Text-Based Factor Models of Equity Prices

David Adelberg

Advisor: Bryan Kelly

Yale University

Abstract. This paper links business news to equity returns through a latent factor model. To estimate the relationship between a high-dimensional set of mostly text-based instruments and latent risk factors, I employ a novel regularization technique that enhances the model's out-of-sample explanatory power. I find that business news captures cross-sectional variation that cannot be explained by the traditional features studied. Such variation is explained by latent systematic risk factors, not the mispricing of news; results are consistent with the efficient markets hypothesis.

1 Introduction

Increases in computational power are making larger, unstructured datasets amenable to analysis. Financial economists are capitalizing on this technological trend by exploring the connection between non-traditional datasets and empirical financial market dynamics. I examine the relationship between the Dow Jones Newswires, a corpus of business news, and US equity returns.

Text differs from traditional datasets in its high dimensionality. Even if the English language contained only 1,000 words and documents were exactly 30 words long, the number of possible documents would approximate the quantity of atoms in the universe (Gentzkow et al., 2017). Dimensionality reduction is essential if one is to make sense of text data.

Since Sharpe proposed the capital asset pricing model (Sharpe, 1964), financial economists have been interested in factor models of equity prices. For instance, the Fama-French (1993) three-factor model posits that equity returns relate to static loadings on observable long-short portfolios based on characteristics such as size. Static loadings on these portfolios are estimated via time series regression (Fama and French, 1993). By contrast, cross-sectional factor models (such as the BARRA factor model) allow factor loadings to vary over time, but they assume that observable firm characteristics are the dynamic loadings. Cross sectional regression of equity returns against these loadings yields the time series of factor returns (Nielsen and Chu Bender, 2010).

Following Kelly, Pruitt, and Su (2017), I estimate a dynamic factor model that is less restrictive. As in fundamental factor models like the BARRA model, I theorize that individual stock returns relate to unobserved risk factor returns; however, factor loadings are also unobserved. These loadings are modeled as an unknown linear function of observed characteristics. This more general specification refrains from making the unrealistic assumption that the systematic risk

exposures of firms do not change over time. In addition, the number of latent risk factors does not limit the number of characteristics. The trade-off is that estimation becomes more complex. Since text has large dimensionality, learning factors from data is essential if one is to make sense of news.

This sort of latent factor model can be estimated using Instrumented Principal Components Analysis, or IPCA (Kelly et al., 2017). Yet IPCA was designed for cases when the number of instruments is small relative to the number of observation dates in the (potentially unbalanced) panel. I will demonstrate empirically that IPCA can perform poorly for high-dimensional instrument sets. I extend IPCA by applying regularization. When we include high-dimensional text-based instruments, regularization leads to large improvements in explanatory power out-of-sample.

Financial economists have long been interested in market efficiency. If markets are semi-strongly informationally efficient, then prices incorporate all information available to the investing public (Fama, 1970). Using a residual bootstrap procedure, Kelly, Pruitt, and Su (2018) find that expected stock returns in excess of systematic risk compensation have no statistically signifiant linear relationship to several dozen traditional financial features. This evidence is consistent with semi-strong informational efficiency. However, non-traditional data such as news could still predict excess returns (alpha). I apply the residual bootstrap to ridge IPCA and find that business news does not predict alphas, consistent with semi-strong informational efficiency.

If markets are efficient, then a set of systematic risk factors explains the panel of equity returns. The number and character of these risk factors has been a longstanding question in financial economics. The estimated text-based latent factor models outperform traditional factor models, suggesting that a complete

description of systematic equity market risks should include factors spanned by business news.

Contribution to the Literature This paper contributes to the literature in three ways. First, I develop Ridge Instrumented Principal Components Analysis (RIPCA), a new econometric method. RIPCA is a technique for estimating dynamic latent factor models. Using a ridge hyperparameter, RIPCA extends IPCA by allowing the econometrician to apply a ridge penalty to selected features. I demonstrate that RIPCA outperforms IPCA for two high-dimensional sets of instruments.

Second, I test the (semi-strong) efficient markets hypothesis in the context of business news. One can use RIPCA to test hypotheses with the help of a residual bootstrap procedure. I examine whether news is related to stock-specific "alpha" unrelated to systematic risk exposures. These tests support the efficient markets hypothesis – I do not find that news predicts excess returns.

Third, I relate business newswires to the cross section of returns and risks. Traditional variables such as accounting metrics derived from a firm's SEC filings and measures of past return are not enough to fully account for systematic risk exposures. I find that business newsires regarding a particular firm inform that firm's exposure to systematic risks.

2 Data

2.1 Fundamentals and Return Data

To study text-based factor models, I worked with features derived from SEC filings, historical stock prices, and business newswires. In this paper, I call accounting-based and return-based instruments "traditional" features. By contrast, I call features derived from newswires "text-based" or "textual." I worked

with the same accounting-based and return-based instruments used in "Characteristics Are Covariances" (Kelly et al., 2018). This dataset contains 40+ features, most of which are derived from accounting data. Other features describe characteristics known to explain variation in asset prices, such as momentum and reversal. The excess returns data is derived from the Center for Research in Security Prices (CRSP) equity data. Features are at monthly frequency.

2.2 Dow Jones Newswires

As we seek to estimate company-specific loadings, it is essential to use a microeconomic news corpus. To study business news, I used the Dow Jones Newswires. Unlike other financial text corpora, these newswires provide their readers with company-level, microeconomically relevant information. By contrast, Wall Street Journal articles primarily communicate macroeconomic news. Prior to analysis, this unstructured dataset was restructured into a tabular format. Unigram and bigram frequencies were calculated for each document. A separate table indicates the firm in question and the publication date for each document.[1]

The corpus contains news from mid-1979 onwards. However, Figure 1 indicates that the number of companies covered increases nearly ten-fold from just over 1,000 firms to nearly 10,000 firms. In Figure 2, we see that the rate of news publication increases significantly in the 1990s. This suggests that this change in company coverage is primarily driven by an increase in the quantity of news rather than a change in reporting priorities. Interestingly, this publication rate has declined in the past decade.

To determine whether typical firms are frequently covered, I visualize how the median number of newswires (per firm) changes over time in Figure 3. In the 1980s and much of the 1990s, the median firm had fewer than twelve news articles per year, or less than one article per month. By contrast, in the 2000s

[1] I thank Leland Bybee for sharing this restructured data.

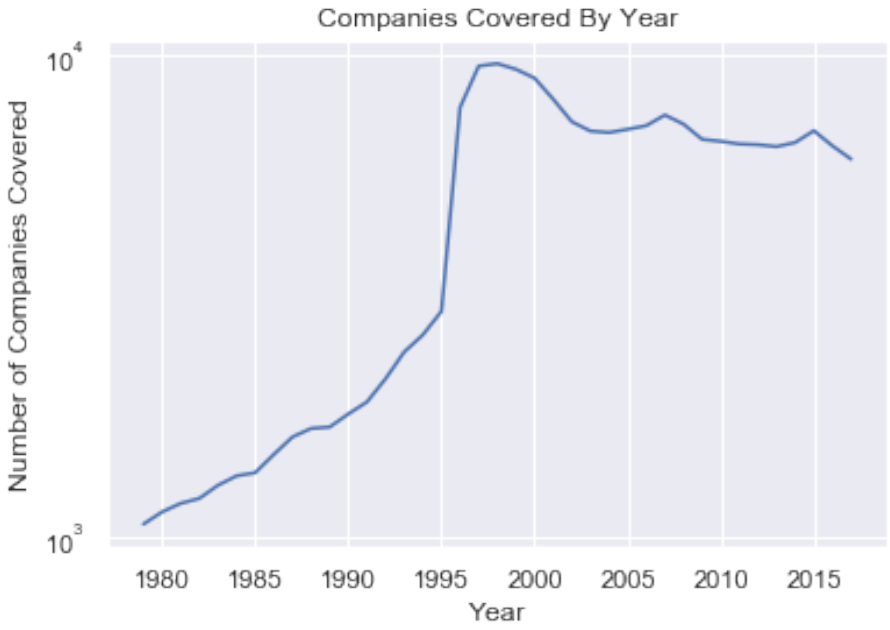

Fig. 1. Firm coverage undergoes large changes in the 1990s.

