formua

```r
options(warn=-1)
suppressPackageStartupMessages(library(data.table))
suppressPackageStartupMessages(library(bit64))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(tidyr))
suppressPackageStartupMessages(library(ggplot2))

if (getwd()=="C:/stuff/stanford") {
  setwd("C:/stuff/stanford/alignment")
}
if(getwd()=="/Users/jakeprasad"){
  setwd("/Users/jakeprasad/alignment")
}
```

It seems our plots aren't giving us expected results. Let's try and figure out why that may be

We know that smoothing is affecting our plots - especially because it affects one term more than another. Maybe instead of subtracting logs, we should move everything inside of the log?

Let's do some thinking math...

For a completely shuffled dataset, we expect alignment to be equal to 0. Let's check if the formula gives us that result.

(A bunch of work done on paper)

This is interesting. It seems that whether alignment is positive or negative depends on if

(ba*nbna) is less than (nba*nba). Let's check what happens in our shuffled datas

```r
  df <- fread('debug/shuffled/TTTTTTTT.csv', header=T)

  test <- ((df$ba*df$nbna)/(df$bna*df$nba))
  test[is.infinite(test)] <- NA
  mean(test, na.rm=T)
```

```
## [1] 0.6776141
```

```r
  df <- fread('debug/shuffled/TTTTTTTT1.csv', header=T)

  test <- ((df$ba*df$nbna)/(df$bna*df$nba))
  test[is.infinite(test)] <- NA
  mean(test, na.rm=T)
```

```
## [1] 0.6871238
```

```r
  df <- fread('debug/shuffled/TTTTTTTT2.csv', header=T)

  test <- ((df$ba*df$nbna)/(df$bna*df$nba))
  test[is.infinite(test)] <- NA
  mean(test, na.rm=T)
```

```
## [1] 0.683967
```

Hmm, we're consistently getting a <1 value, indicating a negative alignment. How do we fix this?

Let's do some more thinkging about what we're trying to calculate. We want to check whether the probability of B saying a given marker is more immediately after A says a given marker.

We're getting a negative alignment on shuffled data because the base frequencies of markers is small.

Let's take a look at the numbers

df$ba = 0df$nbna ~= 1

df$bna => 0, < 1df$nba ~= >0, <1

Since ba*nbna is always going to be approximately 0, we're getting the <1 result.

More thinking…

New Formula

Let's rethink the formula with examples

If A influences B, and A says the marker a lot Before: Afreq = 1, Bfreq = 0 After: Afreq = 0.9, Bfreq = 0.5

If B influences A, and A says the marker a lot Before: Afreq = 1, Bfreq = 0 After: Afreq = 0.5, Bfreq = 0.1

If A influences B, and A doesn't say the marker a lot Before: Afreq = 0, Bfreq = 1 After: Afreq = 0.1, Bfreq = 0.5

If B influences A and A doesn't say the marker a lot Before: Afreq = 0, Bfreq = 1 After: Afreq = 0.5, Bfreq = 0.9

How do we mathematically compare the freqs?

If A influences B, and A says the marker a lot P(A|nB utterance): Afreq = 1 P(B|nA utterance): Bfreq = 0 P(A|B utterance): Afreq = 0.9 P(B|A utterance): Bfreq = 0.5

If B influences A, and A says the marker a lot P(A|nB utterance): Afreq = 1 P(B|nA utterance): Bfreq = 0 P(A|B utterance): Afreq = 0.5 P(B|A utterance): Bfreq = 0.1

If A influences B, and A doesn't say the marker a lot P(A|nB utterance): Afreq = 0 P(B|nA utterance): Bfreq = 1 P(A|B utterance): Afreq = 0.1 P(B|A utterance): Bfreq = 0.5

If B influences A and A doesn't say the marker a lot P(A|nB utterance): Afreq = 0 P(B|nA utterance): Bfreq = 1 P(A|B utterance): Afreq = 0.5 P(B|A utterance): Bfreq = 0.9

And now we don't need to know relative base freqs

If A influences B P(A|nB utterance): Afreq = 1 P(B|nA utterance): Bfreq = 0 P(A|B utterance): Afreq = 0.9 P(B|A utterance): Bfreq = 0.5

If B influences A P(A|nB utterance): Afreq = 1 P(B|nA utterance): Bfreq = 0 P(A|B utterance): Afreq = 0.5 P(B|A utterance): Bfreq = 0.1

If A influences B P(A|nB utterance): Afreq = 0 P(B|nA utterance): Bfreq = 1 P(A|B utterance): Afreq = 0.1 P(B|A utterance): Bfreq = 0.5

If B influences A P(A|nB utterance): Afreq = 0 P(B|nA utterance): Bfreq = 1 P(A|B utterance): Afreq = 0.5 P(B|A utterance): Bfreq = 0.9

And now the coup d'etat

If A influences B P(A|B) - P(A|nB) - = 0.9 - 1 = -0.1 = B's influence on A P(B|A) - P(B|nA) - = 0.5 - 0 = 0.5 = A's infleunce on B

If B influences A P(A|nB) - P(A|B) = 1 - 0.5 = 0.5 = B's influence on A P(B|nA) - P(B|A) = 0 - 0.1 = -0.1 = A's influence on B

If A influences B P(A|nB) - P(A|B) = 0 - 0.1 = -0.1 = B's influence on A P(B|nA) - P(B|A) = 1 - 0.5 = 0.5 = A's influence on B

