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

Latest version here

Alejandro Lopez-Lira*

Warrington College of Business, University of Florida

First Version: August 2018

This Version: December 2021

Abstract

I exploit unsupervised machine learning and natural language processing techniques to elicit the risk factors that firms themselves identify in their annual reports. I quantify the firms' exposure to each identified risk and construct factor-mimicking portfolios that proxy for each undiversifiable source of risk. The portfolios are priced in the cross section and contain information above and beyond commonly used multifactor representations. A model that uses only firm-identified risk factors performs at least as well as traditional factor models when pricing a broad set of assets, despite not using any information from past prices or returns.

[*] I am very grateful to Jules van Binsbergen, Joao Gomes, Nick Roussanov, Itamar Drechsler, Winston Dou, Amir Yaron, Lars Hansen, Seth Pruitt, Thomas Sargent, Alexander Belyakov, Marco Grotteria, Andrew Wu (discussant), Ryan Israelsen (discussant), Simona Abis (discussant), Michael Halling (discussant), Rujian Chen (discussant), Chen Xue (discussant), Jose Penalva (discussant), Diego Garcia (discussant), Jose Luis Montiel Olea (discussant), Kyre Lahtinen (discussant), Yinan Su (discussant), Dacheng Xiu (discussant) as well as participants from numerous conferences for helpful discussions and feedback. I am grateful and acknowledge the financial support provided by the Mack Institute for Innovation Management, the Rodney L. White Center for Financial Research, and The Jacobs Levy Equity Management Center for Quantitative Financial Research. Email: alejandro.lopez-lira@warrington.ufl.edu

Since the introduction of the Capital Asset Pricing Model (CAPM) more than fifty years ago, the asset pricing literature has witnessed tremendous growth in potential additional factors that could explain the cross-section of expected stock returns. The formidable academic challenge has been finding interpretable and economically relevant risk factors. I propose a novel approach to this challenge by eliciting the risk factors that firms themselves identify in their annual reports. I document they contain information above and beyond common factors and characteristics in the literature, and their theoretical and empirical properties make them good candidates to function as explanatory factors.[1]

To accomplish this, I quantify the risks that firms self-identify by applying textual analysis techniques to their financial reports. Firms are required to disclose extensively every risk they face in their annual reports (10-Ks) in the section 'Item 1A Risk Factors'. While the section contains highly detailed information, it is impractical to read every page of every company's report every single year. Hence, instead, I use machine learning to represent each document as the proportion of the risk disclosure that each company allocates to each risk (risk weights). Remarkably, the algorithm reveals the firm-identified risks without any manual input or information from past prices or returns.[2]

To better understand the economics of the risk disclosures, I provide a tractable theoretical model. In the model, investors can sue the firm for not disclosing a particular risk, but because it is costly to file a lawsuit, investors are more likely to sue a firm for not revealing the most impactful risks. Furthermore, the legal system is more likely to dismiss a lawsuit for a given risk if the firm writes more about that risk in its financial reports. Consequently, the model shows rational managers optimally mention a specific risk more frequently when their firm is more exposed to that risk. Hence, the model provides an ex-ante economic connection between risk disclosures and risk factor exposures. Therefore, a priori, the risk

[1] As an example of additional proposed factors see 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

[2] The machine learning algorithm is called Latent Dirichlet Allocation (LDA), Blei, Ng, and Jordan (2003).

weights should have explanatory power in the cross-section of returns.[3]

To study the model's implications, I first document that for firms with similar risk disclosures, their returns are significantly more correlated with each other. Then, I use the risk weights to understand the impact of the disclosures in the first moment of returns by running Fama-Macbeth cross-sectional regressions. I find the risks weights, treated as characteristics, are priced even after controlling for well-known predictive features.

Then, I construct factor-mimicking portfolios for each specific risk. I show the portfolios are priced and drive a significant part of the time-series variation of stock returns. Furthermore, they contain information above and beyond commonly used multifactor representations. For example, a Gibbons, Ross, and Shanken (1989) (GRS) test shows statistically significant evidence that the Fama-French Five-Factor model does not span the set of firms' disclosed risks.[4]

Finally, I treat the firm-identified risks as factors and test their capacity 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.[5] I first study a factor model using the most discussed risks but consider several variations.

The factor model is constructed using exclusively the information disclosed by the firms guided by the theoretical model and, hence, avoids using characteristics previously documented to have predictive power. Furthermore, Cochrane (2005) warns that: "it is probably not a good idea to evaluate economically interesting models with statistical horse races against models that use portfolio returns as factors. Economically interesting models, even if true and perfectly measured, will just equal the performance of their own factor-mimicking portfolios, even in large samples. They will always lose in sample against ad-hoc factor

[3] I provide a more general model in Lopez-Lira (2021) and show the current regulations satisfy the assumptions of the model. Furthermore, there is ample existing empirical evidence in the accounting literature that shows risk disclosures are truthful and informative, e.g., Campbell et al. (2014), Gaulin (2017)

[4] A similar result applies to the models in Hou, Xue, and Zhang (2015) and Stambaugh and Yuan (2017). Naturally, there is variation in the results, and models price some risks better than others.

[5] I choose these portfolios as the test set following the critique of Lewellen, Nagel, and Shanken (2010) I include additional tests using the profitability-investment portfolios and the anomalies portfolios from Stambaugh and Yuan (2017)

models that find nearly ex-post efficient portfolios."

Hence, it is especially surprising that the model appears to have a statistical fit at least as good as the leading models in the literature: the factor models of Fama and French (2015), Stambaugh and Yuan (2017), and Hou, Xue, and Zhang (2015) when considering the broad set of testing assets. Nevertheless, it is essential to reiterate that it is outside of the scope of the paper to run a horse race with existing reduced-form factor models.

Concerning the statistics, using the GRS test,(Gibbons, Ross, and Shanken (1989)) 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% which means 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.[6]

Moreover, because the factors are proxying exclusively for firms' self-identified risks, they provide a natural benchmark for disentangling risk and mispricing. Typically, when a factor model cannot explain a given set of portfolios, it is considered an incomplete factor model, and further modification is recommended. In contrast, whenever the firm-identified risk factors do not explain a portfolio, there is a straightforward interpretation: that portfolio is not spanned by any of the firms' self-disclosed risks, so another economic story is necessary to explain such returns, such as intermediary risks, behavioral factors or limits to arbitrage.

In summary, I use machine learning and the information revealed by the firms in the economy to answer some of the essential questions in asset pricing: What are the significant risks in the economy according to firms themselves? Which ones are systematic? Are they priced? Are they summarized well by existing models? Furthermore, I introduce firm-identified risk factors that perform at least as well as traditional models in a broad set of assets and help disentangle risk from mispricing.

[6] 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 between models.

Related Literature

My paper contributes to two branches of literature: (1) machine learning and text analysis in finance and (2) cross-sectional asset pricing.

I contribute to the recent strand of the literature that employs text analysis to study a variety of finance research questions (e.g., Jegadeesh and Wu (2013), Campbell et al. (2014), Hoberg and Phillips (2016), Gaulin (2017), Baker, Bloom, and Davis (2016), Ke, Kelly, and Xiu (2019), Bybee et al. (2019), Ke, Montiel Olea, and Nesbit (2019)). See Loughran and McDonald (2016) for an excellent review. Some papers employ text analysis to study a specific risk that the researchers have in mind (e.g., Hassan et al. (2019) for political risk; Loughran, McDonald, and Pragidis (2019) for oil risk). I, instead, do not specify any risk ex-ante and instead let them arise naturally from the data using machine learning methods.

The early literature of topic modeling in finance studies the interaction between self-disclosed risks, volatility, and betas, abstaining from studying the pricing of the disclosed risks due to the short time horizon. Israelsen (2014) is one of the first papers in finance that uses topic modeling on the risk disclosures and focuses on the interaction between several disclosed risks, stock-return volatility, and betas of the Fama-French Four-Factor model in the period 2006-2011, using weekly returns. Bao and Datta (2014) explore the interaction between disclosed risk and volatility to showcase their novel topic modeling technique. Israelsen (2014) and Bao and Datta (2014) use topic modeling for the entire period, so the risk weights they use contain look-ahead bias, which I avoid by using an online version of the topic modeling algorithm. Hanley and Hoberg (2018) propose a different way to deal with look-ahead bias and apply the technique to understand emerging risks in the financial sector, although they abstract from asset pricing implications.

Using risk weights with no look-ahead bias and a dataset with a significantly longer time horizon, I answer a completely different set of asset pricing questions compared to the previous literature that uses topic modeling: Which of the underlying risks in the economy are systematic? Are they priced? Are they summarized well by existing models? Can we get interpretable factors that represent economic risk? How much can we explain using the common risks perceived by the firms?

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 study which of the firms self disclosed risks are priced, and construct a factor model using the implied factor-mimicking portfolios.

The factor model I form using the firms' disclosed risks complements the literature in the following ways. First, regarding statistical factor models: while they provide a superior statistical fit, they are not designed to be interpretable, so naturally, it is hard to understand the economics of these factors and whether they represent risk; are generated by behavioral patterns; or represent market inefficiencies, whereas, by design, the factors constructed from the firms' risk disclosures represent economic risk.[7]

Second, regarding empirical factor models: while they succeed in explaining empirically puzzling portfolios (portfolios with $\alpha \neq 0$), they usually do so by iteratively adding (some of) the existing anomalies as risk factors. However, adding previously discovered anomalies as risk factors naturally generates too many factors, what Cochrane (2011) describes as a "factor zoo", and disentangling the genuine risk factors from the anomalies is a complicated endeavor.[8] To further complicate things, there are serious 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. However, since, by construction, all of the factors in the paper, represent risk, it suffices to identify which of these factors are priced and what assets we can price.

Finally, regarding economic theory models: we know from Merton (1973) that the risk

[7] See Kelly, Pruitt, and Su (2019), Kozak, Nagel, and Santosh (2018)

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

premia of every asset depend 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 can affect returns. Asking researchers to identify most of these variables seems like an unworkable task. Firms, however, have a much better understanding of the risks they are facing. Hence, understanding which risks firms face can guide how to improve our theoretical models.

1 Data

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

1.1 CRSP, Compustat and Factor Models

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 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: 2006-2019, 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]) for comparability with the existing literature. 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.

1.2 10-K Annual Reports

Firms disclose in their annual reports the types of risk they are facing.

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 2020 from the EDGAR Database on the SEC's website. The 10-Ks come in many different file formats (.txt, .xml, and .html) and have various styles, so it is 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. Next, I preprocess the documents by removing stop words, lemmatizing, and constructing word bigrams. Finally, I apply online LDA to these documents. I discuss the details of the processing steps in the Appendix.

I use machine learning (LDA) to get two objects: the risks that firms are discussing (risk topics) and (2) how much space each company allocates to discuss each risk (risk weights). The risks topics and the risk weights are available in real-time, and hence any strategy that bets on those specific risks is tradable. I use 25 topics following the literature of topic modeling in financial applications (e.g., Israelsen (2014), Bao and Datta (2014), Hanley and Hoberg (2019)).

The risks topics are called topics in the natural language processing literature, and formally they are probability distributions over words. Intuitively, the documents are projected in the risk topic space. In turn, each document is represented by a distribution over topics, the risk weights. I elaborate on the details of topic modeling in the Appendix.

To illustrate the kind of disclosures that firms make, consider an excerpt from Apple Inc.'s 2010 10-K annual report below. I incorporate suggested labels regarding the type of risk and highlight possible keywords in bold. Note that both labels and keywords are just for illustrative purposes. There is no need to manually label the risks in the paper or define the keywords since the risks 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**

[Insert Figure 1 about here]

Figure 1 shows an example of the disclosures, an excerpt from a specific section of Apple's risk factors, with the total length of that section being ten pages in that annual report. Figure 2 shows the result of applying machine learning: a representation of the document as the proportion of the risk disclosure that the company allocates to each risk (risk weights). Figure 3 shows the International Risk (topic): each risk is described by (a distribution over) words.

[Insert Figure 2 about here]

[Insert Figure 3 about here]

2 Economics of Risk Disclosures

To better understand the economic of the risk disclosures, I provide a tractable theoretical model where I show that managers optimally write longer disclosures for the most critical risk factors because the SEC Regulation and the existing legal doctrine incentivize them to do so.[9]

The model consists of a collection of firms that differ in their risk exposure towards different risk sources. I consider two periods for tractability, but the model is equivalent to a model with infinite symmetric periods because there is no dynamic optimization.

Each firm is subject to multiple risks, and the exposure varies at the firm level. The risks may materialize next period and affect the profits of the company. Formally, firm j's profits at $t+1$ are:

$$\pi_{j,t+1} = x_{j,t+1} - \sum_{i=0}^{I} b_{j,i} R_{i,t+1} + \epsilon_{j,t+1}, \tag{1}$$

where $\pi_{j,t+1}$ denotes firms profits, $x_{j,t+1}$ is a firm fixed effect, $R_{i,t+1}$ is the realization of the i-th risk in the economy with $E[R_i] > 0$, $b_{j,i} >= 0$ is the exposure of firm j to the i-th risk and $\epsilon_{j,t+1}$ is, without loss of generality, a zero mean shock.[10] In what follows, I drop the time subscripts. Let $r_{ij} = b_{i,j}E[R_i]$ the total expected exposure of the firm to each risk. There is a finite number of risks in the economy, I.

In the model, investors have the option to sue the firm for not disclosing a particular risk. However, it is costly to sue the firm. Furthermore, the legal system is more likely to dismiss a lawsuit for a specific risk if the firm writes more about that risk in their financial reports. In the event of this lawsuit passing and the investors winning, the settlement amount decreases the more the firm elaborates on the disclosure and increases if the risk is more relevant.

The managers write the risk disclosures to minimize the sum of the expected cost of lawsuits and the reports' writing. For simplicity, there is no intertemporal discount.[11]

[9] Firms are legally required to discuss "the most significant factors that make the company speculative or risky" (Regulation S–K, Item 105(c), SEC 2005) in a specific section of the 10-K annual reports (Section 1A). They could face legal action if they fail to obey the regulation and be vulnerable to lawsuits from investors. Finally, all of the annual reports are audited. It is stated in the General Accepted Accounting Principles that any material information about the risks that the company faces must be revealed.

[10] The risks $R_{i,t+1}$ in this economy affect negatively the firm on average, contrasting with risks such as TFP shocks in macro models, which are absorbed by $\epsilon_{j,t+1}$ and $x_{j,t+1}$.

[11] There is no maximum limit of pages allocated to risks in the regulation, although since 2020, more than 15 pages require a summary. In practice, we observe a finite number of pages and lawyers help write

Formally, the manager optimizes the following function:

$$min_{\{L_i\}_{i=0}^{I}} \sum_{i=0}^{I} p(L_i)C(L_i, r_{ij}) + h(L_i). \tag{2}$$

$p(L_i)$ is the probability of receiving a lawsuit regarding i-th risk when the section disclosing such risk is of length L_i. $C(L_i, r_{ij})$ is the (expected) cost of settling the lawsuit conditional on the lawsuit proceeding. Finally, $h(L_i)$ is the cost of writing the disclosure. Define $G(L_i, b_j R_i) = p(L_i)C(L_i, b_j R_i)$.

