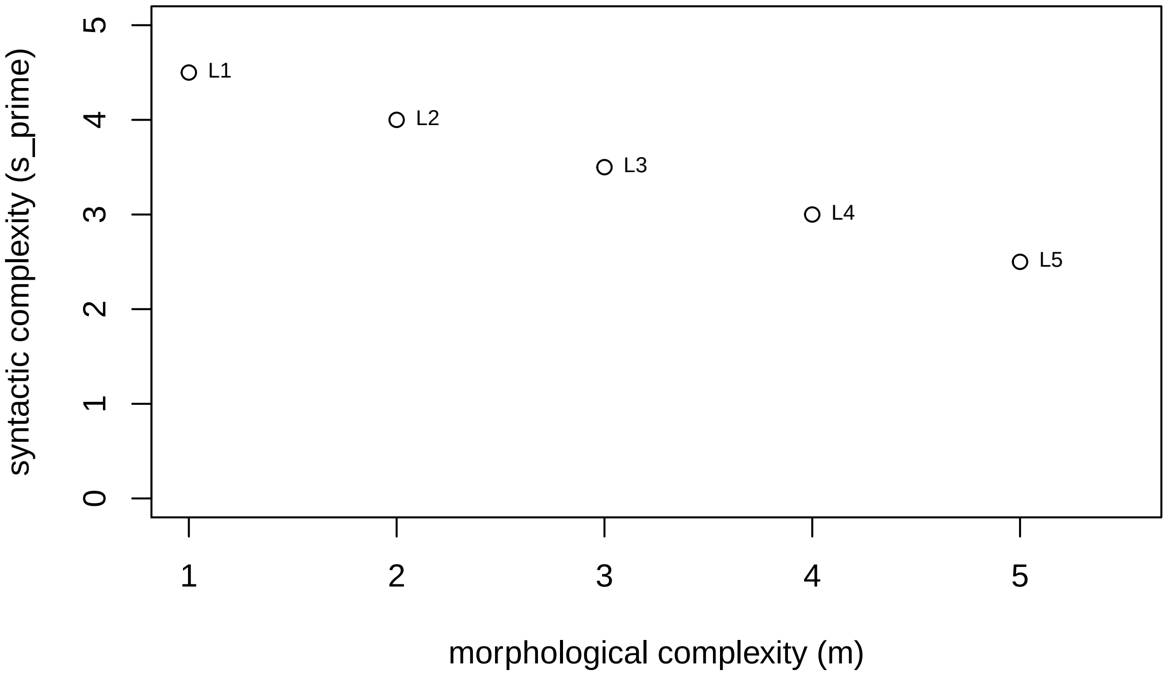
Now, let us “flatten” the curve by decreasing the slope with which the syntactic complexity decreases with morphological complexity to 0.5 (instead of 1 as before). We thus get the following syntactic complexity values, e.g. by squaring them such that $s' = 5 - 0.5m$.

```
language = c("L1", "L2", "L3", "L4", "L5")
m <- c(1, 2, 3, 4, 5)
s_prime <- 5 - 0.5*m
example.df <- data.frame(language, m, s, s_prime)
print(example.df)
```

```
##   language m s s_prime
## 1       L1 1 5     4.5
## 2       L2 2 4     4.0
## 3       L3 3 3     3.5
## 4       L4 4 2     3.0
## 5       L5 5 1     2.5
```

This data set is visualized in the plot below.

```
plot(example.df$m, example.df$s_prime, xlab = "morphological complexity (m)",
     ylab = "syntactic complexity (s_prime)", xlim = c(1, 5.5), ylim = c(0, 5))
text(example.df$m, example.df$s_prime, labels = example.df$language, cex = 0.7, pos = 4)
```



Since the syntactic complexity here decreases less than the morphological complexity increases, we now get differing mean values across the languages.

$$
\begin{aligned}
\overline{L'_1} &= \frac{5.5+1}{2} = 2.75 \\
\overline{L'_2} &= \frac{4+2}{2} = 3 \\
&\qquad\quad etc.
\end{aligned}
\tag{14}
$$

For the condition for a negative Pearson correlation we now get

$$
\begin{aligned}
\sum_{i=1}^{n}(m_i - \overline{m})(s'_i - \overline{s}') = \\
(1-3)(4.5-3.5) + (2-3)(4-3.5) + (3-3)(3.5-3.5) + (4-3)(3-3.5) + (5-3)(2.5-3.5) = \\
-4 - 0.5 + 0 - 0.5 - 2 = -7
\end{aligned}
\tag{15}
$$

