Learning from Text

SoFiE Summer School on Machine Learning & Finance

Asaf Manela

Washington University in St. Louis

July 2018

Motivation

- Digital text is increasingly available to social scientists
  - Newspapers, blogs, regulatory fillings, congressional records ...
- Unlike data often used by economists
  - Text is ultra high-dimensional
  - Phrase counts are sparse
- Statistical learning from text requires
  - Machine learning techniques
  - Scalable algorithms

Outline

- Text as data
- Supervised learning from text
- Text selection
- Example applications

Textual analysis

1. Represent raw text $\mathscr{D}$ as numerical array $\boldsymbol{c}$
2. Map $\boldsymbol{c}$ to predicted value $\hat{\boldsymbol{v}}$ of unknown outcomes $\boldsymbol{v}$
3. Use $\hat{\boldsymbol{v}}$ in subsequent descriptive or causal analysis

Text data is inherently high-dimensional

Bag-of-words representation

Documents $\mathscr{D}$

1: Digital text is available.
2: Text is selected!
⋮

$\Rightarrow$

Document-term matrix $\boldsymbol{c}$

	Digital	text	is	available	Text	selected	…
1:	1	1	1	1	0	0	
2:	0	0	1	0	1	1	
			⋮				⋱

Text data is inherently high-dimensional

Preprocessing reduces dimensionality somewhat, requires careful judgment

Documents $\mathscr{D}$

1: Digital text is available.
2: Text is selected!
⋮

$\Rightarrow$

Document-term matrix $\boldsymbol{c}$

	digit	text	avail	select	…
1:	1	1	1	0	
2:	0	1	0	1	
		⋮			⋱

- Removed stopwords (is), punctuation, lowercased, stemmed

Text data is inherently high-dimensional

Higher order n-grams provide more context, but increase dimension exponentially

Documents $\mathscr{D}$

1: Digital text is available.
2: Text is selected!
$\vdots$

$\Rightarrow$

Document-term matrix $\boldsymbol{c}$

	digit text	text is	is avail	is select	...
1:	1	1	1	0	
2:	0	1	0	1	
		$\vdots$			$\ddots$

Text regression is prone to overfit

- $\boldsymbol{c}_i$ vector of counts in d categories for observation i
  - e.g. c_{ij} is date i newspaper mentions of phrase j ("world war")
- $\boldsymbol{v}_i$ vector of p covariates
  - e.g. intermediary capital ratio, realized variance on date i
- Let $v_{iy} \in \boldsymbol{v}_i$ be a target variable
  - e.g. intermediary capital ratio
- Because $d \gg n$, we cannot run an OLS regression

$$v_{iy} = \beta_0 + [\boldsymbol{c}_i, \boldsymbol{v}_{i,-y}]' \boldsymbol{\beta} + \varepsilon_i$$

Dictionary-based methods

- Simply reduce text to small set of predefined word lists
- Examples:
  - Positive/negative words (Tetlock, 2007)
  - Policy uncertainty word combinations (Baker et al., 2016)
- Inferior to regularized regression (Manela and Moreira, 2017)

It ain't what you don't know that gets you into trouble.
It's what you know for sure that just ain't so.

– Mark Twain?

Dictionary-based methods

- Simply reduce text to small set of predefined word lists
- Examples:
  - Positive/negative words (Tetlock, 2007)
  - Policy uncertainty word combinations (Baker et al., 2016)
- Inferior to regularized regression (Manela and Moreira, 2017)

It ain't what you don't know that gets you into trouble.
It's what you know for sure that just ain't so.

– Mark Twain?

Dictionary-based methods

- Simply reduce text to small set of predefined word lists
- Examples:
  - Positive/negative words (Tetlock, 2007)
  - Policy uncertainty word combinations (Baker et al., 2016)
- Inferior to regularized regression (Manela and Moreira, 2017)

It ain't what you don't know that gets you into trouble.
It's what you know for sure that just ain't so.

– Mark Twain?

Topic models and other latent factor methods

- Reduce text to $k \ll d$ latent factors
- Latent Dirichlet Allocation (LDA)
  - e.g. Jegadeesh and Wu (2017)
- Principal Component Regression (PCR)
  - e.g. Foster et al. (2013)

Regularization

Letting the data speak directly to the problem at hand

- Regularization / penalization of non-zero or large coefficients helps solve *ill-posed* problems

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} L\left(v_{iy}, f\left(\boldsymbol{c}_i, \boldsymbol{v}_{i,-y}\right)\right) + \lambda J\left(f\right)$$

for some

- $L\left(y, f\left(x\right)\right)$ loss function
- $J\left(f\right)$ penalty functions
- $\lambda > 0$ penalty parameter

Support vector regression (Vapnik, 2000)

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} L_\epsilon \left(v_{iy} - \beta_0 - \left[\boldsymbol{c}_i, \boldsymbol{v}_{i,-y} \right]' \boldsymbol{\beta} \right) + \lambda \left| \boldsymbol{\beta} \right|^2$$

where

$$L_\epsilon(r) = \begin{cases} 0 & |r| < \epsilon \\ |r| - \epsilon & \text{otherwise} \end{cases}$$

- Examples:
  - Backcast VIX to 1890 with WSJ (Manela and Moreira, 2017)
  - Predict accruals with 10k's (Frankel et al., 2016)
- Pro: can handle massive feature spaces
- Con: cannot concentrate on individual covariates

Text inverse regression (Taddy, 2013)

- A text inverse regression approach would instead
  1. Regress word counts on covariates

$$\boldsymbol{c}_i = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \boldsymbol{v}_i \qquad \text{(backward regression)}$$

  2. Construct low dimensional projection into v_{iy} direction

$$z_{iy} \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(sufficient reduction projection)}$$