For tractability, I assume the following functional forms, but I show in Lopez-Lira (2021) that the results generalize under minimal assumptions.

The probability of receiving a lawsuit is a decreasing function of L_i, the length of the i-th risk:

$$p(L_i) = (1 + L_i)^{-1}. \tag{3}$$

The cost of a lawsuit if received is a decreasing function of L_i, the length of the i-th risk, an increasing function of $r_{ij} = b_{i,j}E[R_i]$ the total exposure of the firm to each risk and a positive parameter, a.

$$C(L_i, r_{ij}) = (1 + \frac{1}{L_i})(ar_{ij})^3, \ a > 0, \tag{4}$$

and the cost of writing the disclosure is an increasing function of L_i, the length of the i-th risk and a positive parameter d, given by:

$$h(L_i) = \frac{d^3}{2}L_i^2, \ d > 0. \tag{5}$$

Which imply the optimization problem reduces to

the section, which is costly.

$$min_{\{L_i\}_{i=0}^{I}} \sum_{i=0}^{I} p(L_i)C(L_i, r_{ij}) + h(L_i) = min_{\{L_i\}_{i=0}^{I}} \sum_{i=0}^{I} \frac{(ar_{ij})^3}{L_i} + \frac{d^3}{2} L_i^2. \tag{6}$$

With first order conditions:

$$\frac{(ar_{ij})^3}{L_i^{*2}} = d^3 L_i^*. \tag{7}$$

Intuitively, the firm is balancing the marginal benefit of increasing the risk disclosure on the left-hand side of the equality against the marginal cost of increasing the length of the disclosures on the right-hand side of the equality. The marginal benefit includes the decrease in the probability of receiving a lawsuit and the lawsuit's potential cost if received. The marginal cost in practice involves the management team spending time with lawyers and accountants. The second-order conditions guarantee that the first-order conditions define an optimum.

When we simplify:

$$L_i^* = \frac{a}{d} r_{ij}. \tag{8}$$

Each disclosure is linearly increasing in each risk exposure. Furthermore, firms with a higher cost of writing the disclosures, d, report shorter disclosures while firms with more costly lawsuits, characterized by parameter a, allocate more space to this risk. For this particular example, the proportion of space allocated to risk i, l_i is exactly proportional to the impact of each risk:

$$l_i = \frac{r_{ij}}{\sum_k r_{ik}}, \tag{9}$$

and generally the proportion of space allocated to each risk is increasing in the exposure to such risk.

Moreover, the model's prediction are consistent with the existing literature. For example, 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". Furthermore, Gaulin (2017) finds that "managers time their identification of new risk factors and removal of previously identified ones to align with the expected occurrence of future adverse outcomes...[and] firms respond to investor demand in a manner consistent with the litigation shield hypothesis... inconsistent with concerns of uninformative boilerplate or 'copy and paste' disclosure".

3 Risk Topics

Recall that with LDA, we get in real-time all of the common risks that firms discuss and how much space each company allocates discussing each risk. To avoid confusion, I refer to the topics obtained using LDA as risk topics and to the amount of space they allocate to each risk as risk weights. Figure 2 shows an example of the latter. Recall that risk topics are distribution over words, and risk weights are distribution over topics. Hence, LDA is similar to a matrix factorization technique and does not give labels for the risk topics, nevertheless, we can interpret the topics by reading the most frequent words and by looking at which companies discuss the most each risk. Figure 4 shows a general picture of the risks that firms are concerned about, where I show the 25 risk topics extracted from the 10-K annual reports. Figure 5 shows the average percentage of risk disclosures that firms allocate to each risk.

[Insert Figure 4 about here]

[Insert Figure 5 about here]

Figure 6 shows that there is significant interaction between the risk weights and market beta, the book-to-market ratio and the market capitalization. There is no significant correlation between profitability, investment or past returns and any of the risks that firms disclose.

[Insert Figure 6 about here]

Further, the factors in standard pricing models are portfolios of firms, and hence it is natural to wonder what is the implied risk disclosure for the portfolios. That is, if we think of a factor as a hypothetical firm, what would be its risk disclosure. Notice that to understand portfolios this way we need to have the composition of the portfolios. Once we construct factor-mimicking portfolios of each risk, we will be able to study the exposure for an arbitrary portfolio whose returns we have, using standard projection techniques.

Figure 7 shows the implied (time-series average) risk disclosure for the HML factor, constructed by using a weighted average of firms' risk disclosures where the weights are the portfolio weights. The implied HML factor risk disclosure is heavily short international risk, marketing risk, and heavily long oil and property risks. The decline in oil prices explains a significant part of its poor performance during the recent period. Notice the figure does not show the betas with respect to the factor-mimicking portfolios, which look fairly similar, as we will see in Section 5.

[Insert Figure 7 about here]

Figure 8 shows the exercise repeated for the Momentum Factor (portfolios sorted on past performance). There is a clear pattern of no stable relationship with any of the firms' risks: Momentum is moving back and forth between all of the types of risks. We may initially think the effect is mechanical, since Momentum is re-balanced monthly, whereas the risk exposures are stable. However, Momentum could be concentrated in a couple of specific risks, say, in international risk and technology risk. We see it is not the case and, Momentum does not seem to be related to any of the companies' disclosed risks.

[Insert Figure 8 about here]

It is helpful to consider the overlap between the risk weights and industry classifications. The risk weights are a continuous measure and the industries are discrete, so there is no natural comparison. For the easiness of exposition, I consider firms whose risk weights are above 25% to be exposed to a specific risk. Table 1 shows the overlap between firms exposed to the innovation risks and the Hoberg and Phillips (2016) industries. While one of the industries contains 50% overlap, the remaining are dispersed across different clusters. Table 2 shows a similar result for the SIC industries. Roughly half of the firms are in the manufacturing section and half in the services sector. Hence, there is no complete overlap between industries and risk, potentially explaining the failure of using portfolios of industries as factors.

[Insert Table 1 about here]

The explanation for the lack of overlap is straightforward: firms in the same industry can be affected by different risks, for example, consider two firms within the technology sector, one with a significant amount of debt and one without, and hence one is exposed to credit risk while the other one is not. Conversely, two firms can be in entirely different industries but exposed to the same type of risk. For example, consider two multinational corporations exposed to currency risk, but one in the telecommunications industry and another in the manufacturing sector.

[Insert Table 2 about here]

3.1 Correlations

Based on the theoretical model, we should expect companies with similar risks to covary more with each other. Table 3 shows the average, standard deviation and selected percentiles of companies pairwise correlation, risk similarity, book-to-market and size distance. The average realized pairwise correlation for stocks is around 20%. The risk similarity is defined as one minus the pairwise Hellinger distance between the firms' time-series average of the risk weights and its average value is around 0.14 during the sample .[12] The beta exposure is the pairwise product of firms' time-series average betas measured by using yearly time-series regressions of the firms' returns against the market using daily data and its average value is around 1.25. The distance between the log of size and book to market are measured as the pairwise difference between the absolute value of the time-series average of the respective measures and their respective values are 1.05 and 2.23.

[Insert Table 3 about here]

To assess the importance of risk similarities on correlations I run a panel regression of individual stocks' pairwise correlations on the risk similarity, as well as on the beta exposure and controls in a similar spirit as Jotikasthira et al. (2012). Notably, the coefficient on the risk similarity is around 23% regardless of the use of controls. The most extreme interpretation is that for companies with very similar risk weights, their correlation effectively doubles, while a more measured interpretation is that a one standard increase in the risk similarity results in a 3.2% increase in the correlation.

[Insert Table 4 about here]

3.2 Price of Risk

The natural way to measure the price of these risks is using Fama-MacBeth regressions. The controls include betas, book-to-market ratios, size, profitability, and investment. Notice that since the risk weights sum up to one for each company, we cannot include an intercept in

[12] Hellinger distance is defined as $H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$ where $P = (p_1, \ldots, p_k)$ and $Q = (q_1, \ldots, q_k)$ are two discrete probability measures.

the regression since that would induce collinearity.[13]

I include in the regressions only stocks above the 20th percentile of size of the NYSE exchange and exclude microcaps to alleviate any concern about liquidity, see Fama and French (2008).[14] Figure 9 shows the results of the regression without controls and Figure 10 show the results with controls.

Caution is required in interpreting the magnitude of the coefficients. Each coefficient shows the average excess return of a portfolio whose average risk disclosure is concentrated in each risk. However, the average standard deviation for a given risk weight is about 10%, so all of the coefficients should be divided by ten to get a rough sense of the marginal effect of the risk weights in the expected returns.

Innovation related risks carry the highest significant unconditional premium for this period. Since these firms have a low book-to-market the time series regressions when using the Fama-French Five-Factor model as controls will show a significant intercept. Nevertheless, the Sharpe ratios are not be excessively high, since the returns comes with an increase in variance.

Consumer demand and international risk carry a premium when there are additional controls in the regressions. Production and Credit risk are marginally significant. The exposure to China and Oil Risk is not compensated in this period.

Overall, Innovation, Credit and International Risk, carry both a high risk premium and a high covariance, across specifications, despite the addition of controls, suggesting that they provide additional information both for first and second moments of returns.

[Insert Figure 9 about here]

[Insert Figure 10 about here]

[13] Alternatively, we can drop one of the risk weights, but then the interpretation changes significantly.

[14] Results are similar, although the coefficients are naturally slightly smaller when using only the big stocks, stocks above the 50th percentile of size of the NYSE exchange. Results are available in the online Appendix.

4 Portfolios

I use factor mimicking portfolios designed to have unit exposure in one risk and zero in the other risks. The relative risk exposure we get from the topic modeling algorithm is similar to an indicator variable, and hence there is no natural short side. An analogy are industries: there is no 'short' side of the coal industry. Similarly, the opposite of disclosing exposure to demand risk is not disclosing demand risk exposure. Because of the sparsity of the machine learning algorithm, the 'short' side would consist of a well-diversified portfolio extremely similar to the market portfolio.[15]

I use the cross-sectional technique constructed in Fama (1976) and recently described in Back, Kapadia, and Ostdiek (2015) to get the portfolio weights. As in Back, Kapadia, and Ostdiek (2015) and Fama and French (2008) I include only stocks above the 20th size percentile of the NYSE exchange and exclude microcaps to alleviate any concern about liquidity. Kirby (2019) recommends against using a value-weighted procedure, equivalent to a weighted Fama-MacBeth regression using as weights the inverse of the size of the firm, but the results are available in the online Appendix.

Formally, the weights for the portfolio of risk k solve the following problem at every point in time:

$$\min_{w_k} \; w_k' w_k \; s.t. \; w_k' X = e_i,$$

where X is a $n \times K$ matrix, n is the number of stocks, K is the number of risks (25), whose columns are the risk weights, how much time each company spends discussing each risk, each row corresponds to a firm observation, at a given point in time and e_i denotes the (row) i-basis vector of $\mathbb{R}^K$.

The solution is available in analytical form:

[15] As an alternative I consider an indicator variable which is one for the risk that the company discusses the most and value-weight the portfolios. Results are similar and available in the online Appendix.

$$w_k = X'(X'X)^{-1}e_k,$$

or if we collect the portfolio weights for each risk as a column in a Matrix W,

$$W = X'(X'X)^{-1},$$

and notice that $W'X = I_k$ as desired.[16]

Notice that the realized excess returns of the portfolios are the (normalized) slope coefficients on Fama-MacBeth regressions of excess returns on the risk weights. Notice also the portfolios are excess return portfolios. The big advantage of Fama's insight is that we can include additional variables in the X matrix when running the Fama-MacBeth regressions. When we include (ex-ante) market beta for example, we will be forming (ex-ante) beta neutral portfolios. In fact, Back, Kapadia, and Ostdiek (2015) show that getting the portfolio weights using cross-sectional regressions at every time period, and then running time-series regressions is the natural way to correct for the errors-in-variables problem that arises since betas are estimated.

Hence, we get two set of factor mimicking portfolios. I call simply 'firm-identified risk factors' the portfolios that have unit exposure in one risk and zero in the other risks. I call 'orthogonal factors', the portfolios which in addition to having unit exposure in one risk and zero in the other risk, are orthogonal to portfolios formed on ex-ante betas and characteristics.

4.1 Are the risks spanned?

It is natural to wonder about the relationship between the usual models, for example Fama and French (2015) and the firm identified risk factors. We can see in Table 5 that the traditional models do not span the firm identified risk factors. The natural interpretation

[16] The problem can be succinctly written as $min_W \;\sum_i^K e_i'W'We_i$ s.t. $W'X = I_K$.

is that these factors contain additional information, and in fact combining them with the traditional ones leads to greater improvement of the description of cross-section of returns, although at the cost of mixing economic risks with proxies for other factors that affect the stock returns (e.g. Profitability). As a placebo test, the portfolios formed using the Hoberg and Phillips (2016) are spanned.

[Insert Table 5 about here]

[Insert Table 6 about here]

In turn, the firm identified risk factors do not span the traditional ones as Table 6 shows. Table 7 shows that the intercept of the Profitability factor is the only one with a positive significant 'alpha' with respect to the firms identified risk factors and a t-stat of 3.43. The Small Minus Big Factor has a negative 'alpha' with a t-stat of -2.04. The GRS tests and the implied p-values suggest that the factors are not completely in the span of each others.

[Insert Table 7 about here]

Figures 11-14 show the individual t-statistic confirming what we saw in the Fama-MacBeth Regressions: Risks related to innovation contain a significant component unexplained by the Fama-French Five-Factor Model. CAPM in this period performs significantly better than the Five-Factor Model, although there is still a significant unexplained component for the innovation risks.

[Insert Figure 11 about here]

[Insert Figure 12 about here]

[Insert Figure 13 about here]

[Insert Figure 14 about here]

Despite the models not being completely in the span of each others, it is interesting to see the betas between the firm identified risk factors and the standard factors. The firm identified risk factors are excess-return portfolios but not long-short, hence they are naturally correlated with the market portfolio, whereas the orthogonal factors have virtually zero exposure by design as Figure 15 shows.[17]

[Insert Figure 15 about here]

Figure 15 depicts the firm identified risk factors from the perspective of the Five-Factor Model. It shows that Stock-Price Risk covaries positively with the Small-minus-Big Factor, whereas International and China Risk covary negatively with it. Furthermore, the Innovation Risk Cluster covaries negatively with the High-minus-Low Factor and with the Profitability Factor, which explain the higher 'alphas' coming from the Five-Factor Model.

[Insert Figure 15 about here]

Figure 16 depicts the other side of the coin. The Five-Factor Model from the perspective of the firm identified risk factors. It shows that the market-portfolio consist mostly of International risk. Small Minus Big is negatively loaded on International Risk, and positively loaded in Consumer Demand and Production Risk.

