# Risk Factors That Matter: Textual Analysis of Risk Disclosures for the Cross-Section of Returns

Alejandro Lopez-Lira*

The Wharton School, University of Pennsylvania

January 2019

**Abstract**

Using unsupervised machine learning, I introduce interpretable and economically relevant risk factors that characterize the cross-section of returns better than the leading factor models, furthermore, I do not use any information from the past returns to select the risk factors. I exploit natural language processing techniques to identify from the firms' risk disclosures the types of risks that firms face, quantify how much each firm is exposed to each type of risk, and employ the firms' exposure to each type of risk to construct a 4-factor model. The risk factors roughly correspond to Technology and Innovation Risk, Demand Risk, Production Risk and International Risk.

***Keywords*** Cross-Section of Returns, Factor Models, Machine Learning, Big Data, LDA, Text Analysis, NLP

*I am grateful to Joao Gomes, Amir Yaron, Jules van Binsbergen, Nick Roussanov, Winston Dou, Lars Hansen, Thomas Sargent, Seth Pruitt, Alexandr Belyakov and Marco Grotteria for helpful discussions and feedback. I am grateful and acknowledge the financial support provided by the Mack Institute for Innovation Management and the Rodney L. White Center for Financial Research. Please email me for the latest version, comments or suggestions: joselop@wharton.upenn.edu

# 1 Introduction

The goal of most of the empirical cross-sectional asset pricing literature is to explain why different assets earn different returns, usually by specifying a linear factor model for excess returns:

$$r^e_{i,t+1} = \alpha_{i,t} + \beta'_{i,t} f_{t+1} + \epsilon_{i,t+1}, \tag{1}$$

with the important economic restriction[1] that $\alpha_{i,t} = 0$. There is no consensus on how to choose the factors, and researchers generally use one of three different approaches: empirical factors models explicitly designed to overcome anomalies (portfolios of returns that are known to generate $\alpha$), factors constructed from economic theory, and statistical factor models.

All of these approaches, however, are beset with serious concerns,[2] the main ones being data mining, overfitting and p-hacking (that is, finding spurious correlations by repeatedly testing variations of the data); concerns about the theoretical motivations behind factors and characteristics designed to address anomalies in the cross-section of returns;[3] and concerns about the absence of economic interpretability regarding purely statistical factors.[4]

To avoid the usual concerns about factor models, I propose a new way of modelling the cross-section of returns: I use machine learning techniques to extract from the 10-K Annual Reports[5] all of the risks that companies disclose, and use these risks to construct a factor model with economically relevant risk factors.

Importantly, the machine learning approach that I use differs significantly from the usual approaches in statistical factor models or supervised machine learning. A typical statistical factor model uses realized returns to find the factors that best fit the cross-section. In my approach, I

1. In this case $f_{t+1}$ is a vector of excess returns. Additionally we require $E_t[\epsilon_{i,t}] = E_t[e_{i,t+1} f_{t+1}] = 0$ and $E_t[f_{t+1}] = \lambda_t$.

2. See Cochrane (2011), Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016), and Hou, Xue, and Zhang (2017).

3. Fama and French (1993): "The choice of factors, especially the size and book-to-market factors, is motivated by empirical experience. Without a theory that specifies the exact form of the state variables or common factors in returns, the choice of any particular version of the factors is somewhat arbitrary." See also Kelly, Pruitt, and Su (2018).

4. See Kozak, Nagel, and Santosh (2018).

5. While it is reasonable to have concerns about how truthful and informative are the risk disclosures, there exists ample evidence in the accounting literature that risk disclosure are, indeed, useful and informative, see Section 3 for an extensive discussion.

do not use any information from the returns; I use only the information that companies disclose in their annual reports. Hence, the factors that I identify have a clear economic meaning and are fully and easily interpretable, rather than being a linear combination of returns.

I use the risks revealed by the companies to solve the common concerns about factor models in the following way: First, to the best of my knowledge, I am the first to use the 10-K textual data to construct a factor model,[6] so the concerns about data mining or p-hacking should be minimal. Second, instead of defining some risk factors or characteristics that seem subjectively important, I take them directly from the firms, since they are the ones that best understand the risks they face. Finally, the factors unambiguously represent economic risk faced by the companies.

To accomplish this, I characterize the types of risks that public companies consider the most relevant and choose to disclose on their annual reports and quantify how much each company is exposed to each type of risk. Equipped with the per firm proportion of the exposure to the disclosed risks, I sort firms to get portfolios exposed to each specific risk, and form a factor model using the risks considered important by each of the firms. I test the capacity of these factors to price the cross-section of returns using the set of 25 Book-to-Market and 49 Industry Portfolios available from Kenneth French's website.[7]

The main contribution of the paper is using machine learning techniques to discover interpretable and economically relevant risk factors. A second, and equally important contribution, is that the four-factor model has a better statistical fit than the leading models in the literature: the factor models of Fama and French (2015), Stambaugh and Yuan (2017) and Hou, Xue, and Zhang (2015). For example, using the GRS test,[8] which tests the null of no-mispricing ($\alpha_i = 0$), and where lower values of the GRS statistic correspond to lower evidence of mispricing (and higher p-values): with the set of 49 industry portfolios, the GRS statistic is .88, with a corresponding p-value of 68% so we cannot reject the no-mispricing null; compare to the GRS statistic of 1.55 for the Fama and French (2015) model with a p-value of 4.5% in which we can reject the no-mispricing null.[9] The

6. Note, however, that Israelsen (2014) was the first to apply LDA to construct portfolios, and uses them to perform factor analysis of the Fama-French factors

7. I choose these portfolios as the test set following the critique of Lewellen, Nagel, and Shanken (2010).

8. Gibbons, Ross, and Shanken (1989)

9. Additionaly, it succeeds in explaining a large fraction of the time series variation of the cross-section of returns (measured by an average $R^2$ of 63 %, comparable to the 68% average $R^2$ obtained with the Fama and French (2015) Model). However, Lewellen, Nagel, and Shanken (2010) advise against using $R^2$ to compare

results are even sharper when we consider the joint set of 49 industry portfolios, 25 book-to-market portfolios, and 11 anomaly portfolios. See Section 6 for details.

**Table 1:** Percentages of the risk disclosure that Intel Corp. allocates to each risk

| International Risk | Production Risk | Software Services Risk | Internet Risk | Technology and Innovation Risk |
|---|---|---|---|---|
| 0.37 | 0.34 | 0.08 | 0.07 | 0.06 |

The table shows the percentages of the risk disclosure that Intel Corp. allocates to each type of risk in the Section 1A: Risk Factors for their 2010 annual report. The table only shows the five most discussed risks. The values are obtained using Latent Dirichlet Allocation. See Sections 3 and 5 for details

I extract the risks that companies disclose from the 10-K Annual reports using Latent Dirichlet Allocation (LDA),[10] a technique developed in the machine learning literature. LDA is a topic modelling technique that summarizes the risks that firms are concerned about, and how much time each company spends discussing each risk.[11] With the additional observation that when a company spends a longer time discussing a specific risk, they are more exposed to that risk,[12] we get a relative measure of how much each firm is exposed to each type of risk.

With the LDA algorithm we get a set of 25 risks. For parsimony and to be comparable with existing factor models, I select only 4 risks to form the factor model. Ex-ante, the risks reported by the companies could be either idiosyncratic or systematic. I select the systematic risks by choosing the risks that most companies spend more time discussing.[13] Note that no information about the returns is used to select this factors, neither from the test set (the 49 industry portfolios, the 25 book-to-market portfolios and the anomalies), nor from the 25 potential risk factors. I use only information disclosed by the firms so I completely avoid any concern about p-hacking or data mining, in this sense, the tests are actually out-of-sample tests.[14] Nevertheless, the high average returns and Sharpe ratios of the risk factors provide additional evidence that we are indeed

between models.

10. Blei, Ng, and Jordan (2003)

11. See Section 4 for a detailed explanation

12. See Gaulin (2017) and Campbell et al. (2014) for empirical evidence: "the type of risk the firm faces determines whether it devotes a greater portion of its disclosures towards describing that risk type". See Section 3 for an extended discussion.

13. See Section 6 for details

14. See the Appendix for other ways to select the Risk Factors

capturing systematic risks.[15][16]

The factor model I form using the firms' disclosed risks complements the literature in the following ways. First, regarding statistical factor models:[17] while they provide an outstanding statistical fit, it is hard to understand the economics of these factors and whether they represent risk; are generated by behavioral patterns; or represent market inefficiencies (Kozak, Nagel, and Santosh (2018)), whereas by design, the factors constructed from the firms' risk disclosures represent economic risk.

Second, regarding empirical factor models: while they succeed in explaining empirically puzzling portfolios (portfolios with $\alpha \neq 0$) [18], they usually do so by iteratively adding (some of) the existing anomalies as risk factors.[19] However adding anomalies as risk factors naturally generates too many factors, what has been referred a "factor zoo" (Cochrane (2011)), and disentangling the true risk factors from the anomalies is a complicated endeavor.[20] To complicate things further, there are important concerns as to which of these anomalies are significant out-of-sample (Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016)), so adding them as risk factors is at best, risky. Since, by construction, all of the factors in the paper, represent risk, it suffices to identify which of these factors are priced to get a set of risk factors that explains the cross-section.

Finally, regarding economic theory models: we know from Merton (1973) that the risk premia of every asset depends on the covariances of the firms' cash-flows with the market wealth and other state variables that affect the stochastic discount factor (SDF). Any characteristic of the firms that makes their dividends covary with either wealth or state variables would affect returns. Asking researchers to identify most of these variables seems like an unworkable task, and while we usually consider a model successful if it can explain most of the variation in the underlying phenomena, in asset pricing, explaining 90 % of the portfolios still leaves hundreds of anomalies unexplained. Firms, however, have a much better understanding of the risks they are facing. Hence,

15. The Risk Factors roughly correspond to Technology Risk, Demand Risk, Production Risk and International Risk, see Section 5

16. See Table 4.

17. See Kelly, Pruitt, and Su (2018), Kozak, Nagel, and Santosh (2018) and Section 2 Related Literature

18. See for example Fama and French (1992), Fama and French (1993), Fama and French (2015)

19. See for example Fama and French (1992), Fama and French (1993), Fama and French (2015), Hou, Xue, and Zhang (2015), and Stambaugh and Yuan (2017) among many, many others

20. Feng, Giglio, and Xiu (2017) however, provide some hope to succeed in this endeavor.

understanding which risks firms face can provide guidance on how to improve our theoretical models.