  3. Regress target variable on z_{iy} and other covariates

$$v_{iy} = \beta_0 + \left[z_{iy}, \boldsymbol{v}_{i,-y}\right]' \boldsymbol{\beta} + \varepsilon_i \qquad \text{(forward regression)}$$

- $d + p - 1$ dimensional regression reduced to $p + 1$ dimensional!
- z_{iy} summarizes all textual information relevant for prediction

Text inverse regression (Taddy, 2013)

- A text inverse regression approach would instead
  1. Regress word counts on covariates

$$\boldsymbol{c}_i = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \boldsymbol{v}_i \qquad \text{(backward regression)}$$

  2. Construct low dimensional projection into v_{iy} direction

$$z_{iy} \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(sufficient reduction projection)}$$

  3. Regress target variable on z_{iy} and other covariates

$$v_{iy} = \beta_0 + \left[z_{iy}, \boldsymbol{v}_{i,-y}\right]' \boldsymbol{\beta} + \varepsilon_i \qquad \text{(forward regression)}$$

- $d + p - 1$ dimensional regression reduced to $p + 1$ dimensional!
- z_{iy} summarizes all textual information relevant for prediction

Text inverse regression (Taddy, 2013)

- A text inverse regression approach would instead
  1. Regress word counts on covariates

$$\boldsymbol{c}_i = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \boldsymbol{v}_i \qquad \text{(backward regression)}$$

  2. Construct low dimensional projection into v_{iy} direction

$$z_{iy} \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(sufficient reduction projection)}$$

  3. Regress target variable on z_{iy} and other covariates

$$v_{iy} = \beta_0 + \left[z_{iy}, \boldsymbol{v}_{i,-y}\right]' \boldsymbol{\beta} + \varepsilon_i \qquad \text{(forward regression)}$$

- $d + p - 1$ dimensional regression reduced to $p + 1$ dimensional!
- z_{iy} summarizes all textual information relevant for prediction

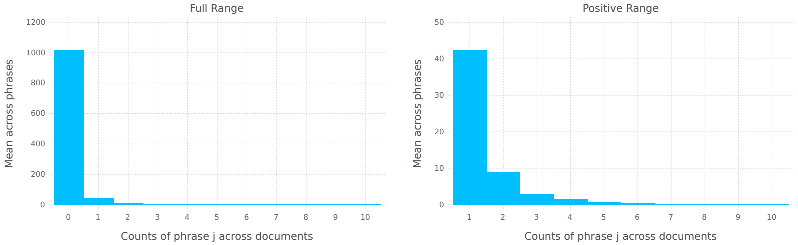
Other methods

- Nonlinear learning methods, like random forests and (deep) neural nets can potentially improve
  - Success may depend on large number of observations
  - Interpretation may be challenging
- See Gentzkow, Kelly, and Taddy (2017a) for a recent survey
- In the remainder I focus on what I find most promising for economics and finance

Text selection (Kelly, Manela, and Moreira, 2018)

- Text is often selected by journalists, speechwriters, and others who cater to an audience with limited attention
- Hurdle Distributed Multiple Regression (HDMR)
  - Highly scalable approach to inference from big counts data
  - Includes an economically-motivated selection equation
  - Especially useful when cover/no-cover choice is separate or more interesting than coverage quantity
- Applications using newspaper coverage for prediction
  1. Backcast intermediary capital ratio (He-Kelly-Manela 2017 JFE)
  2. Forecast macroeconomic series (Stock-Watson 2012 JBES)

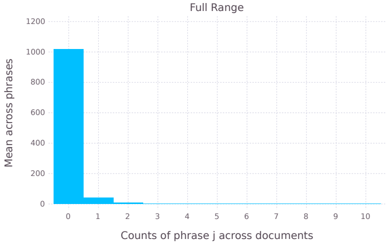
Why would we need a hurdle?

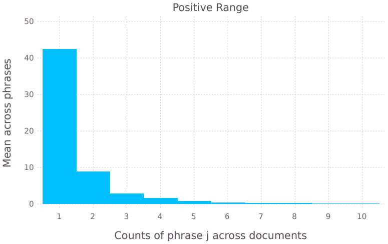


Wall Street Journal, monthly front page text, July 1926 to February 2016

- Statistics: hurdle better describes text data
  - Text data often has many more zeros than predicted by Poisson
- Economics: text is selected
  - Publishers cater to a boundedly rational reader (Gabaix, 2014)
  - Politicians select phrases that resonate with voters (Gentzkow, Shapiro, and Taddy, 2017b)
  - Censored or socially taboo words (Michel et al., 2011)
  - Fixed cost of introducing new terms, low marginal cost

Why would we need a hurdle?

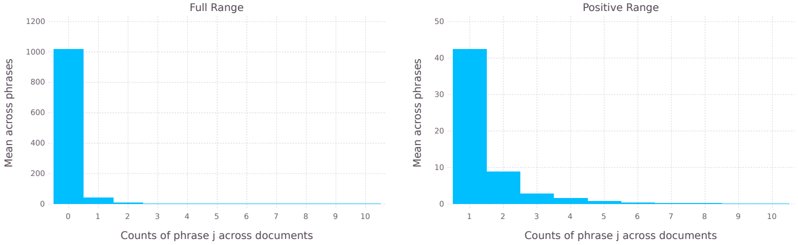




Wall Street Journal, monthly front page text, July 1926 to February 2016

- Statistics: hurdle better describes text data
  - Text data often has many more zeros than predicted by Poisson
- Economics: text is selected
  - Publishers cater to a boundedly rational reader (Gabaix, 2014)
  - Politicians select phrases that resonate with voters (Gentzkow, Shapiro, and Taddy, 2017b)
  - Censored or socially taboo words (Michel et al., 2011)
  - Fixed cost of introducing new terms, low marginal cost

Why would we need a hurdle?