High Minus Low is negatively loaded on Innovation and positively loaded on Oil, Production and Property Risk. RMW (Robust Minus Weak) is negatively related to Innovation Risk, and positively related with International and Demand Risk. CMA (Conservative Minus Aggressive) is positively related only to Property Risk. Momentum is negatively related to Credit and Industrial Risk and slightly positively related to the Innovation Risk Cluster.

[17] We can always rotate the FIRFs so that they are orthogonal to the market portfolio. The projections only makes sense for the the non-orthogonal factors, since by construction the orthogonal factors are orthogonal up to measurement error. The graphs are in Appendix 2 for completeness. See also Israelsen (2014).

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

5 Risk Factors

I select the 4 risks that affect the highest number of firms in 2006 and keep them for the whole sample to avoid 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. 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 Online Appendix.

Despite the model not being designed price the cross-section, it is interesting anyways to compare the performance of the factor model in pricing portfolios of general interest (such as the industry portfolios) and portfolios that are hard for macroeconomic based factors, for example the set of 25 book-to-market portfolios and the anomaly portfolios.[18] Adding more testing portfolios addresses the critique of Lewellen, Nagel, and Shanken (2010).

The table should not be read as a horse-race, other models are there just for comparability, since the models have different objectives, this one, to produce interpretable risk factors that represent economic risks for the firms. I use the GRS test from Gibbons, Ross, and Shanken (1989) and include the performance of the factor models of Fama and French (2015); Stambaugh and Yuan (2017); and Hou, Xue, and Zhang (2015) for benchmark comparison.

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

$$GRS \propto \frac{\alpha' \Sigma^{-1} \alpha}{1 + \mu' \Sigma^{-1} \mu}, \tag{10}$$

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 Sharpe ratio that we get from the factors by adding

[18] See Section 1 for details.

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.

[Insert Table 8 about here]

The firm identified risk factors 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 8.

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 9.

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 9.

[**Insert Table 9 about here**]

As an additional test I consider the anomaly portfolios from Stambaugh and Yuan (2017).[19] 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. Table 9 shows that all the models are able to explain the cross-sectional differences in returns in the anomaly portfolios in the period 2006-2019 mainly because the performance of the anomalies has been declining, especially in the recent period as McLean and Pontiff (2016) document. The average R^2 is the lowest for the models using the firm identified risk factors, indicating that most of the anomalies do not covary significantly with any of the risks that firms are concerned about.

[19] Available on their website.

6 Conclusion

I use machine learning to answer some of the essential questions in asset pricing: What are the fundamental risks in the economy? Which ones are systematic? Are they priced? Are they summarized well by existing models?

I identify the risks that firms consider relevant by letting firms themselves tell us what risks they face. I use natural language processing techniques to extract this information from their annual reports. Furthermore, I introduce interpretable firm identified risk factors (FIRFs) that perform at least as well as traditional models, despite not using any information from past prices or returns.

I provide evidence that firms have a significant understanding of the risks they face. Moreover, the information they reveal is relevant to investors. Using machine learning to understand which risks firms care about can improve our theoretical asset pricing models. Ultimately, this paper shows that text analysis can help identify investors' risk perception and their conditional information set.

References

Back, Kerry, Nishad Kapadia, and Barbara Ostdiek. 2015. "Testing Factor Models on Characteristic and Covariance Pure Plays." *Working Paper,* https://ssrn.com/abstract=2621696.

Baker, Scott R, Nicholas Bloom, and Steven J Davis. 2016. "Measuring Economic Policy Uncertainty*." *The Quarterly Journal of Economics* 131 (4): 1593–1636. ISSN: 0033-5533. https://doi.org/10.1093/qje/qjw024. https://doi.org/10.1093/qje/qjw024.

Bao, Yang, and Anindya Datta. 2014. "Simultaneously Discovering and Quantifying Risk Types from Textual Risk Disclosures." *Manage. Sci.* 60 (6): 1371–1391. ISSN: 0025-1909. https://doi.org/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. https://doi.org/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.* 3:993–1022. ISSN: 1532-4435. http://dl.acm.org/citation.cfm?id=944919.944937.

Bybee, Leland, Bryan T. Kelly, Asaf Manela, and Dacheng Xiu. 2019. "The Structure of Economic News." *SSRN Electronic Journal* (January). ISSN: 1556-5068. https://doi.org/10.2139/ssrn.3446225. https://papers.ssrn.com/abstract=3446225.

Campbell, John L., John L. Campbell, Hsinchun Chen, Hsinchun Chen, Dan S. Dhaliwal, Dan S. Dhaliwal, Hsin-min min Lu, Hsin-min min Lu, Logan B. Steele, and Logan B. Steele. 2014. "The information content of mandatory risk factor disclosures in corporate filings." *Review of accounting studies* (Boston) 19, no. 1 (March): 396–455. ISSN: 1380-6653. https://doi.org/10.1007/S11142-013-9258-3/TABLES/11. https://link.springer.com/article/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. https://doi.org/10.1111/j.1540-6261.1991.tb03750.x. http://dx.doi.org/10.1111/j.1540-6261.1991.tb03750.x.

———. 2005. *Asset Pricing: Revised Edition.* Princeton University Press. ISBN: 9781400829132. https://books.google.co.uk/books?id=20pmeMaKNwsC.

Cochrane, John H. 2011. “Presidential Address: Discount Rates.” *The Journal of Finance* 66 (4): 1047–1108. https://doi.org/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. https://doi.org/10.1111/jofi.12034. http://dx.doi.org/10.1111/jofi.12034.

Fama, Eugene F. 1976. *Foundations of Finance.* Basic Books.

Fama, Eugene F, and Kenneth R French. 2008. “Average Returns, B/M, and Share Issues.” *The Journal of Finance* 63 (6): 2971–2995. https://doi.org/10.1111/j.1540-6261.2008.01418.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2008.01418.x.

———. 1992. “The Cross-Section of Expected Stock Returns.” *The Journal of Finance* 47, no. 2 (June): 427–465. ISSN: 15406261. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.1992.tb04398.x.

———. 1993. “Common risk factors in the returns on stocks and bonds.” *Journal of Financial Economics* 33, no. 1 (February): 3–56. ISSN: 0304-405X. https://doi.org/10.1016/0304-405X(93)90023-5.

———. 2015. “A five-factor asset pricing model.” *Journal of Financial Economics* 116, no. 1 (April): 1–22. ISSN: 0304-405X. https://doi.org/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.”

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. https://doi.org/10.1086/375379. https://doi.org/10.1086/375379.

Hanley, Kathleen Weiss, and Gerard Hoberg. 2018. "Interpretation of Emerging Risks in the Financial Sector." *Forthcoming Review of Financial Studies,* https://ssrn.com/abstract=2792943.

———. 2019. "Dynamic Interpretation of Emerging Risks in the Financial Sector." *The Review of Financial Studies,* ISSN: 0893-9454. https://doi.org/10.1093/rfs/hhz023. https://doi.org/10.1093/rfs/hhz023.

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. https://doi.org/10.1093/qje/qjx045. 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. https://doi.org/10.1093/rfs/hhv059. http://dx.doi.org/10.1093/rfs/hhv059.

Hassan, Tarek A, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. 2019. "Firm-Level Political Risk: Measurement and Effects*." *The Quarterly Journal of Economics* 134 (4): 2135–2202. ISSN: 0033-5533. https://doi.org/10.1093/qje/qjz021. https://doi.org/10.1093/qje/qjz021.

Hoberg, Gerard, and Gordon Phillips. 2016. “Text-Based Network Industries and Endogenous Product Differentiation.” *Journal of Political Economy* 124 (5): 1423–1465. https://doi.org/10.1086/688176. https://doi.org/10.1086/688176.

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.” *Review of Financial Studies* 28 (3): 650–705. ISSN: 0893-9454. https://doi.org/10.1093/rfs/hhu068. https://academic.oup.com/rfs/article-lookup/doi/10.1093/rfs/hhu068.

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

Jegadeesh, Narasimhan, and Di Wu. 2013. “Word power: A new approach for content analysis.” *Journal of Financial Economics* 110 (3): 712–729. ISSN: 0304-405X. https://doi.org/https://doi.org/10.1016/j.jfineco.2013.08.018. http://www.sciencedirect.com/science/article/pii/S0304405X13002328.

Jotikasthira, Chotibhak, Christian Lundblad, Tarun Ramadorai, John Campbell, Cam Harvey, Donghui Li, Ludovic Phalippou, Michael Schill, Ajay Shah, and Dimitri Vayanos. 2012. “Asset Fire Sales and Purchases and the International Transmission of Funding Shocks.” *The Journal of Finance* 67, no. 6 (December): 2015–2050. ISSN: 1540-6261. https://doi.org/10.1111/J.1540-6261.2012.01780.X. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1540-6261.2012.01780.x%20https://onlinelibrary.wiley.com/doi/

abs/10.1111/j.1540-6261.2012.01780.x%20https://onlinelibrary.wiley.com/doi/10.1111/j.1540-6261.2012.01780.x.

Ke, Shikun, José Luis Montiel Olea, and James Nesbit. 2019. "A Robust Machine Learning Algorithm for Text Analysis." *Working Paper.*

Ke, Zheng, Bryan T Kelly, and Dacheng Xiu. 2019. "Predicting Returns with Text Data." *University of Chicago, Becker Friedman Institute for Economics Working Paper,* https://doi.org/http://dx.doi.org/10.2139/ssrn.3074808. https://ssrn.com/abstract=3389884.

Kelly, Bryan T, Seth Pruitt, and Yinan Su. 2019. "Characteristics are covariances: A unified model of risk and return." *Journal of Financial Economics* 134 (3): 501–524. ISSN: 0304-405X. https://doi.org/https://doi.org/10.1016/j.jfineco.2019.05.001. http://www.sciencedirect.com/science/article/pii/S0304405X19301151.

Kirby, Chris. 2019. "Firm Characteristics, Cross-Sectional Regression Estimates, and Asset Pricing Tests." *The Review of Asset Pricing Studies,* ISSN: 2045-9920. https://doi.org/10.1093/rapstu/raz005. https://doi.org/10.1093/rapstu/raz005.

Kogan, Leonid, and Dimitris Papanikolaou. 2014. "Growth Opportunities, Technology Shocks, and Asset Prices." *The Journal of Finance* 69 (2): 675–718. ISSN: 1540-6261. https://doi.org/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. ISSN: 15406261. https://doi.org/10.1111/jofi.12612. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12612%20https://onlinelibrary-wiley-com.ezproxy.library.bi.no/doi/full/10.1111/jofi.12612%20https://onlinelibrary-wiley-com.ezproxy.library.bi.no/doi/abs/10.1111/jofi.12612%20https://onlinelibrary-wiley-com.e.

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. https://doi.org/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. https://doi.org/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. Stroudsburg, PA, USA: Association for Computational Linguistics. https://doi.org/10.3115/1118108.1118117. https://doi.org/10.3115/1118108.1118117.

Lopez-Lira, Alejandro. 2021. "Why do managers disclose risks accurately? Textual analysis, disclosures, and risk exposures." *Economics Letters* 204:109896. ISSN: 01651765. https://doi.org/10.1016/j.econlet.2021.109896. https://linkinghub.elsevier.com/retrieve/pii/S0165176521001737.

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

Loughran, Tim, Bill McDonald, and Ioannis Pragidis. 2019. "Assimilation of oil news into prices." *International Review of Financial Analysis* 63 (May): 105–118. ISSN: 10575219. https://doi.org/http://dx.doi.org/10.2139/ssrn.3074808. https://ssrn.com/abstract=3074808.

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

McLean, R David, and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32. https://doi.org/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, no. 2 (November): 277–309. ISSN: 0304405X. https://doi.org/10.1016/j.jfineco.2004.08.008. http://www.sciencedirect.com/science/article/pii/S0304405X05000735.

Rehurek, Radim, and Petr Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (May): 45–50. ISSN: 2951740867. https://is.muni.cz/publication/884893/en.

Stambaugh, Robert F., and Yu Yuan. 2017. "Mispricing factors." *Review of Financial Studies* 30, no. 4 (April): 1270–1315. ISSN: 0893-9454. https://doi.org/10.1093/RFS/HHW107. https://academic.oup.com/rfs/article/30/4/1270/2965095.

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. https://doi.org/10.1111/j.1540-6261.2005.00725.x. http://dx.doi.org/10.1111/j.1540-6261.2005.00725.x.

7 Tables

Table 1: Number of firms by the Hoberg and Phillips (2016) Fixed 25 - Industry Clusters for firms that are exposed to the Innovation Risk Factor

Industry Cluster Number	Percentage
2	50.92%
13	21.57%
8	15.03%
Other	12.48%

The Table shows the time-series average distribution of firms across the Hoberg and Phillips (2016) fixed 25 industries for the firms whose innovation risk-weights are at least 25%.

Table 2: Number of firms by SIC code for firms that are exposed to the Innovation Risk Factor

2-Digit SIC Code	Industry	Division	Number of firms
35	Manufacturing	Raw Inputs 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 time-series average distribution of firms across the 2-digit SIC industries for the firms whose innovation risk-weights are at least 25%.

Table 3: Descriptive statistics

Statistic	N	Mean	St. Dev.	Pctl(25)	Pctl(75)
Pairwise Correlation	3,347,132	0.20	0.15	0.10	0.30
Risk Simmilarity	3,347,132	0.14	0.14	0.03	0.20
Beta Exposure	3,347,132	1.25	0.41	0.97	1.50
Book-to-Market Distance	3,347,132	1.05	3.20	0.17	0.92
Size Distance	3,347,132	2.23	1.69	0.89	3.21

The table shows the average, standard deviation and selected percentiles of companies pairwise correlation, risk similarity, book-to-market and size distance. Risk similarity is defined as one minus the pairwise Hellinger distance between the firms' time-series average of the risk weights. Beta exposure is the pairwise product of firms' time-series average betas measured by using yearly time-series regressions of the firms' returns against the market using daily data. The distance between the log of size and book to market are measured as the pairwise difference between the absolute value of the time-series average of the respective measures. The period spans 2006–2019

Table 4: Impact of risk similarity on correlation

	Pairwise Correlation		
	(1)	(2)	(3)
Risk Simmilarity	0.230^{***}	0.227^{***}	0.225^{***}
	$t = 3.82$	$t = 3.87$	$t = 3.84$
Beta Exposure		0.142^{***}	0.141^{***}
		$t = 4.54$	$t = 4.30$
Constant	0.176^{***}	0.019^{***}	0.027^{***}
	$t = 13.84$	$t = 5.26$	$t = 6.017$
Observations	$3,210,796$	$3,619,868$	$3,160,469$
Controls	No	No	Yes
Adjusted R^2	0.044	0.103	0.104
Note:			*p<0.1; **p<0.05; ***p<0.01

The table shows the results of a panel regression of realized stock correlation on average risk similarity, as well as on the beta exposure and controls. Risk similarity is defined as one minus the pairwise Hellinger distance between the firms' time-series average of the risk weights. Beta exposure is the pairwise product of firms' time-series average betas measured by using yearly time-series regressions of the firms' returns against the market using daily data. The controls include the distance between the log of size and book to market and a dummy for whether the companies are in the same 3 digits SIC industry classification. The distance between the log of size and book to market are measured as the pairwise difference between the absolute value of the time-series average of the respective measures. The standard errors are bootstrapped. The period spans 2006–2019

Table 5: GRS Test: Are the FIRFs spanned?

