Text Selection

Bryan Kelly
Yale University

Asaf Manela
Washington University in St. Louis

Alan Moreira
University of Rochester

October 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!
⋮

$\Rightarrow$

Document-term matrix 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

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

- $d + p - 1$ 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

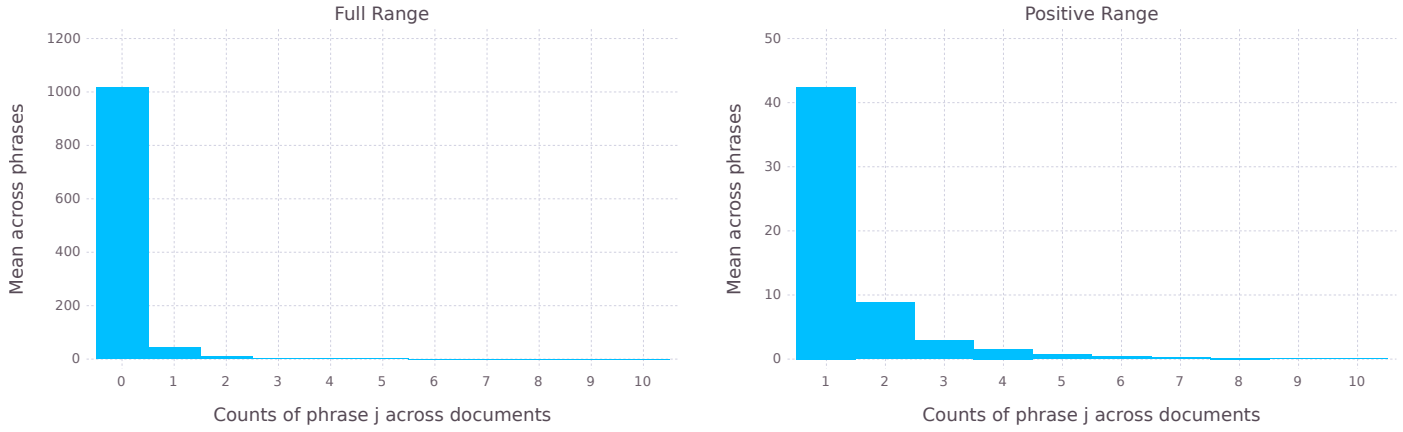
$$z_{iy} \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(sufficient reduction projection)}$$

  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

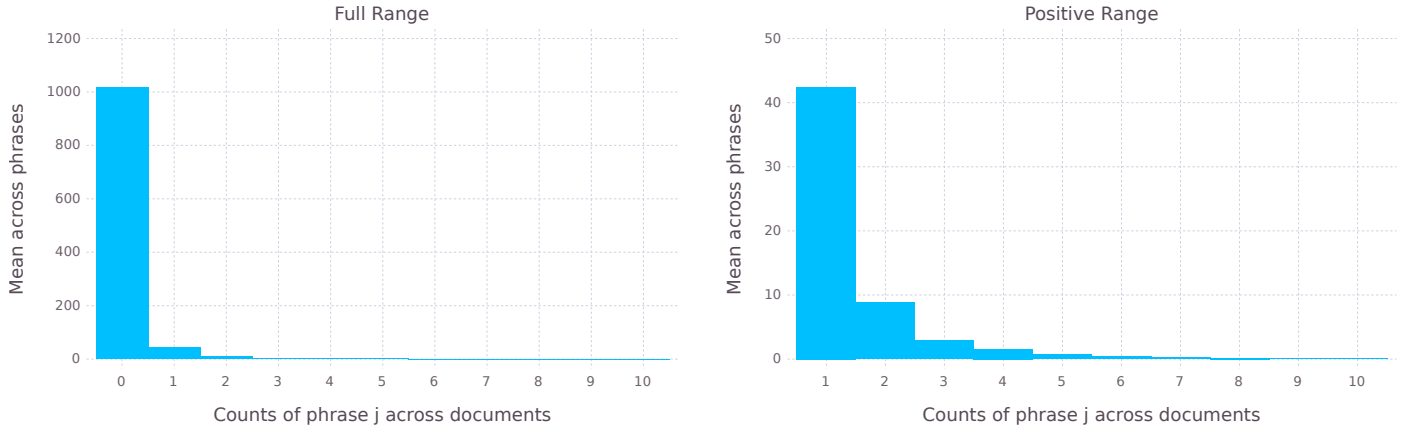
Why would we need a hurdle?



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

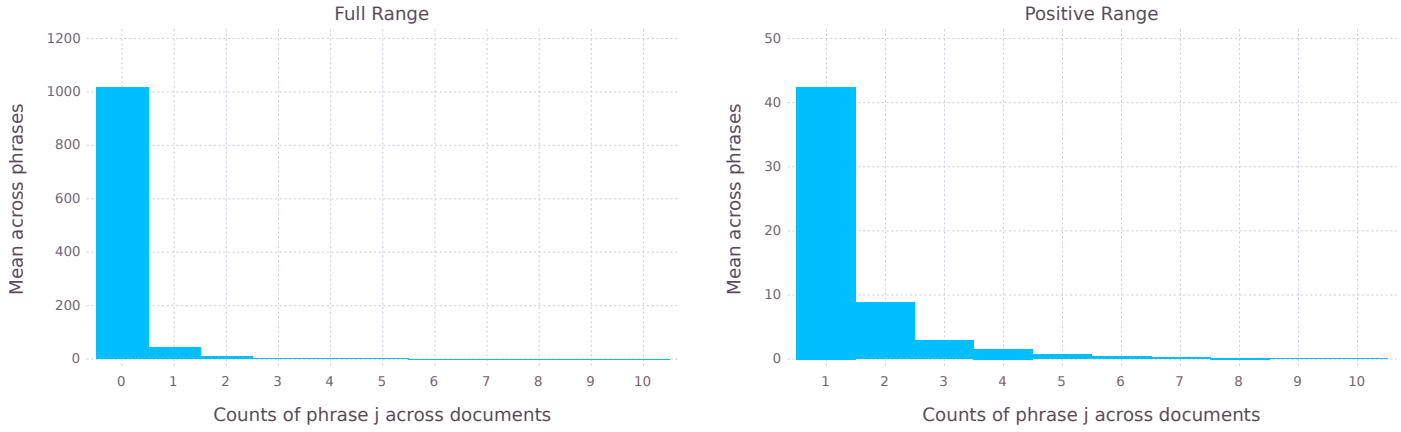
Why would we need a hurdle?



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

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

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

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

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

$$z_{iy}^0 \equiv \sum_j \hat{\delta}_{jy} h_{ij} \qquad z_{iy}^+ \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(SR projections)}$$

Inclusion 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

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

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

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

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

$$z_{iy}^0 \equiv \sum_j \hat{\delta}_{jy} h_{ij} \qquad z_{iy}^+ \equiv \sum_j \hat{\varphi}_{jy} c_{ij} \qquad \text{(SR projections)}$$

Inclusion 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

$$\boldsymbol{h}_i^* = f\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right) + \boldsymbol{\omega}_i \qquad \text{(Inclusion)}$$
$$\boldsymbol{c}_i^* = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \boldsymbol{v}_i \qquad \text{(Repetition)}$$
$$\boldsymbol{c}_i = \boldsymbol{c}_i^* \times \mathbf{1}\left(\boldsymbol{h}_i^* > 0\right) = \boldsymbol{c}_i^* \times \boldsymbol{h}_i \qquad \text{(Observation)}$$

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

$$\underbrace{z_{iy}^0 \equiv \textstyle\sum_j \hat{\delta}_{jy} h_{ij}}_{\text{Inclusion}} \qquad \underbrace{z_{iy}^+ \equiv \textstyle\sum_j \hat{\varphi}_{jy} c_{ij}}_{\text{Repetition}} \qquad \text{(SR projections)}$$

  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*, 1926-2016

Date	Title	Abstract
20080916	AIG Faces Cash Crisis As Stock Dives 61%	American International Group Inc. was
20080916	AIG, Lehman Shock Hits World Markets ...	The convulsions in the U.S. financial s
20080916	Business and Finance	Central banks around the world pumpe
20080916	Keeping Their Powder Dry: Draft Boards ...	The Selective Service System has the a
20080916	Old-School Banks Emerge Atop New ...	Banks are heading "back to basics – to
20080916	World-Wide	Thailand's ruling party chose ousted le
	⋮	

HDMR approach to news implied intermediary capital ratio

- We use HDMR to backcast missing values of ICR with WSJ text
  - + log price dividend ratio (pd_t)
  - + realized variance of financial stocks ($rvfin_t$,$rvfin_{t-1}$)
- 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 an instrument for the inclusion decision
  - Prior attention to an issue may influence its coverage by the press (Boydstun, 2013)
  - We use prior year realized variance of financial stocks ($rvfin_{t-13 \to t-1}$)
  - Assumption: excluded from repetition equation

Predicting ICR with realized variance, pd, and text

- Estimate backward regressions

$$h^*_{tj} = [icr_t, pd_t, rvfin_t, rvfin_{t-1}, rvfin_{t-13\to t-1}]' \boldsymbol{\delta}_j + u_{tj}$$

(Inclusion)

$$c^*_{tj} = \lambda\left([icr_t, pd_t, rvfin_t, rvfin_{t-1}]' \boldsymbol{\varphi}_j\right) + \varepsilon_{tj} > 0$$

(Repetition)

- Regress ICR on $z^+_{ty} \equiv \sum_j \hat{\varphi}_{jy} c_{tj}$, $z^0_{ty} \equiv \sum_j \hat{\delta}_{jy} h_{tj}$ and covariates

