Text Selection

Bryan Kelly
Yale University

Asaf Manela
Washington University in St. Louis

Alan Moreira
University of Rochester

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

This paper

- 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)

Related literature

- We extend machinery developed by Taddy (2012, 2015, 2016) to text selection
  - Layer economically-motivated hurdle / selection equation on his Distributed Multinomial Regression (DMR)
  - Find advantage of HDMR over DMR increases with sparsity
- Provide new tools to literatures in economics and finance
  - Finance and media: Antweiler-Frank (2004), Tetlock (2007, 2011), Fang-Peress (2009), Engelberg-Parsons (2011), Dougal et al (2012), Peress (2014), Manela (2014), Fedyk (2018)
  - Text-based uncertainty: Baker-Bloom-Davis (2016), Manela-Moreira (2017), Hassan et al (2017)
  - Polarization: Gentzkow-Shapiro (2006), Gentzkow-Shapiro-Taddy
- Can better control and learn from high-dimensional content

Text data is inherently high-dimensional

Documents

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

⇒

Document-term matrix $\boldsymbol{c}$

	digital text	text is	is available	is selected	...
1:	1	1	1	0	
2:	0	1	0	1	
		⋮			⋱

Text data is inherently high-dimensional

Documents

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

$\Rightarrow$

Document-term matrix $\boldsymbol{c}$

	digital text	text is	is available	is selected	...
1:	1	1	1	0	
2:	0	1	0	1	
		⋮			⋱

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

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

Text inverse regression

- 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

- 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

- 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

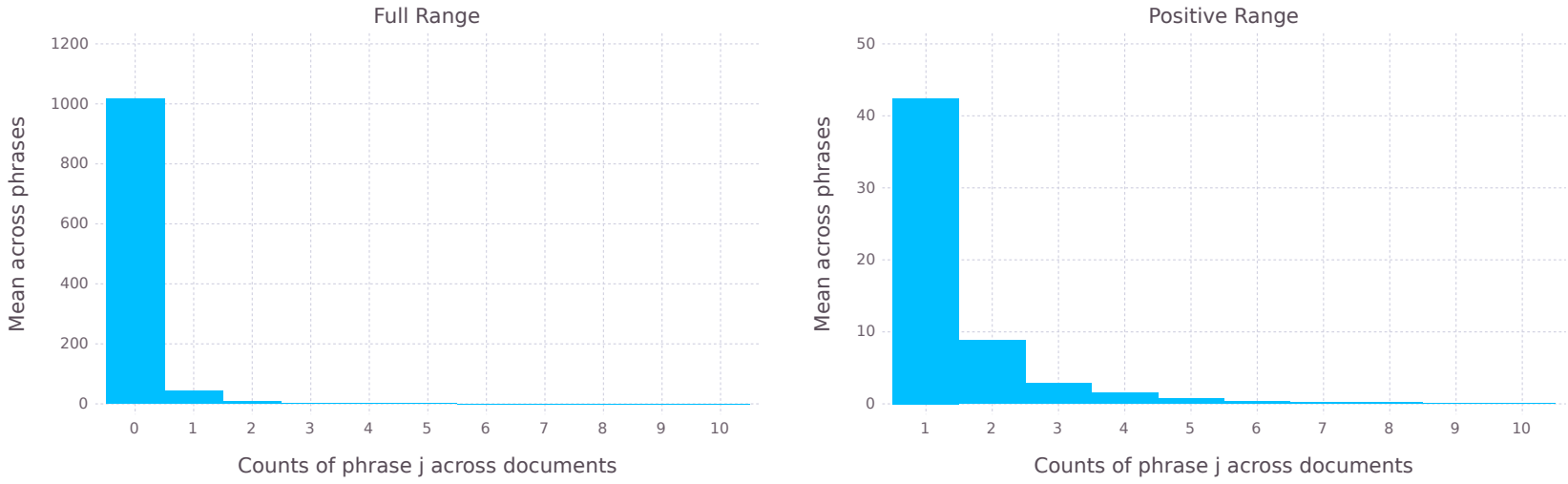
$$z_{iy} \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(sufficient reduction projection)}$$

  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

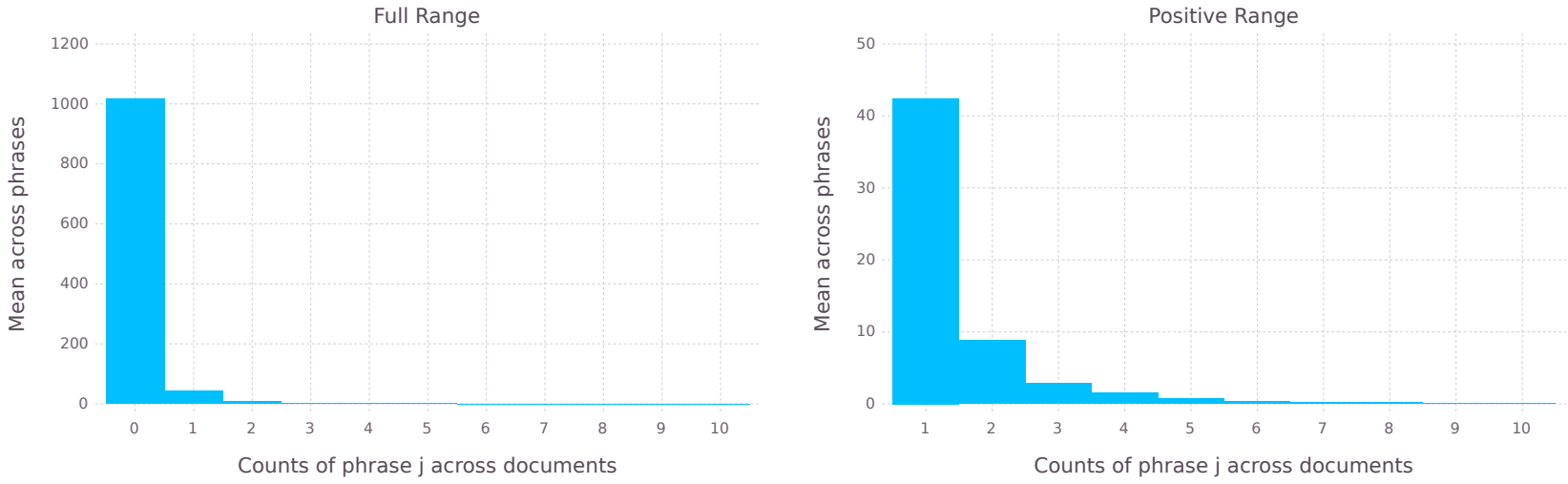
Why would we need a hurdle?