	Firm Identified Risk Factors			Orthogonal Factors			HP Industries		
	GRS	p-value	R^2	GRS	p-value	R^2	GRS	p-value	R^2
CAPM	2.35	0.001	0.48	1.69	0.03	0.46	1.42	0.11	0.63
FF5	5.5	4.1e-11	0.58	2.92	4.46e-05	0.64	1.56	0.07	0.70

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. First row corresponds to using the market portfolio as the unique factor, second row corresponds to the Fama and French (2015) Factor Model. First column shows the result of the pricing of the firm identified risk factors. The second column shows the result of the pricing of the orthogonal factors. The third column shows the result of pricing portfolios using Hoberg and Phillips (2016) industries.

Table 6: GRS Test: Is the Five-Factor Model spanned?

	FF5 Model		
	GRS	p-value	R^2
Firm Identified Risk Factors	6.83	2.6e-7	0.61

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. The row corresponds to all Firm Identified Risk Factors as pricing factors. The column shows the result of the pricing of the Fama and French (2015) Factor Model by the firms identified risk factors.

Table 7: Projection of the Five-Factor Model plus Momentum on the FIRFs

	Intercept	t-stat	R^2
$R_m - R_f$	.08	0.84	0.93
SMB	-0.27	-2.04	0.60
HML	0.10	0.62	0.49
RMW	0.36	3.43	0.42
CMA	0.14	1.31	0.27
Mom	-0.21	-0.65	0.45

The Table shows the estimate of α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the Fama-French Five-Factors, and the pricing factors f^e_{t+1} are the firm identified risk factors. $R_m - R_f$, the excess return on the market, SMB (Small Minus Big), HML (High Minus Low), RMW (Robust Minus Weak), CMA (Conservative Minus Aggressive), Mom (Momentum) are taken from French's website. R^2 is the adjusted coefficient of determination.

Table 8: GRS Test for the 4-Factor FIRFs Model and the Fama–French 5 Factor Model

	49 Industry + 25 B-to-M			49 Industry + 25 B-to-M + 15 α		
	GRS	p-value	R^2	GRS	p-value	R^2
4 FIRFs	1.53	0.03	0.69	2.043	0.007	0.639
Fama-French 5 Factor Model	1.69	0.01	0.76	2.271	0.003	0.731
Mispricing Factors	1.91	0.01	0.76	2.070	0.006	0.724
q-factor Model	1.62	0.02	0.73	2.328	0.002	0.704
All FIRFs	1.8	0.01	0.82	2.007	0.009	0.804
All FIRFs regularized	1.57	0.03	0.79	1.670	0.064	0.761

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 firm identified risk 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), fourth row corresponds to the q-factor model of Hou, Xue, and Zhang (2015), fifth row corresponds to using all 25 of the risks, and sixth row corresponds to using all 25 of the risks and estimating the betas using LASSO regression. 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.

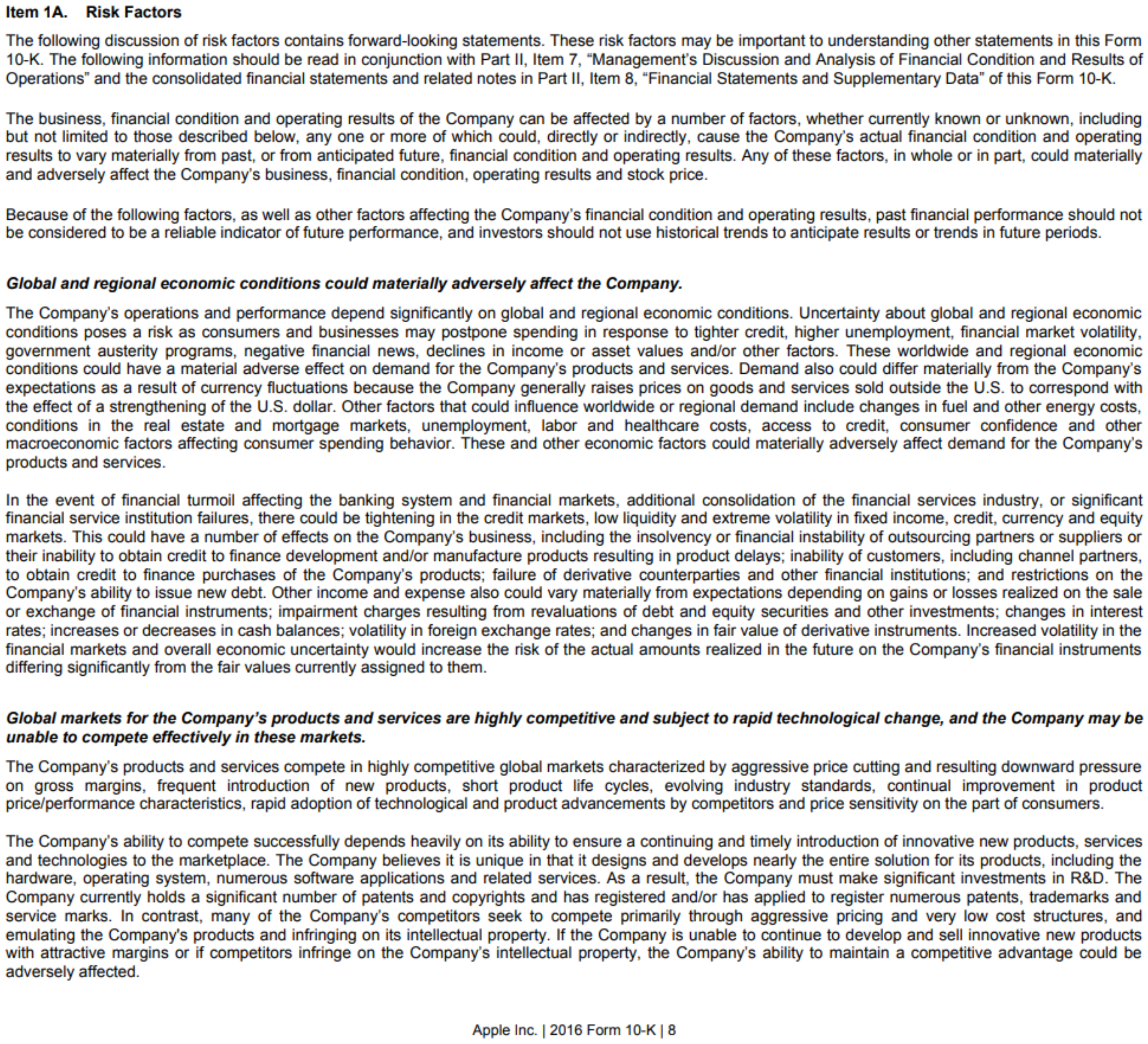
Table 9: GRS Test for the 4-Factor FIRFs Model and the Fama-French 5 Factor Model

	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
FIRFs 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	0.04	0.92	0.68	0.75	0.52
q-factor Model	1.47	0.073	0.67	1.88	0.02	0.92	1.13	0.35	0.43
All FIRFs	1.15	0.29	0.75	2.02	0.008	0.80	1.49	0.15	0.31
All FIRFs regularized	0.90	0.65	0.73	1.70	0.04	0.75	0.90	0.65	0.73

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 firm identified risk 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), fourth row corresponds to the q-factor model of Hou, Xue, and Zhang (2015), fifth row corresponds to using all 25 of the risks, and sixth row corresponds to using all 25 of the risks and estimating the betas using LASSO regression. 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).

8 Figures

Figure 1: Excerpt from Item 1A: Risk Factors in Apple Inc. Annual Report



Item 1A. Risk Factors

The following discussion of risk factors contains forward-looking statements. These risk factors may be important to understanding other statements in this Form 10-K. The following information should be read in conjunction with Part II, Item 7, “Management’s Discussion and Analysis of Financial Condition and Results of Operations” and the consolidated financial statements and related notes in Part II, Item 8, “Financial Statements and Supplementary Data” of this Form 10-K.

The business, financial condition and operating results of the Company can be affected by a number of factors, whether currently known or unknown, including but not limited to those described below, any one or more of which could, directly or indirectly, cause the Company’s actual financial condition and operating results to vary materially from past, or from anticipated future, financial condition and operating results. Any of these factors, in whole or in part, could materially and adversely affect the Company’s business, financial condition, operating results and stock price.

Because of the following factors, as well as other factors affecting the Company’s financial condition and operating results, past financial performance should not be considered to be a reliable indicator of future performance, and investors should not use historical trends to anticipate results or trends in future periods.

Global and regional economic conditions could materially adversely affect the Company.

The Company’s operations and performance depend significantly on global and regional economic conditions. Uncertainty about global and regional economic conditions poses a risk as consumers and businesses may postpone spending in response to tighter credit, higher unemployment, financial market volatility, government austerity programs, negative financial news, declines in income or asset values and/or other factors. These worldwide and regional economic conditions could have a material adverse effect on demand for the Company’s products and services. Demand also could differ materially from the Company’s expectations as a result of currency fluctuations because the Company generally raises prices on goods and services sold outside the U.S. to correspond with the effect of a strengthening of the U.S. dollar. Other factors that could influence worldwide or regional demand include changes in fuel and other energy costs, conditions in the real estate and mortgage markets, unemployment, labor and healthcare costs, access to credit, consumer confidence and other macroeconomic factors affecting consumer spending behavior. These and other economic factors could materially adversely affect demand for the Company’s products and services.

In the event of financial turmoil affecting the banking system and financial markets, additional consolidation of the financial services industry, or significant financial service institution failures, there could be tightening in the credit markets, low liquidity and extreme volatility in fixed income, credit, currency and equity markets. This could have a number of effects on the Company’s business, including the insolvency or financial instability of outsourcing partners or suppliers or their inability to obtain credit to finance development and/or manufacture products resulting in product delays; inability of customers, including channel partners, to obtain credit to finance purchases of the Company’s products; failure of derivative counterparties and other financial institutions; and restrictions on the Company’s ability to issue new debt. Other income and expense also could vary materially from expectations depending on gains or losses realized on the sale or exchange of financial instruments; impairment charges resulting from revaluations of debt and equity securities and other investments; changes in interest rates; increases or decreases in cash balances; volatility in foreign exchange rates; and changes in fair value of derivative instruments. Increased volatility in the financial markets and overall economic uncertainty would increase the risk of the actual amounts realized in the future on the Company’s financial instruments differing significantly from the fair values currently assigned to them.

Global markets for the Company’s products and services are highly competitive and subject to rapid technological change, and the Company may be unable to compete effectively in these markets.

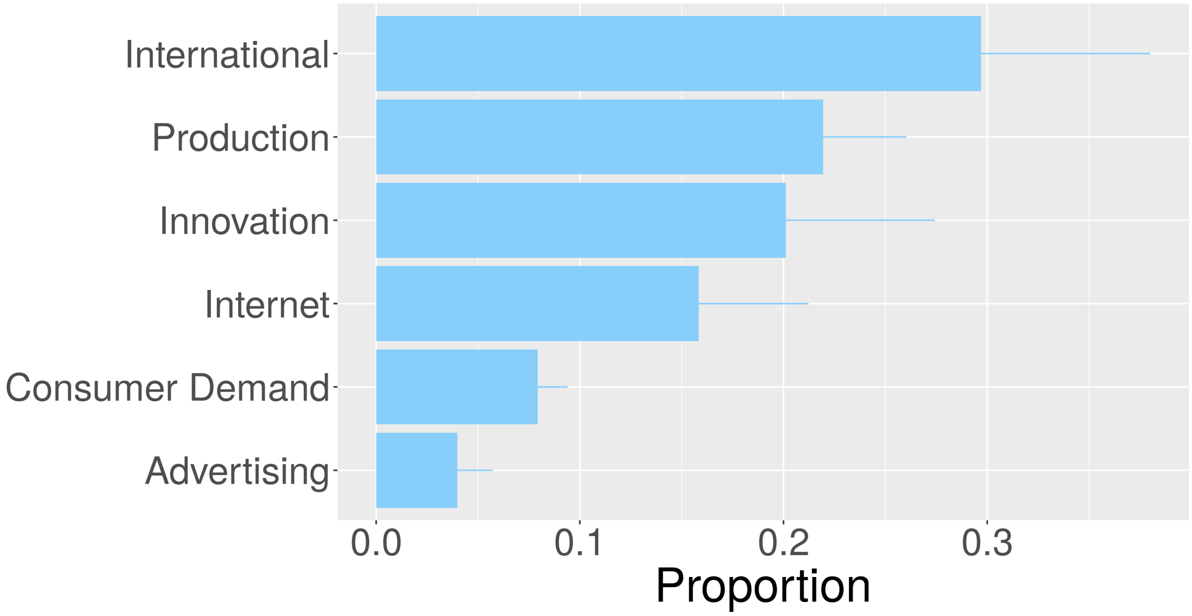
The Company’s products and services compete in highly competitive global markets characterized by aggressive price cutting and resulting downward pressure on gross margins, frequent introduction of new products, short product life cycles, evolving industry standards, continual improvement in product price/performance characteristics, rapid adoption of technological and product advancements by competitors and price sensitivity on the part of consumers.

The Company’s ability to compete successfully depends heavily on its ability to ensure a continuing and timely introduction of innovative new products, services and technologies to the marketplace. The Company believes it is unique in that it designs and develops nearly the entire solution for its products, including the hardware, operating system, numerous software applications and related services. As a result, the Company must make significant investments in R&D. The Company currently holds a significant number of patents and copyrights and has registered and/or has applied to register numerous patents, trademarks and service marks. In contrast, many of the Company’s competitors seek to compete primarily through aggressive pricing and very low cost structures, and emulating the Company's products and infringing on its intellectual property. If the Company is unable to continue to develop and sell innovative new products with attractive margins or if competitors infringe on the Company’s intellectual property, the Company’s ability to maintain a competitive advantage could be adversely affected.