The paper continues as follows: Section 2 provides a (brief) literature review; Section 3 describes the data sets and addresses concerns about the reliability of the annual reports; Section 4 describes extensively the process to recover the types of risks from the annual reports; Section 5 describes the characteristics of these risks; Section 6 describes the risk factors and tests; and Section 7 concludes. The techniques used to extract the risks are fairly new to the finance field, so I encourage the reader to read Section 4 before Section 5 or Section 6.

# 2 Related Literature

My paper makes contributions in two different branches of literature: (1) machine learning and big data methods in finance, and (2) cross-sectional asset pricing.

I contribute to the recent tradition of employing text analysis to study a variety of finance research questions (see Loughran and McDonald (2016) for a systematic review). Some papers employ text analysis to study a specific risk that the researchers have in mind (e.g. Hassan, Hollander, van Lent and Tahoun (2017) for political risk; Hanley and Hoberg (2017) for financial risk). I instead, do not specify any risk ex-ante and instead let them arise naturally from the data using machine learning methods. I build on the research of Israelsen (2014), who uses LDA to perform style analysis between disclosed risks and the Fama-French factors. I complement his paper by using the risks disclosed by firms to form a factor model that explains the cross-section of returns.

I contribute to the literature of big data methods in finance by using unsupervised machine learning to reduce the dimentionality and size of the annual reports. The raw reports occupy hundreds of gigabytes, and the document term matrix is around 10,000x10,000. By using LDA, we get a set of 25 interpretable risk topics, as well as how much company spends discussing each risk. I then reduce the dimentionality even further and focus on 4 systematic risk factors.

My paper is of course related to the large literature on cross-sectional stock returns (see, e.g., Cochrane (1991); Berk, Green, and Naik (1999); Gomes, Kogan, and Zhang (2003); Nagel (2005); Zhang (2005); Livdan, Sapriza, and Zhang (2009);Eisfeldt and Papanikolaou (2013); Kogan and

Papanikolaou (2014)). See Harvey, Liu, and Zhu (2016) for a recent systematic survey. However, to the best of my knowledge, this is the first paper to construct a successful factor model using all of the risks disclosed by the firms.

# 3 Data

I use three sources of data: the 10-Ks Annual Reports, Compustat, and CRSP.

## 3.1 10-K Annual Reports

Firms disclose in their annual reports which types of risk they are facing. There can be some concerns about how true and informative these disclosures are, however, there exists ample evidence that the risk disclosures are, indeed, useful and informative. First, firms are legally required to discuss "the most significant factors that make the company speculative or risky" (Regulation S–K, Item 305(c), SEC 2005) in a specific section of the 10-K annual reports (Section 1A) and could face legal action if they fail to obey the regulation. Additionally, Campbell et al. (2014) find that "the type of risk the firm faces determines whether it devotes a greater portion of its disclosures towards describing that risk type... managers provide risk factor disclosures that meaningfully reflect the risks they face and the disclosures appear to be... specific and useful to investors".

I extract the textual risk factors in Section 1A (mandatory since 2005) of each 10-K Annual Report. I collect the 10-Ks from 2005 to 2018 from the EDGAR database on the SEC's website. The 10-Ks come in many different file formats (.txt., .xml, and .html) and have different formatting, so it is quite challenging to automatically extract the Section 1A-Risk Factors, from the 10-K forms. To do so, I first detect and remove the markup language and then use regular expressions with predefined heuristic rules. I end up with a data set consisting of 79304 documents.

To illustrate the kind of disclosures that firms make, consider the excerpt from Apple Inc.'s 2010 10-K annual report below. I incorporate suggested labels regarding the type of risk, and highlight possible key words in red. Note that both labels and key words are just for illustrative purposes, and there is no need to manually label the risks in the paper or define the keywords, since the risk factors will arise naturally using the LDA algorithm.

- Currency Risk: Demand ... could differ ... since the Company generally raises prices on goods and services sold outside the U.S. to offset the effect of the strengthening of the U.S. dollar change.

- Supplier Risk: The Company uses some custom components that are not common to the rest of the personal computer, mobile communication and consumer electronics industries.

- Competition Risk: Due to the highly volatile and competitive nature of the personal computer, mobile communication and consumer electronics industries, the Company must continually introduce new products

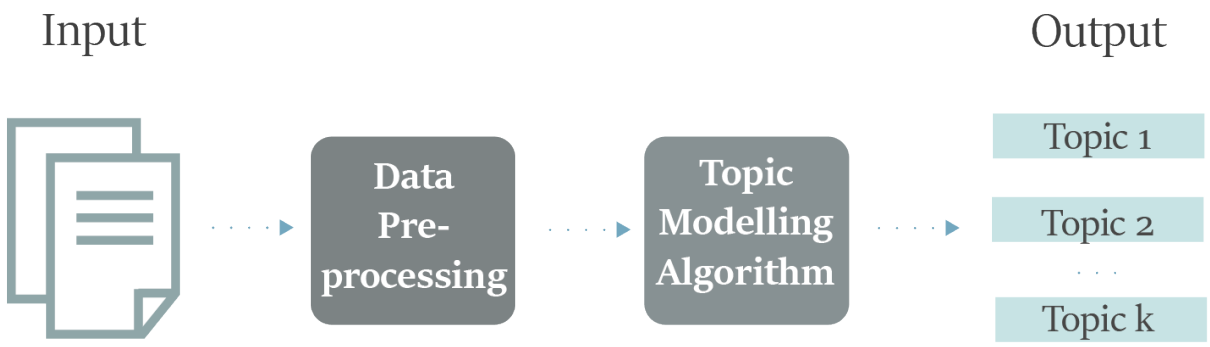
## 3.2 CRSP and Compustat

I follow the usual conventions regarding CRSP and Compustat data. I focus on monthly returns since the disclosures are done annually. For the accounting and return data, I use the the merged CRSP/Compustat database. I use annual firm-level balance sheet data from Compustat due to concerns about seasonality and precision; and monthly returns from CRSP. I use data from the same period as the one where 10-Ks are available: 2005-2018, although not all variables are available for every period. I exclude from the main analysis firms in industries with SIC codes corresponding to the financial industry (SIC in [6000, 7000]).

The Five Factors of Fama and French (2015), the momentum factor, and the one-month Treasury-bill rate come from the French data library on Ken French's website. The Stambaugh and Yuan (2017) factors come from their website. The q-factors of Hou, Xue, and Zhang (2015) come from their website.

# 4 Text Processing

**Figure 1:** Steps for topic modelling




## 4.1 Bag of Words and Document Term Matrix

We need a way to represent text data for statistical purposes. The Bag of Words model achieves this task. Bag of Words considers a text as a list of distinct words in a document and a word count for each word,[21] which implies that each document is represented as a fixed-length vector with length equal to the vocabulary size. Each dimension of this vector corresponds to the count or occurrence of a word in a document. Traditionally, all words are lowercased to reduce the dimension in half.

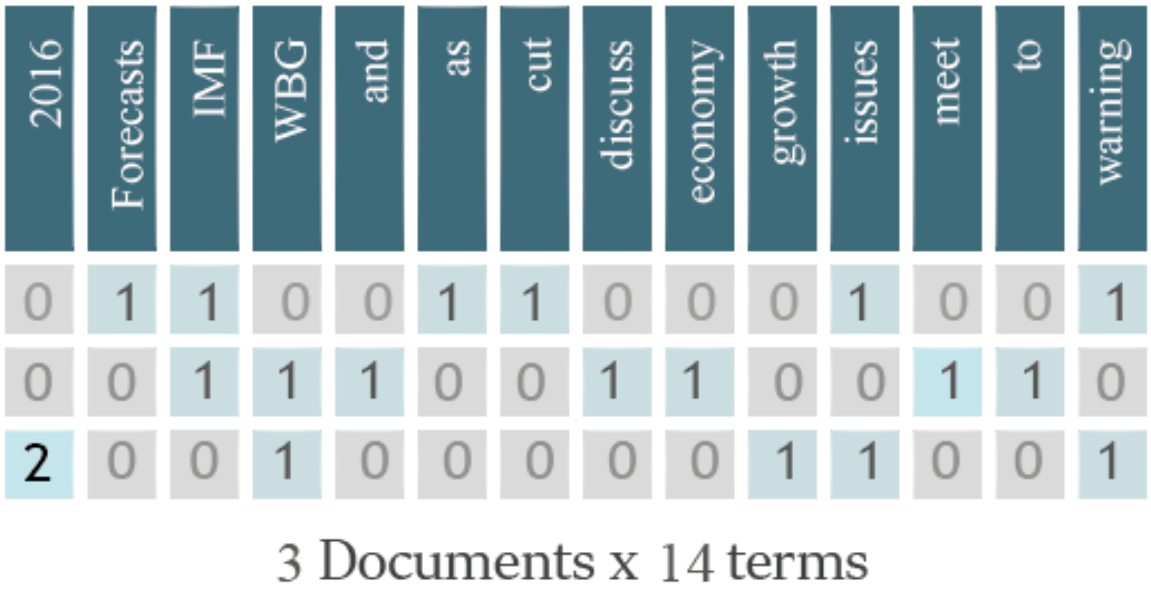
It is called a "bag" of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document. Notice that since we only consider the count, the order of the words is lost. When we consider several documents at a time, we end up with a Document Term Matrix (DTM), see Figure 2 for a simplified example. The DTM is typically highly dimensional (> 10,000 columns), since we consider the space of all words used across all documents; it is also very sparse, since typically documents do not use the whole English vocabulary. Because of the huge dimension of the space, we need a dimensionality reduction technique, such as LDA.

