Testing Alignment Measures for Bias in Toy Cases

Gabe Doyle

July 14, 2015

This document performs a series of tests on toy cases of alignment, so we can test different proposed measures of alignment for bias and variance.

Original metric: subtracting probabilities

Starting off with the original metric (p(B|A) - p(B)). In the toy case, we assume that all users have the same baseline probability, and that alignment is additive. We test on a range of baseline probabilities and alignment strengths; beause the generative model is additive, the estimated value of the alignment should be an unbiased estimate of the "true" alignment based on the model parameter. We also test a range of numbers of examples to see how much variance this estimate has.

```
## Loading required package: data.table
```

```
## Warning: package 'data.table' was built under R version 3.1.2
```

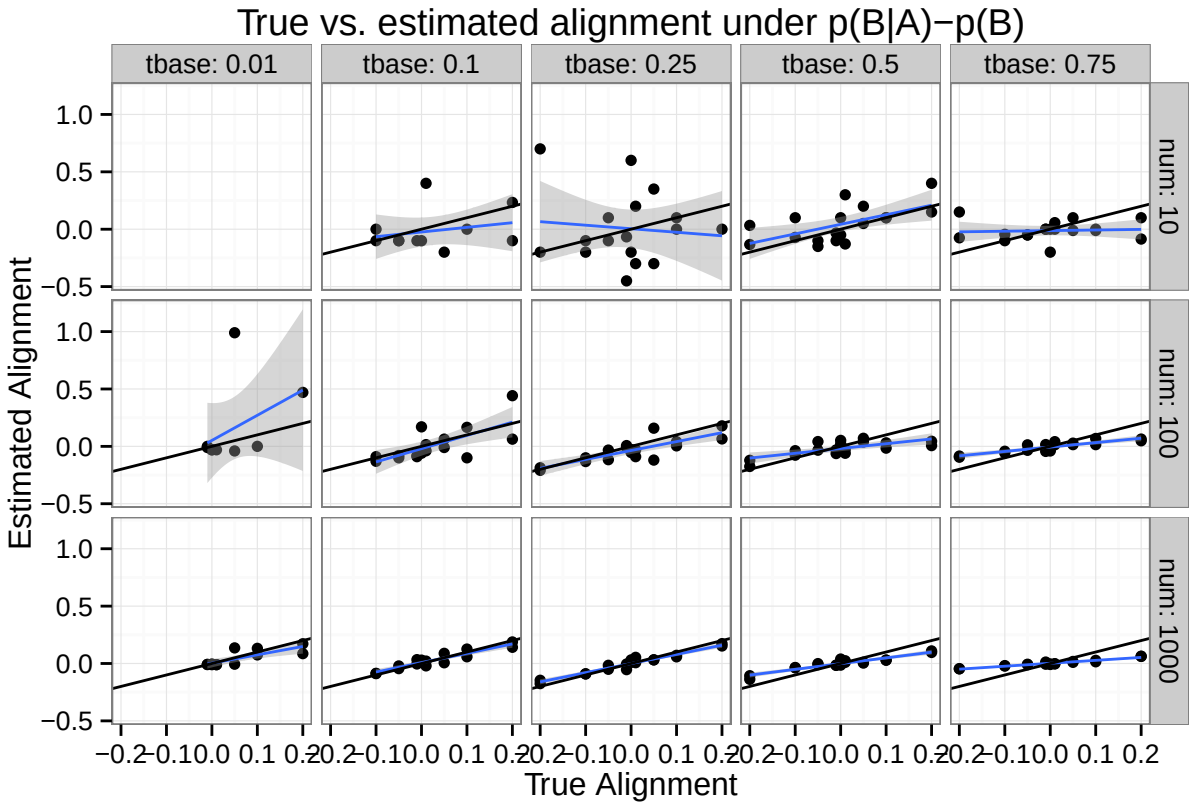
First, we look at the effect of different true baselines (tbase) and number of observed replies (num) on the estimates of alignment under the p(B|A)-p(B) measure. A substantial conservative bias appears even at fairly low baseline rates and becomes stronger as the baseline increases. If the marker occurs in half of all utterances, the estimated alignment is only about half of the true alignment. As the baseline approaches 1, the estimated alignment drops to 0.

