A Bloomberg Professional Services Offering

Content & Data Solutions

Event-Driven Feeds

Sentiment impact on stock prices of news with selected topic codes: Part Two

Ivailo Dimov
Quantitative Research
Bloomberg L.P.

February 2018

Bloomberg

Sentiment impact on stock prices of news with selected topic codes: Part Two

Abstract

We use Latent Semantic Analysis (LSA) with a suitable Independent Component Analysis (ICA) regularization to retrieve latent, interpretable topic code factors in Bloomberg's machine-readable news dataset. We show that although 5% of the LSA factors can explain as much as 70% of the term-document matrix variance, these factors tend to mix different topic categories, with results not readily interpretable. By optimization under ICA regularization, mixed factors can be effectively disentangled, which boosts both explanation power and thematic discoverability.

Contents

Introduction

In Part One of the study, two sets of topic codes were found to have varying sentiment impact:

- Analyst Equity key topic codes "aek": EQUITYKEY, ANA, ANAMOVES, ANACHANGE, ANATGTCHG, ANACUT, ANAHOLD, ANABUY, ANARESU, ANATGTUP, ANARAISE, ANANEW, ANATGTDWN.

  It was found that:
  - Stories tagged by codes in "aek" in general produce greater-than-average sentiment impact.
  - Stories that contain no "aek" codes produce smaller-than-average sentiment impact.
- Controversy topic codes "con": ESGCONTROV, LAW, ESGRES, LITIGATE, LAWPRAC, LAWSUITS, IP, PATENT, CLASS, CALVPOSS.

  It was found that:
  - Stories tagged by codes in "con" in general produce smaller-than-average sentiment impact.
  - Stories that contain no "con" codes produce greater-than-average sentiment impact.

These initial ad hoc findings are encouraging, suggesting that conditioning on topic codes may enhance sentiment performance in quantitative strategies. However, there are practical challenges in further expanding the study to systematically traverse the entire topic code universe. There are, in total, tens of thousands of different topic codes present in the dataset. Each topic code occurs in 5% of the stories on average. The median frequency of occurrence of all topic codes is 0.08%, or, put in another way, half of the topic codes occur in less than 40 out of a million stories.

This high-dimensional, sparse distribution of topic codes invites representing the information in vector space. After appropriate dimension reduction, latent information should be retrieved with a small number of easy-to-interpret principal vectors (factors) explaining majority of the variance.

Overview of methods & results

In this study, we retrieve topic code factors using Latent Semantic Analysis (LSA) with a suitable Independent Component Analysis (ICA) rotation. Details of the techniques are discussed in the Appendix.

Our main results are as follows:

- Both the LSA and the ICA methods build a linear factor model that allows us to decompose any set of topic codes into a linear combination of topic factors. With $K=256$ linear factors, both the LSA and ICA capture around 70% of the variance in our training dataset of $N=500{,}000$ stories from 2015 – each tagged with a set of topic codes from a total of $M=5{,}082$ topics.
- As an example, we inspected the two topic code groups reported in the previous note and found that:
  – The "aek" / analyst revisions topic group is predominantly explained by 16 LSA factors or 3 ICA factors.
  – The "con" / controversial topic group is predominantly explained by 16 LSA factors or 6 ICA factors.
- In general, we find that the ICA factors are an order of magnitude more "localized" than the LSA factors, with each ICA factor described by a much lower number of topic codes than LSA factors. This localization feature of the ICA-enhanced factors has two important implications:
  – The ICA factors are much more interpretable than the LSA factors.
  – The ICA factors occur much less frequently than LSA factors. Therefore, ICA factors tend to do a better job in thematically categorizing stories.

Here, we will discuss the main aspects of both the LSA and ICA methodology while postponing details to the Appendix. In both methods, we start with a term-document $N \times M$ matrix X with non-negative entries. The matrix X is obtained via a suitable *topic2vec* transformation, which allows us to map any given set of topics into an M-dimensional non-negative unit vector. Both methods then build a linear factor model by approximating the term-document matrix as follows:

- The LSA consists of approximating the term-document matrix by its rank-K truncated Singular Value Decomposition (SVD) $X \approx USV^T = U_{(N\times K)}\, S_{(K\times K)}\, V_{(M\times K)}{}^T$ with $U^T U = V^T V = I_K$ and $S = diag(s)$. As elaborated on in the Appendix, the orthonormal K right-SVD components V coincide with the top-K PCA factors of the covariance matrix $X^T X$, while the orthonormal K left-SVD components U correspond to the factor realizations. The variance explained by the k-th PCA factor is s_k^2. As in the PCA method, given a set of topic codes $\tau = \{t_i\}$ with a corresponding M-dimensional term-vector representation x_τ, we can decompose x_τ linearly in terms of the K right-SVD factors:

$$x_\tau = \sum_{k=1}^{K} \beta_{k,\tau} v_k + \varepsilon_\tau$$

To identify which factors have the most explanatory power for the topic set τ, we can then compute the z-score of the k-th factor to be $z_k = \frac{\beta_{k,\tau}}{s_k}$. Those z-scores significantly different from zero identify the significant factors explaining our set of topics. From now on we refer to right-SVD factors V as the *LSA factors*.