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-Taddy, 2017)
  - Censored or socially taboo words (Michel et al, 2011)
  - Fixed cost of introducing new terms, low marginal cost

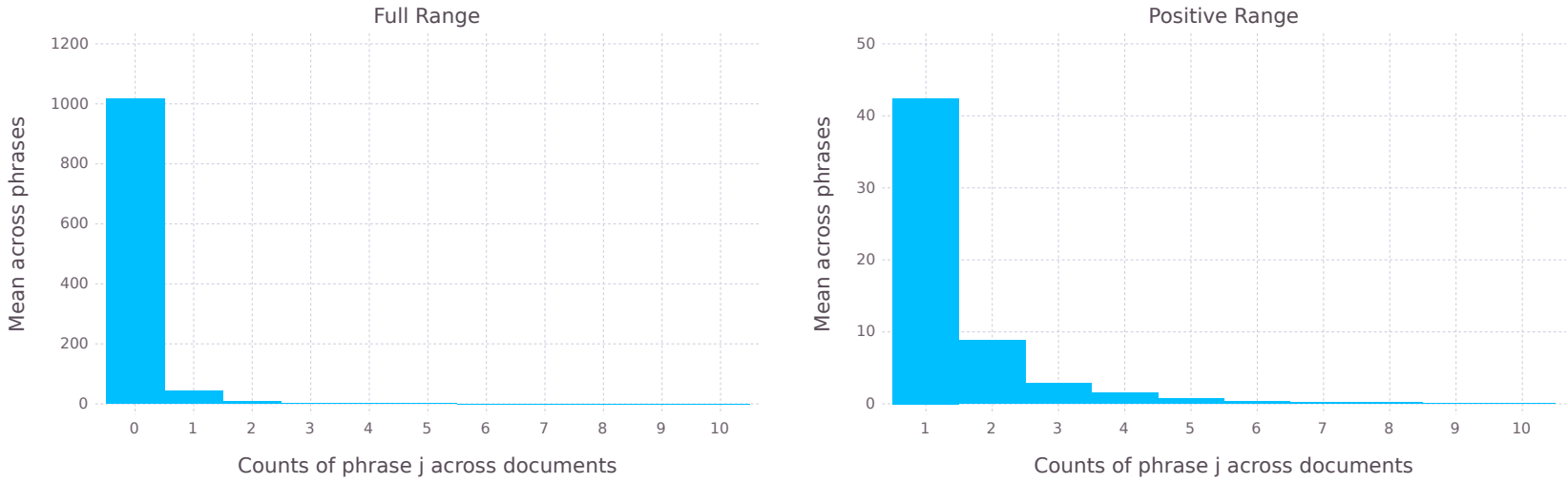
Why would we need a hurdle?



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-Taddy, 2017)
  - 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-Taddy, 2017)
  - 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

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

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

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

  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

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

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

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

  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

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

- $d + p - 1$ 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

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

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

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

  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!

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

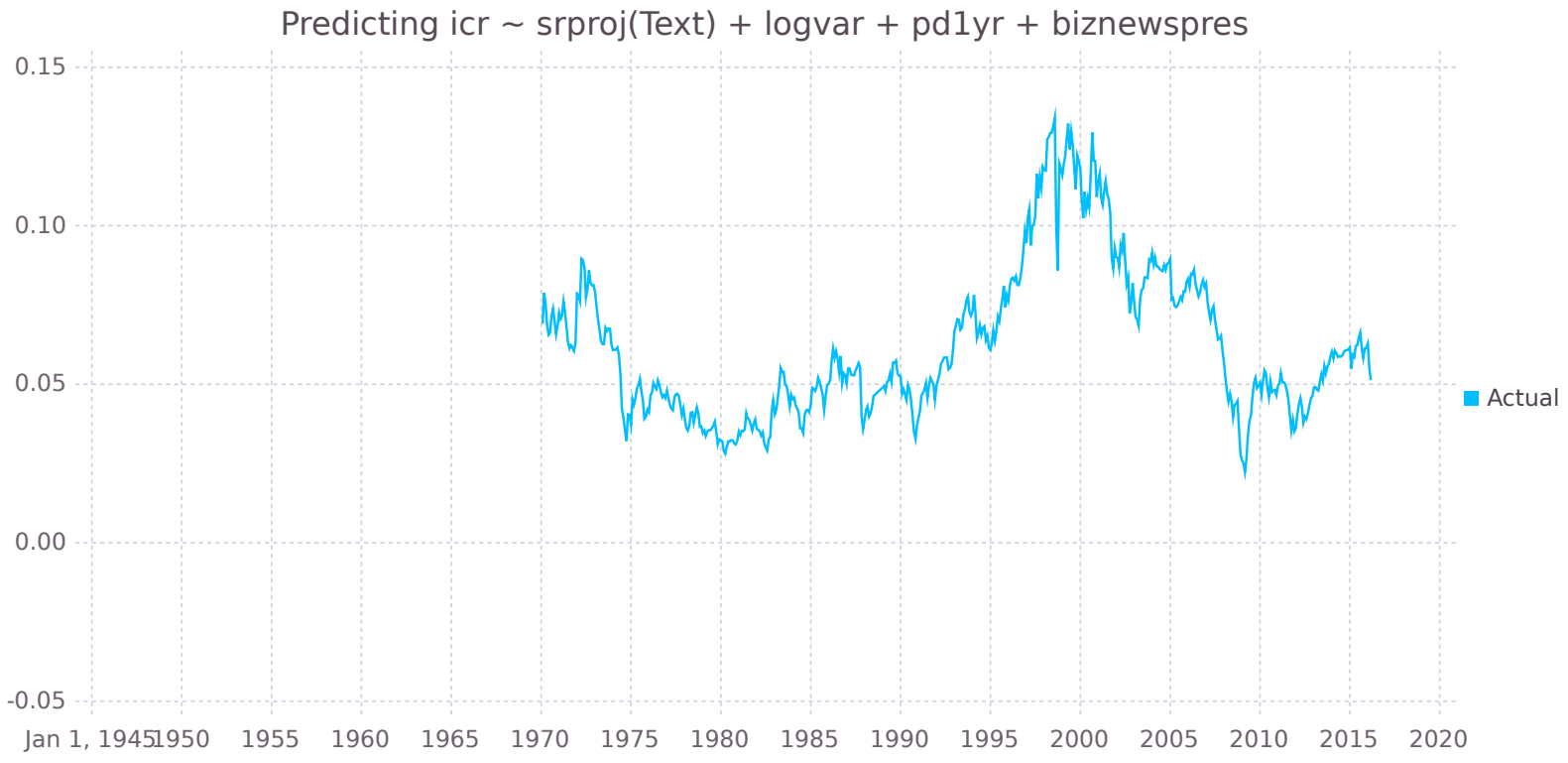
Date	Title	Abstract
2008-09-16	AIG Faces Cash Crisis As Stock Dives 61%	American International Group Inc. was facing a severe cash ...
2008-09-16	AIG, Lehman Shock Hits World Markets ...	The convulsions in the U.S. financial system sent markets ...
2008-09-16	Business and Finance	Central banks around 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-Stromberg, 2007 QJE), starts 1967
  - Business news pressure (Manela, 2014 JFE), starts 1945
  - Assumption: excluded from repetition equation

