Text Selection

Bryan Kelly Asaf Manela Alan Moreira*

May 2019

Abstract

Text data is inherently ultra-high dimensional, which makes machine learning techniques indispensable for textual analysis. Text also tends to be a highly selected outcome—journalists, speechwriters, and others carefully craft messages to target the limited attention of their audiences. We develop an economically motivated high dimensional selection model that improves machine learning from text (and from sparse counts data more generally). Our model is especially useful in cases where the cover/no-cover choice is separate or more interesting than the coverage quantity choice. Our design allows for parallel estimation, making the model highly computationally scalable. We apply our framework to backcast, nowcast, and forecast financial variables using newspaper text, and find that it substantially improves out-of-sample fit relative to alternative state-of-the-art approaches.

Keywords: Text analysis, machine learning, selection model, high dimension forecast, intermediary capital, multinomial regression, hurdle, zero inflation

*Yale University, bryan.kelly@yale.edu; Washington University in St. Louis, amanela@wustl.edu; and University of Rochester, alan.moreira@simon.rochester.edu. We are grateful for helpful comments by Matthew Gentzkow, Hongyi Liu, Jesse Shapiro, Paul Tetlock (discussant) and by seminar participants at École Polytechnique Fédérale de Lausanne, Hebrew University, IDC Herzliya, INSEAD, Ohio State, Indiana University, Tel-aviv University, Rice University, University of Michigan, NBER AP, and the University of Chicago CITE Conference. Computations were performed using the facilities of the Washington University Center for High Performance Computing, which were partially provided through NIH grant S10 OD018091.

1 Introduction

Digital text is increasingly available to social scientists in the form of newspapers, blogs, tweets, regulatory filings, congressional records and more. Two attributes of text differentiate it from other types of data typically used by economists. First, text data are inherently ultra-high dimensional—unique phrases in a corpus (roughly equivalent to the set of variables) often number in the millions. Second, phrase counts are sparse—most phrases have a count of zero in most documents. Due to these attributes, statistical learning from text requires techniques commonly used in machine learning (Gentzkow, Kelly, and Taddy, 2019). Unlike data often used by economists, word or phrase counts in text data are inherently high dimensional because many words are used to describe similar phenomena. Statistical learning from text therefore requires regularization techniques commonly used in machine learning. This paper proposes a new methodology that fully exploits the sparseness of text data. We show that modeling the extensive margin decision to use a particular word reveals information that leads to large improvements in out of sample prediction.

A standard econometric approach for describing counts data is multinomial logistic regression. Unfortunately, this model is computationally intractable in most text-related applications because the number of categories is extremely large. Taddy (2015) shows that one can overcome this dimensionality problem by approximating the multinomial with cleverly shifted independent poisson regressions, one for each word. This *distributed multinomial regression* (DMR) has the great advantage that the poisson regression can be distributed across parallel computing units.

The poisson model has the disadvantage that it provides a poor description of word counts. In particular, there tends to be a much higher proportion of phrases having zero counts than the poisson distribution allows. If one restricts attention to positive counts, however, a (truncated) poisson is a good approximation for the data. This sparsity—the additional probability mass on zero counts—is a feature of many text samples (and is why most text analysis software packages use sparse matrices to store word counts efficiently).[1]

We propose a new regression methodology for text-based regression. Our *hurdle distributed multinomial regression* (HDMR) model consists of two components that are tailored to the twin

[1] Zipf (1935) observed that language follows a power law. Rennie, Shih, Teevan, and Karger (2003) suggest a multinomial naive bayes modification to emulate such power law distributions. Wilbur and Kim (2009) find that in some cases, word "burstiness" is so strong that additional occurrences of a word essentially add no useful information to a classifier.

challenges of high dimensionality and sparsity of text data. To accommodate the dimensionality of text, we build on Taddy's DMR insight of independent phrase-level models. However, we replace each phrase-level poisson regression with a hurdle model that has two parts. The first part is a selection equation. This models the text producer's choice of whether or not to use a particular phrase. The second part is a positive counts model, which describes the choice of how many times a word is used (conditional on being used at all).[2]

Our two-part HDMR model thus generalizes DMR by decomposing language usage decisions into an extensive margin (the selection equation) and intensive margin (the positive counts model). Explicitly modeling the extensive margin of phrase choice has dual advantages. The statistical advantage is that it introduces a modeling component dedicated to capturing the excess probability mass on zero counts. We use the hurdle approach of Mullahy (1986) to model selection because it allows the extensive and intensive components to be estimated independently, and these can also be distributed at essentially no additional computation cost relative to DMR.[3]

The economic advantage of our model is that it adapts the selection methodology of Heckman (1979) to a high dimensional setting. HDMR provides a means for estimating models in which sparsity is of first-order economic importance. In text data, HDMR is particularly useful when an author's choice to cover or not cover a topic is more economically interesting than the choice of coverage intensity. For example, newspaper publishers' extensive margin of coverage is informative about their news production technology and the constrained attention of its audience. The selection decision is a key lever that writers use to signal their ideological type to readers (e.g. Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2006). Politicians carefully select phrases that resonate with voters in congressional speech (Gentzkow, Shapiro, and Taddy, 2019), and fixed costs of using censored or socially taboo words may generate further sparsity (Michel, Shen, Aiden, Veres, Gray, Pickett, Hoiberg, Clancy, Norvig, Orwant, et al., 2011). Sparsity reflects the suboptimality of writing about a particular topic just a little bit: In order for the signal to be comprehensible there must be a minimum amount of exposition, yet the total amount of text cannot exceed the audience's attention span.[4] Gabaix (2014) shows that, when agents are boundedly rational, sparsity emerges

[2] We use L1 regularization (lasso) to further manage model dimensionality.

[3] The hurdle model is widely used, for example, in health economics. Greene (2007) surveys models for counts.

[4] Academic researchers know this well, and therefore tend to use consistent wording to clarify their argument, rather than alternate between synonyms to expound the same contention.

in equilibrium for a wide variety of economic settings under otherwise general conditions.

Our model also serves as a basis for exploring the relationships between phrases and numerical covariates such as macroeconomic state variables or financial markets prices, which enter as conditioning variables for the distribution of phrase counts. First, the model doubles as a dimension reduction method for text. Like DMR, HDMR generates *sufficient reductions* of text by projecting phrase counts onto covariates. Sufficient reductions serve as indices that best summarize the text as it relates to each covariate. HDMR produces two such sufficient reductions per covariate: one for word inclusion (the extensive margin) and the other for repetition (the intensive margin). Second, the model can be flipped in order to predict covariates based on text via an inverse regression. This is useful, for example, to backcast key macro variables that have limited histories or missing data, or for "nowcasting" when text is available in a more timely manner than the forecast target.

To illustrate our methodology, we use HDMR to backcast a measure of equity capitalization in the financial sector using the text of *The Wall Street Journal*. The intermediary capital ratio (*icr*) is the central state variable in the growing literature on intermediary-based asset pricing, and helps explain the behavior of risk premia in a wide array of asset classes (He, Kelly, and Manela, 2017). However, it is only available beginning in 1970. From our long sample of *Wall Street Journal* text, we estimate an *icr* series back to 1927 to investigate the interaction between *icr* and asset prices in the historical period.

We find that HDMR gives substantially improved out-of-sample predictions of *icr* compared to DMR, which indicates that modeling the selection decision helps with forecasting in this context. HDMR also outperforms support vector regression (Vapnik, 2000), which Manela and Moreira (2017) use for text-based backcasting of the VIX stock market volatility index. Unlike support vector regression, both DMR and HDMR can concentrate on individual variables that behave differently from word counts (i.e. non-text control variables), but are useful for prediction. We find that the out-of-sample advantage of HDMR over DMR increases with the sparsity of the text. As we omit more infrequent words from the dictionary—a common approach to ad hoc prefiltering of phrases—the document term matrix becomes denser and the text becomes better described by a selection-free DMR model. At the same time, however, more stringent phrase filters lead to a large deterioration in prediction accuracy from either method. Evidently, less filtering (allowing for more phrases) makes for a better prediction model, and also magnifies the benefits of accounting for text selection via the hurdle model.

The news-implied intermediary capital ratio series provides for more powerful tests that support central predictions of intermediary asset pricing theory. The results show that times when intermediaries are highly capitalized are "good times" when these marginal investors demand a relatively low premium to hold risky assets. Our findings imply that news text of *The Wall Street Journal* provides a strong signal about stock market risk premia, over and above common financial market covariates such as the price-dividend ratio and stock volatility.

In our second application, we ask whether HDMR is able to extract information from text of *The Wall Street Journal* that is useful for forecasting U.S. macroeconomic indicators (beyond that of a benchmark principal components method suggested by Stock and Watson, 2012). We find that the information from text significantly improves over benchmark forecasts for key macroeconomic indicators such as nonfarm payroll employment and housing starts, and that the advantages of text-based information increase as we expand the dimensionality (and along with it, the sparsity) of the text. In a related analysis, we show that text of *The Wall Street Journal* is valuable for nowcasting macroeconomic series, which are released at a lower frequency and with a delay relative to news text.

This paper contributes to a rapidly growing literature that incorporates insights from machine learning into econometrics. Recent work applies prediction algorithms in policy analysis (Kleinberg, Ludwig, Mullainathan, and Obermeyer, 2015; Bajari, Nekipelov, Ryan, and Yang, 2015; Athey, 2017; Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan, 2017; Ludwig, Mullainathan, and Spiess, 2017; Einav, Finkelstein, Mullainathan, and Obermeyer, 2018) and considers parameter uncertainty and causal inference in high dimensional settings (Belloni, Chen, Chernozhukov, and Hansen, 2012; Belloni, Chernozhukov, and Hansen, 2014; Belloni, Chernozhukov, Fernández-Val, and Hansen, 2017; Mullainathan and Spiess, 2017; Athey, Imbens, Pham, and Wager, 2017; Athey, Tibshirani, and Wager, 2019). Our model allows economists interested in analyzing counts data like text, to model selection in the process that generates their data in a flexible, robust and scalable way. We show that the gains from adding structure that is well-grounded in economic theory can substantially improve prediction in high dimensional applications.[5]

[5]See Athey (2018) for a recent survey and Hoberg and Phillips (2016); Hanley and Hoberg (2019); Jiang, Lee, Martin, and Zhou (2019) for recent text analysis applications. We also provide new tools for summarizing high dimensional coverage to a literature studying the media in economics (Gentzkow and Shapiro, 2006; Qin, Strömberg,

Our technology is publicly available via the `HurdleDMR` package for `Julia`, which can be called from many other programming languages like `Python` and `R`. The package allows for computationally efficient distributed estimation of the multiple hurdles over parallel processes, generating sufficient reduction projections, and inverse regressions with selected text. It allows for elastic net type convex combinations of L1 (Lasso) and L2 (Ridge) regularization as in `glmnet` (Friedman, Hastie, and Tibshirani, 2010), and for concave regularization paths as in `gamlr` (Taddy, 2017).

We start Section 2 by presenting the intensive margin of our text selection model. Here we follow the distributed multinomial regression (DMR) model of Taddy (2015). Section 2.1 introduces our main contribution, a model for the extensive margin, which we refer to as Hurdle DMR. Section 2.2 shows how regularization allows our methodology to handle a feature space many times larger the number of observations. Section 2.3 describes how to recover low dimension text indices that track variables of interest. Section 2.4 shows how to use these indices for prediction, and establishes that they are sufficient statistics for the information content of non-text covariates. We end the paper with two applications of our methodology using text of *The Wall Street Journal*. In Section 3 we backcast the intermediary capital ratio and analyze its historical asset pricing properties. In Section 4 we forecast and nowcast a variety of measures of macroeconomic activity.

2 A model for text selection

Let $\boldsymbol{c}_i$ be a vector of counts in d categories for observation i, summing to $m_i = \sum_j c_{ij}$, and let $\boldsymbol{v}_i$ be a p_v-vector of covariates associated with each observation $i = 1 \ldots n$. In a text application, c_{ij} are counts of word or phrase (n-gram) j in document i with attributes $\boldsymbol{v}_i$.[6] An econometrician confronted with modeling counts data may first consider using a multinomial logistic regression:

$$\mathrm{p}\left(\boldsymbol{c}_i | \boldsymbol{v}_i, m_i\right) = MN\left(\boldsymbol{c}_i; \boldsymbol{q}_i, m_i\right) \text{ for } i = 1 \ldots n, \tag{1}$$

$$q_{ij} = \frac{e^{\eta_{ij}}}{\sum_{k=1}^{d} e^{\eta_{ik}}} \text{ for } j = 1 \ldots d, \tag{2}$$

$$\eta_{ij} = \alpha_j + \boldsymbol{v}_i' \boldsymbol{\varphi}_j. \tag{3}$$

and Wu, 2018; Durante and Zhuravskaya, 2018) and finance (Antweiler and Frank, 2004; Tetlock, 2007; Fang and Peress, 2009; Engelberg and Parsons, 2010; García, 2013).

[6]For example, in the context of a finance application, v_i could be the stock market response to a earnings release or the subsequent stock market volatility.

When the number of categories d is very large, as is the case in many natural language processing applications, estimating the parameters of the multinomial, $\boldsymbol{\alpha} = [\alpha_j]$ and $\boldsymbol{\varphi} = [\varphi_{ij}]$, is computationally prohibitive.[7] Equation (2), which makes sure that word probabilities q_{ij} add up to one, is the main barrier to parallelization across categories because every parameter change must be communicated to all other category estimators.

It is well known that the multinomial can be decomposed into independent poissons conditional on the intensities $e^{\eta_{ij}}$, scaled by a poisson for total word counts m_i,

$$MN\left(\boldsymbol{c}_i; \boldsymbol{q}_i, m_i\right) = \frac{\prod_j Po\left(c_{ij}; e^{\eta_{ij}}\right)}{Po\left(m_i; \sum_{j=1}^{d} e^{\eta_{ij}}\right)}. \tag{4}$$

Motivated by this decomposition, Taddy (2015) develops the distributed multinomial regression (DMR), a parallel (independent) poisson plug-in approximation to the multinomial,

$$\mathrm{p}\left(\boldsymbol{c}_i|\boldsymbol{v}_i, m_i\right) = MN\left(\boldsymbol{c}_i; \boldsymbol{q}_i, m_i\right) \approx \prod_j Po\left(c_{ij}; m_i e^{\eta_{ij}}\right). \tag{5}$$

The parameters for each category j can then be estimated independently with negative log likelihood

$$l\left(\alpha_j, \boldsymbol{\varphi}_j|\boldsymbol{c}_j, \boldsymbol{v}\right) = \sum_{i=1}^{n}\left[m_i e^{\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j} - c_{ij}\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right)\right]. \tag{6}$$

Intuitively, each independent poisson intensity $\lambda_{ij} = m_i e^{\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j}$ is shifted to account for the fact that all words are more likely to appear in longer (high m_i) documents. Approximation (5) removes the communication bottleneck of recomputing $\sum_{j=1}^{d} e^{\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j}$ and allows for fast and scalable distributed estimation.

Taddy (2013, 2015) uses the DMR to estimate a low dimensional sufficient reduction projection

$$\boldsymbol{z}_i = \hat{\boldsymbol{\varphi}}'\boldsymbol{c}_i$$

and shows that $\boldsymbol{v}_i$ is independent of $\boldsymbol{c}_i$ conditional on $\boldsymbol{z}_i$. This means that, within this model, $\boldsymbol{z}_i$ is a sufficient statistic that summarizes all of the content that the text has for predicting the

[7] For example, our application in Section A.2 has a vocabulary of more than five hundred thousand phrases, i.e. $d > 500,000$.

covariates$\boldsymbol{v}_i$ (or its individual elements). For example, suppose v_{iy} is the first element of $\boldsymbol{v}_i$, which is available in a subsample, but needs to be predicted for other subsamples. The first step would be to run a multinomial inverse regression of word counts on the covariates $\boldsymbol{v}$ in the training sample to estimate $\hat{\boldsymbol{\varphi}}$. Second, estimate a forward regression (linear or higher order)

$$\mathbb{E}\left[v_{iy}\right] = \beta_0 + \left[z_{iy}, \boldsymbol{v}_{i,-y}, m_i\right]' \boldsymbol{\beta} \tag{7}$$

where $z_{iy} = \sum_j \hat{\varphi}_{jy} c_{ij}$ is the projection of phrase counts in the direction of v_{iy}. Finally, the forward regression can be used to predict v_{iy} using text and the remaining covariates $\boldsymbol{v}_{i,-y}$.

