Text Selection

Bryan Kelly Asaf Manela Alan Moreira*

PRELIMINARY AND INCOMPLETE. PLEASE DO NOT CIRCULATE.

March 2018

Abstract

Text data is inherently high-dimensional, which makes machine learning 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 text 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 backfill central financial variables to historical periods using newspaper coverage, and find that it substantially improves out-of-sample fit relative to alternative state-of-the-art approaches. This advantage increases with the sparsity of the text.

Keywords: Text analysis, machine learning, high dimensional selection, 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 seminar participants at École polytechnique fédérale de Lausanne, Hebrew University, IDC Herzliya, INSEAD, and Tel-aviv University, and of the CITE Conference in Chicago.

1 Introduction

Digital text is increasingly available to social scientists in the form of newspapers, blogs, tweets, regulatory filings, congressional records and more (Gentzkow, Kelly, and Taddy, 2017). 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.

We develop the *hurdle distributed multiple regression* (HDMR), a highly scalable approach to inference from big counts data. An econometrician confronted with modeling counts data, may first consider using a multinomial logistic (softmax) regression, but this approach is computationally intractable in many text-related applications when the number of categories is very large. Taddy (2015) has shown that one can overcome this dimensionality problem by approximating the multinomial with cleverly shifted independent Poisson regressions, one for each word, which can be distributed across parallel computation units. We replace each Poisson regression with a two-part hurdle model: a selection equation to model the text producer's choice whether or not to include a particular word, and a positive counts model for their choice of how many times this word is repeated in a document.

HDMR's main advantage over Taddy (2015)'s *distributed multinomial regression* (DMR) is that it allows us to explicitly model the extensive margin decision of which words to use in a particular body of text. HDMR brings the selection methodology of Heckman (1979) to a high dimensionalit setting. We use the hurdle model of Mullahy (1986) to model selection because its two parts can be estimated independently, and therefore distributed further at essentially no additional cost relative to DMR.

Economics suggests that text data is likely to be sparse. For example, newspaper publishers cater to a boundedly rational reader by selecting the text to concisely describe the most interesting and relevant topics to their readers (Gentzkow and Shapiro, 2006). The basic idea is that there are large cost associated with increasing the dimensionality of the signal space– i.e. you cannot write

about a particular topic just a little bit. In order for the signal to be interpretable there needs to be a minimum amount of context and new terms need to be defined and explained. Gabaix (2014) develops this idea in a model and show that 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.[1]

The fitted model can serve as a basis for exploring the relationships between words and variables of interest, for reducing dimension into a parsimonious set of factors, and for prediction via an inverse regression. Whereas the multinomial inverse regression suggested by Taddy (2013) relies on a single sufficient reduction projection of the counts on each explanatory variable, HDMR produces two such sufficient reductions: one for word inclusion (the extensive margin) and the other for repetition (the intensive margin). While theoretical arguments suggest that the extensive margin would be particularly informative, in the end it comes down to whether modeling this selection decision helps with forecasting. Here we apply this framework to one example to illustrate our methodology.

We apply this framework to backfill a central financial variable to historical periods using *Wall Street Journal* coverage. Specifically, we extend back to 1945 the intermediary capital ratio (ICR), which has recently been shown to explain cross-sectional variation in expected returns across a wide array of asset classes, but is only available starting in 1970 (He, Kelly, and Manela, 2017).

We find that HDMR substantially improves out-of-sample fit both relative to DMR and to support vector regression (Vapnik, 2000) previously-used for this task by Manela and Moreira (2017). 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 and the document term matrix becomes dense, the text is better described by a selection free multinomial model. Importantly, we show that simply omitting infrequent words leads to a large deterioration in prediction accuracy.

[1]The selection decision is also important in contexts where the writer wants to transmit it's type to readers. Politicians carefully select phrases that resonate with voters in congressional speech (Gentzkow, Shapiro, and Taddy, 2017). The fixed cost of using censored or socially taboo words may generate further sparsity (Michel, Shen, Aiden, Veres, Gray, Pickett, Hoiberg, Clancy, Norvig, Orwant, et al., 2011).

Intuitively, a better model for text selection leads to better out-of-sample predictions.

The backfilled news-implied intermediary capital ratio series provides for more powerful tests that support a central prediction of intermediary asset pricing theory—times when intermediaries are highly capitalized are "good times," when these marginal investors demand a relatively low risk premium to hold investment assets. These findings 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 price-dividend ratio and volatility. HDMR provides an efficient way to identify these phrases and their relative weights in a data driven approach while avoiding overfit.

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). We then introduce in Section 2.1 our key contribution, a model for the extensive margin, which we refer as Hurdle DMR. Section 2.2 shows how regularization allows our methodology handle a feature space many times larger the number of observations. Section 2.3 shows how to recover the text-factors that better tracks the variable of interest, i.e. a weighted bag-of-words. Section 2.4 shows how to use these factors for prediction, and how these text-factors they are sufficient statistics for the information content of text with respect to the variable of interest. We end the paper in Section 3 with an application of our methodology to extend a key macroeconomic variable back in time using text data from the *Wall Street Journal*.

2 A model for text selection

Let $\boldsymbol{c}_i$ be a vector of counts in d categories, summing to $m_i = \sum_j c_{ij}$, and let $\boldsymbol{v}_i$ be a p_v-vector of covariates on observation $i = 1 \dots n$. In a text application, c_{ij} are counts of word or phrase (n-gram) j in document i with attributes $\boldsymbol{v}_i$ (author, date, etc.). An econometrician confronted

with modeling counts data, may first consider using a multinomial logistic regression:

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

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

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

When the number of categories d is very large, as is the case in many natural language processing applications[2], estimating the parameters of the multinomial, $\boldsymbol{\alpha} = [\alpha_j]$ and $\boldsymbol{\varphi} = [\varphi_{ij}]$, is computationally costly. 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 $\lambda_{ij} = m_i e^{\eta_{ij}}$ 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 $\sum_{j=1}^{d} e^{\eta_{ij}}$ and allows for fast and scalable distributed estimation.

[2]For example, our application in Section XX has a vocabulary of more than four hundred thousand words, i.e. $d > 400,000$.

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 $\boldsymbol{z}_i$. Put differently, within this model $\boldsymbol{z}_i$ is a sufficient statistic that summarizing all of the content that the text has for predicting $\boldsymbol{v}_i$ (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 multiple regression

In many cases, 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 packages use sparse matrices to store word counts efficiently. As eluded to above, the economics of natural language selection provides many reasons for this sparsity.