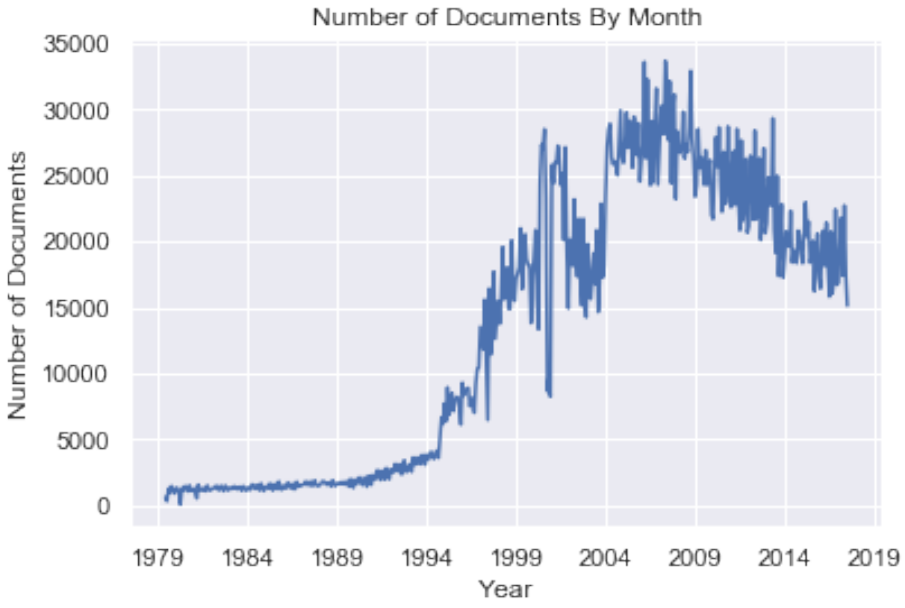

Fig. 2. Document frequency increases rapidly in the 1990s.

this median rate reaches as high as 35 articles per firm-year. Sufficient news is available for most firms.

Fig. 3. The median firm receives more coverage in the 2000s.

If firms covered later in the sample have different characteristics than those covered earlier in the sample, this time series variation in coverage could impact results. To minimize the impact of this issue, I elected to focus on the 1999-present subsample. For this subsample, an average of 2 articles are published per month for each of the 10,000 firms covered. In addition, we saw in Figure 3 that median fims have a sufficient number of newswires, indicating that newswire concentration is moderate.

The corpus has a large vocabulary of 33,127 words. Topic modeling can be done relatively efficiently, so I used the bigram counts for this specific analysis. For the latent factor models, I elected to use the unigram counts because the

unigram vocabulary size is smaller than the bigram vocabulary size. There are 153,355 firm-months in the combined text, accounting, and price dataset with data for all features. As there are many instruments, one must first reduce the dimensionality of the news corpus to tractably analyze this dataset. Otherwise, polynomial time computations such as the matrix inverse would not be feasible.

Naturally, some words occur much more frequently than others. In addition, style drift can cause word frequencies to change over time. To mitigate these issues, I applied the Term Frequency-Inverse Document Frequency (TF-IDF) transform cross-sectionally. Since I divide word frequencies by the probability that an arbitrary document contains that word, the transformed frequencies reflect whether that word occurred more frequently than the cross-sectional average.

Latent Dirichlet Allocation If documents could be summarized with a small number of topics, then the topic weights could be used as textual instruments. I estimated a 10-topic Latent Dirichlet Allocation (LDA) model using bigrams. LDA is a hierarchical generative model in which documents are a mixture of latent topics (Blei et al., 2003).

In Figure 2.2, we examine how topic proportions change over time. We find that some topics (such as topic 6) have fairly stable proportions over time, while other topic proportions trend, indicating style drift.

Firms engage in a wide variety of activities, so one might expect summarizing microeconomic news articles with a small number of topics to be difficult. So, LDA topics should appear highly mixed to a human if the Dow Jones Newswires are topically rich. To understand these topics, I ranked words by informativeness, defined as the quotient of the term-topic proportion and the word frequency. The term-topic proportion is the probability that a particular word comes from that topic. A word is informative for a topic if it occurs infrequently, but is frequently

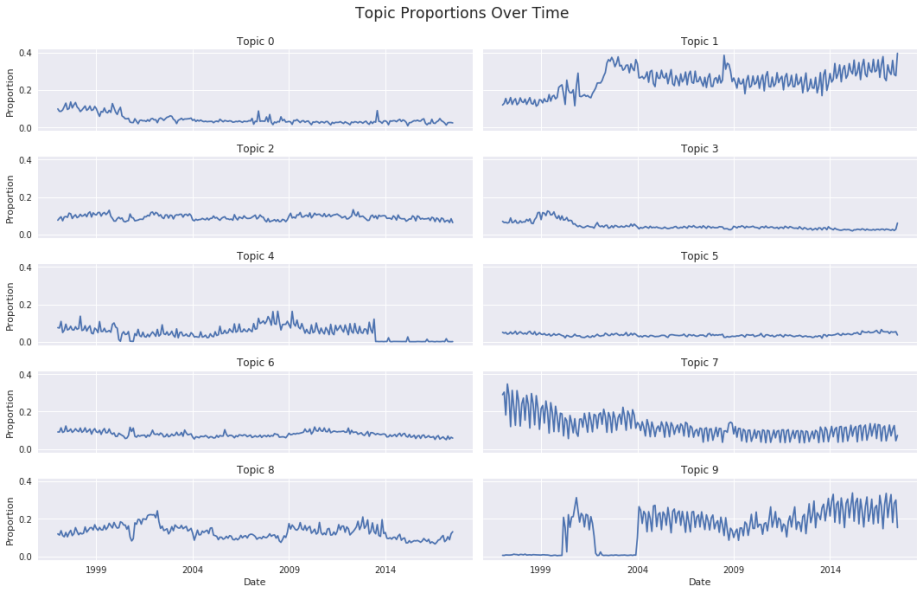

Fig. 4. LDA topic proportions change over time.

found in documents that weight that topic highly. Indeed, when one examines the most informative words (with at least 100 occurrences) for the sample topic displayed in Figure 5, one finds that the topics do not describe a coherent subject. (Informative words for the other topics can be found in Appendix B.) Banking-related terms such as "liquidity market" and "financial website" are alongside unrelated terms such as "russian federation" and "bureau investigation."

```
Topic 0
['liquidity market', 'loan mature', 'financial website', 'game home', 'aim enhance', 'attention de
tail', 'bank portfolio', 'online consumer', 'firm find', 'bureau investigation', 'drive technolog
y', 'fiscal ebitda', 'people eat', 'russian federation', 'online deal', 'fear government', 'global
management', 'company capital', 'statement regard', 'separate file']
```

Fig. 5. The first LDA topic is semantically mixed

If business news could be adequately summarized by LDA, then one could fit a latent factor model using the topic weights as instruments, and no additional

textual features. However, this is not possible; instead, one should build a latent factor model with a high-dimensional set of instruments.

