Get Zipfy With It

Using Zipf's Law to Control for Voluntary Response in Twitter Data

Kellin Rumsey

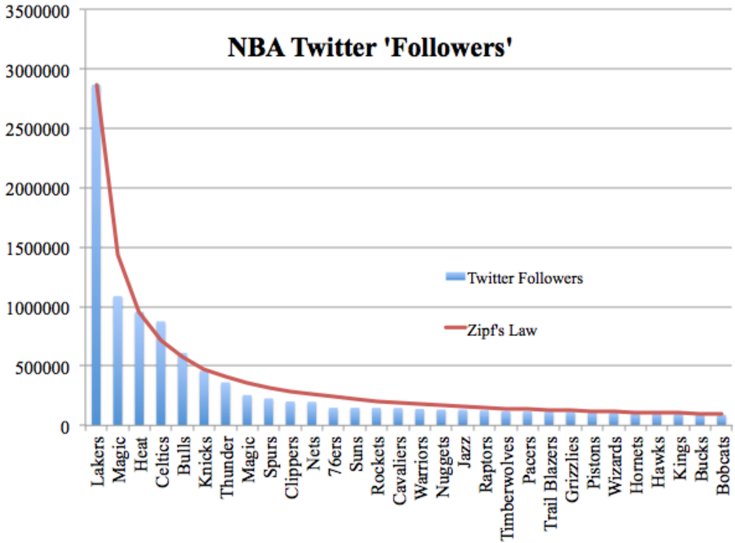
Zipf's Law

- Zipf's Law states that the frequency of X is inversely proportional to it's rank.
- Zipfian Decay: $P(X = x) \propto x^{-\theta}$
- Popularized in 1935 by George Zipf in Linguistics.
- Related to the 80-20 principle
- PMF for x = 0,1,2,...M

$$P(X = x|M, \theta) = \frac{(x + 1)^{-\theta}}{H(M + 1, \theta)} \tag{1}$$

- $H(n, \theta) = \sum_{k=1}^{n} k^{-\theta}$

Zipf's Law


NBA Twitter 'Followers'
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
Twitter Followers
Zipf's Law
Lakers
Magic
Heat
Celtics
Bulls
Knicks
Thunder
Magic
Spurs
Clippers
Nets
76ers
Suns
Rockets
Cavaliers
Warriors
Nuggets
Jazz
Raptors
Timberwolves
Pacers
Trail Blazers
Grizzlies
Pistons
Wizards
Hornets
Hawks
Kings
Bucks
Bobcats

Problem Setting

- Setting: Using twitter data to predict the outcome of the election.
- Collect a bunch of topic-related tweets, and classify the "sentiment" of each one. We assume

$$S_i \sim Bernoulli(\gamma) \tag{2}$$

- Then the MLE is:

$$\hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} S_i \tag{3}$$

Voluntary Response

- Social media data is plagued by Voluntary Response.
- What happens if the sentiment depends on a users "passion".

$$K_i \sim Zipfian(M, \theta) \tag{4}$$

$$S_i | K_i \sim Bernoulli(\gamma(K_i)) \tag{5}$$

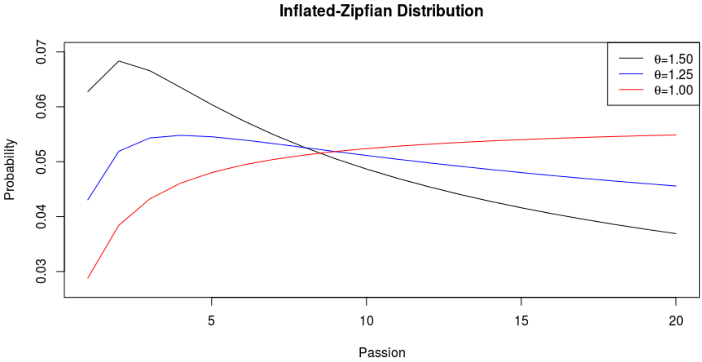
$$S_i \sim Bernoulli(\Gamma) \tag{6}$$

- Our goal is to estimate $\Gamma = E[\gamma(K)]$.
- In theory, just use MLE again... But we cannot obtain a random sample of users, only a random sample of tweets.