2.1 Hurdle distributed multinomial regression

In many cases, and specifically in text applications, the poisson is a poor description of word counts c_{ij}. For example, Figure 1 shows the mean histogram (across documents) for the corpus we use below, which consists of 10,000 two-word phrases (bigrams) appearing in the title and lead paragraph of front page *Wall Street Journal* articles. The left panel shows a substantial mass point at zero that is hard to reconcile with a poisson. The panel on the right shows that if we restrict attention to positive counts, a (truncated) poisson is a reasonable approximation for the data. In our experience, this sparsity is a feature of many text samples, which is why most text analysis software packages use sparse matrices to store word counts efficiently. As alluded to above, the economics of natural language selection provides many reasons for this sparsity. Furthermore, the economics of content selection suggests that relationship between covariates and text will also be better captured by allowing for a separate extensive margin effect– e.g. the decision to start writing about a topic is likely to have more information than writing more about a topic.

To model text selection, we replace the independent poissons with a two part hurdle model for counts c_{ij}, which we label the *hurdle distributed multinomial regression (HDMR)*:

$$h_{ij}^* = \gamma_i + \kappa_j + \boldsymbol{w}_i' \boldsymbol{\delta}_j + \upsilon_{ij}, \tag{8}$$

$$h_{ij} = \mathbf{1}\left(h_{ij}^* > 0\right), \tag{9}$$

$$c_{ij}^* = \lambda\left(\mu_i + \alpha_j + \boldsymbol{v}_i' \boldsymbol{\varphi}_j\right) + \varepsilon_{ij}, \tag{10}$$

$$c_{ij} = \left(1 + c_{ij}^*\right) h_{ij}. \tag{11}$$

The first two equations describe the choice to include ($h_{ij} = 1$) or exclude ($h_{ij} = 0$) word j in document i, often referred to as the model for zeros or participation. This choice depends on observable covariates $\boldsymbol{w}_i \in \mathbb{R}^{p_w}$ and an unobservable υ_{ij}. Equation (10) is the model for word repetition given inclusion in the document, which can depend on the same or other covariates $\boldsymbol{v}_i \in \mathbb{R}^{p_v}$ and an unobservable ε_{ij}. The last equation says that we only observe positive counts for included words.[8]

Let Π_0 denote the discrete density for zeros

$$\mathrm{p}\left(h_{ij} = 0 | \boldsymbol{w}_i\right) = \Pi_0\left(\gamma_i + \kappa_j + \boldsymbol{w}_i' \boldsymbol{\delta}_j\right).$$

Natural choices for Π_0 are the probit and logit binary choice models. Let P^+ denote the model for word repetition, so that conditional on inclusion,

$$\mathrm{p}\left(c_{ij}^* | \boldsymbol{v}_i, h_{ij} = 1\right) = P^+\left(c_{ij}^*; \lambda\left(\mu_i + \alpha_j + \boldsymbol{v}_i' \boldsymbol{\varphi}_j\right)\right).$$

Natural choices for P^+ are the poisson and the negative binomial. Combining terms, the joint density is

$$\mathrm{p}\left(c_{ij} | \boldsymbol{v}_i, \boldsymbol{w}_i\right) = \left[\Pi_0\left(\gamma_i + \kappa_j + \boldsymbol{w}_i' \boldsymbol{\delta}_j\right)\right]^{1 - h_{ij}} \left\{\left[1 - \Pi_0\left(\gamma_i + \kappa_j + \boldsymbol{w}_i' \boldsymbol{\delta}_j\right)\right] P^+\left(c_{ij} - 1; \lambda\left(\mu_i + \alpha_j + \boldsymbol{v}_i' \boldsymbol{\varphi}_j\right)\right)\right\}^{h_{ij}} \tag{12}$$

[8] Our two part model is simpler and faster to estimate than Mullahy (1986)'s hurdle, which models positive counts as drawn from a truncated poisson, as opposed to a regular poisson for *counts in excess of one* (repetitions). In a previous draft we used the truncated poisson and found very similar results.

The negative log likelihood takes a convenient form

$$l(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\varphi},\boldsymbol{\gamma},\boldsymbol{\kappa},\boldsymbol{\delta}|\boldsymbol{c},\boldsymbol{v},\boldsymbol{w}) = \sum_{j=1}^{d} l(\mu_i,\alpha_j,\varphi_j,\gamma_i,\kappa_j,\delta_j|\boldsymbol{c}_j,\boldsymbol{v},\boldsymbol{w}), \tag{13}$$

$$l(\mu_i,\alpha_j,\varphi_j,\gamma_i,\kappa_j,\delta_j|\boldsymbol{c}_j,\boldsymbol{v},\boldsymbol{w}) = l^0(\gamma_i,\kappa_j,\boldsymbol{\delta}_j|\boldsymbol{h}_j,\boldsymbol{w}) + l^+\left(\mu_i,\alpha_j,\boldsymbol{\varphi}_j|\boldsymbol{c}_j,\boldsymbol{v}\right), \tag{14}$$

$$l^0(\gamma_i,\kappa_j,\boldsymbol{\delta}_j|\boldsymbol{h}_j,\boldsymbol{w}) = -\sum_{i|h_{ij}=0}^{n} \log\Pi_0(\gamma_i+\kappa_j+\boldsymbol{w}_i'\boldsymbol{\delta}_j) - \sum_{i|h_{ij}=1}^{n} \log\left[1-\Pi_0(\gamma_i+\kappa_j+\boldsymbol{w}_i'\boldsymbol{\delta}_j)\right], \tag{15}$$

$$l^+\left(\mu_i,\alpha_j,\boldsymbol{\varphi}_j|\boldsymbol{c}_j,\boldsymbol{v}\right) = -\sum_{i|h_{ij}=1}^{n} \log P^+\left(c_{ij}-1;\lambda\left(\mu_i+\alpha_j+\boldsymbol{v}_i'\boldsymbol{\varphi}_j\right)\right). \tag{16}$$

Note that the coefficients γ_i and μ_i introduce dependence across the log likelihood of different words j. To allow for separation across words, and parallelization of the estimation stage, we adapt the argument in Taddy (2015) and use plug in estimators$\hat{\gamma}_i$ and $\hat{\mu}_i$ that approximate the maximum likelihood estimators under certain conditions. We discuss these plug in estimators in Section 2.3 and in Appendix B.

A useful feature of the hurdle is that exclusion ($h_{ij}=0$) is the only source of zero counts. As a result, it decomposes as in (14) into two parts that can be estimated independently, which facilitates further parallelization.[9] Specifically, the parameters that govern inclusion (κ_j, $\boldsymbol{\delta}_j$) only depend on word j indicators $\boldsymbol{h}_j$ and on the covariates $\boldsymbol{w}$, whereas the parameters that govern repetition (α_j, $\boldsymbol{\varphi}_j$) only depend on word repetition $\boldsymbol{c}_j-1>0$ and the covariates $\boldsymbol{v}$ and can be estimated separately in the subsample of with word repetition.

HDMR therefore allows one to estimate text selection in Big Data applications of previously impossible scale, by distributing computation across categories and across the two parts of the selection model.

2.2 Regularization

In many machine learning applications, the feature space (words) is much larger than the number of observations. In such cases, regularization by penalizing nonzero or large $\boldsymbol{\varphi}$ and $\boldsymbol{\delta}$ coefficients is key to avoid overfit. Our results use L_1 regularization separately for each category j and for each

[9] Zero inflation models are alternative approaches that allow for latent $c_{ij}^*=0$, in which case zero count observations could result either from exclusion or from inclusion of zero counts. While this distinction is philosophically interesting, the hurdle is more tractable and faster to estimate.

of the two parts of the hurdle

$$\hat{\kappa}_j, \hat{\boldsymbol{\delta}}_j = \underset{\kappa_j, \boldsymbol{\delta}_j}{\arg\min}\, l^0 \left(\hat{\gamma}_i, \kappa_j, \boldsymbol{\delta}_j | \boldsymbol{h}_j, \boldsymbol{w}\right) + n\lambda^0 \sum_{k=1}^{p_w} |\delta_{jk}| \quad \text{where } \lambda^0 \geq 0, \tag{17}$$

$$\hat{\alpha}_j, \hat{\boldsymbol{\varphi}}_j = \underset{\alpha_j, \boldsymbol{\varphi}_j}{\arg\min}\, l^+ \left(\hat{\mu}_i, \alpha_j, \boldsymbol{\varphi}_j | \boldsymbol{c}_j, \boldsymbol{v}\right) + n^+\lambda^+ \sum_{k=1}^{p_v} |\varphi_{jk}| \quad \text{where } \lambda^+ \geq 0. \tag{18}$$

The penalties λ^0 and λ^+ shrink the loadings toward zero, and because of the Lasso-type L_1 penalties, result in many zero loadings (Tibshirani, 1996).[10] Because the model for positive counts only depends on documents i that include word j, the penalty is normalized by the number of such documents $n^+ \equiv \sum_{i=1}^{n} h_{ij}$. Fast coordinate descent algorithms for these minimization problems have been proposed by Friedman, Hastie, and Tibshirani (2010), which trace out regularization paths of solutions, one for each of a grid of $\lambda' s$, for the class of generalized linear models (GLM, McCullagh and Nelder, 1989). We follow Taddy (2017) in selecting the model that minimizes a corrected AIC, though in relatively modest applications one could use cross validation to select the optimal penalty. To apply the coordinate descent algorithms developed by Friedman, Hastie, and Tibshirani (2010) to our selection model, we frame its two parts as GLMs: a binomial-logit for the inclusion part and a poisson-log for the repetition part (McCullagh and Nelder, 1989).

2.3 Sufficient reduction projections

For simplicity, in what follows we focus on the case where h_{ij} is binomial (bernoulli) distributed

$$\mathrm{p}\left(h_{ij}|\boldsymbol{w}_i\right) = \left[\Pi_0\left(\hat{\gamma}_i + \kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)\right]^{1-h_{ij}} \left[1 - \Pi_0\left(\hat{\gamma}_i + \kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)\right]^{h_{ij}} \tag{19}$$

with a logit link,

$$\log\left(\left(1-\Pi_0\right)/\Pi_0\right) = \kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j + \hat{\gamma}_i. \tag{20}$$

In this case we can show that for a large enough number of categories d, the MLE estimator for γ_i converges to $log\left(\frac{d_i}{d-d_i}\right)$, where $d_i = \sum_{j=1}^{d} h_{ij}$ is the number of categories used in document i. We therefore use $\hat{\gamma}_i = log\left(\frac{d_i}{d-d_i}\right)$ as our plug in estimator. See Appendix B for details.

[10] We focus on Lasso penalties here to simplify the exposition. Our `HurdleDMR` package allows for more general elastic net-type regularization as in `glmnet` (Friedman, Hastie, and Tibshirani, 2010), and for concave regularization paths as in `gamlr` (Taddy, 2017).

The distribution of word repetitions c^*_{ij} is assumed to be poisson,

$$P^+\left(c^*_{ij};\lambda_{ij}\right) = Po\left(c_{ij}-1;\lambda_{ij}\right) = \lambda_{ij}^{c_{ij}-1}e^{-\lambda_{ij}}, \tag{21}$$

with log intensity

$$\log\lambda_{ij} = \alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j + \hat{\mu}_i. \tag{22}$$

Here we follow Taddy (2015) and use $\hat{\mu}_i = \log(m_i - d_i)$ as our plug in estimator. The only difference from Taddy (2015) is that here we are modeling word repetition so the estimator is $\log(m_i - d_i)$ instead of $\log(m_i)$. See Appendix B for details.

In the Appendix, we show how to estimate both parts of the hurdle as a GLM, which allows us to use fast coordinate descent algorithms for regularized estimation as in `glmnet`.

We next show that under these functional forms, the entire empirical content of the text that is useful for predicting a variable in $\boldsymbol{w}$ or $\boldsymbol{v}$, is summarized by two low dimension sufficient statistics.

Result 1. *Assuming a binomial-logit model for inclusion and a poisson-log model for word repetition, the projection $\boldsymbol{\delta h}_i$ is a sufficient statistic for $\boldsymbol{w}_i$ and the projection $\boldsymbol{\varphi c}_i$ is a sufficient statistic for $\boldsymbol{v}_i$, conditional on total counts m_i and vocabulary d_i. Specifically,*

$$\boldsymbol{v}_i, \boldsymbol{w}_i \perp\!\!\!\perp \boldsymbol{h}_i, \boldsymbol{c}_i | \boldsymbol{\delta h}_i, \boldsymbol{\varphi c}_i, m_i, d_i.$$

Proof. That $\boldsymbol{\varphi}(\boldsymbol{c}_i - 1)$ is a sufficient statistic for $\boldsymbol{v}_i$ follows immediately from the DMR case discussed in Section 2. To establish sufficiency of $\boldsymbol{\delta h}_i$, note that the likelihood for counts $\boldsymbol{c}_i$ given observed covariates $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$ can be factored into

$$\mathrm{p}\left(\boldsymbol{c}_i|\boldsymbol{v}_i,\boldsymbol{w}_i\right) = \mathrm{p}\left(\boldsymbol{h}_i \circ \boldsymbol{c}_i|\boldsymbol{v}_i,\boldsymbol{w}_i\right) = \phi(c_i)\psi(h_i)a(w_i,d_i)b(v_i,m_i,d_i)\exp(w_i'\delta h_i + \varphi v_i(c_i - h_i)), \tag{23}$$

where

$$\psi\left(\boldsymbol{h}_i\right) = \prod_{j=1}^{d} e^{\left(log\left(\frac{d_i}{d-d_i}\right)+\kappa_j\right)h_{ij}},$$

$$\phi\left(\boldsymbol{c}_i\right) = \prod_{j|h_{ij}=1}^{d} \exp\left(\left(c_{ij}-1\right)\left(\log\left(m_i-d_i\right)+\alpha_j\right) - log\left(\left(c_{ij}-1\right)!\right)\right),$$

$$a\left(\boldsymbol{w}_i, m_i\right) = \prod_{j=1}^{d} \frac{1}{1 + e^{log\left(\frac{d_i}{d-d_i}\right) + \kappa_j + w_i'\delta_j}},$$

$$b\left(\boldsymbol{v}_i, m_i\right) = \prod_{j|h_{ij}=1}^{d} \exp\left(-(m_i - d_i)e^{\alpha_j + \varphi_j' v_i}\right),$$

and we use the fact that the Hadamard product $\boldsymbol{h}_i \circ \boldsymbol{c}_i$ is equivalent to $\boldsymbol{c}_i$ here. Hence, the usual sufficiency factorization (e.g., Schervish, 1995, Theorem 2.21) applies yielding the stated result. See Appendix C for additional details. □

Result 1 means that once we estimate the HDMR parameters, we can reduce the high-dimension (d) text into low-dimension ($p_v + p_w$) sentiment scores from the text in the direction of the covariates in $\boldsymbol{v}$ or $\boldsymbol{w}$. The projections provide useful summaries of the text, which can be plotted or used as a dimensionality reduction first-step into a more elaborate analysis. As in Taddy (2015), the sufficient statistics $\boldsymbol{\delta h}_i$ and $\boldsymbol{\varphi c}_i$ rely on population parameters, but in practice we use plug-in estimators $\hat{\boldsymbol{\delta}}\boldsymbol{h}_i$ and $\hat{\boldsymbol{\varphi}}\boldsymbol{c}_i$. Whether these provide useful approximations is a setting-dependent empirical matter.

2.4 Inverse regression for prediction

The end goal of many machine learning or natural language processing applications is out-of-sample prediction. Result 1 provides a guide to supervised learning from text via an inverse regression of the text on the target variable and other covariates Taddy (2013). The parameters from the HDMR inverse regression in the training sample are used to form a bivariate sufficient reduction projection of the text. A forward regression (still in the training sample) of the target variable on these projections plus the other covariates is then used to construct the predictor.

More concretely, suppose the target variable $v_{iy} = w_{iy}$ is an element of both $\boldsymbol{v}_i$ and $\boldsymbol{w}_i$. We first construct two univariate sufficient reduction projections $z_{iy}^0 = \boldsymbol{\delta}_y \boldsymbol{h}_{iy}$ and $z_{iy}^+ = \boldsymbol{\varphi}_y \boldsymbol{c}_{iy}$. Because the estimated loadings $\boldsymbol{\delta}_y$ and $\boldsymbol{\varphi}_y$ are partial effects, controlling for the other covariates, $\boldsymbol{w}_{i,-y}$ and $\boldsymbol{v}_{i,-y}$, the projections z_{iy}^0 and z_{iy}^+ correspond to partial associations as well. Conditional on the parameters, z_{iy}^0 contains all the information that is useful for predicting v_{iy} from the selection of words used in the text (the extensive margin). Similarly, z_{iy}^+ contains the incremental predictive information in repeating words within document i (the intensive margin). Intuitively, HDMR can learn separately from both the extensive and intensive margins, and use them for more efficient prediction. We would then estimate a forward regression (linear or higher order)

$$\mathbb{E}\left[v_{iy}\right] = \beta_0 + \left[z^0_{iy}, z^+_{iy}, \boldsymbol{w}_{i,-y}, \boldsymbol{v}_{i,-y}, m_i, d_i\right]' \boldsymbol{\beta} \tag{24}$$

which can be used to predict v_{iy} using text and the remaining covariates $\boldsymbol{w}_{i,-y}$ and $\boldsymbol{v}_{i,-y}$. In the case that the target variable is only an element of $\boldsymbol{w}$ ($\boldsymbol{v}$), one would only use z^0_{iy} (z^+_{iy}) in the forward regression.

