Tensor Factors to Monitor the Co-Movement of Equity Prices

Louiqa Raschid
University of Maryland
louiqa@umiacs.umd.edu

Joe Langsam
University of Maryland
joe@langsam.org

Tharindu Pieris
Lanka Software Foundation
tharindup@opensource.lk

Anushka Bandara
Lanka Software Foundation
anushkab@opensource.lk

ABSTRACT

We identify a set of features that are related to extremes of price changes of individual equities. Our hypothesis is that these extreme features may be used to isolate co-movements of prices for groups of equities, reflecting systematic risk. The equities are classified within industry sectors and we create a three mode tensor to represent the dataset; the dimensions of the three mode tensor correspond to the equity, the industry sector and the day on which the feature occurred. We use a method for non-negative tensor factorization (NOTF) to identify factors or communities that are composed of multiple equities, and / or industry sectors. Our preliminary results indicate that our NOTF approach has the potential to identify such communities of price related features that may experience co-movement across industry sectors and temporal intervals.

Keywords

Price based features; non-negative tensor factorization; systematic risk.

1. INTRODUCTION

Systematic risk research has an overarching goal to identify features that may be used to isolate co-movements of prices for groups of equities. In this project, we identify features associated with extremes of price change of individual equities that may be potential predictors of systematic risk. The features include a (locally) significant maximum price (LMx), a minimum price (LMn) and an (approximate) maximal negative change of price (AMNeg). We focus on a dataset of equities in the S&P 500 Index from 2004 to 2015. Each of the equities is described by its identifier or PERMNO [3] and NAICS industry sector identifier [7].

Our analytical methodology is inspired by tensor factorization of a three mode *occurrence* tensor for each of the time-series, e.g., LMx (Day, PERMNO, NAICS). Each non-

zero entry of an occurrence tensor will represent the occurrence of the specific extrema feature, e.g., a maximum price, LMx (or LMn or AMNeg), for that PERMNO and NAICS code, on that Day. Our expectation is that maxima (or minima or negative slope) for equities within one (or multiple) NAICS sectors may experience co-movement, if these sectors experience or react to some shared systemic stimulus. We may further observe a temporal dimension to the co-movement effect, if the stimulus is very strong.

The existence of co-movement within a NAICS sector, and/or temporal interval, will allow us to decompose the observed occurrence tensor into a low rank tensor. We apply a non-negative tensor factorization (NOTF) method that has been tuned to obtain R non-negative factors, F_r. Each factor F_r represents a community of individual equities (represented by their features), clustering around one or more NAICS sectors or Day(s). The NOTF method will attempt to minimize the occurrence errors in the initial observational tensor representing the LMx (LMn, AMNeg) time-series dataset [9].

Our preliminary results indicate that our NOTF approach has the potential to identify communities of price related features that may reflect co-movement across industry sectors and temporal intervals. Most factors or communities are dominated by a single NAICS sector or Day. As expected, the larger NAICS sectors are dominant over multiple factors. More of the factors for minimal prices (LMn) display temporal co-movement and co-movement across a diversity of NAICS sectors, in comparison to maximal prices (LMx). Further examination of LMn factors identify more complex communities where both a NAICS sector and a Day are dominant. A review of the overlap of multiple LMn factors also identifies more interesting cases with co-movement of features across two NAICS sectors. With a (sampled uniform) representation of equities across all NAICS sectors, we observe price co-movement across equities in almost all sectors.

We note that these preliminary results may be equivalent to results that could have been obtained applying matrix factorization or principle components to a matrix representation of these features. The benefit of using a tensor representation is that it allows us to analyze multi-dimensional tensors. For example, we may consider a four mode tensor, where the fourth dimension identifies the specific extrema feature. This would allow us to study co-movements across multiple features, e.g., AMNeg and LMn. Such challenges will be considered in future work.

	LMx	LMn	AMNeg
2004	5206	4819	3866
2005	4846	5345	4375
2006	5079	4846	4002
2007	4870	4579	3840
2008	4407	4767	3948
2009	4834	4732	4005
2010	4116	4260	3531
2011	4622	4815	4042
2012	4097	4280	3422
2013	4348	4219	3506
2014	4284	4212	3483
2015	4210	4440	3664

Table 1: Count of features (LMx, LMn and AMNeg) for 2004 through 2015.

