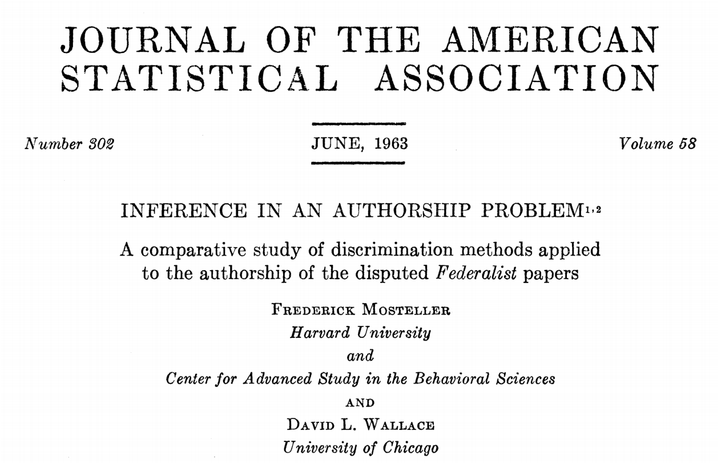
Measuring Rhetoric

Matt Taddy, Chicago Booth

faculty.chicagobooth.edu/matt.taddy/research

History: Text as data

Social science text-as-data from the 1960s:
author identification in the Federalist papers.



JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Number 302 JUNE, 1963 *Volume 58*

INFERENCE IN AN AUTHORSHIP PROBLEM[1,2]

A comparative study of discrimination methods applied to the authorship of the disputed *Federalist* papers

Frederick Mosteller
Harvard University
and
Center for Advanced Study in the Behavioral Sciences
AND
David L. Wallace
University of Chicago

M+W count words in papers by Hamilton and Madison,

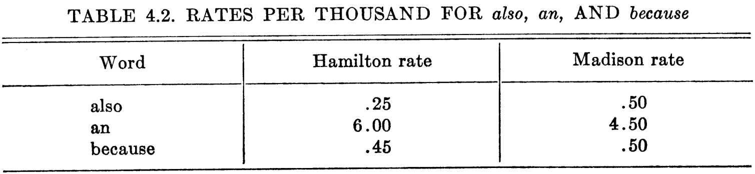


TABLE 4.2. RATES PER THOUSAND FOR *also, an,* AND *because*

Word	Hamilton rate	Madison rate
also	.25	.50
an	6.00	4.50
because	.45	.50

then fit models for counts|author (essentially what I use today!), and use Bayes rule to predict authors|counts on disputed work.

$$\mathrm{p}(\text{Hamilton} \mid \text{text}) \approx \frac{\mathrm{p}(\text{text} \mid \text{Hamilton})}{\mathrm{p}(\text{text} \mid \text{Madison}) + \mathrm{p}(\text{text} \mid \text{Hamilton})}$$

The ‘bag of words’

A ‘word’ is a self-contained meaningful token...

- Actual words: ‘all’, ‘world’, ‘stage’, ‘:-)’, ‘`#textdata`’.
- n-grams: ‘merely players’ (bi), ‘men and women’ (tri)
- complicated clauses: parts of speech, act-of-god.
- user selections on a website, domain ids in browser history

All we do is count them.

The remains state of the art!

Treat tokens for each doc as an i.i.d. sample.

Document i is summarized by counts c_{ij} for tokens $j = 1...d$.

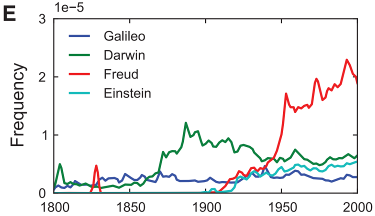
Dumb but works: extra rules aren't worth their complexity.

Text as data in Social Science

There's been an explosion of interest from social scientists.

Until very recently, one used pre-defined dictionaries.

Picking words: culturomics, Michel et al, Science 2011.



Psychosocial dictionaries, such as Harvard GI in *Tetlock 2007, Giving Content to Investor Sentiment* and others:

able, abundant, accept vs abandon, abrupt, absurd.

Topic Models

Techniques from stats and ML are beginning to filter through and researchers are estimating relationships *from the data*.

A large area of research has developed around *topic models*

$$\mathbf{c}_i \sim \mathrm{MN}(\omega_{i1}\boldsymbol{\theta}_1 + \ldots + \omega_{iK}\boldsymbol{\theta}_K, m_i)$$

a multinomial with probabilities $\sum_k \omega_{ik}\boldsymbol{\theta}_k$ and size $m_i = \sum_j c_{ij}$.