3 Application I: Backcasting the intermediary capital ratio

A rapidly growing body of work finds empirical support for intermediary asset pricing theories (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014; Adrian, Etula, and Muir, 2014; Haddad and Muir, 2018; Baron and Muir, 2018; He and Krishnamurthy, 2018; Koijen and Yogo, 2019). In particular He, Kelly, and Manela (2017) find that a simple two-factor model that includes the excess stock market return and the aggregate capital ratio of major financial intermediaries—primary dealers—can explain cross-sectional variation in expected returns across a wide array of asset classes. They also present preliminary results on return predictability (time-series regressions), but their conclusions are limited by a relatively short time-series that starts in 1970. Prior to 1970, most primary dealers were private, which precludes a calculation of their capital ratio.

We conjecture that as a publication catering to investors, text that appears on the front page of the *Wall Street Journal* would be informative about the aggregate state of the intermediary sector. Dire language on financial intermediaries' failure is used to cover unfolding crises like the financial crisis of 2008, the LTCM liquidity crisis following Russia's default in 1998, and the failure of important dealers like Drexel Burnham Lambert in 1990.

3.1 Data

Our text data includes all titles and lead paragraphs that appear on the front page of the *Wall Street Journal* from July 1926 to February 2016. We include the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming. We aggregate the data to the monthly frequency so that $\boldsymbol{c}_t$ are phrase counts observed during month t.

Figure 1 shows the mean histogram for phrase counts in this sample. The left panel shows that the entire range is highly sparse (has many zeros). The right panel omits zero counts, and shows that a truncated (at zero) poisson distribution is a reasonable approximation for the positive range of counts.

We match this data with the monthly intermediary capital ratio icr_t of He, Kelly, and Manela (2017).[11] This ratio is our prediction target and is therefore the first element of both covariate vectors $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$. We additionally include in both, two natural covariates that are likely to be correlated with the icr_t: the log price dividend ratio (pd_t) from CRSP, and the realized variance of financial stocks ($rvfin$) over the same month and the prior one. Table 1 reports summary statistics for these variables.

3.2 Newspaper coverage choice

Our selection model is parametrically identified and therefore technically does not require that different variables be used in the inclusion and repetition equation. However, Heckman (1979) selection models are known to be nonparametrically identified if a continuous variable enters the selection equation but can be excluded from the second equation (Gallant and Nychka, 1987; Heckman and MaCurdy, 1986). Proving such a result in our setting can be useful, but left for future work. Motivated by their insight, we seek an instrument for word inclusion.

Boydstun (2013) suggests that prior attention to an issue may influence its coverage by the press. The idea is that once fixed costs such as journalist travel and familiarization with an issue have been incurred, a marginal article is easier to produce. To capture prior attention to the financial sector, we add prior year realized variance of financial stocks ($rvfin_{t-13\to t-1}$) as an explanatory variable in the model for inclusion alone, excluding it from the repetition equation. This choice assumes that after conditioning on the intermediary capital ratio, the price-dividend ratio, current and lagged monthly $rvfin$, and on phrase j being included in the *Journal*, the number of times

[11] The icr is available at http://apps.olin.wustl.edu/faculty/manela/data.html

this phrase is repeated does not depend on the prior year's volatility of financial stocks.[12]

3.3 Sparsity and out-of-sample fit

A key choice in the preprocessing stage of many text analyses, is to omit words or phrases that rarely appear in the sample. For example, we may keep the X most frequent phrases. From the vantage point of our selection model, this choice is important. If the "cleansed" word counts matrix $\boldsymbol{c}$ is highly dense because phrases that often do not appear in the text are excluded from the analysis, then the benefit of modeling the extensive margin is likely to be low. Therefore, we assess the improvement in out-of-sample fit as a function of the number of most frequent phrases kept in the sample.

Figure 2 compares the out-of-sample root mean squared error from a 10-fold cross validation exercise. It compares HDMR with DMR, which is provided with the same covariates and text, and with a linear regression of the target on the same covariates but without the text. Both DMR and HDMR improve considerably over the No Text benchmark and reduce the error by 46 percent (from 1.3 to 0.7 percentage points). We can see that when only a few hundred words are considered in vocabulary, both DMR and HDMR generate similar improvements, but as rarer phrases are included in the vocabulary, selection plays a bigger role, and the benefit from using HDMR increases. The advantage is hump-shaped, and eventually, as rarely-used phrases enter, the out-of-sample fit of both text models suffers.

Because our prediction exercise involves using data in one time period to predict out-of-sample in a different time, cross-validation with random folds may be misleading when both the text and target variable are persistent. For example, if the model relies heavily on the fact that the phrase "subprime mortgage" appears often in the period around the 2008 financial crisis when intermediary capital was low, but not in earlier parts of the sample, then random cross-validation, which would likely include observations around the same period in the test subsamples, may give an overly optimistic measure of out-of-sample fit.

Figure 3, therefore, uses pseudo out-of-sample prediction, starting with the latter half of the sample and rolling back with wider training subsamples, predicting one observation at a time. Even

[12] We have experimented with business news pressure (Manela, 2014), a variant of television news pressure (Eisensee and Stromberg, 2007), that is available for a shorter sample starting 1967, and found similar out-of-sample fit improvements.

though this validation approach results in somewhat wider confidence intervals, the results are quite similar to random cross-validation. Table 2 focuses on the optimal model by cross-validation, the one that uses the 10,000 most frequent phrases. We find that regardless of the validation method, HDMR provides a significant reduction in out-of-sample root-mean-squared-error.

3.4 News-implied intermediary capital ratio, 1927–2016

Having established that our model produces good out-of-sample fit, we use it to backcast the intermediary capital ratio back to the June 1927, the first month when a full year of financial volatility is available. Figure 4a shows that the intermediary capital ratio predicted by HDMR closely follows the actual one in the period when the latter is available, 1970–2016. Financial variance and the price-dividend ratio, which alone can explain much of the variation in the icr, provide a back-bone for the predictor, as can be seen from the No Text benchmark. DMR, HDMR, and SVR, all use these covariates plus the text to improve prediction, but generate somewhat different time-series. For example, the HDMR predicted values appear lower than those of DMR and feature more negative spikes in the capital ratio. The SVR predictor makes clear that it does not find the variation in $rvfin$ and pd important.

Figure 4b zooms in on the Great Depression period. With the exception of the SVR predicted series, all predictors indicate that financial intermediaries were substantially undercapitalized between 1929 and 1939. Figure 4c shows a similar pattern, for the more recent Great Recession period, with all predictors showing a sharp fall in the aggregate capital ratio of financial intermediaries, that matches the actual icr's behavior. Interestingly, financial variables without the text overstate the recovery of the icr starting in 2010. Compared with HDMR, DMR understates the icr's recovery between 2012 and 2013. To better understand the source of this difference, we next focus on individual phrases and their importance.

3.5 Which phrases are pivotal for out-of-sample fit?

To better understand the improvements in out-of-sample fit that HDMR generates, we report in Table 3 the phrases whose removal from the corpus causes the largest deterioration in out-of-sample root-mean-squared-error.[13] We report results separately for the two validation methods discussed

[13] Gentzkow, Shapiro, and Taddy (2019) identify partisan phrases with a similar approach.

earlier in Section 3.3 because our intuition is that time-series persistence can lead to overstatement of out-of-sample fit using random cross-validation.

A related concern is that language changes over time, so the phrases associated with movements in the *icr* in the fitted sample, are not useful its prediction in the earlier, 1926–1969 period. For example, `subprime mortgage` is unlikely to relate to appear in the *Wall Street Journal* prior to the 2007–2009 financial crisis. But the important question for our purposes is whether HDMR can avoid overfitting to such phrases. Manela and Moreira (2017, Section 2.3) analyze this possibility in detail for the same corpus and a different financial variable as target. They find that while language does change over time, the deterioration it induces in out-of-sample fit is quite modest, even when backcasting farther to 1890. Examining the list of pivotal phrases can shed light on this issue.

Table 3, Panel (a) shows that some of the predictive ability of the text comes phrases, such as the the positively associated `jobless marri` capture fairly robust economic fundamentals. Other phrases such as `barack obama`, a US president elected at the peak of the 2008 financial crisis, show up as negatively correlated with the *icr*, even though they are unlikely to be useful for its prediction before 2008. Interestingly, his main opponent during the 2012 election, `mitt romney,` is negatively associated too, suggesting that political uncertainty as opposed to specific policies may be the culprit (Pastor and Veronesi, 2012; Baker, Bloom, and Davis, 2016; Manela and Moreira, 2017; Hassan, Hollander, van Lent, and Tahoun, 2017).

Panel (b) uses pseudo out-of-sample rolling validation instead, and shows as conjectured, greater focus on robust fundamentals that are likely to be relevant over the entire sample. Pivotal phrases have to do with government policy (`tax report`, `washington wire`) and economic conditions (`busi bulletin`, `labor letter`). We therefore find this validation approach more likely to approximate true out-of-sample fit.

3.6 Focusing on a single phrase for intuition

For a better intuitive understanding of how inverse regression loadings translate into forward regression prediction, we next focus on a single phrase, `financi crisi`. We expect front page reports of financial crises to be a negative signal about the capital ratio of the intermediary sector.

The backward hurdle regression estimates in the first two columns of Table 4a show that the *icr* is indeed negatively correlated with repeated mentions of `financi crisi`, but also that the mere inclusion of this phrase on the front page is a strong negative signal, conditional on the price-dividend ratio and financial volatility. The positive coefficient on $rvfin_{t-13\to t-1}$ means that above average prior year financial volatility is followed by higher financial crisis coverage on the front page. Specifically, a one standard deviation in this instrument relative to its mean implies a 22% increase in the log odds ratio for a financial crisis inclusion (from 0.17 to 0.21). The last column shows that a poisson regression (DMR) treats inclusion and repetition as a single object, and assigns a relatively low weight to this instrument compared to contemporaneous financial volatility (both variables are in annual variance units).

These coefficients are used to construct the two sufficient reduction projections, $z^0_{ty} = \boldsymbol{\delta}_y \boldsymbol{h}_{ty}$ and $z^+_{ty} = \boldsymbol{\varphi}_y \boldsymbol{c}_{ty}$, and plugged into a forward regression of the *icr* on the these and the remaining covariates, as described in Section 2.4:

$$y_t = b_0 + b_z z^+_{ty} + b_s z^0_{ty} + b_v \boldsymbol{v}_{t,-y} + b_m m_t + b_d d_t + \varepsilon_t.$$

The contribution of a single phrase j to the predicted value is therefore

$$\hat{y}_{tj} = b_z \varphi_{jy} \left(c_{tj}/m_t\right) + b_s \delta_{jy} \left(h_{tj}/m_t\right).$$

Table 4b reports the forward regression coefficients' products with those of the backward regression, $b_z\varphi_{jy}$ and $b_s\delta_{jy}$ for HDMR, and contrasts it with the corresponding single coefficient product of DMR. We can see that much of the contribution of `financi crisi` to the predicted value in HDMR comes from the extensive margin. A different way to see this is by looking at the time series $\hat{y}_{tj}$, which appears in Figure 5. A single mention of financial crises is all it takes for HDMR to predict a lower intermediary capital, with repeated mentions having a much lower effect.

3.7 Time-varying risk premia and the intermediary capital ratio

A central prediction of the intermediary asset pricing model (He and Krishnamurthy, 2012, 2013) is that times when intermediaries are highly capitalized are "good times," when these marginal investors demand a relatively low risk premium to hold investment assets. Preliminary such time-series predictability regression reported in He, Kelly, and Manela (2017) support this prediction, but the short time-series used there limits the power of these tests.

The news-implied intermediary capital ratio allows us to test this prediction in a larger sample that goes back to 1927. Return predictability tests are reported in Table 5 for the monthly, quarterly and annual horizons. Because such regressions use overlapping observations, we use the standard Hodrick (1992) correction to the standard errors. For each horizon, the first column show that restricting attention to the actual *icr*, which is available only starting 1970, yields a negative but statistically insignificant coefficient. By contrast, we find that the news-implied $\widehat{icr}$, which is available over a much longer period, is significantly negative both over the postwar sample and the full sample. The point estimates are fairly consistent across horizons and subsamples and imply that a one standard deviation increase in the *icr* predicts a 4–5 percentage points lower market premium.

To understand better whether the predictive ability comes from covariates that are known predictors like price-dividend ratio or from the text, we regress future stock market excess returns at various horizons on lagged sufficient reduction projections z_{t-1}^{0}, z_{t-1}^{+}, and the covariates. We find that the inclusion projection z_{t-1}^{0} is a strong predictor of future market returns in the postwar sample, over and above the price-dividend ratio, and that the repetition projection z_{t-1}^{+} is statistically significant only in the full sample.

The results imply that there is a set of phrases whose inclusion on the front page of the *Journal* provides a strong signal about stock market risk premia, over and above the valuation ratio (pd). HDMR provides an efficient way to identify these phrases and their relative weights in a data driven approach while avoiding overfit.

4 Application II: Forecasting key macroeconomic indicators

Stock and Watson (2012) compare forecasts from various forecasting methods designed for a large number of orthogonal predictors with dynamic factor model (DFM) forecasts using a U.S. macroeconomic dataset with 143 quarterly variables spanning 1960–2008. They find that for most series, DFM-5 forecasts, which are based on a simple linear regression of the target on the top 5 principal components of the lagged variables, are superior to shrinkage forecasts. A large literature has explored ways of improving on their results (e.g. Stock and Watson, 2011, 2016; Kim and Swanson, 2014; Kelly and Pruitt, 2013, 2015; Bitto and Frühwirth-Schnatter, 2018; Chudik, Kapetanios, and Pesaran, 2018; Boot and Nibbering, 2019; Cepni, Güney, and Swanson, 2019).

Our question is different. We ask whether the text that appears in the *Wall Street Journal* contains additional information that is useful for forecasting these macroeconomic indicators beyond that of the DFM-5 benchmark, and whether HDMR can extract such information from the text.

4.1 Data

Because we are interested in forecasting, we focus on more recent data for which we have a much richer body of text—the full text of all *Wall Street Journal* articles that appear on the front page of the from January 1990 to December 2010. We include two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming, and aggregate the data to the monthly frequency so that $\boldsymbol{c}_t$ are phrase counts observed during month t.

Figure 6 shows the mean histogram for phrase counts in this sample. It shows that the mean distribution of phrase counts is now much less concentrated at zero, even when we include the top 100,000 most frequent phrases. Note that some individual phrases still exhibit many more zeros than implied by a poisson.

We match the text with the monthly macroeconomic indicators dataset made available for replication by Stock and Watson (2012). Following their categorization and normalization, while restricting the analysis to monthly series, we use 92 lower-level disaggregated series to compute principal components, and use 12 headline indicators as prediction targets. Table 6 reports summary statistics for these target variables.

4.2 Methods

We compare the out-of-sample root mean squared error for HDMR against the DFM-5 benchmark and against DMR with the same data. We assess out-of-sample fit via (i) random cross-validation and via (ii) pseudo out-of-sample rolling forward predictions (starting with half the available time-series as training sample and forecasting one month ahead). The latter approach is especially appealing given recent evidence of time-varying predictability (Farmer, Schmidt, and Timmermann, 2019).

Let $Y^{\tau}_{t+\tau}$ denote the variable to be forecasted in a τ-period ahead forecast $Y^{\tau}_{t+\tau}$. In each training sample (fold) we use the following methods for fitting.

For DFM-5, we simply form principal components of the entire sample once, and keep the top 5. We expect this to give DFM-5 a slight advantage over the competing methods. We then run an ordinary least squares regression of the target on the PCs:

$$Y^{\tau}_{t+\tau} = \beta_0 + \left[pc^1_t, \ldots, pc^5_t\right]' \boldsymbol{\beta} + \varepsilon_{t+\tau} \tag{25}$$

For HDMR, we use these same 5 PCs as well as the target as explanatory variables in inverse regressions (8) and (10). We then form sufficient reduction projections in the direction of the target. These projections summarize all the information in the text from phrase inclusion (z^0_{tY}) and repetition (z^+_{tY}) that is useful for predicting the target $Y^{\tau}_{t+\tau}$ after controlling for the PCs. We then run a forward regression (24) of the target on the sufficient reduction projections and the PCs, still using only training sample data:

$$Y^{\tau}_{t+\tau} = \beta_0 + \left[z^0_{tY}, z^+_{tY}, pc^1_t, \ldots, pc^5_t, m_i, d_i\right]' \boldsymbol{\beta} + \varepsilon_{t+\tau} \tag{26}$$

For DMR, we follow essentially the same procedure as for HDMR, with the same variables used to explain the text, and use its single sufficient reduction projection in the forward regression.