2. DATASET AND FEATURES

We consider the end-of-day prices of individual equities that comprise the S&P 500 Index [3]. We use the time period January 1 2004 December 31, 2015. The value for end-of-day price (PRC) can be discontinuous when there is a split of the equity, etc. To overcome that limitation, we compute a continuous PseudoPrice value. Consider RET: Holding Period Return; it is the change in the total value of investment in an equity over some period of time (days) per unit dollar of initial investment. RET(d) is computed based on a purchase at the most recent time previous to d when the equity had a valid price. Usually, (d-1) is the previous day. PseudoPrice is defined as follows: Given a return value RET_d on some day d, and an initial price P_{d-1}, a pseudo price P_d for day d is a linear interpolate $= (1+RET_d)*P_{d-1}$. We use the price P_0 from January 1 2004 to initialize the computation.

We apply Kalman filtering to smooth the PsuedoPrice time-series [8]. We then extract the following features: 1. A (locally) significant maximum price (LMx). 2. A locally significant minimum price (LMn). 3. An (approximate) maximal negative change of price (AMNeg). We parameterize these features using a time interval within which the PseudoPrice is a maximum or a minimum, or where the rate of change is most negative. We report on results for a time interval of 20 days, and where the maximal and minimal values were more than twice the standard deviation from the mean; the mean for the entire dataset and for the current year were considered and did not show much difference for the outcome of NOTF. We note that the Kalman filter implementation [8] assumed that the observation time-series was stationary. This assumption, and the sensitivity of the NOTF method to the chosen time interval may impact NOTF results.

Features can be viewed via the Karsha PseudoPrice Visual Analytics server [6].

Table 1 reports on the annual count of features. We note that one pattern is to observe the following sequence in the PseudoPrice: (LMx → AMNeg → LMn). However, there are some cases where we observe the following sequences of consecutive minima or extrema features: (LMx → LMx) or (LMn → LMn). Consequently, the count of LMx and LMn are not always balanced over each annual interval. We further note that on average there appears to be approximately one LMx and one LMn per equity per month.

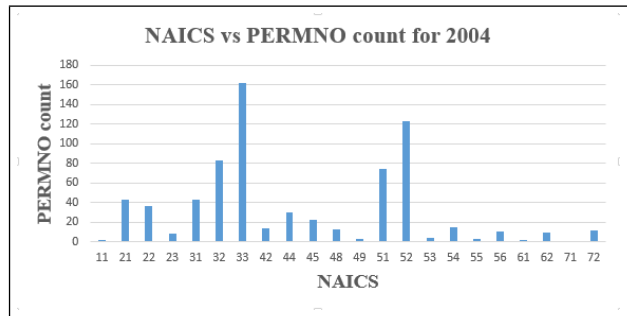Figure 1 is a histogram (count of PERMNO) for each of

Figure 1: Histogram of equities (count of PERMNO) for each NAICS industry sector.

the 24 NAICS industry sectors. We observe that Manufacturing (33) and Finance and Insurance (52) are dominant. This will also be reflected in our results where these two NAICS sectors dominate multiple factors or communities.

3. METHODOLOGY AND RELATED WORK

Consider a time-series of local maxima (LMx), or minima (LMn), or (approximately) maximal negative slope ($AMNeg$), for each equity (identified by PERMNO), and its NAICS code. The aggregate time-series over all equities, for example LMx, can be represented by a three mode observational *occurrence* tensor $LMx(PERMNO, NAICS, Day)$. Characteristics of this occurrence tensor include that it has positive values, i.e., an occurrence value of [0|1]. Given an entry of value 1 in this tensor, the indices for the three modes, Day, PERMNO and NAICS, will correspond to the equity and NAICS sector that has an occurrence of the extrema feature (LMx, LMn or AMNeg) on that day. We note that the dimensions of this occurrence tensor are not independent, i.e., each equity is assigned to a single NAICS sector. A sensitivity analysis could determine if this dependency has an impact on the results of our NOTF tensor factorization.

Our expectation is that maxima (or minima or negative slope) for equities within one (or multiple) NAICS sectors may experience a co-movement effect, if these sectors experience or react to some shared systemic stimulus. We may further observe a temporal dimension to the co-movement effect, if the stimulus is very strong. The existence of co-movement within a NAICS sector, and/or temporal interval, will allow us to decompose the observed occurrence tensor into a low rank tensor using our NOTF solution. Each factor would represent a community of features (LMx or LMn or AMNeg) of individual equities, within some NAICS sector(s), and occurring on some Day(s).