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
ggplot(df[is.finite(ealign),],aes(x=talign,y=ealign)) + geom_point() + geom_smooth(method='lm') + geom_a
  facet_grid(num ~ tbase,labeller=label_both) +
  labs(title='True vs. estimated alignment under p(B|A)-p(B)',y='Estimated Alignment',x='True Alignment
  theme_bw()
```



Why is this measure conservative and dependent on the baseline? Because the baseline is constructed from ALL of B's replies to A, regardless of whether A used the marker. Since $p(B) = p(B|notA)p(notA) + p(B|A)p(A)$, as $p(A)$ increases, $p(B)$ approaches $p(B|A)$ proportionately. Note that this equation also motivates our later use of p(B|notA) as the "baseline".

Furthermore, even in Twitter and child-directed speech, where messages are short, for many of the relevant markers, a 25-50% baseline would be unsurprising (*the*, for instance). This will become worse if we move to word categories.

First revised estimator: p(B|notA) as baseline

Our first change to the alignment measure is to replace the baseline $p(B)$ with $p(B|notA)$. For this first step, we retain the subtractive measure and only change the baseline: alignment is now $p(B|A) - p(B|notA)$. We will move to log-odds space next.

We encounter a new problem here in that $p(B|notA)$ may be undefined. Under the old metric, we did not have this problem because we would only calculate alignment if the original speaker A used the marker at least once. Here, knowing A used the marker is no guarantee that A also didn't use the marker, which is the denominator of $p(B|notA)$. So we also test add-lambda smoothing.

```
require(data.table)

dfinitialized <- F

aprobs <- c(.01,.25,.5,.75,.99) #true p(B|A)
notaprobs <- c(.01,.25,.5,.75,.99) #true p(B|notA)
```

```
numexampleses <- c(10,100,1000)
numruns <- 20
smoothings <- c(0,.01,.1,1)
#numexamples <- 1000

for (i in seq(1,numruns)) {
  #print(i)
for (smoothing in smoothings) {
for (aprob in aprobs) {
  for (notaprob in notaprobs) {
    base <- mean(c(aprob,notaprob)) #setting p(A) as mean of p(B|A) and p(B|notA)
    #base <- notaprob               #alternately, setting p(A) as p(B|notA), assuming absence has no pr
    for (numexamples in numexampleses) {
      speaker <- runif(numexamples)<base
      replier <- runif(numexamples)<(speaker*aprob)+((1-speaker)*notaprob)
      eaprob <- (sum(speaker&replier)+smoothing)/(sum(speaker)+2*smoothing)
      enotaprob <- (sum((1-speaker)&replier)+smoothing)/(sum(1-speaker)+2*smoothing)
      newline <- c(aprob,notaprob,aprob-notaprob,numexamples,
                   sum(speaker&replier),sum((1-speaker)&replier),sum(speaker&(1-replier)),sum((1-speake
                   eaprob,enotaprob,eaprob-enotaprob,smoothing)

      tempdf <- as.data.table(t(newline))
      setnames(tempdf,names(tempdf),
                      c('taprob','tnotaprob','talign','num',
                        'tt','ft','tf','ff',
                        'eaprob','enotaprob','ealign','smoothing'))