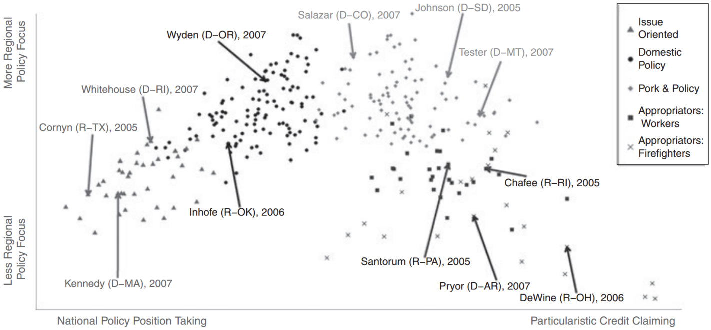
(Latent Dirichlet Allocation; Blei, Ng, Jordan 2003)

This is a factor model for count data.

Topics provide low-D structure, which the SS interprets.

Especially common in PoliSci; King, Grimmer, Quinn, ...

FIGURE 1 A Typology of Home Styles in the U.S. Senate



Grimmer 2013: fit latent topics in press releases
(e.g., ‘Iraq’, ‘Infastructure’) then investigate who uses what topic.

Structured topic models

The basic topic model finds *dominant sources of variation* in **C**.

In SocSci, this is often not what we're seeking (needle + haystack).

There is a huge industry on extensions to topic models that push the topics to be relevant or interpretable for specific questions.

supervised TM, dynamic TM, structural TM, IR TM ...

These model weights (ω) and topics (θ) as functions of covariates.

Lots of good and interesting work.

Multinomial Regression

Instead of jumping straight to latent structure, perhaps we can answer our questions by regressing the text onto observables '**v**'.

Massive response logistic regressions:

$$\mathbf{c}_i \sim \mathrm{MN}(\mathbf{q}_i, m_i) \;\; \text{with} \;\; q_{ij} = e^{\eta_{ij}} / \sum_l e^{\eta_{il}}$$

$\eta_{ij} = \alpha_j + \mathbf{v}_i'\boldsymbol{\varphi}_j$ is a 'log intensity' $\approx \mathrm{E}\log(c_{ij}/m_i)$

This is a regression like any other.

We will be estimating *partial correlations*, can build *random effects* and *interactions* into **v**, ... all our familiar regression ideas apply.

MN Inverse Regression + rejoinder, 2013. Political Sentiment on Twitter, 2013. Distributed MN Regression, 2015.

Distributed Multinomial Regression

A regression like any other, except the response is super HD.

We approximate the MN likelihood with *independent* Poissons:

$$c_{ij} \sim \mathrm{Po}(\ m_i e^{\eta_{ij}}\ )$$

$\Rightarrow$ you can estimate each regression fully independently!

This works because MN dependence is *only induced by totals*.

DMR is equivalent to MN logit in a variety of simple examples, and is shown empirically to perform well in more complex settings.

Everything in distribution: estimation, penalization, selection ...

More precisely, start from the Poisson:

$$c_{ij} \overset{ind}{\sim} \mathrm{Pois}\left(\exp\left[\mu_i + \eta_{ij}\right]\right)$$

where μ_i is a 'verbosity' nuisance parameter.

This model leads to

$$\Pr\left(\mathbf{c}_i \mid m_i\right) = \frac{\prod_j \mathrm{Po}\left(c_{ij}; \exp\left[\mu_i + \eta_{ij}\right]\right)}{\mathrm{Po}\left(m_i; \sum_l \exp\left[\mu_i + \eta_{il}\right]\right)} = \mathrm{MN}(\mathbf{c}_i;\ \mathbf{q}_i, m_i)$$

Thus, given m_i, Poisson and MN imply the same model.

DMR fixes $\hat{\mu}_i = \log m_i$, so LHD factorizes to independent Poissons.

More generally: for Big Data, consider using plug-in [marginal] estimates of parameters about which you have little uncertainty. Focus computation on the bits that are hard to measure.

Yelp Reviews

We'll illustrate using publicly available review data from Yelp.

- $n = 215{,}879$ reviews on 11,535 businesses by 43,873 users.
- taken around Phoenix AZ on January 19, 2013.
- $d = 13{,}938$ words in more than 20 reviews.

The reviews are marked with review, business, and user attributes: number of stars, user and business star averages, business type (333 overlapping), and the number of funny/useful/cool votes.

Each *word-j* intensity regression equation has

$$\eta_{ij} = \alpha_j + \mathbf{a}_i'\varphi_j^a + \mathbf{b}_i'\varphi_j^b$$

where we've resolved the meta-data attributes $\mathbf{V} = [\ \mathbf{A}\ \mathbf{B}\ ]$ into variables of primary interest and those viewed as controls

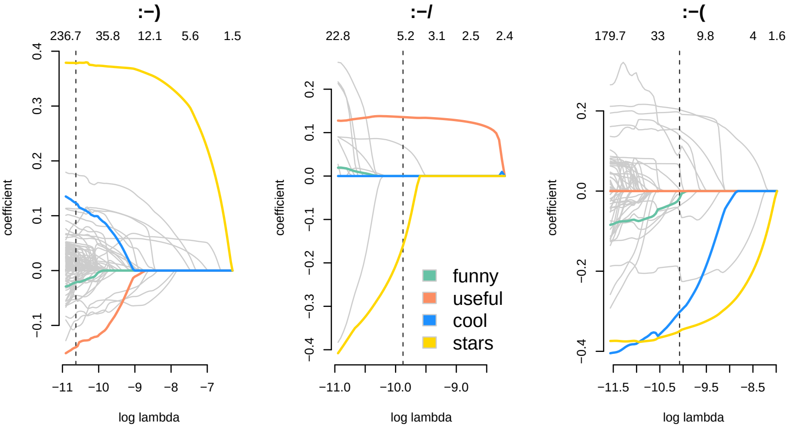
a: star and vote counts; 11 dimensional.

b: 400 categories and 11,500 business random effects.