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 multiple regression (HDMR)*:

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

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

$$c_{ij}^* = \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right) + \varepsilon_{ij} > 0, \tag{10}$$

$$c_{ij} = h_{ij}c_{ij}^*. \tag{11}$$

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 repetition of positive counts 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.

Let Π_0 denote the discrete density for zeros

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

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

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

Natural choices for P^+ are the Poisson and the negative binomial, truncated at zero. Combining terms, noting that $h_{ij} = \mathbf{1}\left(c_{ij} > 0\right)$, the joint density is

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

The negative log likelihood takes a convenient form

$$l\left(\boldsymbol{\alpha}, \boldsymbol{\varphi}, \boldsymbol{\kappa}, \boldsymbol{\delta}|\boldsymbol{c}, \boldsymbol{v}, \boldsymbol{w}\right) = \sum_{j=1}^{d} l\left(\alpha_j, \varphi_j, \kappa_j, \delta_j|\boldsymbol{c}_j, \boldsymbol{v}, \boldsymbol{w}\right), \tag{12}$$

$$l\left(\alpha_j, \boldsymbol{\varphi}_j, \kappa_j, \boldsymbol{\delta}_j | \boldsymbol{c}_j, \boldsymbol{v}, \boldsymbol{w}\right) = l^0\left(\kappa_j, \boldsymbol{\delta}_j | \boldsymbol{h}_j, \boldsymbol{w}\right) + l^+\left(\alpha_j, \boldsymbol{\varphi}_j | \boldsymbol{c}_j, \boldsymbol{v}\right), \tag{13}$$

$$l^0\left(\kappa_j, \boldsymbol{\delta}_j | \boldsymbol{h}_j, \boldsymbol{w}\right) = -\sum_{i|h_{ij}=0}^{n} \log \Pi_0\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right) - \sum_{i|h_{ij}=1}^{n} \log\left[1 - \Pi_0\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)\right], \tag{14}$$

$$l^+\left(\alpha_j, \boldsymbol{\varphi}_j | \boldsymbol{c}_j, \boldsymbol{v}\right) = -\sum_{i|h_{ij}=1}^{n} \log P^+\left(c_{ij}; \lambda\left(\alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j\right)\right). \tag{15}$$

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 (13) into two parts that can be estimated independently, which facilitates further parallelization.[3] Specifically, the parameters that govern inclusion (κ_j and $\boldsymbol{\delta}_j$) only depend on word j indicators $\boldsymbol{h}_j$ and on the covariates $\boldsymbol{w}$, whereas the parameters that govern repetition (α_j and $\boldsymbol{\varphi}_j$) only depend on positive word counts $\boldsymbol{c}_j > 0$ and the covariates $\boldsymbol{v}$ and can be estimated separately in the subsample of positive j counts.

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 of the two parts of the hurdle

$$\hat{\kappa}_j, \hat{\boldsymbol{\delta}}_j = \underset{\kappa_j, \boldsymbol{\delta}_j}{\arg\min}\, l^0\left(\kappa_j, \boldsymbol{\delta}_j | \boldsymbol{h}_j, \boldsymbol{w}\right) + n\lambda^0 \sum_{k=1}^{p_w} |\delta_{jk}| \quad \text{where } \lambda^0 \geq 0, \tag{16}$$

$$\hat{\alpha}_j, \hat{\boldsymbol{\varphi}}_j = \underset{\alpha_j, \boldsymbol{\varphi}_j}{\arg\min}\, l^+\left(\alpha_j, \boldsymbol{\varphi}_j | \boldsymbol{c}_j, \boldsymbol{v}\right) + n^+\lambda^+ \sum_{k=1}^{p_v} |\varphi_{jk}| \quad \text{where } \lambda^+ \geq 0. \tag{17}$$

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).[4] Because the model for positive counts

[3] 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.

[4] 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

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. In Appendix B we show how to specify the two parts of the hurdle as GLMs.

2.3 Sufficient reduction projections

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

$$\mathrm{p}\left(h_{ij}|\boldsymbol{w}_i\right) = \left[\Pi_0\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)\right]^{1-h_{ij}}\left[1-\Pi_0\left(\kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j\right)\right]^{h_{ij}} \tag{18}$$

with a logit link,

$$\log\left(\left(1-\Pi_0\right)/\Pi_0\right) = \kappa_j + \boldsymbol{w}_i'\boldsymbol{\delta}_j + \log m_i. \tag{19}$$

The distribution of the latent positive counts c_{ij}^* is assumed to be Positive Poisson,

$$\mathrm{p}\left(c_{ij}|\boldsymbol{v}_i, h_{ij}=1\right) = Po^+\left(c_{ij};\lambda_{ij}\left(\boldsymbol{v}_i, m_i\right)\right) = \frac{Po\left(c_{ij};\lambda_{ij}\left(\boldsymbol{v}_i, m_i\right)\right)}{1-Po\left(0;\lambda_{ij}\left(\boldsymbol{v}_i, m_i\right)\right)} = \frac{\lambda_{ij}^{c_{ij}}e^{-\lambda_{ij}}}{c_{ij}!\left(1-e^{-\lambda_{ij}}\right)}, \tag{20}$$

with log intensity

$$\log\lambda_{ij} = \alpha_j + \boldsymbol{v}_i'\boldsymbol{\varphi}_j + \log m_i. \tag{21}$$

We shift both (19) and (21) by the log of the total number of words in the document m_i to account for the fact that both probability of both inclusion and repetition increase in a larger document. In the Appendix, we show how to estimate both parts of the hurdle as a GLM, which allow 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 positive Poisson-log model for*

paths as in `gamlr` (Taddy, 2017).

positive counts, 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. Specifically,

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

Proof. That $\boldsymbol{\varphi c}_i$ is a sufficient statistic for $\boldsymbol{v}_i$ as in the DMR case follows from the fact that a sufficient statistic for a distribution is also sufficient for its truncated version (Tukey, 1949). 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

$$\mathrm{p}\left(\boldsymbol{c}_i | \boldsymbol{v}_i, \boldsymbol{w}_i\right) = \mathrm{p}\left(\boldsymbol{h}_i \circ \boldsymbol{c}_i | \boldsymbol{v}_i, \boldsymbol{w}_i\right) = \phi\left(\boldsymbol{c}_i\right) \psi\left(\boldsymbol{h}_i\right) a\left(\boldsymbol{w}_i, m_i\right) b\left(\boldsymbol{v}_i, m_i\right) \exp\left\{\boldsymbol{w}_i' \boldsymbol{\delta h}_i + \boldsymbol{v}_i' \boldsymbol{\varphi c}_i\right\}, \quad (22)$$

where

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

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

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

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

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. □

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 pipeline. 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 $z^0_{iy} = \boldsymbol{\delta}_y \boldsymbol{h}_{iy}$ 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^0_{iy} and z^+_{iy} correspond to partial associations as well. Conditional on the parameters, z^0_{iy} 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)

$$\mathbb{E}\left[v_{iy}\right] = \beta_0 + \left[z^0_{iy}, z^+_{iy}, \boldsymbol{w}_{i,-y}, \boldsymbol{v}_{i,-y}, m_i\right]' \boldsymbol{\beta} \tag{23}$$

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 target variable is included only in $\boldsymbol{w}$ ($\boldsymbol{v}$), one would only use z^0_{iy} (z^+_{iy}) in the forward regression.