- In the ICA step, we suitably rotate the *left*-SVD factors in the previous section, thus approximating $X \approx U_I S_I V_I^T$ where $U_I = UA$ with $A^T A = I_K$. As derived in the Appendix, the diagonal matrix S_I is "variance-preserving," i.e., $tr(S^2) = tr(S_I^2)$, and the columns of V_I are unit vectors so that $diag(V_I^T V_I) = I_K$. Therefore, as in the previous point, we can interpret the columns of V_I as factors and we can write a linear model for every set of topic codes in terms of these *ICA factors*. One key difference between the ICA and the LSA factors is that the ICA factors in our transformation are not expected to be orthogonal. However, for our dataset, we find that they are nearly so. In other words, the off-diagonal entries of the ICA factor covariance $V_I^T V_I$ are small and no two ICA factors are collinear.

Why LSA factors are not easily interpretable

In our analysis we find that $K=256$ factors (i.e., the top 5% of the factors) explain about 70% of the variance. In Fig. 1 we plot the square singular values (left) and the explained variance (right).

Note that aside from the top-2 factors there is no clear separation in terms of the square singular values. In terms of the variance explained, the separation is even less. In fact, the top factor explains only 4.5% of the total variance.

In Fig. 2 we plot the top-40 components of the second LSA factor sorted by magnitude. Each component corresponds to a topic code – which is also plotted on the figure.

Topics whose component has the same sign tend to occur together in the same stories. Entries of opposite sign tend not to occur together. Therefore, the above factor seems to be composed of stories containing the following codes: US, MSCINAMER, G7MEMB, G10MEMB, MSCIWORLD, NORTHAM, ALLSTATES, PADDIST, SPREGIONS, FINNEWS, WORLD, ACEXCLUDE, DEVECO, HEADS, PADD1, SRCRANK1, BGOVBILLGO, BGOVCODES, EDGSDR, CFDOCS, FILINGS, BONDWIRES, FORM4, PADD5, USWE, USMA, IBS, CA, CMP but not the following codes SRCRANK3, SRCRANK2, CPNYCNT1, TEC, MISC, TMT, TLS, WRLS, MOBILE, HAR, INTERNET. Essentially this factor corresponds to tech/mobile single-company news of source rank 2-3 vs. several other categories such as global financial, Edgar filings, oil districts and government news.

Although there does seem to be non-trivial structure in the composition of this factor, its interpretability is challenging. For example, it is not clear how this threshold of significance should be set for every topic in the factor. In fact, had the threshold instead been 15 topics, one would've missed all the tech-related topics altogether. However, with 40 topics, there are several categories of topic codes that appear together but don't seem to have a clear relationship (i.e., Edgar filings, U.S. and macro news, and oil districts). Fundamentally, the issue is that the magnitude of the components in most LSA factors seems to decay quite smoothly, so there is no clear separation between significant and non-significant topics.

In the next section we will present a solution to this problem using ICA regularization.