Another subtle disadvantage of the Bag of Words model, is that it breaks multi-word concepts such as "real state" into "real" and "state", which have to be rejoined later, since counting those

21. Manning, Raghavan, and Schütze (2008)

words separately will produce different results than counting the multi-word concept.

**Figure 2:** Example of a very simple document term matrix




## 4.2 Preprocessing

It is common to preprocess the raw text in several steps in order to make the topics more interpretable and to reduce the dimension. The purpose is to reduce the vocabulary to a set of terms that are most likely to reveal the underlying content of interest, and thereby facilitate the estimation of more semantically meaningful topics.

I remove common English words ("the", "and", "or", etc.) and additional terms that do not convey any meaning or are considered legal warnings in the 10-K ("materially adverse", "no assurance", etc.) in order to extract only risks from the text. See the appendix for a full list and a detailed explanation.

Some words represent the same underlying concept. For example, "copy", "copied", and "copying"; all deal with either a thing made to be similar or identical to another or to make a similar or identical version of. The model might treat them differently, so I strip such words to their core. We can achieve this by either stemming or lemmatization, which are fundamental text processing

methods for text in the English language.

Stemming helps to create groups of words that have similar meanings and works based on a set of rules, such as remove "ing" at the ends of words.[22] Different types of stemmers are available in standard text processing software such as NLTK (Loper and Bird (2002)), and within the stemmers there are different versions such as PorterStemmer, LancasterStemmer and SnowballStemmer. The disadvantages of stemming is that it cannot relate words that have different forms based on grammatical constructs, for example: "is", "am", and "be" all come from the same root verb, "to be", but stemming cannot prune them to their common form. Another example: the word "better" should be resolved to good, but stemmers would fail to do that. With stemming, there is lot of ambiguity that may cause several different concepts to appear related. For example, "axes" is both a plural form of "axe" and "axis". By chopping of the "s", there is no way to distinguish between the two.

Lemmatization in linguistics is the process of grouping together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form(Manning, Raghavan, and Schütze (2008)). In order to relate different inflectional forms to their common base form, it uses a knowledge base called WordNet. With the use of this knowledge base, lemmatization can convert words that have a different form and cannot be solved by stemmers, for example converting "are" to "be". The disadvantages of lemmatization are that it is slower compared to stemming, however, I use lemmatization to preserve meaning and make the topics more understandable.

Phrase Modeling is another useful technique whose purpose is to (re)learn combinations of tokens that together represent meaningful multi-word concepts. We can develop phrase models by looking for words that co-occur (i.e., appear one after another) together much more frequently than you would expect them to by random chance. The formula to determine whether two tokens $A$ and $B$ constitute a phrase is:

$\frac{count(A,B)-count_{min}}{count(A)*count(B)} * N \geq threshold$ , where:

- $count(A)$ is the number of times token $A$ appears in the corpus

22. Manning, Raghavan, and Schütze (2008)

- $count(B)$ is the number of times token $B$ appears in the corpus
- $count(A, B)$ is the number of times the tokens $A$ and $B$ appear in the corpus in that order
- $N$ is the total size of the corpus vocabulary
- $count_{min}$ is a parameter to ensure that accepted phrases occur a minimum number of times
- *threshold* is a parameter to control how strong of a relationship between two tokens the model requires before accepting them as a phrase

With phrase modeling, named entities will become phrases in the model (so new york would become new_york). We also would expect multi-word expressions that represent common concepts, but are not named entities (such as real state) to also become phrases in the model.

## 4.3 Dictionary methods

The most common approach to text analysis in economics relies on dictionary methods, in which the researcher defines a set of words of interest and then computes their counts or frequencies across documents. However, this method has the disadvantage of subjectivity from the researcher perspective, since someone has to pick the words. Furthermore, it is very hard to get the full list of words related to one concept and the dictionary methods assume the same importance or weight for every word. Since the purpose of the paper is to extract the risks that managers consider important with minimum researcher input, dictionary methods are unsatisfactory.

Furthermore, dictionary methods have other disadvantages, as noted by Hansen, McMahon, and Prat (2018):

> For example, to measure economic activity, we might construct a word list which includes "growth". But clearly other words are also used to discuss activity, and choosing these involves numerous subjective judgments. More subtly, "growth" is also used in other contexts, such as in describing wage growth as a factor in inflationary pressures, and accounting for context with dictionary methods is practically very difficult.

For the purpose of studying the cross-section of returns, the problem is similar to picking which characteristics are important for the returns. The dictionary methods would be equivalent to manually picking which characteristics would enter a regression. The following algorithm, Topic Modelling, is akin to automatic selection methods, such as LASSO (Tibshirani (1996)).

## 4.4 Topic Models

A topic model is a type of statistical model for discovering a set of topics that describe a collection of documents based on the statistics of the words in each document, and the percentage that each document allocates to each topic. Since in this case, the documents are the risk disclosures from the annual statements and they only concern risks, the topics discovered will correspond to different types of risks.

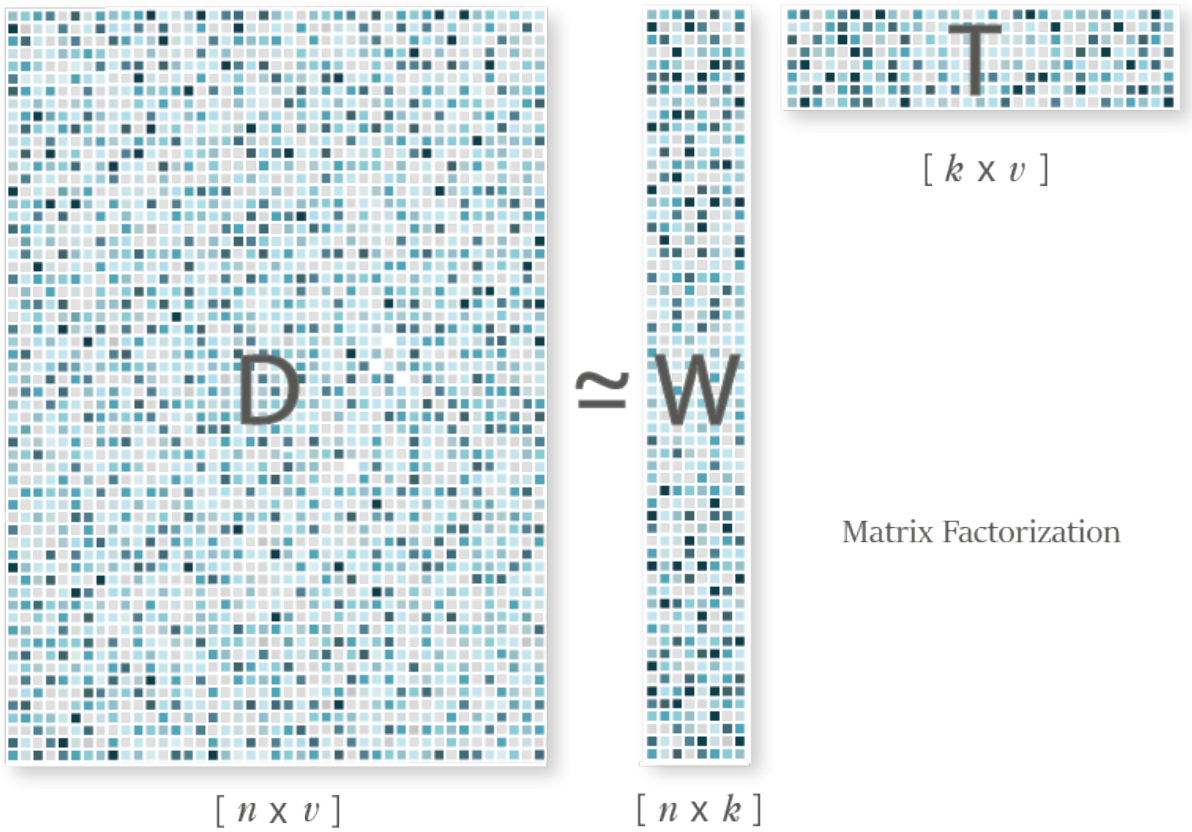
Intuitively, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently. For example: "internet" and "users" will appear more often in documents produced by firms in the technology sector; "oil", "natural gas" and "drilling" will appear more frequently in documents produced by firms in the oil industry, while "company" and "cash" would appear similarly in both.

A document typically concerns multiple topics, or in this case risks, in different proportions; thus, in a company risk disclosure that is concerned with 20% about financial risks and 20% about internet operations, the risk report would approximately have around 8 times more technology words than financial words.

Because of the large number of firms in the stock market, the amount of time to read, categorize and quantify the risks disclosed by every firm is simply beyond human capacity, but topic models are capable of identifying these risks.

The most common topic model currently in use is the LDA model proposed by Blei, Ng, and Jordan (2003). The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. The interaction between the observed documents and the hidden topic structure is manifested in the probabilistic generative process associated with LDA.