3 Application: Backfilling the intermediary capital ratio

Recent empirical work finds empirical support for intermediary asset pricing theories (He and Krishnamurthy, 2013; Brunnermeier and Sannikov, 2014). 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, 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 preclude 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 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 counts 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 $\boldsymbol{c}_t$ includes 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).[5] This ratio is our prediction target and is therefore the first element of both covariates 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 over annualized dividend ratio pd_t from CRSP, and log realized variance rv_t which is the sum of squared daily market returns over month t. Table 1 reports summary statistics for all variables.

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 there is a continuous variable

[5]The ICR is available on Asaf Manela's website: http://apps.olin.wustl.edu/faculty/manela/data.html

enters the selection equation but can be excluded from 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.

Eisensee and Stromberg (2007) suggest news pressure as an instrument for the television coverage of natural disasters, which is available starting 1967. Because we wish to backfill the ICR as far back as possible, we use a variant of news pressure suggested in Manela (2014), which is the mean monthly number of pages in the first section (A) of the *Wall Street Journal.* The idea for this selection shifter is that on high news pressure days, such as election periods or major sporting events, the first section is larger to account for both more news and more ad space. On such days, we expect the average phrase to be crowded out of the front page, so that on the margin, the probability of inclusion is diminished. We therefore include the monthly average number of pages in section A, $biznewspres_t$ in the inclusion equation (in $\boldsymbol{w}_t$). By excluding it from the repetition equation we assume that the editor mostly picks topics that are newsworthy given both news pressure and its budget constraint, but that conditional on topic inclusion, the number of words per article is relatively constant. This is obviously an approximation, and unlikely to be universally true. To the extent it is wrong, we would expect this exclusion restriction to hurt the out-of-sample performance of our model.

3.2 Sparsity and out-of-sample fit

A key choice in the data cleaning 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 used 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 about 50

percent (from 1.4 to 0.7 percentage points). We can see that when only a few hundred words are included in the sample and the text, both DMR and HDMR generate a similar improvement, but as rarer phrases are included in the sample, and selection play a bigger role, the benefit from using HDMR increases. The advantage is hump-shaped, and eventually, as rarely used phrases enter the sample, the out-of-sample fit of both text models suffers.

Because our data involves using data in one time period to predict out-of-sample in a different time, cross-validation with random fold selection 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 "sub-prime mortgage" appears often in the period around the 2008 financial crisis where intermediary capital was low, but not in the 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 a serial variant of cross-validation, where instead of randomly splitting the sample into folds, we split it into 10 serial folds. The first fold is 1970-01 to 1974-07, then 1974-08 to 1979-02, and so on. Consistent with our concern, all models perform somewhat worse when evaluated based on serial cross-validation fit. Interestingly, we find that the advantage of HDMR over DMR is somewhat larger in this case.

3.3 News-implied intermediary capital ratio, 1945–2016

Having established that our model can produce quite good out-of-sample fit, we use it to backfill the intermediary capital ratio back to the January 1946, the first month when *biznewspres* is available. Figure 4 shows that the intermediary capital ratio predicted by HDMR closely follows the actual one in the period when the actual one exists, 1970–2016. The realized variance and 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 rv and pd important.

3.4 Explaining the text

To better understand what HDMR does, it is useful to examine phrase loadings on the covariates. Because these coefficient matrices are high-dimensional, we simply report in Table 3 the phrases with the top most positive and negative predicted values on each variable. Where the predicted values are given by each phrase loading multiplied by the square-root of the mean count for the phrase, i.e. $\bar{c}_j$. This predicted value better reflect the relative importance of a particular phrase as it downweights phrases that are very rare, and therefore matter little even if they have high loadings (Airoldi and Bischof, 2016; Taddy, 2016).

Regularization results in over 57 percent of all coefficients being exactly zero and helps HDMR avoid overfit. Some phrases, such as the positively associated "nation recoveri" and the negatively associated "hedg fund," capture fairly robust features of the data. Still, many 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.

Realized variance (rv) is higher than average when the front page mentions the market prices of commodities ("bushel wheat"), fixed income securities ("yr trea") and stocks ("stock nyse"). The price-dividend ratio (pd) is higher when technology is covered prominently ("technolog journal" "nasdaq"), but low when coverage focuses on employment ("labor letter", "jobless marri", "hour earn").

Loadings on the inclusion instrument ($biznewspres$) suggest that the first section of the *Journal* is thicker when the federal government generates news ("washington wire") as would be the case during election periods. During these periods, coverage of commodities is crowded out of the front page ("bushel wheat", "futur barrel", "commod oil").

3.5 Focusing on a single phrase for intuition

For a better intuitive understanding of how these 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 realized variance and the price-dividend. The negative coefficient on *biznewspres* means that above average business news pressure crowds out financial crisis coverage from the front page. The last column shows that a Poisson regression (DMR) treats inclusion and repetition as a single object, and does not assign any weight to news pressure, even though it is included in the (regularized) regression.

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:

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

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 predict 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, whereas DMR does not separate inclusion from repetition.

3.6 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 backfilled news-implied intermediary capital ratio allows us to test this prediction in a larger sample that spans the entire post-war period. 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^0_{t-1}, z^+_{t-1}, and covariates rv_{t-1}, pd_{t-1} and $biznewspres_{t-1}$. Because such regressions use overlapping observations, we use the standard Hodrick (1992) correction to the standard errors. For comparison, we report similar regressions where the single DMR projection z^{dmr}_{t-1} is used instead. Results are reported in Table 5.

We find that the inclusion projection z^0_{t-1} is a strong predictor of future market returns, over and above the price-dividend ratio, but that the repetition projection z^+_{t-1} is only marginally statistically significant, as is the case for z^{dmr}_{t-1}.