2.3 Data Transformation and Dimensionality Reduction

Random Projection We have just discussed the difficulty of compressing documents to a small number of topic weights. Still, it would be desirable to reduce the dimensionality of the 30,000+ word vocabulary to a smaller size so in-memory computation is possible. As the dataset is quite large, a technique like Principal Component Analysis (PCA) would be computationally expensive. Instead, I applied a gaussian random projection to the transformed word frequencies. We premultiply the feature matrix by a random matrix in order to randomly project data to a 1,000 dimensional space. Each of the elements of this matrix was independently drawn from a normal distribution.

By the Johnson-Lindenstrauss lemma, this class of embedding nearly preserves distances with high probability. In addition, this embedding is nearly orthogonal (Bingham and Mannila, 2001). Furthermore, the TF-IDF transformation reduces disparities in word count frequency, so the features should have similar magnitudes. So, the random embedding is likely to preserve most of the useful information in the news corpus.

After merging these projected, transformed term frequencies with traditional features, we are left with an instrument set large enough to capture most of the relevant information yet small enough for computational tractability. Furthermore, this transformed dataset fits in the RAM (8GB) of a MacBook Pro. The price paid to achieve this 30x reduction in dimensionality is severe mixing of features. Later, we will see that this random mixing complicates interpretation of text-based factors. Future work could consider sparse dimensionality reduction techniques that might aid in the interpretation of results.

3 A Dynamic Latent Factor Model of Equity Prices

3.1 Model

As discussed above, equity returns are assumed to depend on latent factor returns, dynamic loadings on latent factors, and idiosyncratic noise. Following Kelly, Pruitt and Su (2017), I write:

$$y_t = \beta_{t-1} f_t + \eta_t$$
$$\beta_{t-1} = Z_{t-1}\Gamma + \nu_{t-1}$$

In other words, the vector of equity returns on a particular date can be decomposed into the sum of a vector representing the systematic risk contribution ($\beta_{t-1} f_t$) and an idiosyncratic risk vector η_t. The matrix of dynamic loadings is a linear function of observable instruments ($Z_{t-1}\Gamma$), plus noise (ν_{t-1}). Substituting the dynamic loadings model into the returns model, we obtain

$$y_t = Z_{t-1}\Gamma f_t + \epsilon_t$$

This specification assumes that expected returns depend solely on systematic risk compensation. However, it is easy to admit instrument-dependent alphas: one can simply restrict one of the factors to have a constant return of 1.

3.2 Estimation with Few Instruments

Initialization of Parameters We obtain an asymptotically consistent estimator of Γ and f_t by minimizing the sum of squared errors. Unfortunately, there is no analytical solution to this optimization problem (Kelly et al., 2018). How-

ever, an approximate solution is available. Following Kelly, Pruitt and Su (2018), consider the realized returns of characteristic-managed portfolios:

$$x_{t+1} = Z_t' r_{t+1}$$

Each managed portfolio purchases and shorts stocks according the value of its associated instrument. For instance, a portfolio managed on the basis of time series momentum would purchase (or short) a quantity of stock proportionate to its historical return.

If $Z_t' Z_t$ were constant, then Γ would consist of the first K eigenvectors of $\sum_t x_t x_t'$ (Kelly et al., 2018). As stock characteristics do in fact change over time, this estimator is inexact; still, it can be used to initialize a more accurate estimator.

Estimation Kelly, Pruitt, and Su (2018) propose an alternating least squares (ALS) algorithm for estimation of this model:

1. First, initialize Γ as described above.
2. Repeat until convergence:
   (a) Using the latest estimate of Γ, solve for the latent factor returns at each point in time via ordinary least squares.
   (b) Using the latest estimate of the f_t, solve for Γ via ordinary least squares.

As each ordinary least squares (OLS) subproblem is easily solved with efficient algorithms based on the singular value decomposition (Golub and Reinsch, 1970), this IPCA algorithm for dynamic latent factor models is not much slower than estimation of static factor models such as principal components analysis (Kelly et al., 2018).

3.3 Ridge IPCA: Estimation with Many Instruments

When the number of instruments is large relative to the number of observations, the IPCA estimator exhibits high variance. The above algorithm relies on OLS for estimation of Γ, but OLS is inaccurate when the number of regressors is large. I modify IPCA to handle this high-dimensional case, and call this extension ridge IPCA (RIPCA).

We begin by rewriting the composite equation for stock returns, splitting instruments into a low-dimensional subset (e.g. traditional instruments) and a high-dimensional subset (e.g. text-based instruments).

$$\begin{aligned}
\beta_{t-1} &= Z_{t-1}\Gamma + \eta_{t-1} \\
\beta_{t-1} &= Z^l_{t-1}\Gamma^l + Z^h_{t-1}\Gamma^h + \eta_{t-1} \\
r_{t+1} &= Z_t\Gamma f_{t+1} + \epsilon_{t+1} \\
r_{t+1} &= (Z^l_t\Gamma^l + Z^h_t\Gamma^h)f_{t+1} + \epsilon_{t+1}
\end{aligned}$$

We still perform alternating linear regression to solve for the parameters. However, we minimize a penalized objective function:

$$\min_{\Gamma,F} \Sigma_{t=1}^{T-1}(r_{t+1} - (Z^l_t\Gamma^l + Z^h_t\Gamma^h)f_{t+1})'(r_{t+1} - (Z^l_t\Gamma^l + Z^h_t\Gamma^h)f_{t+1}) + \lambda vec(\Gamma_h)'vec(\Gamma_h)$$

This ridge objective function penalizes the squared weights on text-based features. The ridge penalty parameter λ can be chosen by cross-validation outside of the alternating regression loop. We evaluate values of λ using total R^2, which is the proportion of variance explained by the latent factor model, including the

fitted latent factor returns (Kelly et al., 2018). For clarity, here is the RIPCA algorithm:

1. Initialize a vector Λ consisting of candidate ridge penalty parameters.
2. Initialize an empty list R that will contain the squared error for each ride penalty.
3. Estimate Γ using the approximate SVD-based algorithm described above. Store this estimate.
4. Randomly split the dates into k groups.
5. For each λ in Λ:
   (a) Initialize an empty list R_i to store model R^2's.
   (b) For each date group:
      i. Use data for all dates except those in the current date group.
      ii. Assign the SVD-based estimate of Γ to Γ.
      iii. Repeat until convergence:
         A. Update factor returns using OLS.
         B. Update Γ using ridge regression.
      iv. Compute the model's total R^2 for the held-out data. Append this to R_i.
   (c) Append the mean of R_i to R.
6. Let λ^* be the value of λ yielding the highest cross-validated total R^2.
7. Using λ^* and the full training dataset, re-estimate the latent factor model M.
   (a) Recalculate Γ using the SVD-based approximate algorithm.
   (b) Repeat until convergence:
      i. Update factor returns using OLS.
      ii. Update Γ using ridge regression.
8. Return the model M.

This longer procedure wraps the original IPCA algorithm, modified with ridge regression, inside line search and cross-validation loops. RIPCA is significantly slower than standard IPCA for two reasons:

1. The hyperparameter line search and cross-validation steps cause the runtime of RIPCA to have two additional linear multiplicative terms.
2. OLS is replaced by ridge regression, which runs more slowly than OLS.

Despite this drawback, RIPCA makes it feasible to fit latent factor models with many instruments that generalize out-of-sample. I will now present empirical results demonstrating that RIPCA sometimes generalizes better than IPCA, especially when the number of features is large.