**Figure 3:** Intuition for Topic Modelling




## 4.5 LDA

In LDA each document can be described by a (probability) distribution over topics and each topic can be described by a (probability) distribution over words. In matrix algebra terms, we are factorizing the term-document matrix $D$ into a matrix $W$ mapping words to topics, and a matrix $T$ mapping topics to words, similar to the factorization used in Principal Component Analysis, see Figure 3. In this way, LDA reduces the dimensionality of each document, from thousands of words, to the number of topics (25 in our case). However, LDA retains most of the information about the individual word counts, since the topics themselves are probability distribution over words

Formally, LDA is a Bayesian factor model for discrete data that considers a fixed latent set

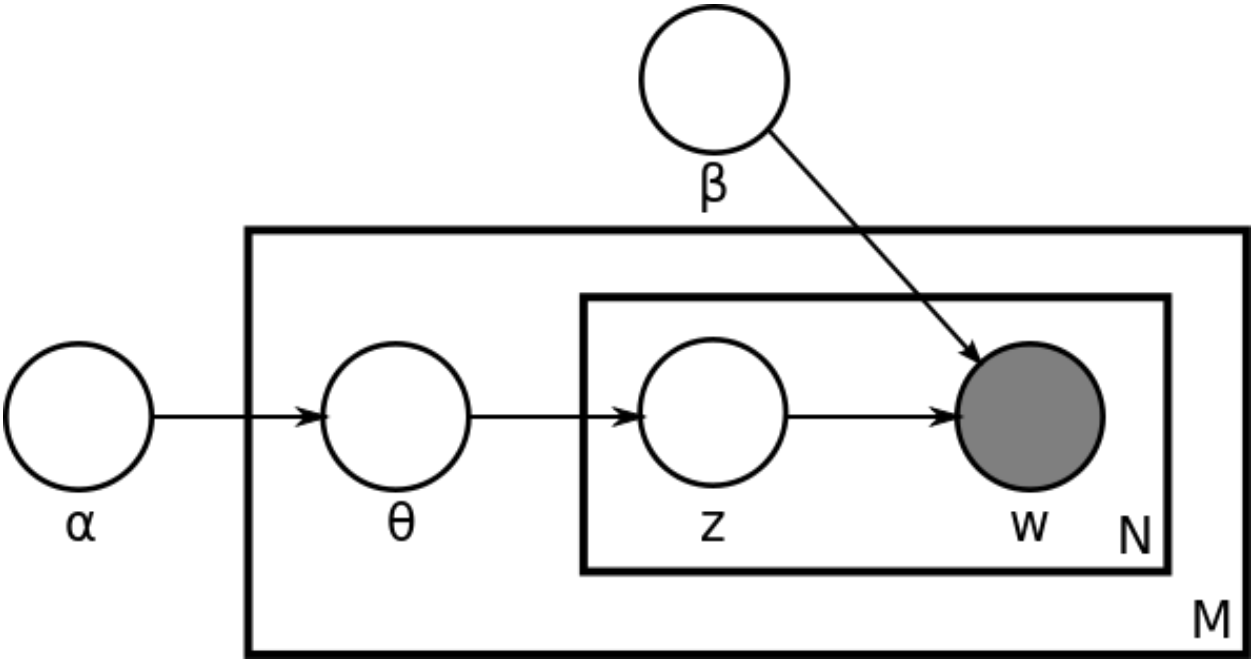
of topics. Suppose there are D documents that comprise a corpus of texts with V unique terms. The K topics (in this case, risk types), are probability vectors $\beta_k \in \Delta_{V-1}$ over the V unique terms in the data, where $\Delta_M$ refers to the M-dimensional simplex. By using probability distributions, we allow the same term to appear in different topics with potentially different weights. We can think of a topic as a weighted word vector that puts higher mass in words that all express the same underlying theme.[23]

In LDA, each document is described by a distribution of topics that appear in the document, so each document d has its own distribution over topics given by $\theta_d$ (in our case, how much each company discusses each type of risk). Within a given document, each word is influenced by two factors, the topics proportions for that document, $\theta_{dk}$, and the probability measure over the words within the topics. Formally, the probability that a word in document d is equal to the nth term is $p_{dn}\theta_d^k$.

It is easier to frame LDA in the language of graphical models, see Figure 4. Where M is the set of all the documents; N is the number of words per document. Inside the rectangle N we see w: the words observed in document i, z: the random topic for the jth word for document i, $\theta$: the topic distribution for document i. $\alpha$: the prior distribution over topics intuitively controls the sparsity of topics within a document (i.e. how many topics we need to describe a document). $\beta$ the prior distribution of words within a topic controls how sparse the topics are in terms of words (i.e. how many words we need to describe a topic). There is a trade-off between the sparsity of the topics, i.e. how specialize they are, and the number of topics.

23. See Blei, Ng, and Jordan (2003) and Hansen, McMahon, and Prat (2018)

**Figure 4:** LDA Graphical Model




### 4.5.1 Number of topics

The number of topics is a hyperparameter in LDA. Ideally, there should be enough topics to be able to distinguish between themes in the text, but not so many that they lose their interpretability. In this case 25 topics accomplish this task, and is consistent with the numbers used in the literature (Israelsen (2014), Bao and Datta (2014)).

There are technical measures such as perplexity or predictive likelihood to help determine the optimal number of topics from a statistical point of view. These measures are rarely used however, because these metrics are not correlated with human interpretability of the model and prescribe a very high number of topics, whereas for topic models, we care about getting interpretable topics (which correspond to the type of risks).

In the case of risk disclosures, a low number ($< 20$) cannot capture the diversity in risk disclosure (e.g. mixes airplanes with hospitals), and a high number ($> 50$) starts capturing very specific industry risks. Another issue is that with a large number of topics, very few firms will have significant exposure to each risk, and hence portfolios exposed to some risks will be poorly diversified. I set the number of topics equal to 25 because since I have around 2500 firm observations per year,

in the case where risks are divided uniformly, we can have at least (ex-ante) 100 firms that are very exposed to each topic.

A natural challenge is then to further reduce the extracted risks into a lower number of portfolios for the cross-section. See Section 6 Portfolios for more details.

### 4.5.2 Estimation

The estimation of the posterior parameters is done using the open-source software Gensim (Řehůřek and Sojka (2010)) which runs on Python. Gensim uses an online Variational Bayes algorithm. Because of the huge size of the collection of annual reports, the use of online algorithms allows us to not load every document into the RAM memory and hence we can estimate the model in a normal laptop. See the Appendix and Hoffman, Bach, and Blei (2010) for details.

# 5 Risk Topics and Risk Factors

To avoid confusion, I refer to the topics obtained using LDA as Risk Topics and to the portfolios formed using these risks as Risk Factors. Hence, the Risk Topics are distribution over words and the Risk Factors are portfolios of stocks. It is important to remember that LDA does not give us labels for the topics, but nevertheless the topics are easily interpretable since they are characterized by the most frequent words as we can see from Figure 5.

**Figure 5:** Risk Topics


In this word cloud of the risks that firms face, a bigger font corresponds to a bigger weight for that word within each topic. See Section 4 for details on the procedure

We can get a general picture of the risks that firms are facing from Figure 5, where I present the 25 risk topics extracted from the 10-K annual reports. Recall from Section 4 that a topic is a probability distribution over words. I discuss here only the risks that I will use for the factor model, the ones that affect most firms: Technology Risk, Production Risk, International Risk and Demand Risk. See the Appendix for the rest.

## 5.1 Portfolios

I construct portfolios exposed to each risk in the following way: Every year, I consider the firms that allocate more than 25% of their last risk disclosure available at June discussing that risk, and value-weight their returns. The results are robust to the choice of the threshold and I discuss other approaches in the Appendix.

To construct excess return portfolios, I substract the risk-free rate instead of the traditional

long-short portfolios. I do so because the short side of the long-short portfolios would consist of firms spend 0% of their risk disclosure discussing a specific risk, but these firms must be allocating their risk disclosures to other risk types,[24] so if we short portfolios with 0% for a specific risk types, the short side would consist of a mixture of risks in an unknown proportion, which would result in not understanding which risk the portfolio is representing. An alternative construction involving long-short portfolios is discussed in the Appendix.

## 5.2 Risk Factor Selection

I select the 4 risks that affect the highest number of firms in 2006 and keep them for the whole sample to avoid data-mining and look-ahead bias. Firms spend on average 36% of their risks disclosures discussing these 4 risks, and allocate the remaining 64% to the other 21 risks. I discuss them extensively in the following section. Briefly, the risk factors correspond to Innovation Risk, Demand Risk, Production Risk and International Risk. I explore other dynamic approaches to select the factors in the Appendix. See Table 2.

**Table 2:** Average proportion of the risk disclosures allocated to each risk for the most discussed risks in the year 2006

| Technology Risk | Production Risk | International Risk | Demand Risk | Total |
|---|---|---|---|---|
| 0.11 | 0.09 | 0.08 | 0.08 | 0.36 |

The table shows the cross-sectional average of each firms distribution over topics for the annual reports of 2006, but only for the four most mentioned topics. See Section 4 for details.

The number of firms in each factor is enough to have a diversified portfolio as we can see from Table 2. The firms that spend discussing more than 25% across the four risk topics, are about half of the firms in the sample.

We can see from Table 4 that the risk factors (selected using the importance to the firms) have high average returns and Sharpe ratios, despite the fact that no information from past or future returns is used. Since we know from economic theory that the most important risks have a high price of risk, this constitutes preliminary evidence that we are indeed capturing systematic risks and we will be able to use these risk factors to explain the cross-section. We will, indeed, see in

24. I discard firms with no risk disclosures.

**Table 3:** Number of firms heavily exposed to each risk

| Year | Technology Risk | Production Risk | International Risk | Demand Risk | Percentage of Total Firms |
|---|---|---|---|---|---|
| 2006 | 413 | 364 | 343 | 264 | 0.54 |
| 2007 | 442 | 354 | 324 | 245 | 0.52 |
| 2008 | 388 | 270 | 356 | 223 | 0.50 |
| 2009 | 355 | 305 | 412 | 215 | 0.51 |
| 2010 | 300 | 275 | 387 | 211 | 0.48 |
| 2011 | 285 | 266 | 422 | 221 | 0.49 |
| 2012 | 258 | 252 | 468 | 202 | 0.50 |
| 2013 | 261 | 237 | 452 | 205 | 0.49 |
| 2014 | 248 | 215 | 479 | 203 | 0.48 |
| 2015 | 230 | 197 | 493 | 196 | 0.46 |
| 2016 | 213 | 171 | 505 | 205 | 0.44 |

The table shows the number of firms that spend more than 25% of the time discussing each topic. See the text for details.

Section 6 that the factor model formed with these four risks performs significantly better than the most used models.