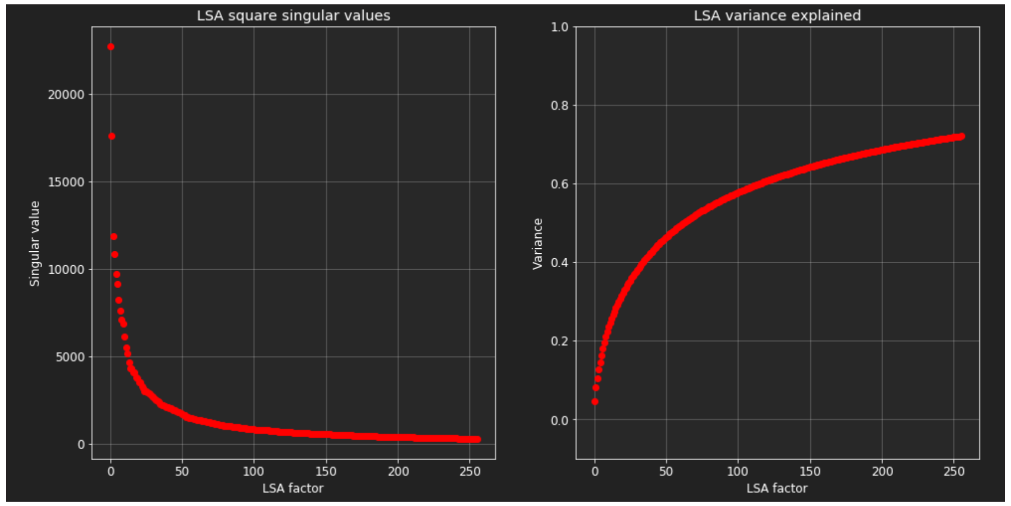

Figure 1 – LSA square singular values (left) and variance explained (right)

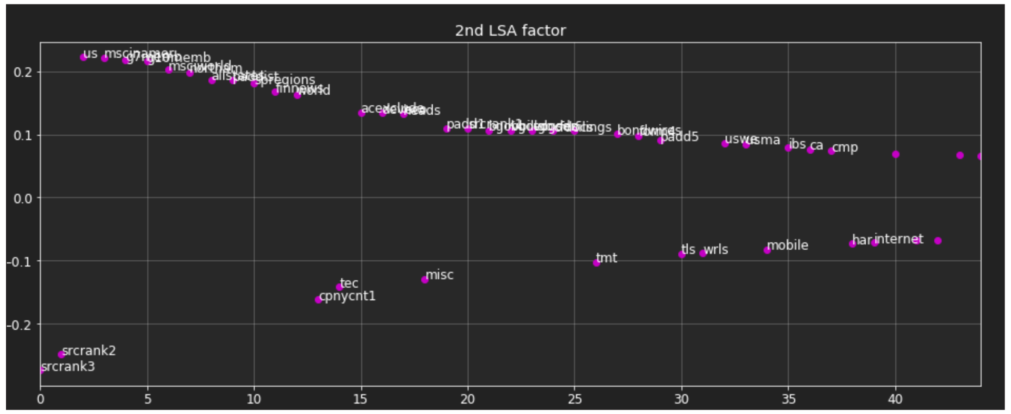

Figure 2 – The top 40 components of the second LSA factor sorted by magnitude

Applying ICA regularization to achieve better localization & interpretability

The top three ICA factors are plotted in Fig. 3. Note that most of the topics in each factor contribute nearly zero, so there are only a handful of significant topics. We find similar behavior for all ICA factors.

More precisely, in order to quantify the amount of significant components in each, we introduce the concept of *participation*. Suppose v is a vector in an N-dimensional Euclidean space and furthermore suppose v has unit norm, $|v|^2=1$. Then the participation $P(v)$ and the fractional participation $p(v)$ of v are defined as:

$$P(v) = \frac{1}{\sum_{i=1}^{N} v_i^4}, \qquad p(v) \equiv \frac{1}{N} P(v), \qquad |v|^2=1$$

For an N-dimensional unit with L non-zero components of the same magnitude (i.e., each equal to $1/\sqrt{L}$), the participation is $P(v)=L$ and the fractional participation is $p(v)=L/N$. Moreover, because

$$P(v) = \frac{1}{\sum_{i=1}^{N} v_i^4} = \frac{1}{NE[v_i^4]} = \frac{1}{N(Var(v_i^2)+E[v_i^2]^2)} = \frac{1}{N(Var(v_i^2)+1/N^2)}$$

we see that maximal participation is achieved when the variance of the square of all vector entries is zero – in other words, v is a vector whose components equal $\pm 1/\sqrt{L}$ each. The maximal participation in this case is $P(v)=N$. Therefore, $P(v)$ can be interpreted as the effective number of non-zero components of v, while $p(v)$ is the effective fraction of non-zero components of v.