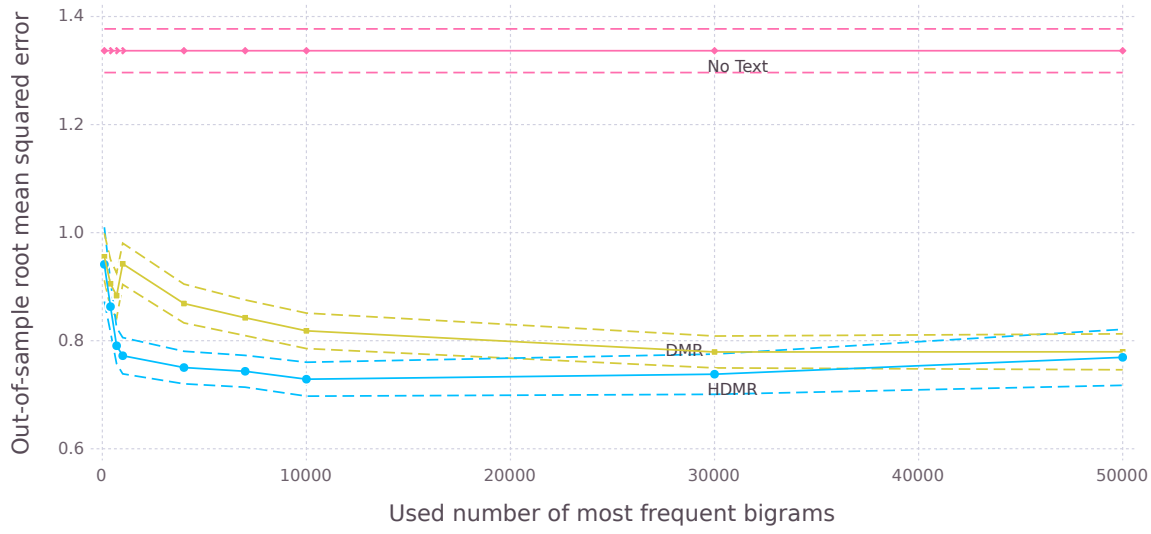
$$icr_t = \left[z^+_{ty}, z^0_{ty}, pd_t, rvfin_t, rvfin_{t-1}, rvfin_{t-13\to t-1}, m_t\right]' \boldsymbol{\beta} + v_t$$

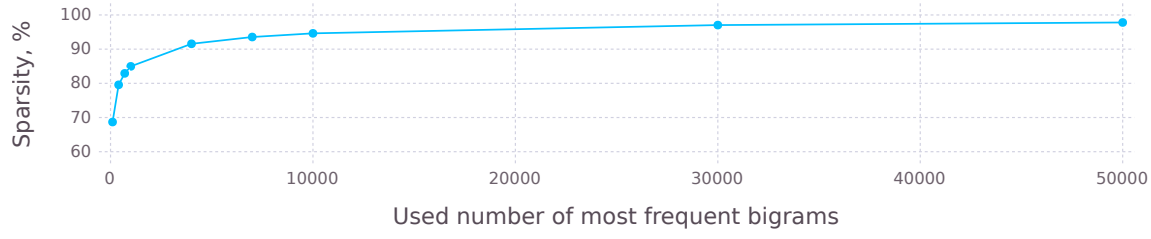
(forward regression)

- Predict out-of-sample
  - Cross-validation with 10 random folds
  - Pseudo out-of-sample rolling regressions
  - Report root mean squared error (RMSE)

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

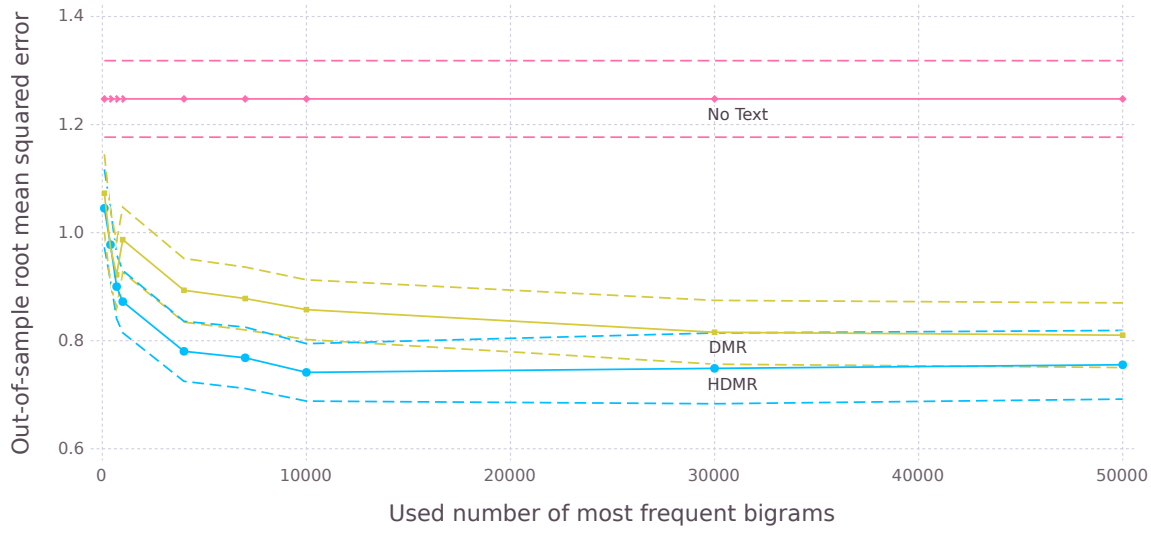
HDMR's out-of-sample fit advantage changes with text sparsity (10-fold cross validation)

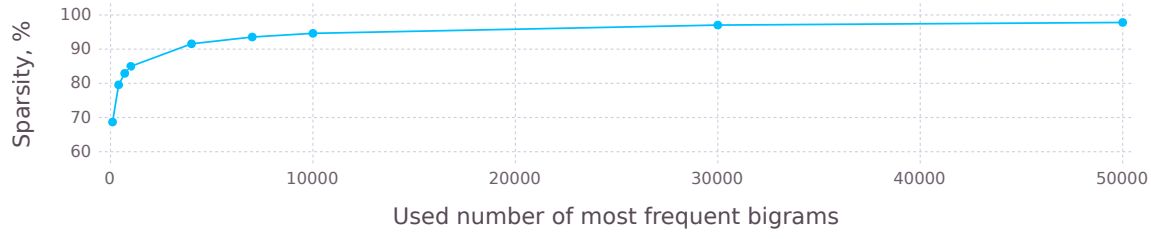




Out-of-sample prediction of ICR with text and covariates

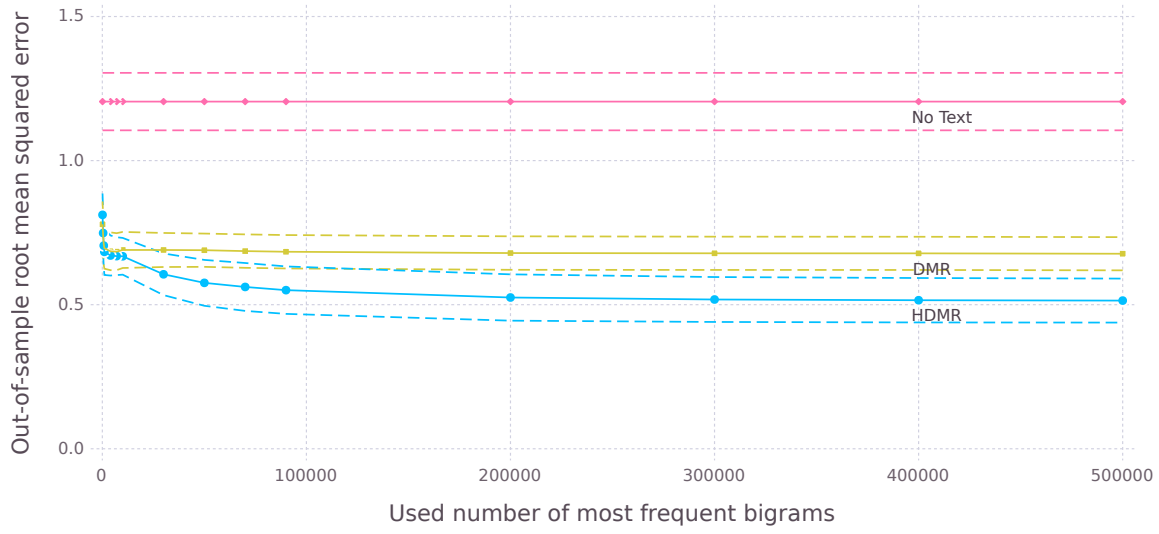
HDMR's out-of-sample fit advantage changes with text sparsity (Pseudo out-of-sample)