Estimate parameters to minimize the penalized Poisson deviance

$$\hat{\alpha}_j, \hat{\boldsymbol{\varphi}}_j = \mathrm{argmin}\left\{ l(\alpha_j, \boldsymbol{\varphi}_j) + n\lambda \left[\sum_k \omega_{jk}^a |\varphi_{jk}^a| + \frac{1}{\tau} \sum_k \omega_{jk}^b |\varphi_{jk}^b| \right] \right\}$$

one-step estimator paths for concave regularization, 2015.



Poisson regression regularization paths under relative weight $\tau = 2$. AICc selection is marked.

Resolving correlated effects

Bigger τ gives fewer but *cleaner* nonzero terms.

	τ	$\lvert\hat{\varphi}\rvert_0$	**top ten words by loading**
+stars	marg		*great love amaz favorite deliciou best awesome alway perfect excellent*
	2	8440	*unmatch salute :-)) prik laurie pheonix trove banoffee exquisite sublime*
	20	3077	*heaven perfection gem divine amaz die superb phenomenal fantastic deliciousnes*
	200	508	*gem heaven awesome wonderful amaz fantastic favorite love notch fabulou*
-stars	marg		*not worst ask horrib minut rude said told would didn*
	2	8440	*rude livid disrespect disgrace inexcusab grosse incompet audacity unmelt acknowl*
	20	3077	*rude incompet unaccept unprofession inedib worst apolog disrespect insult acknowl*
	200	508	*worst horrib awful rude inedib terrib worse tasteles disgust waste*
funny	marg		*you that know like your yelp ... what don who*
	2	6508	*dimsum rue reggae acne meathead roid bong crotch peni fart*
	20	1785	*bitch shit god dude boob idiot fuck hell drunk laugh*
	200	120	*bitch dear god hell face shit hipst dude man kidd*
useful	marginal		*that yelp you thi know biz-photo like all http ://*
	2	5230	*fiancee rife dimsum maitre jpg poultry harissa bureau redirect breakdown*
	20	884	*biz-photo meow harissa www bookmark :-/ http :// (?), tip*
	200	33	*www http :// com factor already final immediate ask hope*

I think $\tau = 20$ strikes a good balance here.

Sufficient Reduction

What is the funny/useful/cool content of a review?

Coefficients $\mathbf{\Phi}$ are a linear map from text to attribute space.

They provide a *sufficient reduction*. For example,

$$\mathbf{a}_i \perp\!\!\!\perp \mathbf{c}_i \mid \mathbf{\Phi}^a \mathbf{c}_i, \mathbf{b}_i, m_i$$

where $\mathbf{\Phi}^a$ are loadings relevant to our ‘primary interest’ covariates.

In words: the 11 dimensional $\mathbf{z}_i = \mathbf{\Phi}^a \mathbf{c}_i$ contains all the information in the text that is *directly* relevant to $\mathbf{a}_i$, controlling for $\mathbf{b}_i$ and m_i.

Funniest and most useful 50-100 word review, as voted by Yelp users
(votes normalized by square root of review age).

I use to come down to Coolidge quite a bit and one of the cool things I use to do was come over here and visit the ruins. A great piece of Arizona history! Do you remember the Five C's? Well, this is cotton country. The Park Rangers will tell you they don't really know how old the ruins are, but most guess at around 600 years plus. But thanks to a forward thinking US Government, the ruins are now protected by a 70 foot high shelter. Trust me, it comes in handy in July and August, the two months I seem to visit here most. LOL. I would also recommend a visit to the bookstore. It stocks a variety of First Nation history, as well as info on the area. http://www.nps.gov/cagr/index.htm. While you are in Coolidge, I would recommend the Gallopin' Goose for drinks or bar food, and Tag's for dinner. Both are great!

50-100 word review with the most funny content,
as measured by SR projection $z_{\texttt{funny}} = \hat{\phi}'_{\texttt{funny}}\mathbf{c}$.

Dear La Piazza al Forno: We need to talk. I don't quite know how to say this so I'm just going to come out with it. I've been seeing someone else. How long? About a year now. Am I in love? Yes. Was it you? It was. The day you decided to remove hoagies from your lunch menu, about a year ago, I'm sorry, but it really was you...and not me. Hey... wait... put down that pizza peel... try to stay calm... please? [Olive oil container whizzing past head] Please! Stop throwing shit at me... everyone breaks up on social media these days... or haven't you heard? Wow, what a Bitch!

most funny by $z_{\texttt{funny}}/m$: *Holy Mother of God*

50-100 word review with the most useful content,
as measured by SR projection $z_{\texttt{useful}} = \hat{\phi}'_{\texttt{useful}}\mathbf{c}$.