4 Comparing RIPCA and IPCA

4.1 Expanding the Traditional Instrument Set

To empirically assess the efficacy of RIPCA, I initially exclude text-based features from my analysis, so that this unorthodox dataset does not confound the comparison. I generate a larger dataset with an expanded feature set in order to explore the effect of regularization with a larger number of instruments. For each traditional feature x_i and for $p = 1..P$, I create a new feature by standardizing x_i^p to have mean 0 and standard deviation 1.

The impact of regularization is smaller when the model has fewer factors; as the size of one's factor model increases, one should be more inclined to regularize. If regularization improves results for this small model, then it is likely to yield even greater benefits when estimating a larger, more realistic model. For this section, I apply regularization to all of the features and only use three latent factors. To quantify out-of-sample performance, I measure total R^2 for the post-2008 period. Models are trained via RIPCA using data prior to 2008. Five-fold cross-validation is used to select the ridge hyperparameter.

Results In the scatterplot below, we compare the total out-of-sample R^2 achieved by RIPCA and IPCA for $P = 1..8$:

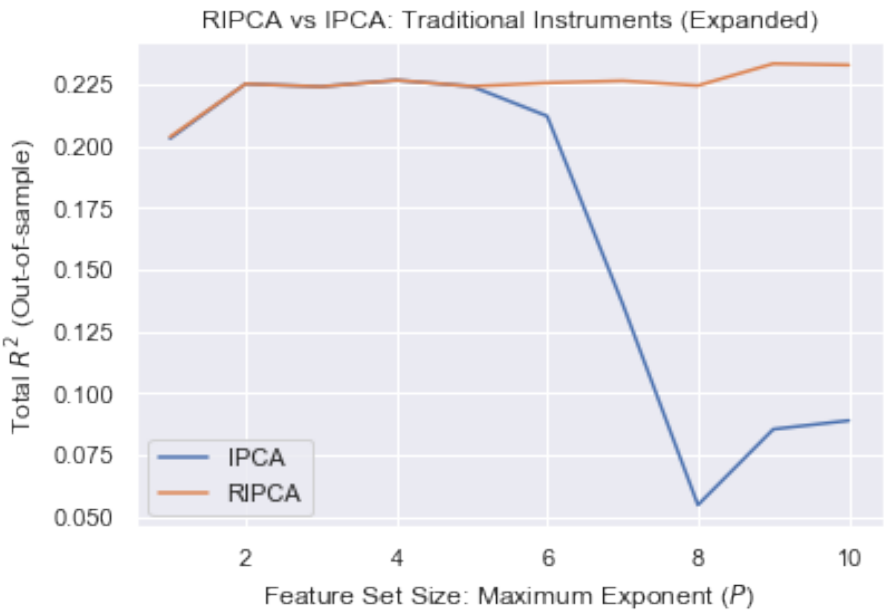

Fig. 6. RIPCA outperforms IPCA for large P

Comparing the orange RIPCA line with the blue IPCA line in Figure 6, we find that regularization achieves little to no improvement in total R^2 when the feature set is small. When the hyperparameter optimization step sets $\lambda = 0$, the two lines overlap. However, we witness a dramatic divergence in out-of-sample performance for $P >= 6$. When the number of instruments exceeds around 250, RIPCA performs meaningfully better than IPCA.

4.2 Quantifying the Impact of Regularization with Text-Based Features

Having shown that regularization produces meaningful benefits for an expanded feature set, I will now repeat the above analysis for the full set of instruments, including text-based features. I no longer apply power transformations to the traditional features. Here, regularization is only applied to the text-based features. Above, we found that $IPCA$ works well for the raw traditional features ($P = 1$), consistent with Kelly, Pruitt, and Su (2018). Regularization is unnecessary for the associated parameters. In addition, there are far more text-based features than traditional features. Furthermore, we expect such features, which randomly average specific transformed word frequencies, to have smaller weights than the traditional features. So, it makes sense to regularize only the parameters associated with text data.

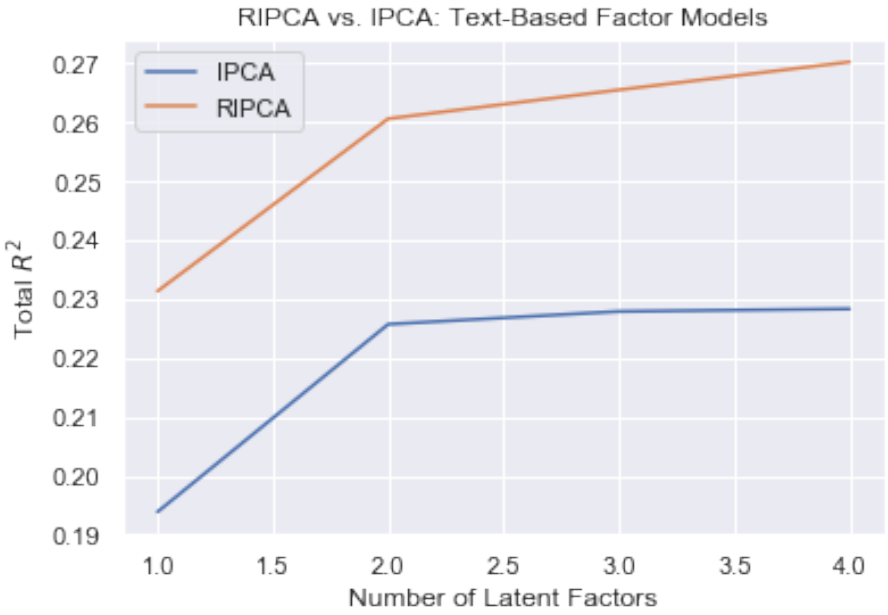

Fig. 7. RIPCA outperforms IPCA for text-based factor models.

In Figure 7, we compare the out-of-sample total R^2 of latent text-based factor models estimated with RIPCA and IPCA. Since the number of instruments exceeds 1,000, it is not surprising that RIPCA outperforms IPCA. With just one latent factor, RIPCA (orange) yields a total R^2 around 3.7% higher than IPCA (blue). With four factors, the incremental value of RIPCA over IPCA exceeds 4.1%.

The difference in the slopes of these two curves is also interesting. While IPCA extracts two interesting text-based factors, it struggles to identify additional systematic risk factors that explain meaningful variation. By contrast, the third and fourth RIPCA factors explain 1% of variance.

5 Testing the Efficient Markets Hypothesis with RIPCA

5.1 Hypothesis Testing with RIPCA

Above, we saw how RIPCA can be applied to business news for the purpose of better explaining equity price dynamics. RIPCA can also be used to test theories such as the efficient markets hypothesis. I will now outline how this can be done.

If markets are not semi-strongly efficient with respect to the traditional and text-based instruments, then investors can generate alpha (in excess of systematic risk compensation) by investing in stocks with favorable values of the instruments. To quantify such inefficiency, we can estimate a RIPCA model that includes "alphas." This can be modeled by a predictable factor that always realizes a return of 1.

$$r_{i,t+1} = \alpha_{i,t} + \beta_{i,t} f_{t+1} + \epsilon_{i,t+1}$$

To test the hypothesis of zero instrumented alphas, we examine whether the estimated loadings for the predictable factor are statistically distinguishable from zero (jointly):

$$H_0 : \Gamma_\alpha = 0$$
$$H_a : \Gamma_\alpha \neq 0$$

Following Kelly, Pruitt, and Su (2018), we test this hypothesis using a Wald-like test statistic:

$$W_\alpha = \hat{\Gamma_\alpha}' \hat{\Gamma_\alpha}$$

First, we estimate the latent factor model using RIPCA, requiring one of the factors to be 1 for all time periods. Then, we compute the returns of "managed portfolios."

$$\begin{aligned} x_{t+1} &= Z_t' r_{t+1} \\ &= (Z_t' Z_t)\Gamma_\alpha + (Z_t' Z_t)\Gamma_\beta f_{t+1} + d_{t+1} \end{aligned}$$

Rather than resampling individual stock idiosyncratic returns, we instead resample these managed portfolios' fitted idiosyncratic returns $\hat{d}_t$.