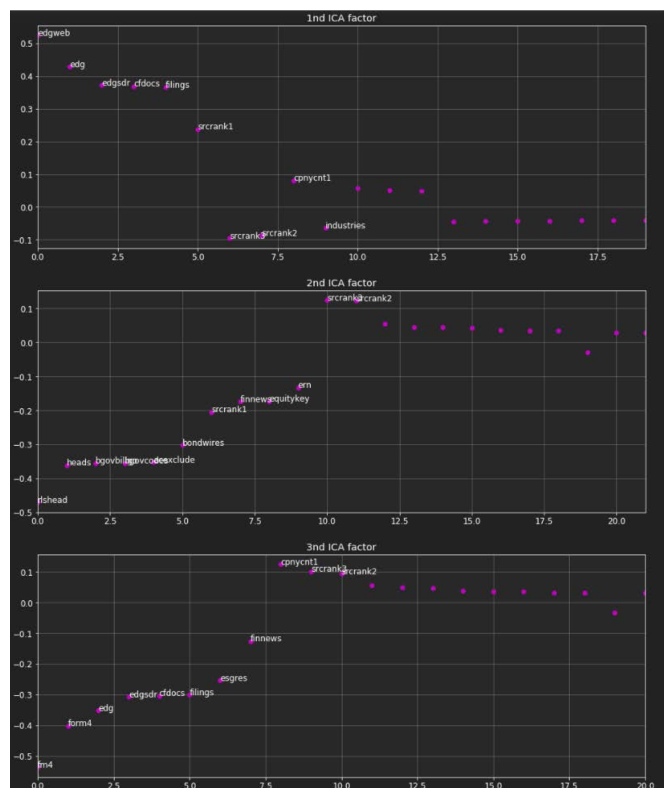

Figure 3 – The three ICA factors – each with entries sorted by magnitude

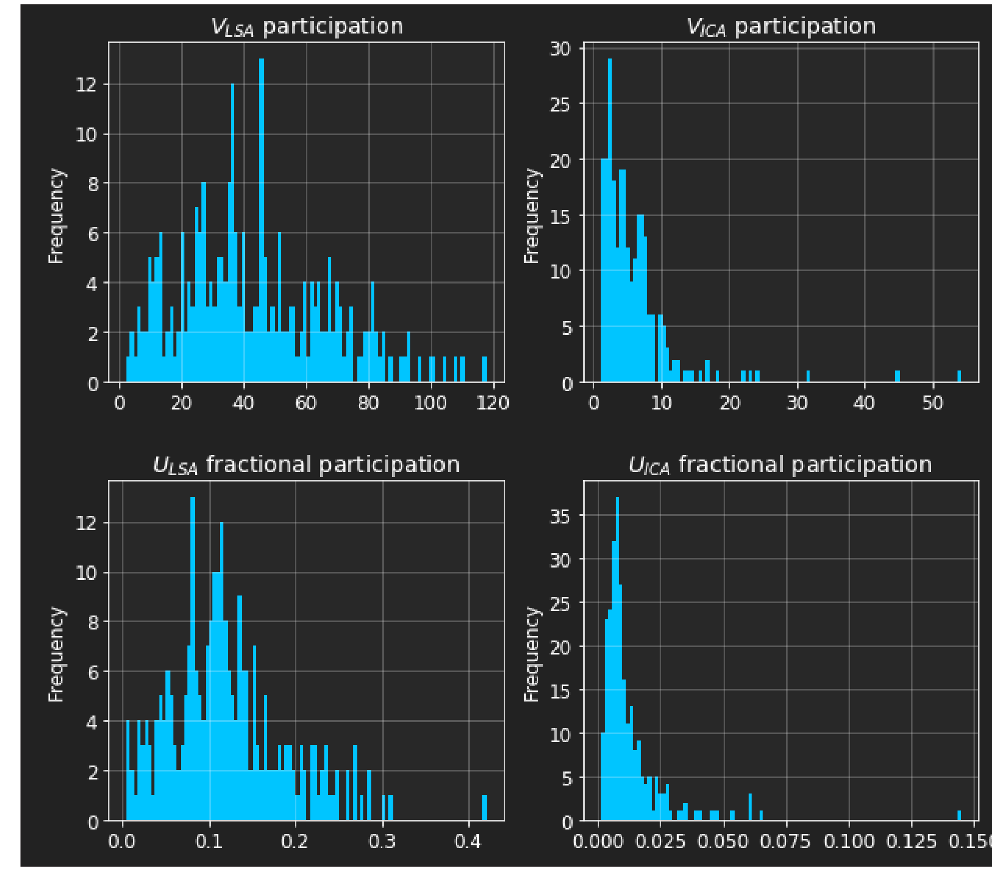

Figure 4 – (top) LSA and ICA factor participation. The mean ICA participation of 5.9 topics is much lower than the mean LSA participation of 43.5 topics. (bottom) LSA and ICA fractional participation. The mean fractional participation of 1.3% of the ICA factors is much lower than the mean fractional participation of 12.1% of the LSA factors, implying that each ICA factor is significant for only a small fraction of the stories.

In Fig. 4 we show the participation distribution of the LSA factors vs. that of the ICA factors. We also show the fractional participation distribution of the LSA factor realizations (i.e., left-SVD components) vs. that of the ICA factors realizations (i.e., left-ICA components).