**Table 4:** Selected characteristics of the Risk Factor Portfolios and the Market Portfolio for the period 2007-2018

| | Technology Risk | Production Risk | International Risk | Demand Risk | Market Portfolio |
|---|---|---|---|---|---|
| Mean | 0.96 | 1.13 | 0.73 | 0.86 | 0.70 |
| Sd. | 5.58 | 6.27 | 4.12 | 4.29 | 4.41 |
| Corr. with Mkt | 0.89 | 0.87 | 0.97 | 0.81 | 1.00 |
| Annualized Sharpe Ratio | 0.59 | 0.62 | 0.62 | 0.69 | 0.55 |

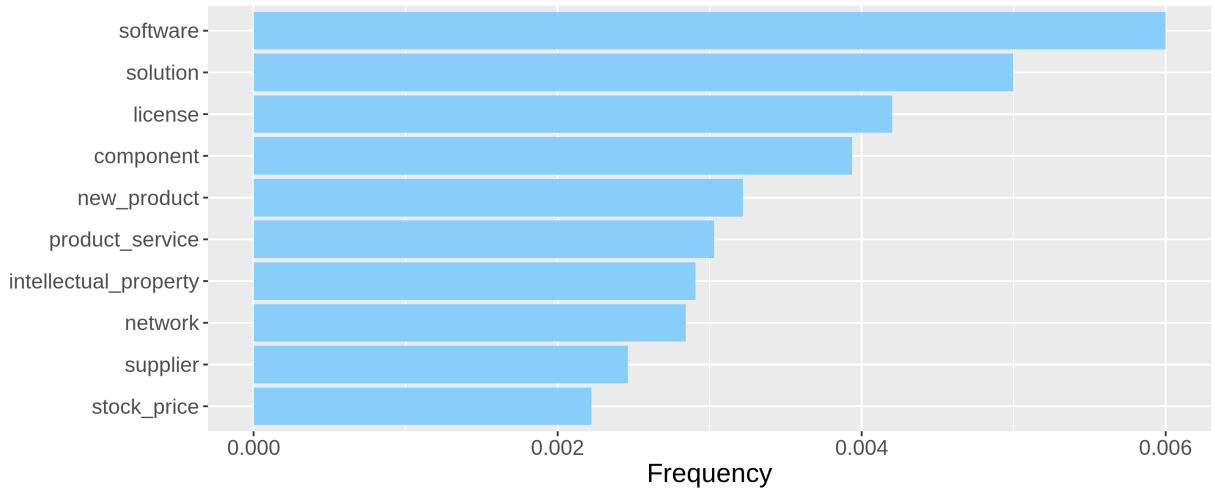
The table shows the monthly average excess return; the monthly standard deviation; the correlation with the market portfolio; and the annualized Sharpe ratio for the selected portfolios. See the text for details.

Next, I discuss with great detail the most important risk topics for the firms and the properties of the risk factors.

## 5.3 Technology and Innovation Risk

The most discussed risk topic, the Technology Risk Topic is characterized by words that have a direct relation to technology and innovation, such as: "software", "new product", "intellectual property", "network"; as we can see from Figure 6. We can see in Table 5 that when we inspect the biggest companies that spend more than 25% of their risk disclosures commenting about the Technology Risk Topic, we see companies that spend a lot of resources in technology: Microsoft,

**Figure 6:** Technology Risk Topic




Distribution of the 10 most frequent words for the Technology and Innovation Risk Topic

Oracle, Cisco, HP, among others.

**Table 5:** Biggest 10 Companies that are exposed more than 25% to the Technology and Innovation Risk Factor

| Company Name | Market Value (Millions) |
|---|---|
| MICROSOFT CORP | 354392 |
| ORACLE CORP | 166066 |
| CISCO SYSTEMS INC | 144516 |
| QUALCOMM INC | 81885 |
| EMC CORP/MA | 49896 |
| HP INC | 48628 |
| ADOBE SYSTEMS INC | 45530 |
| ILLUMINA INC | 28136 |
| VMWARE INC -CL A | 23870 |
| ELECTRONIC ARTS INC | 19873 |

The table shows the biggest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the Technology Risk Topic. See the text for details.

At this point it is natural to wonder whether the risk factors are just capturing industry specific risks, however, it is not the case. We can see from Table 6 that the SIC industries cannot fully capture the relationship between risks and industries. Although as expected the firms that load on the Technology Risk are concentrated mostly in the electronic, computing and business services sectors; the SIC codes are too rigid and put half of these firms in the manufacturing division and

the other half in the services division. The main reason being that industry classification is about what the business of the firm is, so HP and Oracle will look very different in that perspective, one selling computers and the other selling software services, however, they share similar risks, mainly technical challenges and innovation.

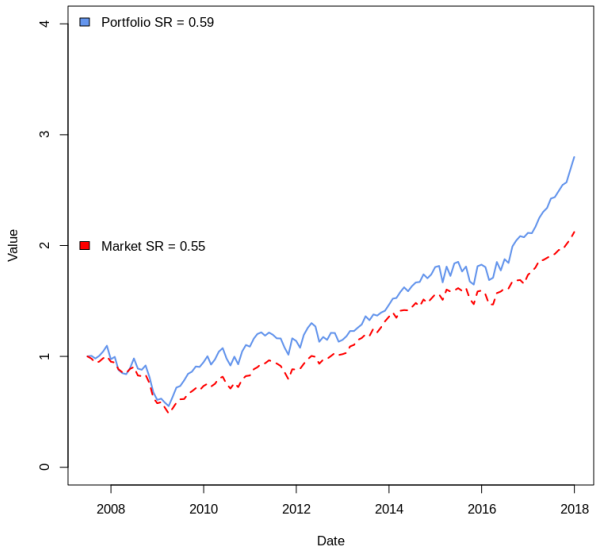
**Table 6:** Number of firms by SIC code for firms that are exposed to the Technology Risk Factor

| 2-Digit SIC Code | Industry | Division | Number of firms |
|---|---|---|---|
| 35 | Manufacturing | Industrial and Commercial Machinery and Computer Equipment | 43 |
| 36 | Manufacturing | Electronic and other Electrical Equipment and Components, except Computer Equipment | 58 |
| 38 | Manufacturing | Measuring, Analyzing, and Controlling Instruments; Photographic, Medical and Optical Goods; Watches and Clocks | 18 |
| 73 | Services | Business Services | 82 |

The table shows the number of firms by SIC code for the firms that allocate more than 25% of their risk disclosure to the Technology Risk Topic. The number of firms is taken at June 2016. I only present the SIC codes for which the number of firms is higher than 15. See the text for details.

The Technology Risk Factor has a Sharpe ratio similar to the market in the period: .59, see Table 4. We can see from Figure 7 that while it fluctuates very much with the business cycles, especially during the financial crisis, the Technology Risk Factor consistently performs better than the market portfolio, but has a higher variance.

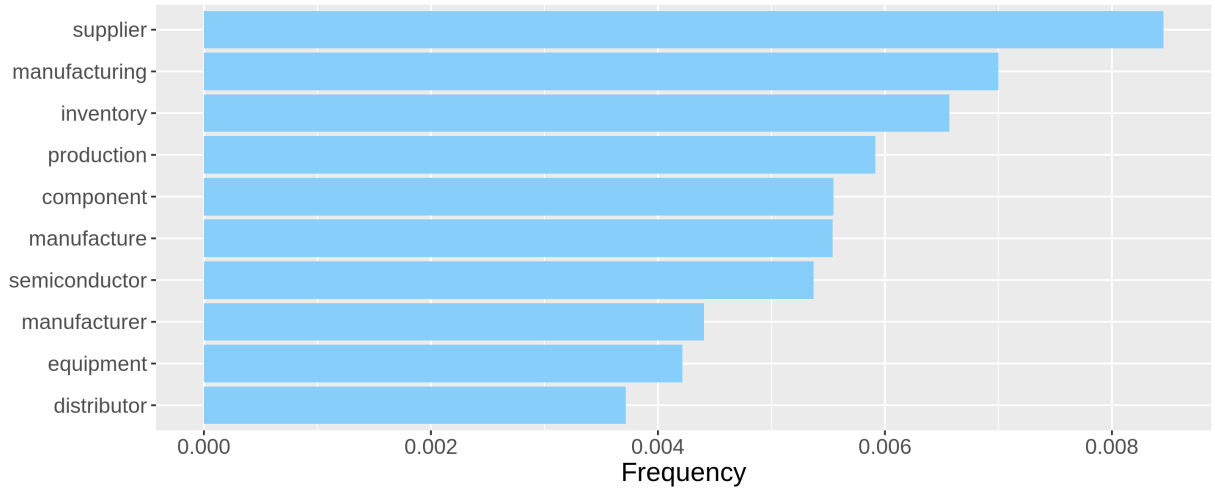
**Figure 7:** Cumulative Return of the Technology and Innovation Risk Factor




Cumulative return of investing one dollar in June 2007. Blue: Technology and Innovation Risk Factor, Red: Market Return, SR: Sharpe Ratio

## 5.4 Production Risk

**Figure 8:** Production Risk Topic




Distribution of the 10 most frequent words for the Production Risk Topic

The Production Risk Topic is characterized by words that have a direct relation to production,

such as: "supplier", "manufacturing", "inventory", "component"; as we can see from Figure 8. We can see in Table 7 that when we inspect the biggest companies that spend more than 25% of their risk disclosures commenting about the Production Risk Topic, we can see companies whose production process seems to be very relevant for the business: Intel, Nvidia, Under Armour, among others.