Next, we draw B "wild" residual bootstrap samples:

$$\tilde{d}^b_{t+1} = q^b_1 \hat{d}_{q^b_2}$$
$$\tilde{x}^b_t = (Z'_t Z_t)\Gamma_\alpha + (Z'_t Z_t)\Gamma_\beta f_{t+1} + \tilde{d}^b_{t+1}$$

For each date, we draw a time index q^b_2 uniformly at random. We multiply the associated fitted residual by a random variable q_1 distributed as a unit Student's t with 5 degrees of freedom. This "wild" bootstrap step improves statistical efficiency in the presence of heteroskedasticity (Gonalves and Kilian, 2004). By construction, these bootstrapped samples are drawn from a model satisfying the null (the efficient markets hypothesis).

For each bootstrap sample, we re-estimate the factor model using RIPCA and compute $\tilde{W}^b_\alpha$. If $\hat{P}(\tilde{W}^b_\alpha > W_\alpha)$ is less than 0.05, then we reject the null hypothesis of market efficiency.

5.2 Testing the Efficient Markets Hypothesis with a Text-Based Factor Model

I now apply the procedure described above to test whether business news informs returns in excess of systematic risk compensation. For $K = 1..3$, I test whether a text-based factor model with K latent factors is consistent with the efficient markets hypothesis. I draw 30 bootstrap samples per model, for a total of 90 samples. These tests take around 12 hours to run on a personal computer. Results are summarized in Table 1. When we allow for two or more factors, we fail to reject the null hypothesis of market efficiency. These tests suggest that investors cannot generate "alpha" using word frequency data; at best, they can increase their portfolios' loadings on compensated systematic text-based risk factors.

K	p-value
1	0.0
2	0.3
3	0.6

Table 1. Results are consistent with the efficient markets hypothesis for $K >= 2$.

6 A Text-Based Factor Model of Equity Prices

6.1 Quantifying the Explanatory Power of Business Newswires

The hypothesis tests above suggest that one cannot generate alpha using business newswire word frequencies. I will now investigate whether news is useful for understanding systematic risk exposures. I estimate two sets of latent factor models: for the first set, I only use traditional instruments; for the second set, I also include text-based instruments. Each set contains models with 1-4 factors. I fit these models to the full panel of observed returns data, selecting hyperparameters associated with good cross-validation results. I then compute total R^2 for all of the models and summarize the results in Figure 8.

Regardless of the number of factors selected, newswire data enhances explanatory power. With one latent factor, newswires explain an extra 1% of variance; with four factors, newswires account for around 2% of variance. In summary, newswire data is useful for explaining equity price dynamics.

6.2 Properties of Text-Based Factors

I now estimate a larger text-based factor model with eight latent factors and investigate the properties of these factors. In Figure 9, we examine the proportion of total variance explained by each of the latent factors. Unsurprisingly, the first latent factor explains far more variance than the other factors. This is the primary "market" risk factor. Higher-index factors explain less variation.

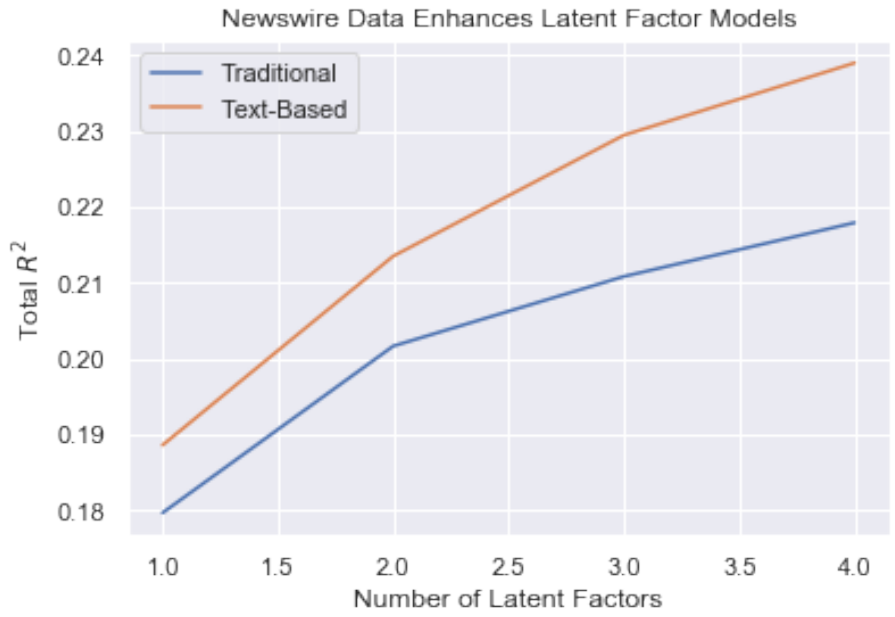

Fig. 8. Factor models with text-based instruments outperform traditional factor models.

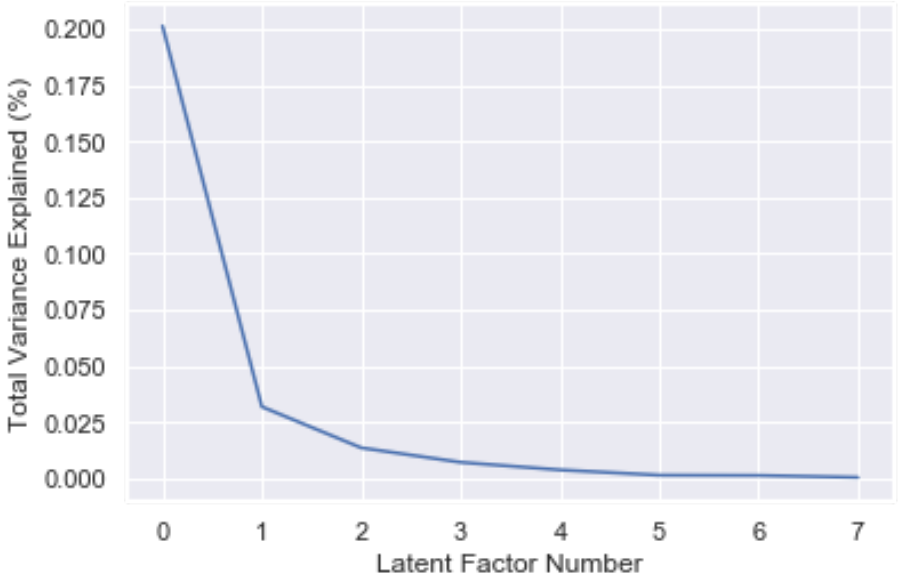

Fig. 9. The first few latent factors capture most of the explained variance.

I will now analyze the returns of these eight factors. (Annualized means, standard deviations, and sharpe ratios can be found in Appendix A.) I begin by plotting training and test factor sharpe ratios in Figure 10. Most of the factors experienced positive out-of-sample returns; in addition, factors with larger training sharpe ratios tended to realize a higher sharpe ratio in the test period.

Fig. 10. Training sharpe ratios of factors persist in the test set.