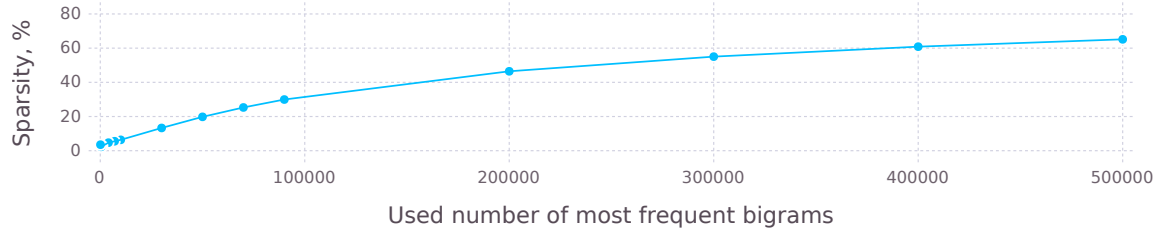




Denser text: HDMR's advantage increases with sparsity

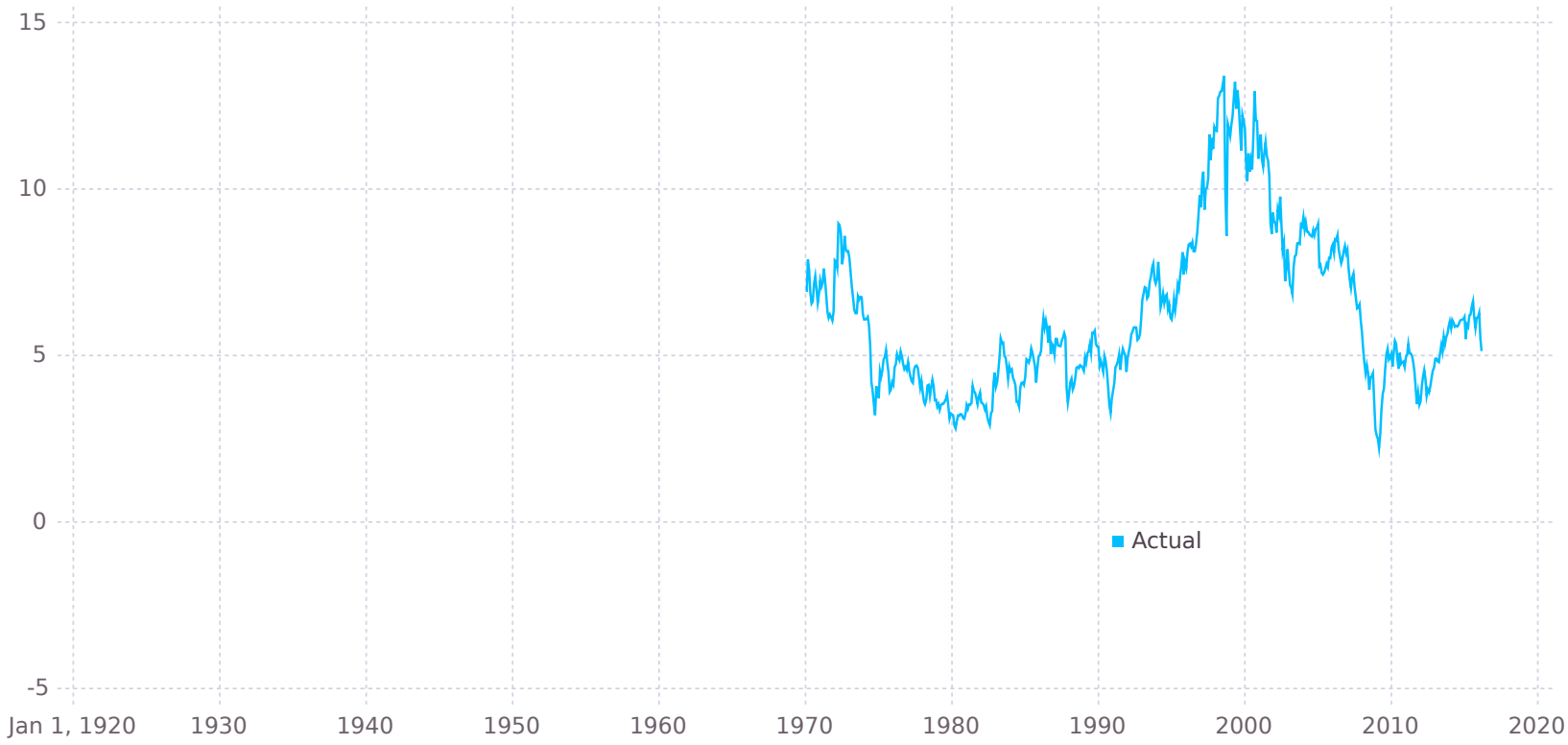
Full WSJ monthly phrase counts, January 1990 to December 2010