Hence, this condition is still fullfilled, and we have a negative Pearson correlation. In fact, the Pearson correlation is still perfect, i.e. -1, as can be double-checked with the cor() function.

```
cor(example.df$m, example.df$s_prime, method = "pearson")
```

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

In order to evaluate the Spearman condition (equation (6)), we first create a table with the rankings of complexity measurements in ascending order.

```
language = c("L1", "L2", "L3", "L4", "L5")
m <- c(1, 2, 3, 4, 5)
s <- 6 - 1*m
s_prime <- 5 - 0.5*m
```

```r
rank_m <- c(1, 2, 3, 4, 5)
rank_s_prime <- c(5, 4, 3, 2, 1)
example.df <- data.frame(language, m, s, s_prime, rank_m, rank_s_prime)
print(example.df)
```

```
##   language m s s_prime rank_m rank_s_prime
## 1       L1 1 5     4.5      1            5
## 2       L2 2 4     4.0      2            4
## 3       L3 3 3     3.5      3            3
## 4       L4 4 2     3.0      4            2
## 5       L5 5 1     2.5      5            1
```

We can already see in the table that the ranking order involving s' is equivalent to s. Hence, we get the exact same result as before.

$$\frac{6\sum_{i=1}^{n}(\text{rank}(m_i)-\text{rank}(s'_i))^2}{n(n^2-1)} = 2 \tag{16}$$

We thus also still have a perfect negative Spearman correlation.

This result can be double-checked with the R function cor().

```r
cor(example.df$m, example.df$s_prime, method = "spearman")
```

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

Summary

We have thus found a simple transformation of our syntactic complexity values which yields the exact same perfect negative correlations as the untransformed syntactic complexity values for both the Pearson and the Spearman correlation. However, while in the case of the original syntactic complexity values (s) the means across languages are also the same, in the scenario of transformed values (s'), they are not. This proofs that not even perfect negative Pearson and Spearman correlations strictly entail equality of means, i.e.

$$(r_{ms} < 0), (\rho_{ms} < 0) \nvdash (\overline{L_j} = \overline{L_k}). \tag{17}$$

Open Issues

Standardization of data

In the proof above we have used two slightly differing linear transformations to derive s and s' from m. A general formular for linear transformations is $y = a + bx$. In the first case we used a = 6 and b = -1, and in the second case a = 5 and b = -0.5. It is noteworthy in this context that standardization of the sample values neutralizes the effect of linear transformations on the mean values of languages.

Standardization typically involves centering (which is defined as subtracting the mean value from all individual values of a sample), and scaling (which is defined as dividing the values of a sample by the standard deviation of the sample.). For instance, centering of the morphological complexity values is generally defined as

$$m^{centered} = m - \overline{m}, \tag{18}$$

and scaling is defined as

$$m^{scaled} = \frac{m}{\sigma^m}, \tag{19}$$

where σ^m is the standard deviation of the morphological complexity values.

Standardization then typically involves both centering and scaling, such that we have

$$m^{standardized} = \frac{m - \overline{m}}{\sigma^m}, \tag{20}$$

As an example, let us standardize the m, s, and s' sample values from above:

```
m_standardized <- (m-mean(m))/sd(m)
s_standardized <- (s-mean(s))/sd(s)
s_prime_standardized <- (s_prime-mean(s_prime))/sd(s_prime)
example.df <- data.frame(m, s, s_prime, m_standardized, s_standardized, s_prime_standardized)
example.df
```

```
##   m s s_prime m_standardized s_standardized s_prime_standardized
## 1 1 5     4.5     -1.2649111      1.2649111            1.2649111
## 2 2 4     4.0     -0.6324555      0.6324555            0.6324555
## 3 3 3     3.5      0.0000000      0.0000000            0.0000000
## 4 4 2     3.0      0.6324555     -0.6324555           -0.6324555
## 5 5 1     2.5      1.2649111     -1.2649111           -1.2649111
```

Remember that each row of this data frame corresponds to values of a different languages. As we have pointed out above, the mean values per language given m and s' are different, i.e.

```
(m + s_prime)/2
```

```
## [1] 2.75 3.00 3.25 3.50 3.75
```

However, this effect of the linear transformation from s to s' on the mean values of the languages is neutralized if we standardize both m and s_prime, i.e.

```
(m_standardized + s_prime_standardized)/2
```

```
## [1] 0 0 0 0 0
```

Thus, due to standardization of the data, the mean values across languages are again equal, despite the linear transformation of on s.