In Figure 11, we use a paired barplot to compare training and test sharpe ratios for the eight factors. Since factors were orthogonalized using PCA, the sample standard deviation for these factors is biased downwards and the sample sharpe ratio is biased upwards; so, the high training sharpe ratio for these factors is unsurprising. However, the high sharpe ratios in the test set are unexpected. The four factors with the largest test sharpe ratios are factors 4-8. These factors

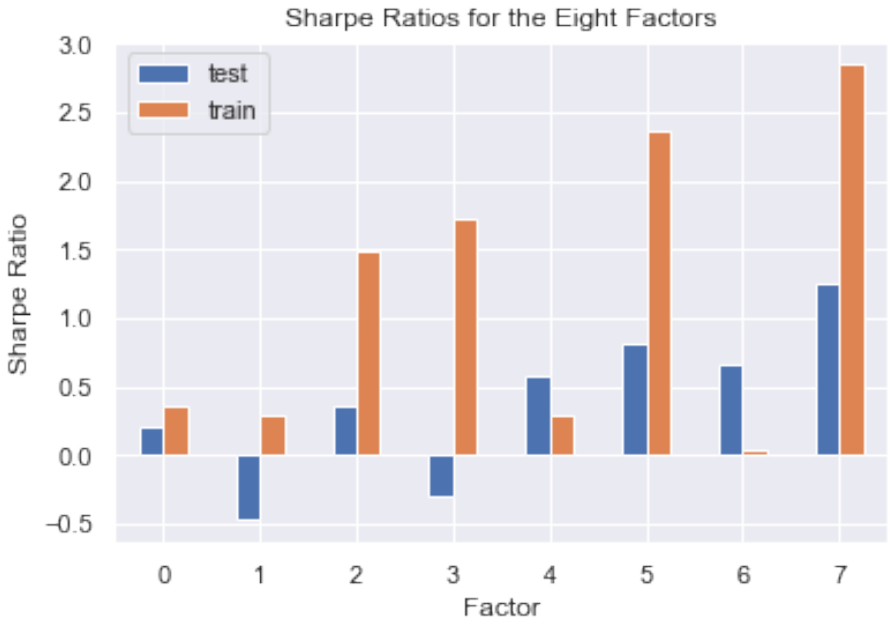

Fig. 11. The high-index factors tend to have higher sharpe ratios.

have an average test sharpe ratio of around 0.83. High-sharpe strategies are concentrated in the high-index factors.

Recall that the gap in total R^2 between RIPCA and IPCA for text-based factor models increased as the size of the number of latent factors increases. The slow growth in total R^2 for IPCA indicates that RIPCA better captures high-index factors. So, investors who would like to diversify into high-sharpe text-based factors would be advised to prefer RIPCA over IPCA for their factor model.

6.3 Interpreting the Loadings

To interpret the text-based factor model, it would be helpful to know which words have large loadings. Fortunately, it is easy to compute these loadings using the random projection matrix R:

$$\Gamma_{words} Z = \Gamma^{tb}_{features}(RZ)$$

$$\Gamma_{words} = \Gamma^{tb}_{features} R$$

To compute the loadings on individual words, we multiply the loading matrix (for the textual features only) by the random projection matrix.

In Appendix C, I list the words with the largest absolute loadings for an 8-factor model. Note that these loadings are applied to cross-sectionally TF-IDF transformed word frequencies; due to style drift, loadings on raw unigram frequencies change over time. Latent factors were orthogonalized and sorted using PCA.

We find that the most important words for each factor are difficult to categorize semantically. Our text-based factor models perform well out-of-sample, so this is not due to overfitting. Rather, this is an expected drawback of the dimensionality reduction technique employed. Using a random projection matrix, we reduced the number of textual instruments thirty-fold. The resulting projected instruments mix many unrelated words, some of which are company names.

Nonetheless, one can attempt to construct a story from these factor loadings. The first factor seems to have a number of words related to systematic balance sheet shocks. The loadings on the words "patent," "re-examination," "claim," "indebt," "usdbrr" (the exchange rate), "boughtdeal," and "appease" are consistent with this interpretation. Uncertainty regarding government interaction with business could generate a systematic risk reflected by these words. Alternatively, these words could simply reflect systematic market risk: in boom times, companies invest in patents and make deals. The second factor heavily weights words related to technology, innovation and growth, such as "electronic," "network," "semiconductor," "hazardous," "cagr," "vasomedical," "adhesive," and

"system." Perhaps this factor reflects a systematic innovation risk. Alternatively, this factor could measure the returns of growth stocks relative to value stocks. The third factor seems to describe systematic risks related to the oil industry and perhaps environmentalism. Terms like "drill," "oil," "gas," "lighthouse," "rig," "hydro," "activism," "inlet," and "archipelago" have large weights, supporting this interpretation. In addition, this factor could measure oil price exposure.

As discussed above, these interpretations are highly speculative because each of these factors mixes a number of concepts. Future work can examine whether sparse dimensionality reduction techniques can produce more interpretable factor loadings.

6.4 Optimal Text-Based Factor Portfolios

Using text-based factors, we can construct portfolios that aim to optimize return relative to risk. I use the intuitive yet powerful Markowitz mean-variance optimization framework (Markowitz, 1952). For the sharpe-maximizing return target μ^{tang}, we minimize the portfolio's variance:

$$\begin{aligned} x^* &= \min_{x:\mu^T x >= \mu^{tang}, 1^T x = 1} x^T \Sigma x \\ &= (1^T \Sigma^{-1} \mu)^{-1} \Sigma^{-1} \mu \end{aligned}$$

To estimate this portfolio of (dynamic) factors, I plug the training mean return vector and covariance matrix for these factors into the formula above. For this analysis, I train with RIPCA using pre-2005 returns and test using the rest of the dataset. This allows us to investigate the portfolio's performance during the financial crisis.

In Figure 12, I plot backtested cumulative returns of the estimated Markowitz factor portfolio. For simplicity, I assume that the portfolio is 100% financed and

leveraged to a volatility of 10%. The portfolio is rebalanced monthly with zero transaction costs.

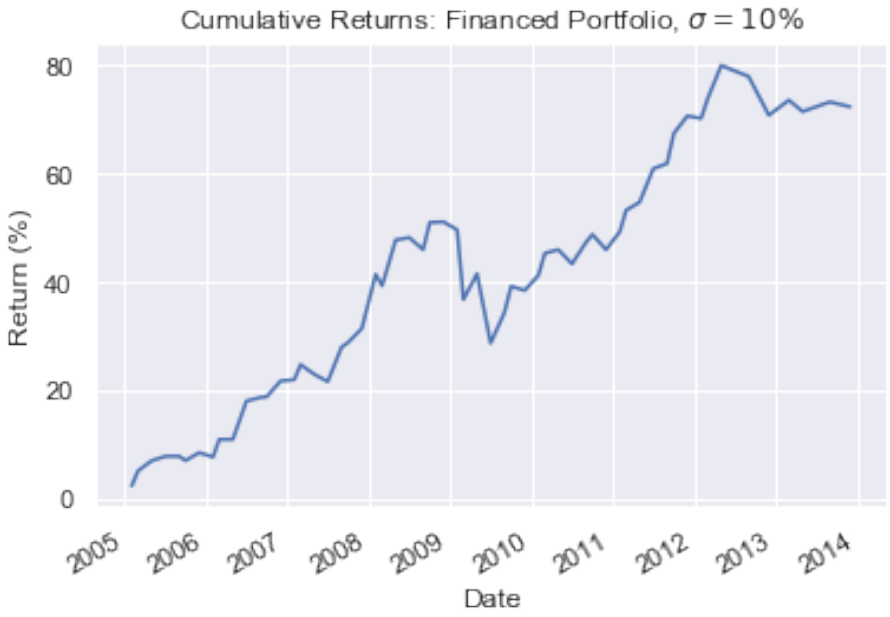

Fig. 12. An optimized latent factor portfolio performs well throughout the backtest period.