For each of the three models we use the predicted values from the forward regression, but applied to out-of-sample validation observations.

4.3 Forecasting results

Table 7 reports out-of-sample RMSE for HDMR, relative to DFM-5, which uses only the PCs without the text, and relative to DMR with the same data. Lower reported ratios indicate larger improvements from using HDMR. We find significant improvements in out-of-sample forecasts for a few variables at the monthly horizon ($\tau = 1$). For example, in Panel (a), total nonfarm payroll employment sees a 17% (18%) improvement from using HDMR relative to DFM-5 and housing starts sees a 32% (31%) improvement using random cross validation (rolling validation).

We find that the text appearing in the *Journal* is more informative about longer horizons, where HDMR generates significant improvements in most forecasts. For example, using the text improves industrial production and employment forecasts by about 20–30% at the annual horizon ($\tau = 12$). Newspaper coverage also considerably improves asset pricing forecasts at this longer horizon, generating large reductions in RMSE for interest rates, exchange rates, and stocks.

We further find that the advantage of HDMR over DMR increases as we increase the dimensionality of the text and use less frequent and more sparse phrases. Because of the computational cost of this exercise, we simply compare a restricted sample with the most frequent 10,000 phrases in Panel (a) with a larger vocabulary of the most frequent 100,000 phrases in Panel (b). We find, for example, that even though the quarterly inflation (CPI-ALL) forecasts of HDMR are comparable to those of DMR when we use 10,000 phrases, they are substantially better when we use 100,000 phrases. Similar improvements can be seen for M1 and consumer expectations. These findings suggest that HDMR can better learn from high dimensional sparse data like business newspaper coverage, which is selected based on newsworthiness.

4.4 Nowcasting results

Nowcasting is a related but distinct strand of the forecasting literature, which focuses on predicting activity that occurs now but that will only be reported later. A growing body of work starting with Giannone, Reichlin, and Small (2008) and surveyed in Bańbura, Giannone, Modugno, and Reichlin (2013) compares different methods for nowcasting macroeconomic indicators using lagged and contemporaneous macroeconmic series that is available before the actual measurements are released publicly.

We can evaluate whether the text of the *Journal* is informative about the present, by lagging one month both the target and the PCs so that the predictive forward regression becomes

$$Y^{\tau}_{t-1+\tau} = \beta_0 + \left[z^0_{tY}, z^+_{tY}, pc^1_{t-1}, \ldots, pc^5_{t-1}\right]' \boldsymbol{\beta} + \varepsilon_{t-1+\tau}. \tag{27}$$

We then follow the same out-of-sample validation procedure of the forecasting exercise. But now, the question we ask is whether the text of the *Journal* that is reported over month t is informative about macroeconomic activity over the same month Y^1_t, or starting in the same month, over and above the information in month $t-1$ macroeconomic indicators that are publicly available by month t.

Table 8 shows that this body of text, and our approach to learning from it in particular, are highly valuable for nowcasting. The improvements in out-of-sample fit are mostly greater than in the forecasting exercise of Table 7. These suggest that nowcasting with text may offer substantial gains over using macroeconmic series alone. We leave that question for future research, as well as whether our methods can improve upon state-of-the-art nowcasting models that use several factor lags, daily data, or mixed frequency VAR, and that deal with the jagged edge of macroeconomic news in realtime.

5 Conclusion

Text data is inherently high-dimensional, which makes machine learning regularization techniques natural tools for its analysis. Text is often selected by journalists, speechwriters, and others who cater to an audience with limited attention.

We develop an economically-motivated high dimensional selection model that can improve machine learning from text in particular and from sparse counts data more generally. Our highly scalable approach to modeling coverage selection is especially useful in cases where the cover/no-cover choice is separate or more interesting than the coverage quantity choice.

We apply this framework to backcast a central financial variable to historical periods using newspaper coverage, and to nowcast and forecast key macroeconomic indicators. We find that it substantially improves out-of-sample fit relative to alternative state-of-the-art approaches, and that this advantage increases with the sparsity of the text.

A Robustness

A.1 Alternative text regressions

Table 2a focuses on the optimal model by cross-validation, the one that uses the 10,000 most frequent phrases and compares HDMR to several benchmarks. For each model we report the measure of fit with and without the text, and the change in the measure of fit.

The first benchmark is DMR, which is provided with the same covariates and text. The improvement from modeling selection with HDMR is a 13 percent reduction in out-of-sample root mean squared error, from 82 to 71 basis points. HDMR provides a 45 percent improvement relative to the No Text benchmark that only uses the other covariates to predict.

The second benchmark model is a "fabricated" variant of HDMR (FHDMR) which adds $h_{ij} = \mathbf{1}(c_{ij} > 0)$ indicators to the text counts matrix $\boldsymbol{c}$ and then runs DMR as usual with $\tilde{\boldsymbol{c}} = [\boldsymbol{c}\ \ \boldsymbol{h}]$. If all that HDMR was doing is allow for a nonlinearity of the counts matrix, we would expect FHDMR to do just as well. Instead we find that it generates an 79 basis point RMSE, which is only slightly better that the 82 bp of DMR.

The last benchmark we consider is support vector regression (SVR), which Manela and Moreira (2017) use for a similar backcasting purpose. We follow their approach to calibrating the SVR meta-parameters. Even though we standardize both text and covariates to unit variance, SVR still cannot concentrate on the covariates, which provide first order information on our prediction target. SVR with text improves on an SVR without text, but its 126 basis points error rate is much larger than that of HDMR.

Table 2b reports pseudo-out-sample rolling back validation results, and finds similar improvements in out-of-sample fit from using HDMR.

A.2 Denser text

For a shorter time-series, 1990 to 2010, we can assess HDMR with much denser text—the full *Wall Street Journal.* Figure 7 shows that in this sample, the advantage from using the richer body of text is larger, as it attains lower out-of-sample error rates. These results, however, could also be driven by the different time period. What does seem like a robust conclusion from this comparison is that the advantage of HDMR over DMR increases with the sparsity of the text, which is plotted in the bottom panel. With $d = 500,000$ phrases, the counts matrix is just over 60 percent zeros, and HDMR reduces out-of-sample root mean squared error by 58 percent (121 to 51 basis points) relative to the No Text benchmark, and by 24 percent (68 to 51 bp) relative to DMR.

B What does HDMR approximate?

The method presented in this paper is developed with the goal of computational efficiency. With this goal in mind we model the intensive margin of the word counts using a poisson distribution. Here we show that our HDMR procedure can be thought as maximizing a likelihood that approximates a mixture of d bernoulis and a multinomial. The bernoulis determine the extensive margin, i.e., whether a given word is used in a given document, and the multinomial determines the intensive margin, i.e. how much a word is used in a given document.

Taddy (2015) shows that we can factorize a poisson distribution for the distribution of word counts in a document as a poisson distribution for the total number of words in the document, and a multinomial distribution for the count of each word given the todal word count:

$$p(c_i) = \prod_j Po(c_{ij}, e^{\eta_{ij}}) = MN(c_i; q_i, m_i)Po(m_i; \Lambda_i).$$

Furthermore, he shows that if η_{ij} is of the form $\eta_{ij} = \mu_i + \alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j$, and μ_i in the poisson model is estimated at its conditional MLE, $\mu_i^* = log\left(\frac{m_i}{\sum_j e^{\eta_{ij}}}\right)$, then the coefficients estimates $\boldsymbol{\varphi}_j$ and α_j are the same regardless of whether the distribution is conditioned on m_i. This means that conditional on μ_i^*, estiamting the poisson model is equivalent to estimating the multinomial model.

However, choosing $\mu_i^* = log\left(\frac{m_i}{\sum_j e^{\eta_{ij}}}\right)$ is computationally costly because it makes the d poisson models depend on each other, and therefore makes it impossible to parellize the model estimation across words, which is essential to make it useful for any Big Data application. Taddy (2015) shows particular cases where μ_i^* degenerates to $\hat{\mu_i} = log(m_i)$, and shows that in more general cases the choice $\hat{\mu_i}$ generalizes well in out-of-sample validation.

We now apply this insight above to our setting and show what our hurdle model approximates. We start from the mixture of d binomials for the extensive margin of each word, and a multinomial distribution for the excess word counts of each word. We then show that we obtain the joint distribution associated with (12).

Lets start with the word count c_i distribution conditional on word count totals m_i and document vocabulary size d_i. Define $h_i = c_i \geq 1$, where the inequality is evaluated word by word, so h_i is d by 1 vector of ones and zeros. Thus h_i defines the vocabulary of document i as it identifies all words used in the document. Note that $1'h_i = d_i \leq d$, i.e., each document vocabulary is smaller than the universal vocabulary. A document vocabulary is given by the mixture of d bernoullis with success probabilities given by Π_{ij} for word j in document i. The bernoullis determine whether $h_{ij} = 0$ or $h_{ij} = 1$,i.e., if word j is part of the used vocabulary in document i. This implies that the expected vocabulary size is $E[\sum_{j=1}^{d} h_{ij}] = \sum_{j=1}^{d} \Pi_{ij}$. This holds approximately for large enough universal vocabulary d because each draw is independent , i.e., the law of large numers implies $E[\sum_{j=1}^{d} h_{ij}] = \sum_{j=1}^{d} \Pi_{ij} \approx \sum_{j=1}^{d} h_{ij} = d_i$.

We can then represent the binomial distribution as

$$p(h_i) = \prod_{j=1}^{d} B(h_{ij}, \Pi_{ij})$$

with $\Pi_{ij} = 1/(1 + e^{-\gamma_i - \epsilon_{ij}})$, this implies the log-likelihood ratio is,

$$log\left(\frac{\Pi_{ij}}{1-\Pi_{ij}}\right) = log\left(\frac{1/(1+e^{-\gamma_i-\epsilon_{ij}})}{1-1/(1+e^{-\gamma_i-\epsilon_{ij}})}\right) = log\left(\frac{1/(1+e^{-\gamma_i-\epsilon_{ij}})}{(e^{-\gamma_i-\epsilon_{ij}})/(1+e^{-\gamma_i-\epsilon_{ij}})}\right) = \gamma_i + \epsilon_{ij},$$

where γ_i captures document specific vocabulary richness, and ϵ_{ij} capture the relative probability of usage of word j in document i. So for example, a news paper that talks about very broad set of topics will have large γ_i but ϵ_{ij} are all close around zero. While a finance journal will perhaps have a low γ_i but very disperse ϵ_{ij} with some words being significantly more likely to be used and other significantly less likely.

We now show that in similar spirit as Taddy (2015) we can use an approximation to motivate a plug in estimator for γ_i, which is document specific, and in general would require all the words to be jointly estimated. We start by doing a linear approximation of the probability Π_{ij} as function of ϵ_{ij},

$$\Pi_{ij} = 1/(1+e^{-\gamma_i-\epsilon_{ij}}) \approx 1/(1+e^{-\gamma_i}) + 1/(1+e^{-\gamma_i})^2 e^{-\gamma_i}\epsilon_{ij},$$

we then add probabilities within document and across words,

$$\begin{aligned}\sum_{j=1}^{d}\Pi_{ij} &\approx \sum_{j=1}^{d} 1/(1+e^{-\gamma_i}) + \sum_{j=1}^{d} 1/(1+e^{-\gamma_i})^2 e^{-\gamma_i}\epsilon_{ij} \\ &\approx d/(1+e^{-\gamma_i}) + 1/(1+e^{-\gamma_i})^2 e^{-\gamma_i}\sum_{j=1}^{d}\epsilon_{ij} \\ &\approx d/(1+e^{-\gamma_i}),\end{aligned}$$

where we use the the universal vocabulary is large and the the law of large numbers applies to get $\sum_{j=1}^{d}\epsilon_{ij} \approx 0$. Finally we use our earlier result that $\sum_{j=1}^{d}\Pi_{ij} \approx d_i$ to get

$$\hat{\gamma_i} = log\Big(\frac{d_i}{d-d_i}\Big),$$

which is our plug-in estimator.

We then use a standard logit approximation and obtain the following log-likelihood ratios,

$$log\left(\frac{\Pi_{ij}}{1-\Pi_{ij}}\right) = \gamma_i + \epsilon_{ij} = log\Big(\frac{d_i}{d-d_i}\Big) + \epsilon_{ij} = log\Big(\frac{d_i}{d-d_i}\Big) + \kappa_j + w_i'\delta_j,$$

where we model the word-document relative word usage ϵ_{ij} as a function of a word specific component κ_j that captures the fact that some words are just more likely to be used unconditionally, and the term $w_i'\delta_j$ which captures conditional variation in word usage as a function of our covariates w_i.

We descibed so far the joint distribution for the extensive margin of word usage. Now we describe the the intensive margin, which we model with a multinomial. Specifically the multinomial distribution describes the distribution of word repetitions. Define $c^*_{ij} = c_{ij} - h_{ij}$, the vector of repetitions, $m^*_i = m_i - d_i$ the total amount of word repetitions in document i and $q^*_{ij} = \frac{h_{ij}q_{ij}}{\sum_{j=1}^{d} h_{ij}q_{ij}}$ is the share of word counts for word j in document i. Note that $q^*_{ij} = 0$ if $h_{ij} = 0$ since this distribution is conditional on the vocabulary h_i.Then given the document vocabulary h_i, the distribution of word counts is

$$\mathrm{p}\left(c_i^*|\boldsymbol{v}_i, m_i^*, h_i, d_i\right) = \frac{m_i^*!}{\prod_{j=1}^{d} c_{ij}^*!} \prod_{j=1}^{d} \left(q_{ij}^*\right)^{c_{ij}^*} = MN\left(c_i^*; q_i^*, m_i^*\right)$$

We now follow Taddy (2015) and factorize the joint distribution implied by several independent poissons in terms of a multinomial distribution for the words counts and a poisson distribution for the total word count, and get

$$MN\left(c_{ij}^*; q_i^*, m_i^*\right) Po(m_i^*; \Lambda_i) = \prod_{j|h_{ij}=1} Po(c_{ij}^*; e^{\eta_{ij}}), with\ \eta_{ij} = \alpha_j + \mu_i + \varphi_j' v_i,$$

where μ_i picks up a document specific tendency to repetition when they use any word, α_j picks up a word specific tendency to repetition when used in any document, and $\varphi_j' v_i$ picks up the conditional association between word repetition and the covariates v_i.

Using Taddy (2015) proposed plug-in estimator for the poisson distribution of total word counts $\hat{\mu_i} = log(m_i^*) = log(m_i - d_i)$,we obtain that the multinomial distribution for word repetitions can be represented by the product of independent poissons,

$$MN\left(c_{ij}^*; q_i^*, m_i^*\right) \approx \prod_{j|h_{ij}=1} Po(c_{ij}^*; m_i^* e^{\alpha_j + \varphi_j' v_i \cdot}).$$

All together this implies our hurdle model that we bring to the data:

$$\prod_{j=1}^{d} B(h_{ij}, \Pi_{ij}) MN\left(c_{ij} - h_{ij}; \frac{h_{ij} q_{ij}}{\sum_{j=1}^{d} h_{ij} q_{ij}}, m_i - d_i\right) Po(m_i - d_i; \Lambda_i) \approx \prod_{j=1}^{d} \Pi_{ij}{}^{1-h_{ij}} \left(\Pi_{ij} Po(c_{ij} - h_{ij}; (m_i - d_i) e^{\alpha_j + \varphi_j' v_i \cdot})\right)^{h_{ij}},$$

where $\Pi_{ij} = \frac{1}{1 + \frac{d - d_i}{d_i} e^{-\kappa_j - w_i' \delta_j}}$.