Wall Street Journal, monthly front page text, July 1926 to February 2016

- Statistics: hurdle better describes text data
  - Text data often has many more zeros than predicted by Poisson
- Economics: text is selected
  - Publishers cater to a boundedly rational reader (Gabaix, 2014)
  - Politicians select phrases that resonate with voters (Gentzkow, Shapiro, and Taddy, 2017b)
  - Censored or socially taboo words (Michel et al., 2011)
  - Fixed cost of introducing new terms, low marginal cost

Text selection model

With sparse text, extensive margin may be more informative than intensive margin

- We suggest a text selection model instead
  1. Two part text selection model for counts

$$\boldsymbol{h}_i^* = f\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right) + \boldsymbol{\omega}_i \qquad \text{(Inclusion)}$$

$$\boldsymbol{c}_i^* = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \boldsymbol{v}_i \qquad \text{(Repetition)}$$

$$\boldsymbol{c}_i = \boldsymbol{c}_i^* \times \mathbf{1}\left(\boldsymbol{h}_i^* > 0\right) = \boldsymbol{c}_i^* \times \boldsymbol{h}_i \qquad \text{(Observation)}$$

  2. Construct two low dimensional projections into v_{iy} $(= w_{iy})$

$$z_{iy}^0 \equiv \sum_j \hat{\delta}_{jy} h_{ij} \qquad \text{(SR projection for inclusion)}$$

$$z_{iy}^+ \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(SR projection for repetition)}$$

  3. Regress target variable on z_{iy}^+, z_{iy}^0 and other covariates

$$v_{iy} = \beta_0 + \left[z_{iy}^0, z_{iy}^+, \boldsymbol{w}_{i,-y}, \boldsymbol{v}_{i,-y}\right]' \boldsymbol{\beta} + \varepsilon_i \qquad \text{(forward regression)}$$

- $d + p - 1$ dimensional regression reduced to $p + 2$ dimensional!

Text selection model

With sparse text, extensive margin may be more informative than intensive margin

- We suggest a text selection model instead
  1. Two part text selection model for counts

$$\boldsymbol{h}_i^* = f\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right) + \boldsymbol{\omega}_i \quad \text{(Inclusion)}$$

$$\boldsymbol{c}_i^* = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \boldsymbol{v}_i \quad \text{(Repetition)}$$

$$\boldsymbol{c}_i = \boldsymbol{c}_i^* \times \mathbf{1}\left(\boldsymbol{h}_i^* > 0\right) = \boldsymbol{c}_i^* \times \boldsymbol{h}_i \quad \text{(Observation)}$$

  2. Construct two low dimensional projections into v_{iy} $(= w_{iy})$

$$z_{iy}^0 \equiv \sum_j \hat{\delta}_{jy} h_{ij} \quad \text{(SR projection for inclusion)}$$

$$z_{iy}^+ \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \quad \text{(SR projection for repetition)}$$

  3. Regress target variable on z_{iy}^+, z_{iy}^0 and other covariates

$$v_{iy} = \beta_0 + \left[z_{iy}^0, z_{iy}^+, \boldsymbol{w}_{i,-y}, \boldsymbol{v}_{i,-y}\right]' \boldsymbol{\beta} + \varepsilon_i \quad \text{(forward regression)}$$

- $d + p - 1$ dimensional regression reduced to $p + 2$ dimensional!

Text selection model

With sparse text, extensive margin may be more informative than intensive margin

- We suggest a text selection model instead
  1. Two part text selection model for counts

$$\boldsymbol{h}_i^* = f\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right) + \boldsymbol{\omega}_i \quad \text{(Inclusion)}$$

$$\boldsymbol{c}_i^* = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \boldsymbol{v}_i \quad \text{(Repetition)}$$

$$\boldsymbol{c}_i = \boldsymbol{c}_i^* \times \mathbf{1}\left(\boldsymbol{h}_i^* > 0\right) = \boldsymbol{c}_i^* \times \boldsymbol{h}_i \quad \text{(Observation)}$$

  2. Construct two low dimensional projections into v_{iy} $(= w_{iy})$

$$z_{iy}^0 \equiv \sum_j \hat{\delta}_{jy} h_{ij} \quad \text{(SR projection for inclusion)}$$

$$z_{iy}^+ \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \quad \text{(SR projection for repetition)}$$

  3. Regress target variable on z_{iy}^+, z_{iy}^0 and other covariates

$$v_{iy} = \beta_0 + \left[z_{iy}^0, z_{iy}^+, \boldsymbol{w}_{i,-y}, \boldsymbol{v}_{i,-y}\right]' \boldsymbol{\beta} + \varepsilon_i \quad \text{(forward regression)}$$

- $d + p - 1$ dimensional regression reduced to $p + 2$ dimensional!