Apple Inc. | 2016 Form 10-K | 8

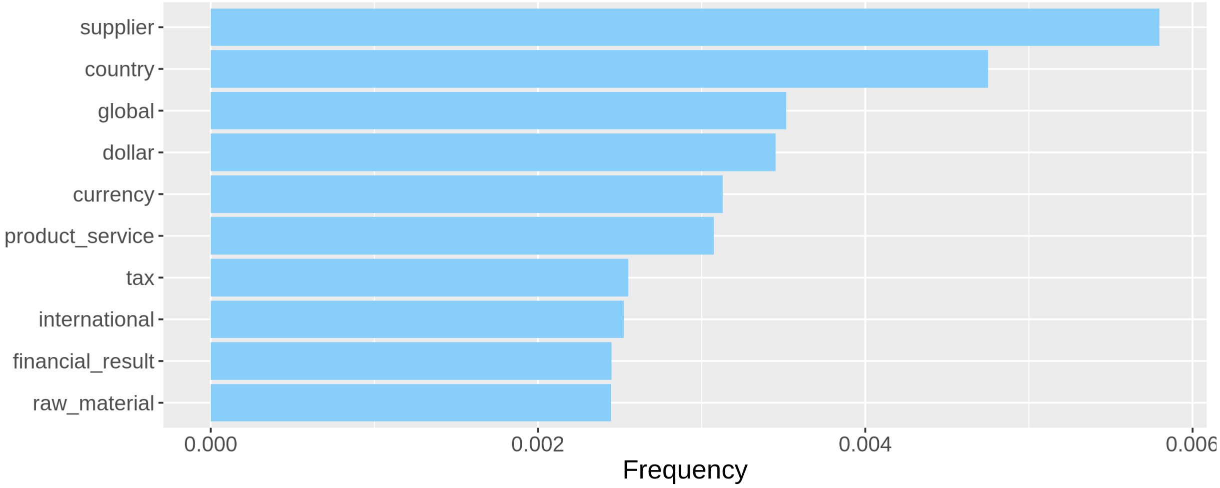
The Figure shows the first page out of ten of Item 1A: Risk Factors in Apple Inc. 10-K 2016 annual report. The document is available on the SEC EDGAR database.

Figure 2: Percentage of the risk disclosure that Apple Inc. allocates to each risk



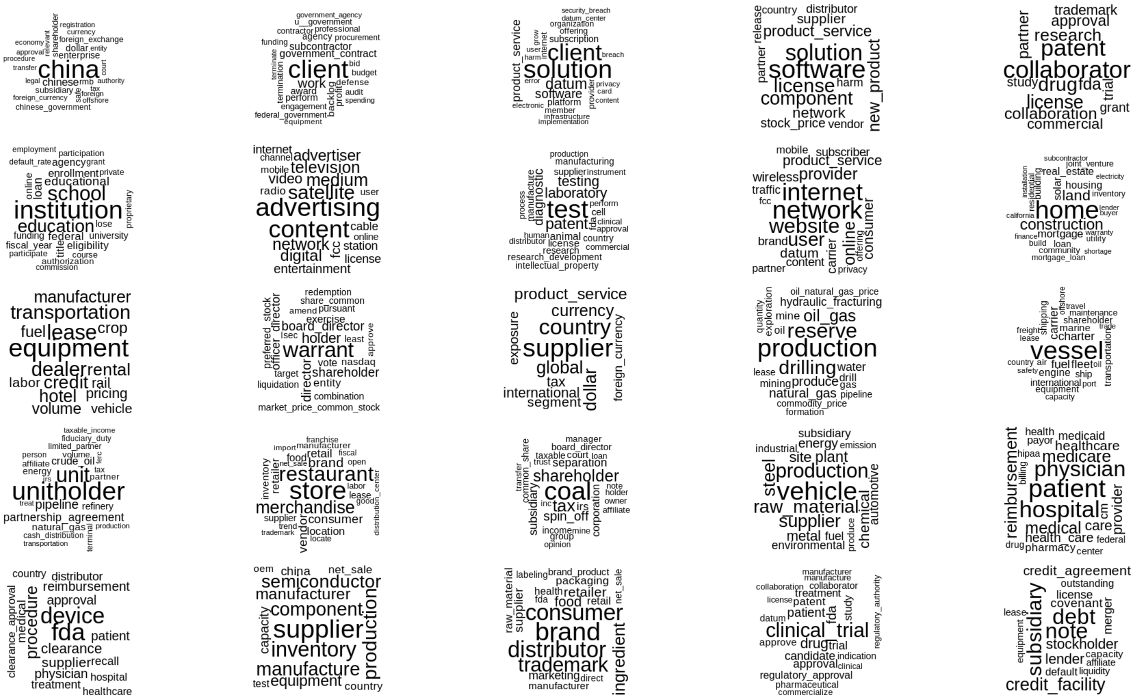
The table shows the average percentages of the risk disclosure that Apple Inc. allocates to each type of risk in the Section 1A: Risk Factors for their annual reports. The table only shows the five most discussed risks. The values are obtained using Latent Dirichlet Allocation.

Figure 3: International Risk Topic



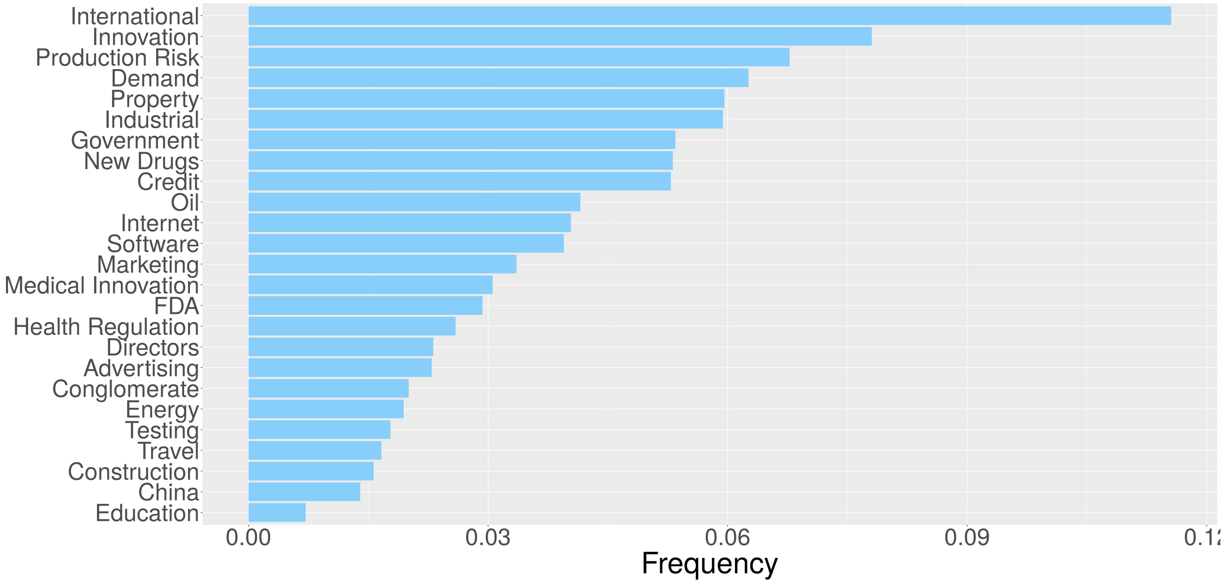
Distribution of the 10 most frequent words for the International Risk Topic (excluding ‘company’)

Figure 4: Risk Topics



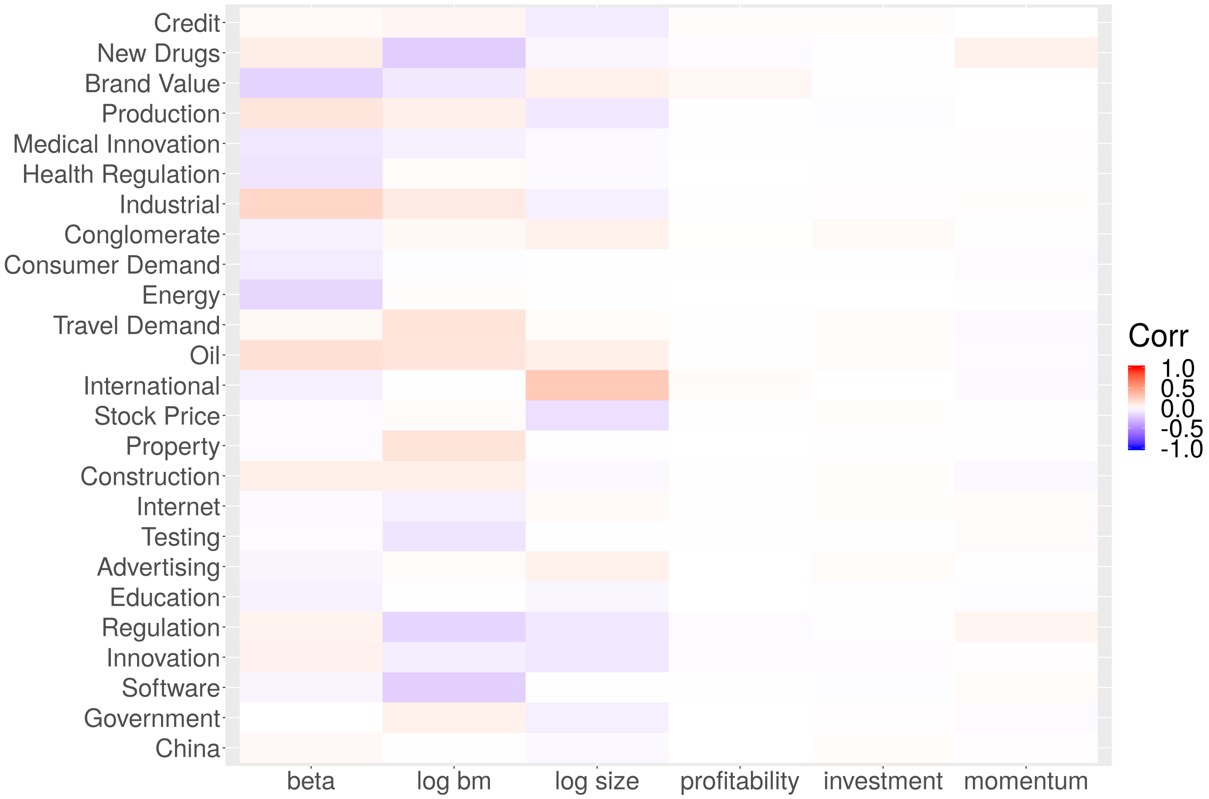
In this word cloud of the risks that firms face, a bigger font corresponds to a bigger weight for that word within each topic. See Section 1.2 for details on the procedure

Figure 5: Average percentage of the risk disclosure that firms allocates to each risk



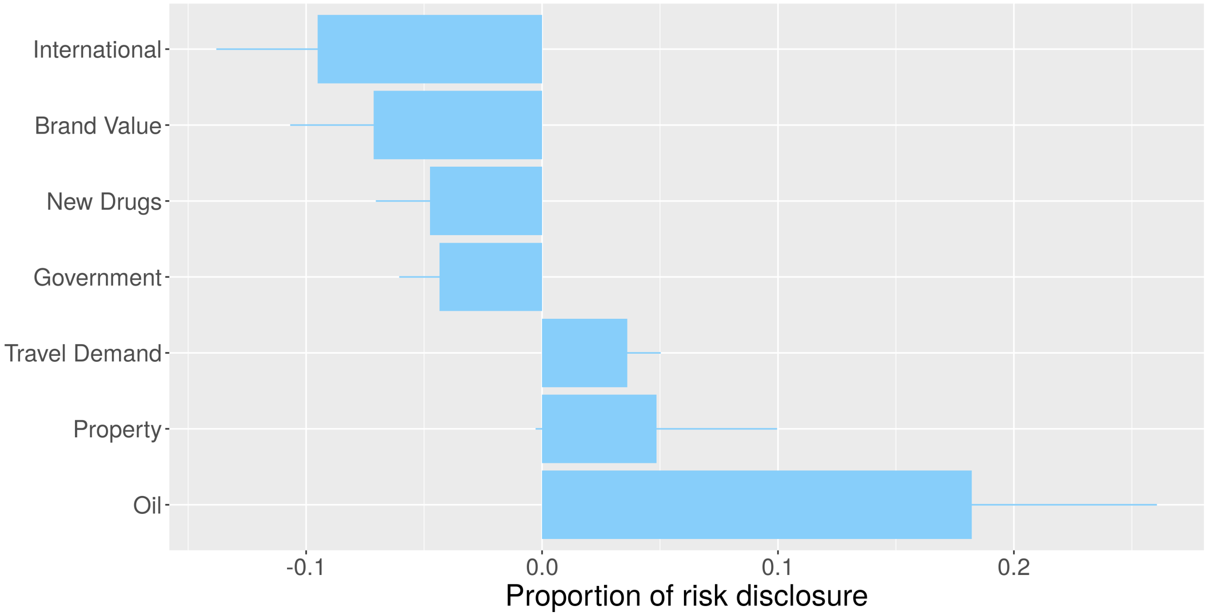
The figure shows the cross-sectional and time-series average of the percentage of the risk disclosure that firms allocates to each type of risk in the Section 1A: Risk Factors for the years 2006-2018. The values are obtained using Latent Dirichlet Allocation. See Sections 1 and 3 for details

Figure 6: Correlation of the risk weights with market beta, book-to-market, size, profitability, investment and past returns.



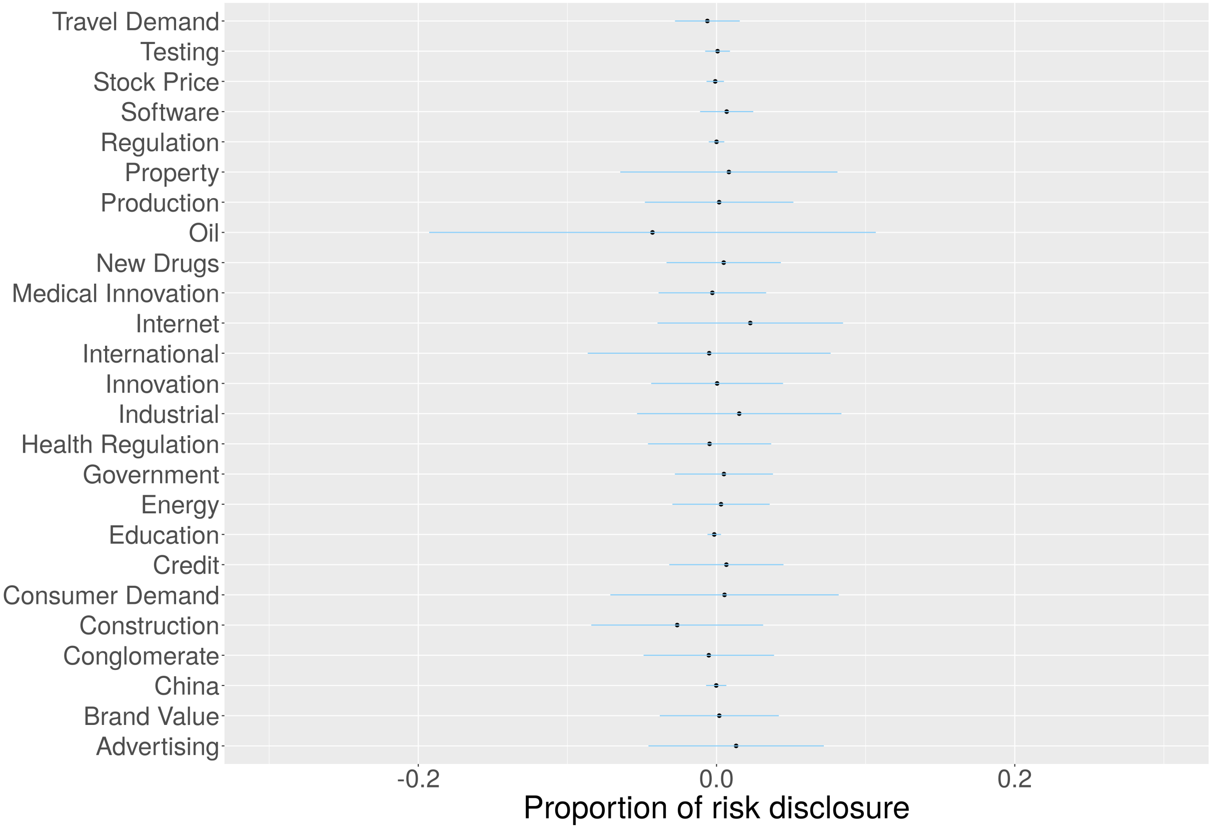
The Figure shows correlation between risk weights and some common predictors of the cross-section of returns. The sample period is 2006-2019. The predictors include yearly rolling window betas calculated with daily returns; and book-to-market ratios, size, profitability, and investment calculated as in Fama and French (2015). The data comes from the merged CRSP/Compustat database and the 10-K reports. The risk weights are calculated as in Section 1.2.