We found Sprouts shortly after moving to town. There's a nice selection of Groceries & Vitamins. It's like a cheaper, smaller version of Whole Foods. [biz-photo] [biz-photo] We shop here at least once a week. I like their selection of Peppers....I like my spicy food! [biz-photo][biz-photo][biz-photo] Their freshly made Pizza isn't too bad either. [biz-photo] Overall, it's a nice shopping experience for all of us. Return Factor - 100%

most useful by $z_{\texttt{useful}}/m$: *Ask for Nick!*

The SR projections are based on *partial correlations*.

E.g., compare the correlation matrices

*attributes (***v***)*

	f	u	c	$\star$
funny	1	0.7	0.8	0
useful	0.7	1	0.9	0
cool	0.8	0.9	1	0
stars	0	0	0	1

*text projections (***z***)*

	f	u	c	$\star$
funny	1	-0.1	-0.7	-0.4
useful	-0.1	1	0.1	-0.2
cool	-0.7	0.1	1	0.5
stars	-0.4	-0.2	0.5	1

SR projections make great inputs to prediction algorithms: MNIR.

Measuring Segregation in High Dimensions

with Matt Gentzkow and Jesse Shapiro

There's a big lit on measuring change in polarization/segregation.

For example:

- The d-vector $\mathbf{c}_{tr} = [c_{tr1} \dots c_{trd}]$ counts the number of members of race r in each neighborhood j at time t.
  Or, compare across cities/countries/school districts/etc.
- Segregation measure maps counts $\{\mathbf{c}_{tr}\}_{r=1}^{R}$ into scalar s_t
- Use $s_1 \dots s_T$ to answer questions like

  "Has this city become more segregated over time?"

Today's talk is about how to build s_t when d is really big.

Current indices are derived from sets of stated desirable properties.

Example: Atkinson index (Frankel and Volij 2010)

$$s_t = 1 - \sum_j \sqrt{\frac{c_{t0j}}{m_{t0}} \; \frac{c_{t1j}}{m_{t1}}}$$

for two groups (e.g., races) $r = 0$ and $r = 1$, where $m_{tr} = \sum_j c_{trj}$.

This builds on an earlier literature that provides isolation, dissimilarity, mutual information, gini, and other indices.

Are we after properties of the sample or of the DGP?

That is, if the data are consistent with race-blind assignment do we want to say that segregation is low?

e.g., Cortese et al. (1976), Carrington and Troske (1997) both do.

Think of Atkinson as

$$\hat{s}_t = 1 - \sum_j \sqrt{\hat{p}_{t0j} \ \hat{p}_{t1j}}$$

where $\mathbf{p}_{trj}$ is the true probability that a member of r lives in j at t.

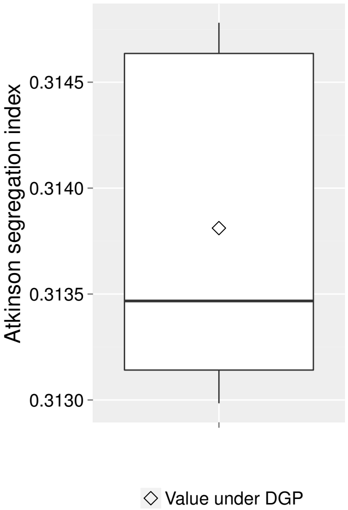
This distinction is unimportant if $\hat{s}_t \approx s_t$.

It is so for residential segregation because zipcodes are large.

But estimation bias becomes very important as units get smaller...

Number of Districts Small

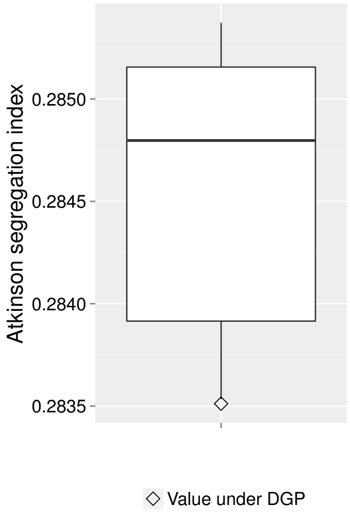
1 million students



100 districts.

