Ask EDGAR:
The Informational Content of Mutual Fund Prospectuses

Simona Abis and Anton Lines

Columbia Business School
Motivation

- How do investors allocate their capital across different asset managers?
  - Prior research mostly explained their decisions looking at past returns and other "hard information", delivering low explanatory power.
  - The SEC has been urging investors to be weary of past performance and to read prospectuses carefully when making investment decisions.

Mutual Fund Investing: Look at More Than a Fund's Past Performance

May 8, 2007

You can't open a newspaper or read a magazine without seeing ads promoting the stellar performance of "hot" mutual funds. But past performance is not as important as you may think, especially the short-term performance of relatively new or small funds. As with any investment, a fund's past performance is no guarantee of its future success. Over the long-term, the success (or failure) of your investment in a fund also will depend on factors such as:

So, look at more than the fund's past performance when making your investment decisions. Read the fund's prospectus and shareholder reports, and consider these tips:

https://www.sec.gov/reportspubs/investor-publications/investorpubsmfperformhtm.html
Motivation continued...

- **Are prospectuses informative?**
  - Investors also have access to “soft information”
    - e.g. marketing materials, regulatory disclosures, in-person meetings
  - Prospectuses are the main document containing funds information
    - likely proxying for other forms of communication
  - The SEC has been investing in
    - Educating investors about prospectuses
    - Monitoring their fair disclosure

**BUT**

- Disclosure requirements have been shown to be very costly for funds
- The SEC has not systematically shown their usefulness
- Investors might not be paying attention to them
This Paper

- **Question:** are prospectuses informative above and beyond what can be learned from “hard information”?
  - **DATA:**
    - Collect and categorize all textual information disclosed by mutual funds through the EDGAR system
    - Developed a comprehensive parsing algorithm which allows for automatic collection, mapping and parsing of historical filings
    - Focus on active equity mutual funds in the US
  - **DESCRIPTIVE:**
    - Characterize prospectuses’ content by conditioning on the same regulatory questions across funds and over time
      - Length, complexity, sentiment
      - Clustering (unsupervised learning)
  - **ANALYSIS:**
    - Supervised learning to predict funds’ likelihood to incur in agency-like behaviours (an example: risk-shifting)
    - Work in progress:
      → Prediction of likelihood of legal actions
      → Unsupervised learning to predict future return distribution
cross-sectional differences in mandatory disclosures

mutual fund prospectuses
We'll apply supervised and unsupervised learning to extract signals from mutual fund prospectuses

risk-shifting behavior
We'll use supervised learning applied to mutual fund prospectuses to predict funds risk-shifting behavior


We'll apply supervised and unsupervised learning to extract signals from mutual fund prospectuses

Risk shifting: Huang, Siau and Zhang (2007); Teelock, Sear, Tschantz, Macskassy, Mackensy (2007); Teelock, Sear, Tschantz, Macskassy, Mackensy, Mackensy (2007); Teelock, Sear, Tschantz, Macskassy, Mackensy, Mackensy (2007).
Roadmap

1. Introduction
2. Data
3. Descriptive Analysis
4. Empirical Analysis
5. Future Research

- Work in progress...
- Risk Shifting
- Clustering
- Length, Complexity, Sentiment

6. Conclusion
Prospectuses Availability

- Funds are required to publish prospectuses regularly
- There are clear guidelines regarding the information these should contain
  - Funds can be sued by the SEC for misrepresenting their behavior
- Prospectuses are publicly available through the EDGAR system since 1995
  - Sophisticated investors can automate access using the FTP of the SEC
  - Retail investors might also access prospectuses of selected funds through their online brokerage accounts or by post
Prospectuses Description

- They are divided in sections addressing different regulatory questions e.g.:
  - Principal Investment Strategies (PIS)
  - Principal Risks (PIR)
- The content, writing style and length of different sections vary substantially
  - Crucial to condition on sections when comparing text cross-sectionally
- Regulatory requirements can be satisfied in just a few sentences
  - Some funds choose to write substantially more

E.g.: Vanguard - JAG Large Cap Growth Fund

**Principal Investment Strategies**

The Fund invests primarily in common stocks of U.S. companies that the Fund’s advisor believes have strong earnings and revenue growth potential. Under normal conditions, the Fund will invest at least 80% of the Fund’s net assets plus any borrowings for investment purposes in large cap stocks defined as stocks of companies with market capitalizations of at least $8 billion.