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 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 backfill central financial variables to historical periods using newspaper coverage, and find that it substantially improves out-of-sample fit relative to alternative state-of-the-art approaches. 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 percent change in the measure of fit (%Δ).

The first benchmark is DMR, which is provided with the same covariates and text. The improvement from modeling selection with HDMR is a 10 percent reduction in out-of-sample root mean squared error, from 83 to 75 basis points. This is 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 83 basis point RMSE, which almost identical to the 82 bp of DMR.

The last benchmark we consider is support vector regression (SVR), which Manela and Moreira (2017) use for a similar backfilling 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 considerably on an SVR without text, but its 99 basis points error rate is much larger than that of HDMR.

Figure 2b reports serial cross-validation results, and finds an even larger improvement 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 6 shows that the mean distribution of phrase counts is now much less concentrated at zero, even when we include the top 500,000 most frequent phrases. Note that some individual phrases still exhibit many more zeros than implied by a Poisson.

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. This results 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 56 percent (121 to 53 basis points) relative to the No Text benchmark, and by 19 percent (65 to 53 bp) relative to DMR.

B Estimation Details

To apply the coordinate descent algorithms developed by Friedman, Hastie, and Tibshirani (2010) to our selection model, we frame the hurdle model as a GLM. Its first part, the model for zeros, is a simple Binomial-logit (McCullagh and Nelder, 1989). Its second part, however, is a Positive Poisson (truncated at zero). We next show how to map the Positive Poisson to a GLM.

B.1 Positive Poisson as a GLM

The Positive (zero-truncated) Poisson density with intensity $\lambda = e^{\eta}$ is

$$p(y;\lambda) = \frac{\lambda^{y} e^{-\lambda}}{y!\left(1-e^{-\lambda}\right)} = \exp\left(y\eta - b(\eta) - \log y!\right),$$

with

$$b(\eta) = \log\left(e^{\lambda}-1\right) = \log\left(e^{e^{\eta}}-1\right).$$

The linear predictor μ is

$$\mu = E[y|\lambda] = b'(\eta) = \frac{\lambda}{1-e^{-\lambda}} = \frac{e^{\eta}}{1-e^{-e^{\eta}}} = g^{-1}(\eta).$$

For GLM we need η to be a one-to-one function $g(\cdot)$ of the linear predictor μ

$$\eta = \alpha + v'\varphi = g(\mu)$$

The following Lemma proves useful

Lemma 1. *Suppose* λ *such that* $e^{\lambda} \neq 1$, $\mu \neq 0$, *and*

$$\mu = \frac{\lambda}{1 - e^{-\lambda}}$$

then λ *is a one-to-one function* $f(\cdot)$ *of* μ *given by*

$$\lambda = f(\mu) = \mu + \mathcal{W}\left(-e^{-\mu}\mu\right),$$

where $\mathcal{W}$ *is Lambert's W function (also known as the omega function or product log).*

As a corollary,

$$\eta = g(\mu) = \log\left(\mu + \mathcal{W}\left(-e^{-\mu}\mu\right)\right). \tag{24}$$

Fast algorithms for calculating Lambert's W are readily available.

The derivative of the linear predictor w.r.t η is

$$\frac{d\mu}{d\eta} = \mu\left[1 - \mu e^{-\lambda}\right] = \mu\left[1 - \mu e^{-e^{\eta}}\right].$$

The log likelihood of each realized observation y as a function of μ is

$$\ell(\mu, y) = \log p(y; f(\mu)) = \begin{cases} y \log f(\mu) - \log\left(e^{f(\mu)} - 1\right) - \log y! & y > 1, \mu > 1 \\ -\infty & y > 1, \mu \downarrow 1 \\ \log\left[-\mathcal{W}\left(-e^{-\mu}\mu\right)\right] = \log\mu - f(\mu) & y = 1, \mu > 1 \\ 0 & y = 1, \mu \downarrow 1 \end{cases}$$

and deviance is

$$2\left[l_i(y, y) - l_i(\mu, y)\right] = 2 \times \begin{cases} y \log f(y) - \log\left(e^{f(y)} - 1\right) - y \log f(\mu) + \log\left(e^{f(\mu)} - 1\right) & y > 1, \mu > 1 \\ +\infty & y > 1, \mu \downarrow 1 \\ f(\mu) - \log\mu & y = 1, \mu > 1 \\ 0 & y = 1, \mu \downarrow 1 \end{cases} \tag{25}$$

Because for large y or μ, the exponents in (25) blow up, it is numerically better to evaluate the $\log(e^x - 1)$ terms as $x + \log(1 - e^{-x})$.

References

Airoldi, Edoardo M., and Jonathan M. Bischof, 2016, Improving and evaluating topic models and other models of text, *Journal of the American Statistical Association* 111, 1381–1403.

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

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

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.

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

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

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

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.

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.

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

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.

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

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.

——— , 2016, Comment on improving and evaluating topic models and other models of text, *Journal of the American Statistical Association* 111, 1403–1405.

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

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

Tukey, John W, 1949, Sufficiency, truncation and selection, *Annals of Mathematical Statistics* 20, 309–311.

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

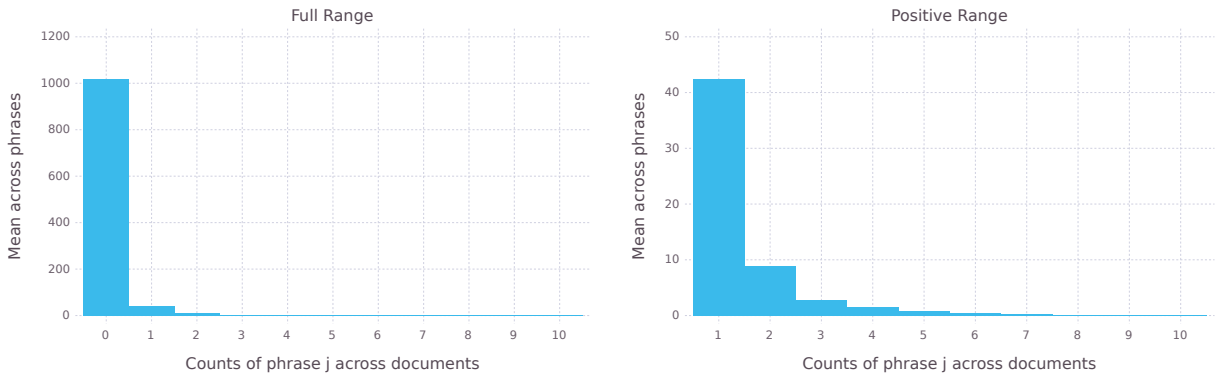

Figure 1: Mean distribution of WSJ front page articles monthly phrase counts

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. 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. 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.