Number of Districts Medium

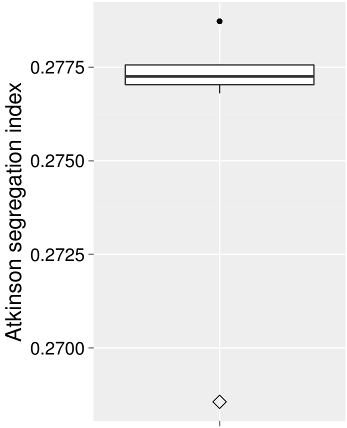
1 million students



1000 districts

Number of Districts Large

1 million students

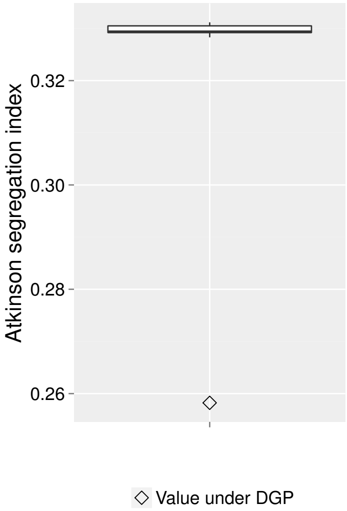


10,000 districts

◇ Value under DGP

Number of Districts Very Large

1 million students



100,000 districts

This paper

We build a utility-based model of assignment to groups, then define a segregation measure within this model.

⇒ a Big-response-dimension multinomial logistic regression.

- Use penalization to control finite-sample behavior
- Implement distributed estimation for scalability

We control for 100s of covariates and 1000s of random effects.

Applications to polarization/segregation in

- congressional text (gop vs. dem, south vs. north)
- internet browsing (white vs. black)
- grocery store purchases (college educated vs. not)

How has polarization in political speech changed over time?

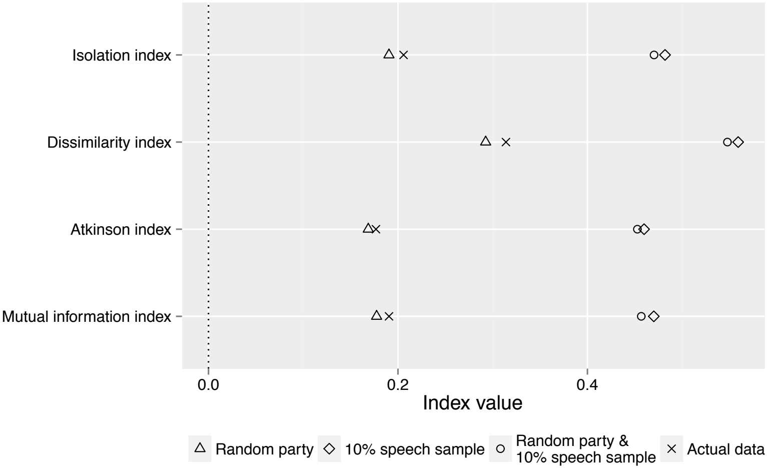
The concept here is of speakers across parties:

- using different words to describe the same thing
  `tax.cut`/`tax.break`, `war.on.terror`/`war.in.iraq`
- choosing to focus on different substantive topics
  `stem.cell`, `african.american`, `soldier.sailor`.

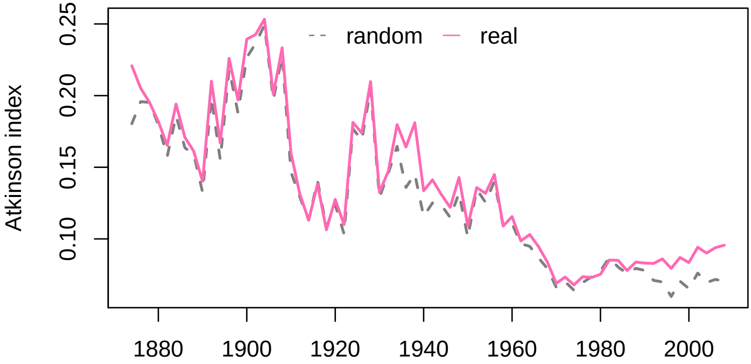
And doing this because of party membership or ideology.

Say $r_{it} = 1$ if speaker i is republican at t, 0 if democrat.

Existing Approaches

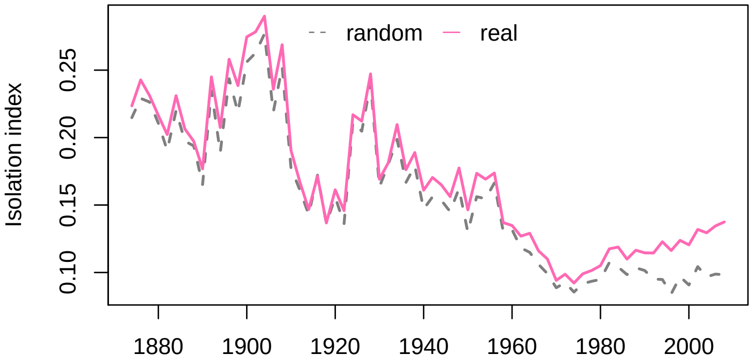


Atkinson



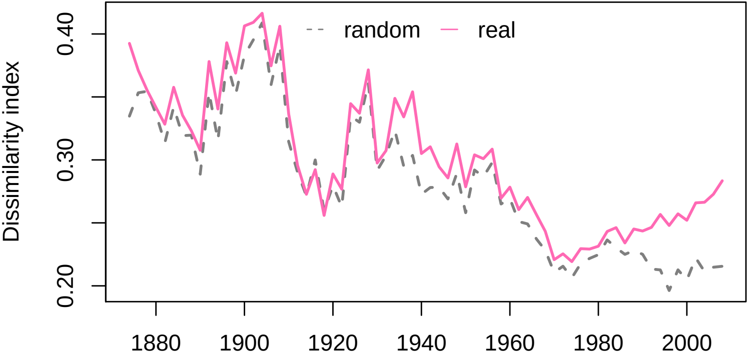
$$\hat{s}_t = 1 - \sum_j \sqrt{\hat{p}_{t1j} \; \hat{p}_{t0j}}$$