    if (dfinitialized) {
    df <- rbind(df,tempdf)
    } else {
      df <- tempdf
      dfinitialized <- T
    }
    }
  }
}
}
}
```

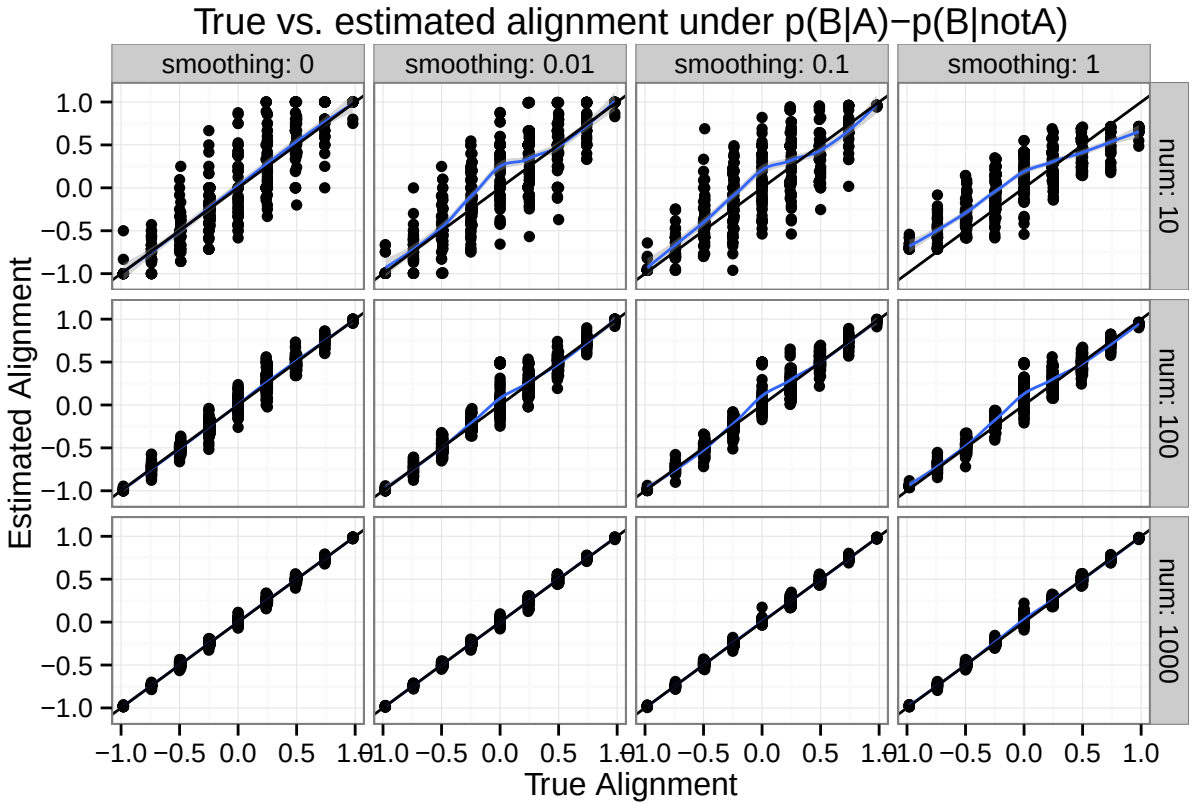
First, the rate of undefined values (FALSE=undefined) over different numbers of observed replies when no smoothing. (Any add-lambda smoothing fixes all of these).