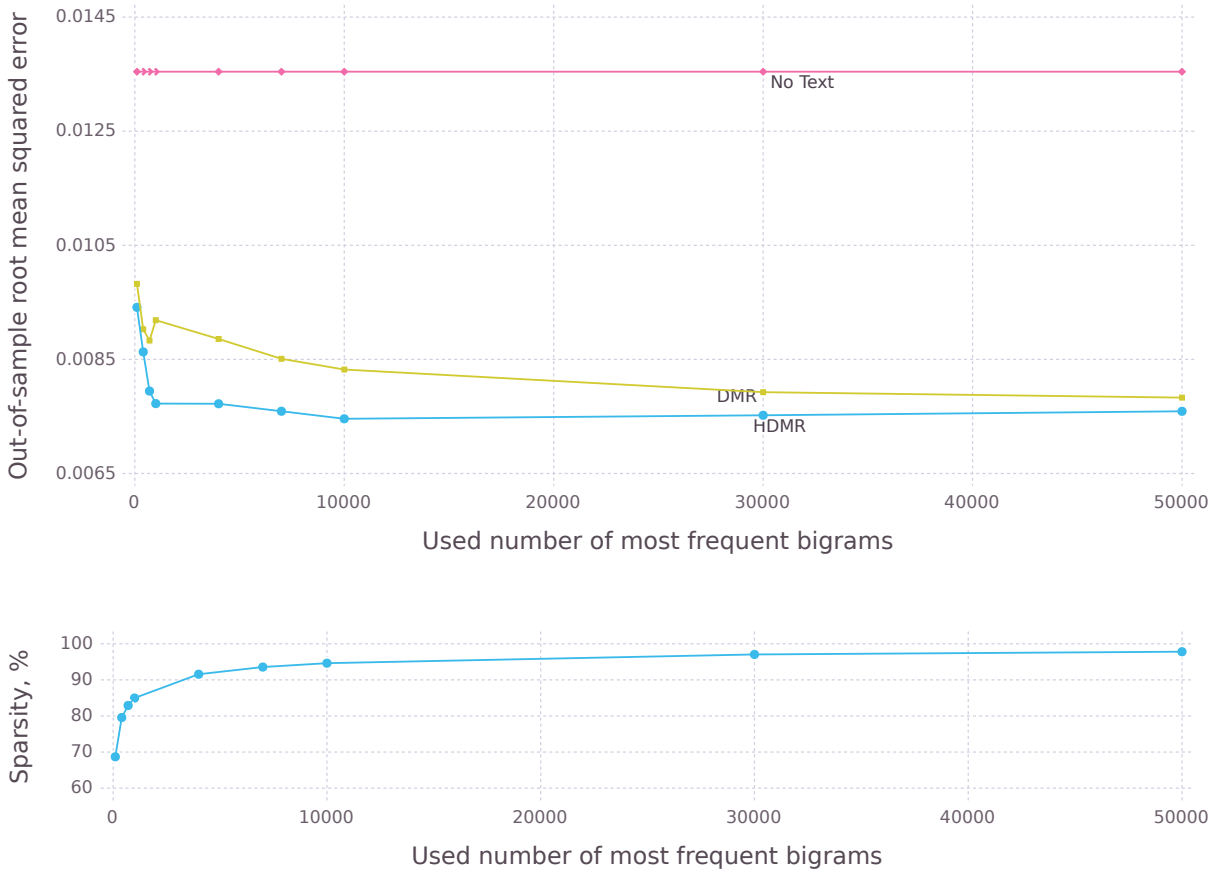

Figure 2: Predicting the intermediary capital ratio with text and covariates

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 log realized variance (rv_t), the log price dividend ratio (pd_t), business news pressure ($biznewspres_t$), 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 multiple 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. The bottom panel shows how sparsity increases with this choice, where sparsity is the fraction of zero phrase counts in the corpus.