Figure 7: Implied average percentage of the risk disclosure that the High-minus-low (book-to-market) factor allocates to each risk



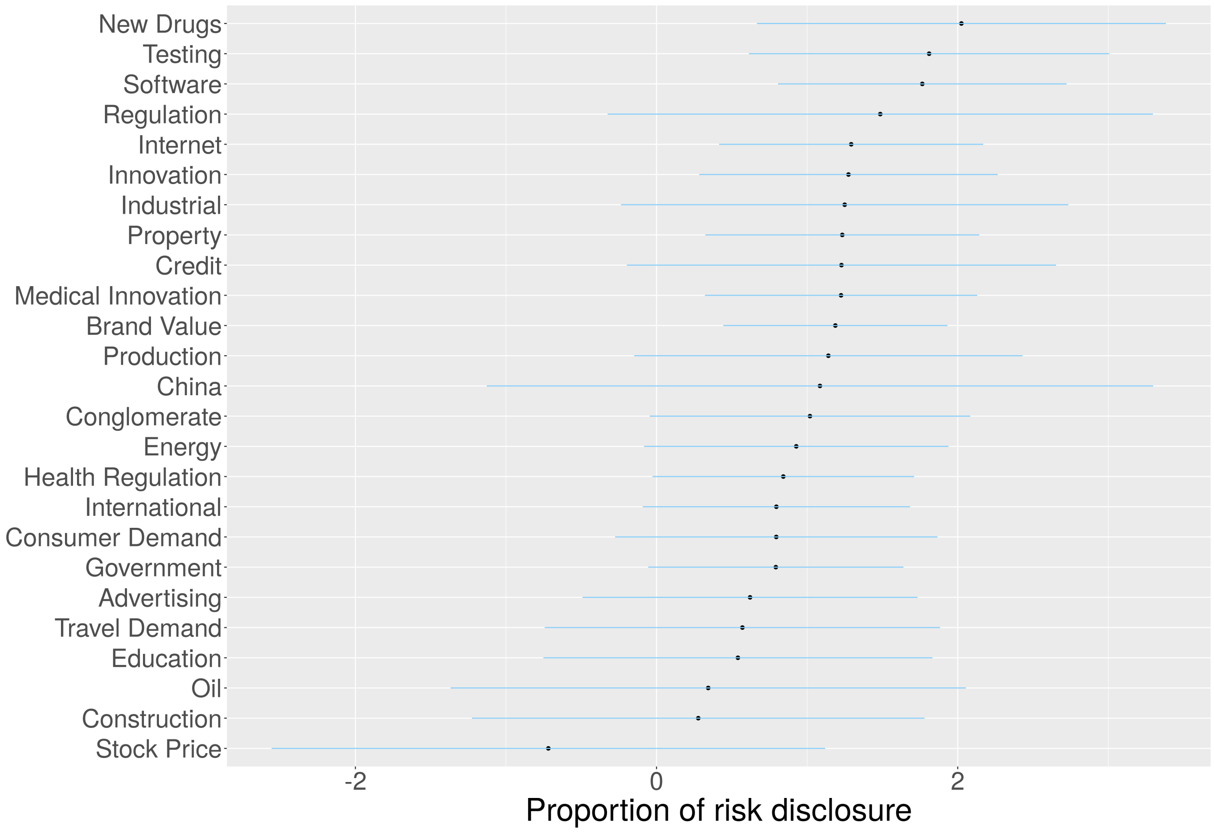
The figure shows the time-series average of the percentage of the weighted proportions of the risk disclosures that firms allocates to each type of risk in the Section 1A: Risk Factors for the years 2006-2018. The weights correspond to the weights in the High-minus-low (book-to-market) factor. The lines are one standard deviation from the average value. The values are obtained using Latent Dirichlet Allocation. See Sections 1 and 3 for details

Figure 8: Implied average percentage of the risk disclosure that the Momentum factor allocates to each risk



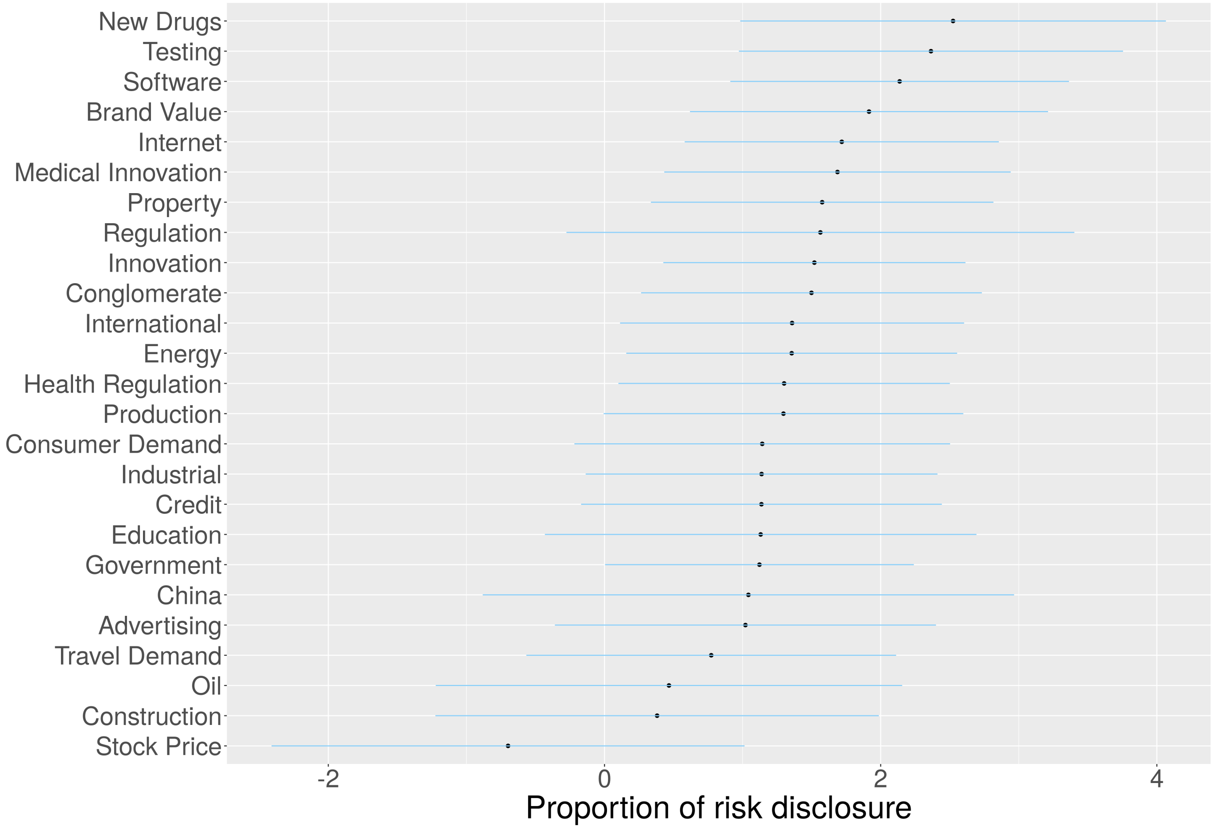
The figure shows the time-series average of the percentage of the weighted proportions of risk disclosures that firms allocates to each type of risk in the Section 1A: Risk Factors for the years 2006-2018. The weights correspond to the weights in the Momentum factor. The lines are one standard deviation from the average value. The values are obtained using Latent Dirichlet Allocation. See Sections 1 and 3 for details

Figure 9: Coefficients of the Risk Weights in a Cross-Sectional Predictive Regression



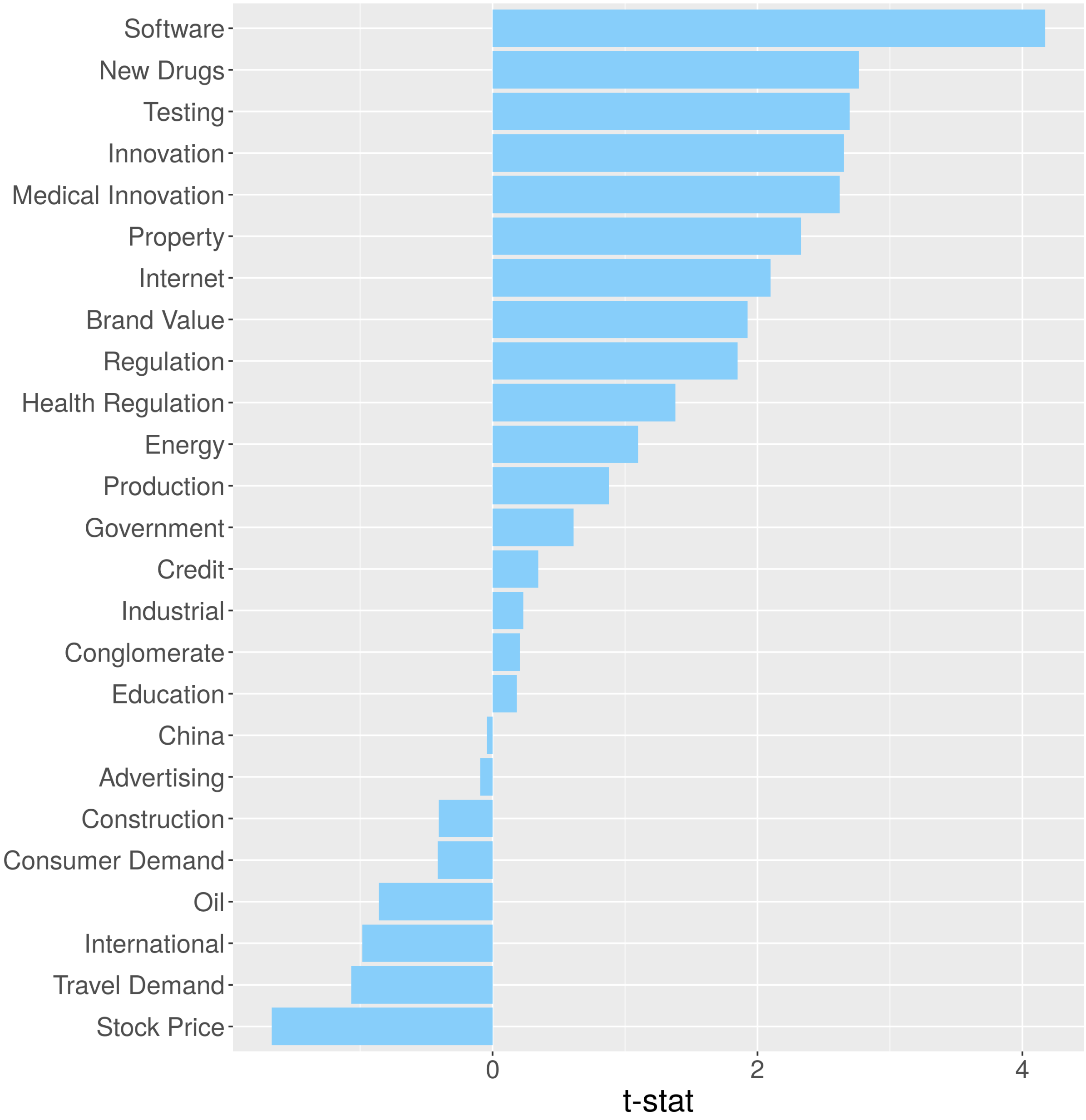
The Figure shows the coefficients associated with the risk weights in a Fama–MacBeth cross-sectional regression of next month returns on the risk weights of the form: $r^e_{i,t+1} = b'_t\theta_{i,t}$. The dot represents the coefficient and the bars represent 1.96 standard deviations.

Figure 10: Coefficients of the Risk Weights in a Cross-Sectional Predictive Regression with Controls



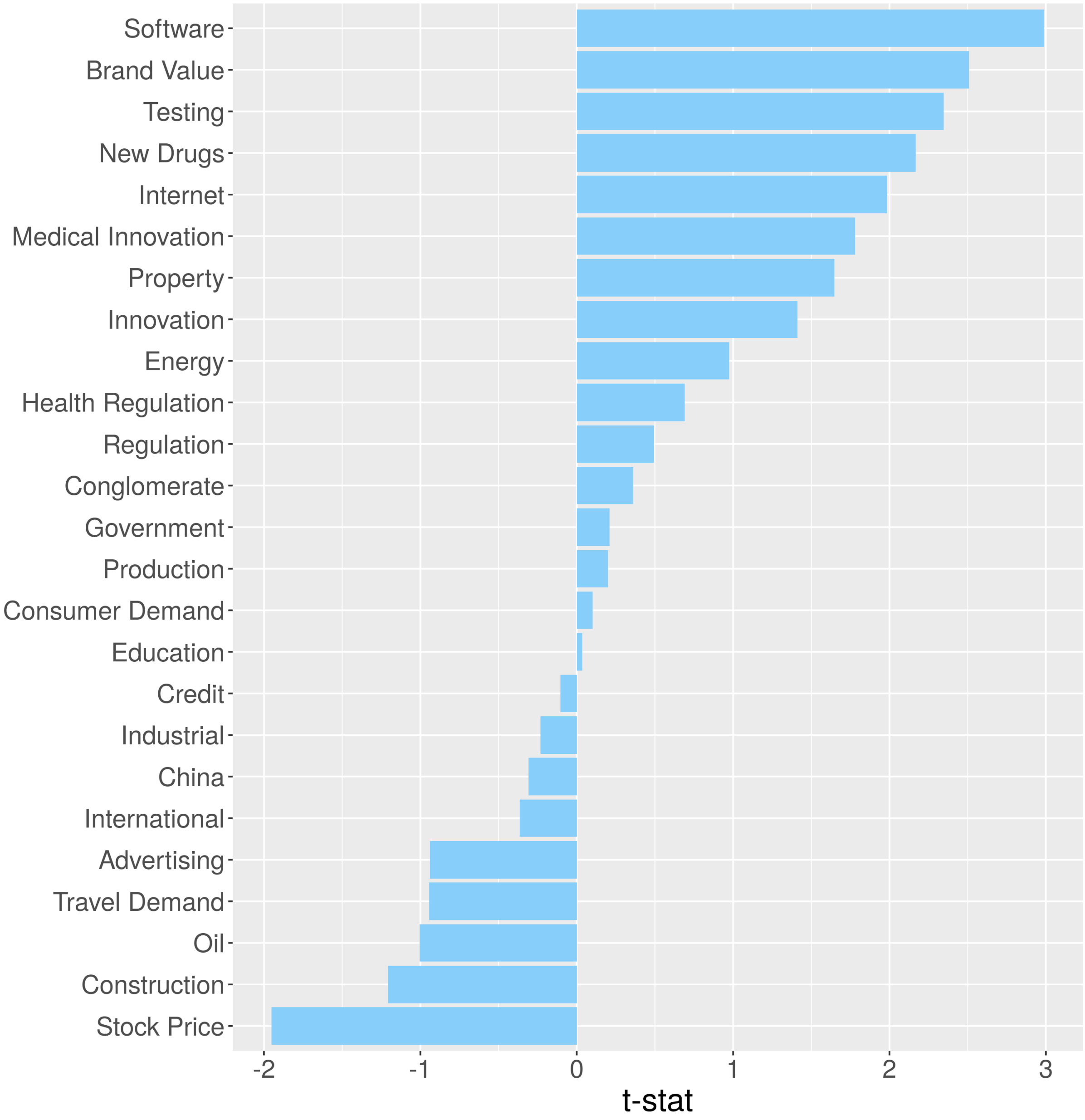
The Figure shows the coefficients associated with the risk weights in a Fama–MacBeth cross-sectional regression of next month returns on the risk weights of the form: $r^e_{i,t+1} = b'_t\theta_{i,t} + \gamma'_t x_{i,t}$. The controls include log of book-to-market, log of size, market beta, profitability and investment. The dot represents the coefficient and the bars represent 1.96 standard deviations.