If B influences A P(A|nB) - P(A|B) = 0 - 0.5 = -0.5 = B's influence on A P(B|nA) - P(B|A) = 1 - 0.9 = 0.1 = A's influence on B

Therefore, if abs(P(B|A) - P(B|nA)) > abs(P(A|B) - P(A|nB)), A influences B

Let's see what the formula looks like:

P(A|nB) = anb/(anb+nanb)

P(A|B) = ab/(ab+nab)

P(B|nA) = bna/(bna+nbna)

P(B|A) = ba/(ba+nba)

Therefore,

abs(ba/(ba+nba) - bna/(bna+nbna)) - abs(ba/(ba+bna) - nba/(nba+nbna))

TESTS

Let's think about how it'd work on shuffled data with example numbers:

ba = 1 nba = 5 bna = 5 nbna = 14

abs(ba/(ba+nba) - bna/(bna+nbna)) - abs(ba/(ba+bna) - nba/(nba+nbna)) = 0

It works (on shuffled data at least)

How about for an expected positive alignment?

Let's say that A influences B more than B influences A. And let's also suppose A says the marker a lot and B doesn't. There are two things to take into acount here:

1) A says the marker more than B says the marker => P(a|nb) > P(b|na) == nba/(nba+nbna) - bna/(bna+nbna) > 0
2) A influences B => P(b|a) > P(b|na) == ba/(ba+nba) - bna/(bna+nbna) > 0
3) We're using valid frequencies ba + nba + bna + nbna == 1

Let's use new numbers

ba = 0.5 nba = 0.8 bna = 0.1 nbna = 0.20

abs(ba/(ba+nba) - bna/(bna+nbna)) - abs(ba/(ba+bna) - nba/(nba+nbna)) > 0

= -0.003663004

```
ba = 0.3
nba = 0.4
bna = 0.1
nbna = 0.2

nba/(nba+nbna) - bna/(bna+nbna) > 0
```

```
## [1] TRUE
```

```r
ba/(ba+nba) - bna/(bna+nbna) > 0
```

```
## [1] TRUE
```

```r
ba + nba + bna + nbna == 1
```

```
## [1] TRUE
```

```r
abs(ba/(ba+nba) - bna/(bna+nbna)) - abs(ba/(ba+bna) - nba/(nba+nbna)) > 0
```

```
## [1] TRUE
```

Positive alignment =)

Let's say that A influences B more than B influences A. And let's also suppose B says the marker more than A does:

1) A says the marker less than B says the marker => $P(b|na) > P(a|nb)$ == bna/(bna+nbna) - nba/(nba+nbna) > 0
2) A influences B $P(b|na) > P(b|a)$ == bna/(bna+nbna) - ba/(ba+nba) > 0
3) We're using valid frequencies ba + nba + bna + nbna == 1

```r
ba = 0.05
nba = 0.1
bna = 0.70
nbna = 0.15

bna/(bna+nbna) - nba/(nba+nbna) > 0
```

```
## [1] TRUE
```

```r
bna/(bna+nbna) - ba/(ba+nba) > 0
```

```
## [1] TRUE
```

```r
ba + nba + bna + nbna == 1
```

```
## [1] TRUE
```

```r
abs(ba/(ba+nba) - bna/(bna+nbna)) - abs(ba/(ba+bna) - nba/(nba+nbna)) > 0
```

```
## [1] TRUE
```

Positive alignment =)

Let's say that B influences A more than A influences B. And let's also suppose B says the marker more than A does:

1) B says the marker more than A says the marker => $P(b|na) > P(a|nb)$ == bna/(bna+nbna) - nba/(nba+nbna) > 0
2) B influences A $P(a|b) > P(a|nb)$ == ba/(ba+bna) - nba/(nba+nbna) > 0
3) We're using valid frequencies ba + nba + bna + nbna == 1

```r
ba = 0.3
nba = 0.1
bna = 0.3
nbna = 0.3

bna/(bna+nbna) - nba/(nba+nbna) > 0
```

```
## [1] TRUE
```

```r
ba/(ba+bna) - nba/(nba+nbna) > 0
```

```
## [1] TRUE
```

```r
ba + nba + bna + nbna == 1
```

```
## [1] TRUE
```

```r
abs(ba/(ba+nba) - bna/(bna+nbna)) - abs(ba/(ba+bna) - nba/(nba+nbna)) < 0
```

```
## [1] TRUE
```

Negeative alignment =)

Let's say that B influences A more than A influences B. And let's also suppose A says the marker more than B does:

1) A says the marker more than B says the marker => P(a|nb) > P(b|na) == nba/(nba+nbna) - bna/(bna+nbna) > 0
2) B influences A P(a|nb) > P(a|b) == nba/(nba+nbna) - ba/(ba+bna) > 0
3) We're using valid frequencies ba + nba + bna + nbna == 1

```r
ba = 0.1
nba = 0.5
bna = 0.1
nbna = 0.3

nba/(nba+nbna) - bna/(bna+nbna) > 0
```

```
## [1] TRUE
```

```r
nba/(nba+nbna) - ba/(ba+bna) > 0
```

```
## [1] TRUE
```

```r
ba + nba + bna + nbna == 1
```

```
## [1] TRUE
```

```r
abs(ba/(ba+nba) - bna/(bna+nbna)) - abs(ba/(ba+bna) - nba/(nba+nbna)) < 0
```

```
## [1] TRUE
```

Negeative alignment =)