C Sufficient Reduction Projection

We now show how these approximations allow us to recover the maximum likelihood estiamtes for the model parameters even if we separate the big estimation problem in multiple independent problems. Formally, under the conditions of the approximation described above we can factorize the log-lihehood of the joint word distribution as follows:

$$
\begin{aligned}
p(c_i|m_i,d_i) &= \prod_j \left[1/(e^{\gamma_i+\epsilon_{ij}}+1)\right]^{1-h_{ij}} \left[e^{\gamma_i+\epsilon_{ij}}/(e^{\gamma_i+\epsilon_{ij}}+1) Po(c_{ij}-h_{ij};(m_i-d_i)e^{\eta_{ij}})\right]^{h_{ij}} \\
&= \prod_j \frac{1}{(1+e^{\gamma_i+\epsilon_{ij}})} \prod_j e^{(\gamma_i+\epsilon_{ij})h_{ij}} \prod_j [Po(c_{ij}-h_{ij};(m_i-d_i)e^{\eta_{ij}})]^{h_{ij}} \\
&= \prod_j \frac{1}{(1+e^{log\left(\frac{d_i}{d-d_i}\right)+\kappa_j+w_i'\delta_j})} \prod_j e^{(log\left(\frac{d_i}{d-d_i}\right)+\kappa_j+w_i'\delta_j)h_{ij}} \prod_j [Po(c_{ij}-h_{ij};(m_i-d_i)e^{\eta_{ij}})]^{h_{ij}}.
\end{aligned}
$$

This has two components: the bernoulli component that determines each document vocabulary and the Poisson that determines each document use of this vocabulary. Focusing on the bernoulli component we can write,

$$
\begin{aligned}
\prod_j \frac{1}{(1+e^{log\left(\frac{d_i}{d-d_i}\right)+\kappa_j+w_i'\delta_j})} \prod_j e^{(log\left(\frac{d_i}{d-d_i}\right)+\kappa_j+w_i'\delta_j)h_{ij}} &= \prod_j \frac{1}{(1+e^{log\left(\frac{d_i}{d-d_i}\right)+\kappa_j+w_i'\delta_j})} \prod_j e^{\left(log\left(\frac{d_i}{d-d_i}\right)+\kappa_j\right)h_{ij}} \prod_j e^{w_i'\delta_j h_{ij}} \\
&= a(w_i,d_i)\psi(h_i) \prod_{j=1}^{d} e^{w_i'\delta_j h_{ij}}
\end{aligned}
$$

where $\psi(h_i) = \prod_{j=1}^{d} e^{\left(log\left(\frac{d_i}{d-d_i}\right)+\kappa_j\right)h_{ij}}$ and $a(w_i,d_i) = \prod_{j=1}^{d} \left(1+e^{log\left(\frac{d_i}{d-d_i}\right)+\kappa_j+w_i'\delta_j}\right)^{-1}$. We can similarly factorize the intensive margin component modeled with the poisson distribution as

$$
\begin{aligned}
&\prod_j [Po(c_{ij}-h_{ij};(m_i-d_i)e^{\eta_{ij}})]^{h_{ij}} \\
&= \prod_{j=1|h_{ij=1}}^{d} \frac{((m_i-d_i)e^{\eta_{ij}})^{c_{ij}-1} e^{-(m_i-d_i)e^{\eta_{ij}}}}{(c_{ij}-1)!} \\
&= \prod_{j=1|h_{ij=1}}^{d} \exp\left((c_{ij}-1)\log(m_i-d_i)+(c_{ij}-1)\eta_{ij}\right) \exp\left(-(m_i-d_i)e^{\eta_{ij}}\right) \exp\left(-log\left((c_{ij}-1)!\right)\right) \\
&= \prod_{j=1|h_{ij=1}}^{d} \exp\left(-(m_i-d_i)e^{\alpha_j+\varphi_j' v_i}\right) \prod_{j=1|h_{ij=1}}^{d} \exp\left((c_{ij}-1)\left(\log(m_i-d_i)+\alpha_j\right)-log\left((c_{ij}-1)!\right)\right) \prod_{j=1|h_{ij=1}}^{d} \exp\left((c_{ij}-1)(\varphi_j' v_i)\right).
\end{aligned}
$$

Therefore the joint distribution can be factored in as

$$
p(c_i|v_i,w_i) = \phi(c_i)\psi(h_i)a(w_i,d_i)b(v_i,m_i,d_i)\exp(w_i'\delta h_i + \varphi v_i(c_i-h_i)),
$$

where

$$b(v_i, m_i, d_i) = \prod_{j=1|h_{ij=1}}^{d} \exp\left(-(m_i - d_i)e^{\alpha_j + \varphi_j' v_i}\right)$$

$$\phi(c_i) = \prod_{j=1|h_{ij=1}}^{d} \exp\left((c_{ij} - 1)\left(\log\left(m_i - d_i\right) + \alpha_j\right) - log\left((c_{ij} - 1)!\right)\right).$$

This expression is exactly the joint likelihood that our HDMR procedure maximizes.

References

Adrian, Tobias, Erkko Etula, and Tyler Muir, 2014, Financial intermediaries and the cross-section of asset returns, *Journal of Finance* 69, 2557–2596.

Antweiler, Werner, and Murray Z. Frank, 2004, Is all that talk just noise? The information content of Internet stock message boards, *Journal of Finance* 59, 1259–1293.

Athey, Susan, 2017, Beyond prediction: Using big data for policy problems, *Science* 355, 483–485.

——— , 2018, The impact of machine learning on economics, in *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press).

——— , Guido Imbens, Thai Pham, and Stefan Wager, 2017, Estimating average treatment effects: Supplementary analyses and remaining challenges, *American Economic Review* 107, 278–81.

Athey, Susan, Julie Tibshirani, and Stefan Wager, 2019, Generalized random forests, *The Annals of Statistics* 47, 1148–1178.

Bańbura, Marta, Domenico Giannone, Michele Modugno, and Lucrezia Reichlin, 2013, Now-Casting and the Real-Time Data Flow, in *Handbook of Economic Forecasting*, vol. 2 . pp. 195–237 (Elsevier).

Bajari, Patrick, Denis Nekipelov, Stephen P. Ryan, and Miaoyu Yang, 2015, Machine learning methods for demand estimation, *American Economic Review* 105, 481–85.

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

Baron, Matthew, and Tyler Muir, 2018, Intermediaries and Asset Prices: Evidence from the U.S., U.K., and Japan, 1870-2016, Working paper.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen, 2012, Sparse models and methods for optimal instruments with an application to eminent domain, *Econometrica* 80, 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Christian Hansen, 2017, Program evaluation and causal inference with high-dimensional data, *Econometrica* 85, 233–298.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, 2014, Inference on treatment effects after selection among high-dimensional controls, *The Review of Economic Studies* 81, 608–650.

Bitto, Angela, and Sylvia Frühwirth-Schnatter, 2018, Achieving shrinkage in a time-varying parameter model framework, *Journal of Econometrics*.

Boot, Tom, and Didier Nibbering, 2019, Forecasting using random subspace methods, *Journal of Econometrics*.

Boydstun, Amber E., 2013, *Making the News: Politics, the Media, and Agenda Setting* (University of Chicago Press).

Brunnermeier, Markus K., and Yuliy Sannikov, 2014, A macroeconomic model with a financial sector, *American Economic Review* 104, 379–421.

Cepni, Oguzhan, I. Ethem Güney, and Norman R. Swanson, 2019, Nowcasting and forecasting GDP in emerging markets using global financial and macroeconomic diffusion indexes, *International Journal of Forecasting* 35, 555–572.

Chudik, Alexander, George Kapetanios, and M. Hashem Pesaran, 2018, A One Covariate at a Time, Multiple Testing Approach to Variable Selection in High-Dimensional Linear Regression Models, *Econometrica* 86, 1479–1512.

Diebold, Francis X, and Robert S Mariano, 1995, Comparing Predictive Accuracy, *Journal of Business & Economic Statistics* 13, 134–144.

Durante, Ruben, and Ekaterina Zhuravskaya, 2018, Attack When the World Is Not Watching? US News and the Israeli-Palestinian Conflict, *Journal of Political Economy* 126, 1085–1133.

Einav, Liran, Amy Finkelstein, Sendhil Mullainathan, and Ziad Obermeyer, 2018, Predictive modeling of US health care spending in late life, *Science* 360, 1462–1465.

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

Engelberg, Joseph, and Christopher A. Parsons, 2010, The Causal Impact of Media in Financial Markets, *Journal of Finance*.

Fang, Lily, and Joel Peress, 2009, Media Coverage and the Cross-section of Stock Returns, *Journal of Finance* 64, 2023–2052.

Farmer, Leland, Lawrence Schmidt, and Allan Timmermann, 2019, Pockets of Predictability, Working Paper.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani, 2010, Regularization paths for generalized linear models via coordinate descent, *Journal of statistical software* 33, 1.

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

Gallant, A Ronald, and Douglas W Nychka, 1987, Semi-nonparametric maximum likelihood estimation, *Econometrica* pp. 363–390.

García, Diego, 2013, Sentiment during recessions, *Journal of Finance* 68, 1267–1300.

Gentzkow, Matthew, Bryan T. Kelly, and Matt Taddy, 2019, Text as Data, *Journal of Economic Literature*.

Gentzkow, Matthew, Jesse Shapiro, and Matt Taddy, 2019, Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech, *Econometrica* Status: forthcoming.

Gentzkow, Matthew, and Jesse M. Shapiro, 2006, Media bias and reputation, *Journal of Political Economy* 114, pp. 280–316.

Giannone, Domenico, Lucrezia Reichlin, and David Small, 2008, Nowcasting: The real-time informational content of macroeconomic data, *Journal of Monetary Economics* 55, 665–676.

Greene, William, 2007, Functional form and heterogeneity in models for count data, *Foundations and Trends in Econometrics* 1, 113–218.

Haddad, Valentin, and Tyler Muir, 2018, Do Intermediaries Matter for Aggregate Asset Prices?, Working paper.

Hanley, Kathleen Weiss, and Gerard Hoberg, 2019, Dynamic Interpretation of Emerging Risks in the Financial Sector, *The Review of Financial Studies*.

Hassan, Tarek A., Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, 2017, Firm-Level Political Risk: Measurement and Effects, SSRN Scholarly Paper ID 2838644 Social Science Research Network Rochester, NY.

He, Zhiguo, Bryan Kelly, and Asaf Manela, 2017, Intermediary asset pricing: New evidence from many asset classes, *Journal of Financial Economics* 126, 1–35.

He, Zhiguo, and Arvind Krishnamurthy, 2012, A model of capital and crises, *The Review of Economic Studies* 79, 735–777.

———, 2013, Intermediary asset pricing, *American Economic Review* 103, 732–770.

———, 2018, Intermediary Asset Pricing and the Financial Crisis, *Annual Review of Financial Economics* 10, 173–197.

Heckman, James J., 1979, Sample selection bias as a specification error, *Econometrica* 47, 153–161.

Heckman, James J, and Thomas E MaCurdy, 1986, Labor econometrics, *Handbook of econometrics* 3, 1917–1977.

Hoberg, Gerard, and Gordon Phillips, 2016, Text-Based Network Industries and Endogenous Product Differentiation, *Journal of Political Economy* 124, 1423–1465.

Hodrick, Robert J, 1992, Dividend yields and expected stock returns: Alternative procedures for inference and measurement, *Review of Financial Studies* 5, 357–386.

Jiang, Fuwei, Joshua Lee, Xiumin Martin, and Guofu Zhou, 2019, Manager sentiment and stock returns, *Journal of Financial Economics* 132, 126–149.

Kelly, Bryan, and Seth Pruitt, 2013, Market expectations in the cross-section of present values, *Journal of Finance* 68, 1721–1756.

———, 2015, The three-pass regression filter: A new approach to forecasting using many predictors, *Journal of Econometrics* 186, 294–316.

Kim, Hyun Hak, and Norman R. Swanson, 2014, Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence, *Journal of Econometrics* 178, 352–367.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, 2017, Human decisions and machine predictions, *The quarterly journal of economics* 133, 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer, 2015, Prediction policy problems, *American Economic Review* 105, 491–95.

Koijen, Ralph S. J., and Motohiro Yogo, 2019, A Demand System Approach to Asset Pricing, *Journal of Political Economy* Status: forthcoming.

Ludwig, Jens, Sendhil Mullainathan, and Jann Spiess, 2017, Machine learning tests for effects on multiple outcomes, *arXiv preprint arXiv:1707.01473*.

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

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

McCullagh, Peter, and James A Nelder, 1989, *Generalized Linear Models* (Chapman & Hall).

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.

Mullainathan, Sendhil, and Andrei Shleifer, 2005, The market for news, *American Economic Review* 95, 1031–1053.

Mullainathan, Sendhil, and Jann Spiess, 2017, Machine learning: an applied econometric approach, *Journal of Economic Perspectives* 31, 87–106.

Pastor, Lubos, and Pietro Veronesi, 2012, Uncertainty about government policy and stock prices, *Journal of Finance* 67, 1219–1264.

Qin, Bei, David Strömberg, and Yanhui Wu, 2018, Media Bias in China, *American Economic Review* 108, 2442–2476.

Rennie, Jason D., Lawrence Shih, Jaime Teevan, and David R. Karger, 2003, Tackling the poor assumptions of naive bayes text classifiers, in *Proceedings of the 20th international conference on machine learning (ICML-03)* pp. 616–623.

Schervish, Mark J, 1995, *Theory of statistics* (Springer Science & Business Media).

Stock, James H., and Mark Watson, 2011, Dynamic factor models, *Oxford handbook on economic forecasting.*

Stock, James H., and Mark W. Watson, 2012, Generalized Shrinkage Methods for Forecasting Using Many Predictors, *Journal of Business & Economic Statistics* 30, 481–493.

——— , 2016, Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics, in *Handbook of macroeconomics*, vol. 2 . pp. 415–525 (Elsevier).

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

——— , 2015, Distributed multinomial regression, *Annals of Applied Statistics* 9, 1394–1414.

——— , 2017, One-step estimator paths for concave regularization, *Journal of Computational and Graphical Statistics* pp. 1–12.

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

Tibshirani, Robert, 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

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

Wilbur, W. John, and Won Kim, 2009, The ineffectiveness of within-document term frequency in text classification, *Information Retrieval* 12, 509–525.

Zipf, George Kingsley, 1935, The psychology of language, *Houghton-Mifflin.*

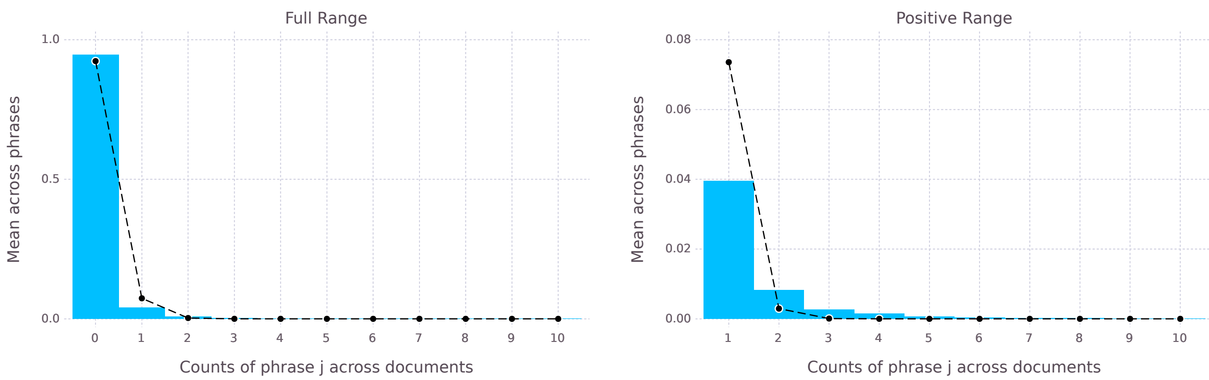


Figure 1: Mean distribution of WSJ front page articles monthly phrase counts

Notes: The figure shows the mean histogram for phrases that appear in the title or lead paragraph of front page *Wall Street Journal* articles, aggregated to form a monthly sample from July 1926 to February 2016. We construct the mean histogram by first calculating a histogram for each phrase across documents, and then averaging over phrases and normalizing to unit scale. The left panel shows that the entire range is highly sparse (has many zeros). The right panel omits zero counts, and shows that a poisson density fitted to the entire range (dashed line) is a poor description of the positive range. The corpus includes the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming. Including less frequent phrases makes the corpus sparser and the pattern above more pronounced.