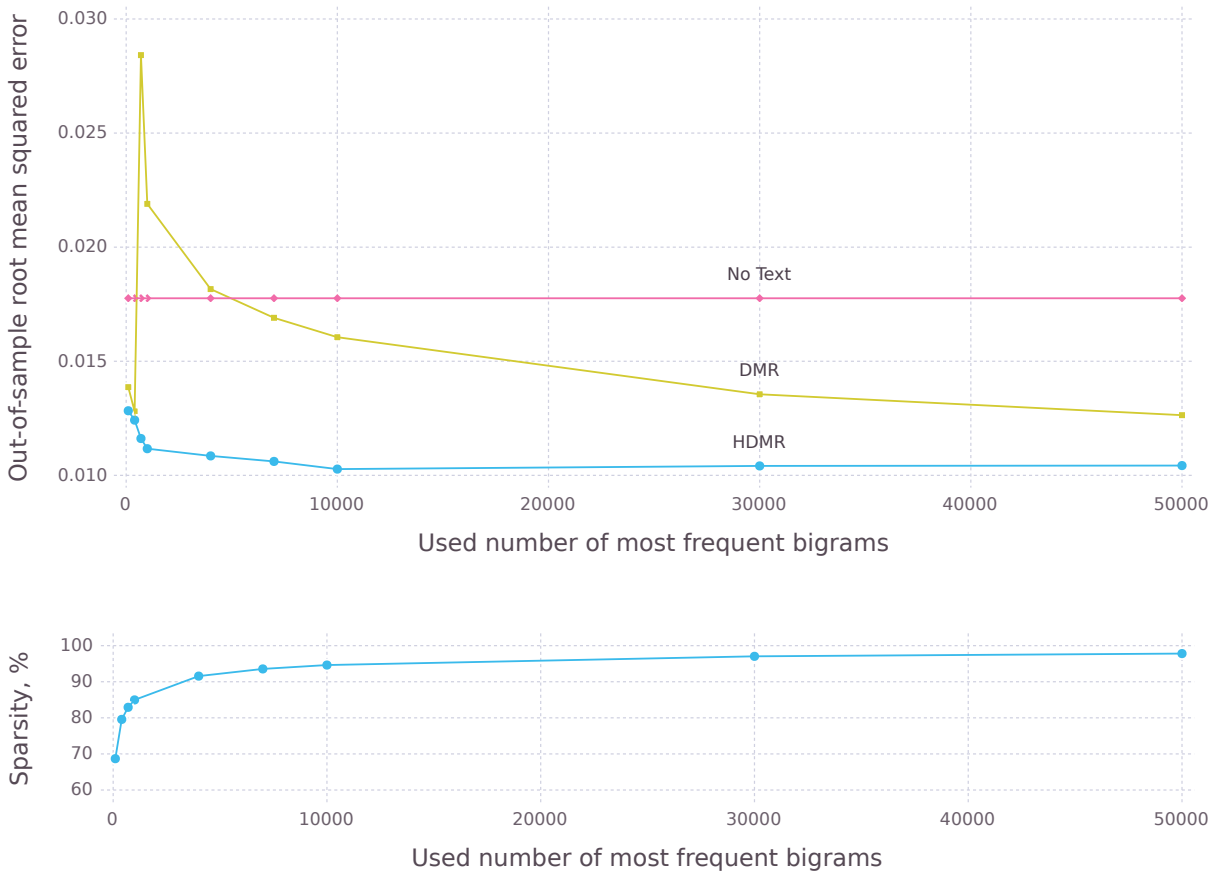

Figure 3: Predicting the intermediary capital ratio with text and covariates: Serial cross-validation

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 log realized variance (rv_t), the log price dividend ratio (pd_t), business news pressure ($biznewspres_t$), 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 with serial folds, each constituting a distinct sub-period. Our proposed model, the hurdle distributed multiple 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. The bottom panel shows how sparsity increases with this choice, where sparsity is the fraction of zero phrase counts in the corpus.

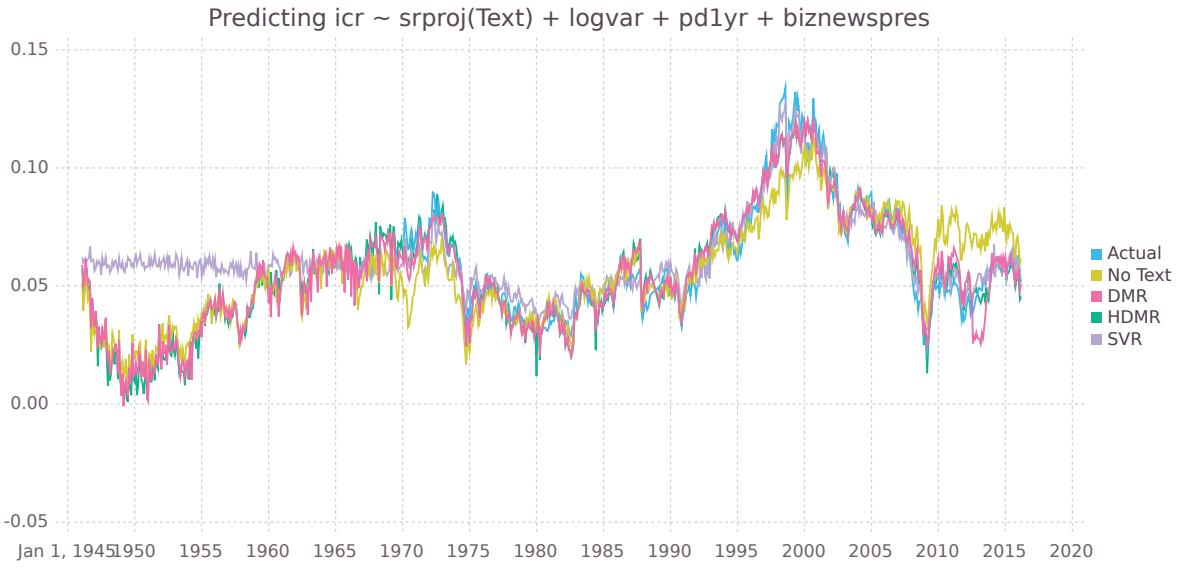

Figure 4: Backfilling the intermediary capital ratio with text and covariates

Notes: The figure shows the predicted intermediary capital ratio ($\widehat{icr}_t$) using log realized variance (rv_t), the log price dividend ratio (pd_t), and monthly WSJ front page phrase counts, over the extended sample, January 1946 to February 2016. The intermediary capital ratio is only available starting January 1970. Our proposed model, the hurdle distributed multiple regression (HDMR) is compared with two 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).

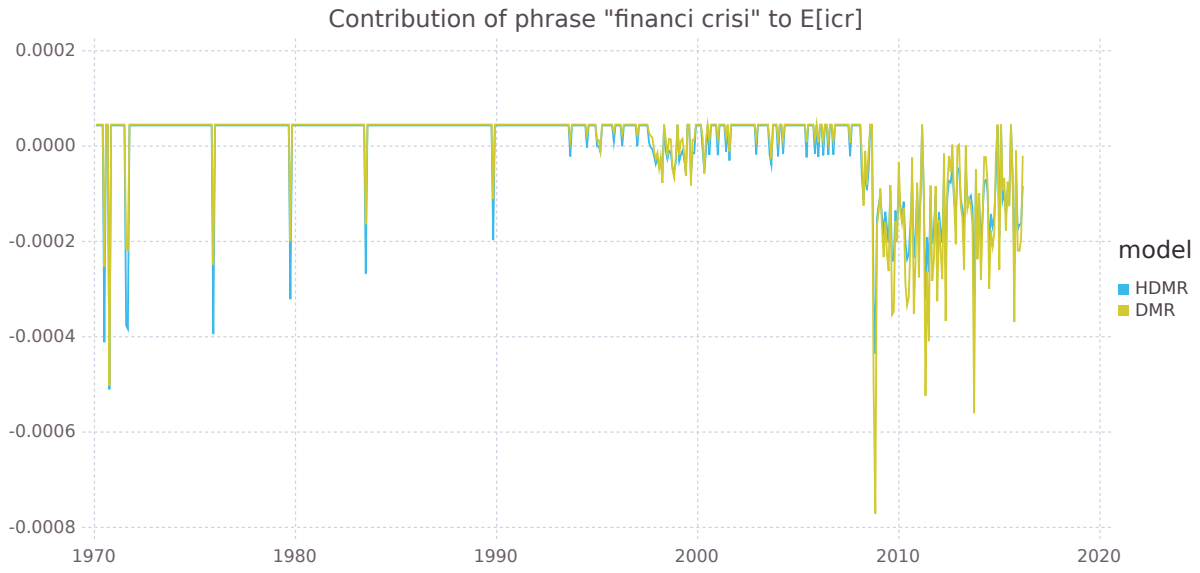

Figure 5: Focusing on a single phrase for intuition

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 multiple 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.