Vectors and matrices are denoted by lowercase and capital letters, respectively. Higher mode tensors are denoted by Euler script letters. Consider a three mode tensor $X \in \mathbb{R}^{N_1 \times N_2 \times N_3}$. Denote the vector outer product by $\circ$. The CANDECOMP/PARAFAC(CP) decomposition [2, 4] factorizes a tensor into a sum of R rank-one tensors as $X = [A, B, C] = \sum_{r=1}^{R} a_r \circ b_r \circ c_r$. This decomposition was further extended with an efficient algorithm for non-negative occurrence tensor factorization (NOTF) [9]. Each rank-one tensor $F_r = a_r \circ b_r \circ c_r$ captures a community of equities, within one or multiple NAICS sectors, on specific days, and $a_r \geq 0, b_r \geq 0, c_r \geq 0$ represent the weights for PERMNO, NAICS and Day, respectively, in the community. We refer

	LMx	LMn
Top 5 NAICS		
Manufacturing (33)	1243	1166
Finance and Insurance (52)	905	831
Manufacturing (32)	594	553
Information (51)	497	488
Mining (21)	315	
Manufacturing (31)		275
Top 3 Days		
Oct 06 2004	174	
Mar 05 2004	137	
Jun 30 2004	136	
Mar 24 2004		175
May 10 2004		148
May 17 2004		146

Table 2: The Top 5 NAICS sectors and Top 3 Days based on the count of features (LMx and LMn) for 2004.

the reader to [9] for a detailed discussion of the NOTF solution. NOTF was implemented in Matlab using the Tensor toolbox [1].

4. EVALUATION RESULTS

In this section, we present preliminary observations of the communities obtained from applying NOTF to the LMx and LMn occurrence tensors. We first summarize factor characteristics, for the two features, for a baseline year 2004. Next we explore some potentially interesting individual factors. We then consider a crisis year and compare to the baseline 2004. We conclude this section with a discussion on the prediction errors obtained from the NOTF solution. Factors can be viewed via the Karsha NOTF Visual Analytics server [5].

4.1 Summary of Factors for Baseline 2004

We report on a summary of 20 factors for the baseline year 2004. While we did not complete a detailed sensitivity analysis, we note that as we increase the count of factors, the characteristics of the factors are conserved, i.e., factors that appear with a choice of 15 are conserved if we increase the count of factors to 20 or 25.

Table 2 reports on the Top 5 NAICS sectors and Top 3 Days for the feature dataset (maximal count) for 2004. We expect that these Top 5 NAICS sectors and Days will have a dominant effect on the factors for 2004. Thus, our summary of factors will try to identify additional NAICS sectors and Days that also play an important role in the 2004 factors.

The NAICS sectors associated with the 20 LMx and LMn communities for 2004 are shown in Figures 2 and 3, respectively. The Y axis identifies the community and the X axis corresponds to the NAICS sector. Our first observation is that each factor is typically dominated by a single NAICS sector. This confirms that there is indeed price co-movement within equities in the same industry sector. As expected, the Top NAICS sectors based on feature count dominate. We observe that Manufacturing (33) is dominant in 7 LMx communities, 5 LMn communities and 10 AMNeg communities. Finance and Insurance (52) is dominant in 4 LMx, 2 LMn and 5 AMNeg communities.
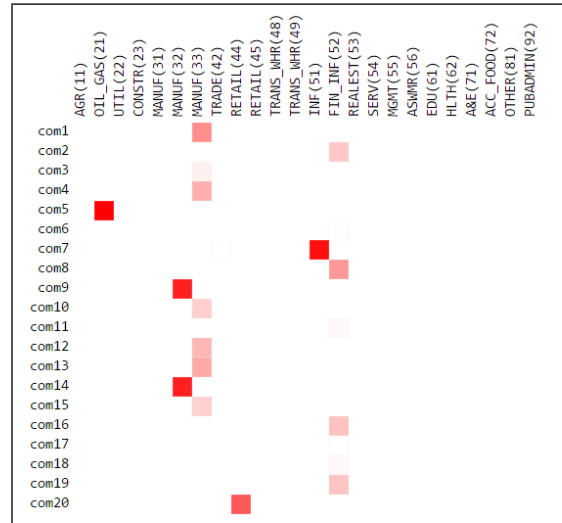