It is important to note that the participation of the ICA factors is an order of magnitude lower than that of the LSA factors, which implies that:

1. Each ICA factor is composed of a much smaller effective number of topics than each LSA factor and is, therefore, much more interpretable.
2. Each ICA factor is significant for a much smaller fraction of the stories than each LSA factor

In other words, ICA factors exhibit much less mixing in explaining each story.

Examples

As an exercise, let us see what ICA factors explain a given set of topic codes. In Fig. 5, compare the LSA vs. the ICA factors explaining a story tagged by the topic code 49ERS. Clearly the ICA method is far superior in identifying that 49ERS is related to NFL topics. The method also makes clear that 49ERS stories tend to be not related to any corporate filing topics.

In Fig. 6, we compare the LSA vs. the ICA factors explaining a story that contains all the controversial topics "con." Again, the ICA method identifies far fewer factors as relevant. In fact, plotting the factor with the biggest explanatory power in the bottom of Fig. 6, we see that, indeed, most of the topic codes of that factor are related to controversial topic codes. We find similar behavior for the remaining factors.

In Fig. 7, we compare the LSA vs. the ICA factors explaining a story that contains all the analyst topics "aek." This time, three ICA factors mainly explain most of the variance of an "aek" story, while the LSA factors are nearly 20. A closer look at these three ICA factors at the bottom of the figure reveals that one of these factors corresponds to analyst rating changes (ANACHANGE), the other factor corresponds to analyst target price changes (ANATGTCHG) and the third one corresponds to analyst new ratings (ANANEW).

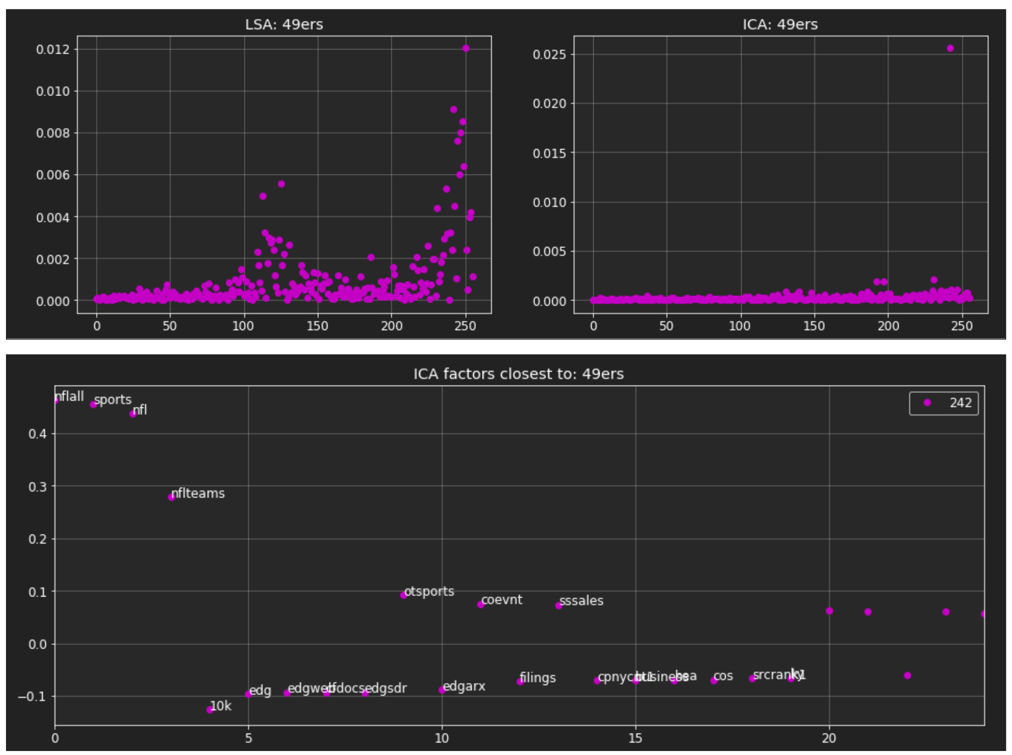

Figure 5 – (top) Factors related to topic code 49ERS. Although there are about 40 LSA factors explaining this topic code, there is predominantly one ICA factor explaining 49ERS. (bottom) ICA factor 242 mostly explaining 49ERS. The topic codes frequently occurring in this factor are all related to NFL and SPORTS but not related to any Edgar filing-related news.