The advisor’s employs a bottom-up, quantitatively-derived buy discipline to identify stocks the advisor believes have superior earnings and revenue growth characteristics. The cornerstone of the advisor’s investment process is a proprietary multi-factor model that scores several thousand equity securities according to a variety of weighted factors measuring earnings and revenue growth, valuation, size and relative strength. The sell discipline is designed to eliminate portfolio holdings with inferior price performance and deteriorating earnings and revenue growth factors.

The Fund actively trades its portfolio investments, which may lead to higher transaction costs that may affect the Fund’s performance.

**Principal Risks of Investing in the Fund**

As with any mutual fund, there is no guarantee that the Fund will achieve its objective. Investment markets are unpredictable and there will be certain market conditions where the Fund will not meet its investment objective and will lose money. The Fund’s net asset value and returns will vary and you could lose money on your investment in the Fund and those losses could be significant.

The following summarizes the principal risks of investing in the Fund. These risks could adversely affect the net asset value, total return and the value of the Fund and your investment.

- **Equity Securities Risks.** Common stocks are subject to market risks that affect the value of the Fund. Factors such as interest rate levels, market conditions, and political events may adversely affect equity prices.

- **Management Risk.** The Portfolio Manager’s judgments about the attractiveness, value and potential appreciation of particular stocks, options or other securities in which the Fund invests or sells short may prove to be incorrect and there is no guarantee that the Portfolio Manager’s
The EDGAR Mutual Fund database includes over 1 million filings. The historical data is highly unstructured. Isolating single sections for all funds over time is complex.

Types of information:

- Traditional fund-level data (returns, AUM, fees, etc)

- "Hard information" (CRSP/Thomson Mutual Fund Database)

- "Soft information" (EDGAR)

A separate variable for each section of the prospectus (P1S and P1R)

We have a total of 40,000 prospectuses correctly parsed (1994-2006 work in progress...)

It produces reliable information as of 2006:

We parsed both prospectuses and N-SAR filings.

The parsing job has so far been applied to US active equity mutual funds.

The EDGAR Mutual Fund database includes over 1 million filings.

We parsed both prospectuses and N-SAR filings.

Parser
Roadmap

1. Introduction
2. Data
3. Descriptive Analysis
   - Length, Complexity, Sentiment
   - Clustering
4. Empirical Analysis
   - Risk Shifting
   - Work in progress...
5. Future Research
6. Conclusion
PIS vs. PIR

- Strategy descriptions (PIS) are substantially shorter than Risk ones (PIR)
- But they are harder to understand - Dale Chall Score:
  \[
  \begin{cases}
  3.6365 \frac{\text{nonDaleChalCount}}{\text{wordCount}} > 0.5 \\
  0 \quad \text{otherwise}
  \end{cases}
  + 15.79 \frac{\text{nonDaleChalCount}}{\text{wordCount}} + 0.0496 \frac{\text{wordCount}}{\text{sentCount}}
  \]
PIS vs. PIR

- Strategy descriptions (PIS) are substantially shorter than Risk ones (PIR)

- But they are harder to understand - Dale Chall Score:
  \[
  \begin{cases}
  3.6365 & \frac{\text{nonDaleChalCount}}{\text{wordCount}} > 0.5 \\
  0 & \text{otherwise}
  \end{cases}
  + 15.79 \times \frac{\text{nonDaleChalCount}}{\text{wordCount}} + 0.0496 \frac{\text{wordCount}}{\text{sentCount}}
  \]

![Dale Chall Score Chart]

- Score School Grade Comprehension
  - <= 4.9: 4th grade or lower
  - 5.0-5.9: 5th or 6th grade
  - 6.0-6.9: 7th or 8th grade
  - 7.0-7.9: 9th or 10th grade
  - 8.0-8.9: 11th or 12th grade
  - 9.0-9.9: college student
  - >= 10: college graduate

-abis and Lines

Columbia University
Using Loughran and McDonald sentiment word lists we find that PIR contain:

- A higher frequency of Negative words
- A higher frequency of Uncertainty and Litigious and Constraining words
- A lower frequency of Positive words

**Sentiment**, \( t_i = T_i \cdot T_i^0 + v_i + w_i + u_i \cdot e_i^t \)

\begin{center}
\begin{tabular}{cccccc}
 & 24716 & 24716 & 24716 & 24716 & 24716 \\
N & 58.34 & -12.29 & 0.495*** & 0.953*** & 0.252*** \\
 & (-58.34) & (-12.29) & 0.953*** & 0.495*** & 0.252*** \\
 & **0.252*** & **0.495*** & **0.953*** & **0.252*** & **0.495*** \\
\end{tabular}
\end{center}