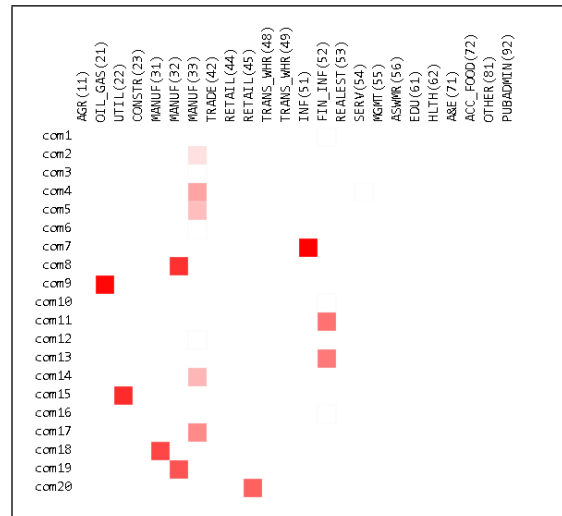**Figure 2: NAICS sectors associated with 20 LMx communities for 2004.**

Figure 3: NAICS sectors associated with 20 LMn communities for 2004.

Several additional NAICS sectors are represented across LMn and AMNeg communities; Oil and Gas (21), Utilities (22), Manufacturing (31), Manufacturing (32), Retail (45) and Information (51) each dominate one LMn community. Similar sectors including Retail (44) are present in AMNeg communities There is less diversity across LMx communities; Oil and Gas (21), Manufacturing (32), Retail (45) and Information (51) are present. The key observation appears to be that equities within multiple NAICS sectors exhibit co-movement with respect to minimal prices and maximal negative change of prices; hence we see multiple NAICS sectors dominate LMn and AMNeg communities. There appears to be less co-movement across these NAICS sectors for maximal prices. While we expect the large NAICS sectors, Manufacturing (33) and Finance and Insurance (52) to dominate more communities, the dominance of Information (51) over 3 AMNeg communities is notable.

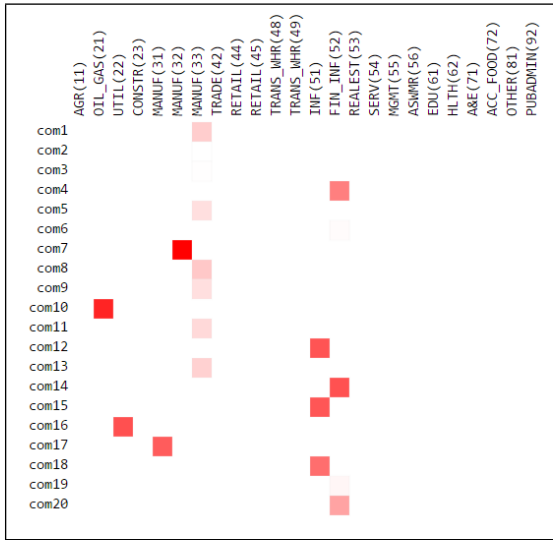Table 3 presents the Day(s) and the corresponding weight

Figure 4: NAICS sectors associated with 20 AMNeg communities for 2004.

Factor	Day	weight	NAICS	weight
LMx F_7	72 days	low		
			Information (51)	8.3076
LMn F_4	Feb 04	16649		
			Manufact (33)	2.9609

Figure 5: Communities associated with multiple NAICS sectors and Days in 2004.

(importance of the Day) for the 20 LMx and LMn factors. The differences are fairly striking. Only four LMx communities are dominated by Day(s) and there are eight significant Day(s) in these communities. In contrast, there are twelve LMn communities dominated by twenty five Day(s). Further, four of these communities, F_3, F_6, F_{12} and F_{17}, have very high weights associated with individual Day(s). The key observation, which is not unduly surprising, is that minimal prices (LMn) exhibit more co-movement within a temporal interval in comparison to maximal prices (LMx).

