

News Co-Occurrences: 关注同时出现在新闻中的股票

Original 全网Quant都在看 量化投资与机器学习 2021-06-08 16:51

收录于合集

#深度研读系列

21个 >



量化投资与机器学习公众号独家解读

量化投资与机器学习公众号 *QIML Insight*——**深度研读系列** 是公众号今年全力打造的一档**深度、前沿、高水准**栏目。

公众号**遴选**了各大期刊最新论文，按照理解和提炼的方式为读者呈现每篇论文最精华的部分。QIML希望大家能够读到可以成长的量化文章，愿与你共同进步！

[第一期](#) | [第二期](#) | [第三期](#) | [第四期](#) | [第五期](#)

本期遴选论文

来源：Journal of Risk and Financial Management 19 March 2019

作者：Yi Tang、Yilu Zhou、Marshall Hong

标题：News Co-Occurrences, Stock Return Correlations, and Portfolio Construction Implications

核心观点

- 股票同时出现在新闻的频率与股票市值、股票波动及分析师覆盖度之间存在明显的关联性。
- 个股之间的相关性随着在新闻中同时出现频率的增加而增加。
- 个股在新闻中同时出现频率可以用于预测未来个股之间的相关性，从而应用与风险模型。


随着NLP技术的发展，新闻分析数据在量化投资中的应用的场景越来越丰富。本篇论文从新闻中同时出现不同股票（News Co-Occurrences）的角度出发，去验证其所包含的经济学含义，并探索其在量化投资中的应用。

同时出现在一篇新闻的股票之间是否有某种程度上的关联？同时出现的频率与股票关注度之间的变化是否有关系？是否会对股票之间的相关性产生影响？这些都是作者试图在文中探索的问题。

作者主要采用了线性回归的方式进行实证分析，涉及的数据及相关指标说明如下：

- 数据时间范围：2007年5月-2016年12月
- 股票范围：S&P1500
- 计算准则：月度指标至少需要24个月的数据、日度指标至少需要15天的数据
- 所有指标都在月末计算

文中涉及的其他指标的说明：

指标	解释	
BETA	过去60个月的月度收益率与基准进行回归	$R_{i,t} = \alpha_i + \beta_i MKT_t + \varepsilon_{i,t},$
ME	市值	
CVRG	月度分析师覆盖	
IVOL	日度收益与Fama因子回归后，残差的标准差（年化）	$R_{i,d} = \alpha_i + \beta_i R_{m,d} + \gamma_i SMB_d + \varphi_i HML_d + \varepsilon_{i,d},$
CORR	两个股票某月日度收益的相关系数	
IND	SIC行业分类哑变量	
CS	供应链关系哑变量，1有供应关系，0没有供应链关系	
GEO	地址位置哑变量，1在同一地区，0不在同一地区	
ASV	Google异常搜索指数，当天搜索量SVI与前t-21至t-260工作日搜索量均值的相对变化率。	$ASV_{i,d} = \frac{SVI_{i,d} - \overline{SVI_{i,(d-260,d-21)}}}{\overline{SVI_{i,(d-260,d-21)}}},$  量化投资与机器学习

统计分析

作者首先对不同分组的股票的相关指标做了统，一共分为三组：

- COC=1：当月至少和别的股票出现在同一篇新闻的所有股票
- COC=0：当月未曾和别的股票出现在同一篇新闻的股票
- All stocks：S&P500所有股票

对以上三组股票分别计算2007年5月至2016年12月每月末截面上各指标的均值，然后再计算时序上每月均值的平均值，计算结果如下表所示，可以看出：

- 第一列Pi表示，每个月，平均有47%的股票至少和其他股票同时出现在至少一篇新闻里。
- 和别的股票同时出现在一篇新闻的股票（COC=1）跟从未和别的股票出现在同一篇新闻的股票（COC=0）相比具有更低的风险（BETA及IVOL更低）、更高的市值（ME）及更高的分析师覆盖（CVGR），且与其他股票之间的相关性也更高（CORR）。

Table 1. Descriptive statistics.