\( t \) statistics in parentheses

- \* \( p<0.10 \)
- \** \( p<0.05 \)
- \*** \( p<0.01 \)

\( (1) \ (2) \ (3) \ (4) \ (5) \)

\( T_i \) if Section Type \( = 1 \) and time = time of fund

\( P_i \) if Section Type \( = 1 \) and time = time of fund

A lower frequency of Positive words

A higher frequency of Uncertainty and Litigious and Constraining words

A higher frequency of Negative words

A higher frequency of Positive words of Negative words
Clustering

- Unsupervised learning methods ("clustering") can discover groupings of similar prospectuses
  - Use 4 different clustering algorithms:
    - DBSCAN
    - Mean-Shift
    - K-Means
    - Gaussian Mixture
Stability and Interpretability of Clusters

June 2013 PIR Cluster 9: "Derivative Risk"

June 2013 PIR Cluster 15: "Currency Risk"
Roadmap

1. Introduction
2. Data
3. Descriptive Analysis
   - Length, Complexity, Sentiment
4. Empirical Analysis
   - Clustering
   - Risk Shifting
   - Work in progress...
5. Future Research
6. Conclusion
Risk-Shifting

- A vast literature has shown that funds may suddenly increase their risk in order to obtain a short-term boost in returns
  - Related to career concerns and agency problems
  - Found to have a negative impact on performance

- Holdings-based measure from Huang, Sialm and Zhang:
  - \( RS_{i,t} = \sigma_{f,t}^H - \sigma_{f,t}^R \)
  - \( RH_{f,t} = \omega_{f,t} R_t = \sum_{i=1}^{n} \omega_{f,t}^i R_t^i \)
  - \( \text{Var}(RH_{f,t}) = \text{Var}(\omega_{f,t} R_t) = \omega_{f,t} \Sigma \omega \)
We predict future risk-shifting using the information disclosed ex-ante by funds in their prospectuses. We use “hard information” as a benchmark prediction. We split the sample into two randomly selected sub-samples of unique documents. The other used to assess performance (\textit{Testing Sample} \( N=178-3388 \))\textsuperscript{1} One used for model estimation (\textit{Training Sample} \( N=4102-7905 \)).

**Outcome variables:**
- Deciles based on the risk-shifting distribution

**Predictors:**
- All predictors matched at prior windows: (-48m, -36m, -24m, -12m)
- Hard information: AUM, fees, returns, turnover-ratio
- Soft information: tf-idf matrix of unique PIS and PIR
- text statistics (word count, complexity, dictionary frequencies)

Assess model goodness by looking at:

\[
\text{Recall} = \frac{TP}{TP + FN} \quad \text{and} \quad \text{Precision} = \frac{TP}{TP + FP}
\]

\textit{Recall} \& \textit{Precision} for \( \text{Recall} \geq \text{Recall} \).
Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
  - Train 6 machine learning algorithms using training sample & Verify model accuracy using testing sample.
Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
  
  - Train 6 machine learning algorithms using training sample & Verify model accuracy using testing sample.
  
  - Deciles prediction:

| Testing-sample predictability of deciles using information at 24 months prior window |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Model: Only hard information    | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  | precision | recall | support |
| 1                               | 51  | 30  | 11  | 9   | 11  | 8   | 14  | 10  | 13  | 14  | 0.18      | 0.3    | 171     |
| 2                               | 26  | 46  | 23  | 32  | 23  | 22  | 16  | 26  | 29  | 23  | 0.16      | 0.17   | 266     |
| 3                               | 35  | 47  | 48  | 47  | 18  | 29  | 32  | 25  | 31  | 31  | 0.19      | 0.14   | 343     |
| 4                               | 28  | 33  | 43  | 58  | 41  | 37  | 39  | 28  | 25  | 25  | 0.18      | 0.16   | 357     |
| 5                               | 25  | 34  | 37  | 52  | 44  | 39  | 31  | 42  | 26  | 29  | 0.17      | 0.12   | 359     |
| 6                               | 26  | 14  | 28  | 41  | 39  | 50  | 40  | 26  | 29  | 18  | 0.19      | 0.16   | 311     |
| 7                               | 18  | 16  | 16  | 22  | 37  | 31  | 37  | 35  | 36  | 40  | 0.14      | 0.13   | 288     |
| 8                               | 29  | 31  | 17  | 24  | 21  | 19  | 25  | 34  | 43  | 27  | 0.12      | 0.13   | 270     |
| 9                               | 26  | 22  | 11  | 21  | 9   | 17  | 20  | 33  | 33  | 26  | 0.11      | 0.15   | 218     |
| 10                              | 13  | 17  | 18  | 16  | 11  | 12  | 20  | 13  | 29  | 47  | 0.17      | 0.24   | 196     |
Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
  - Train 6 machine learning algorithms using training sample & Verify model accuracy using testing sample.
  - Deciles prediction:

![Testing-sample predictability of deciles using information at 24 months prior window](image-url)
Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
  
  - Train 6 machine learning algorithms using training sample & Verify model accuracy using testing sample.
  
  - Deciles prediction:

```
<table>
<thead>
<tr>
<th>Abis and Lines</th>
<th>Columbia University</th>
</tr>
</thead>
<tbody>
<tr>
<td>14/21</td>
<td></td>
</tr>
</tbody>
</table>
```

### Testing-sample predictability of deciles using information at 24 months prior window

<table>
<thead>
<tr>
<th>Model: All textual and hard information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
Risk-Shifting Model Selection

- We build supervised learning models to predict future risk-shifting
  - Train 6 machine learning algorithms using training sample & Verify model accuracy using testing sample.
  - Deciles prediction:

![Testing-sample predictability of deciles using information at 24 months prior window](image)

- Problem: low power in explaining middle deciles
- Solution: binary outcome var (1 for top & bottom 10%, 0 otherwise)
  - Use stratified sampling to split between training and testing
  - Down-sample majority/Up-sample minority class while training
  - Use the most accurate algorithm: Random Forest with down-sampling
Risk-Shifting Prediction Accuracy

- Testing prediction of binary risk-shifting variable
  - \( \text{Recall} = \frac{TP}{TP + FN} \) & \( \text{Precision} = \frac{TP}{TP + FP} \)

<table>
<thead>
<tr>
<th>Only Hard Information</th>
<th>Only Prospectuses</th>
<th>All textual and hard information</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1530</td>
<td>1848</td>
</tr>
<tr>
<td></td>
<td>882</td>
<td>564</td>
</tr>
<tr>
<td></td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td>0.77</td>
</tr>
<tr>
<td>1</td>
<td>136</td>
<td>108</td>
</tr>
<tr>
<td></td>
<td>231</td>
<td>259</td>
</tr>
<tr>
<td></td>
<td>0.21</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>support</td>
<td>2412</td>
<td>367</td>
</tr>
</tbody>
</table>
Risk-Shifting Prediction Accuracy

- Testing prediction of binary risk-shifting variable
  - \( \text{Recall} = \frac{TP}{TP + FN} \) & \( \text{Precision} = \frac{TP}{TP + FP} \)
Risk-Shifting Insights

Features Importance and Holdings Concentration
Beyond risk shifting, can we utilize supervised learning to predict likelihood of future lawsuits?

Collecting legal action data against mutual funds from N-SAR and ATY filings

Use cluster assignment to predict future return distribution

Is textual similarity associated with similar fund returns?
Roadmap

1 Introduction

2 Data

3 Descriptive Analysis

4 Empirical Analysis

5 Future Research

6 Conclusion

Work in progress...

Risk Shifting

Clustering

Length, Complexity, Sentiment

Descriptive Analysis
Regulatory Changes:

Relate textual measures to fund growth one year after inception.

Distinguish between institutional and retail investors' assets.

Relate signals extracted from the prospectuses to funds size/growth distribution.

Utilize the model to derive testable predictions regarding the distribution of funds size.

Solve a model of rational inattention with investors with different levels of financial literacy who learn from prospectuses.

Future Research

1995: Introduction of online EDGAR distribution system
1998 (Rule 421): Readability act
1999: Increase disclosure requirements in PIS
2004: More frequent disclosure

Abis and Lines

Columbia University 20/21
Roadmap

1 Introduction
2 Data
3 Descriptive Analysis
4 Empirical Analysis
5 Future Research
6 Conclusion

Work in progress...
Risk Shifting
Clustering
Length, Complexity, Sentiment
Conclusion

Results suggest that prospectuses contain meaningful information about:

- Funds' trustworthiness
- The shape of funds' future return distribution

It remains an open question (that we plan to answer) whether investors are paying attention to this information!