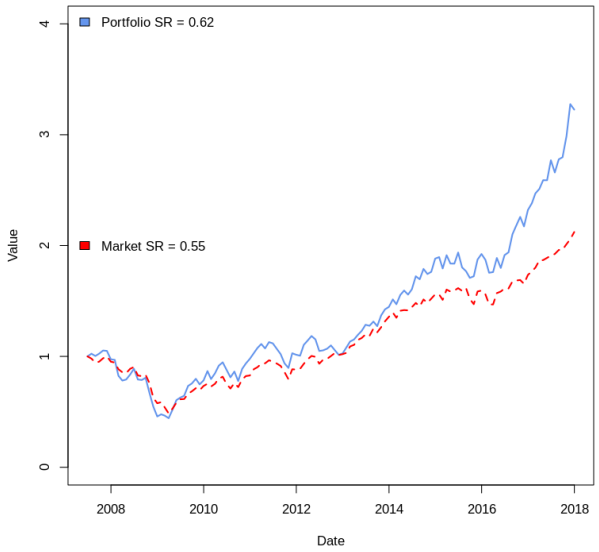
**Table 7:** Biggest 10 Companies that are exposed more than 25% to the Production Risk

| Company Name | Market Value (Millions) |
|---|---|
| INTEL CORP | 162776 |
| TEXAS INSTRUMENTS INC | 55428 |
| APPLIED MATERIALS INC | 19453 |
| ANALOG DEVICES | 18761 |
| WESTERN DIGITAL CORP | 18037 |
| UNDER ARMOUR INC | 17419 |
| MICRON TECHNOLOGY INC | 17050 |
| SKYWORKS SOLUTIONS INC | 16025 |
| NVIDIA CORP | 15787 |
| SEAGATE TECHNOLOGY PLC | 14984 |

The table shows the biggest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the Production Risk Topic. See the text for details.

The Production Risk Factor has a high Sharpe ratio: .62. We can see from Figure 9 that while it is very exposed to business cycles, especially during the financial crisis, but it has a greater average return, 1.13 % monthly to compensate for this risk.

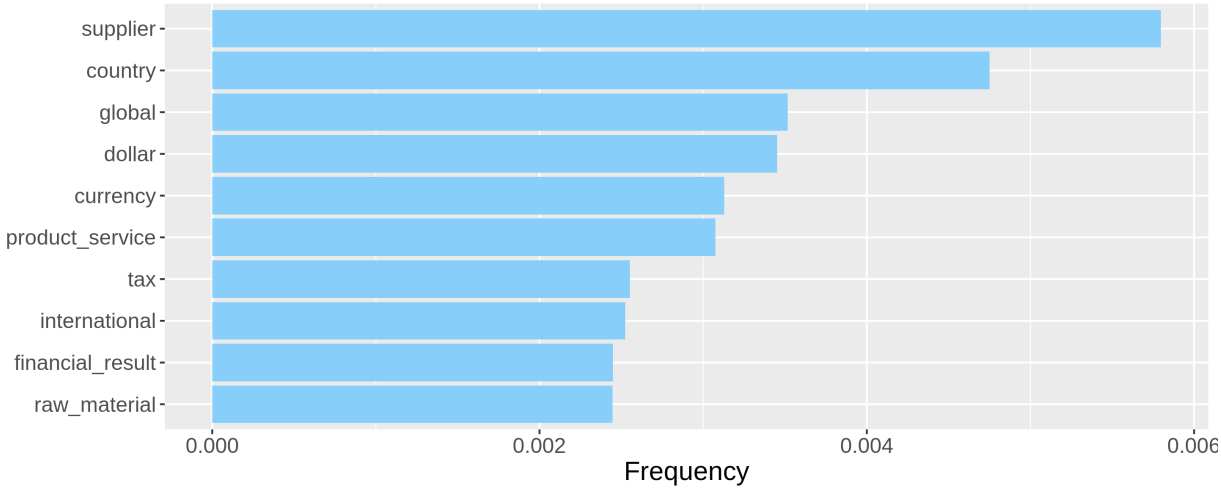
**Figure 9:** Cumulative Return of the Production Risk Factor




Cumulative return of investing one dollar in June 2007. Blue: Production Risk Factor, Red: Market Return, SR: Sharpe Ratio

## 5.5 International Risk

**Figure 10:** International Risk Topic




Distribution of the 10 most frequent words for the International Risk Topic (excluding "company")

The International Risk Topic is characterized by words that have a direct relation to international concerns, such as: "currency", "dollar", "global", "country"; as we can see from Figure 10. We can see in Table 8 that when we inspect the biggest companies that spend more than 25% of their risk disclosures commenting about the International Risk Topic, we can see companies that operate in global markets: Apple Inc, Exxon Mobile, Procter & Gamble, Coca-Cola, among others.

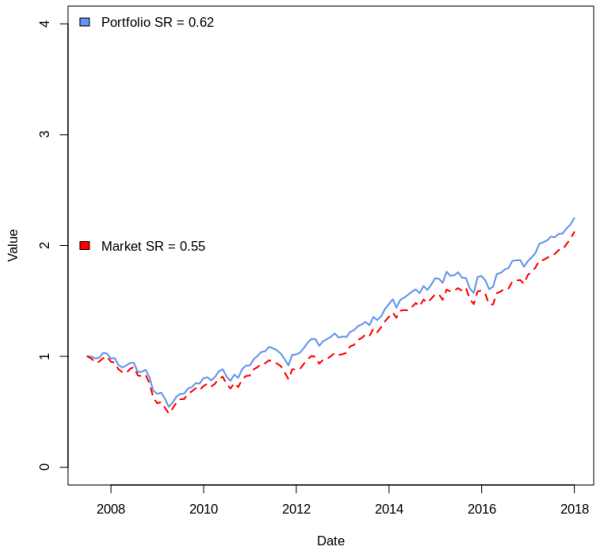
**Table 8:** Biggest 10 Companies that are exposed more than 25% to the International Risk

| Company Name | Market Value (Millions) |
|---|---|
| APPLE INC | 615336 |
| EXXON MOBIL CORP | 323960 |
| PROCTER & GAMBLE CO | 212388 |
| AT&T INC | 211447 |
| PFIZER INC | 199329 |
| COCA-COLA CO | 185759 |
| CHEVRON CORP | 169378 |
| ORACLE CORP | 166066 |
| INTEL CORP | 162776 |
| MERCK & CO | 146899 |

The table shows the biggest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the International Risk Topic. See the text for details.

The International Risk Factor is very positively correlated with the market portfolio (.97), having just a slightly higher mean and a lower variance means it has a higher Sharpe Ratio: .62. We can see from Figure 11 that while it indeed fluctuates with the business cycles, but suffers less during declines and recovers better. Because the International Risk Factor is almost a linear function of the market portfolio we do not need to incorporate it in the factor model to get the average risk premium as is usually the case in the literature.

**Figure 11:** Cumulative Return of the International Risk Factor

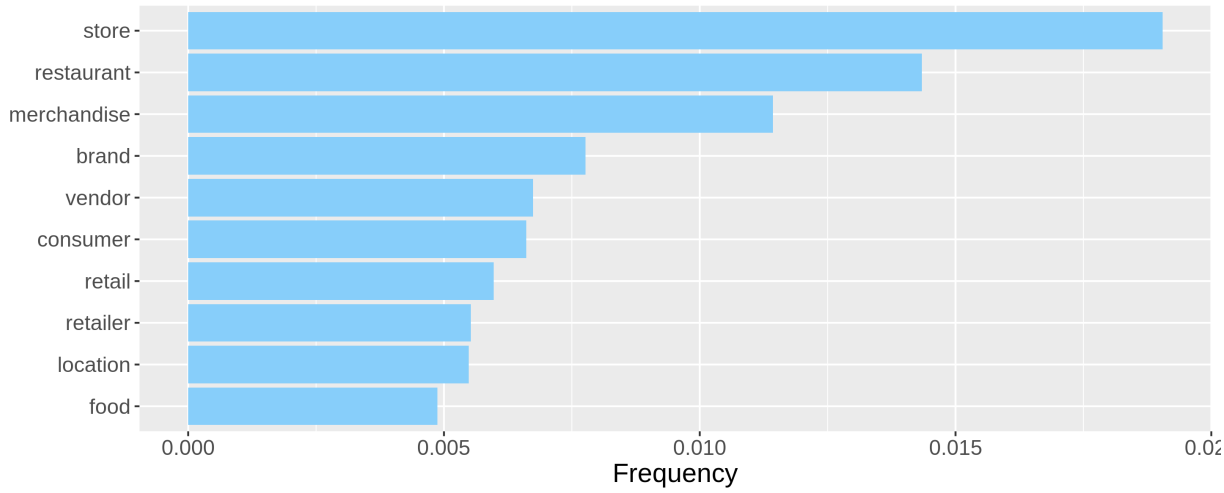



Cumulative return of investing one dollar in June 2007. Blue: International Risk Factor, Red: Market Return, SR: Sharpe Ratio

## 5.6 Demand Risks

The Demand Risk Topic is characterized by words that have are related to demand and sales, such as: "consumer", "store", "retail", and "merchandise"; as we can see from Figure 12. We can see in Table 9 that when we inspect the biggest companies that spend more than 25% of their risk disclosures commenting about the Demand Risk Topic, we see the companies that focus on the consumption sector: Wal-Mart, Home-Depot, McDonald's, Starbucks.

**Figure 12:** Demand Risk Topic




Distribution of the 10 most frequent words for the Demand Risk Topic

**Table 9:** Biggest 10 Companies that are exposed more than 25% to Demand Risks

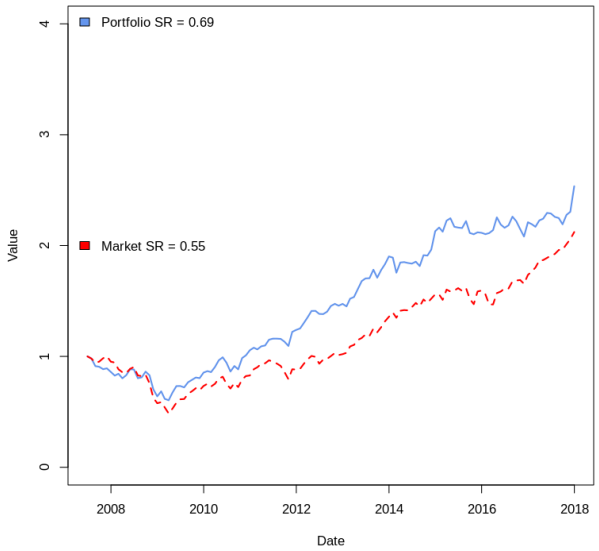
| Company Name | Market Value (Millions) |
|---|---|
| WALMART INC | 209830 |
| HOME DEPOT INC | 157452 |
| MCDONALD'S CORP | 107129 |
| NIKE INC | 92880 |
| STARBUCKS CORP | 84413 |
| LOWE'S COMPANIES INC | 65211 |
| COSTCO WHOLESALE CORP | 61335 |
| TJX COMPANIES INC | 47267 |
| TARGET CORP | 43613 |
| KROGER CO | 37529 |

The table shows the biggest firms by market capitalization measured in June 2016 that allocate more than 25% of their risk disclosure to the Demand Risk Topic. See the text for details.