4.2 Isolating Interesting Factors

A majority of the factors are dominated either by a single NAICS sector or a community of one (or more) Day(s). However, we did isolate some factors corresponding to a *complex* community comprising a dominant NAICS sector and one or more Days. Figure 5 illustrates LMx factor F_7 where Information (51) is dominant. This factor occurs over a large number of Days and is consistent with our previous observation that maximal prices (LMx factors) exhibit less co-movement within a temporal interval. In contrast, LMn factor F_4 occurs on a single day, February 04 2004; Manufacturing (33) is dominant. We note that February 04 was not one of the Top 3 days for LMn features; hence a simple statistical summary would not have isolated this factor.

Figure 6 illustrates a community that is created through the overlap of two factors. Factor F_5 is associated with Manufacturing (33) and spans two days. Factor F_{13} is associated with Finance and Insurance (52) and spans four days. August 06 2004 is common to both factors resulting in a

LMx Community	Day	weight
F_6	Jan 26	248.34
	Mar 5	248.23
	Apr 1	248.29
	Nov 12	248.17
F_{11}	Apr 1	237.93
	Nov 15	237.96
F_{17}	Mar 5	1011.5
F_{18}	Sep 7	429.55

LMn Community	Day	weight
F_1	Feb 24	496.97
	Mar 24	496.97
	May 10	496.97
	Aug 9	496.97
	Oct 25	496.97
F_2	Sep 27	49343
F_3	Aug 12	21258000
	Oct 25	21259000
	Dec 7	21259000
F_4	Feb 4	16649
F_5	Mar 11	3062.1
	Aug 6	3062.1
F_6	May 17	349220000
	Aug 12	349130000
F_{10}	Mar 24	98010
	Aug 13	98010
F_{11}	Apr 20	18718
	Jun 14	18718
	Sep 23	18718
F_{12}	Mar 23	7542100
	Aug 12	7544200
F_{14}	Aug 13	51435
F_{16}	May 10	3307.7
F_{17}	Mar 24	5984600
	May 10	5984600

Table 3: Days associated with 20 LMx and LMn communities for 2004.

complex community reflecting price co-movement across 2 NAICS sectors within a temporal interval. As before, we note that August 06 was not one of the Top 3 days for LMn features and a simple statistical summary would not have isolated this factor.

Isolating additional interesting factors involving complex communities of multiple NAICS sectors and Days requires extensions to our NOTF methodology and fine-tuning our experiment protocol. One issue is that two NAICS sectors, Manufacturing (33) and Finance and Insurance (52), dominate the equities in the dataset, and hence they dominate the NOTF factors. As a first step to fine-tuning the experiment protocol, and reducing the dominant influence of these two sectors, we (uniformly sampled) ten equities from each of the twenty four NAICS sectors. We then applied our NOTF solution to yield 20 factors.

Figure 7 illustrates the NAICS sectors associated with twenty LMx factors (post sampling) for 2004. The contrast with Figure 2 (without sampling) is remarkable. We observe that there is significant price co-movement across all NAICS sectors and sixteen NAICS sectors are associated with these

Factor	Day	weight	NAICS	weight
LMn F_5	Mar 11	3062		
	Aug 06	3062		
			Manufacturing (33)	2.1
LMn F_{13}	Feb 05	22		
	May 17	22		
	Aug 06	22		
	Sep 27	22		
			Finance&Ins (52)	4.3

Figure 6: Communities across multiple factors and associated with multiple NAICS sectors and Days in 2004.

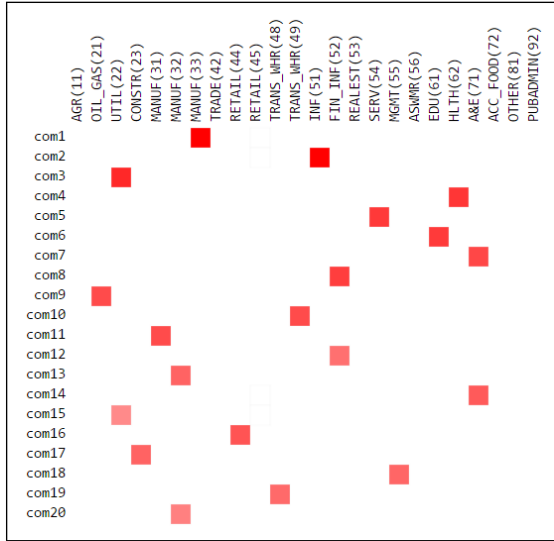**Figure 7: Impact of sampling: NAICS sectors associated with 20 LMx communities for 2004.**

Figure 8: NAICS sectors associated with 20 LMn communities for 2008.