Hurdle distributed multiple regression (HDMR)

- Scale of text data requires convenient functional forms
- DMR uses independent Poissons to approximate the multinomial, one for each phrase
- We replace these Poissons with Hurdles (Mullahy, 1986)
- Hurdle model decomposes into two independent regressions
  1. Inclusion coefs. estimated from coverage indicators $\boldsymbol{h}_j$ and covariates $\boldsymbol{w}_i$
  2. Repetition coefs. estimated from positive counts $\boldsymbol{c}_j$ and covariates $\boldsymbol{v}_i$
- Can be distributed further!
- Lasso (L_1) regularization for both parts to avoid overfit

Selection bias

- Coefficients are biased if we use DMR on selected text data
- Severe bias if omitted variable in $\boldsymbol{w}$ is correlated with $\boldsymbol{v}$
- For example, suppose:
  - FIFA World Cup crowds out financial news (limited attention)
  - ... and reduces market vol (traders watch it too)
  - Omitting it would yield biased effect of vol on financial news

Intermediary capital ratio (ICR)

- Intermediary asset pricing
  - Theory (Brunnermeier-Pedersen 2009 RFS, He-Krishnamurthy 2013 AER; Brunnermeier-Sannikov, 2014 AER)
  - Evidence (Adrian-Etula-Muir, 2014 JF; He-Kelly-Manela, 2017 JFE; Muir, 2017 QJE; Haddad-Muir, 2018)
- He-Kelly-Manela (2017 JFE):
  - Intermediary capital ratio (ICR) is the aggregate market capital ratio of NY Fed primary dealers
  - Innovations to the ICR price many asset classes
  - Suggestive results on predictive ability limited by short time-series starting 1970
- Can we backcast the ICR using historical newspaper text?
- Does high ICR predict low future market returns?

Data

Front-page titles and abstracts of the *Wall Street Journal*, 1890-2016

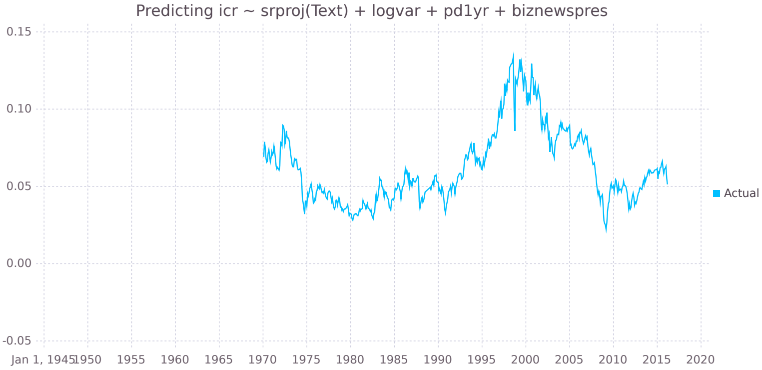
Date	Title	Abstract
2008-09-16	AIG Faces Cash Crisis As Stock Dives 61%	American International Group Inc. was facing a severe cash ...
2008-09-16	AIG, Lehman Shock Hits World Markets ...	The convulsions in the U.S. financial system sent markets ...
2008-09-16	Business and Finance	Central banks around the world pumped cash into money ...
2008-09-16	Keeping Their Powder Dry: Draft Boards ...	The Selective Service System has the awkward task of ...
2008-09-16	Old-School Banks Emerge Atop New ...	Banks are heading "back to basics – to, if you like, the core ...
2008-09-16	World-Wide	Thailand's ruling party chose ousted leader Thaksin's ...

HDMR approach to news implied intermediary capital ratio

- We use HDMR to backcast missing values of ICR with WSJ text + realized vol + price-dividend ratio
- Heckman selection models are non-parametrically identified
  - If a continuous variable enters the selection equation but can be excluded from second equation (Gallant-Nychka, 1984)
  - Proving such a result can be useful, but left for future work
- We seek a shifter for the inclusion decision
  - News pressure (Eisensee and Stromberg, 2007), starts 1967
  - Business news pressure (Manela, 2014), starts 1945
  - Assumption: excluded from repetition equation

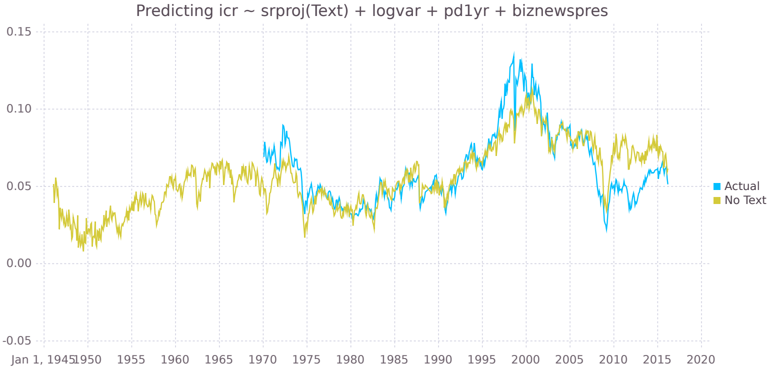
News implied intermediary capital ratio

ICR is available only since 1970 because dealers used to be private



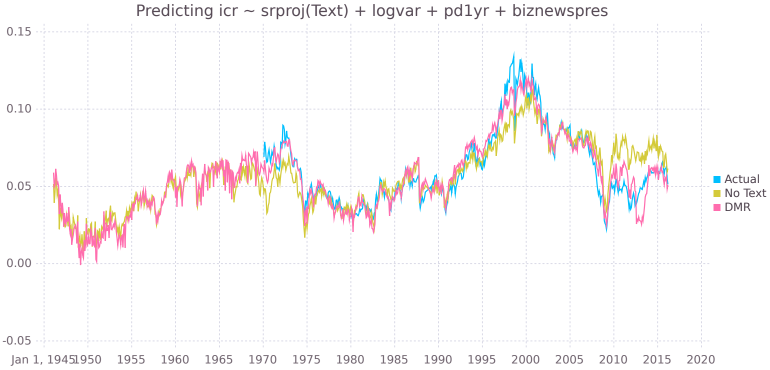
News implied intermediary capital ratio

First stab may be to fit using realized variance and price-dividend ratio without text