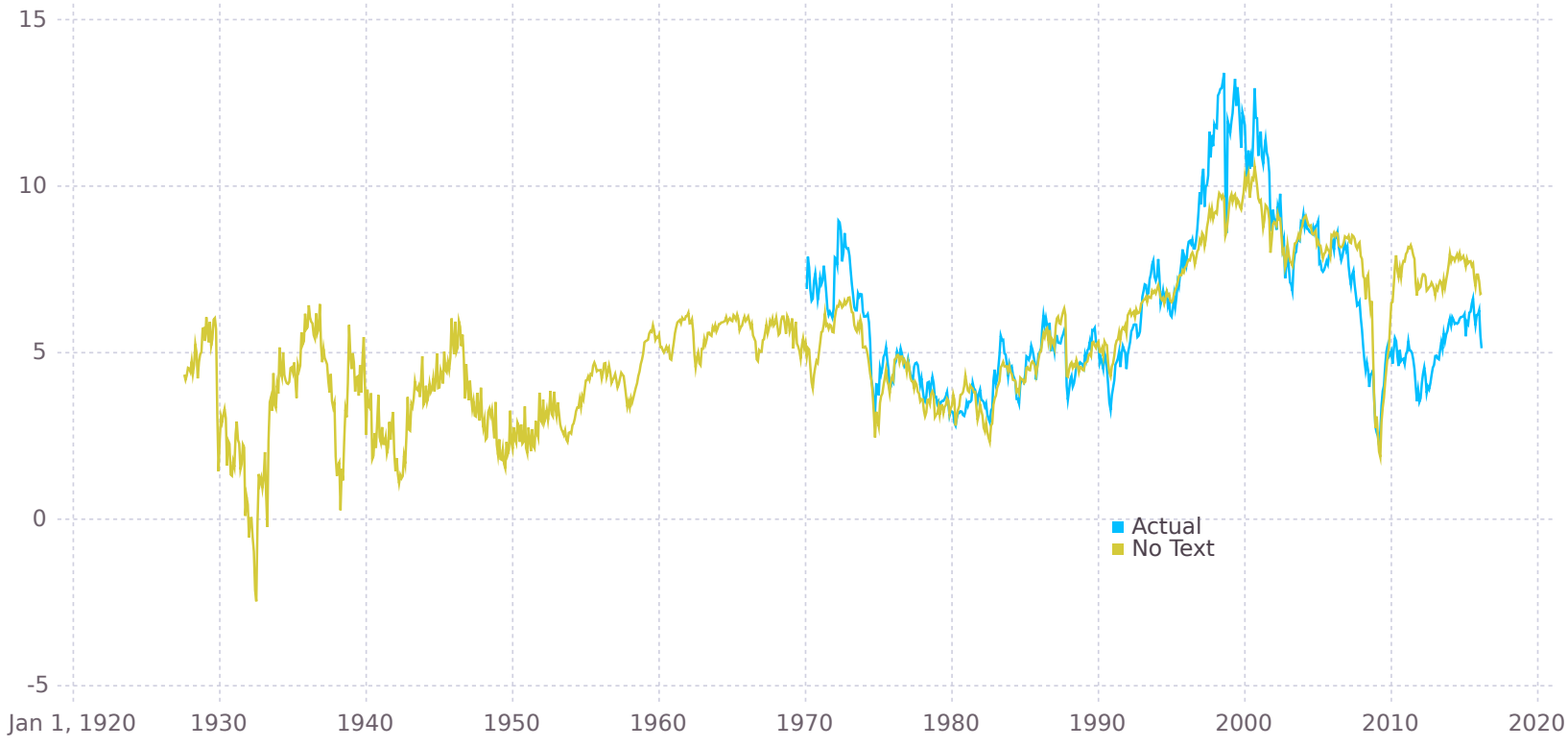
News implied intermediary capital ratio

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



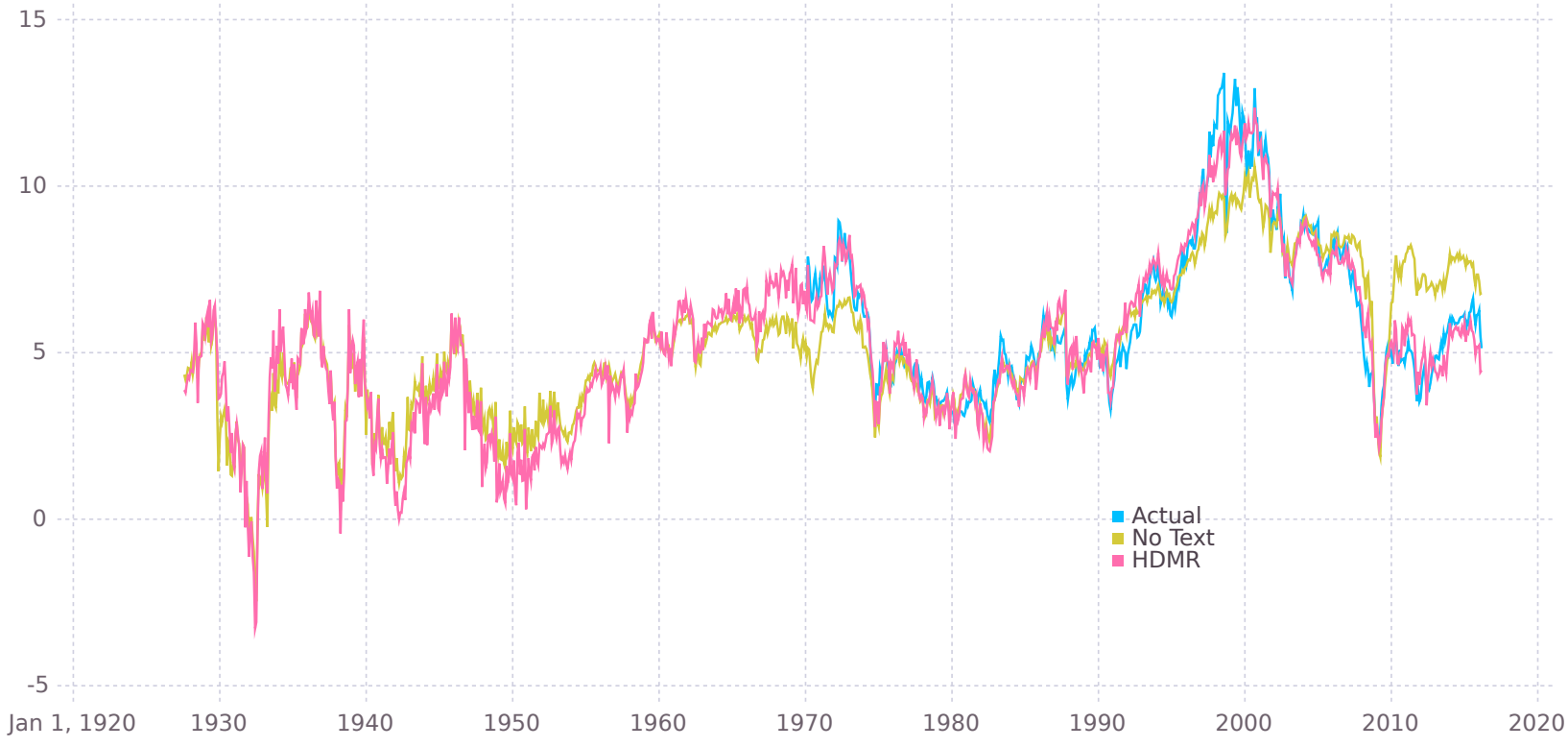
News implied intermediary capital ratio

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



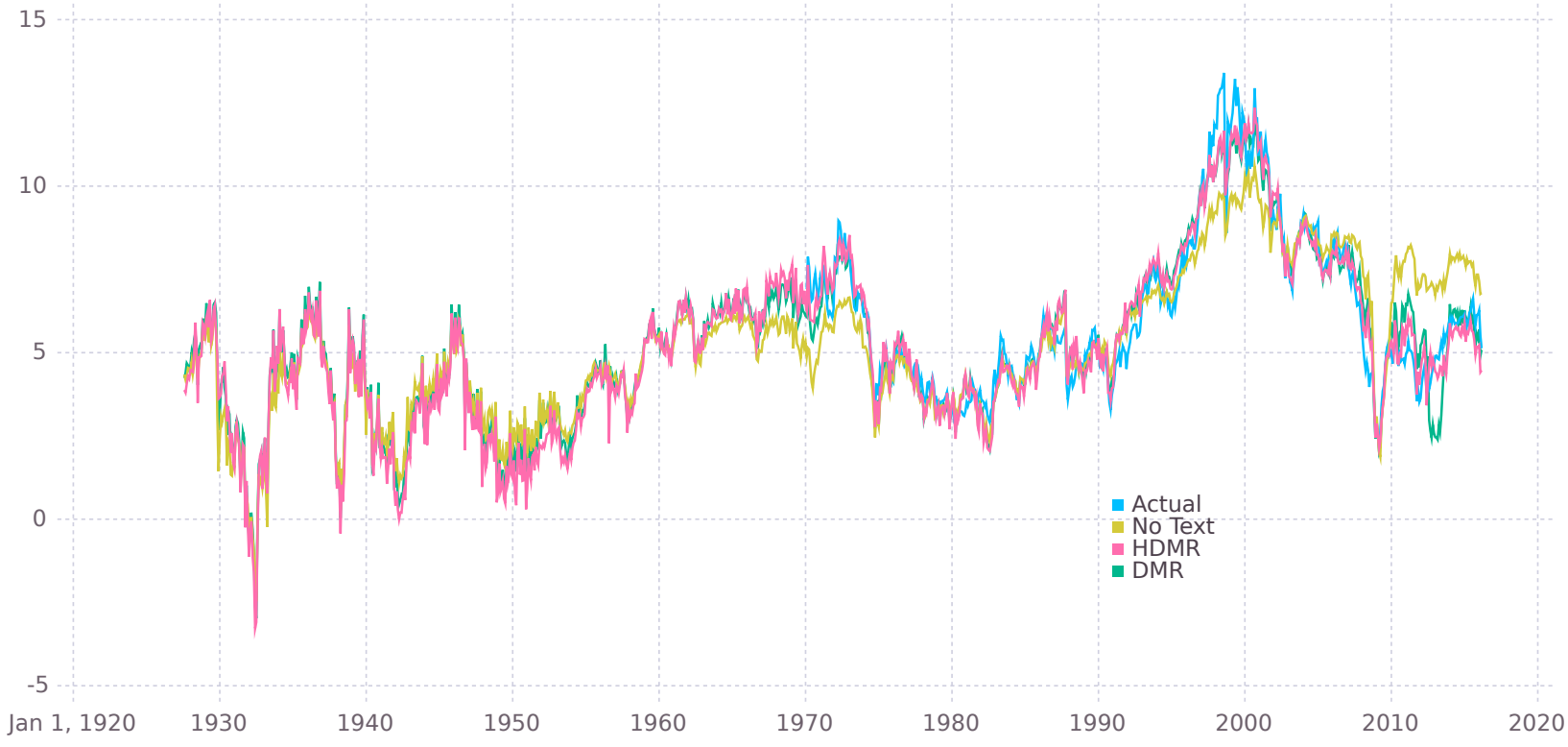
News implied intermediary capital ratio

HDMR gives a different predicted series exploiting text inclusion and repetition



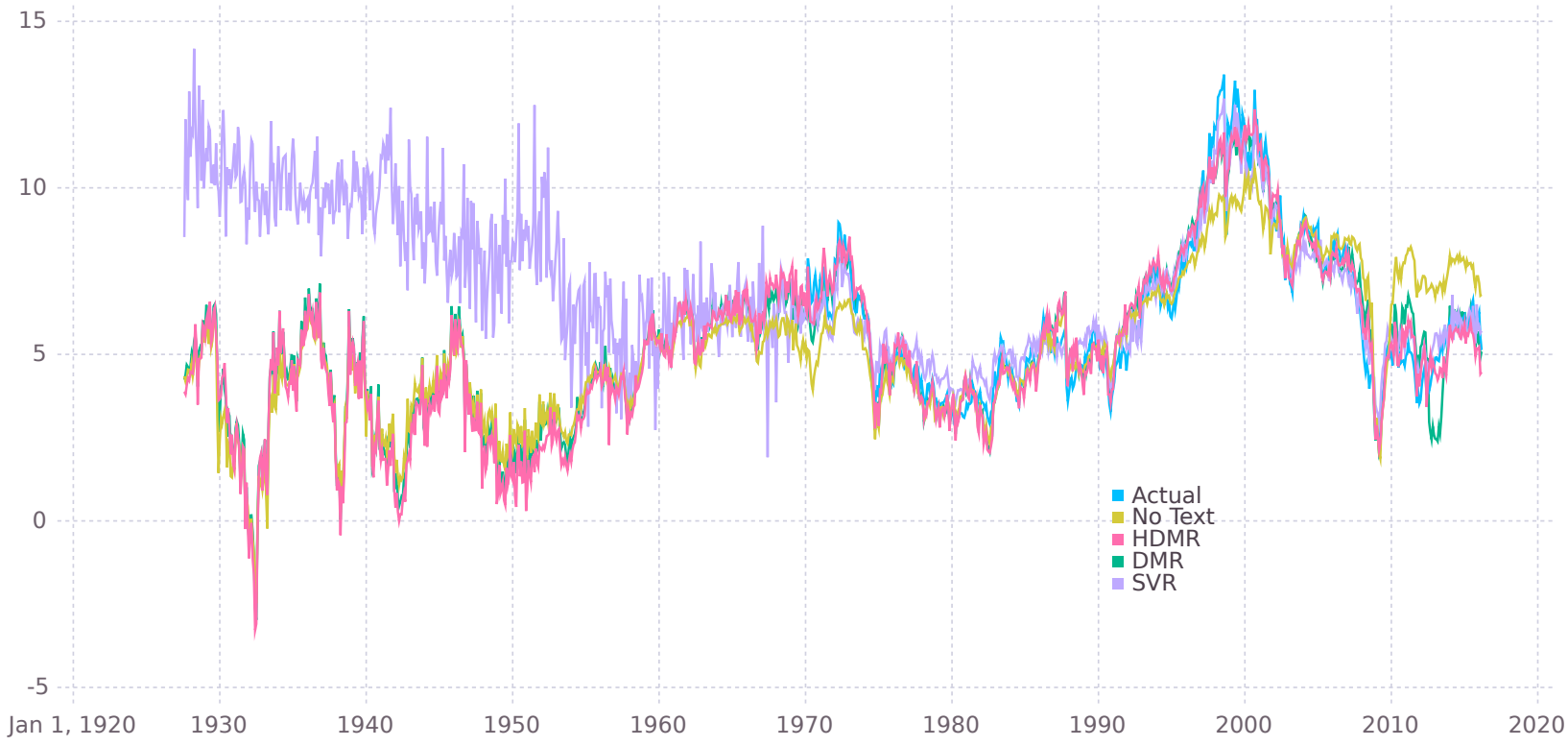
News implied intermediary capital ratio

DMR uses same information as HDMR but does not separate inclusion from repetition



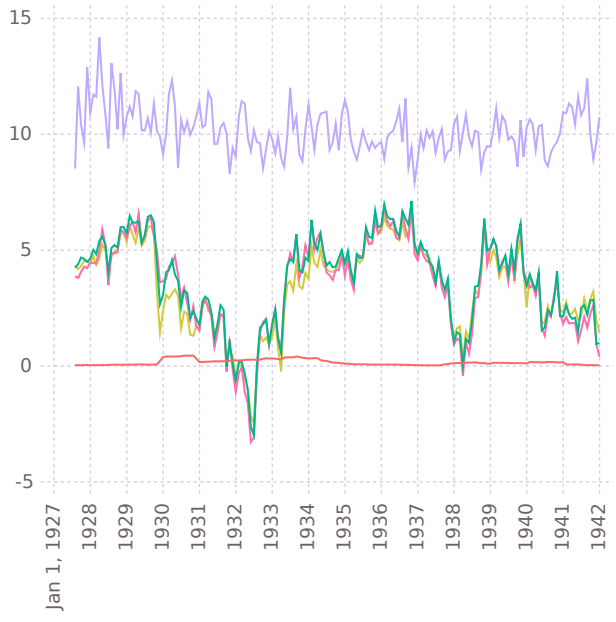
News implied intermediary capital ratio

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

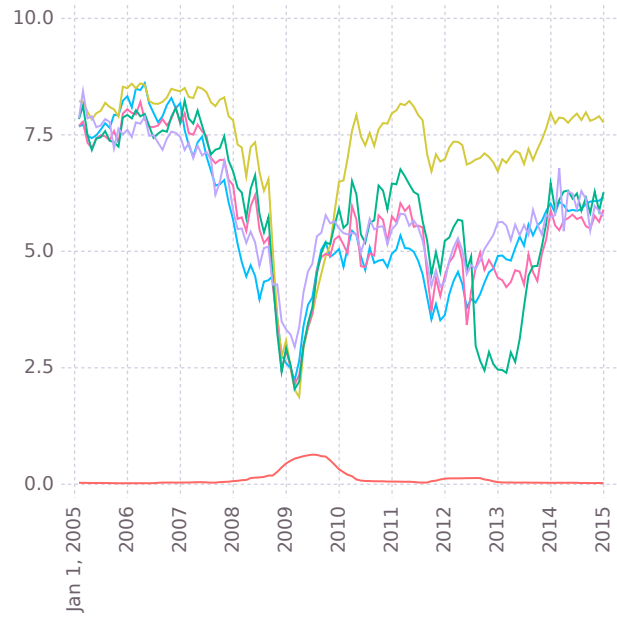


News implied intermediary capital ratio

Great Depression intermediaries were insolvent. Great Recession was almost as bad.