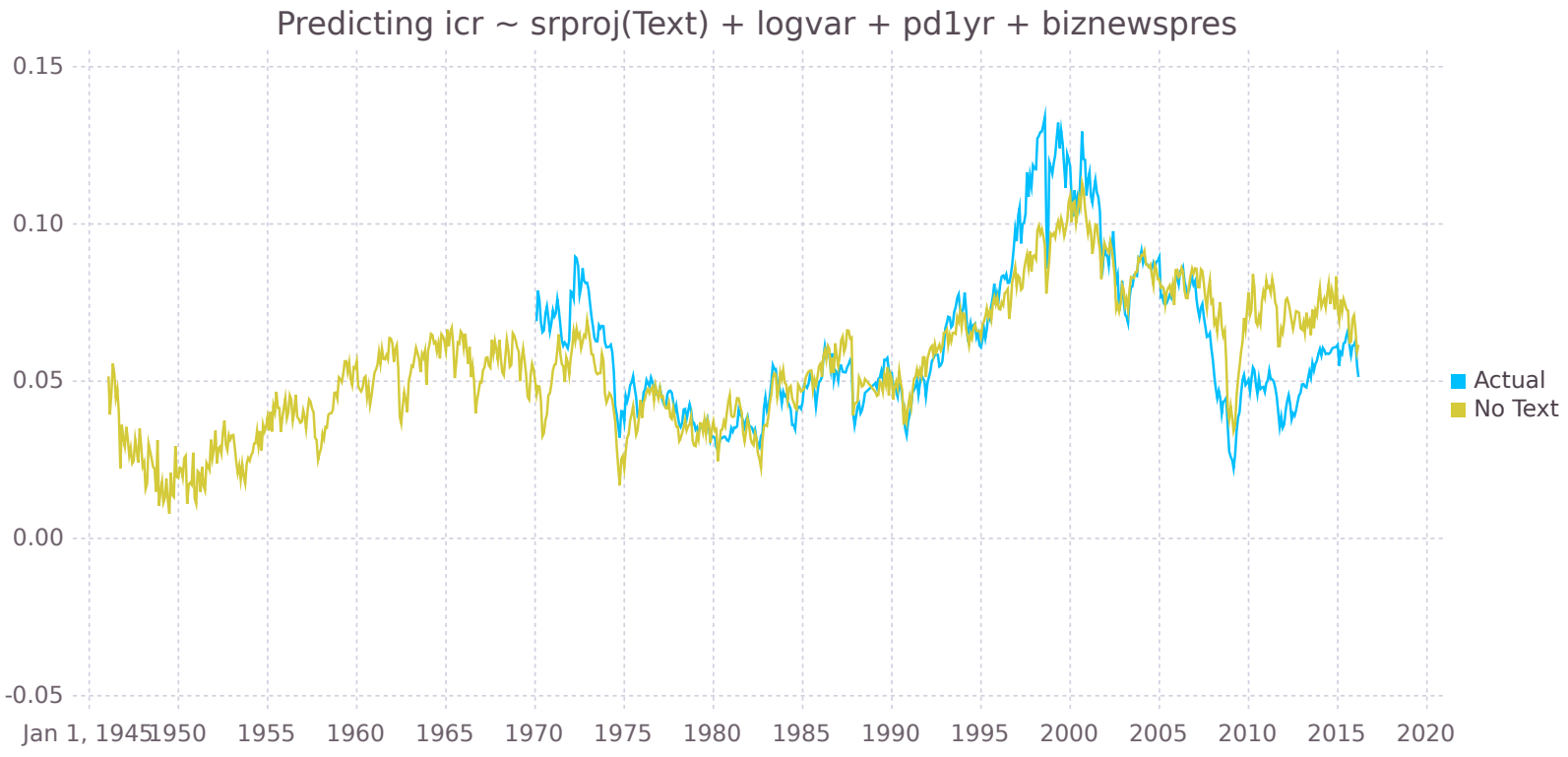
News implied intermediary capital ratio

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



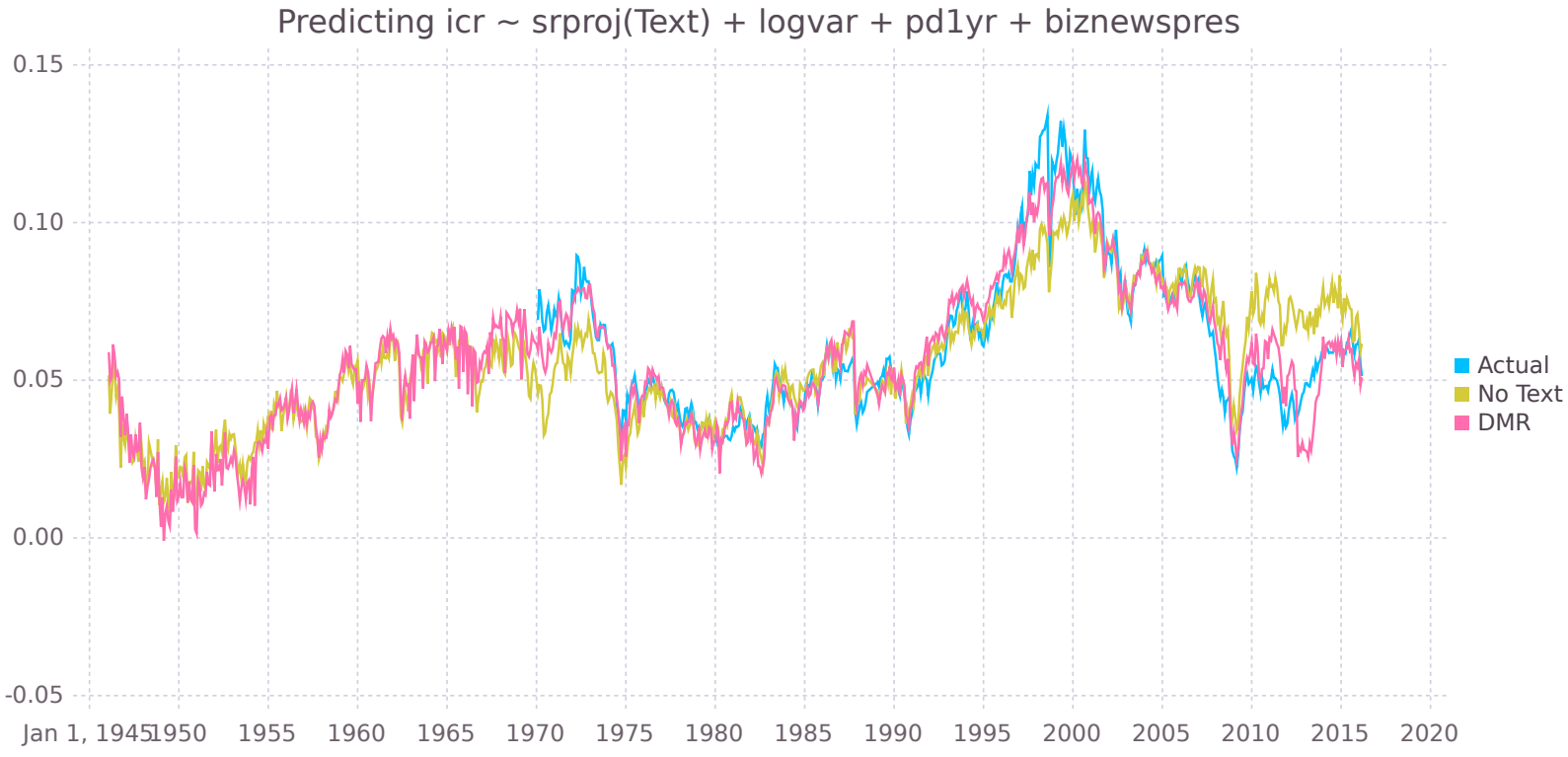
News implied intermediary capital ratio

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



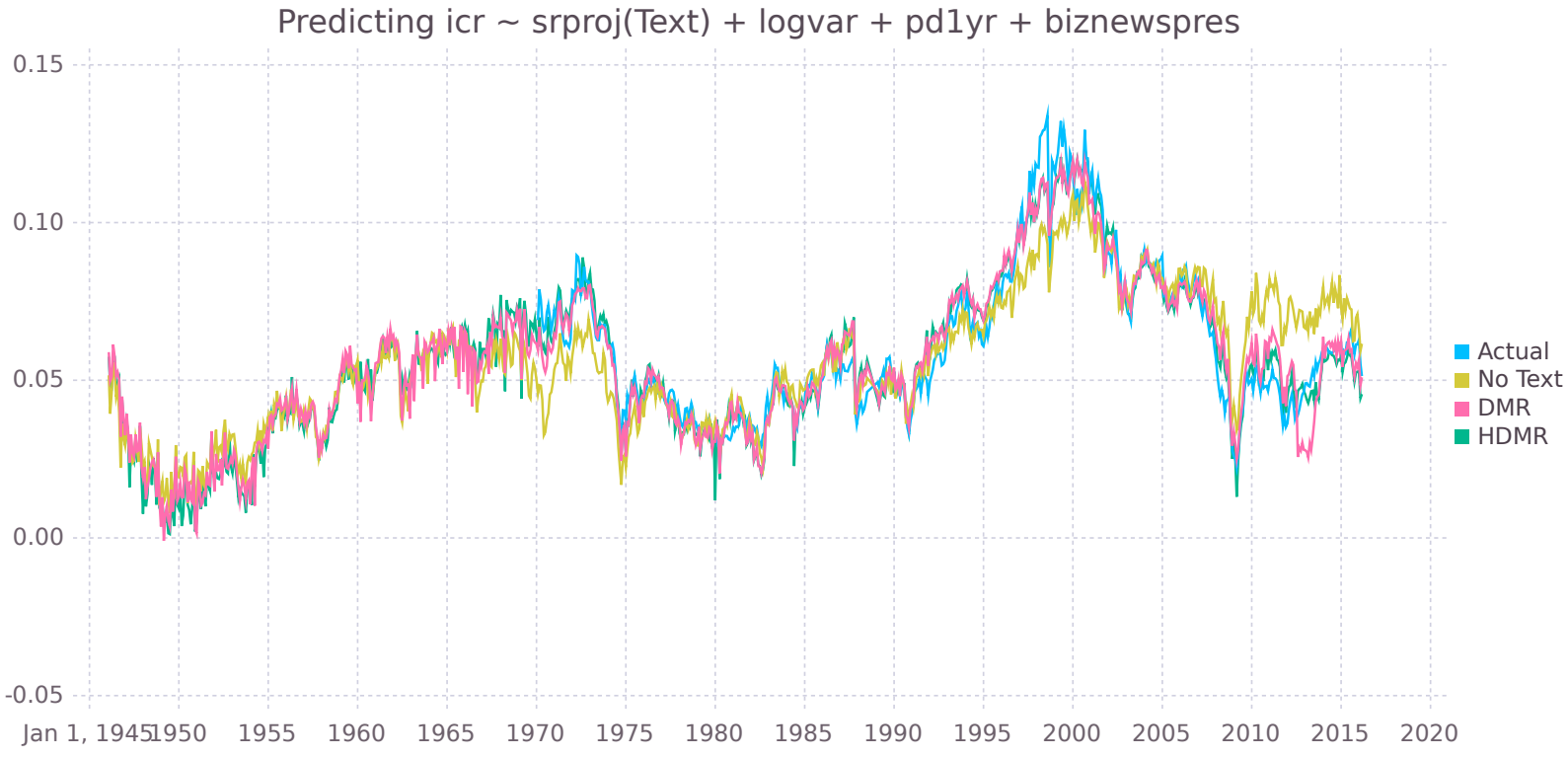
News implied intermediary capital ratio