Isolation



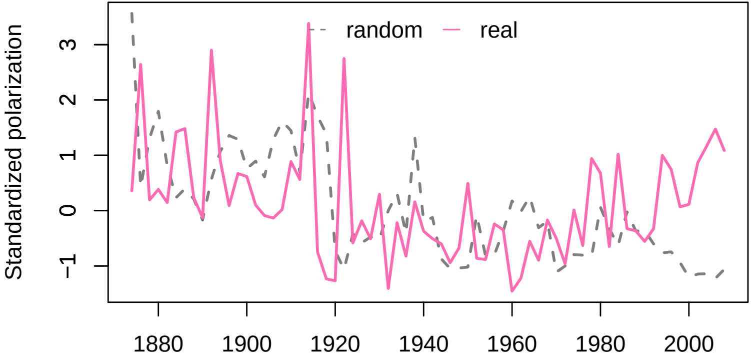
$$\hat{s}_t = \sum_j (\hat{p}_{t1j} - \hat{p}_{t0j}) \frac{c_{t0j}}{c_{t1j} + c_{t0j}}$$

Dissimilarity



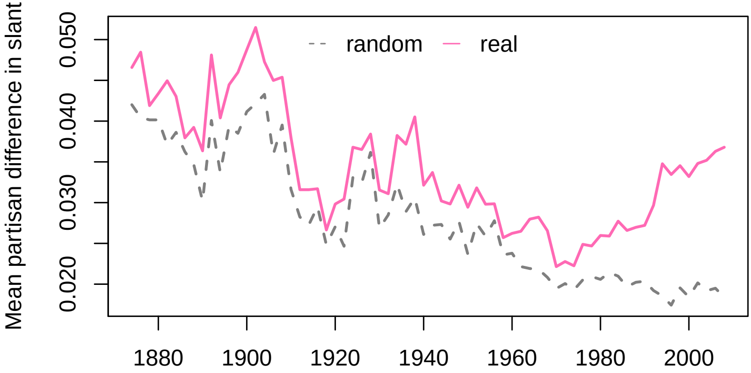
$$\hat{s}_t = \frac{1}{2} \sum_j \left| \hat{p}_{t1j} - \hat{p}_{t0j} \right|$$

Brookings



Jensen et al. (2010)

Slant



Gentzkow and Shapiro (2010)

Instead, we'll fit our big Multinomial logit model:

Each individual i in period t
makes m_{it} choices over units j to maximize utility

$$\eta_{itj} + \varepsilon_{itj} = \alpha_{jt} + \mathbf{u}'_{it}\boldsymbol{\gamma}_{jt} + \varphi'_{jt} r_{it} + \varepsilon_{ijt}$$

where:

- α_{jt} is unit-specific utility intercept
- $\mathbf{u}_{it}$ are covariates and $\boldsymbol{\gamma}_{jt}$ are associated loadings
- $r_{it} \in \{0, 1\}$ is an indicator for group membership and φ_{jt} are associated loadings.
- ε_{ijt} is T1EV random utility component

Motivated by congressional example define the partisanship z_{it} of individual i at time t (after observing $m_{it} = \sum_j c_{itj}$ choices) as

$$z_{it} = \boldsymbol{\varphi}_t' \mathbf{c}_{it} / m_{it},$$

utility gain to $r = 1$ relative to an $r = 0$ from individual i's choices.

z_{it} is also a model-based sufficient statistic for group membership:

$$r_{it} \perp\!\!\!\perp \mathbf{c}_{it} \mid z_{it}, \mathbf{u}_{it}, m_{it}.$$

'Sufficient projection' z_{it} contains all info in $\mathbf{c}_{it}$ relevant to r_i.

Finally, we can measure segregation (polarization) as difference in mean partisanship between those with $r = 1$ and those with $r = 0$.

