Textual Analysis Methodology

We rely on new tools provided by metaHeuristica LLC to gather and parse 10-Ks, compute LDA, and to provide tools for reviewing content. MetaHeuristica uses natural language processing to parse and enrich textual data, and its pipeline employs “Chained Context Discovery”. (See Cimiano (2010) for details.) This approach leverages an empirical ontology to normalize the content and map it to a standardized form. Data is then stored in a high speed database.

Our approach corresponds to the information analytic approach in Demski (1980). The words in the specific sections of the 10-K are the underlying data. This data is too high dimensional to be analyzed in its raw form and has to be transformed into lower dimensional signals to be useful in research.

Because the sections of the 10-K are large, for example, MD&As contain over fifty thousand distinct words, it is difficult to use the text directly to examine content and test more refined theories. Accordingly, we use Latent Dirichlet Allocation (LDA) to reduce dimensionality to a more manageable level.

We use LDA to identify the topics discussed in each of the relevant sections of the 10-K and to examine the granularity of the textual data. LDA can also reveal the economic channels through which content in the section explains variation in the co-movement of stock prices. LDA was developed by Blei, Ng and Jordan (2003) from an underlying model in which each document is assumed to be generated from a probability distribution over topics.

The text corresponding to each topic in the document is assumed to be generated using the distribution of vocabulary associated with the chosen topic. LDA algorithmically derives both a measure of how much text in each document corresponds to each topic, and the topic vocabularies for each topic (these are the two key LDA data structures, which we describe below). LDA has been used extensively in computational linguistics, is replicable, and does not require researcher prejudgment in that the researcher is not required to make assumptions about specific topics to be found in the document, or about the associated word distributions. An influential early use is by Griffiths and Steyvers (2004) who assess the content of 28,154 abstracts from the National Academy of Science.

Each LDA topic can be defined as a probability distribution over individual words. For example, the word "mortgage" should occur with a higher probability in a discussion of financing risk than in a discussion of risk management. To gain basic intuition, consider a stylized assumption that there are a fixed number of $T$ topics that firms address in their RFs. Potential topics might include firm liquidity issues, litigation, competitive pressures, profitability, and investment strategy. When discussing each topic, managers draw words from topic-specific vocabularies. Although readers of 10-Ks will expect many of these listed topics to appear, we do not specify any topics ex-ante, as they are determined algorithmically by LDA.

More formally, let $R$ be the number of Risk Factors (RF) sections in our corpus. Further assume that there are $W$ unique individual words in the union of all Rs ($W$ is likely to be around 50,000 in our case). We then can represent each firm $i$’s RF as a long $W$ dimensional vector of words $w_i = \{w_1, w_2, \ldots, w_W\}$. Taking a specific RF, the probability $P(w_i)$ that a word $w_i$ appears in that document depends both on the probabilities with which the word is used in each of the $T$ topics and on the probability that the topic appears in the given RF:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j)P(z_i = j)$$

where the variable $z_i$ is a latent variable referring to the topic, $j=1,\ldots,T$ from which the word $w_i$ is drawn. $P(z_i = j)$ is the probability that a word is drawn from topic $j$, and $P(w_i \mid z_i = j)$ is the probability that a word used in topic $j$ is the word $w_i$. In our example, if the word $w_i = \text{mortgage}$, we expect that $P(w_i \mid z_i = \text{financing risk}) > P(w_i \mid z_i = \text{risk management})$. The probability that the RF contains a discussion of financing $P(z_i = \text{financing risk})$ depends on the extent to which the particular document loads on this topic.
We now briefly summarize the foundation used by models such as LDA for identifying the \( T \) topics that appear in a corpus (see Griffiths and Steyvers (2004) for more details). Noting that there are \( W \) unique words used in the RFs, we represent \( P(w|z) \) with a set of \( T \) multinomial distributions \( \phi \) over the set of words so that \( P(w|z = j) = \phi^T_w \). We further represent \( P(z = j) \) with a set of \( R \) multinomial distributions \( \theta \), one for each RF, over the \( T \) topics, so that for a word in a RF, the probability that it comes from topic \( j \) is \( P(z = j) = \theta^T_j \). This problem of characterizing the topics that appear in a corpus is equivalent to using equation (1) to maximize \( P(w|\phi) \) by choosing appropriate \( \phi \) and \( \theta \).

LDA puts further structure on the problem by assuming that the RFs are generated by picking a distribution over topics \( \theta \), from a Dirichlet distribution, so that for each RF \( P(z) \) comes from a Dirichlet distribution, and the distribution of \( \phi \) also has a Dirichlet prior. Each RF is then generated as in equation (1) by picking topics \( j \) from the \( P(z) \), and then picking a word from that topic using \( P(w|z_i = j) \) for a given \( \phi^j \).

When extended to an entire corpus, the estimation problem is to maximize

\[
P(w|\theta, \alpha) = \int P(w|\phi, \theta)P(\theta, \alpha)d\theta
\]

where \( P(\phi) \) is a Dirichlet distribution with parameter \( \alpha \). As discussed in Griffiths and Steyvers (2004), this does not have a closed form solution but can be computed using Gibbs sampling (see Casella and George (1992) for more details).

On a more practical level, the first data structure LDA generates for the researcher describes the distribution of topics discussed in each RF. These firm-year-specific distributions are commonly referred to as "topic loadings". For LDA based on 50 topics, LDA thus generates a vector of length 50 for each RF in our sample, which scores each document on the extent to which it discusses each of the 50 topics. This data structure is a numerical reduced-dimension summary of the aggregate content of the RFs. It has reduced dimensionality because it summarizes each document using a vector of length 50, whereas raw RFs have a dimensionality likely exceeding 50,000, which is the number of unique words in the corpus of RFs.

The second data structure LDA generates is a set of word-frequency distributions for each topic. For LDA based on 50 topics, this data structure contains 50 word lists with corresponding word frequencies. Each topic is thus described by a probability distribution of individual words.

We use these two LDA-generated data structures to determine which topic loadings in the RF are related to the co-movement of stock prices. Those topics that load most significantly and positively on the covariance of returns are potential candidates for emerging risks.

LDA requires one critical input from the researcher: the number of topics \( T \) to be generated. To maintain parsimony, we focus on 25 topics (although we consider other numbers for robustness). The choice of 25 topics reflects the multi-faceted nature of RF text. As RF is known to contain some boiler plate content and some industry content, the researcher can manually inspect the topics and reduce the set to those that are most likely "economic topics".

One limitation of the LDA-generated data is that the researcher must interpret the economic theme. Because the topic structure changes from year to year and the model only produces a few key phrases or words, we are faced with the challenge of consistently labelling themes across different years. In order to do so, we have utilized a tool within Metaheuristic, the Semantic Vector that allows us to take key words from a particular LDA theme and find related words. By doing so, we are able to better establish the economic interpretation of the LDA themes and we are able to generate highly transparent themes that are also stable over time, and flexible when the researcher needs to further refine the granularity of the risk factor analysis.
References:


Demski, Joel S., 1980, Information Analysis, Addison-Wesley.