DMR gives a different predicted series exploiting the text



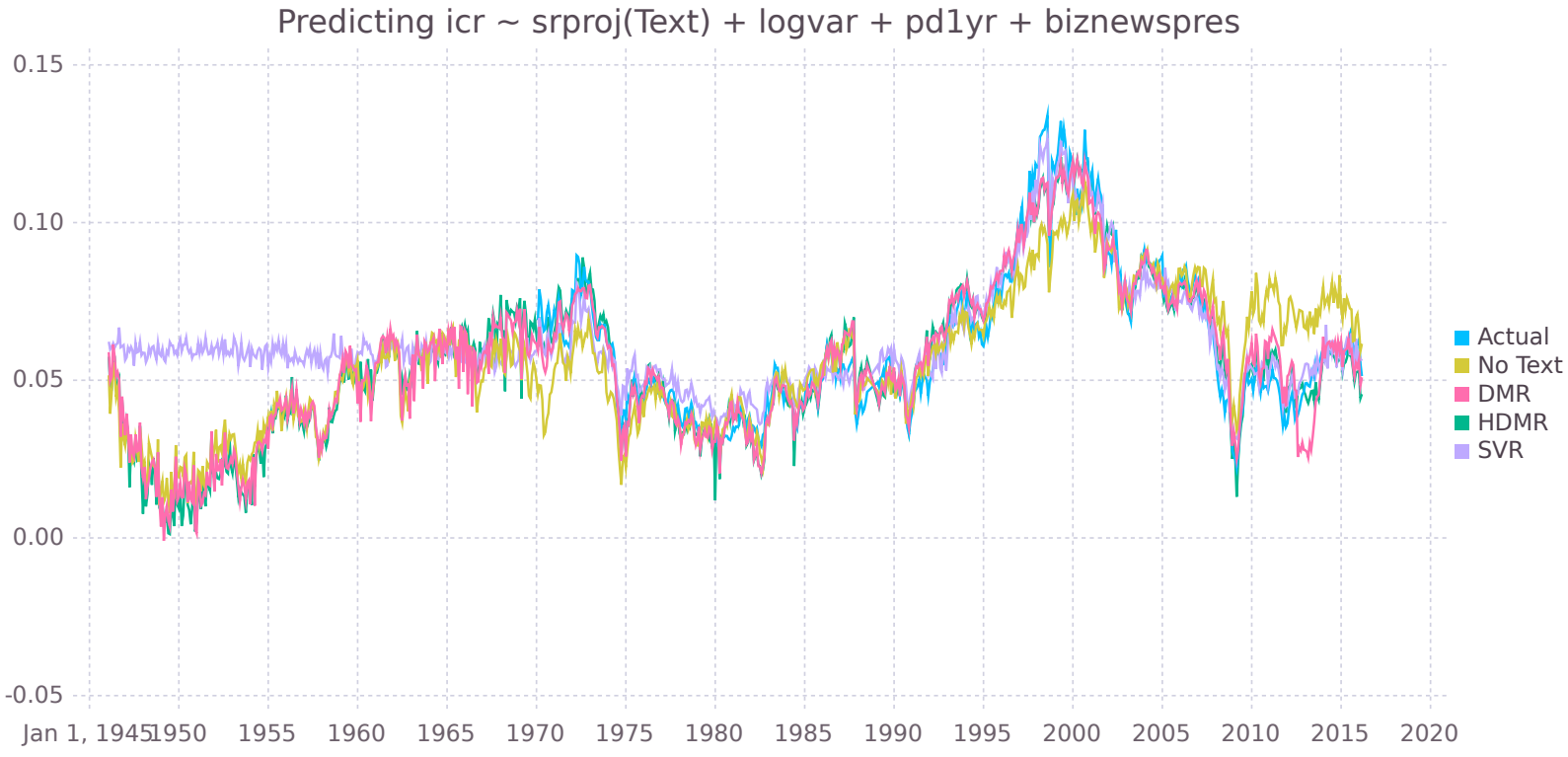
News implied intermediary capital ratio

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



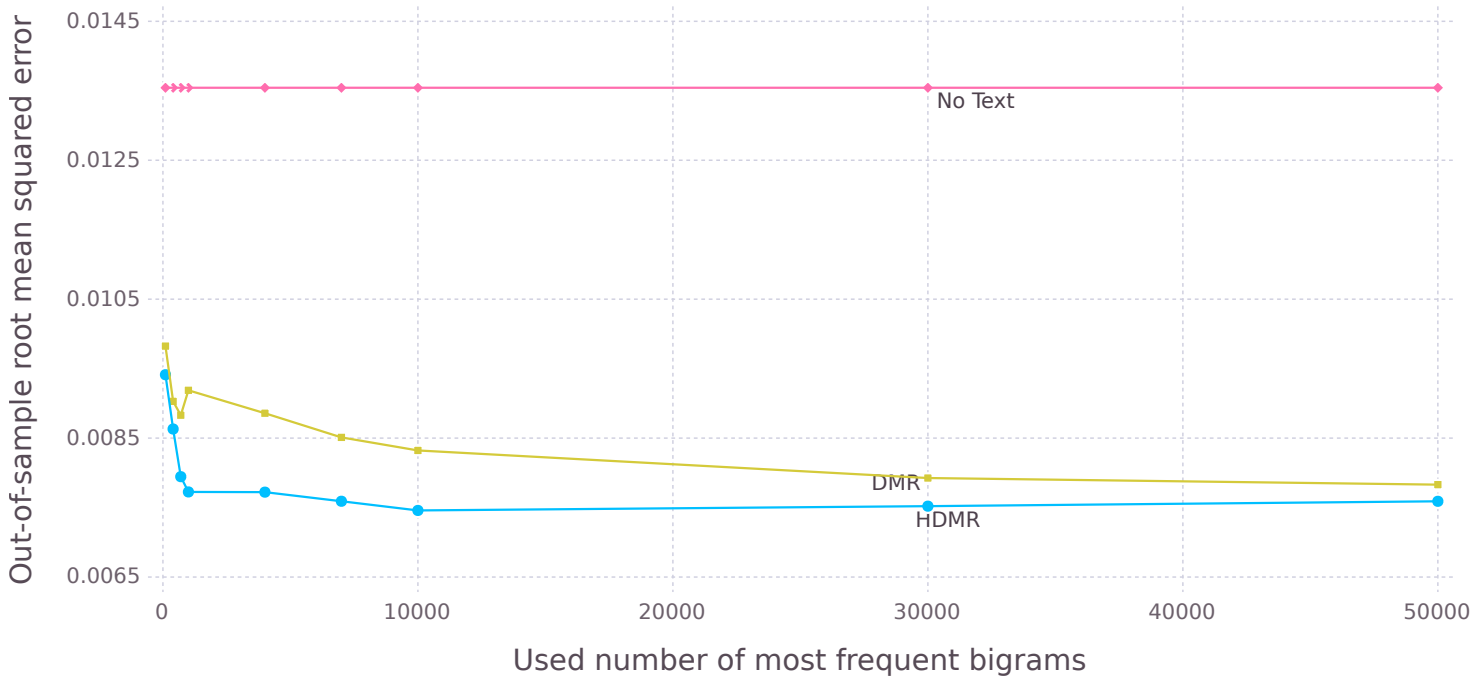
News implied intermediary capital ratio

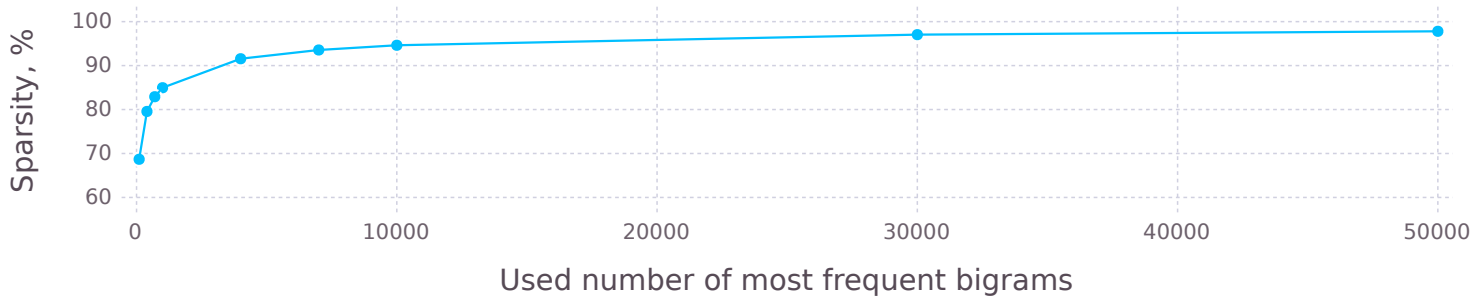
Support Vector Regression of Manela-Moreira (2017) cannot concentrate on covariates