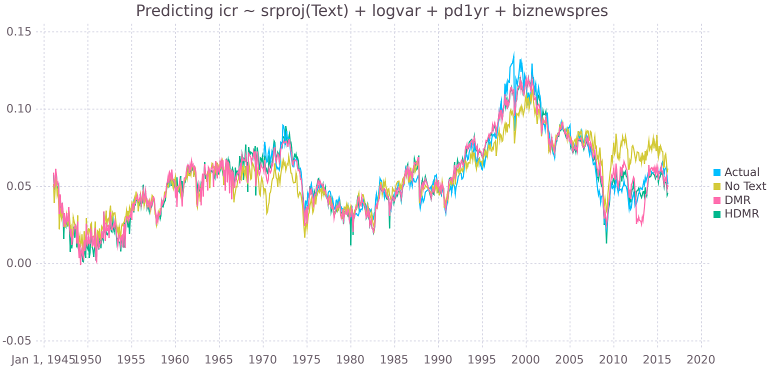
News implied intermediary capital ratio

DMR gives a different predicted series exploiting the text



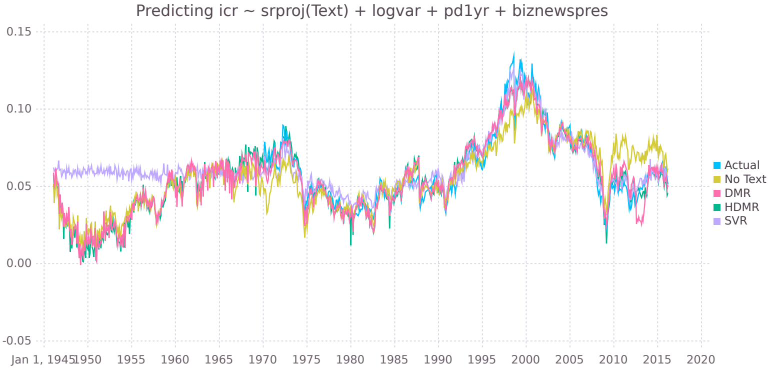
News implied intermediary capital ratio

HDMR uses same information as DMR but separates extensive from intensive margin



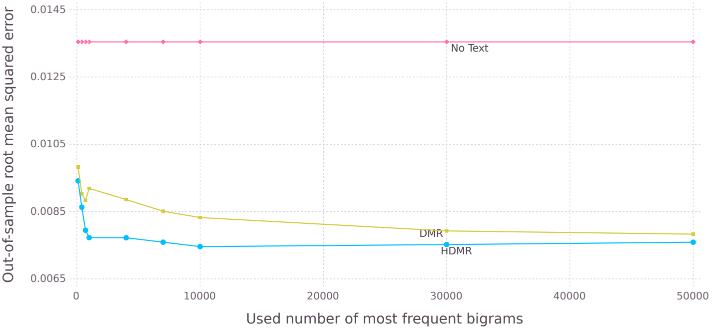
News implied intermediary capital ratio

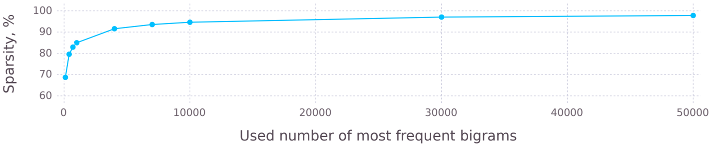
Support Vector Regression of Manela-Moreira (2017) cannot concentrate on covariates



Out-of-sample prediction of ICR with text and covariates

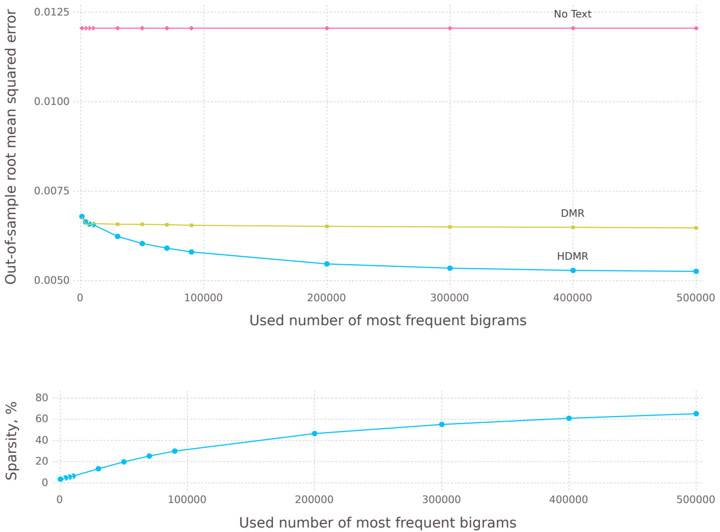
HDMR's out-of-sample fit advantage changes with text sparsity





Denser text: HDMR's advantage increases with sparsity

Full WSJ monthly phrase counts, January 1990 to December 2010



News-implied ICR predicts market returns

Consistent with He-Krishnamurthy (2013), high ICR means low risk premium

Dep. Var:	$r^{em}_{t\to t+1}$		$r^{em}_{t\to t+3}$		$r^{em}_{t\to t+6}$		$r^{em}_{t\to t+12}$	
z^0_{t-1}	-0.05		-0.05		-0.04		-0.04	
	(-2.49)		(-2.93)		(-2.49)		(-2.58)	
z^+_{t-1}	0.07		0.10		0.05		0.04	
	(1.28)		(2.13)		(1.38)		(0.89)	
z^{dmr}_{t-1}		-0.02		-0.02		-0.02		-0.02
		(-1.69)		(-1.48)		(-1.49)		(-1.77)
rv_{t-1}	0.04	0.02	0.23	0.19	0.20	0.18	0.10	0.10
	(0.22)	(0.10)	(1.37)	(1.15)	(1.32)	(1.16)	(1.04)	(0.92)
pd_{t-1}	-1.37	-1.15	-1.42	-1.18	-1.32	-1.16	-1.26	-1.11
	(-3.19)	(-2.74)	(-3.23)	(-2.73)	(-3.17)	(-2.74)	(-2.79)	(-2.43)
$NewsPressure_{t-1}$	0.03	0.01	0.02	0.00	0.00	-0.01	0.00	-0.01
	(1.02)	(0.43)	(0.57)	(-0.13)	(-0.05)	(-0.56)	(-0.16)	(-0.63)
R-squared, %	1.46	0.96	4.81	3.12	7.96	6.20	13.81	11.34
Obs	834	834	832	832	829	829	823	823