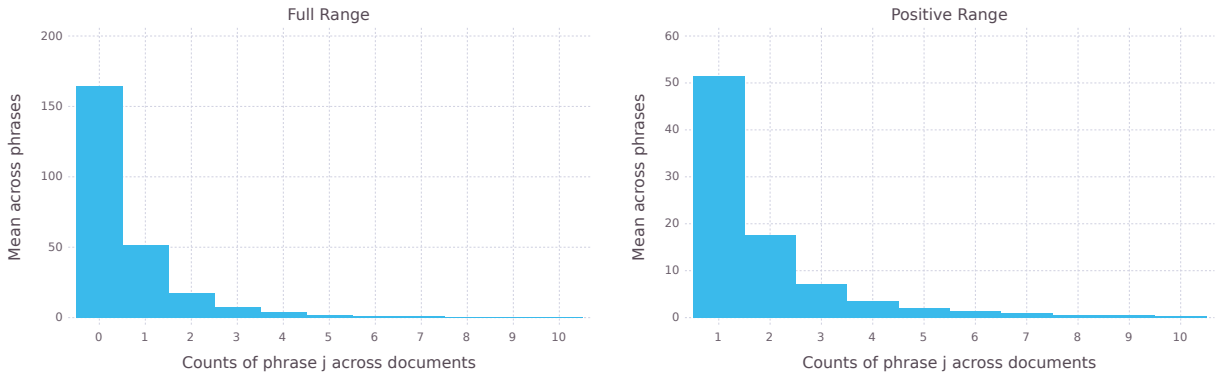

Figure 6: Mean distribution of full WSJ monthly phrase counts

Notes: The figure shows the 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. The corpus includes the 500,000 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.

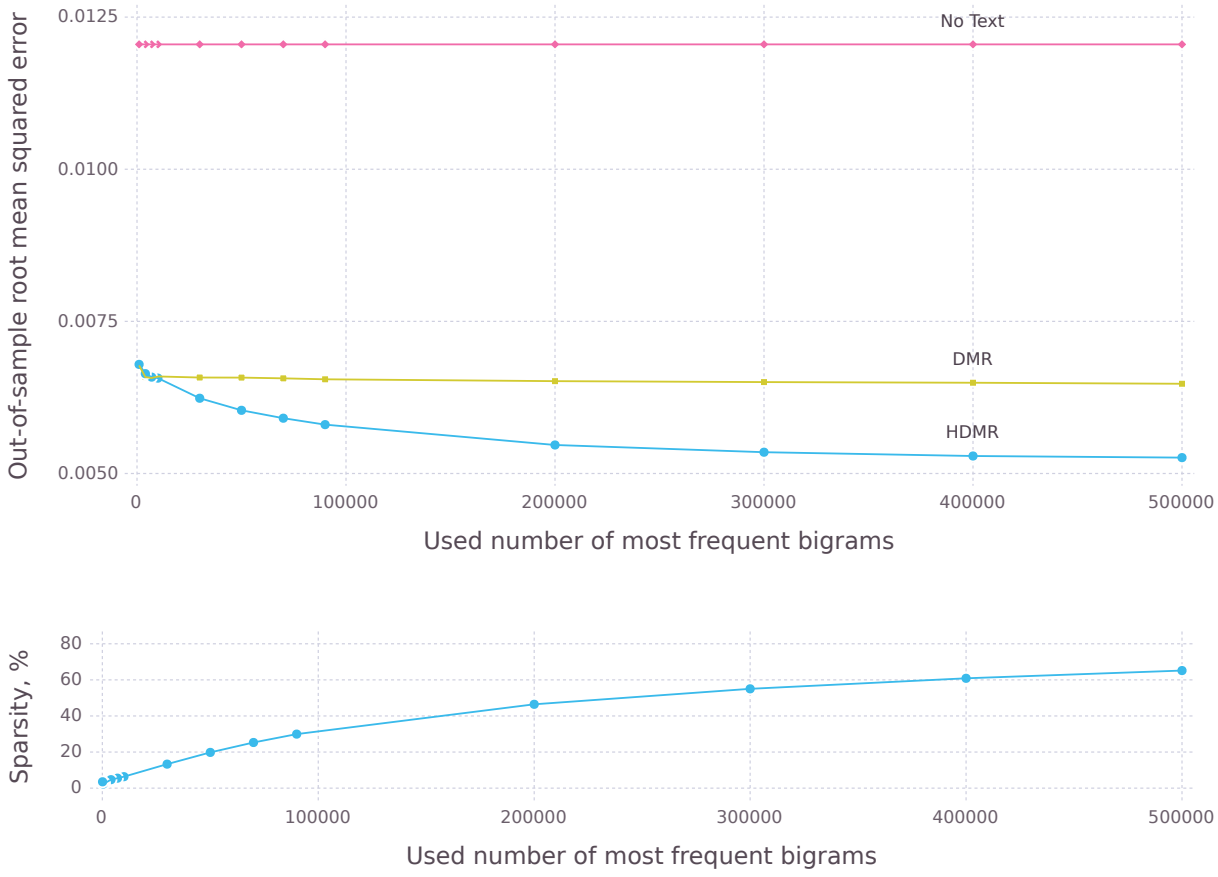

Figure 7: Predicting the intermediary capital ratio with denser text and covariates

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 log realized variance (rv_t), the log price dividend ratio (pd_t), business news pressure ($biznewspres_t$), 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 multiple 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. 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

Variable	Mean	Std	Min	p10	Median	p90	Max	Obs	Available
icr	0.06	0.02	0.02	0.04	0.06	0.10	0.13	557	197001–201605
rv	-4.26	1.05	-7.18	-5.47	-4.40	-2.88	-0.47	1079	192607–201605
pd	3.44	0.40	2.21	2.96	3.39	4.02	4.56	1075	192611–201605
biznewspres	16.84	5.92	5.28	8.90	16.30	25.68	33.60	839	194601–201605

Notes: Reported are summary statistics for variables in the monthly sample from July 1926 to May 2016. Intermediary capital ratio (icr_t) is the aggregate ratio of market equity to market equity plus book debt of New York Fed primary dealers. The log price over past year dividends (pd_t) is from CRSP. The log realized variance rv_t is the sum of squared daily market returns over month t. Business news pressure ($biznewspres_t$) is the mean monthly number of pages in the first section (A) of the *Wall Street Journal*.