Voluntary Response

- When we find topic-related tweets, we are sampling from an "Inflated-Zipfian Distribution".

$$P(X = x) \propto x(x+1)^{-\theta} \quad (7)$$



Voluntary Response

- What we are actually sampling.

$$\mathcal{K}_i \sim \textit{Inflated-Zipfian}(M, \theta) \tag{8}$$

$$\mathcal{S}_i | \mathcal{K}_i \sim \textit{Bernoulli}(\gamma(\mathcal{K}_i)) \tag{9}$$

$$\mathcal{S}_i \sim \textit{Bernoulli}(\Gamma_2) \tag{10}$$

- But we are trying to estimate Γ... not Γ_2, and they can be very different.
- Solution: Each time we find a topic-related tweet do two things.
  1. Classify the tweet and find it's sentiment.
  2. Look at the users most recent M tweets, and see how many are also related to the topic.

Our Solution

- Let's take a closer look at Γ.

$$\Gamma = E[\gamma(K)] = \sum_{k=0}^{N} \gamma(k) \frac{(k+1)^{-\theta}}{H(N+1,\theta)} \tag{11}$$

- We can break Γ into two parts.

$$\Gamma = \frac{1}{H}\gamma(0) + (1 - \frac{1}{H})\Gamma^* \tag{12}$$

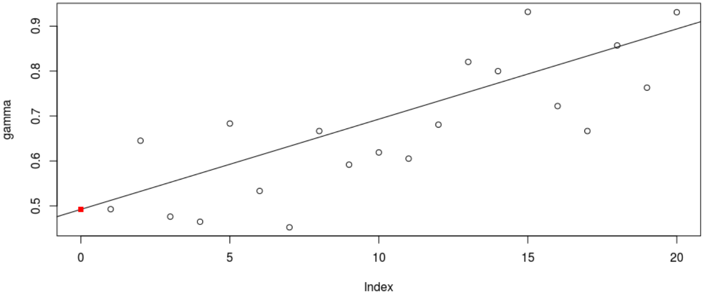
- We can construct an ubiased estimator for Γ^*.

$$\hat{\Gamma}^* = \frac{\sum \mathcal{S}_i \mathcal{K}_i^{-1}}{\sum \mathcal{K}_i^{-1}} \tag{13}$$

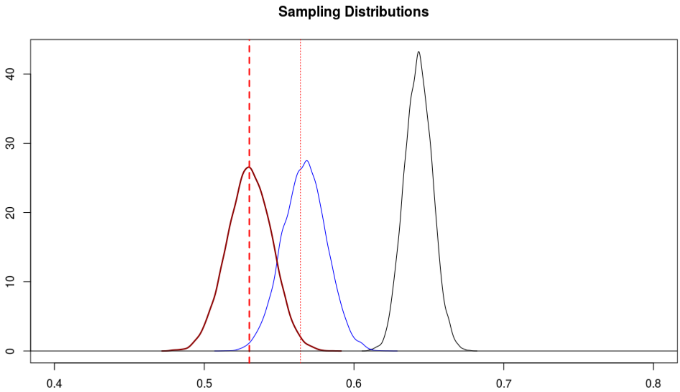
- We *may* be able to estimate $\gamma(0)$ with statistical learning (Regression).

Simulation Study

Estimating $\hat{\gamma}(0)$



Comparing Estimators

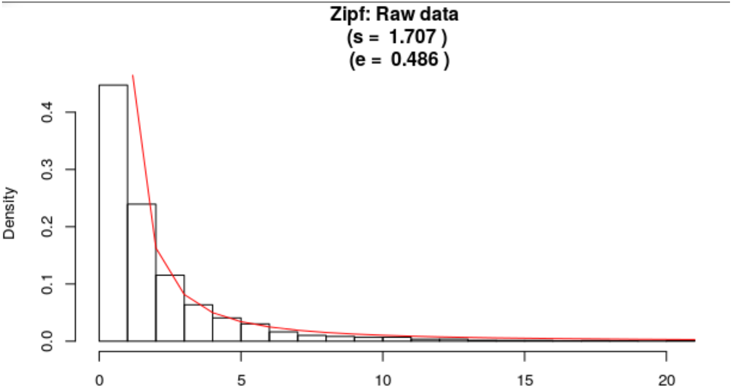


Simulation Study Summary

- We assume that there is a relationship between passion and sentiment.
  - If not, our estimator will still not work, but in this case the naive estimator might be okay.
- We assume that Zipf's Law applies to the data.
  - The method is flexible, we can easily choose a different decay model.
  - Zipfian Distribution has proven to be more reasonable for this kind of data.
- **We assume that we make no misclassification error.**
  - In practice, we estimate that our misclassification error was as high as 25%.
  - It may be possible to model the misclassification errors. Point of possible future research.