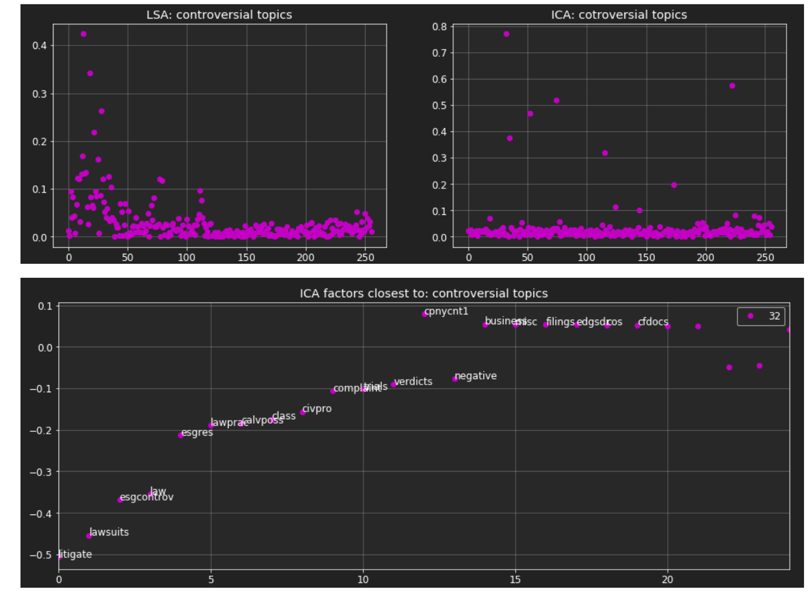

Figure 6 – (top) Factors related to the controversial topic codes in the previous section. (bottom) The ICA factor most related to the controversial topic codes.

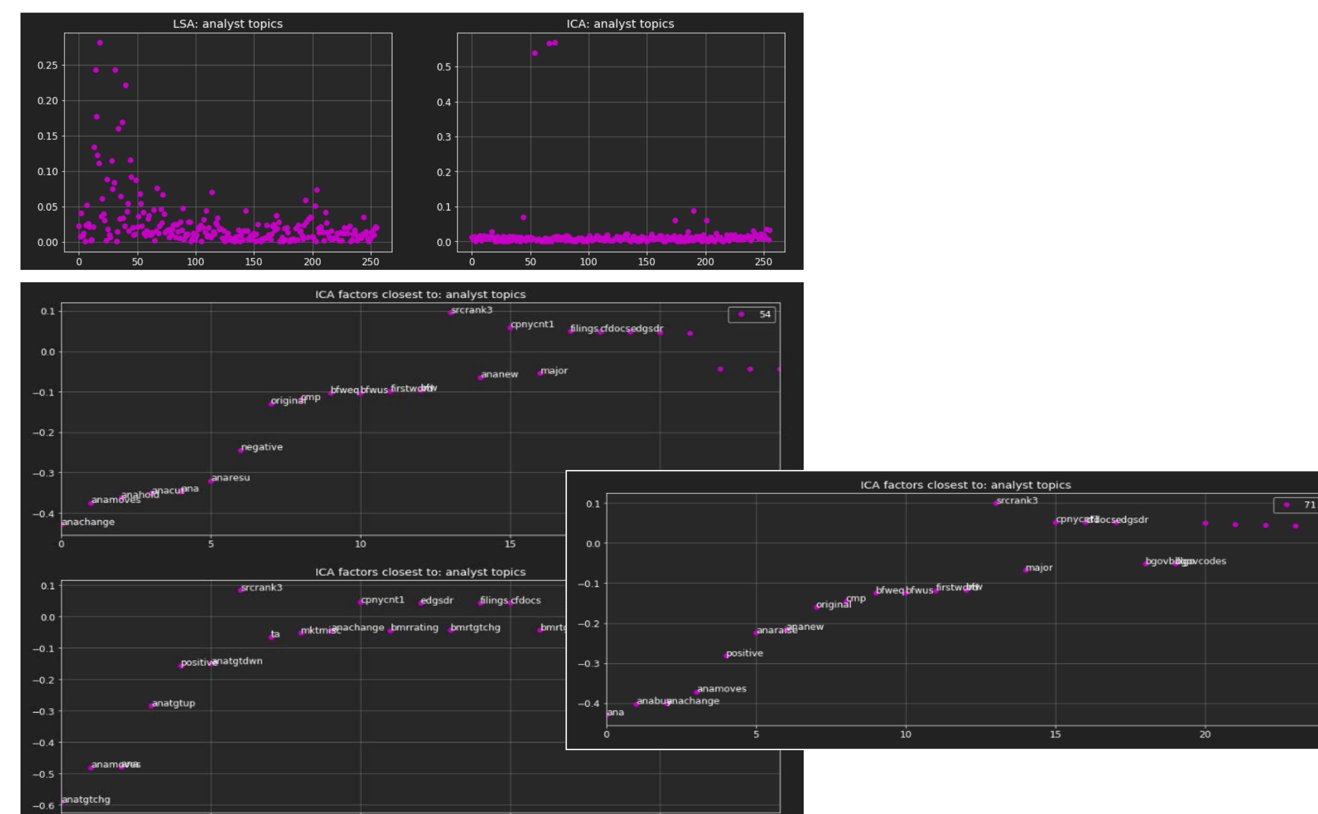

Figure 7 – (top) Factors related to the analyst-related topic codes in the previous section. (bottom 3) The top-3 ICA factors related to analyst stories, corresponding to analyst rating changes (first), analyst target revisions (second) and analyst initiating coverage/buy revisions (third plot).