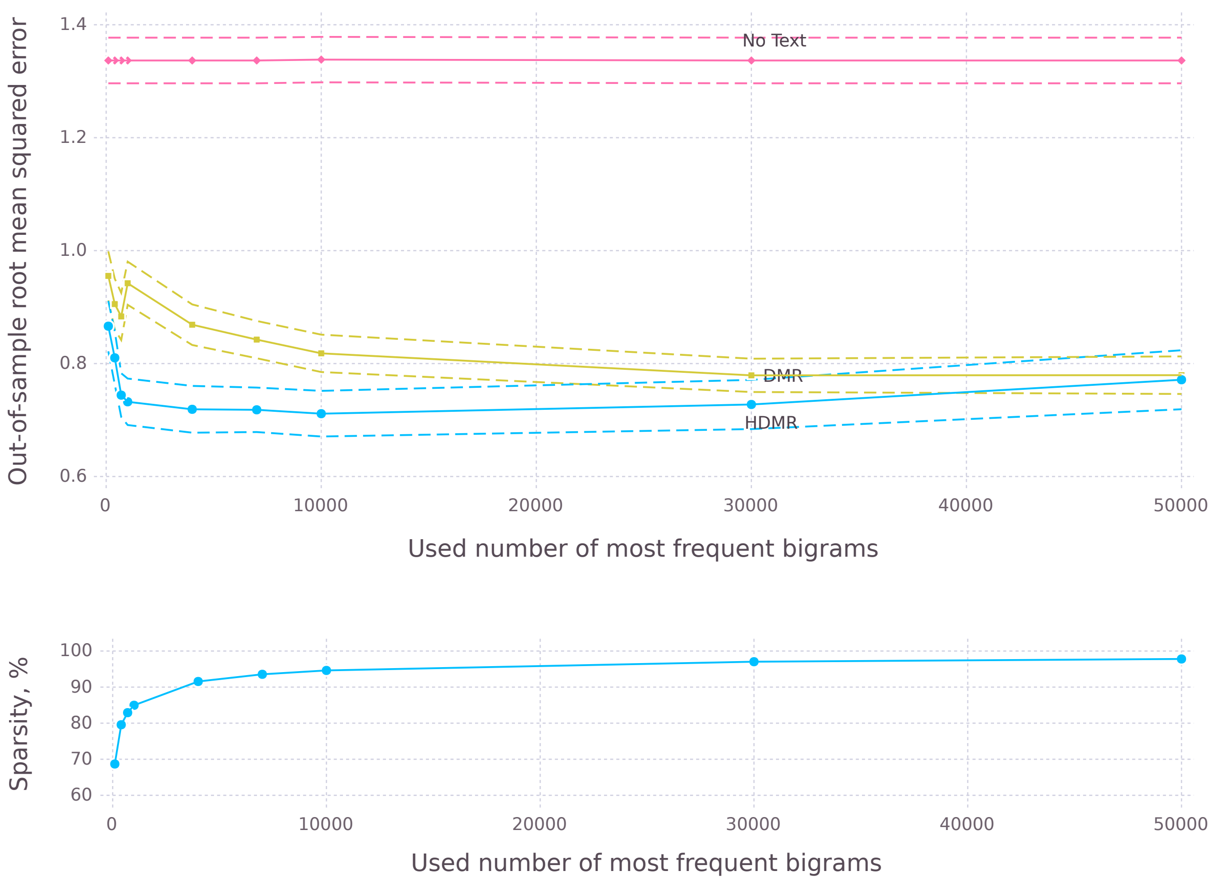


Figure 2: Predicting the intermediary capital ratio with text and covariates: 10-fold cross validation

Notes: The top panel reports out-of-sample root mean squared error from a 10-fold cross validation exercise that tries to predict the intermediary capital ratio (icr_t) using the log price dividend ratio (pd_t), realized variance of financial stocks ($rvfin$) over the same month, over the prior month, and over the prior year, and monthly WSJ front page phrase counts, over the subsample when the capital ratio is available, January 1970 to February 2016. Our proposed model, the hurdle distributed multinomial regression (HDMR) is compared with two benchmarks: (a) The distributed multinomial regression (DMR), which is provided with the same covariates and text, is a state-of-the-art approach to prediction with high-dimensional text, and (b) a linear regression of the target on the same covariates without the text (No Text). The figure shows how the advantage of HDMR in terms of out-of-sample fit changes as a function of the number of most frequent phrases included in the corpus. Dashed lines indicate the 95% confidence interval. The bottom panel shows how sparsity increases with this choice, where sparsity is the fraction of zero phrase counts in the corpus.

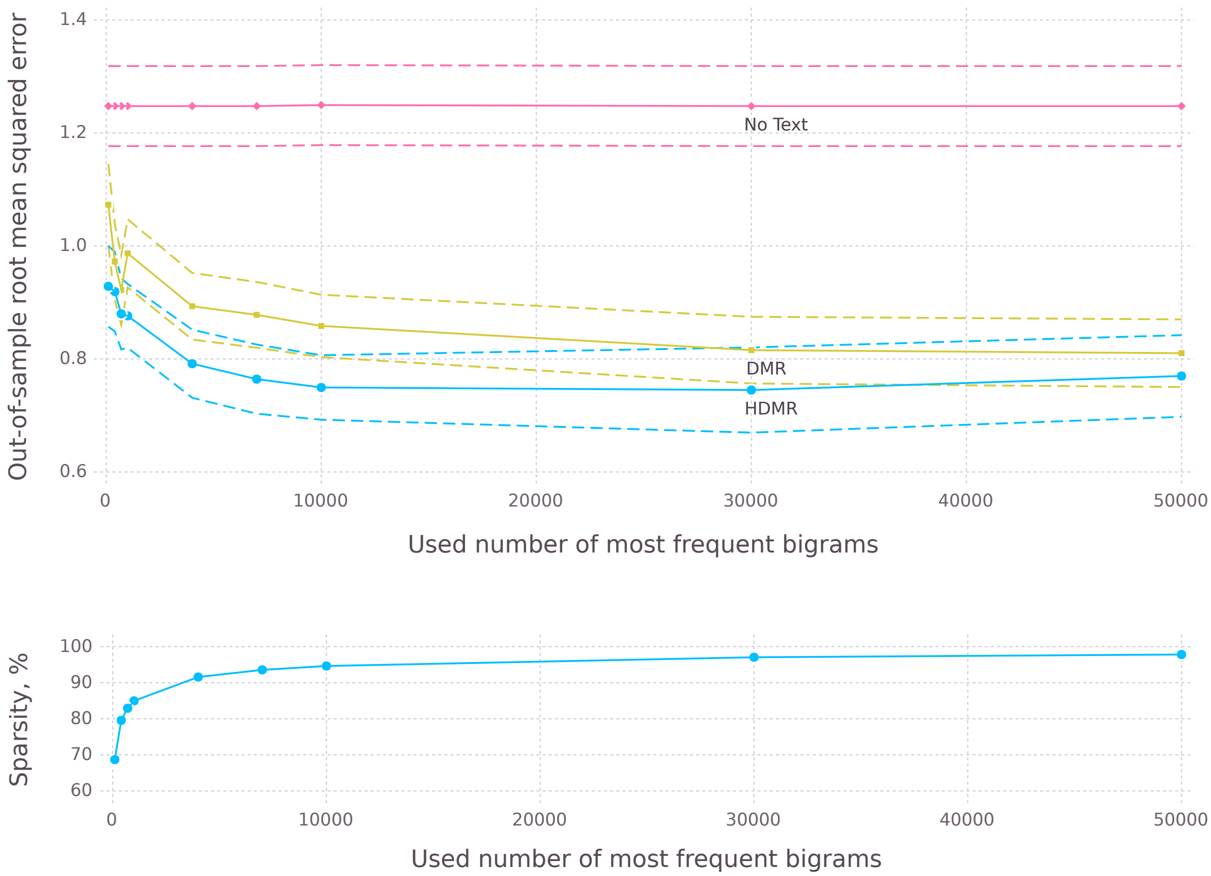
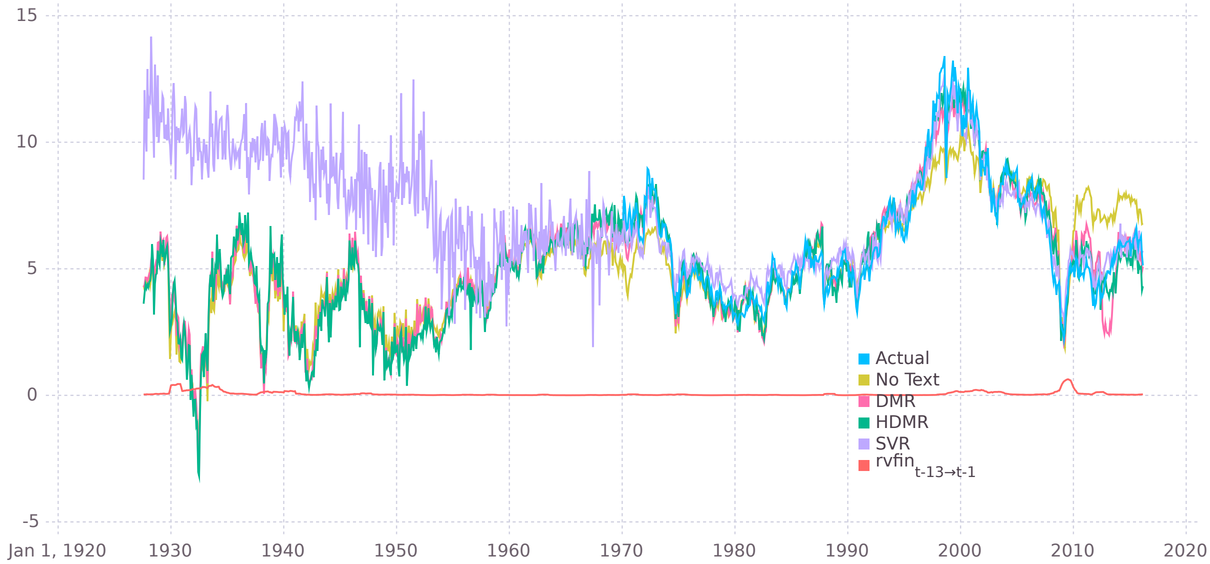
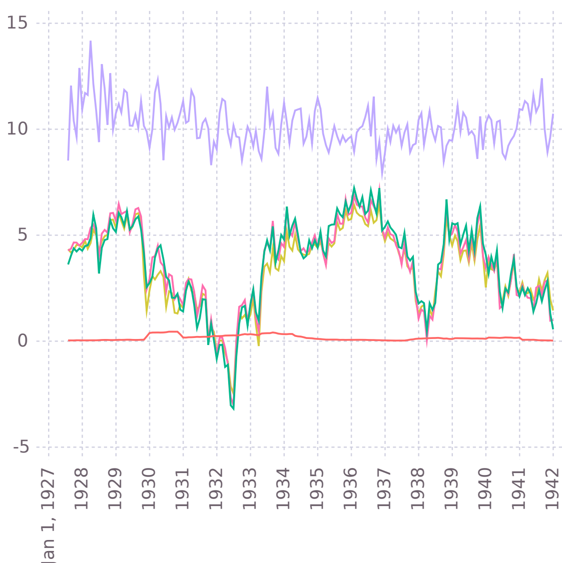


Figure 3: Predicting the intermediary capital ratio with text and covariates: Pseudo out-of-sample

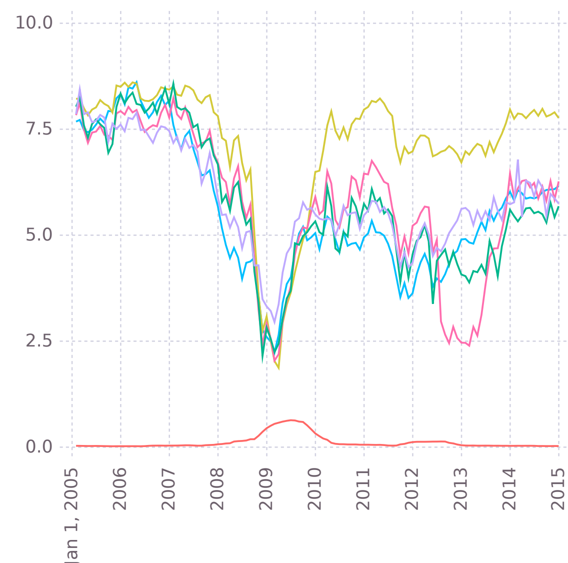
Notes: The top panel reports out-of-sample root mean squared error for predicting the intermediary capital ratio (icr_t) using the log price dividend ratio (pd_t), realized variance of financial stocks ($rvfin$) over the same month, over the prior month, and over the prior year, and monthly WSJ front page phrase counts, over the subsample when the capital ratio is available, January 1970 to February 2016. Unlike the random folds used before for validation, here we assess fit of a pseudo out-of-sample rolling back prediction exercise, starting with the later half of the sample, predicting one observation earlier, then extending the training sample by one earlier observation and rolling backward to assess fit in a backcasting exercise. Our proposed model, the hurdle distributed multinomial regression (HDMR) is compared with two benchmarks: (a) The distributed multinomial regression (DMR), which is provided with the same covariates and text, is a state-of-the-art approach to prediction with high-dimensional text, and (b) a linear regression of the target on the same covariates without the text (No Text). The figure shows how the advantage of HDMR in terms of out-of-sample fit changes as a function of the number of most frequent phrases included in the corpus. Dashed lines indicate the 95% confidence interval. The bottom panel shows how sparsity increases with this choice, where sparsity is the fraction of zero phrase counts in the corpus.



(a) Full sample



(b) Great Depression



(c) Great Recession

Figure 4: Backcasting the intermediary capital ratio with text and covariates

Notes: The figure shows the predicted intermediary capital ratio ($\widehat{icr_t}$) using the log price dividend ratio (pd_t), realized variance of financial stocks ($rvfin$) over the same month, over the prior month, and over the prior year, and monthly WSJ front page phrase counts, over the extended sample, June 1927 to February 2016. The intermediary capital ratio is only available starting January 1970. Our proposed model, the hurdle distributed multinomial regression (HDMR), which excludes prior year financial stocks variance ($rvfin_{t-13 \to t-1}$, bottom line) from the repetition equation, is compared with three benchmarks: (a) distributed multinomial regression (DMR, Taddy, 2015), which is provided with the same covariates and text, (b) support vector regression (SVR), and (c) linear regression of the target on the same covariates without the text (No Text).

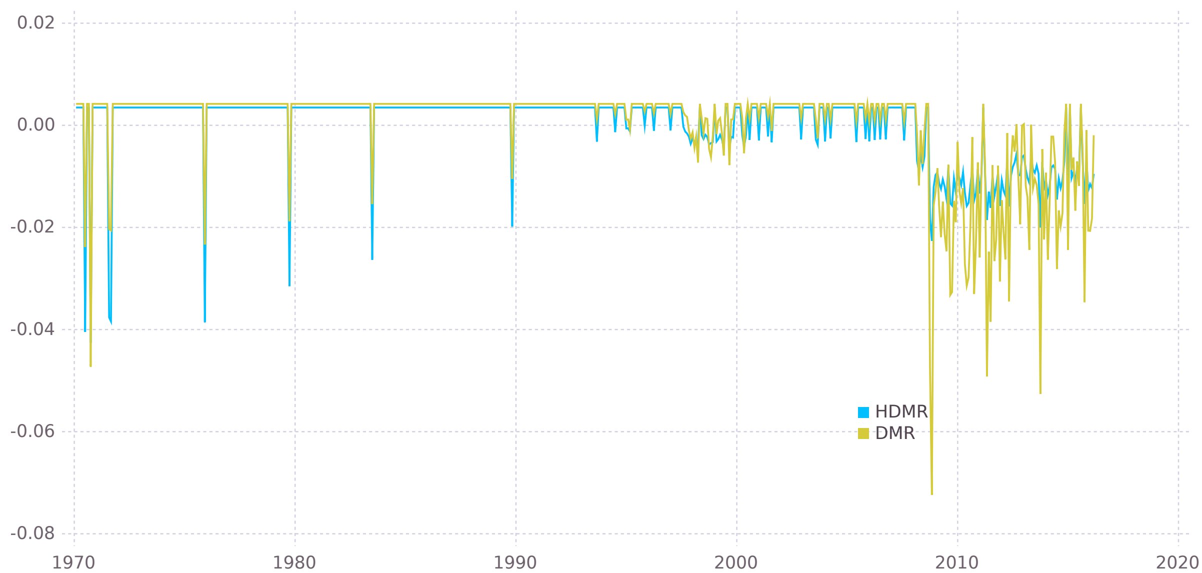


Figure 5: Focusing on the phrase "financial crisis" for intuition

Notes: The figure shows the predicted intermediary capital ratio ($\widehat{icr_t}$) due only to a single stemmed phrase, "financi crisi." Our proposed model, the hurdle distributed multinomial regression (HDMR) gives more weight to the mere inclusion of this phrase on the front page of the *Wall Street Journal*, as opposed to its repeated use. Distributed multinomial regression (DMR) estimates, which does not break the variation into inclusion versus repetition, are shown for comparison.

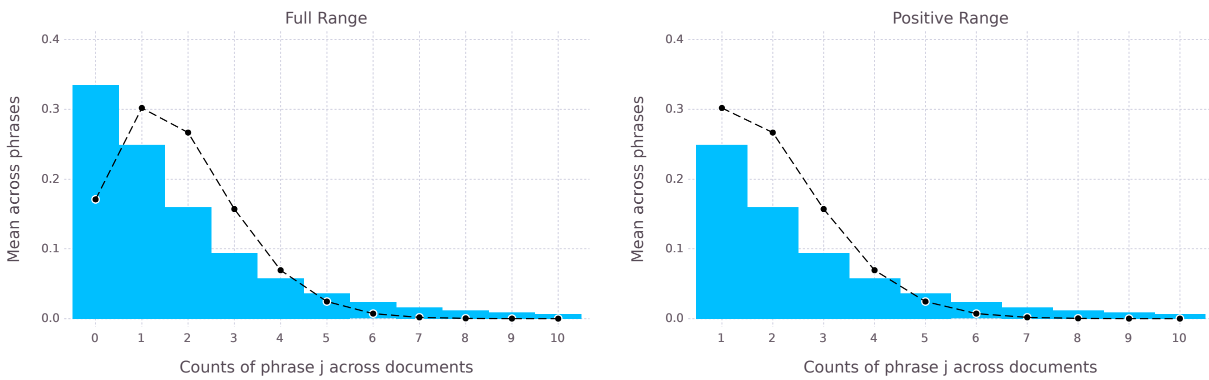


Figure 6: Mean distribution of full WSJ monthly phrase counts

Notes: The figure shows the normalized mean histogram for phrases that appear in the title or body of all *Wall Street Journal* articles, aggregated to form a monthly sample from January 1990 to December 2010. We construct the mean histogram by first calculating a histogram for each phrase across documents, and then averaging over phrases and normalizing to unit scale. The corpus includes the 100,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming.