Great Depression



Great Recession

News-implied ICR predicts market returns

Consistent with He-Krishnamurthy (2013), 1σ higher ICR means 4.8pp lower risk premium

	$r^{em}_{t\to t+1}$			$r^{em}_{t\to t+3}$			$r^{em}_{t\to t+12}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
icr_t	-1.29			-1.38			-1.22		
	(0.97)			(1.07)			(1.03)		
$\widehat{icr}_t$		-2.05***			-2.11***			-2.19***	
		(0.74)			(0.77)			(0.80)	
z^0_t			-43.93*			-42.68**			-43.66**
			(23.82)			(21.32)			(19.51)
z^+_t			41.20			42.85			38.58
			(71.65)			(63.02)			(59.36)
pd_t			-13.92***			-14.68***			-15.19***
			(4.95)			(5.14)			(5.19)
$rvfin_{t-1\to t}$			-95.23***			-43.59*			-8.82
			(27.52)			(23.45)			(7.70)
$rvfin_{t-2\to t-1}$			43.94			6.23			7.83
			(29.65)			(21.34)			(7.22)
$rvfin_{t-13\to t-1}$			44.63			42.31*			11.47
			(31.76)			(23.71)			(20.88)
N	552	841	841	552	841	841	544	833	833
Adjusted R^2	0.14	0.78	1.99	0.84	2.53	3.92	3.01	10.56	12.93

Hodrick (1992) standard errors are in parentheses

News-implied ICR predicts market returns

Similar magnitudes comparing postwar to full sample, but early sample has less text

	$r^{em}_{t\to t+1}$			$r^{em}_{t\to t+3}$			$r^{em}_{t\to t+12}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
icr_t	-1.29			-1.38			-1.22		
	(0.97)			(1.07)			(1.03)		
$\widehat{icr}_t$		-1.92**			-2.05*			-2.14**	
		(0.81)			(1.23)			(0.95)	
z^0_t			-31.56			-34.28			-37.49*
			(28.07)			(25.32)			(20.82)
z^+_t			-7.95			-2.64			-11.51
			(84.44)			(71.63)			(68.51)
pd_t			-13.10**			-13.87*			-14.51**
			(5.37)			(7.64)			(5.91)
$rvfin_{t-1\to t}$			-27.43			-7.76			-4.73
			(17.91)			(18.51)			(8.28)
$rvfin_{t-2\to t-1}$			29.31			0.57			-0.93
			(19.18)			(16.82)			(8.42)
$rvfin_{t-13\to t-1}$			-30.74			-17.60			-21.89
			(27.38)			(40.99)			(37.44)
N	552	1,062	1,061	552	1,062	1,061	544	1,054	1,053
Adjusted R^2	0.14	0.43	0.42	0.84	1.50	1.48	3.01	6.64	8.43

Hodrick (1992) standard errors are in parentheses

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

Variable	Sparsity	Top positive	Top negative
icr_t^0	0.577	busi bulletin, tax report, labor letter, washington wire, presid clinton	barack obama, presid barack, week output,
pd_t^0	0.587	barack obama, insid journal, peopl familiar, gold comex, comex troy	labor letter, busi bulletin, tax report, washi
$rvfin_{t-1\to t}^0$	0.694	wall street, white hous, steel buy, presid clinton, chief execut	secretari christoph, tutsi rebel, wire wall, se
$rvfin_{t-2\to t-1}^0$	0.725	wall street, week mar, rwandan refuge, eastern ukrain, peopl familiar	wire wall, titl front, secretari perri, pictur c
$rvfin_{t-13\to t-1}^0$	0.652	chief execut, presid barack, steel buy, yr trea, trea yld	secretari christoph, dole ks, yitzhak rabin, v
icr_t^+	0.860	ounc dow, lopez obrador, presid clinton, clan leader, bosnia muslim	yr treasuri, treasuri yld, wsj research, c c, b
pd_t^+	0.880	c c, yr treasuri, treasuri yld, avail headlin, bond yr	presid clinton, barrel dow, lopez obrador, w
$rvfin_{t-1\to t}^+$	0.883	residenti construct, & unfil, temporari help, earn busi, israel plo	week aug, shimon pere, minist john, haiti n
$rvfin_{t-2\to t-1}^+$	0.893	intern telephon, week aug, telegraph corp, clan leader, fix incom	temporari help, radovan karadz, & paperbo

- Predicted icr shaped by WSJ front page mentions of
  - Business and economic news
  - Government policy
  - Asset prices
  - Conditional on contemporaneous valuation ratios and financial volatility
- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

Variable	Sparsity	Top positive	Top negative
icr_t^0	0.577	busi bulletin, tax report, labor letter, washington wire, presid clinton	barack obama, presid barack, week output,
pd_t^0	0.587	barack obama, insid journal, peopl familiar, gold comex, comex troy	labor letter, busi bulletin, tax report, washi
$rvfin_{t-1\to t}^0$	0.694	wall street, white hous, steel buy, presid clinton, chief execut	secretari christoph, tutsi rebel, wire wall, se
$rvfin_{t-2\to t-1}^0$	0.725	wall street, week mar, rwandan refuge, eastern ukrain, peopl familiar	wire wall, titl front, secretari perri, pictur c
$rvfin_{t-13\to t-1}^0$	0.652	chief execut, presid barack, steel buy, yr trea, trea yld	secretari christoph, dole ks, yitzhak rabin, v
icr_t^+	0.860	ounc dow, lopez obrador, presid clinton, clan leader, bosnia muslim	yr treasuri, treasuri yld, wsj research, c c, b
pd_t^+	0.880	c c, yr treasuri, treasuri yld, avail headlin, bond yr	presid clinton, barrel dow, lopez obrador, w
$rvfin_{t-1\to t}^+$	0.883	residenti construct, & unfil, temporari help, earn busi, israel plo	week aug, shimon pere, minist john, haiti n
$rvfin_{t-2\to t-1}^+$	0.893	intern telephon, week aug, telegraph corp, clan leader, fix incom	temporari help, radovan karadz, & paperbo

- Predicted icr shaped by WSJ front page mentions of
  - Business and economic news
  - Government policy
  - Asset prices
  - Conditional on contemporaneous valuation ratios and financial volatility
- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

Variable	Sparsity	Top positive	Top negative
icr_t^0	0.577	busi bulletin, tax report, labor letter, washington wire, presid clinton	barack obama, presid barack, week output,
pd_t^0	0.587	barack obama, insid journal, peopl familiar, gold comex, comex troy	labor letter, busi bulletin, tax report, washi
$rvfin_{t-1\to t}^0$	0.694	wall street, white hous, steel buy, presid clinton, chief execut	secretari christoph, tutsi rebel, wire wall, se
$rvfin_{t-2\to t-1}^0$	0.725	wall street, week mar, rwandan refuge, eastern ukrain, peopl familiar	wire wall, titl front, secretari perri, pictur c
$rvfin_{t-13\to t-1}^0$	0.652	chief execut, presid barack, steel buy, yr trea, trea yld	secretari christoph, dole ks, yitzhak rabin, v
icr_t^+	0.860	ounc dow, lopez obrador, presid clinton, clan leader, bosnia muslim	yr treasuri, treasuri yld, wsj research, c c, b
pd_t^+	0.880	c c, yr treasuri, treasuri yld, avail headlin, bond yr	presid clinton, barrel dow, lopez obrador, w
$rvfin_{t-1\to t}^+$	0.883	residenti construct, & unfil, temporari help, earn busi, israel plo	week aug, shimon pere, minist john, haiti n
$rvfin_{t-2\to t-1}^+$	0.893	intern telephon, week aug, telegraph corp, clan leader, fix incom	temporari help, radovan karadz, & paperbo

- Predicted icr shaped by WSJ front page mentions of
  - Business and economic news
  - Government policy
  - Asset prices
  - Conditional on contemporaneous valuation ratios and financial volatility
- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

Variable	Sparsity	Top positive	Top negative
icr_t^0	0.577	busi bulletin, tax report, labor letter, washington wire, presid clinton	barack obama, presid barack, week output,
pd_t^0	0.587	barack obama, insid journal, peopl familiar, gold comex, comex troy	labor letter, busi bulletin, tax report, washi
$rvfin_{t-1\to t}^0$	0.694	wall street, white hous, steel buy, presid clinton, chief execut	secretari christoph, tutsi rebel, wire wall, se
$rvfin_{t-2\to t-1}^0$	0.725	wall street, week mar, rwandan refuge, eastern ukrain, peopl familiar	wire wall, titl front, secretari perri, pictur c
$rvfin_{t-13\to t-1}^0$	0.652	chief execut, presid barack, steel buy, yr trea, trea yld	secretari christoph, dole ks, yitzhak rabin, v
icr_t^+	0.860	ounc dow, lopez obrador, presid clinton, clan leader, bosnia muslim	yr treasuri, treasuri yld, wsj research, c c, b
pd_t^+	0.880	c c, yr treasuri, treasuri yld, avail headlin, bond yr	presid clinton, barrel dow, lopez obrador, w
$rvfin_{t-1\to t}^+$	0.883	residenti construct, & unfil, temporari help, earn busi, israel plo	week aug, shimon pere, minist john, haiti n
$rvfin_{t-2\to t-1}^+$	0.893	intern telephon, week aug, telegraph corp, clan leader, fix incom	temporari help, radovan karadz, & paperbo