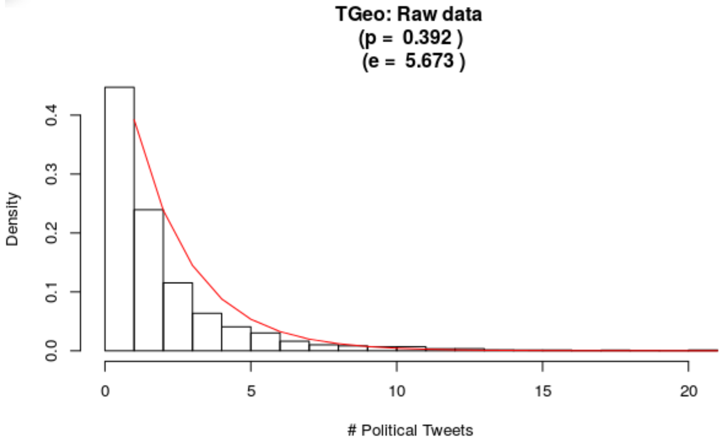
Twitter Data

- Early on, we collected a random sample of $\approx$ 7,000 tweets from NYC.
- For each unique user in the sample, we pulled their last 20 tweets and counted how many were "political".
- Using Metropolis-Hastings, we were able to estimate θ under the Zipf's Law Assumption.



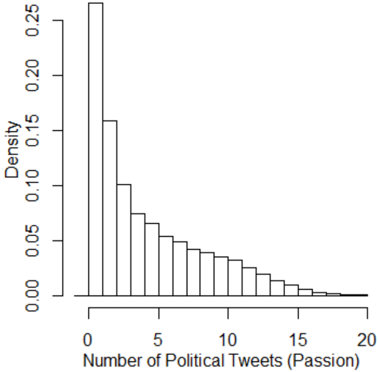
Twitter Data

Compare to a Truncated Geometric Distribution.



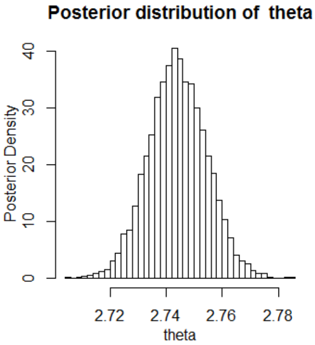
Twitter Data

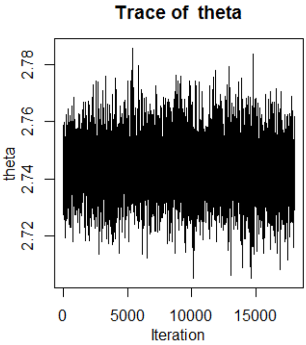
- More recently, we collected $\approx$ 19,000 political tweets. For each we classify sentiment and passion from last 20 tweets.
- We assume these are drawn from an *inflated* decay distribution.



Twitter Data

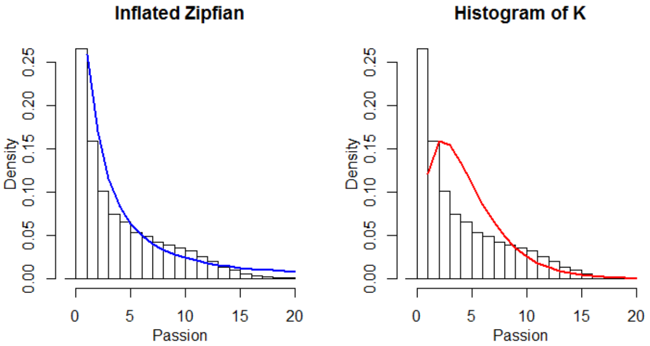
- Inflated-Zipfian Distribution fit's this nicely. $\theta = 2.744$





Twitter Data

- Inflated-Zipfian Distribution fit's this nicely. $\theta = 2.744$.
- Inflated-Truncated-Geometric, not so much. $\theta = 1.534$.