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

	Frequent phrases with the most positive loadings
icr^0	labor letter, busi bulletin, tax report, washington wire, gross net, nation recoveri, presid clinton, trend take, job trend, life job
rv^0	barrel dow, bushel wheat, yr trea, trea yld, dow jone, aig spot, futur dj, outstand stock, aig futur, stock nyse
pd^0	barack obama, yr trea, trea yld, technolog journal, insid journal, journal insid, nasdaq amp, commod oil, aig futur, weekend journal
np^0	washington wire, busi bulletin, tax report, labor letter, report steel, journal special, substanti gain, rubber tire, presid reagan, life job
icr^+	confer washington, ounc dow, miami beach, presid clinton, survey found, bosnia muslim, clinton health, presid slobodan, serb croat, ba
rv^+	west africa, amp unfil, falun gong, stock market, republican guard, composit index, john mccain, dow jone, jone industri, abu dhabi
pd^+	c c, avail headlin, yr treasuri, treasuri yld, bond yr, eastern ukrain, amp unfil, announc week, al qaeda, moammar gadhafi

	Frequent phrases with the most negative loadings
icr^0	barack obama, presid barack, obama administr, substanti gain, unit total, bodi stock, al qaeda, compar januari, regist co, hedg fund
rv^0	busi bulletin, tax report, washington wire, substanti gain, labor letter, chase nation, press build, bodi stock, chicago rock, amp unfil
pd^0	labor letter, busi bulletin, tax report, washington wire, jobless marri, steel product, factori shipment, hour earn, lead indic, busi failur
np^0	bushel wheat, futur barrel, commod oil, jone aig, aig futur, barrel dow, barack obama, dj aig, al qaeda, chicago rock
icr^+	yr treasuri, construct spend, wsj research, treasuri yld, euro zone, c c, bond yr, announc week, marshal field, avail headlin
rv^+	intern aid, american cyanamid, construct spend, survey found, buyer market, vote presid, sharp contrast, harrisburg pa, talk aim, unit
pd^+	presid clinton, barrel dow, fiscal cliff, hurrican katrina, confer washington, bosnia muslim, wire clinton, orang counti, survey found, serb

- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008
- Realized variance high when front page mentions commodities, fixed income, stocks
- Commodities coverage is crowded out by news pressure

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

	Frequent phrases with the most positive loadings
icr^0	labor letter, busi bulletin, tax report, washington wire, gross net, nation recoveri, presid clinton, trend take, job trend, life job
rv^0	barrel dow, bushel wheat, yr trea, trea yld, dow jone, aig spot, futur dj, outstand stock, aig futur, stock nyse
pd^0	barack obama, yr trea, trea yld, technolog journal, insid journal, journal insid, nasdaq amp, commod oil, aig futur, weekend journal
np^0	washington wire, busi bulletin, tax report, labor letter, report steel, journal special, substanti gain, rubber tire, presid reagan, life job
icr^+	confer washington, ounc dow, miami beach, presid clinton, survey found, bosnia muslim, clinton health, presid slobodan, serb croat, ba
rv^+	west africa, amp unfil, falun gong, stock market, republican guard, composit index, john mccain, dow jone, jone industri, abu dhabi
pd^+	c c, avail headlin, yr treasuri, treasuri yld, bond yr, eastern ukrain, amp unfil, announc week, al qaeda, moammar gadhafi

	Frequent phrases with the most negative loadings
icr^0	barack obama, presid barack, obama administr, substanti gain, unit total, bodi stock, al qaeda, compar januari, regist co, hedg fund
rv^0	busi bulletin, tax report, washington wire, substanti gain, labor letter, chase nation, press build, bodi stock, chicago rock, amp unfil
pd^0	labor letter, busi bulletin, tax report, washington wire, jobless marri, steel product, factori shipment, hour earn, lead indic, busi failur
np^0	bushel wheat, futur barrel, commod oil, jone aig, aig futur, barrel dow, barack obama, dj aig, al qaeda, chicago rock
icr^+	yr treasuri, construct spend, wsj research, treasuri yld, euro zone, c c, bond yr, announc week, marshal field, avail headlin
rv^+	intern aid, american cyanamid, construct spend, survey found, buyer market, vote presid, sharp contrast, harrisburg pa, talk aim, unit
pd^+	presid clinton, barrel dow, fiscal cliff, hurrican katrina, confer washington, bosnia muslim, wire clinton, orang counti, survey found, serb

- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008
- Realized variance high when front page mentions commodities, fixed income, stocks
- Commodities coverage is crowded out by news pressure

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

	Frequent phrases with the most positive loadings
icr^0	labor letter, busi bulletin, tax report, washington wire, gross net, nation recoveri, presid clinton, trend take, job trend, life job
rv^0	barrel dow, bushel wheat, yr trea, trea yld, dow jone, aig spot, futur dj, outstand stock, aig futur, stock nyse
pd^0	barack obama, yr trea, trea yld, technolog journal, insid journal, journal insid, nasdaq amp, commod oil, aig futur, weekend journal
np^0	washington wire, busi bulletin, tax report, labor letter, report steel, journal special, substanti gain, rubber tire, presid reagan, life job
icr^+	confer washington, ounc dow, miami beach, presid clinton, survey found, bosnia muslim, clinton health, presid slobodan, serb croat, ba
rv^+	west africa, amp unfil, falun gong, stock market, republican guard, composit index, john mccain, dow jone, jone industri, abu dhabi
pd^+	c c, avail headlin, yr treasuri, treasuri yld, bond yr, eastern ukrain, amp unfil, announc week, al qaeda, moammar gadhafi