Sample	π	BETA	ME	IVOL	CVRG	FREQ	TF_{μ}	TF_{\max}	CORR	$CORR_{coc=1}$
COC = 1	--	1.19	15,699	23.03	13	16	2	8	0.34	0.41
All stocks	47	1.23	9951	25.40	11	--	--	--	0.33	--
COC = 0	--	1.26	4728	27.45	9	--	--	--	0.32	--

For each month over the period May 2007–December 2016, we calculated the cross-sectional means of a set of stock characteristics, including market beta (BETA), market capitalization (ME, in millions of dollars), annualized idiosyncratic volatility of daily stock returns (IVOL, in percentage terms), number of financial analysts covering a stock (CVRG), the number of different news articles that a stock co-occurred with other stocks (FREQ), the mean (TF_{μ}) and maximum (TF_{\max}) number of different news articles in which the same pair of stocks co-occurred, the correlation coefficient of two stocks' daily returns (CORR), and the correlation coefficient of daily returns for two stocks that co-occurred in news article ($CORR_{coc}$). We then averaged the cross-sectional means across time. This table reports the time-series averages of the cross-sectional means for three samples of (1) stocks that occurred with other stocks in the same news articles in a month (denoted "COC = 1"), (2) S&P 1500 stocks (denoted "S&P 1500"), and (3) stocks that did not co-occur with any other stocks in news articles in a month (denoted "COC = 0"). The First column (π) reports the average percentage of the S&P 1500 stocks that co-occurred in news articles in a month.

News Co-Occurrences与股票特征之间的关系

News Co-Occurrences截面的变动

作者采用Fama-MacBeth的方法对以下两个等式进行回归分析（先截面回归，再算回归系数在时序上的均值）

$$LNTF_{ij,t} = \lambda_{0,t} + \lambda_{1,t}IND_{ij,t-1} + \lambda_{2,t}CS_{ij,t-1} + \lambda_{3,t}GEO_{ij,t-1} + \lambda_{4,t}LNTF_{ij,t-1} + \varepsilon_{ij,t}$$

$$LNTF_{ij,t} = \lambda_{0,t} + \lambda_{1,t}IND_{ij,t-1} + \lambda_{2,t}CS_{ij,t-1} + \lambda_{3,t}GEO_{ij,t-1} + \lambda_{4,t}LNTF_{ij,t-1} + \gamma_{1,t}\overline{BETA}_{t-1} + \gamma_{2,t}\overline{SIZE}_{t-1} + \gamma_{3,t}\overline{IVOL}_{t-1} + \gamma_{4,t}\overline{CVRG}_{t-1} + \varepsilon_{ij,t}$$

相关变量的解释：

指标	解释
$LNTF_{ij,t}$	t月中，对（1+股票i,j同时出现的新闻的数量）取对数
$IND_{ij,t-1}$	t-1月，股票i,j是否属于同一个行业
$CS_{ij,t-1}$	t-1月，股票i,j是否存在供应链关系
$GEO_{ij,t-1}$	t-1月，股票i,j是否在同一注册地
\overline{BETA}_{t-1}	t-1月，股票i和j的BETA均值
\overline{SIZE}_{t-1}	t-1月，股票i和j的SIZE均值
\overline{IVOL}_{t-1}	t-1月，股票i和j的IVOL均值
\overline{CVRG}_{t-1}	t-1月，股票i和j的CVRG均值

量化投资与机器学习

下表给出了回归的结果，其中Model1对应等式4，Model2对应等式5。其中Model1中，IND、CS及GEO的回归系数分别是0.073、0.098及0.032，且在置信度99%的区间里均显著。这意味着处于同一个行业，存在供应链关系或在同一个地区的股票有更高的概率出现在同一篇新闻中。即使在Model2中控制了其他变量（包括BETA、SIZE、

IVOL及CVRG)，IND、CS及GEO的回归结果与Model1相比基本没受影响。同时也可以看出，同时出现在新闻的数量与BETA成负相关，与SIZE和CVRG呈正相关，这个结果与表1的结果保持一致。

News Co-Occurrences的拆解

作者用LNTFP和LNTFR分别表示模型的拟合值和残差。每个月，分别计算LNTFP及LNTFR的均值和标准差，再计算时序上的统计值。在表2的B部分，Expected表示拟合值LNTFP，Shock表示残差LNTFR。可以看出，Model1和Model2的结果非常类似。再后续的分析应用中，作者选取了更完整的Model2。

Table 2. News co-occurrence and stock characteristics.

Panel A. Explaining News Occurrences									
Model	IND	CS	GEO	LTF	BETA	SIZE	IVOL	CVRG	Adj. R ²
(1)	0.073 (11.10)	0.098 (7.19)	0.032 (5.70)	0.307 (51.37)					0.157
(2)	0.073 (12.20)	0.091 (6.59)	0.035 (6.64)	0.307 (54.80)	-0.012 (-3.01)	0.005 (3.40)	0.002 (0.73)	0.001 (1.70)	0.165
Panel B. Descriptive Statistics for Components of News Co-Occurrences									
	Model (1)		Model (2)						
	Expected	Shock	Expected	Shock					
Mean	1.016	0.000	1.017	0.000					
Std. dev.	0.036	0.048	0.034	0.048					