Estimating Γ

- We can consider $\Gamma_{true} = 0.5158$.
- The Naive Estimator just ignores the VR bias, and takes the mean.

$$\hat{\Gamma}_N = \frac{1}{19000}\sum_{i=1}^{19000} \mathcal{S}_i = 0.4782 \tag{14}$$

- Recall our strategy.

$$\Gamma = \frac{1}{H}\gamma(0) + (1 - \frac{1}{H})\Gamma^* \tag{15}$$

- We need to estimate $\hat{\gamma}(0)$ and $\hat{\Gamma}^*$.

Estimating Γ

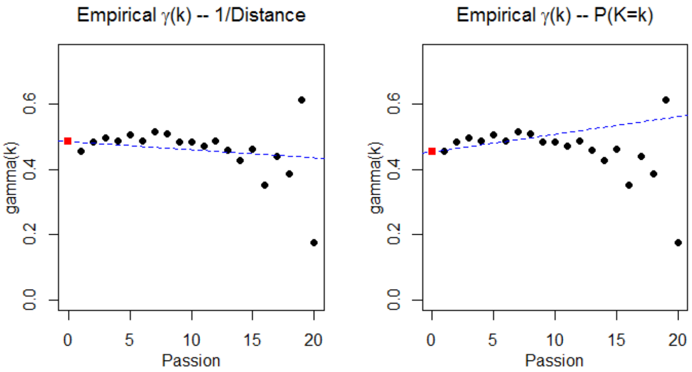
- We can obtain an estimate for Γ^* by weighting each person's contribution by the inverse passion. We call this the Inverse-Passion Adjustment (IPA).

$$\hat{\Gamma}^* = \frac{\sum_{i=1}^{19000} \mathcal{S}_i \mathcal{K}_i^{-1}}{\sum_{i=1}^{19000} \mathcal{K}_i^{-1}} \tag{16}$$

- This estimator is actually unbiased for Γ^*.
  - Proof pending.
  - Only checked for Zipfian Decay.

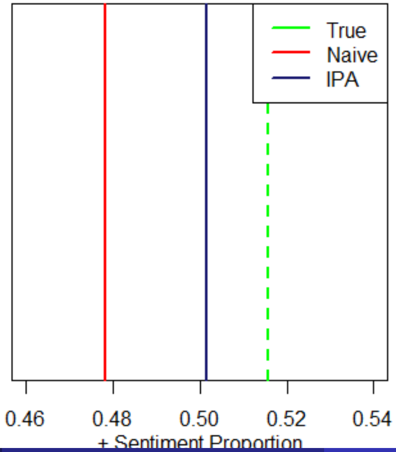
Estimating $\hat{\gamma}(0)$

- Our ability to estimate $\gamma(0)$ accurately depends heavily on the problem. We must be careful here.



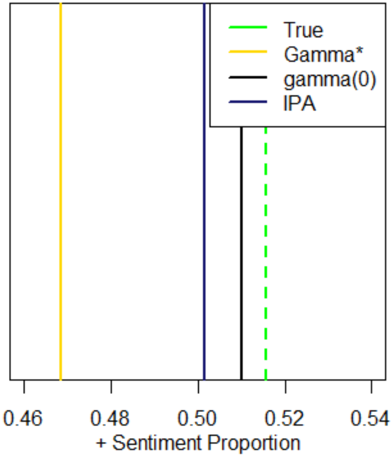
Results

- Although heavily dependent on choice of weighted regression, our estimator is able to reduce some of this bias.



Results

- What's actually happening?



Conclusions

- Simulation study shows that under certain conditions, the Classical (Naive) estimator can be heavily influenced by VR bias.
- Simulation study shows that our estimator can (in theory) eliminate this bias.
- The application to real data showed several limitations to the method.
  - Needs truly big data. Twitter's limitations make this difficult.
  - Possibly hurt by large misclassification error. We should improve the classifiers, and consider including binomial errors into the model.
  - As expected, $ga\hat{m}ma(0)$ may be impossible to fit reliably in many circumstances.