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

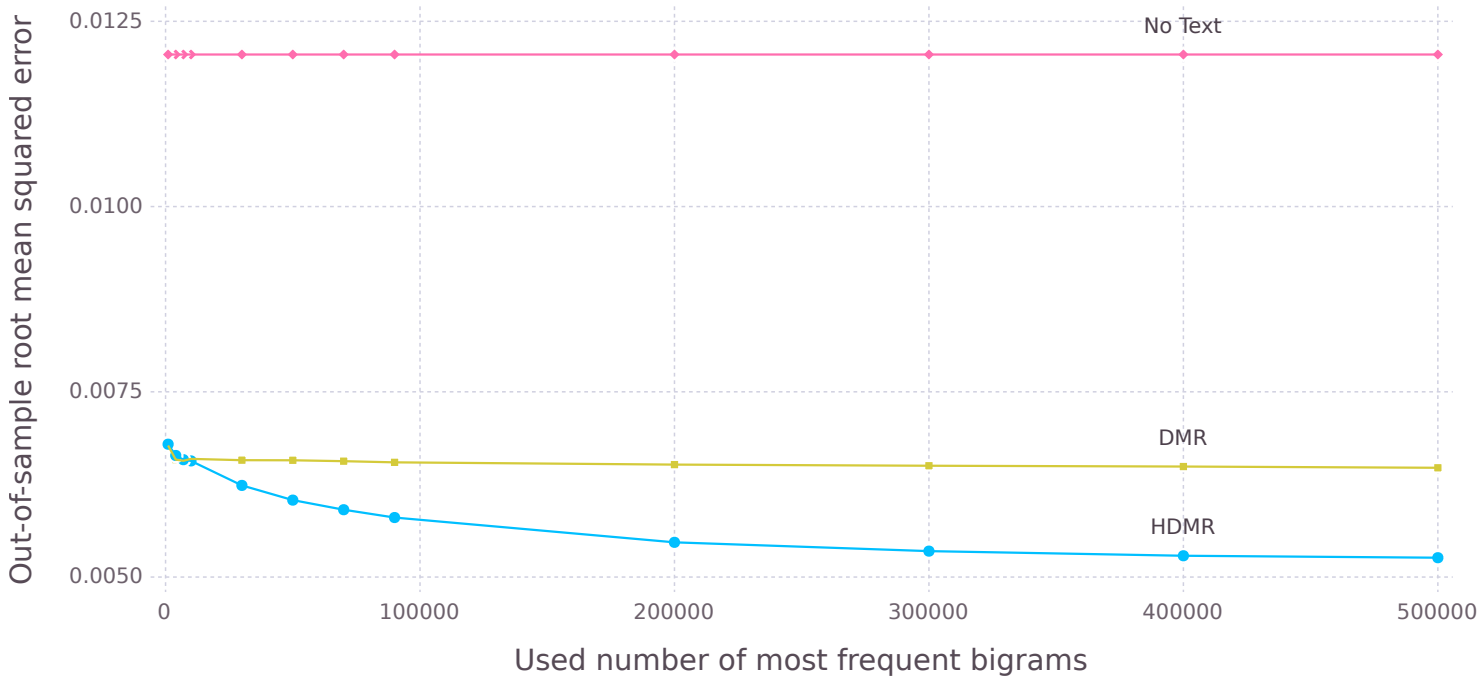
HDMR's out-of-sample fit advantage changes with text sparsity

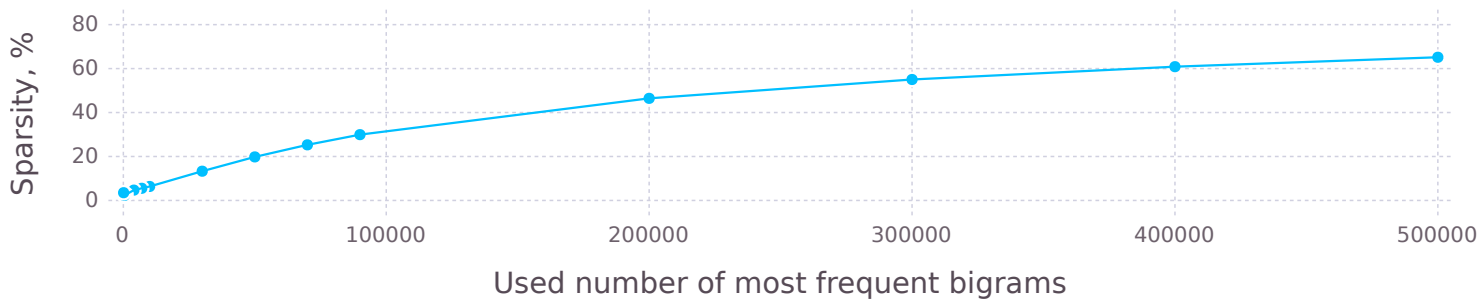




Denser text: HDMR's advantage increases with sparsity

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, serl

- 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, serl

- 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, serl

- 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, serl

- 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, serl

- 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

Does newspaper coverage forecast macroeconomic series?