	Frequent phrases with the most negative loadings
icr^0	barack obama, presid barack, obama administr, substanti gain, unit total, bodi stock, al qaeda, compar januari, regist co, hedg fund
rv^0	busi bulletin, tax report, washington wire, substanti gain, labor letter, chase nation, press build, bodi stock, chicago rock, amp unfil
pd^0	labor letter, busi bulletin, tax report, washington wire, jobless marri, steel product, factori shipment, hour earn, lead indic, busi failur
np^0	bushel wheat, futur barrel, commod oil, jone aig, aig futur, barrel dow, barack obama, dj aig, al qaeda, chicago rock
icr^+	yr treasuri, construct spend, wsj research, treasuri yld, euro zone, c c, bond yr, announc week, marshal field, avail headlin
rv^+	intern aid, american cyanamid, construct spend, survey found, buyer market, vote presid, sharp contrast, harrisburg pa, talk aim, unit
pd^+	presid clinton, barrel dow, fiscal cliff, hurrican katrina, confer washington, bosnia muslim, wire clinton, orang counti, survey found, serb

- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008
- Realized variance high when front page mentions commodities, fixed income, stocks
- Commodities coverage is crowded out by news pressure

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

	Frequent phrases with the most positive loadings
icr^0	labor letter, busi bulletin, tax report, washington wire, gross net, nation recoveri, presid clinton, trend take, job trend, life job
rv^0	barrel dow, bushel wheat, yr trea, trea yld, dow jone, aig spot, futur dj, outstand stock, aig futur, stock nyse
pd^0	barack obama, yr trea, trea yld, technolog journal, insid journal, journal insid, nasdaq amp, commod oil, aig futur, weekend journal
np^0	washington wire, busi bulletin, tax report, labor letter, report steel, journal special, substanti gain, rubber tire, presid reagan, life job
icr^+	confer washington, ounc dow, miami beach, presid clinton, survey found, bosnia muslim, clinton health, presid slobodan, serb croat, ba
rv^+	west africa, amp unfil, falun gong, stock market, republican guard, composit index, john mccain, dow jone, jone industri, abu dhabi
pd^+	c c, avail headlin, yr treasuri, treasuri yld, bond yr, eastern ukrain, amp unfil, announc week, al qaeda, moammar gadhafi

	Frequent phrases with the most negative loadings
icr^0	barack obama, presid barack, obama administr, substanti gain, unit total, bodi stock, al qaeda, compar januari, regist co, hedg fund
rv^0	busi bulletin, tax report, washington wire, substanti gain, labor letter, chase nation, press build, bodi stock, chicago rock, amp unfil
pd^0	labor letter, busi bulletin, tax report, washington wire, jobless marri, steel product, factori shipment, hour earn, lead indic, busi failur
np^0	bushel wheat, futur barrel, commod oil, jone aig, aig futur, barrel dow, barack obama, dj aig, al qaeda, chicago rock
icr^+	yr treasuri, construct spend, wsj research, treasuri yld, euro zone, c c, bond yr, announc week, marshal field, avail headlin
rv^+	intern aid, american cyanamid, construct spend, survey found, buyer market, vote presid, sharp contrast, harrisburg pa, talk aim, unit
pd^+	presid clinton, barrel dow, fiscal cliff, hurrican katrina, confer washington, bosnia muslim, wire clinton, orang counti, survey found, serb

- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008
- Realized variance high when front page mentions commodities, fixed income, stocks
- Commodities coverage is crowded out by news pressure

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

	Frequent phrases with the most positive loadings
icr^0	labor letter, busi bulletin, tax report, washington wire, gross net, nation recoveri, presid clinton, trend take, job trend, life job
rv^0	barrel dow, bushel wheat, yr trea, trea yld, dow jone, aig spot, futur dj, outstand stock, aig futur, stock nyse
pd^0	barack obama, yr trea, trea yld, technolog journal, insid journal, journal insid, nasdaq amp, commod oil, aig futur, weekend journal
np^0	washington wire, busi bulletin, tax report, labor letter, report steel, journal special, substanti gain, rubber tire, presid reagan, life job
icr^+	confer washington, ounc dow, miami beach, presid clinton, survey found, bosnia muslim, clinton health, presid slobodan, serb croat, ba
rv^+	west africa, amp unfil, falun gong, stock market, republican guard, composit index, john mccain, dow jone, jone industri, abu dhabi
pd^+	c c, avail headlin, yr treasuri, treasuri yld, bond yr, eastern ukrain, amp unfil, announc week, al qaeda, moammar gadhafi

	Frequent phrases with the most negative loadings
icr^0	barack obama, presid barack, obama administr, substanti gain, unit total, bodi stock, al qaeda, compar januari, regist co, hedg fund
rv^0	busi bulletin, tax report, washington wire, substanti gain, labor letter, chase nation, press build, bodi stock, chicago rock, amp unfil
pd^0	labor letter, busi bulletin, tax report, washington wire, jobless marri, steel product, factori shipment, hour earn, lead indic, busi failur
np^0	bushel wheat, futur barrel, commod oil, jone aig, aig futur, barrel dow, barack obama, dj aig, al qaeda, chicago rock
icr^+	yr treasuri, construct spend, wsj research, treasuri yld, euro zone, c c, bond yr, announc week, marshal field, avail headlin
rv^+	intern aid, american cyanamid, construct spend, survey found, buyer market, vote presid, sharp contrast, harrisburg pa, talk aim, unit
pd^+	presid clinton, barrel dow, fiscal cliff, hurrican katrina, confer washington, bosnia muslim, wire clinton, orang counti, survey found, serb

- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008
- Realized variance high when front page mentions commodities, fixed income, stocks
- Commodities coverage is crowded out by news pressure

Focus on a single phrase for intuition

“financial crisis” is crowded out at high $NewsPressure$ times

Backward regressions			
	HDMR		DMR
	Inclusion	Repetition	
intercept	–16.02	–8.06	–13.70
icr	–60.08	–33.41	–58.87
rv	0.48	0.15	0.26
pd	3.89	1.20	3.01
NewsPressure	–0.02		0.00

⇒

Forward regressions		
	HDMR	DMR
Repetition	–0.03	–0.05
Inclusion	–0.04	

Focus on a single phrase for intuition

“financial crisis” is crowded out at high $NewsPressure$ times

Backward regressions

	HDMR		DMR
	Inclusion	Repetition	
intercept	–16.02	–8.06	–13.70
icr	–60.08	–33.41	–58.87
rv	0.48	0.15	0.26
pd	3.89	1.20	3.01
NewsPressure	–0.02		0.00

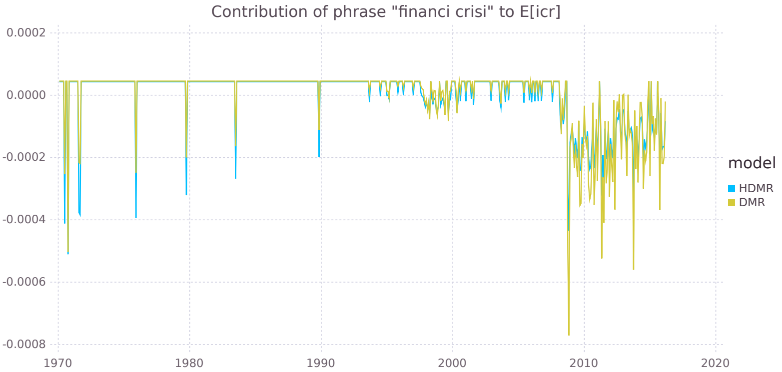
$\Rightarrow$

Forward regressions

	HDMR	DMR
Repetition	–0.03	–0.05
Inclusion	–0.04	

Focus on a single phrase for intuition

"financial crisis" on the front page is bad news for dealers, regardless of repetition



Does newspaper coverage forecast macroeconomic series?

- Stock-Watson (2012) show that macro forecasts of a simple dynamic factor model (DFM-5) are hard to beat
- We use their data + WSJ text to forecast 1–4 months ahead
- Findings:
  - Substantial OOS RMSE improvement using text with HDMR relative to DFM-5 for macroeconomic fundamentals
    - Nonfarm payroll employment forecast is 23–44% better
    - Housing starts forecast is 45–52% better
  - WSJ text is not helping predict asset prices directly (stocks, treasuries, currencies)
  - Advantage of HDMR increases with sparsity of the text
  - Stronger results for nowcasting

Conclusion

- Text provides a relatively untapped source of data
- Incorporating structural economic restrictions into machine learning methods can improve out-of-sample prediction
- Hurdle Distributed Multiple Regression (HDMR)
  - Highly scalable approach to inference from big counts data
  - Includes an economically-motivated selection equation
  - Useful where extensive margin is interesting or more important than intensive margin
  - Publicly available as a Julia package: `HurdleDMR`

References

Baker, Scott R, Nicholas Bloom, and Steven J Davis, 2016, Measuring economic policy uncertainty, *Quarterly Journal of Economics* 131, 1593–1636.

Eisensee, Thomas, and David Stromberg, 2007, News droughts, news floods, and u.s. disaster relief, *Quarterly Journal of Economics* 122, 693–728.

Foster, Dean P, Mark Liberman, and Robert A Stine, 2013, Featurizing text: Converting text into predictors for regression analysis, Working paper.

Frankel, Richard, Jared Jennings, and Joshua Lee, 2016, Using unstructured and qualitative disclosures to explain accruals, *Journal of Accounting and Economics* 62, 209–227.

Gabaix, Xavier, 2014, A sparsity-based model of bounded rationality, *Quarterly Journal of Economics* 129, 1661–1710.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy, 2017a, Text as data, Working Paper 23276, National Bureau of Economic Research.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy, 2017b, Measuring polarization in high-dimensional data: Method and application to congressional speech, Technical report, National Bureau of Economic Research.

Jegadeesh, Narasimhan, and Di Andrew Wu, 2017, Deciphering fedspeak: The information content of FOMC meetings, Working paper.

Kelly, Bryan, Asaf Manela, and Alan Moreira, 2018, Text selection, Working paper.

Manela, Asaf, 2014, The value of diffusing information, *Journal of Financial Economics* 111, 181–199.

Manela, Asaf, and Alan Moreira, 2017, News implied volatility and disaster concerns, *Journal of Financial Economics* 123, 137–162.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al., 2011, Quantitative analysis of culture using millions of digitized books, *Science* 331, 176–182.

Mullahy, John, 1986, Specification and testing of some modified count data models, *Journal of econometrics* 33, 341–365.

Taddy, Matt, 2013, Multinomial inverse regression for text analysis, *Journal of the American Statistical Association* 108, 755–770.

Tetlock, Paul C., 2007, Giving content to investor sentiment: The role of media in the stock market, *Journal of Finance* 62, 1139–1168.

Vapnik, N. Vladimir, 2000, *The Nature of Statistical Learning Theory* (Springer-Verlag, New York.).