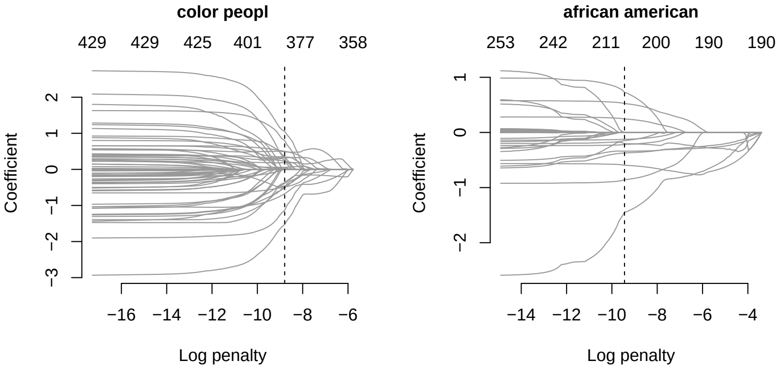
Model: $\mathbf{c}_{it} \sim \mathrm{MN}(\mathbf{p}_{it}, m_{it})$ with $p_{itj} = e^{\eta_{itj}} / \sum_l e^{\eta_{itl}}$.

r_{it} indicates gop/dem and φ_{tj} moves 'smoothly' in t.

Partisanship is defined as segregation in speech by party that is not explainable by our set of controls.

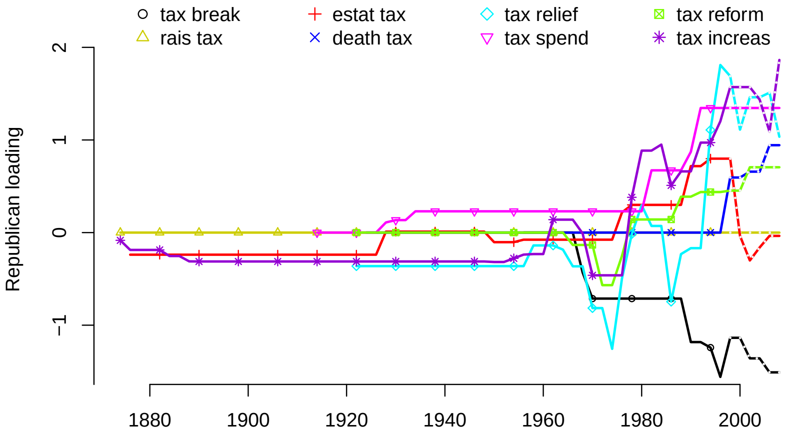
- $\mathbf{u}_{it}$ contain state, chamber, indicator for majority party.
- We allow region effects to vary with time
- check robustness to speaker random effects.

Poisson regression regularization paths



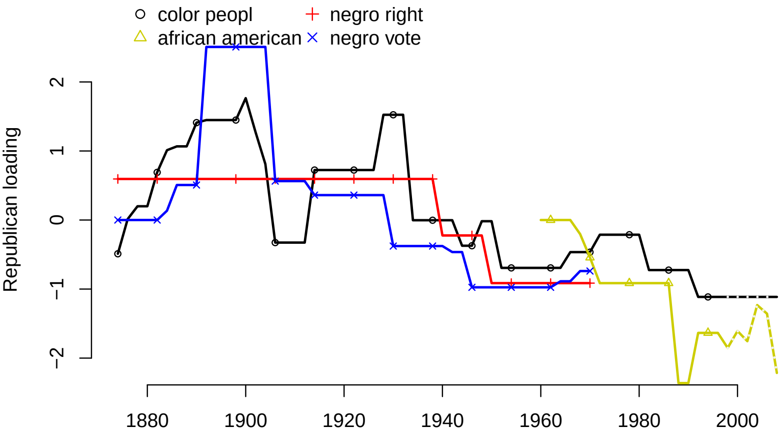
We've run a separate Poisson regression for each phrase.
The code runs a MapReduce routine using `dmr` for R.
BIC selection occurs within reducer, and is marked here.

Dynamic Phrase-Party Loadings: Tax



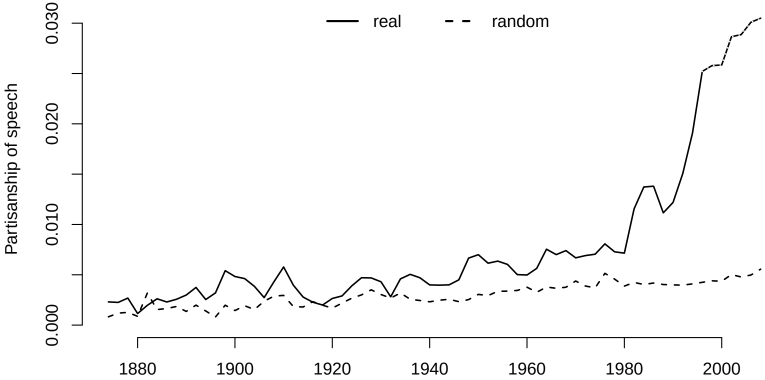
The resulting fit has φ_{tj} changing as a step function in t.

Dynamic Phrase-Party Loadings: Race



For this example, partisanship is robust to fixing $\varphi_{tj} = \varphi_j$.