- Predicted icr shaped by WSJ front page mentions of
  - Business and economic news
  - Government policy
  - Asset prices
  - Conditional on contemporaneous valuation ratios and financial volatility
- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

Variable	Sparsity	Top positive	Top negative
icr_t^0	0.577	busi bulletin, tax report, labor letter, washington wire, presid clinton	barack obama, presid barack, week output,
pd_t^0	0.587	barack obama, insid journal, peopl familiar, gold comex, comex troy	labor letter, busi bulletin, tax report, washi
$rvfin_{t-1\to t}^0$	0.694	wall street, white hous, steel buy, presid clinton, chief execut	secretari christoph, tutsi rebel, wire wall, se
$rvfin_{t-2\to t-1}^0$	0.725	wall street, week mar, rwandan refuge, eastern ukrain, peopl familiar	wire wall, titl front, secretari perri, pictur c
$rvfin_{t-13\to t-1}^0$	0.652	chief execut, presid barack, steel buy, yr trea, trea yld	secretari christoph, dole ks, yitzhak rabin, v
icr_t^+	0.860	ounc dow, lopez obrador, presid clinton, clan leader, bosnia muslim	yr treasuri, treasuri yld, wsj research, c c, b
pd_t^+	0.880	c c, yr treasuri, treasuri yld, avail headlin, bond yr	presid clinton, barrel dow, lopez obrador, w
$rvfin_{t-1\to t}^+$	0.883	residenti construct, & unfil, temporari help, earn busi, israel plo	week aug, shimon pere, minist john, haiti n
$rvfin_{t-2\to t-1}^+$	0.893	intern telephon, week aug, telegraph corp, clan leader, fix incom	temporari help, radovan karadz, & paperbo

- Predicted icr shaped by WSJ front page mentions of
  - Business and economic news
  - Government policy
  - Asset prices
  - Conditional on contemporaneous valuation ratios and financial volatility
- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

Variable	Sparsity	Top positive	Top negative
icr_t^0	0.577	busi bulletin, tax report, labor letter, washington wire, presid clinton	barack obama, presid barack, week output,
pd_t^0	0.587	barack obama, insid journal, peopl familiar, gold comex, comex troy	labor letter, busi bulletin, tax report, washi
$rvfin_{t-1\to t}^0$	0.694	wall street, white hous, steel buy, presid clinton, chief execut	secretari christoph, tutsi rebel, wire wall, se
$rvfin_{t-2\to t-1}^0$	0.725	wall street, week mar, rwandan refuge, eastern ukrain, peopl familiar	wire wall, titl front, secretari perri, pictur c
$rvfin_{t-13\to t-1}^0$	0.652	chief execut, presid barack, steel buy, yr trea, trea yld	secretari christoph, dole ks, yitzhak rabin, v
icr_t^+	0.860	ounc dow, lopez obrador, presid clinton, clan leader, bosnia muslim	yr treasuri, treasuri yld, wsj research, c c, b
pd_t^+	0.880	c c, yr treasuri, treasuri yld, avail headlin, bond yr	presid clinton, barrel dow, lopez obrador, w
$rvfin_{t-1\to t}^+$	0.883	residenti construct, & unfil, temporari help, earn busi, israel plo	week aug, shimon pere, minist john, haiti n
$rvfin_{t-2\to t-1}^+$	0.893	intern telephon, week aug, telegraph corp, clan leader, fix incom	temporari help, radovan karadz, & paperbo

- Predicted icr shaped by WSJ front page mentions of
  - Business and economic news
  - Government policy
  - Asset prices
  - Conditional on contemporaneous valuation ratios and financial volatility
- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008

Explaining the text with ICR-related covariates

WSJ front page monthly, January 1970 to February 2016

Variable	Sparsity	Top positive	Top negative
icr_t^0	0.577	busi bulletin, tax report, labor letter, washington wire, presid clinton	barack obama, presid barack, week output,
pd_t^0	0.587	barack obama, insid journal, peopl familiar, gold comex, comex troy	labor letter, busi bulletin, tax report, washi
$rvfin_{t-1\to t}^0$	0.694	wall street, white hous, steel buy, presid clinton, chief execut	secretari christoph, tutsi rebel, wire wall, se
$rvfin_{t-2\to t-1}^0$	0.725	wall street, week mar, rwandan refuge, eastern ukrain, peopl familiar	wire wall, titl front, secretari perri, pictur c
$rvfin_{t-13\to t-1}^0$	0.652	chief execut, presid barack, steel buy, yr trea, trea yld	secretari christoph, dole ks, yitzhak rabin, v
icr_t^+	0.860	ounc dow, lopez obrador, presid clinton, clan leader, bosnia muslim	yr treasuri, treasuri yld, wsj research, c c, b
pd_t^+	0.880	c c, yr treasuri, treasuri yld, avail headlin, bond yr	presid clinton, barrel dow, lopez obrador, w
$rvfin_{t-1\to t}^+$	0.883	residenti construct, & unfil, temporari help, earn busi, israel plo	week aug, shimon pere, minist john, haiti n
$rvfin_{t-2\to t-1}^+$	0.893	intern telephon, week aug, telegraph corp, clan leader, fix incom	temporari help, radovan karadz, & paperbo

- Predicted icr shaped by WSJ front page mentions of
  - Business and economic news
  - Government policy
  - Asset prices
  - Conditional on contemporaneous valuation ratios and financial volatility
- Some phrases capture fairly robust features of the data
- Others are unlikely to be useful for prediction before 2008

Focus on a single phrase for intuition

1σ increase in past year financial vol increases "financial crisis" inclusion odds by 30%

	HDMR		DMR
	Repetition	Inclusion	
Intercept	-9.94	-18.20	-16.66
icr_t	-0.35	-0.59	-0.61
pd_t	1.55	3.76	3.49
$rvfin_{t-1\to t}$	1.23	0.85	1.44
$rvfin_{t-2\to t-1}$	-0.56	1.07	-0.54
$rvfin_{t-13\to t-1}$		2.80	1.26

Backward regressions

⇒

	HDMR	DMR
Repetition	-2.66	-4.73
Inclusion	-4.51	

Forward regressions

Focus on a single phrase for intuition

1σ increase in past year financial vol increases "financial crisis" inclusion odds by 30%

	HDMR		DMR
	Repetition	Inclusion	
Intercept	-9.94	-18.20	-16.66
icr_t	-0.35	-0.59	-0.61
pd_t	1.55	3.76	3.49
$rvfin_{t-1\to t}$	1.23	0.85	1.44
$rvfin_{t-2\to t-1}$	-0.56	1.07	-0.54
$rvfin_{t-13\to t-1}$		2.80	1.26

Backward regressions

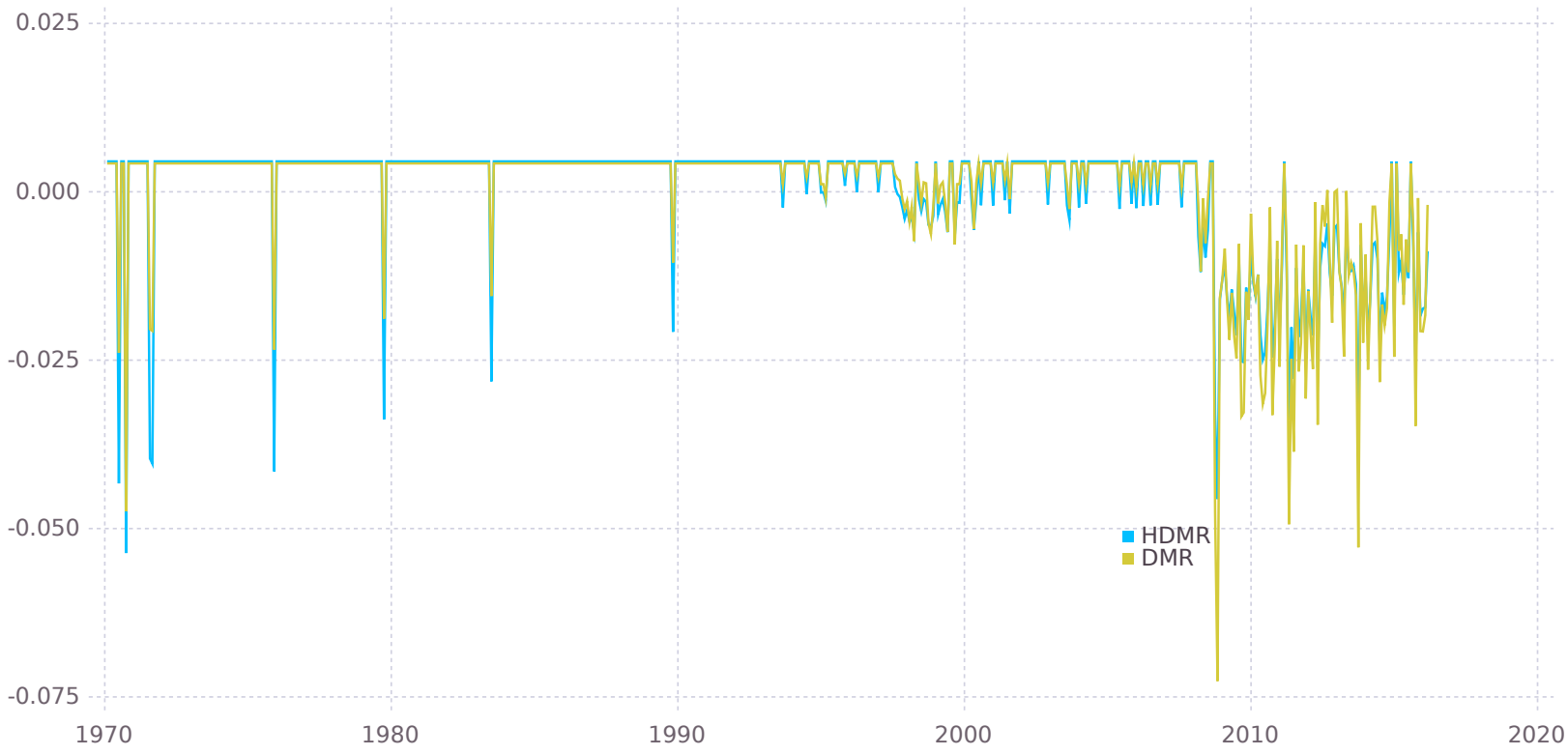
$\Rightarrow$

	HDMR	DMR
Repetition	-2.66	-4.73
Inclusion	-4.51	

Forward regressions

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 are hard to beat

$$Y_{t+h}^h = \beta_0 + \left[pc_t^1, \ldots, pc_t^5\right]' \boldsymbol{\beta} + \varepsilon_{t+h} \qquad \text{(DFM-5)}$$