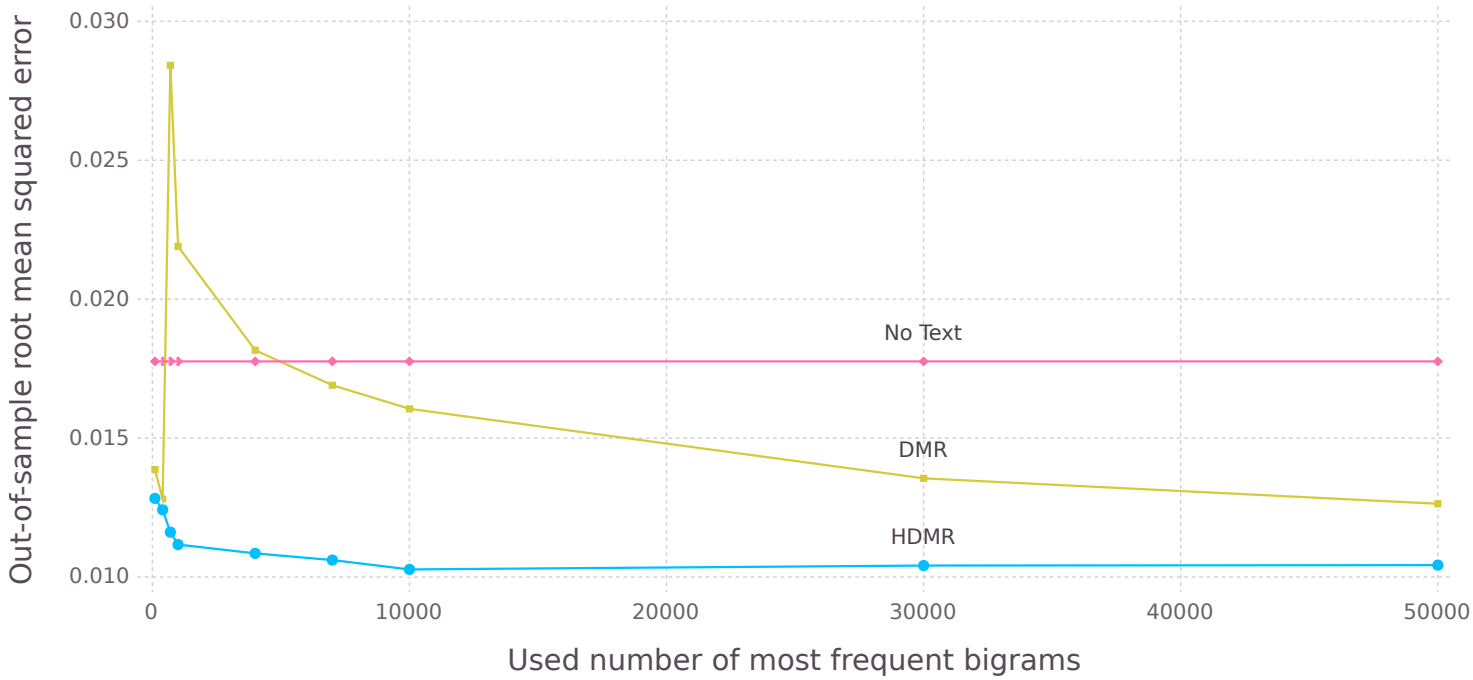
- Stock-Watson (2012) show that macro forecasts of a simple dynamic factor model (DFM-5) are hard to beat
- We use their data + WSJ text to forecast 1–4 months ahead
- Findings:
  - Substantial OOS RMSE improvement using text with HDMR relative to DFM-5 for macroeconomic fundamentals
    - Nonfarm payroll 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

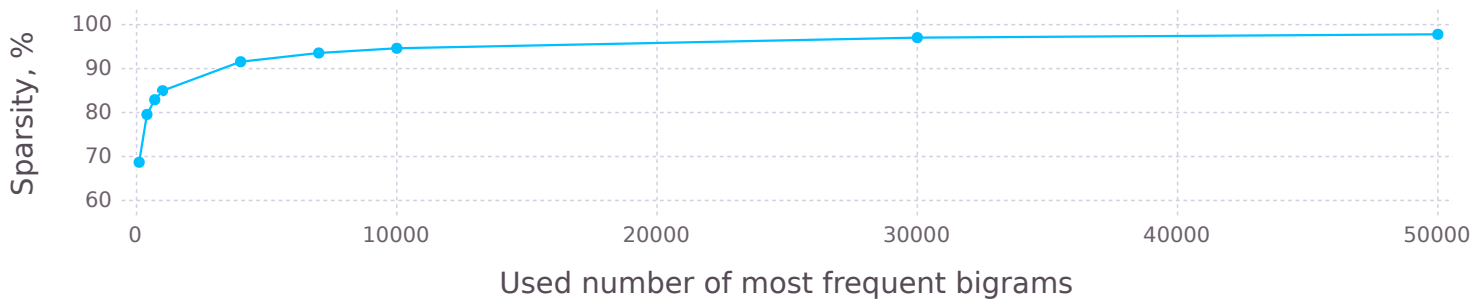
Conclusion

- 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
- Applications using newspaper coverage for prediction
  1. Backcast intermediary capital ratio
  2. Forecast macroeconomic series

Predicting ICR with text and covariates

HDMR's advantage is larger with serial cross-validation





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" 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