量化投资与机器学习

News Co-Occurrences与投资者关注度之间的关系

为了研究News Co-Occurrences与投资者关注度之间的关系，作者采用了两个模型，等式6和7的区别是，等式6中News Co-Occurrences直接用LNTF表示。等式7中，News Co-Occurrences用两个变量LNTFP和LNTFR表示，它们各自有自己的回归系数，这样做就可以看出是LNTFP更重要还是LNTFR更重要。

$$\overline{ASV}_{ij,t} = \lambda_{0,t} + \lambda_{1,t}LNTF_{ij,t} + \varepsilon_{ij,t}$$

$$\overline{ASV}_{ij,t} = \lambda_{0,t} + \lambda_{1,t}LNTFP_{ij,t} + \lambda_{2,t}LNTFR_{ij,t} + \varepsilon_{ij,t}$$

下表3给出了以上两个模型的回归结果，可以看出LNTF、LNTFP及LNTFR的回归系数均显著，但可以看出LNTFR相比LNTFP来的更显著，说明异常的News Co-Occurrences更能引起投资者的关注。

Table 3. News co-occurrence and investor attention.

LNTF	LNTFP	LNTFR
0.004 (2.41)		
	-0.003 (-0.85)	0.005 (3.95)

量化投资与机器学习

News Co-Occurrences VS 股票之间的相关性：同步性

作者通过以下两个模型，验证股票之间的相关性与News Co-Occurrences的关系。大部分变量在上文解释过。这里在重复下， $CORR_{ij,t}$ 表示在t月，股票i与j月度收益率的相关系数。

$$CORR_{ij,t} = \lambda_{0,t} + \lambda_{1,t}LNTF_{ij,t} + \lambda_{2,t}\overline{ASV}_{ij,t} + \lambda_{3,t}(\overline{ASV}_{ij,t} \times LNTF_{ij,t}) + \gamma_t CORR_{ij,t-1} + \varepsilon_{ij,t}$$

$$CORR_{ij,t} = \lambda_{0,t} + \lambda_{1,t}LNTFP_{ij,t} + \lambda_{2,t}LNTFR_{ij,t} + \lambda_{3,t}\overline{ASV}_{ij,t} + \lambda_{4,t}(\overline{ASV}_{ij,t} \times LNTFP_{ij,t}) + \lambda_{5,t}(\overline{ASV}_{ij,t} \times LNTFR_{ij,t}) + \gamma_t CORR_{ij,t-1} + \varepsilon_{ij,t}$$

以上两个模型的主要区别是，模型8使用了LNTF，用以整体判断News Co-Occurrences与股票之间的相关性是否有关系。模型9分别使用了LNTFP和LNTFR，就可以知道是LNTFP还是LNTFR与CORR的关联性更大。

表4给出了回归的结果：

其中Model(2)对应是等式8的回归结果，可以看出，ASV与ASV*LNTF的回归系数并不显著。且相比Model(1)，LNTF及CORR的回归结果基本无变化。总体可以看出，News Co-Occurrences与股票之间的相关性存在显著的关联性。

其中Model(6)对应是等式9的回归结果，可以看出，相对LNTFR，LNTFP的回归系数更显著，说明长期的LNTFP与股票之间的相关性的关联程度更大。

Table 4. Contemporaneous relation between return correlation and news co-occurrence.

Model	Intercept	LNTF	LNTFP	LNTFR	ASV	ASV × LNTF	ASV × LNTFP	ASV × LNTFR	CORR	Adj. R ²
(1)	0.270 (18.74)	0.016 (8.87)							0.308 (26.51)	0.098
(2)	0.269 (18.69)	0.017 (9.08)			-0.011 (-0.98)	0.005 (0.57)			0.308 (26.36)	0.100
(5)	0.209 (13.06)		0.078 (12.72)	0.003 (1.78)					0.304 (26.29)	0.103
(6)	0.208 (12.91)		0.079 (12.57)	0.003 (2.09)	-0.016 (-0.73)		-0.005 (-0.27)	0.011 (1.21)	0.304 (26.09)	0.105