- We use their data + WSJ text to forecast 1–12 months ahead

$$Y_{t+h}^h = \beta_0 + \left[z_{tY}^0, z_{tY}^+, pc_t^1, \ldots, pc_t^5\right]' \boldsymbol{\beta} + \varepsilon_{t+h} \qquad \text{(HDMR)}$$

Main 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 helps predict asset prices directly (stocks, treasuries, currencies) in quarterly/annual horizon but not monthly
- Advantage of HDMR increases with sparsity of the text
- Stronger results for nowcasting

Significant improvements in out-of-sample forecasting

HDMR RMSE relative to DFM-5: WSJ full text, 10,000 bigrams

Y^h_{t+h}	$h=1$	$h=3$	$h=12$
IP: total	0.984	0.897**	0.775***
	(0.372)	(0.042)	(0.000)
Emp: total	0.881**	0.828***	0.785***
	(0.025)	(0.001)	(0.000)
U: all	0.956	0.825***	0.709***
	(0.226)	(0.003)	(0.000)
HStarts: Total	0.716***	0.658***	0.618***
	(0.000)	(0.000)	(0.000)
PMI	0.838***	0.852***	0.735***
	(0.000)	(0.000)	(0.000)
CPI-ALL	1.110	1.064	1.030
	(1.000)	(1.000)	(0.939)
Real AHE: goods	0.985	0.902***	0.629***
	(0.328)	(0.003)	(0.000)
FedFunds	0.952*	0.842***	0.677***
	(0.068)	(0.000)	(0.000)
M1	1.100	1.102	0.990
	(1.000)	(0.999)	(0.340)
Ex rate: avg	1.065	0.997	0.881***
	(0.999)	(0.462)	(0.001)
S&P 500	1.041	0.941*	0.793***
	(0.888)	(0.085)	(0.000)
Consumer expect	1.110	1.043	0.980
	(1.000)	(0.874)	(0.337)

Diebold-Mariano (1995) p-values are in parentheses

- Text is informative about future
  - Short and long run fundamentals
  - Long run fundamentals and prices
- Advantage of HDMR increases with text sparsity
- Text is also useful for nowcasting

Significant improvements in out-of-sample forecasting

HDMR RMSE relative to DFM-5: WSJ full text, 100,000 bigrams

Y^h_{t+h}	$h=1$	$h=3$	$h=12$
IP: total	0.976	0.864***	0.619***
	(0.313)	(0.003)	(0.000)
Emp: total	0.769***	0.690***	0.506***
	(0.000)	(0.000)	(0.000)
U: all	0.986	0.819***	0.530***
	(0.412)	(0.001)	(0.000)
HStarts: Total	0.546***	0.485***	0.519***
	(0.000)	(0.000)	(0.000)
PMI	0.829***	0.842***	0.798***
	(0.000)	(0.000)	(0.000)
CPI-ALL	1.012	1.051	1.081
	(0.757)	(0.984)	(0.998)
Real AHE: goods	1.045	0.985	0.644***
	(0.971)	(0.319)	(0.000)
FedFunds	0.996	0.925**	0.701***
	(0.451)	(0.028)	(0.000)
M1	1.040	1.029	1.058
	(1.000)	(0.984)	(0.994)
Ex rate: avg	1.073	1.009	0.738***
	(0.999)	(0.641)	(0.000)
S&P 500	1.054	0.881**	0.663***
	(0.918)	(0.010)	(0.000)
Consumer expect	1.061	1.097	0.852***
	(1.000)	(0.990)	(0.002)

Diebold-Mariano (1995) p-values are in parentheses

- Text is informative about future
  - Short and long run fundamentals
  - Long run fundamentals and prices
- Advantage of HDMR increases with text sparsity
- Text is also useful for nowcasting

Significant improvements in out-of-sample nowcasting

HDMR RMSE relative to DFM-5: WSJ full text, 100,000 bigrams

Y^h_{t+h}	$h=1$	$h=3$	$h=12$
IP: total	0.975	0.850***	0.626***
	(0.167)	(0.000)	(0.000)
Emp: total	0.771***	0.694***	0.524***
	(0.000)	(0.000)	(0.000)
U: all	0.982	0.819***	0.542***
	(0.302)	(0.000)	(0.000)
HStarts: Total	0.527***	0.509***	0.475***
	(0.000)	(0.000)	(0.000)
PMI	0.851***	0.866***	0.847***
	(0.000)	(0.000)	(0.000)
CPI-ALL	0.962*	1.022	1.098
	(0.082)	(0.904)	(0.994)
Real AHE: goods	1.047	1.010	0.622***
	(0.910)	(0.610)	(0.000)
FedFunds	0.974	0.888***	0.728***
	(0.310)	(0.009)	(0.000)
M1	1.072	1.050	1.064
	(1.000)	(0.998)	(0.991)
Ex rate: avg	1.080	1.025	0.743***
	(1.000)	(0.763)	(0.000)
S&P 500	1.040	0.902**	0.700***
	(0.890)	(0.016)	(0.000)
Consumer expect	1.045	1.088	0.851***
	(0.998)	(0.991)	(0.000)

Diebold-Mariano (1995) p-values are in parentheses

- Text is informative about future
  - Short and long run fundamentals
  - Long run fundamentals and prices
- Advantage of HDMR increases with text sparsity
- Text is also useful 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

Backcasting application

Summary statistics

Variable	Mean	Std	Min	p10	Median	p90	Max	Obs	Available
Phrase counts, c_{tj}	0.086	0.379	0.000	0.000	0.003	0.114	4.576	1075	192607–201602
Phrase indic. h_{tj}	0.054	0.212	0.000	0.000	0.002	0.089	1.000	1075	192607–201602
icr	6.236	2.399	2.230	3.616	5.574	9.578	13.400	557	197001–201605
pd	3.442	0.402	2.213	2.960	3.394	4.017	4.564	1075	192611–201605
$rvfin_{t-1\to t}$	0.061	0.144	0.002	0.006	0.022	0.133	2.059	1079	192607–201605
$rvfin_{t-12\to t}$	0.061	0.094	0.004	0.010	0.026	0.159	0.636	1068	192706–201605

Macro forecasting

Summary statistics

Variable	Mean	Std	Min	p10	Median	p90	Max	Obs	Available
Phrase counts, c_{tj}	2.971	2.732	0.178	0.590	2.342	6.190	16.909	252	199001–201012
Phrase indic. h_{tj}	0.680	0.409	0.032	0.161	0.860	0.998	1.000	252	199001–201012
IP: total	0.855	3.157	-16.004	-2.697	1.067	4.348	11.810	602	195901–200902
Emp: total	0.606	1.075	-4.177	-0.679	0.769	1.718	5.830	602	195901–200902
U: all	0.004	0.183	-0.700	-0.200	0.000	0.200	0.900	602	195901–200902
HStarts: Total	7.307	0.236	6.192	6.991	7.327	7.602	7.822	602	195901–200902
PMI	52.867	6.927	29.400	44.100	53.500	60.780	72.100	603	195901–200903
CPI-ALL	0.000	1.075	-5.282	-1.162	-0.003	1.091	7.018	602	195902–200903
Real AHE: goods	0.274	1.228	-4.401	-1.113	0.255	1.560	5.883	602	195901–200902
FedFunds	0.002	0.371	-1.560	-0.420	0.010	0.380	1.600	602	195901–200902
M1	0.012	2.184	-10.505	-2.344	-0.000	2.407	7.479	601	195902–200902
Ex rate: avg	-0.170	5.953	-21.103	-8.241	0.038	6.892	21.174	601	195901–200901
S&P 500	1.796	14.410	-91.153	-14.428	2.798	16.943	45.355	603	195901–200903
Consumer expect	-0.053	3.975	-16.500	-4.600	-0.200	4.600	22.500	603	195901–200903