The Demand Risk Factor has the highest Sharpe ratio among the factors considered: .69. It is

also the least correlated with the Market Portfolio.[25] We can see from 13 that it did worse than the market portfolio during the financial crisis, but outperformed the market during the recovery and stayed above it consistently.

**Figure 13:** Demand Risk Factor




Cumulative return of investing one dollar in June 2007. Blue: Demand Risk Factor, Red: Market Return, SR: Sharpe Ratio

Next, I discuss the power of the risk factors to characterize the cross-section of stock returns.

# 6 Factor Model Comparisons

I test the capacity of the risk factors formed with the risks that affect most firms to price the cross-section of returns. I call the risk factors "text-based factors" for clarity; and use the set of 49 industry portfolios, the set of 25 book-to-market portfolios, and the set of 11 anomaly portfolios to perform the tests following the critique of Lewellen, Nagel, and Shanken (2010).[26] I use the

25. Because the portfolios are value-weighted and the short side corresponds to the risk free rate, it naturally induces high positive correlation with the Market Portfolio

26. See Section 3 for details.

GRS test from Gibbons, Ross, and Shanken (1989). I include the performance of the factor models of Fama and French (2015); Stambaugh and Yuan (2017); and Hou, Xue, and Zhang (2015) for comparison.

Recall that the GRS statistic is a measure of whether $\alpha_i = 0$ and that:

$$GRS \propto \frac{\alpha' \Sigma^{-1} \alpha}{1 + \mu' \Sigma^{-1} \mu}, \tag{2}$$

which we understand as a weighted and normalized sum of the squared alphas, divided by 1 plus the Sharpe ratio of the factors. Intuitively, if the test portfolios are spanned by the factors, we cannot increase the maximum sharp ratio that we get from the factors by adding the test portfolios and $\alpha_i = 0$.

High values of the GRS statistic are indicative of high mispricing errors ($|\alpha_i| \gg 0$), and low values are indicative of low mispricing errors ($\alpha_i \sim 0$). The null hypothesis in the GRS test is that the model is correct: there is no mispricing, the GRS statistic is small and $\alpha_i = 0$, hence, when the p-value is low we have strong evidence against the model and when the p-value is high, there is less evidence to reject the model. Lewellen, Nagel, and Shanken (2010) advice against the use of the average $R^2$ to make comparisons between factor models.

**Table 10:** GRS Test for the 4-Factor Text-Based Model and the Fama-French 5 Factor Model

| | 49 Industry + 25 B-to-M | | | 49 Industry + 25 B-to-M + 15 $\alpha$ | | |
|---|---|---|---|---|---|---|
| | GRS | p-value | $R^2$ | GRS | p-value | $R^2$ |
| Text-based 4 Factor Model | 1.52 | 0.061 | 0.69 | 2.09 | 0.018 | 0.64 |
| Fama-French 5 Factor Model | 1.85 | 0.012 | 0.76 | 3.05 | 0.001 | 0.72 |
| Mispricing Factors | 1.67 | 0.044 | 0.76 | 2.47 | 0.006 | 0.73 |
| q-factor Model | 1.81 | 0.024 | 0.75 | 2.48 | 0.005 | 0.71 |

The table shows the result of the GRS test: high values of the GRS statistic are indicative of high mispricing errors and generate a lower p-value, which is evidence against the fit of the model since the null is of no-mispricing. Lewellen, Nagel, and Shanken (2010) advice against the use of the average $R^2$ to make comparisons between factor models. First row corresponds to the text-based factors presented in the paper, second row corresponds to the Fama and French (2015) Factor Model, third row corresponds to the Anomaly Factors of Stambaugh and Yuan (2017) and fourth row corresponds to the q-factor model of Hou, Xue, and Zhang (2015). I perform the test on the joint set of 49 industry portfolios and 25 book-to-market portfolios available on Kennet French's website in the first column, and include the set of 11 long-short anomaly portfolios of Stambaugh and Yuan (2017) in the second column. See the text for details.

The text-based factor model constructed using the firms' disclosed risks is the best when we consider all portfolios jointly: the 49 industry portfolios, the 25 book-to-market portfolios, and the 11 anomaly portfolios. For the joint set of 25 book-to-market and 49 industry portfolios: The GRS statistic that measures whether $\alpha_i = 0$ is 1.52, lower than the GRS statistic of 1.85 for the Fama and French (2015) Model, and implies a p-value of 6.1%, so there is limited evidence against the model and $\alpha_i = 0$, hence, there is little evidence of mispricing; for comparison, the p-value for the Fama and French (2015) Model is 1.2%, that is, we can reject the null hypothesis that $\alpha_i = 0$ and there is evidence of mispricing. In short, the 4-factor model describes significantly better the joint set of 25 book-to-market and 49 industry portfolios than the leading factor models. The result is even sharper when we include the anomaly portfolios. See Table 10.

The model has an statistical fit significantly better than the factor models of Fama and French (2015); Stambaugh and Yuan (2017) and Hou, Xue, and Zhang (2015) in the set of 49 industry portfolios. Crucially, it explains the cross-sectional variation of returns: the GRS statistic that measures whether $\alpha_i = 0$ is .88, significantly lower than the GRS statistic of 1.55 for the Fama and French (2015) Model, and implies a p-value of 68%, that is, we cannot reject the null hypothesis that $\alpha_i = 0$, so there is little evidence of mispricing; for comparison, the p-value for the Fama and French (2015) Model is 4.4%, we can reject the null hypothesis that $\alpha_i = 0$ and there is stronger evidence of mispricing. In short, the GRS test says that the 4-factor model describes extremely well the set of expected returns of the 49 industry portfolios, especially compared to the factor models of Fama and French (2015), Stambaugh and Yuan (2017) and Hou, Xue, and Zhang (2015). See Table 11.

Surprisingly, the model has an statistical fit slightly better than the factor models of Fama and French (2015) and Hou, Xue, and Zhang (2015) in the test of the 25 book-to-market portfolios despite their inclusion of a book-to-market factor. The GRS statistic that measures whether $\alpha_i = 0$ is 1.83, slightly lower than the GRS statistic of 1.91 for the Fama and French (2015) Model. The factor model of Stambaugh and Yuan (2017) actually performs better, consistent with their evidence that book-to-market is not a proxy for risk, but rather for mispricing. Unfortunately, and as expected from the previous literature, there is evidence of mispricing since the p-values are low for all of the models, recall that lower p-values imply there is more evidence against the models.

See Table 11.

**Table 11:** GRS Test for the 4-Factor Text-Based Model and the Fama-French 5 Factor Model

| | 49 Industry Portfolios | | | 25 Book-to-Market Portfolios | | | 15 Anomaly Portfolios | | |
|---|---|---|---|---|---|---|---|---|---|
| | GRS | p-value | $R^2$ | GRS | p-value | $R^2$ | GRS | p-value | $R^2$ |
| Text-based 4 Factor Model | 0.88 | 0.679 | 0.63 | 1.83 | 0.019 | 0.8 | 1.34 | 0.21 | 0.21 |
| Fama-French 5 Factor Model | 1.55 | 0.045 | 0.68 | 1.91 | 0.013 | 0.94 | 1.12 | 0.35 | 0.43 |
| Mispricing Factors | 1.22 | 0.223 | 0.68 | 1.70 | 0.037 | 0.92 | 0.68 | 0.75 | 0.52 |
| q-factor Model | 1.47 | 0.073 | 0.67 | 1.88 | 0.017 | 0.92 | 1.13 | 0.35 | 0.43 |

The table shows the result of the GRS test: high values of the GRS statistic are indicative of high mispricing errors and generate a lower p-value, which is evidence against the fit of the model since the null is of no-mispricing. Lewellen, Nagel, and Shanken (2010) advice against the use of the average $R^2$ to make comparisons between factor models. First row corresponds to the text-based factors presented in the paper, second row corresponds to the Fama and French (2015) Factor Model, third row corresponds to the Anomaly Factors of Stambaugh and Yuan (2017) and fourth row corresponds to the q-factor model of Hou, Xue, and Zhang (2015). First and second columns correspond to the set of 49 industry portfolios and 25 book-to-market portfolios available on Kennet French's website, third column corresponds to the set of 11 long-short anomaly portfolios available of Stambaugh and Yuan (2017). See the text for details.

As an additional test I consider the anomaly portfolios from Stambaugh and Yuan (2017).[27] In this case no model is rejected by this dataset, with the p-values being 21%, 43%, 75% and 35%; and there is no strong evidence of mispricing. Naturally, the model of Stambaugh and Yuan (2017) performs best in these portfolios. A possible interpretation of the result is that most of these anomalies cannot be mapped to firms' risks and instead can be indicative of behavioral biases, market inefficiencies or be related to the SDF in dimensions other than risks that firms face. See Table 11.

# 7 Conclusion

I introduce risk factors that unambiguously represent economic risk for the firms, are interpretable, are taken directly from the companies, and characterize the cross-section of returns better than leading factor models. I apply Unsupervised Machine Learning and Natural Language Processing techniques to extract the risks that firms face from their annual reports, and provide evidence that the firms' risk disclosures are useful and informative. However, unlike supervised machine learning or statistical factor models, I do not use any information of the past returns to select the

27. Available on their website

factors. Hence, I address several concerns of the literature: data mining, overfitting, p-hacking, and concerns about the economic relevance and interpretability of the factors.