We find that the financed 10%-volatility portfolio achieves attractive returns over a nearly 9-year period. More interesting than the level of the returns is their time series pattern: the strategy performs well during the quant crisis of 2007 and the financial crisis of 2008. In addition, returns are positive during the pre-crisis and post-crisis regimes.

7 Conclusion

In this essay, I describe RIPCA, a new method that performs well for challenging dynamic latent factor model estimation problems. I demonstrate that RIPCA

outperforms IPCA on two financial datasets when the number of instruments becomes large. I also use RIPCA to investigate the efficient markets hypothesis in the context of text data. I examine whether business newswires inform returns in excess of systematic risk. I find that business newswires do not predict "alpha." This result contributes additional evidence to the debate on financial market efficiency. Nonetheless, business newswire data enhances the performance of latent factor models of equity prices. Furthermore, the estimated text-based factors have attractive standalone return properties; a portfolio of these factors also performs well in a backtest.

While business news adds to the explanatory power of a traditional latent factor model, the random projection dimensionality reduction technique employed in this paper produces semantically opaque factor loadings. Further work could investigate whether a sparse dimensionality reduction algorithm can produce more interpretable loadings. In addition, one could simply use the raw newswire data in conjunction with a distributed RIPCA algorithm and a cluster of computers.

With RIPCA, researchers can more effectively connect large sets of instruments to latent systematic risk factors. One could also use RIPCA to investigate the relationship between other high-dimensional datasets, such as Twitter tweets, and financial markets. More generally, RIPCA could facilitate the application of "big data" to latent factor models.

Bibliography

Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 245–250, New York, NY, USA. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as Data. NBER Working Papers 23276, National Bureau of Economic Research, Inc.

Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420.

Gonalves, S. and Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, 123(1):89 – 120.

Kelly, B., Pruitt, S., and Su, Y. (2018). Characteristics are covariances: A unified model of risk and return. Working Paper 24540, National Bureau of Economic Research.

Kelly, B. T., Pruitt, S., and Su, Y. (2017). Instrumented principal component analysis. *SSRN Electronic Journal*.

Markowitz, H. (1952). Portfolio selection*. *The Journal of Finance*, 7(1):77–91.

Nielsen, F. and Chu Bender, J. (2010). The fundamentals of fundamental factor models (june 2010). *SSRN Electronic Journal*.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk*. *The Journal of Finance*, 19(3):425–442.

A Latent Factors: Risk and Return

Factor	Statistic	mean	stdev	sharpe
0	test	0.059823	0.279563	0.213989
	train	0.326257	0.888101	0.367365
1	test	-0.114281	0.241221	-0.473760
	train	0.102585	0.352923	0.290671
2	test	0.063963	0.175207	0.365068
	train	0.342527	0.229984	1.489352
3	test	-0.049606	0.171034	-0.290034
	train	0.287880	0.166281	1.731285
4	test	0.047284	0.081998	0.576644
	train	0.035891	0.119901	0.299338
5	test	0.046470	0.057535	0.807689
	train	0.170119	0.071844	2.367901
6	test	0.038770	0.057782	0.670958
	train	0.002526	0.068329	0.036962
7	test	0.032117	0.025708	1.249302
	train	0.104653	0.036606	2.858893

B Most Important Words for the LDA Topics

Here, I list the most important words for all ten LDA topics.

```
Topic 1
['mfc tsx', 'teck resource', 'retain earn', 'spe process', 'cent reallowance', 'investor snap', 'deal complet', 'engi
neer capability', 'kvh industry', 'renown brand', 'logo register', 'edit david', 'john mill', 'easi manage', 'subscri
ber latest', 'expire extend', 'forego credit', 'material oral', 'report bas', 'run window']


Topic 2
['maintain sale', 'firm dedicate', 'likelihood time', 'baker gmt', 'lloyd london', 'claim arise', 'range rest', 'veda
nta resource', 'manufacture center', 'meet israeli', 'move prompt', 'boost lend', 'roger communication', 'usddem deut
sche', 'price surrender', 'accordance current', 'specialize service', 'peso source', 'statement oral', 'evaluate oper
ate']


Topic 3
['hole return', 'overnight trader', 'report compile', 'largest hold', 'elect leader', 'receive fix', 'held hear', 'in
dependent confirm', 'political sensitive', 'lost straight', 'weight asset', 'ease restriction', 'jeffery group', 'dea
d bomb', 'prop pfd', 'newly launch', 'rough share', 'enjoy benefit', 'left game', 'rate whitehall']


Topic 4
['vance mun', 'spirit airline', 'expeditor intl', 'target cite', 'negligent result', 'tse london', 'mount concern',
'dai vol', 'relation jim', 'invest prospectus', 'material implied', 'hold zmh', 'brookfield ppty', 'sach downgrade',
'part dai', 'mkt vwo', 'toyear yield', 'rosetta resource', 'deloitte consult', 'hit streak']


Topic 5
['file factor', 'annualize cost', 'amex final', 'disease announce', 'update detail', 'eventual lead', 'emergency mee
t', 'include discuss', 'equity condense', 'adjust sale', 'div internet', 'study showe', 'randomize doubleblind', 'lim
it statement', 'group manag', 'moody affiliate', 'publication intend', 'min division', 'approv market', 'announce int
end']


Topic 6
['unit pric', 'announcement original', 'plain exploration', 'transportation relate', 'equity bas', 'gmp security', 's
erv president', 'represent judgment', 'settle chg', 'team season', 'move swift', 'join rank', 'intend vigorous', 'fal
interest', 'export ton', 'reserve barrel', 'obligation account', 'import rose', 'statement mak', 'trader estimate']


Topic 7
['hold rev', 'secur net', 'ral dollar', 'uncertainty predict', 'complet expansion', 'moody aaa', 'moody standard', 'c
linton expect', 'figure parenthese', 'applicable report', 'recur expense', 'earlier add', 'chg corporate', 'loan nonp
erform', 'sun hydraulic', 'debt capita', 'severance relate', 'vary applicable', 'book period', 'rate entity']


Topic 8
['weapon inspector', 'schwab restrict', 'seat fill', 'agreement fil', 'source respect', 'detail follow', 'knight trad
e', 'active pursue', 'russian force', 'traveler property', 'card issu', 'website moody', 'creditlink note', 'add expe
ct', 'group urg', 'newly form', 'trader call', 'bas good', 'conjunction financial', 'borrow capacity']


Topic 9
['report representation', 'product cent', 'item present', 'tighten cycle', 'borrow share', 'thoma title', 'symbol exc
h', 'dilute total', 'reynold american', 'discuss result', 'state texa', 'teacher ret', 'price quot', 'careful evaluat
e', 'rate tighten', 'risk pos', 'davi president', 'bas primarily', 'restat credit', 'jame title']
```

C Latent Factor Loadings on Words

For completeness, I include latent factor loadings on specific words.