Factor	Day	weight	NAICS	weight
LMx F_1	Feb 01	89		
	Nov 04	89		
			Finance & Ins (52)	5.4
LMx F_2	Feb 01	7		
			Manufacturing (33)	5
LMx F_3	Nov 04	5.4		
			Manufacturing (33)	3.8

Figure 9: Communities across three LMx factors and associated with two NAICS sectors and two days in 2008.

twenty factors. Very small NAICS sectors including Construction (23), Transportation and Warehousing (48) and (49), Services (54), Management (55), Education (61) and Health (62), are each associated with a factor. Arts and Entertainment (71) is associated with two factors.

4.3 Factors from a Crisis Year

The NAICS sectors associated with the 20 LMn communities for the crisis year 2008 are shown in Figure 8. We note that there are 10 NAICS sectors associated with LMx factors and 12 NAICS sectors (shown in the figure) associated with LMn sectors, in 2004. This is in contrast to the 6 NAICS sectors associated with LMx factors, and the 8 NAICS sectors associated with LMn sectors, in 2004. We note that Trade (42), Retail (44) and (45), and Services (54) are the smaller sectors that show co-movement of LMn factors. This observation of co-movement across a diversity of sectors seems consistent with a crisis year.

Thirteen LMx factors and eleven LMn factors show temporal co-movement, and are associated with one to four days. The details are not shown due to a lack of space. We note that in 2004, only four LMx factors showed temporal co-movement. This too appears to be consistent with a crisis year. The weights associated with the days for factors in 2008 appear to be much lower, in comparison to the weights
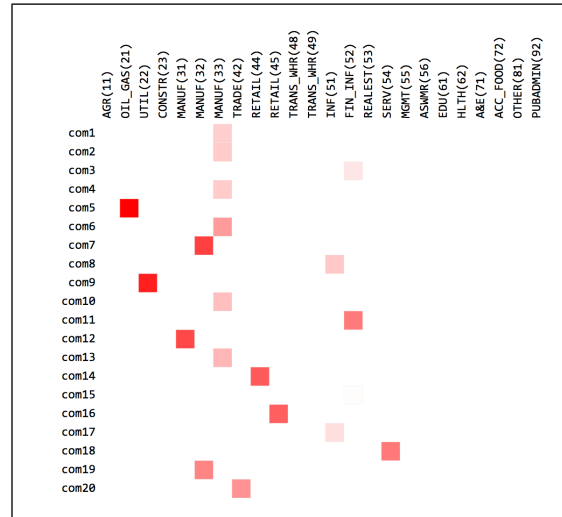
in Table 3 from 2004, and requires further study.

Finally, we end this brief discussion of crisis year 2008 with Figures 9 and 10. Three LMx factors combine to form a complex co-movement community over two NAICS sectors and across two days in Figure 9. Similarly, four LMn factors combine to form a complex co-movement community over three NAICS sectors and across three days in Figure 10.

4.4 Prediction Error

Consider some observational tensor O, corresponding to LMx, LMn or AMNeg, for some year. Consider the reconstructed tensor X obtained from the application of NOTF to O with R factors; $X = \sum_{r=1}^{R} a_r \circ b_r \circ c_r$. The prediction error is a measure of how well X can predict some element of O.

To determine the prediction error, for each pair (Day, PERMNO), we compare the tensor element from O and X. 1. If both O and X are 0, or if O is 1 and X is a non-zero value, greater than some threshold value, then we determine that there is no prediction error. 2. If O is 0 and X is a non-zero value, greater than some threshold value, Or if O is 1 and X is 0, then this is counted as an error for the corresponding Day index.