```
tdf <- df[smoothing==0,]
table(is.finite(tdf$ealign),tdf$num)
```

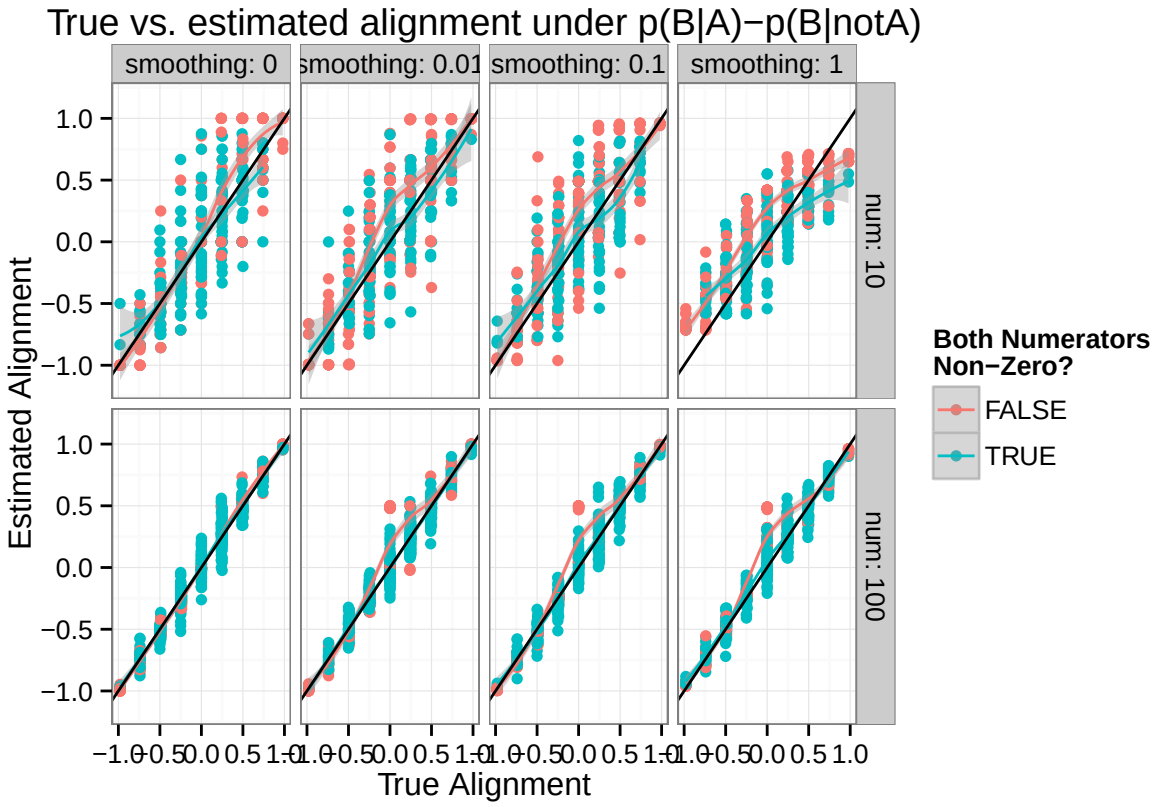
```
##
##          10 100 1000
##   FALSE  67  13    0
##   TRUE  433 487  500
```

Now, we plot the true vs. estimated alignment $p(B|A) - p(B|notA)$. We encounter a new problem due to the smoothing, and it's a bad one. There is a hump of phantom alignments in the middle of the low-replies runs

when smoothing is implemented. In addition, while in the high-replies limit there is no bias, in low-reply settings, the floor/ceiling alignment conservatism bias still appears.

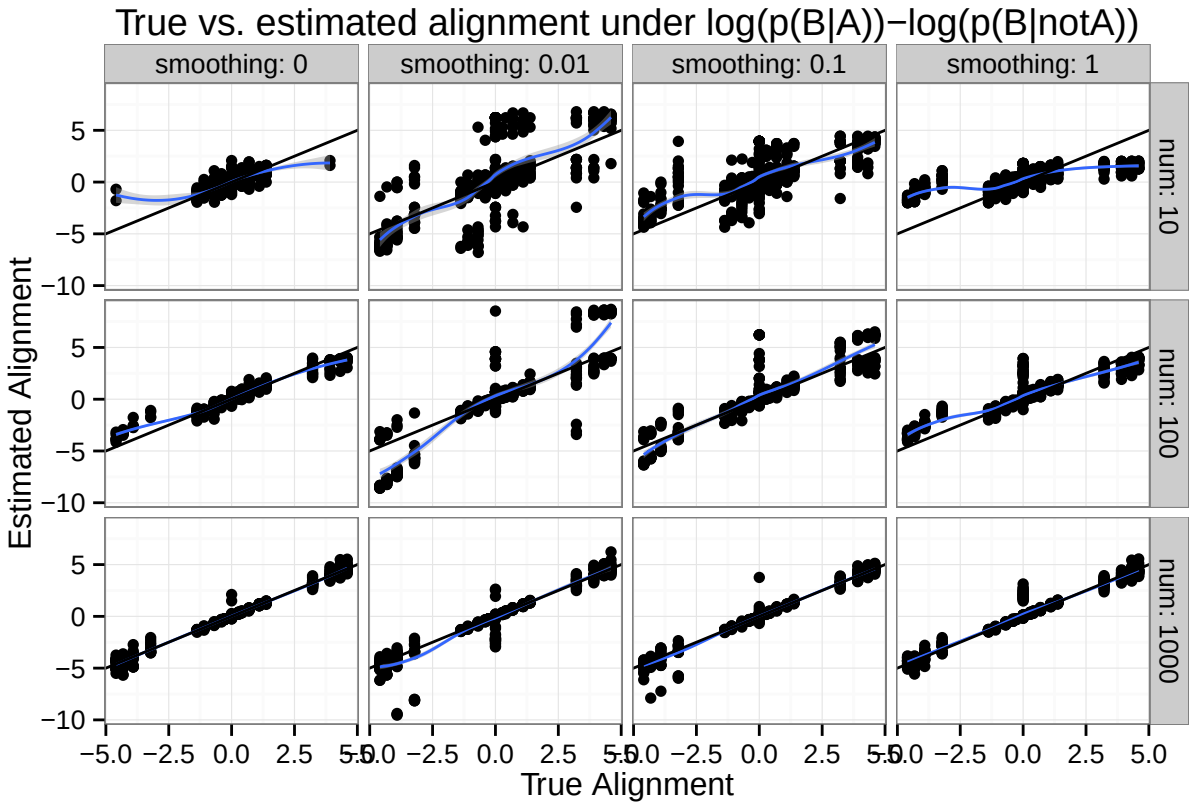


Looking briefly at the smoothing hump, we see that the estimated alignment is flawed when either numerator is zero – when we see B using the marker either only after A uses it or only after A doesn't use it. Those are the sorts of cases that smoothing is supposed to improve, so it looks like we have inappropriate smoothing for this measure.



Second revised estimator: log-odds with not-A baseline

We use the same data as above, just calculating the baseline differently. Again, in the high-reply limit, we see convergence to an unbiased estimate of the true alignment. When the number of replies is small, though, we see a conservative bias induced from sparsity and/or smoothing, and some weird behavior for small alignments (e.g., the sigmoidal behavior around alignment=0 in smoothing=.01,num=10).



Summary

So, in summary, we find that the original metric from the Danescu-Niculescu-Mizil et al papers is conservatively biased in its estimates of alignment. This bias remains even as the number of replies increases. Furthermore, the strength of the alignment bias is dependent on the baseline probability of a marker's use, with larger biases from more common markers.

Our first proposed change to alignment is to change the baseline calculation. This change appears to solve some problems. The alignment estimate is unbiased in the limit (large number of replies) and is not dependent on the rate of baseline usage. However, it requires smoothing (or exclusion of rare markers) and our current smoothing methods introduce odd behavior that may cause non-aligning repliers to appear to be aligning. I'm still having trouble identifying the source of this deviation.

Our second proposed change is to move from additive alignment to multiplicative by moving into log-odds space. This removes the floor/ceiling effects on alignment and thus can handle a wider range of baseline usage probabilities (potentially useful for accurate alignment on content words). This method is also unbiased in the limit, but still requires smoothing. Smoothing again induces odd behavior when there are few examples of the marker, and again may lead to large phantom alignments – although these phantom alignments appear approximately equally likely to be positive or negative, as opposed to the clear positive bias in the previous measure's phantom alignments.

In conclusion, I favor the log-odds alignment measure, with markers getting binned so as to avoid low baseline probabilities. We may also want to bin users (e.g., by follower counts or verified status on Twitter, by adults vs. children in CHILDES) in order to decrease the number of low-reply speaker-replier-marker triplets. We should work on improving the smoothing method and/or use a fairly strong threshold on number of replies per triplet to limit the sparsity-driven errors.