Conclusions

We have found that LSA coupled with ICA regularization can be used as an effective way to retrieve localized, interpretable latent factors for the news topic codes. This paves the way for systematic study of how topic codes influence sentiment impact, a looming puzzle and foundational problem that holds the key to fully unlock the potential of sentiment-driven alpha strategies.

A following study will present the empirical results of using ICA factors to enhance sentiment-driven strategy performance.

Appendix

A.1 Latent Semantic Analysis (LSA) Model

LSA is a standard method in discovering structure in a set of text documents. Given a set of N documents, the method consists of the following steps:

1. Map each document to an M-dimensional vector using a bag-of-words approach within the TF-IDF representation. In other words, for the i-th document d_i of a corpus D of size N, $d_i \epsilon D=\{d_i\}_{i\epsilon 1...N}$, we associate an M-dimensional vector v_i, where M is the number of unique topic codes $\{t_j\}_{j\epsilon 1...M}$ so that the entire document corpus is represented by an $N \times M$ matrix $X=X_{ij}=v_i(j)$. The ij-th entry of this matrix is then defined by $X_{ij}=tf(t_j,d_i) \times idf(t_j)$ where the binary *term-frequency* $tf(t_j,d_i)=1$ if t_j occurs in d_i and 0 otherwise and the *inverse-domain-frequency* is defined $idf(t_j)=\log\frac{1+N}{1+n_j}+1$ with n_j the number of documents in which the term t_j appears. Note that the IDF part of the transformation above ensures that frequently occurring topic codes have a lower contribution to the vector representation of each document – otherwise they will dominate the top components of the SVD factors. That way the less generic but often informative topic codes will have an enhanced contribution to the TF-IDF matrix X and hence its top factors.

2. Perform the truncated Singular Value Decomposition of the TF-IDF matrix so as to let $X=X_{N\times M} \approx U_{N\times K} S_{K\times K} V_{M\times K}{}^T$. Here $K \ll min(N,M)$; in our setup we have N=500,000, M=5,082 and K=256. Note that the K columns of both U and V form an orthonormal basis so that $U^T U=V^T V=I_K$. The matrix S is diagonal, $S=diag(s)$. In order to achieve reasonable speeds of the SVD calculation, we use a rank-K SVD random projection algorithm as in Halko et al. 2011.

3. Note that the left-singular components, i.e., the K-columns of $U_{N\times K}$, are essentially the top-K principal components of XX^T, while the right-singular components, i.e., the K-columns of $V_{M\times K}$, are the top-K principal components of $X^T X$. The sum of the square singular values, i.e., $tr(S^2)$, equals the total variance explained by the top-K principal components. As in Section 3, we will refer to the right-SVD components as the LSA factors and to the left-SVD components as the LSA factor realizations. For a given story x_i corresponding to the i-th row of X, the i-th row of U is a length-K vector whose k-th entry can be interpreted as the k-th factor realization for this story. The k-th column of V will then be interpreted as the composition of the k-th factor.

A.2 Building the Independent Component Analysis Model

As we suggested in Section 3, the LSA factors suffer from lack of interpretability as they tend to mix categories of topics that are often not related. This, however, is not unexpected. Even if each story were, indeed, described by a linear model of a set of independent topic factors, LSA would not be able to distinguish those factors that have a similar variance contribution. To see this, note that rank-K SVD is the optimal rank-K approximation in the Frobenius norm [Golub & Van Loan]:

$$U,S,V=\underset{\substack{U,U^TU=I_K\\ V,V^TV=I_K\\ S-diag}}{argmin}|X-USV^T|_F^2$$

If all factors had the same variance contribution, then S would be proportional to the identity matrix. In this case the U and V are not unique and can be rotated by an arbitrary orthogonal matrix A (e.g., $A^T A=I_K$) so that $\tilde{U}=UA$ and $\tilde{V}=VA$ are also a solution. More generally, if S contains a subspace of Q *degenerate* eigenvalues (e.g., Q factors with the same variance contribution), then the solution to the above optimization problems are defined up to a rotation within the Q – dimensional degenerate subspace. In practice, even if the variance contribution between factors is not the same, any idiosyncratic noise or finite-sample measurement error would make factors close to each other indistinguishable. Therefore, LSA factors would tend to mix topic categories that occur with roughly the same frequency.