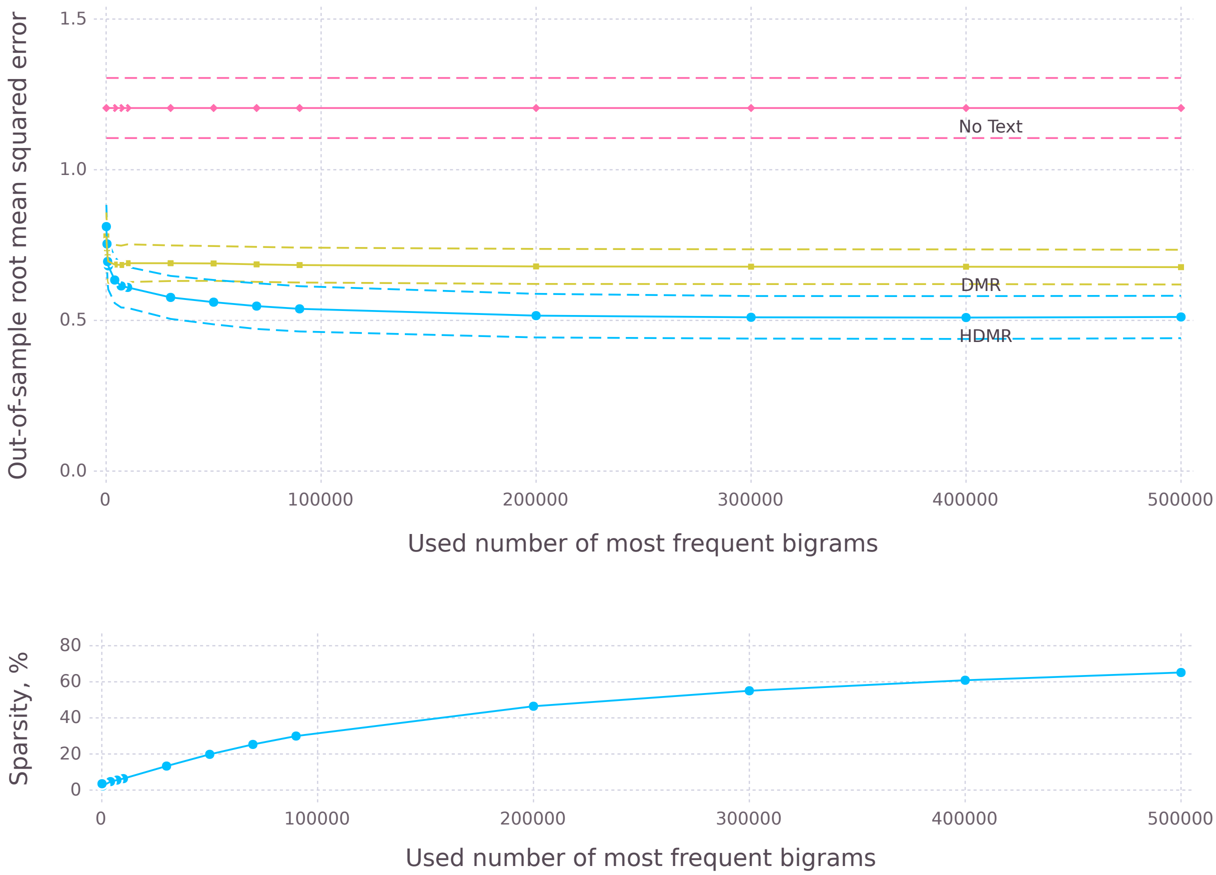


Figure 7: Predicting the intermediary capital ratio with denser text and covariates

Notes: The top panel reports out-of-sample root mean squared error from a 10-fold cross validation exercise that tries to predict the intermediary capital ratio (icr_t) using the log price dividend ratio (pd_t), realized variance of financial stocks ($rvfin$) over the same month, over the prior month, and over the prior year, and all monthly WSJ phrase counts, over the subsample when this text is available, January 1990 to December 2010.Our proposed model, the hurdle distributed multinomial regression (HDMR) is compared with two benchmarks: (a) The distributed multinomial regression (DMR), which is provided with the same covariates and text, is a state-of-the-art approach to prediction with high-dimensional text, and (b) a linear regression of the target on the same covariates without the text (No Text). The figure shows how the advantage of HDMR in terms of out-of-sample fit changes as a function of the number of most frequent phrases included in the corpus. Dashed lines indicate the 95% confidence interval. The bottom panel shows how sparsity increases with this choice, where sparsity is the fraction of zero phrase counts in the corpus.

Table 1: Summary statistics: Backcasting application

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

Notes: Reported are summary statistics for the *Wall Street Journal* front page articles text and for variables in the monthly sample from July 1926 to May 2016. The corpus includes the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming. To summarize the text, we report the mean of per phrase statistics, for counts c_{tj} and for inclusion indicators, $h_{tj} \equiv 1_{\{c_{tj}>\}}$. Intermediary capital ratio (icr_t) is the aggregate ratio of market equity to market equity plus book debt of New York Fed primary dealers in percentage terms. The log price over past year dividends (pd_t) is from CRSP. Realized variance ($rvfin$) is the annualized daily variance of financial stock returns over the prior month ($t-1 \to t$) or year ($t-12 \to t$).

Table 2: Predicting the intermediary capital ratio with text and covariates

(a) Cross-validation with 10 random folds

	Out-of-sample			In-sample		
Model	Text	No Text	Difference	Text	No Text	Difference
HDMR	0.711	1.338	-0.627	0.573	1.322	-0.750
	(0.021)	(0.021)	(0.020)	(0.002)	(0.002)	(0.002)
DMR	0.818	1.338	-0.520	0.716	1.322	-0.606
	(0.017)	(0.021)	(0.022)	(0.003)	(0.002)	(0.004)
FHDMR	0.789	1.338	-0.549	0.673	1.322	-0.649
	(0.016)	(0.021)	(0.020)	(0.002)	(0.002)	(0.003)
SVR	1.256	1.334	-0.078	0.641	1.326	-0.685
	(0.042)	(0.020)	(0.046)	(0.005)	(0.002)	(0.005)

(b) Pseudo out-of-sample rolling back

	Out-of-sample			In-sample		
Model	Text	No Text	Difference	Text	No Text	Difference
HDMR	0.750	1.249	-0.500	0.600	1.341	-0.741
	(0.029)	(0.036)	(0.030)	(0.002)	(0.003)	(0.002)
DMR	0.858	1.249	-0.391	0.770	1.341	-0.570
	(0.028)	(0.036)	(0.022)	(0.003)	(0.003)	(0.002)
FHDMR	0.828	1.249	-0.422	0.732	1.341	-0.609
	(0.028)	(0.036)	(0.024)	(0.003)	(0.003)	(0.002)
SVR	1.911	0.956	0.955	0.633	1.358	-0.725
	(0.055)	(0.042)	(0.065)	(0.001)	(0.003)	(0.003)

Notes: Reported is in- and out-of-sample root mean squared error (RMSE) for predicting the intermediary capital ratio (icr_t) using the log price dividend ratio (pd_t), realized variance of financial stocks ($rvfin$) over the same month, over the prior month, and over the prior year. Models with text additionally include monthly WSJ front page phrase counts, over the subsample when the capital ratio is available, January 1970 to February 2016. The corpus includes the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming. Panel (a) uses 10 random folds for validation while Panel (b) uses pseudo out of sample prediction, starting with the latter half of the sample and rolling back one observation at a time. Our proposed model, the hurdle distributed multinomial regression (HDMR), which excludes prior year $rvfin$ from the repetition equation, is compared with three benchmarks:(a) distributed multinomial regression (DMR, Taddy, 2015), which is provided with the same covariates and text, (b) a "fabricated" variant of HDMR which adds $h_{ij} = \mathbf{1}(c_{ij} > 0)$ indicators to the text counts matrix $\boldsymbol{c}$ and then runs DMR (FHDMR), and (c) support vector regression (SVR). For each model we report RMSE with and without the text, the change in the measure of fit. Standard errors are in parentheses.

Table 3: Pivotal phrases for predicting the intermediary capital ratio out-of-sample

(a) Cross-validation with 10 random folds

Phrase	Δ OOS RMSE	δ	φ	Mean counts	Mean inclusions
barack obama	0.006	-7.050	-0.166	1.224	0.195
jobless marri	0.005	1.583	0.460	0.291	0.260
busi bulletin	0.004	2.338	0.195	4.863	0.602
instal credit	0.004	2.515	0.000	0.020	0.020
industri output	0.003	0.776	0.000	0.221	0.219
euro zone	0.003	-1.169	-1.677	0.508	0.132
penn central	0.003	1.375	2.024	0.063	0.042
presid nixon	0.003	1.210	0.528	0.291	0.123
aluminum output	0.003	1.329	0.753	0.206	0.177
labor letter	0.003	2.090	0.177	3.273	0.535

(b) Pseudo out-of-sample rolling back

Phrase	Δ OOS RMSE	δ	φ	Mean counts	Mean inclusions
busi bulletin	0.075	2.338	0.195	4.863	0.602
tax report	0.065	1.949	0.254	4.929	0.604
labor letter	0.063	2.090	0.177	3.273	0.535
washington wire	0.046	1.458	0.184	5.430	0.618
steel product	0.023	0.923	0.567	0.577	0.512
factori shipment	0.022	0.796	0.036	0.629	0.530
hour earn	0.022	0.987	0.000	0.570	0.524
jobless marri	0.019	1.583	0.460	0.291	0.260
and unfil	0.019	-0.188	0.000	0.219	0.215
lead indic	0.018	1.109	-0.163	0.492	0.391

Notes: The table reports the top 10 phrases whose removal from the corpus causes the largest deterioration in out-of-sample root-mean-squared-error (Δ OOS RMSE), when predicting the intermediary capital ratio (icr_t) using text and the log price dividend ratio (pd_t), realized variance of financial stocks ($rvfin$) over the same month, over the prior month, and over the prior year. The text includes monthly WSJ front page phrase counts, over the subsample when the capital ratio is available, January 1970 to February 2016. Panel (a) uses 10 random folds for validation while Panel (b) uses pseudo out of sample prediction, starting with the latter half of the sample and rolling back one observation at a time. We also report full sample HDMR coefficients on icr_t for phrase inclusion (δ) and repetition (φ), and mean counts and inclusions across observations. The corpus includes the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming.

Table 4: Focusing on the phrase "financial crisis" for intuition

(a) Backward regressions

	HDMR		DMR
	Repetition	Inclusion	
Intercept	-10.73	-9.60	-16.65
icr_t	-0.43	-0.60	-0.61
pd_t	2.04	4.01	3.49
$rvfin_{t-1\to t}$	1.41	0.90	1.44
$rvfin_{t-2\to t-1}$	-0.41	1.00	-0.54
$rvfin_{t-13\to t-1}$		2.66	1.26

(b) Forward regressions

	HDMR	DMR
Repetition	-1.11	-4.71
Inclusion	-5.27	

Notes: Panel (a) reports backward HDMR coefficient estimates for the (stemmed) phrase "financi crisi" on the covariates, which excludes prior year financial stocks volatility ($rvfin$) from the repetition equation. Panel (b) reports the forward regression coefficient products with those of the backward regression, $b_z\varphi_{jy}$ and $b_s\delta_{jy}$ for HDMR, and contrasts it with the corresponding single coefficient product of DMR. The hurdle distributed multinomial regression (HDMR) gives more weight to the mere inclusion of this phrase on the front page of the *Wall Street Journal*, as opposed to its repeated use. Distributed multinomial regression (DMR) estimates, which does not break the variation into inclusion versus repetition, are shown for comparison. The corpus includes the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming.

Table 5: Time-varying risk premia and the news-implied intermediary capital ratio

(a) Postwar sample, 194601–201602

	$r^{em}_{t\to t+1}$			$r^{em}_{t\to t+3}$			$r^{em}_{t\to t+12}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
icr_t	-3.10			-3.33			-2.94		
	(2.34)			(2.57)			(2.49)		
$\widehat{icr}_t$		-4.59***			-4.69**			-5.00**	
		(1.77)			(1.84)			(1.95)	
z^0_t			6.44***			6.75***			6.41***
			(2.44)			(1.87)			(1.65)
z^+_t			-3.06			-2.87			-3.13
			(2.06)			(2.02)			(1.90)
pd_t			-9.61***			-10.13***			-10.13***
			(2.50)			(2.34)			(2.19)
$rvfin_{t-1\to t}$			-9.86***			-4.60**			-1.04
			(2.79)			(2.34)			(0.78)
$rvfin_{t-2\to t-1}$			4.32			0.47			0.63
			(3.01)			(2.15)			(0.76)
$rvfin_{t-13\to t-1}$			3.78			3.60**			1.24
			(2.43)			(1.79)			(1.74)
N	552	841	841	552	841	841	544	833	833

(b) Full sample, 192707–201602

	$r^{em}_{t\to t+1}$			$r^{em}_{t\to t+3}$			$r^{em}_{t\to t+12}$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
icr_t	-3.10			-3.33			-2.94		
	(2.34)			(2.57)			(2.49)		
$\widehat{icr}_t$		-4.31**			-4.51			-4.90**	
		(1.99)			(3.03)			(2.45)	
z^0_t			2.57			2.70			3.50
			(2.45)			(3.84)			(2.73)
z^+_t			-3.05			-3.25*			-3.73**
			(2.30)			(1.91)			(1.81)
pd_t			-5.96***			-6.35			-6.84**
			(2.25)			(4.08)			(2.81)
$rvfin_{t-1\to t}$			-4.19			-1.36			-0.99
			(2.60)			(2.59)			(1.29)
$rvfin_{t-2\to t-1}$			4.24			0.08			-0.15
			(2.78)			(2.45)			(1.31)
$rvfin_{t-13\to t-1}$			-3.54			-2.35			-2.98
			(2.67)			(3.58)			(3.57)
N	552	1,062	1,061	552	1,062	1,061	544	1,054	1,053

Notes: Reported are time-series predictability regression estimates of future stock market excess returns ($r^{em}_{t\to t+\tau}$) at one to twelve month horizons on the intermediary capital ratio (icr_t) in the short sample in which it is available, on the news-implied $\widehat{icr}_t$ that is available for a much longer time-series. We additionally decompose the $\widehat{icr}_t$ into the sufficient reduction projections z^0_t, z^+_t that summarize the text, the log price dividend ratio (pd_t), realized variance of financial stocks ($rvfin$) over the same month, over the prior month, and over the prior year. Explanatory variables are demeaned and scaled to unit variance. The corpus includes the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Hodrick (1992) standard errors are in parentheses.

Table 6: Summary statistics: Macroeconomic forecasting dataset

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	0.873	3.225	-16.153	-2.760	1.081	4.423	12.172	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	7.306	0.236	6.190	6.990	7.326	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	1.089	-5.401	-1.187	-0.003	1.099	7.175	602	195902–200903
Real AHE: goods	0.295	1.318	-4.700	-1.187	0.272	1.688	6.274	602	195901–200902
FedFunds	0.002	0.371	-1.560	-0.420	0.010	0.380	1.600	602	195901–200902
M1	0.012	2.188	-10.512	-2.347	-0.000	2.414	7.485	601	195902–200902
Ex rate: avg	-0.172	6.017	-21.276	-8.339	0.038	6.961	21.450	601	195901–200901
S&P 500	1.807	14.500	-91.237	-14.479	2.821	17.014	45.411	603	195901–200903
Consumer expect	-0.053	3.975	-16.500	-4.600	-0.200	4.600	22.500	603	195901–200903

Notes: Reported are summary statistics for the *Wall Street Journal* full text and for target variables in the monthly sample of Stock and Watson (2012). The corpus includes the 100,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming. To summarize the text, we report the mean of per phrase statistics, for counts c_{tj} and for inclusion indicators, $h_{tj} \equiv 1_{\{c_{tj}>\}}$. For each of the 12 categories of variables we use the headline variable as the prediction target. The transformed series are generally first differences of logarithms (growth rates) for real quantity variables, first differences for nominal interest rates, and second differences of logarithms (changes in rates of inflation) for price series.