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

(a) Cross-validation with 10 *random* folds

	RMSE						R-squared					
	Out-of-sample			In-sample			Out-of-sample			In-sample		
Model	Text	No Text	%Δ	Text	No Text	%Δ	Text	No Text	%Δ	Text	No Text	%Δ
HDMR	0.007	0.014	-44.926	0.006	0.013	-54.009	0.904	0.685	32.094	0.935	0.691	35.239
DMR	0.008	0.014	-38.550	0.007	0.013	-47.006	0.881	0.685	28.671	0.913	0.691	32.141
FHDMR	0.008	0.014	-39.793	0.007	0.013	-49.716	0.886	0.685	29.368	0.922	0.691	33.392
SVR	0.010	0.020	-51.760	0.006	0.013	-54.683	0.832	0.280	197.295	0.936	0.689	35.870

(b) Cross-validation with 10 *serial* folds

	RMSE						R-squared					
	Out-of-sample			In-sample			Out-of-sample			In-sample		
Model	Text	No Text	%Δ	Text	No Text	%Δ	Text	No Text	%Δ	Text	No Text	%Δ
HDMR	0.010	0.018	-42.148	0.006	0.013	-53.586	0.818	0.458	78.852	0.935	0.700	33.560
DMR	0.016	0.018	-9.606	0.007	0.013	-47.533	0.557	0.458	21.676	0.918	0.700	31.000
FHDMR	0.015	0.018	-17.967	0.007	0.013	-50.036	0.635	0.458	38.763	0.925	0.700	32.097
SVR	0.017	0.592	-97.155	0.006	0.013	-54.114	0.513	-601.287	-100.085	0.937	0.699	34.048

Notes: Reported are in- and out-of-sample root mean squared error (RMSE) and R-squared from a 10-fold cross validation exercise that tries to predict the intermediary capital ratio (icr_t) using log realized variance (rv_t), the log price dividend ratio (pd_t), business news pressure ($biznewspres_t$), and monthly WSJ front page phrase counts, over the subsample when the capital ratio is available, January 1970 to February 2016. Panel (a) uses random folds for validation while Panel (b) uses serial folds, each constituting a distinct sub-period. Our proposed model, the hurdle distributed multiple regression (HDMR) 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 the measure of fit with and without the text, and the percent change in the measure of fit (%Δ).

Table 3: Explaining the text with intermediary capital ratio-related covariates

(a) Frequent phrases with the most positive loadings

Variable	Sparsity	Top positive
icr^0	0.571	labor letter, busi bulletin, tax report, washington wire, gross net, nation recoveri, presid clinton, trend take, job trend, life job
rv^0	0.590	barrel dow, bushel wheat, yr trea, trea yld, dow jone, aig spot, futur dj, outstand stock, aig futur, stock nyse
pd^0	0.579	barack obama, yr trea, trea yld, technolog journal, insid journal, journal insid, nasdaq amp, commod oil, aig futur, weekend journal
$biznewspres^0$	0.637	washington wire, busi bulletin, tax report, labor letter, report steel, journal special, substanti gain, rubber tire, presid reagan, life job
icr^+	0.840	confer washington, ounc dow, miami beach, presid clinton, survey found, bosnia muslim, clinton health, presid slobodan, serb croat, barrel dow
rv^+	0.838	west africa, amp unfil, falun gong, stock market, republican guard, composit index, john mccain, dow jone, jone industri, abu dhabi
pd^+	0.861	c c, avail headlin, yr treasuri, treasuri yld, bond yr, eastern ukrain, amp unfil, announc week, al qaeda, moammar gadhafi

(b) Frequent phrases with the most negative loadings

Variable	Sparsity	Top negative
icr^0	0.571	barack obama, presid barack, obama administr, substanti gain, unit total, bodi stock, al qaeda, compar januari, regist co, hedg fund
rv^0	0.590	busi bulletin, tax report, washington wire, substanti gain, labor letter, chase nation, press build, bodi stock, chicago rock, amp unfil
pd^0	0.579	labor letter, busi bulletin, tax report, washington wire, jobless marri, steel product, factori shipment, hour earn, lead indic, busi failur
$biznewspres^0$	0.637	bushel wheat, futur barrel, commod oil, jone aig, aig futur, barrel dow, barack obama, dj aig, al qaeda, chicago rock
icr^+	0.840	yr treasuri, construct spend, wsj research, treasuri yld, euro zone, c c, bond yr, announc week, marshal field, avail headlin
rv^+	0.838	intern aid, american cyanamid, construct spend, survey found, buyer market, vote presid, sharp contrast, harrisburg pa, talk aim, unit sale
pd^+	0.861	presid clinton, barrel dow, fiscal cliff, hurrican katrina, confer washington, bosnia muslim, wire clinton, orang counti, survey found, serb croat

Notes: The table reports the 20 phrases with the most positive or most negative loadings estimated in an HDMR of monthly WSJ front page phrase counts on the intermediary capital ratio (icr_t), log realized variance (rv_t), the log price dividend ratio (pd_t), business news pressure ($biznewspres_t$), over the subsample when the capital ratio is available, January 1970 to February 2016. We multiply each phrase loading by the square-root of the mean count for the phrase $\bar{c}_j$ before sorting to get terms that are both associated with the variable and relatively frequent. Superscript 0 and + indicate, respectively, variable loadings in the inclusion and repetition equations. Sparsity is the fraction of phrase loadings that are exactly zero. 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 a single phrase for intuition: "financial crisis"

(a) Backward regressions

	HDMR		DMR
	Repetition	Inclusion	
Intercept	-8.06	-16.02	-13.70
icr	-33.41	-60.08	-58.87
rv	0.15	0.48	0.26
pd	1.20	3.89	3.01
biznewspres		-0.02	0.00

(b) Forward regressions

	HDMR	DMR
Repetition	-0.03	-0.05
Inclusion	-0.04	

Notes: Panel (a) reports backward HDMR coefficient estimates for the (stemmed) phrase "financi crisi" on the covariates, which exclude *biznewspres* from the repetition equation (the model for positive counts). Panel (b) 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. The hurdle distributed multiple 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

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

Notes: Reported are time-series predictability regression estimates of future stock market excess returns at one to twelve months horizon on lagged sufficient reduction projections z^0_{t-1}, z^+_{t-1} that summarize the text, the log realized variance (rv_{t-1}), log price-dividend ratio (pd_{t-1}) and business news pressure ($biznewspres_{t-1}$). For comparison, we report similar regressions where the single DMR projection z^{dmr}_{t-1} is used instead. The corpus includes the 10,000 most frequent two-word phrases (bigrams) in separate sentences, after removing case, stopwords, nonletters and Porter stemming. Hodrick (1992) standard errors are in parentheses.