One way to disentangle mixed factors measured by LSA is to regularize the above optimization procedure so as to pick a preferred rotation even when factors are nearly degenerate. ICA provides precisely such regularization procedure. Given the K-truncated SVD decomposition $\approx USV^T$, our methodology produces an ICA-enhanced decomposition $X \approx U_I S_I V_I^T$ as follows:

1. Perform an ICA transformation on U. In other words, find a K-dimensional orthogonal matrix A, $A^T A=I$ such that the absolute excess kurtosis of $\tilde{U}=UA^T \equiv \{\tilde{u}_i\}_{i=1,...,K}$ is maximized:

$$A=\underset{A,A^TA=I_K}{argmin}|kurt(\tilde{U})|=\underset{A,A^TA=I_K}{argmin}\left|\frac{E[\tilde{u}^4]}{E[\tilde{u}^2]}-3\right|=\underset{A,A^TA=I_K}{argmin}|E[\tilde{u}^4]-3|$$

In arriving at the last equality, we have used the fact that $\tilde{U}$ is unitary so that $tr(\tilde{U}^T\tilde{U})=KE[\tilde{u}^2]=K=const.$ Since the excess kurtosis of a Gaussian distribution is zero, the above optimization maximizes the amount of non-Gaussianity of $\tilde{U}$ (see, for example, Hyvärinen et al., Ch 8). If $\tilde{U}$ is fat-tailed so that $kurt(\tilde{U})>0$, we can drop the absolute value from the above optimization and rewrite it as:

$$A=\underset{A,A^TA=I_K}{argmin}E[\tilde{u}^4]=argmin\sum_{k=1}^{K}1/P(\tilde{u}_k)$$

It is clear from the above reformulation that for fat-tailed distributions maximizing kurtosis of $\tilde{U}$ is equivalent to minimizing the participation ratio of the resulting factors. We, indeed, find that the excess kurtosis of the LSA factors is predominantly positive, which explains why we observe a much lower participation ratio for the ICA components as in Fig. 6.

2. Let $U_I=\tilde{U}$ and compute S_I, and V_I as a function of S, V and A. To do this, we start from the truncated SVD decomposition of X and rewrite it as a function of the ICA rotation as follows:

$$X \approx USV^T=U_I ASV^T=U_I SS^{-1}ASV^T \equiv U_I S\tilde{V}^T$$

where $\tilde{V}^T \equiv S^{-1}ASV^T$. Note that $\tilde{V}$ is not a unitary matrix whose columns are not even unit vectors, e.g., $diag(\tilde{V}^T\tilde{V}) \neq I_K$. However, it is easy to see that $S^2\tilde{V}^T\tilde{V}$ has the same trace as S^2 $tr(S^2\tilde{V}^T\tilde{V})=tr(S^2S^{-1}AS^2A^TS^{-1})=tr(S^2A^TA)=tr(S^2)$ which means that renormalizing the columns of $\tilde{V}$ to have unit length is a variance-preserving operation. By letting $N \equiv diag(\tilde{V}^T\tilde{V}^{1/2}$, $S_I \equiv SN$ *and* $V_I^T \equiv N^{-1}\tilde{V}^T$ we therefore have:

$$U_I S\tilde{V}^T=U_I SNN^{-1}\tilde{V}^T \equiv U_I S_I V_I^T,\ tr(S_I^2)=tr(S^2)$$

References

Golub, G., Van Loan, C. Matrix Computations. John Hopkins University Press 2012.

Halko, N., Martinsson, P. G. and Tropp, J. A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review* 53(2), 217-288. Python implementation of the algorithm can be found in sklearn.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent Component Analysis*. Wiley 2001.

Take the next step.

For additional information, press the <HELP> key twice on the Bloomberg Terminal®.

Beijing
+86 10 6649 7500

Dubai
+971 4 364 1000

Frankfurt
+49 69 9204 1210

Hong Kong
+852 2977 6000

London
+44 20 7330 7500

Mumbai
+91 22 6120 3600

New York
+1 212 318 2000

San Francisco
+1 415 912 2960

São Paulo
+55 11 2395 9000

Singapore
+65 6212 1000

Sydney
+61 2 9777 8600

Tokyo
+81 3 3201 8900

bloomberg.com/professional