Macro forecasting

WSJ full text, 10,000 bigrams

Months forward:	$h=1$				$h=3$				$h=12$			
Folds:	Random		Rolling		Random		Rolling		Random		Rolling	
Y_{t+h}^h	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR
IP: total	0.984	0.960	0.976	0.991	0.897**	0.892**	0.899***	0.865***	0.775***	0.793***	0.799***	0.785**
	(0.372)	(0.118)	(0.291)	(0.409)	(0.042)	(0.012)	(0.009)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Emp: total	0.881**	0.858***	0.926*	0.941	0.828***	0.831***	0.874***	0.889**	0.785***	0.778***	0.771***	0.826**
	(0.025)	(0.009)	(0.078)	(0.153)	(0.001)	(0.000)	(0.007)	(0.018)	(0.000)	(0.000)	(0.000)	(0.004)
U: all	0.956	0.948	0.975	1.007	0.825***	0.843***	0.753***	0.802***	0.709***	0.755***	0.691***	0.753**
	(0.226)	(0.189)	(0.296)	(0.559)	(0.003)	(0.010)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
HStarts: Total	0.716***	0.653***	0.783***	0.766***	0.658***	0.636***	0.736***	0.773***	0.618***	0.571***	0.815***	0.760**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PMI	0.838***	0.815***	0.967	0.980	0.852***	0.839***	0.880**	0.922*	0.735***	0.751***	0.888*	0.922
	(0.000)	(0.000)	(0.271)	(0.377)	(0.000)	(0.000)	(0.013)	(0.084)	(0.000)	(0.000)	(0.051)	(0.132)
CPI-ALL	1.110	1.096	1.116	1.188	1.064	1.091	1.054	1.056	1.030	1.020	1.041	1.003
	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(1.000)	(0.992)	(0.927)	(0.939)	(0.819)	(0.933)	(0.536)
Real AHE: goods	0.985	0.996	1.030	1.024	0.902***	0.919***	0.989	1.028	0.629***	0.627***	0.884**	0.902*
	(0.328)	(0.441)	(0.789)	(0.793)	(0.003)	(0.006)	(0.388)	(0.744)	(0.000)	(0.000)	(0.025)	(0.056)
FedFunds	0.952*	0.918***	0.994	1.056	0.842***	0.835***	0.960	1.011	0.677***	0.665***	0.840***	0.926
	(0.068)	(0.005)	(0.446)	(0.836)	(0.000)	(0.001)	(0.215)	(0.573)	(0.000)	(0.000)	(0.006)	(0.146)
M1	1.100	1.082	1.170	1.172	1.102	1.094	1.199	1.134	0.990	1.001	1.057	1.113
	(1.000)	(1.000)	(1.000)	(1.000)	(0.999)	(1.000)	(1.000)	(1.000)	(0.340)	(0.510)	(0.906)	(0.986)
Ex rate: avg	1.065	1.059	1.097	1.175	0.997	0.971	1.034	1.080	0.881***	0.835***	0.867**	0.863**
	(0.999)	(1.000)	(0.966)	(0.999)	(0.462)	(0.123)	(0.709)	(0.908)	(0.001)	(0.000)	(0.015)	(0.013)
S&P 500	1.041	1.021	1.089	1.003	0.941*	0.940*	0.939	0.895***	0.793***	0.785***	0.739***	0.736**
	(0.888)	(0.762)	(0.963)	(0.541)	(0.085)	(0.062)	(0.123)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)
Consumer expect	1.110	1.159	1.180	1.672	1.043	1.022	1.102	1.143	0.980	0.916**	0.894**	0.850**
	(1.000)	(1.000)	(1.000)	(1.000)	(0.874)	(0.780)	(0.990)	(0.993)	(0.337)	(0.036)	(0.031)	(0.003)

Macro forecasting

WSJ full text, 100,000 bigrams

Months forward:	$h=1$				$h=3$				$h=12$			
Folds:	Random		Rolling		Random		Rolling		Random		Rolling	
Y^h_{t+h}	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR
IP: total	0.976	0.961	1.025	1.000	0.864***	0.884**	0.900**	0.857***	0.619***	0.775***	0.680***	0.769**
	(0.313)	(0.176)	(0.718)	(0.496)	(0.003)	(0.016)	(0.014)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Emp: total	0.769***	0.832***	0.933	0.920*	0.690***	0.803***	0.811***	0.873***	0.506***	0.737***	0.607***	0.782**
	(0.000)	(0.005)	(0.121)	(0.073)	(0.000)	(0.000)	(0.000)	(0.005)	(0.000)	(0.000)	(0.000)	(0.000)
U: all	0.986	0.945	1.023	0.992	0.819***	0.824***	0.921**	0.785***	0.530***	0.705***	0.672***	0.726**
	(0.412)	(0.184)	(0.677)	(0.441)	(0.001)	(0.008)	(0.030)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
HStarts: Total	0.546***	0.647***	0.537***	0.764***	0.485***	0.626***	0.499***	0.762***	0.519***	0.561***	0.529***	0.758**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PMI	0.829***	0.774***	0.990	0.994	0.842***	0.787***	0.982	0.935	0.798***	0.724***	1.001	0.929
	(0.000)	(0.000)	(0.433)	(0.466)	(0.000)	(0.000)	(0.379)	(0.153)	(0.000)	(0.000)	(0.509)	(0.169)
CPI-ALL	1.012	1.152	1.065	1.246	1.051	1.130	1.032	1.107	1.081	1.039	1.067	1.034
	(0.757)	(1.000)	(1.000)	(1.000)	(0.984)	(1.000)	(0.936)	(0.995)	(0.998)	(0.926)	(0.966)	(0.847)
Real AHE: goods	1.045	1.012	1.158	1.036	0.985	0.905***	1.017	1.000	0.644***	0.595***	0.849***	0.888**
	(0.971)	(0.653)	(1.000)	(0.845)	(0.319)	(0.001)	(0.685)	(0.499)	(0.000)	(0.000)	(0.006)	(0.040)
FedFunds	0.996	0.906***	1.214	1.036	0.925**	0.805***	1.012	0.975	0.701***	0.602***	0.694***	0.846**
	(0.451)	(0.002)	(1.000)	(0.745)	(0.028)	(0.000)	(0.605)	(0.325)	(0.000)	(0.000)	(0.000)	(0.008)
M1	1.040	1.169	1.074	1.411	1.029	1.158	1.134	1.327	1.058	1.005	1.118	1.135
	(1.000)	(1.000)	(1.000)	(1.000)	(0.984)	(1.000)	(1.000)	(1.000)	(0.994)	(0.579)	(0.998)	(0.993)
Ex rate: avg	1.073	1.080	1.064	1.265	1.009	0.968	1.112	1.130	0.738***	0.805***	0.884**	0.851**
	(0.999)	(1.000)	(0.922)	(1.000)	(0.641)	(0.122)	(0.968)	(0.973)	(0.000)	(0.000)	(0.031)	(0.008)
S&P 500	1.054	1.031	1.100	1.013	0.881**	0.932*	0.916**	0.876***	0.663***	0.767***	0.632***	0.716**
	(0.918)	(0.801)	(0.980)	(0.629)	(0.010)	(0.056)	(0.043)	(0.003)	(0.000)	(0.000)	(0.000)	(0.000)
Consumer expect	1.061	1.206	1.065	1.675	1.097	1.036	1.121	1.166	0.852***	0.909**	0.968	0.853**
	(1.000)	(1.000)	(0.997)	(1.000)	(0.990)	(0.862)	(0.999)	(0.997)	(0.002)	(0.035)	(0.305)	(0.007)

Macro nowcasting

WSJ full text, 100,000 bigrams

Months forward:	$h=1$				$h=3$				$h=12$			
Folds:	Random		Rolling		Random		Rolling		Random		Rolling	
Y^h_{t+h}	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR	HDMR	DMR
IP: total	0.975	0.942**	1.039	0.971	0.850***	0.872***	0.893***	0.855***	0.626***	0.771***	0.701***	0.765**
	(0.167)	(0.020)	(0.824)	(0.249)	(0.000)	(0.001)	(0.010)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Emp: total	0.771***	0.820***	0.938	0.873***	0.694***	0.779***	0.830***	0.844***	0.524***	0.726***	0.633***	0.783**
	(0.000)	(0.000)	(0.134)	(0.009)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)
U: all	0.982	0.943	1.035	0.946	0.819***	0.818***	0.949	0.778***	0.542***	0.684***	0.705***	0.714**
	(0.302)	(0.106)	(0.793)	(0.122)	(0.000)	(0.001)	(0.112)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
HStarts: Total	0.527***	0.641***	0.521***	0.750***	0.509***	0.619***	0.509***	0.752***	0.475***	0.560***	0.507***	0.755**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
PMI	0.851***	0.801***	0.981	0.980	0.866***	0.792***	0.976	0.931	0.847***	0.725***	1.023	0.942
	(0.000)	(0.000)	(0.371)	(0.380)	(0.000)	(0.000)	(0.344)	(0.140)	(0.000)	(0.000)	(0.653)	(0.202)
CPI-ALL	0.962*	1.151	1.044	1.440	1.022	1.135	1.042	1.133	1.098	1.054	1.079	1.030
	(0.082)	(1.000)	(0.984)	(1.000)	(0.904)	(1.000)	(0.965)	(0.999)	(0.994)	(0.975)	(0.984)	(0.817)
Real AHE: goods	1.047	1.023	1.210	1.047	1.010	0.931**	1.052	0.965	0.622***	0.583***	0.823***	0.889**
	(0.910)	(0.695)	(1.000)	(0.869)	(0.610)	(0.022)	(0.927)	(0.197)	(0.000)	(0.000)	(0.001)	(0.038)
FedFunds	0.974	0.889***	1.180	1.004	0.888***	0.785***	1.026	0.954	0.728***	0.609***	0.764***	0.879**
	(0.310)	(0.005)	(1.000)	(0.527)	(0.009)	(0.000)	(0.727)	(0.209)	(0.000)	(0.000)	(0.000)	(0.030)
M1	1.072	1.271	1.091	1.558	1.050	1.138	1.116	2.005	1.064	0.996	1.111	1.110
	(1.000)	(1.000)	(1.000)	(1.000)	(0.998)	(1.000)	(1.000)	(1.000)	(0.991)	(0.459)	(0.997)	(0.986)
Ex rate: avg	1.080	1.109	1.042	1.192	1.025	0.973	1.099	1.132	0.743***	0.804***	0.911*	0.865**
	(1.000)	(0.998)	(0.840)	(0.999)	(0.763)	(0.181)	(0.948)	(0.976)	(0.000)	(0.000)	(0.079)	(0.012)
S&P 500	1.040	1.020	1.073	1.024	0.902**	0.929**	0.930*	0.896**	0.700***	0.766***	0.671***	0.731**
	(0.890)	(0.725)	(0.935)	(0.715)	(0.016)	(0.035)	(0.085)	(0.011)	(0.000)	(0.000)	(0.000)	(0.000)
Consumer expect	1.045	1.234	1.030	1.453	1.088	1.033	1.097	1.220	0.851***	0.909**	0.991	0.879**
	(0.998)	(1.000)	(0.929)	(1.000)	(0.991)	(0.866)	(0.997)	(1.000)	(0.000)	(0.022)	(0.445)	(0.021)