Table 7: Forecasting macroeconomic series

(a) *Wall Street Journal* full text, 10,000 most frequent phrases

Months forward:	$\tau = 1$				$\tau = 3$				$\tau = 12$			
Folds:	Random		Rolling		Random		Rolling		Random		Rolling	
$Y^{\tau}_{t+\tau}$ \ Benchmark:	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR
IP: total	0.986	1.026	0.953	0.963*	0.856***	0.961**	0.827***	0.957***	0.697***	0.879***	0.668***	0.852***
	(0.395)	(0.801)	(0.127)	(0.086)	(0.005)	(0.033)	(0.000)	(0.006)	(0.000)	(0.000)	(0.000)	(0.000)
Emp: total	0.831***	0.967	0.817***	0.867***	0.763***	0.917***	0.751***	0.844***	0.679***	0.873***	0.640***	0.780***
	(0.001)	(0.110)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.000)	(0.000)
U: all	0.954	1.005	0.979	0.975	0.805***	0.954**	0.735***	0.911***	0.641***	0.849***	0.621***	0.825***
	(0.181)	(0.584)	(0.328)	(0.266)	(0.000)	(0.040)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)
HStarts: Total	0.677***	1.038	0.694***	0.906***	0.633***	0.997	0.658***	0.852***	0.624***	1.093	0.773***	1.018
	(0.000)	(0.943)	(0.000)	(0.000)	(0.000)	(0.439)	(0.000)	(0.000)	(0.000)	(0.998)	(0.000)	(0.885)
PMI	0.838***	1.028	1.029	1.043	0.853***	1.018	1.009	1.091	0.714***	0.952***	0.822***	0.893***
	(0.000)	(0.837)	(0.687)	(0.917)	(0.000)	(0.770)	(0.558)	(0.998)	(0.000)	(0.009)	(0.003)	(0.000)
CPI-ALL	1.090	0.996	1.134	0.957*	1.062	0.975	1.105	1.046	1.026	1.006	1.103	1.105
	(1.000)	(0.390)	(1.000)	(0.074)	(1.000)	(0.196)	(1.000)	(0.914)	(0.951)	(0.651)	(0.994)	(0.999)
Real AHE: goods	0.968	0.972	1.033	1.021	0.889***	0.964	0.992	0.961	0.616***	0.984	0.844***	0.938***
	(0.230)	(0.211)	(0.780)	(0.729)	(0.007)	(0.177)	(0.429)	(0.107)	(0.000)	(0.269)	(0.007)	(0.000)
FedFunds	0.914***	0.996	1.042	0.986	0.812***	0.974	1.038	1.025	0.647***	0.972	0.869**	0.943*
	(0.000)	(0.442)	(0.776)	(0.312)	(0.000)	(0.194)	(0.722)	(0.774)	(0.000)	(0.178)	(0.026)	(0.070)
M1	1.101	1.018	1.179	1.006	1.109	1.014	1.236	1.090	1.004	1.003	1.068	0.959
	(1.000)	(0.964)	(1.000)	(0.602)	(0.998)	(0.707)	(1.000)	(0.999)	(0.550)	(0.556)	(0.932)	(0.144)
Ex rate: avg	1.066	1.008	1.088	0.922***	1.011	1.041	1.005	0.929**	0.891***	1.067	0.849**	0.982
	(1.000)	(0.696)	(0.953)	(0.005)	(0.640)	(0.974)	(0.535)	(0.019)	(0.002)	(0.999)	(0.012)	(0.246)
S&P 500	1.030	1.008	1.072	1.067	0.937*	0.995	0.934	1.044	0.750***	0.955**	0.692***	0.939***
	(0.816)	(0.811)	(0.945)	(0.995)	(0.076)	(0.366)	(0.108)	(0.955)	(0.000)	(0.010)	(0.000)	(0.005)
Consumer expect	1.096	0.944***	1.233	0.828***	1.032	1.010	1.110	0.971	0.953	1.040	0.865**	1.017
	(1.000)	(0.006)	(1.000)	(0.000)	(0.858)	(0.639)	(0.994)	(0.212)	(0.140)	(0.987)	(0.014)	(0.707)

(b) *Wall Street Journal* full text, 100,000 most frequent phrases

Months forward:	$\tau = 1$				$\tau = 3$				$\tau = 12$			
Folds:	Random		Rolling		Random		Rolling		Random		Rolling	
$Y^{\tau}_{t+\tau}$ \ Benchmark:	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR
IP: total	0.985	1.025	0.984	0.985	0.835***	0.944**	0.847***	0.990	0.607***	0.784***	0.641***	0.834***
	(0.401)	(0.747)	(0.337)	(0.312)	(0.000)	(0.027)	(0.000)	(0.297)	(0.000)	(0.000)	(0.000)	(0.000)
Emp: total	0.762***	0.914**	0.884***	0.961	0.676***	0.841***	0.776***	0.887***	0.499***	0.676***	0.568***	0.727***
	(0.000)	(0.021)	(0.006)	(0.151)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
U: all	0.942	0.996	0.976	0.983	0.796***	0.965	0.880***	1.113	0.545***	0.772***	0.664***	0.912***
	(0.111)	(0.446)	(0.310)	(0.348)	(0.000)	(0.242)	(0.000)	(0.995)	(0.000)	(0.000)	(0.000)	(0.000)
HStarts: Total	0.555***	0.858***	0.528***	0.689***	0.512***	0.819***	0.512***	0.672***	0.504***	0.899**	0.521***	0.687***
	(0.000)	(0.005)	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.029)	(0.000)	(0.000)
PMI	0.808***	1.044	0.969	0.972	0.840***	1.067	0.937	1.002	0.800***	1.106	0.915*	0.983
	(0.000)	(0.948)	(0.306)	(0.186)	(0.000)	(0.995)	(0.133)	(0.529)	(0.000)	(0.990)	(0.054)	(0.341)
CPI-ALL	1.061	0.923***	1.054	0.846***	1.063	0.941***	1.027	0.930**	1.072	1.033	1.052	1.020
	(1.000)	(0.000)	(0.999)	(0.000)	(1.000)	(0.001)	(0.907)	(0.011)	(0.998)	(0.845)	(0.973)	(0.756)
Real AHE: goods	1.000	0.988	1.070	1.042	0.982	1.085	1.026	1.026	0.643***	1.082	0.818***	0.923***
	(0.503)	(0.336)	(0.962)	(0.874)	(0.283)	(0.990)	(0.739)	(0.828)	(0.000)	(1.000)	(0.002)	(0.000)
FedFunds	0.954**	1.053	1.166	1.123	0.880***	1.094	1.006	1.032	0.734***	1.218	0.726***	0.858***
	(0.047)	(0.929)	(1.000)	(0.993)	(0.000)	(0.988)	(0.546)	(0.815)	(0.000)	(1.000)	(0.000)	(0.000)
M1	1.039	0.887***	1.090	0.776***	1.024	0.883***	1.121	0.844***	1.036	1.032	1.070	0.941*
	(1.000)	(0.000)	(1.000)	(0.000)	(0.990)	(0.000)	(1.000)	(0.000)	(0.897)	(0.880)	(0.990)	(0.066)
Ex rate: avg	1.060	0.982	1.086	0.850***	1.082	1.118	1.059	0.933**	0.756***	0.939***	0.837***	0.984
	(1.000)	(0.205)	(0.984)	(0.000)	(1.000)	(1.000)	(0.895)	(0.047)	(0.000)	(0.007)	(0.006)	(0.259)
S&P 500	1.086	1.052	1.049	1.032	0.898**	0.962**	0.913**	1.042	0.631***	0.823***	0.614***	0.857***
	(0.999)	(0.978)	(0.888)	(0.919)	(0.028)	(0.031)	(0.028)	(0.956)	(0.000)	(0.000)	(0.000)	(0.000)
Consumer expect	1.028	0.851***	1.048	0.707***	1.064	1.027	1.054	0.905***	0.899***	0.988	0.970	1.137
	(1.000)	(0.000)	(0.986)	(0.000)	(0.988)	(0.806)	(0.954)	(0.009)	(0.010)	(0.369)	(0.296)	(1.000)

Notes: Reported are out-of-sample RMSE for HDMR that uses the lagged text plus the top 5 PCs of the Stock and Watson (2012) monthly macroeconomic series, relative to DFM-5, which uses only these PCs without the text, and relative to DMR with the same data. Diebold and Mariano (1995) p-values testing the null hypothesis that the RMSE of HDMR is larger than the benchmark's are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Nowcasting macroeconomic series

(a) *Wall Street Journal* full text, 10,000 most frequent phrases

Months forward:	$\tau = 1$				$\tau = 3$				$\tau = 12$			
Folds:	Random		Rolling		Random		Rolling		Random		Rolling	
$Y^{\tau}_{t+\tau}$ \ Benchmark:	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR
IP: total	0.956**	1.013	0.947*	0.980	0.853***	0.963**	0.802***	0.929***	0.706***	0.893***	0.653***	0.838***
	(0.033)	(0.859)	(0.061)	(0.209)	(0.000)	(0.027)	(0.000)	(0.000)	(0.000)	(0.005)	(0.000)	(0.000)
Emp: total	0.804***	0.947**	0.769***	0.868***	0.747***	0.915***	0.747***	0.862***	0.686***	0.888***	0.652***	0.797***
	(0.000)	(0.015)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.007)	(0.000)	(0.000)
U: all	0.962	1.020	0.928*	0.989	0.820***	0.974***	0.759***	0.942**	0.657***	0.890***	0.639***	0.866***
	(0.202)	(0.955)	(0.067)	(0.358)	(0.000)	(0.009)	(0.000)	(0.034)	(0.000)	(0.006)	(0.000)	(0.000)
HStarts: Total	0.655***	1.013	0.669***	0.885***	0.636***	1.020	0.651***	0.857***	0.621***	1.102	0.755***	1.001
	(0.000)	(0.631)	(0.000)	(0.000)	(0.000)	(0.788)	(0.000)	(0.000)	(0.000)	(1.000)	(0.000)	(0.512)
PMI	0.853***	1.016	1.038	1.059	0.850***	1.011	1.033	1.119	0.729***	0.965	0.842***	0.897***
	(0.000)	(0.669)	(0.756)	(0.961)	(0.000)	(0.711)	(0.692)	(1.000)	(0.000)	(0.114)	(0.008)	(0.000)
CPI-ALL	1.027	0.949**	1.171	0.936**	1.060	0.976**	1.115	1.028	1.044	1.018	1.069	1.062
	(0.970)	(0.015)	(1.000)	(0.014)	(0.998)	(0.038)	(1.000)	(0.890)	(0.990)	(0.847)	(0.977)	(0.997)
Real AHE: goods	1.036	1.019	1.051	1.008	0.910**	0.964	0.987	1.007	0.597***	0.971	0.829***	0.924***
	(0.797)	(0.872)	(0.880)	(0.627)	(0.027)	(0.175)	(0.383)	(0.580)	(0.000)	(0.216)	(0.002)	(0.000)
FedFunds	0.923*	1.023	1.069	1.063	0.831***	1.023	1.040	1.079	0.641***	0.951**	0.858**	0.795***
	(0.051)	(0.809)	(0.906)	(0.967)	(0.000)	(0.750)	(0.755)	(0.979)	(0.000)	(0.020)	(0.015)	(0.000)
M1	1.125	0.974	1.264	0.840***	1.101	1.019	1.180	0.623***	1.010	1.028	1.056	1.005
	(1.000)	(0.237)	(1.000)	(0.000)	(1.000)	(0.943)	(1.000)	(0.000)	(0.618)	(0.951)	(0.919)	(0.589)
Ex rate: avg	1.043	0.980*	1.055	0.942**	1.018	1.039	1.090	1.010	0.879***	1.058	0.918	1.053
	(0.949)	(0.086)	(0.852)	(0.023)	(0.685)	(0.957)	(0.922)	(0.635)	(0.002)	(1.000)	(0.121)	(0.971)
S&P 500	1.012	0.991	1.032	1.024	0.939*	1.000	0.948	1.037	0.763***	0.970	0.715***	0.901***
	(0.654)	(0.374)	(0.776)	(0.806)	(0.082)	(0.496)	(0.185)	(0.874)	(0.000)	(0.129)	(0.000)	(0.003)
Consumer expect	1.171	0.995	1.265	0.936**	1.007	0.990	1.115	0.948	0.942**	1.028	0.895**	1.018
	(1.000)	(0.409)	(1.000)	(0.014)	(0.644)	(0.219)	(0.994)	(0.106)	(0.031)	(0.886)	(0.043)	(0.728)

(b) *Wall Street Journal* full text, 100,000 most frequent phrases

Months forward:	$\tau = 1$				$\tau = 3$				$\tau = 12$			
Folds:	Random		Rolling		Random		Rolling		Random		Rolling	
$Y^{\tau}_{t+\tau}$ \ Benchmark:	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR	DFM-5	DMR
IP: total	0.956**	1.015	0.986	1.013	0.813***	0.932**	0.831***	0.974*	0.608***	0.788***	0.645***	0.844***
	(0.012)	(0.734)	(0.336)	(0.640)	(0.000)	(0.024)	(0.000)	(0.098)	(0.000)	(0.000)	(0.000)	(0.000)
Emp: total	0.764***	0.931**	0.879***	1.006	0.666***	0.854***	0.782***	0.923***	0.521***	0.717***	0.589***	0.753***
	(0.000)	(0.015)	(0.004)	(0.569)	(0.000)	(0.001)	(0.000)	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)
U: all	0.942*	0.998	0.955	1.008	0.788***	0.963	0.880***	1.131	0.544***	0.795***	0.692***	0.970
	(0.098)	(0.469)	(0.154)	(0.581)	(0.000)	(0.197)	(0.001)	(0.999)	(0.000)	(0.000)	(0.000)	(0.144)
HStarts: Total	0.534***	0.832***	0.523***	0.698***	0.516***	0.833***	0.503***	0.670***	0.454***	0.811***	0.499***	0.661***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)
PMI	0.822***	1.025	0.963	0.981	0.853***	1.075	0.947	1.019	0.817***	1.127	0.932*	0.992
	(0.000)	(0.757)	(0.266)	(0.286)	(0.000)	(0.984)	(0.187)	(0.732)	(0.000)	(1.000)	(0.090)	(0.418)
CPI-ALL	0.969*	0.843***	1.061	0.739***	1.048	0.923***	1.054	0.928***	1.062	1.010	1.052	1.021
	(0.065)	(0.000)	(0.997)	(0.000)	(0.978)	(0.000)	(0.993)	(0.005)	(0.991)	(0.664)	(0.971)	(0.783)
Real AHE: goods	1.024	1.000	1.109	1.057	1.022	1.096	1.057	1.085	0.633***	1.088	0.810***	0.914***
	(0.806)	(0.501)	(0.996)	(0.937)	(0.725)	(0.993)	(0.927)	(0.997)	(0.000)	(0.998)	(0.001)	(0.000)
FedFunds	0.945*	1.064	1.131	1.123	0.870***	1.107	1.016	1.067	0.713***	1.169	0.764***	0.887***
	(0.090)	(0.996)	(1.000)	(0.996)	(0.000)	(0.999)	(0.634)	(0.960)	(0.000)	(1.000)	(0.000)	(0.010)
M1	1.084	0.854***	1.121	0.713***	1.048	0.922***	1.108	0.624***	1.042	1.046	1.055	0.953*
	(0.999)	(0.000)	(1.000)	(0.000)	(1.000)	(0.000)	(1.000)	(0.000)	(0.904)	(0.922)	(0.961)	(0.065)
Ex rate: avg	1.034	0.949**	1.031	0.862***	1.091	1.121	1.052	0.927**	0.764***	0.950**	0.872**	1.009
	(0.936)	(0.027)	(0.795)	(0.000)	(0.983)	(1.000)	(0.873)	(0.029)	(0.000)	(0.010)	(0.026)	(0.645)
S&P 500	1.018	0.996	0.996	0.974	0.911**	0.980	0.895**	0.999	0.667***	0.870***	0.642***	0.866***
	(0.672)	(0.443)	(0.460)	(0.191)	(0.017)	(0.158)	(0.012)	(0.485)	(0.000)	(0.000)	(0.000)	(0.000)
Consumer expect	1.030	0.834***	1.013	0.695***	1.038	1.005	1.044	0.852***	0.910***	1.001	0.991	1.129
	(0.980)	(0.000)	(0.747)	(0.000)	(0.887)	(0.554)	(0.934)	(0.001)	(0.006)	(0.509)	(0.435)	(1.000)

Notes: Reported are out-of-sample RMSE for HDMR that uses the contemporaneous text plus the top 5 PCs of the Stock and Watson (2012) monthly macroeconomic series, relative to DFM-5, which uses only these PCs without the text, and relative to DMR with the same data. Diebold and Mariano (1995) p-values testing the null hypothesis that the RMSE of HDMR is larger than the benchmark's are in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.