Factor 0		Factor 1	
term		term	
idc	0.731940	electronic	0.323850
interdigital	0.714344	network	0.313959
patent	0.576599	kosovar	0.220692
reexamination	0.538327	sepr	0.215262
claim	0.344501	accretive	0.207714
mcic	0.298112	perjury	0.207671
espirito	0.295541	nrgy	0.205598
hollow	0.285888	regent	0.204793
annualise	0.282220	semiconductor	0.202433
bavaria	0.280075	amende	0.201818
trmk	0.278730	westpac	0.200344
cbh	0.272440	hazardous	0.200211
indebt	0.272079	logistic	0.199949
trt	0.270419	overturn	0.197501
burgeon	0.269009	banque	0.197373
jdcom	0.268884	additional	0.196876
boughtdeal	0.268772	proxim	0.196322
ike	0.266254	cagr	0.196265
visual	0.264726	assess	0.196014
usdbrr	0.263821	vasomedical	0.195313
argo	-0.245938	enhancer	-0.190860
fairfax	-0.248111	semen	-0.193043
carney	-0.249933	adhesive	-0.193896
wayfair	-0.252187	brunei	-0.195267
miyako	-0.254580	laird	-0.195348
appease	-0.254786	sekb	-0.195614
chaparral	-0.256669	dead	-0.196592
pbf	-0.258850	warner	-0.199054
comml	-0.260105	ipv	-0.199153
reference	-0.261103	delphi	-0.199198
stumble	-0.262501	thqi	-0.199258
bunny	-0.262949	analyst	-0.199642
sunstone	-0.262984	hov	-0.200605
persian	-0.263058	materiality	-0.200686
vps	-0.265582	brat	-0.202166
sworn	-0.266314	corpo	-0.210134
state	-0.276498	margot	-0.216352
satisfactorily	-0.284186	olmert	-0.220184
transwitch	-0.284836	vki	-0.242260
kaufman	-0.286309	system	-0.246526
Name: 0, dtype: float64		Name: 1, dtype: float64	

Factor 2

term	
dril	0.366050
oil	0.307113
gas	0.224958
tcf	0.201237
lighthouse	0.185377
began	0.185179
karl	0.185036
performance	0.179255
eclipsys	0.178169
antm	0.177812
rig	0.176772
azumi	0.176405
hydro	0.175367
friedrich	0.173481
spi	0.172430
astrom	0.171705
indbusiness	0.171522
money	0.169988
ncf	0.169906
retailer	0.168879
activism	-0.162030
juicy	-0.162114
lucia	-0.164117
remission	-0.164420
fastenal	-0.164492
novelli	-0.165585
omn	-0.166301
reconstitution	-0.166383
dba	-0.168242
sandridge	-0.170095
downe	-0.170521
network	-0.171157
inlet	-0.171807
bloomfield	-0.173084
anderson	-0.176138
observance	-0.179856
solari	-0.181082
archipelago	-0.186661
boj	-0.191068
fico	-0.193179

Name: 2, dtype: float64

Factor 3

term	
oil	0.204016
product	0.180014
unleash	0.172947
ochziff	0.171211
viisage	0.167746
symantec	0.158602
electronic	0.156888
enforcement	0.150497
legacy	0.148063
assertive	0.144348
network	0.144171
ssti	0.143569
tiscali	0.142959
privatelyheld	0.142871
parson	0.142750
anss	0.141541
yoon	0.141403
dril	0.141199
mti	0.138848
blogdeal	0.138830
fdacleared	-0.141716
qtd	-0.142549
primetass	-0.144415
florida	-0.144483
securitas	-0.145115
jol	-0.145839
bls	-0.147186
gmt	-0.148913
glenn	-0.149501
segregate	-0.150485
environmentalist	-0.151721
enskilda	-0.153331
siddiqi	-0.153741
bki	-0.153985
provider	-0.154824
sandridge	-0.160277
sale	-0.161691
repeatcorrect	-0.162611
jda	-0.165273
loss	-0.170678

Name: 3, dtype: float64

Factor 4

term	
atmi	0.104478
defuse	0.101162
research	0.096317
withdrawal	0.094099
arab	0.093126
nfs	0.092362
gas	0.090399
lwr	0.090367
investorowned	0.089513
wpc	0.088078
cii	0.088054
convergent	0.087373
talkeurope	0.086774
refi	0.086201
meridien	0.086173
amdoc	0.085964
exercisable	0.085314
veracruz	0.085104
supplying	0.084829
rate	0.084826
shipowner	-0.085429
alpine	-0.085737
emulex	-0.088401
poorer	-0.088994
bookrun	-0.089298
revocation	-0.089525
radiometric	-0.089840
shr	-0.090211
interphase	-0.090457
bavarian	-0.091320
hutcheson	-0.094283
structure	-0.094342
single	-0.094694
company	-0.097401
eurdem	-0.102768
xffi	-0.108552
share	-0.112276
system	-0.120811
technology	-0.136846
network	-0.182957

Name: 4, dtype: float64

Factor 5

term	
stock	0.105162
aso	0.095713
supplier	0.093712
company	0.091525
optionality	0.090833
share	0.088052
industry	0.083350
net	0.080414
headon	0.079847
planar	0.079842
lobbied	0.077905
terex	0.074196
lynx	0.073144
africanamerican	0.072901
spc	0.072476
earthmov	0.072405
perth	0.072303
customer	0.072128
loss	0.072047
handler	0.071481
ecology	-0.071494
aws	-0.071817
indefinite	-0.072081
expansion	-0.072814
tie	-0.072856
resecuritization	-0.073351
statewide	-0.073560
patent	-0.073868
scheringpl	-0.074208
viisage	-0.074922
divine	-0.075562
riga	-0.076087
trc	-0.078217
reap	-0.080502
cwtr	-0.080713
symantec	-0.081486
hanna	-0.083092
samestore	-0.084152
store	-0.096202
system	-0.105192

Name: 5, dtype: float64

Factor 6

term	
sale	0.096112
earn	0.080587
store	0.076467
time	0.071725
gable	0.065994
infusion	0.064357
sie	0.062972
usefulness	0.055970
semiconductor	0.055732
von	0.054986
filenet	0.054733
patent	0.054006
automobile	0.053485
salvage	0.053304
wabash	0.052854
lagardere	0.052428
shaanxi	0.052298
hugin	0.051813
gandhi	0.051705
integram	0.051452
micromuse	-0.051726
cinematic	-0.052198
column	-0.052686
kopin	-0.052991
halfdozen	-0.053006
samerestaurant	-0.053405
bunny	-0.053437
dynamex	-0.054536
arabica	-0.055243
pay	-0.056003
mandalay	-0.057510
ass	-0.058368
realnetwork	-0.058454
att	-0.060242
shiloh	-0.060343
mcgee	-0.062954
affirmation	-0.064314
blindness	-0.065917
altria	-0.066135
fil	-0.071169

Name: 6, dtype: float64

Factor 7

term	
oil	0.059503
compar	0.055151
loss	0.051935
market	0.051652
catz	0.049970
gmt	0.046286
wood	0.046179
stream	0.046145
energy	0.045085
loan	0.044726
markup	0.044399
trade	0.044378
zucker	0.042027
meta	0.041878
credit	0.041766
sympathizer	0.041650
suit	0.040740
administrative	0.040654
service	0.040373
headline	0.040364
collar	-0.038638
defector	-0.039059
disposal	-0.039174
knxa	-0.039218
discovery	-0.039511
outfitter	-0.040101
writeoff	-0.040220
expansion	-0.040741
xechem	-0.040754
qualifi	-0.041539
cfr	-0.041620
refresh	-0.041684
holde	-0.042010
aviva	-0.042240
taxcut	-0.042405
annoy	-0.043463
esq	-0.045562
holderseq	-0.046617
ges	-0.046676
udi	-0.050753

Name: 7, dtype: float64