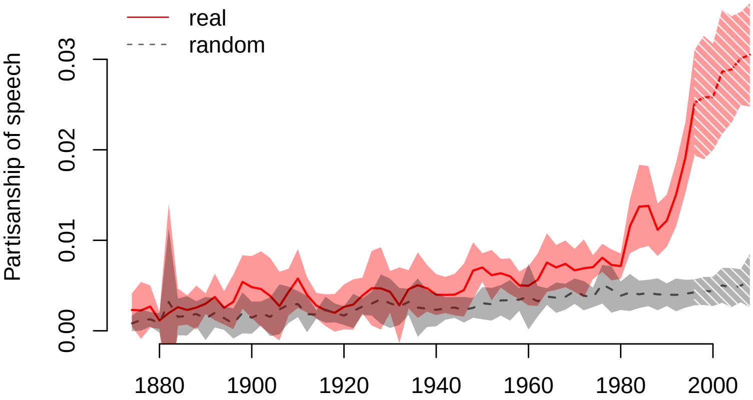
Polarization Results: Baseline Specification



We take $\bar{z}_{rt} = \frac{1}{n_{rt}} \sum_{i:r_i=r} z_{it}$ for each party in each session, and the difference is our partisanship index.

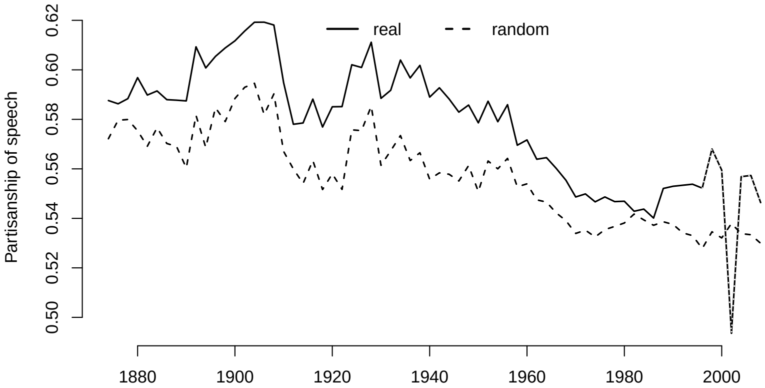
Partisanship of rhetoric has exploded since around 1980.

Nonparametric Bootstrap



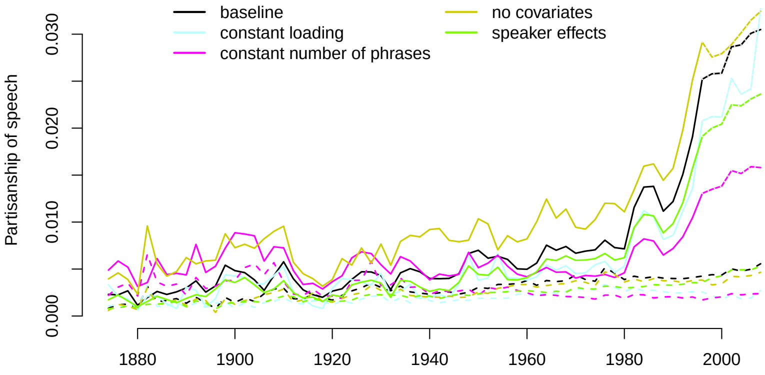
1980 is also when real partisanship becomes more than $2SE$ away from that for random permutations.

Penalization is a necessary ingredient



These are the results corresponding to the end of each lasso path (most complex model)

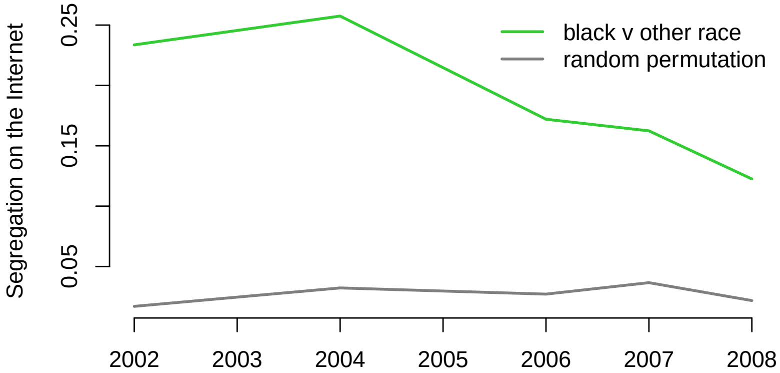
Robustness



With BIC penalty selection, the same shape holds under a variety of specifications. But notice: the controls do make a difference.

Another application: racial segregation on the internet

ComScore browser histories on the websites used in Gentzkow + Shapiro (2011), making visit-counts a function of r_{it} black/white.



Segregation online is high, but has been dropping since 2004.

wrap up

Big picture: Give regression a chance!

Everything here – random effects, synthetic controls, lasso variable selection, utility interpretations – is common in regression.

For example, Segregation Econometrics:

We've written down a model for choices,
and defined segregation in terms of that model.

We can specify time dynamics and covariates inside this model.

The same techniques that allow machine learners to avoid overfit in prediction can be used to recover representative model fits.

Thanks!