Figure 11: 'Alphas' of the firm identified risk factors with respect to the Fama-French Five-Factor Model (t-stats)



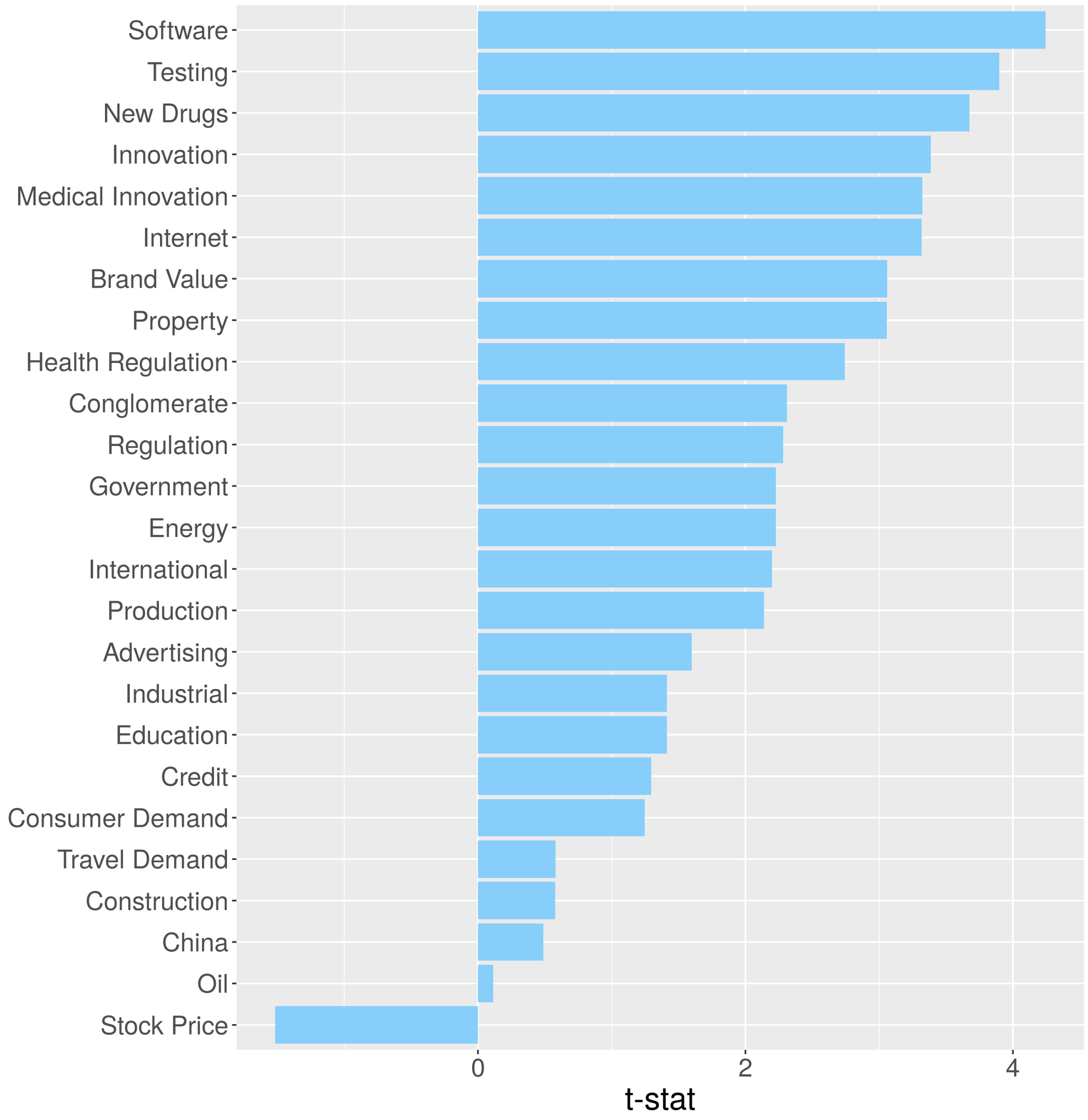
The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the firm identified risk factors, and the pricing factors f^e_{t+1} are the Fama-French Five-Factors. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 12: 'Alphas' of the firm identified risk factors with respect to CAPM (t-stats)



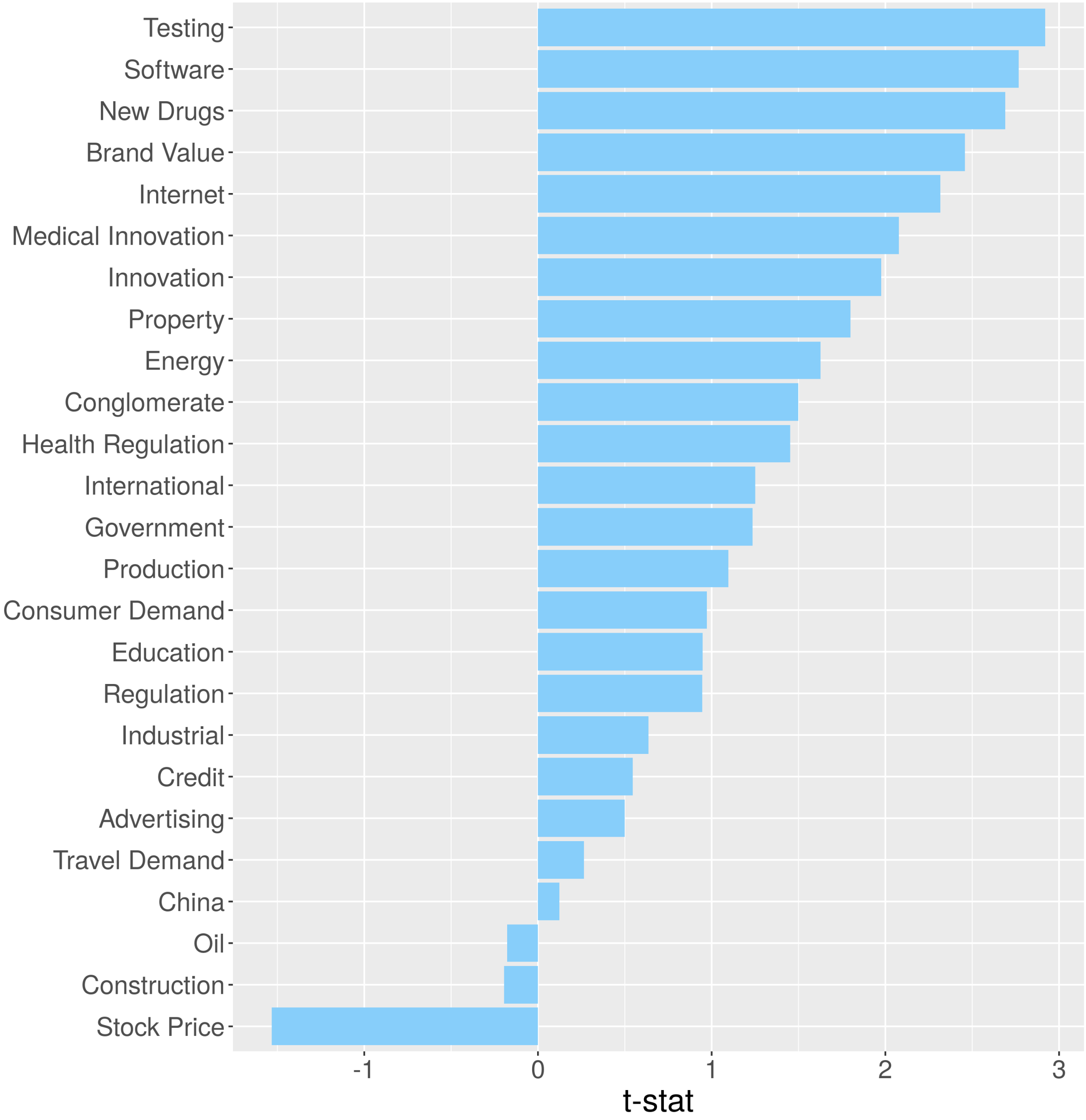
The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the firm identified risk factors, and the single pricing factor f^e_{t+1} is the excess return of the market portfolio. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 13: 'Alphas' of the orthogonal factors with respect to the Fama-French Five-Factor Model (t-stats)



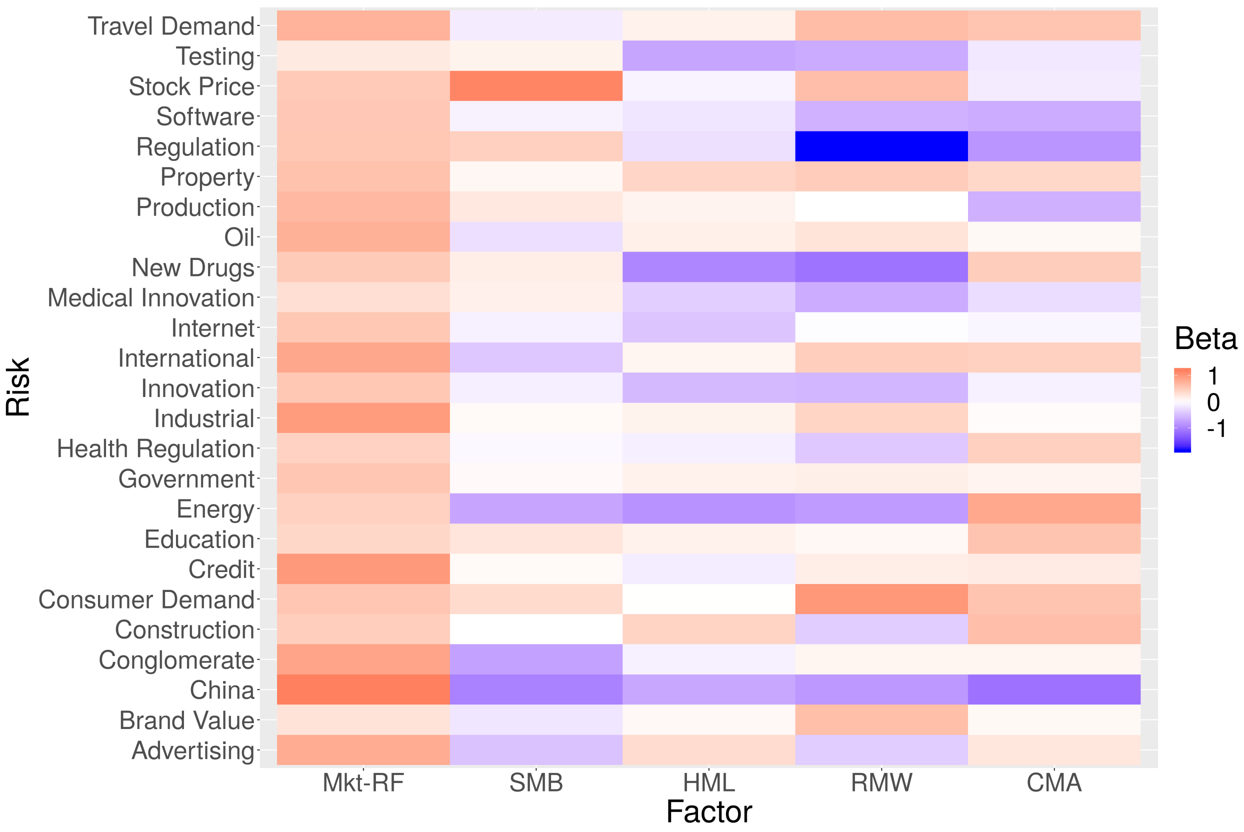
The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the orthogonal factors, and the pricing factors f^e_{t+1} are the Fama-French Five-Factors. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 14: 'Alphas' of the orthogonal factors with respect to CAPM (t-stats)



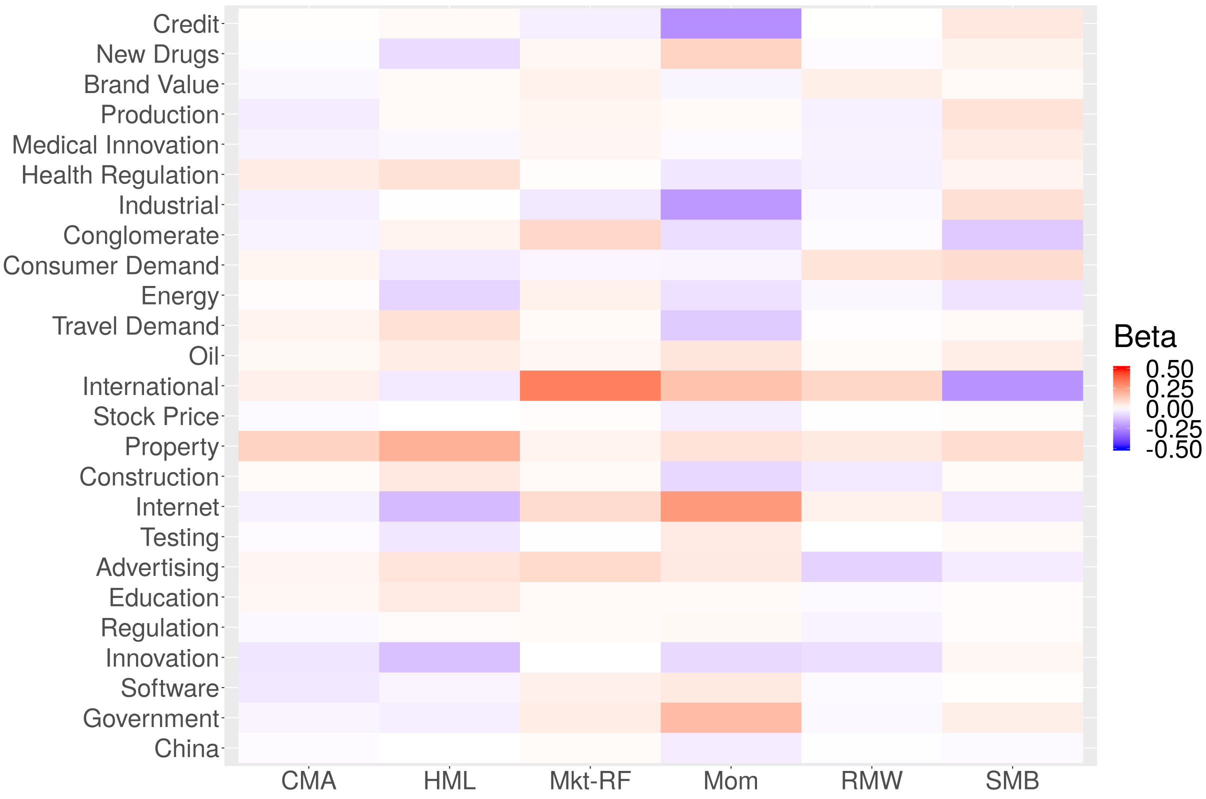
The Figure shows the t-stats of the coefficients α_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the orthogonal factors, and the single pricing factor f^e_{t+1} is the excess return of the market portfolio. The standard errors are adjusted for heteroskedasticity and autocorrelation.