Figure 11 represents the error in predicting the LMx time-series for 2004, using $R = 20$ factors. Recall that there were 5206 observations in the LMx time-series for 252 days in

Factor	Day	weight	NAICS	weight
LMn F_3	Mar 10	5.7		
	Oct 27	5.7		
	Nov 20	5.7		
			Finance & Ins (52)	1.0
LMx F_7	Mar 10	10.5		
	Nov 20	10.5		
			Manufacturing (32)	8
LMx F_{10}	Oct 27	149		
			Manufacturing (33)	2.7
LMx F_{13}	Mar 10	6		
			Manufacturing (33)	3.1

Figure 10: Communities across four LMn factors and associated with three NAICS sectors and three days in 2008.
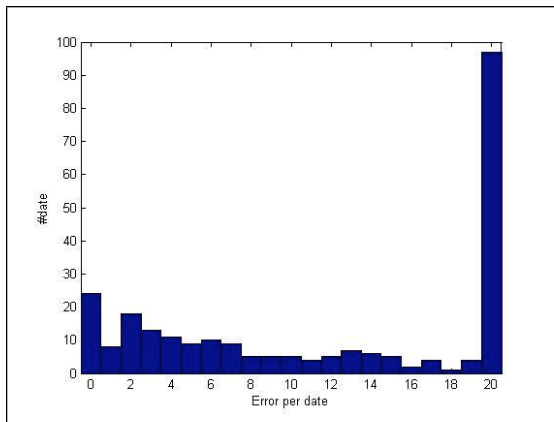

Figure 11: NAICS sectors associated with 20 LMx communities for 2004.

2004. We observe that 60+ days have 0 to 3 prediction errors. 92 Days have from 4 to 19 errors. Approximately 90+ Days have 20+ errors.

We note that our preliminary experiments had an exploratory focus and we were not trying to fine-tune our method to reduce prediction error. As a simple example, a majority of errors are false positive errors. The prediction does not appear to be very sensitive to the chosen threshold; this requires further study.

5. CONCLUSION AND FUTURE WORK

We identified features associated with extremes of price change of individual equities including a (locally) significant maximum price (LMx), a minimum price (LMn) and an (approximate) maximal negative change of price (AM-Neg). We focus on a dataset of equities in the S&P 500 Index from 2004 to 2015; each is described by its identifier or PERMNO and NAICS industry sector identifier. The dimensions of a corresponding three mode tensor correspond to the PERMNO, the NAICS code and the day on which the feature occurred. We use a method for non-negative tensor factorization (NOTF) to identify factors or communities that are composed of multiple equities, and / or industry sectors and/or temporal intervals.

Our preliminary results indicate that our NOTF approach

has the potential to identify such communities of price related features that may experience co-movement across NAICS industry sectors and temporal intervals. Many communities are dominated by a single NAICS sector or a temporal interval. More of the factors for minimal prices (LMn) display temporal co-movement and co-movement across a diversity of NAICS sectors, in comparison to maximal prices (LMx). Further examination of factors identify more complex communities, e.g., including both a NAICS sector and a temporal interval. For crisis year 2008, there is a display of co-movement across a greater diversity of sectors, in comparison to the baseline year 2004, for both maximal (LMx) and minimal (LMn) prices. We also observe more complex communities across more factors.

Further research is needed to isolate complex factors, e.g., a co-movement that starts in one community and moves on to another community, or co-movement communities that repeat over the temporal dimension. We may additionally consider a four mode tensor, where the fourth dimension identifies the specific extrema feature. This would allow us to study co-movements across multiple features, e.g., AM-Neg and LMn. The occurrence tensor can also be extended to include additional attributes, e.g., the capital gain or loss experienced by an equity prior to the LMx or LMn, respectively.

7. REFERENCES

[1] B. W. Bader and T. G. Kolda. Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4):635–653, December 2006.

[2] J. D. Carroll and J.-J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of *eckart-young* decomposition. *Psychometrika*, 35(3):283–319, 1970.

[3] Center for research in security prices. http://www.crsp.com/.

[4] R. A. Harshman. Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. 1970.

[5] Karsha: Visualization of notf factors. https://karsha.umiacs.umd.edu/LM-NOTF/.

[6] Karsha: Pseudoprice for equities in the s & p 500 index from 2004 to 2015. https://karsha.umiacs.umd.edu/Price_Pattern/aggregate.html.

[7] North american industry classification system. https://www.naics.com/.

[8] Wso2 siddhi extensions documentation. https://docs.wso2.com/display/SIDDHIEXTENSIONS/.

[9] Z. Xu, F. Huang, L. Raschid, and T. Goldstein. Non-negative factorization of the occurrence tensor from financial contracts. 2016.