The Risk Factors roughly correspond to Technology and Innovation Risk, Demand Risk, Production Risk and International Risk. The 4-factor model constructed with the firms' risk disclosures consistently outperforms the Five Factor Model of Fama and French (2015): The 4-factor model is not rejected in the set of 49 industry Portfolios using the GRS test, which is not the case for the Five Factor Model of Fama and French (2015), and even outperforms Fama and French's Model in the set of 25 book-to-market portfolios. The model also outperforms the factor models of Stambaugh and Yuan 2017 and Hou, Xue, and Zhang 2015 when we consider the set of 49 industry portfolios and the set of 25 book-to-market portfolios.

In short, I provide evidence that firms have a significant understanding of the risk they are facing; the information they provide is important to investors; and the information revealed by the firms can provide guidance on how to improve our theoretical asset pricing models.

# References

Bao, Yang, and Anindya Datta. 2014. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures." *Manage. Sci.* (Institute for Operations Research)(the Management Sciences (INFORMS), Linthicum, Maryland, USA) 60, no. 6 (June): 1371–1391. ISSN: 0025-1909. doi:10.1287/mnsc.2014.1930. https://doi.org/10.1287/mnsc.2014.1930.

Berk, Jonathan B., Richard C. Green, and Vasant Naik. 1999. "Optimal Investment, Growth Options, and Security Returns." *The Journal of Finance* 54 (5): 1553–1607. ISSN: 1540-6261. doi:10.1111/0022-1082.00161. http://dx.doi.org/10.1111/0022-1082.00161.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *J. Mach. Learn. Res.* 3 (March): 993–1022. ISSN: 1532-4435. http://dl.acm.org/citation.cfm?id=944919.944937.

Campbell, John L., Hsinchun Chen, Dan S Dhaliwal, Hsin min Lu, and Logan B. Steele. 2014. "The information content of mandatory risk factor disclosures in corporate filings" [in English (US)]. *Review of Accounting Studies* 19, no. 1 (March): 396–455. ISSN: 1380-6653. doi:10.1007/s11142-013-9258-3.

Cochrane, John H. 1991. "Production-Based Asset Pricing and the Link Between Stock Returns and Economic Fluctuations." *The Journal of Finance* 46 (1): 209–237. ISSN: 1540-6261. doi:10.1111/j.1540-6261.1991.tb03750.x. http://dx.doi.org/10.1111/j.1540-6261.1991.tb03750.x.

Cochrane, John H. 2011. "Presidential Address: Discount Rates." *The Journal of Finance* 66 (4): 1047–1108. doi:10.1111/j.1540-6261.2011.01671.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6261.2011.01671.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01671.x.

Eisfeldt, Andrea L., and Dimitris Papanikolaou. 2013. "Organization Capital and the Cross-Section of Expected Returns." *The Journal of Finance* 68 (4): 1365–1406. ISSN: 1540-6261. doi:10.1111/jofi.12034. http://dx.doi.org/10.1111/jofi.12034.

Fama, Eugene F., and Kenneth R. French. 1992. "The Cross-Section of Expected Stock Returns." *The Journal of Finance* 47 (2): 427–465. ISSN: 1540-6261. doi:10.1111/j.1540-6261.1992.tb04398.x. http://dx.doi.org/10.1111/j.1540-6261.1992.tb04398.x.

———. 1993. "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics* 33 (1): 3–56. ISSN: 0304-405X. doi:https://doi.org/10.1016/0304-405X(93)90023-5. http://www.sciencedirect.com/science/article/pii/0304405X93900235.

———. 2015. "A five-factor asset pricing model." *Journal of Financial Economics* 116 (1): 1–22. ISSN: 0304-405X. doi:http://dx.doi.org/10.1016/j.jfineco.2014.10.010. http://www.sciencedirect.com/science/article/pii/S0304405X14002323.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu. 2017. "Taming the Factor Zoo."

Gaulin, Maclean Peter. 2017. "Risk Fact or Fiction: The Information Content of Risk Factor Disclosures." *Dissertation.*

Gibbons, Michael R., Stephen A. Ross, and Jay Shanken. 1989. "A Test of the Efficiency of a Given Portfolio." *Econometrica* 57 (5): 1121–1152. ISSN: 00129682, 14680262. http://www.jstor.org/stable/1913625.

Gomes, João, Leonid Kogan, and Lu Zhang. 2003. "Equilibrium Cross Section of Returns." *Journal of Political Economy* 111 (4): 693–732. doi:10.1086/375379. eprint: https://doi.org/10.1086/375379. https://doi.org/10.1086/375379.

Hansen, Stephen, Michael McMahon, and Andrea Prat. 2018. "Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*." *The Quarterly Journal of Economics* 133 (2): 801–870. doi:10.1093/qje/qjx045. eprint: /oup/backfile/content_public/journal/qje/133/2/10.1093_qje_qjx045/1/qjx045.pdf. http://dx.doi.org/10.1093/qje/qjx045.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. 2016. "... and the Cross-Section of Expected Returns." *The Review of Financial Studies* 29 (1): 5–68. doi:10.1093/rfs/hhv059. eprint: /oup/backfile/content_public/journal/rfs/29/1/10.1093_rfs_hhv059/2/hhv059.pdf. http://dx.doi.org/10.1093/rfs/hhv059.

Hoffman, Matthew, Francis R. Bach, and David M. Blei. 2010. "Online Learning for Latent Dirichlet Allocation." In *Advances in Neural Information Processing Systems 23,* edited by J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, 856–864. Curran Associates, Inc. http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf.

Hou, Kewei, Chen Xue, and Lu Zhang. 2015. "Digesting Anomalies: An Investment Approach." *The Review of Financial Studies* 28 (3): 650–705. doi:10.1093/rfs/hhu068. eprint: /oup/backfile/content_public/journal/rfs/28/3/10.1093/rfs/hhu068/3/hhu068.pdf. +%20http://dx.doi.org/10.1093/rfs/hhu068.

———. 2017. *Replicating Anomalies.* Working Paper, Working Paper Series 23394. National Bureau of Economic Research, May. doi:10.3386/w23394. http://www.nber.org/papers/w23394.

Israelsen, Ryan D. 2014. "Tell It Like It Is: Disclosed Risks and Factor Portfolios." *Working paper.*

Kelly, Bryan, Seth Pruitt, and Yinan Su. 2018. "Characteristics Are Covariances: A Unified Model of Risk and Return," Working Paper Series, no. 24540 (April). doi:10.3386/w24540. http://www.nber.org/papers/w24540.

Kogan, Leonid, and Dimitris Papanikolaou. 2014. "Growth Opportunities, Technology Shocks, and Asset Prices." *The Journal of Finance* 69 (2): 675–718. ISSN: 1540-6261. doi:10.1111/jofi.12136. http://dx.doi.org/10.1111/jofi.12136.

Kozak, SERHIY, STEFAN Nagel, and SHRIHARI Santosh. 2018. "Interpreting Factor Models." *The Journal of Finance* 73 (3): 1183–1223. doi:10.1111/jofi.12612. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12612. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12612.

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken. 2010. "A skeptical appraisal of asset pricing tests." *Journal of Financial Economics* 96 (2): 175–194. ISSN: 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2009.09.001. http://www.sciencedirect.com/science/article/pii/S0304405X09001950.

Livdan, Dmitry, Horacio Sapriza, and Lu Zhang. 2009. "Financially Constrained Stock Returns." *The Journal of Finance* 64 (4): 1827–1862. ISSN: 1540-6261. doi:10.1111/j.1540-6261.2009.01481.x. http://dx.doi.org/10.1111/j.1540-6261.2009.01481.x.

Loper, Edward, and Steven Bird. 2002. "NLTK: The Natural Language Toolkit." In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1,* 63–70. ETMTNLP '02. Philadelphia, Pennsylvania: Association for Computational Linguistics. doi:10.3115/1118108.1118117. https://doi.org/10.3115/1118108.1118117.

Loughran, TIM, and BILL McDonald. 2016. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research* 54 (4): 1187–1230.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* New York, NY, USA: Cambridge University Press. ISBN: 0521865719, 9780521865715.

McLean, R. David, and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32. doi:10.1111/jofi.12365. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.12365. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12365.

Merton, Robert C. 1973. "An Intertemporal Capital Asset Pricing Model." *Econometrica* 41 (5): 867–887. ISSN: 00129682, 14680262. http://www.jstor.org/stable/1913811.

Nagel, Stefan. 2005. "Short sales, institutional investors and the cross-section of stock returns." *Journal of Financial Economics* 78 (2): 277–309. ISSN: 0304-405X. doi:https://doi.org/10.1016/j.jfineco.2004.08.008. http://www.sciencedirect.com/science/article/pii/S0304405X05000735.

Řehůřek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora" [in English]. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks,* 45–50. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May.

Stambaugh, Robert F., and Yu Yuan. 2017. "Mispricing Factors." *The Review of Financial Studies* 30 (4): 1270–1315. doi:10.1093/rfs/hhw107. eprint: /oup/backfile/content_public/journal/rfs/30/4/10.1093_rfs_hhw107/2/hhw107.pdf. +%20http://dx.doi.org/10.1093/rfs/hhw107.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288. ISSN: 00359246. http://www.jstor.org/stable/2346178.

Zhang, Lu. 2005. "The Value Premium." *The Journal of Finance* 60 (1): 67–103. ISSN: 1540-6261. doi:10.1111/j.1540-6261.2005.00725.x. http://dx.doi.org/10.1111/j.1540-6261.2005.00725.x.

Hanley , Kathleen Weiss and Hoberg, Gerard, Dynamic Interpretation of Emerging Risks in the Financial Sector (February 28, 2018). Available at SSRN: https://ssrn.com/abstract=2792943 or http://dx.doi.org/10.2139/ssrn.2792943

Hassan, Tarek A. and Hollander, Stephan and van Lent, Laurence and Tahoun, Ahmed, Firm-Level Political Risk: Measurement and Effects (December 2017). Available at SSRN: https://ssrn.com/abstract=28 or http://dx.doi.org/10.2139/ssrn.2838644