Figure 15: Betas of the firm identified risk factors with respect to the Fama-French Five-Factor Model



The Figure shows the estimate of β_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the firm identified risk factors, and the pricing factors f^e_{t+1} are the Fama-French Five-Factors.

Figure 16: Betas of the Fama-French Five-Factor Model (plus momentum) with respect to the firm identified risk factors



The Figure shows the estimate of β_i in regressions of the form: $r^e_{i,t+1} = \alpha_i + \beta_i f^e_{t+1} + \epsilon_{i,t+1}$, $\alpha_i = 0$. The excess returns $r^e_{i,t+1}$ are the Fama-French Five-Factors, and the pricing factors f^e_{t+1} are the firm identified risk factors.

9 Appendix

9.1 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,[20] 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 18 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.[21]

[Insert Figure 17 about here]

[Insert Figure 18 about here]

[20] Manning, Raghavan, and Schütze (2008)

[21] Another subtle disadvantage of the Bag-of-Words model, is that it breaks multi-word concepts such as "real estate" into "real" and "estate", which have to be rejoined later, since counting those words separately will produce different results than counting the multi-word concept.

9.2 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] 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

[22] Manning, Raghavan, and Schütze (2008). 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.

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
- $count(B)$ 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 consecutively
- 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 estate) to also become phrases in the model.

9.3 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)).

9.4 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.

[Insert Figure 19 about here]

9.5 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 19. 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 Δ_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 θ_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, θ_{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 20. 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, θ: the topic distribution for document i. α: 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). β 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

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

a trade-off between the sparsity of the topics, i.e. how specialize they are, and the number of topics.

[Insert Figure 20 about here]

9.5.1 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. I use the technical measure of topic coherence and out of sample log likelihood to help determine the optimal number of topics. In this case 25 topics accomplish this task, and is consistent with the numbers used in the literature of topic modeling in finance applications (Israelsen (2014), Bao and Datta (2014), Hanley and Hoberg (2019)).

A natural challenge is then to further reduce the extracted risks into a lower number of portfolios for the cross-section. See Section 4 for more details.

9.5.2 Estimation

The estimation of the posterior parameters is done using the open-source software Gensim (Rehurek 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. Because it is an online algorithm, the estimation is performed on a rolling basis. As new risk disclosures arrive, the risk topics get updated, and we get a new set of weights (the projection of the documents on the topic space).

10 Appendix Figures

Figure 17: Steps for topic modelling

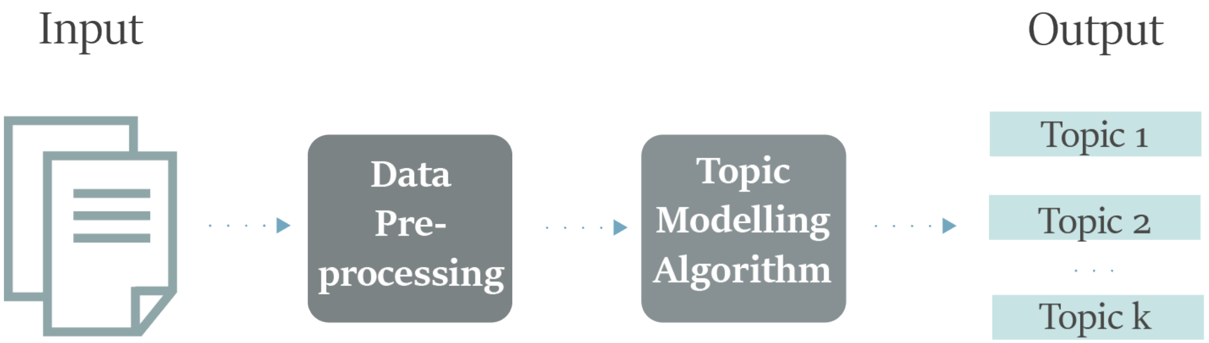
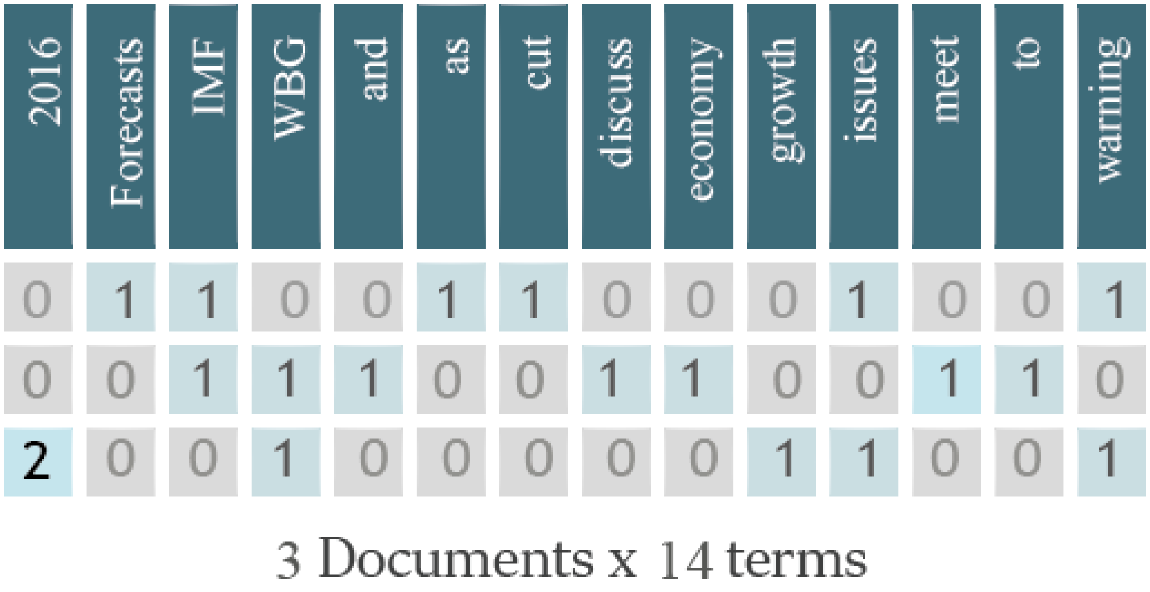


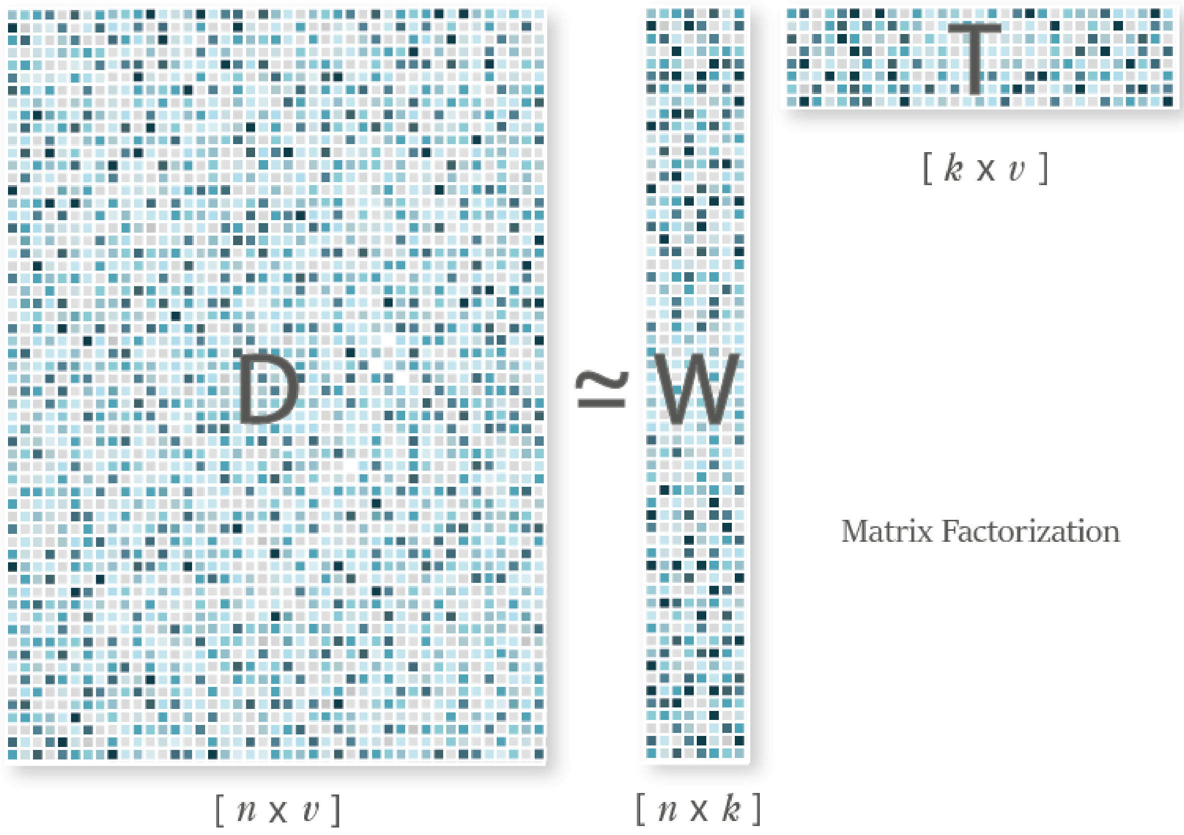
Figure 18: Example of a very simple document term matrix



2016	Forecasts	IMF	WBG	and	as	cut	discuss	economy	growth	issues	meet	to	warning
0	1	1	0	0	1	1	0	0	0	1	0	0	1
0	0	1	1	1	0	0	1	1	0	0	1	1	0
2	0	0	1	0	0	0	0	0	1	1	0	0	1

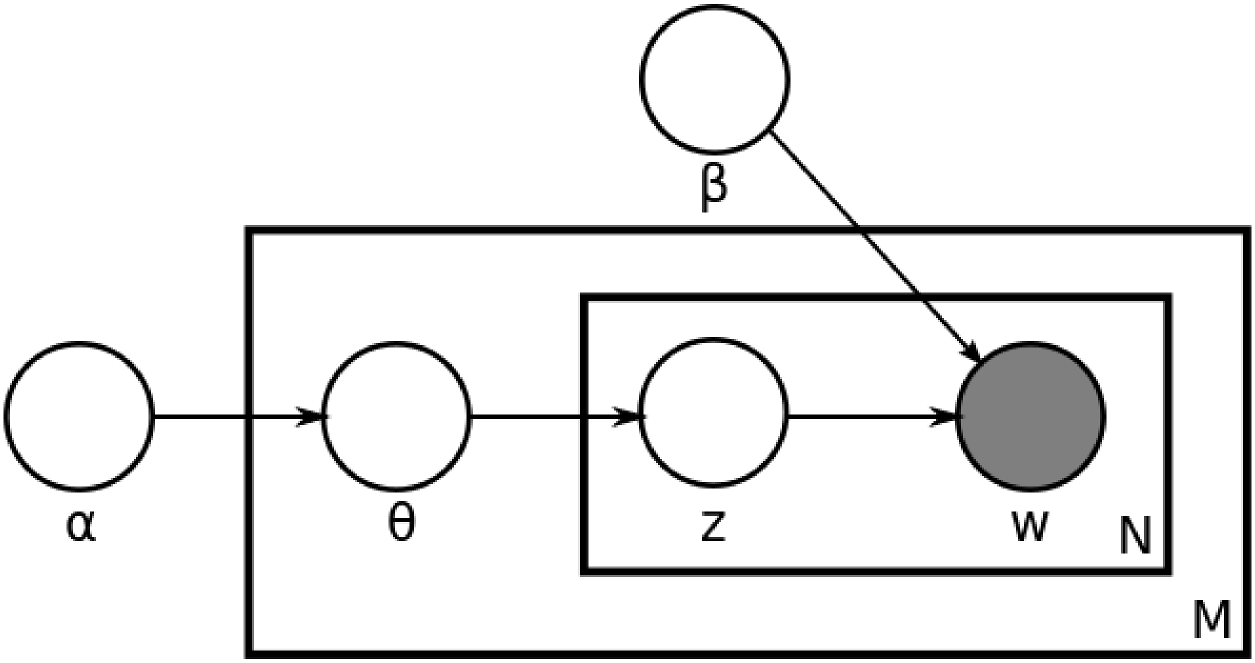
3 Documents x 14 terms

Figure 19: Intuition for Topic Modelling



The figure shows the intuition for topic modeling. The Matrix D is the Document-Term Matrix with dimensions n x v, n is the number of documents and v is the number of terms. The matrix is intuitively decomposed into two matrices: Matrix T and Matrix W. Matrix T has dimensions k x v, where k is the number of topics and v is the number of terms. Each row in Matrix T sums up to one and all the elements are non-negative. Hence, each topic is a distribution over words. Matrix W has dimensions n x k, where k is the number of topics and n is the number of documents. Each row in Matrix W sums up to one and all the elements are non-negative. Hence, its rows are distributions over topics, risk weights.

Figure 20: LDA Graphical Model



LDA in the language of graphical models. 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, θ: the topic distribution for document i. α: 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). β 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).

11 Online Appendix

Full online appendix available by request: alejandro.lopez-lira@warrington.ufl.edu