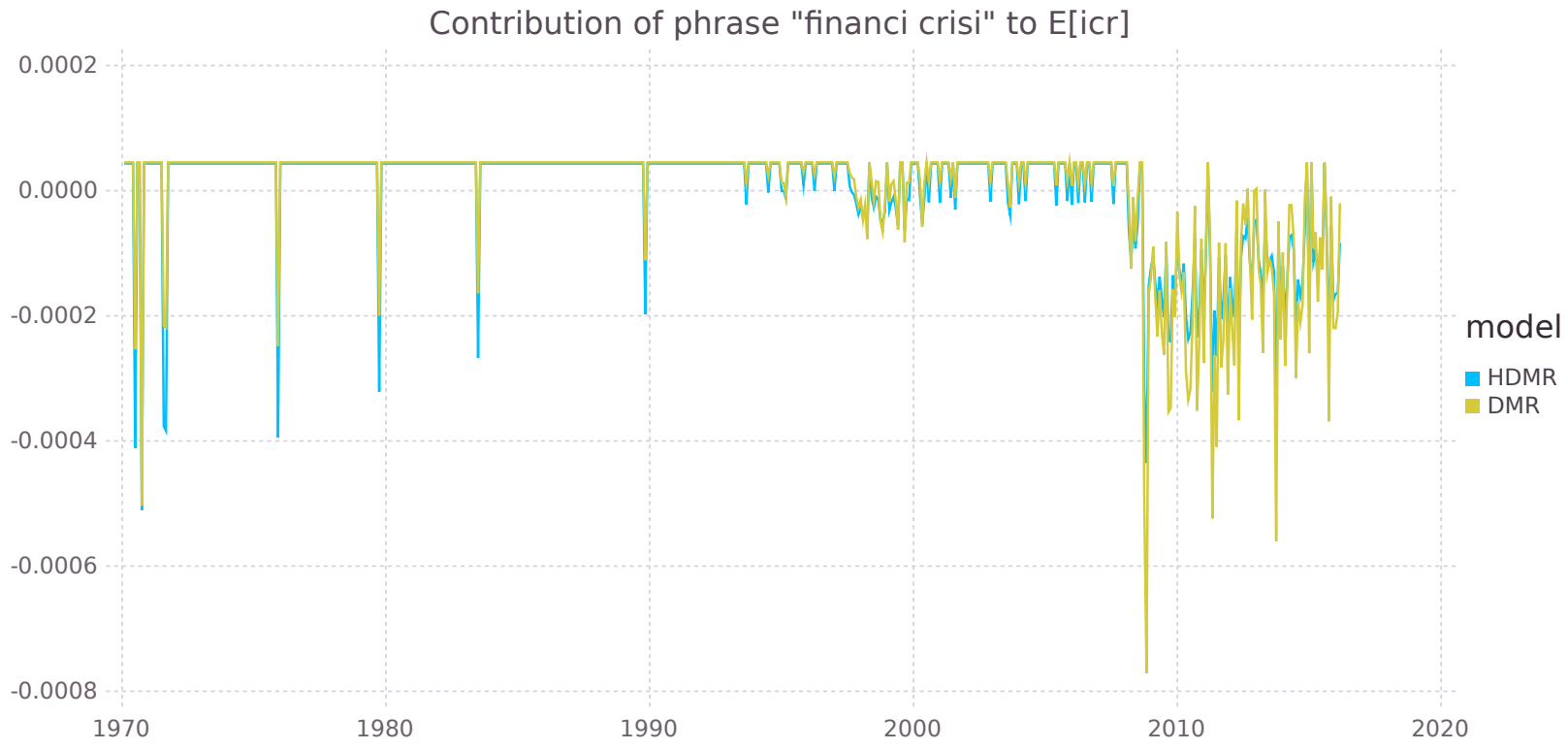
$\Rightarrow$

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



Macro forecasting

WSJ full text, 10,000 bigrams

Months forward:	$h = 1$				$h = 2$				$h = 4$			
Folds:	Random		Serial		Random		Serial		Random		Serial	
Y_{t+h}^h	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR
IP: total	0.984	0.960	1.019	0.975	0.942	0.911	1.038	0.963	0.874	0.876	0.997	0.962
Emp: total	0.881	0.858	0.988	0.953	0.830	0.836	0.983	0.953	0.810	0.827	0.964	0.963
U: all	0.956	0.948	1.109	1.065	0.857	0.869	1.066	1.094	0.812	0.826	1.126	1.076
HStarts: Total	0.716	0.653	0.898	0.877	0.684	0.637	0.885	0.864	0.668	0.624	0.844	0.853
PMI	0.838	0.815	1.144	1.146	0.830	0.826	1.137	1.152	0.832	0.836	1.135	1.151
CPI-ALL	1.110	1.096	1.062	1.025	1.080	1.107	1.025	1.039	1.113	1.093	1.072	1.048
Real AHE: goods	0.985	0.996	0.974	0.987	0.933	0.948	0.979	0.960	0.842	0.852	0.940	0.943
FedFunds	0.952	0.918	1.074	1.064	0.877	0.854	1.101	1.097	0.827	0.812	1.096	1.135
M1	1.100	1.082	1.121	1.069	1.132	1.113	1.102	1.053	1.103	1.111	1.118	1.126
Ex rate: avg	1.065	1.059	1.113	1.139	1.030	1.009	1.148	1.135	0.966	0.941	1.126	1.117
S&P 500	1.041	1.021	1.032	1.025	0.984	0.970	1.003	1.027	0.927	0.924	1.006	1.021
Consumer expect	1.110	1.159	1.042	1.096	1.052	1.041	1.087	1.046	1.041	1.007	1.070	1.040

Macro forecasting

WSJ full text, 100,000 bigrams

Months forward:	$h=1$				$h=2$				$h=4$			
Folds:	Random		Serial		Random		Serial		Random		Serial	
Y_{t+h}^h	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR
IP: total	0.976	0.961	1.012	0.990	0.895	0.902	0.980	0.983	0.833	0.864	0.963	0.984
Emp: total	0.769	0.832	0.912	0.957	0.722	0.808	0.890	0.956	0.665	0.800	0.873	0.966
U: all	0.986	0.945	1.102	1.068	0.883	0.858	1.128	1.100	0.772	0.802	1.097	1.089
HStarts: Total	0.546	0.647	0.818	0.878	0.499	0.627	0.783	0.866	0.480	0.619	0.768	0.853
PMI	0.829	0.774	1.136	1.171	0.852	0.782	1.157	1.173	0.840	0.789	1.140	1.162
CPI-ALL	1.012	1.152	1.031	1.077	1.030	1.156	1.032	1.093	1.089	1.147	1.049	1.080
Real AHE: goods	1.045	1.012	1.036	0.989	1.034	0.946	1.064	0.952	0.921	0.841	0.973	0.928
FedFunds	0.996	0.906	1.153	1.055	0.930	0.825	1.151	1.081	0.906	0.772	1.078	1.114
M1	1.040	1.169	1.038	1.167	1.032	1.220	1.012	1.103	1.052	1.156	1.045	1.124
Ex rate: avg	1.073	1.080	1.104	1.166	1.059	1.012	1.138	1.157	0.968	0.930	1.102	1.112
S&P 500	1.054	1.031	1.049	1.049	0.963	0.966	1.014	1.031	0.843	0.915	0.985	1.021
Consumer expect	1.061	1.206	1.013	1.122	1.079	1.063	1.096	1.093	1.102	1.014	1.118	1.074