量化投资与机器学习

上一部分，我们用当期的CORR与当期的News Co-Occurrences进行回归，检验它们的同步关联性。这一次，我们用当期的News Co-Occurrences与后面K期的CORR进行回归，检验News Co-Occurrences对CORR的预测性。

$$\begin{aligned}
 CORR_{ij,t+k} &= \lambda_{0,t} + \lambda_{1,t} LNTF_{ij,t} + \gamma_t CORR_{ij,t} + \varepsilon_{ij,t} \\
 CORR_{ij,t+k} &= \lambda_{0,t} + \lambda_{1,t} LNTF_{ij,t} + \lambda_{2,t} \overline{ASV}_{ij,t} + \lambda_{3,t} (\overline{ASV}_{ij,t} \times LNTF_{ij,t}) + \gamma_t CORR_{ij,t} + \varepsilon_{ij,t} \\
 CORR_{ij,t+k} &= \lambda_{0,t} + \lambda_{1,t} LNTFP_{ij,t} + \lambda_{2,t} LNTFR_{ij,t} + \gamma_t CORR_{ij,t} + \varepsilon_{ij,t} \\
 CORR_{ij,t+k} &= \lambda_{0,t} + \lambda_{1,t} LNTFP_{ij,t} + \lambda_{2,t} LNTFR_{ij,t} + \lambda_{3,t} \overline{ASV}_{ij,t} + \lambda_{4,t} (\overline{ASV}_{ij,t} \times LNTFP_{ij,t}) \\
 &\quad + \lambda_{5,t} (\overline{ASV}_{ij,t} \times LNTFR_{ij,t}) + \gamma_t CORR_{ij,t} + \varepsilon_{ij,t}
 \end{aligned}$$

下表5给型的回归结果其中PanelA对应等式10，PanelB对应等式11，PanelC对应等式12，PanelD对应等式13。果，不同的K，表示不同的预测间隔，如K=2，表示用当月的News Co-Occurrences预测未来2个月后的CORR。主要结论如下：News Co-Occurrences能够显著预测未来个股之间的相关性CORR，且长期的均值LNTFP比短期的变动LNTFR具有更强的预测性，且不随着预测间隔的增加出现衰减。

Table 5. Predictive relation between return correlation and news co-occurrence.

Panel A. Results from Model (1)				
	Intercept	LNTF	CORR	Adj. R ²
k = 1	0.286 (19.43)	0.014 (7.01)	0.299 (26.93)	0.096
k = 2	0.275 (17.56)	0.014 (6.55)	0.309 (27.31)	0.100
k = 3	0.282 (16.75)	0.017 (7.23)	0.279 (26.48)	0.080
k = 4	0.290 (16.65)	0.017 (7.68)	0.264 (26.55)	0.075
k = 5	0.283 (16.49)	0.018 (7.37)	0.277 (25.35)	0.081
k = 6	0.289 (16.55)	0.021 (8.05)	0.251 (25.04)	0.068
k = 7	0.291 (17.17)	0.019 (7.75)	0.253 (24.25)	0.068
k = 8	0.283 (15.81)	0.019 (9.25)	0.274 (25.54)	0.079
k = 9	0.297 (17.36)	0.019 (8.45)	0.245 (24.48)	0.066
k = 10	0.302 (17.58)	0.018 (7.55)	0.239 (22.53)	0.063
k = 11	0.290 (18.55)	0.019 (8.46)	0.265 (28.97)	0.078
k = 12	0.298 (17.49)	0.017 (7.65)	0.243 (24.86)	0.064

Panel B. Results from Model (2)

	Intercept	LNTF	ASV	ASV \times LNTF	CORR	Adj. R ²
$k = 1$	0.286 (19.49)	0.015 (7.07)	-0.012 (-1.03)	0.005 (0.62)	0.299 (26.93)	0.098
$k = 2$	0.275 (17.61)	0.014 (6.47)	-0.004 (-0.38)	0.004 (0.46)	0.309 (27.31)	0.101
$k = 3$	0.281 (16.71)	0.017 (7.58)	-0.010 (-0.92)	0.008 (0.94)	0.279 (26.48)	0.081
$k = 4$	0.290 (16.64)	0.017 (7.76)	-0.008 (-0.58)	0.005 (0.49)	0.264 (26.55)	0.077
$k = 5$	0.283 (16.55)	0.018 (7.30)	0.004 (0.38)	-0.004 (-0.44)	0.277 (25.35)	0.083
$k = 6$	0.288 (16.56)	0.021 (8.22)	-0.017 (-1.40)	0.014 (1.47)	0.251 (25.04)	0.070
$k = 7$	0.290 (17.16)	0.020 (7.67)	-0.023 (-1.91)	0.010 (1.09)	0.253 (24.25)	0.070
$k = 8$	0.282 (15.76)	0.020 (9.07)	-0.016 (-1.30)	0.005 (0.50)	0.274 (25.54)	0.082
$k = 9$	0.295 (17.27)	0.019 (8.61)	-0.015 (-1.38)	0.004 (0.52)	0.245 (24.48)	0.068
$k = 10$	0.301 (17.52)	0.018 (7.64)	-0.019 (-1.73)	0.013 (1.57)	0.239 (22.53)	0.065
$k = 11$	0.288 (18.46)	0.020 (8.52)	-0.023 (-2.03)	0.014 (1.73)	0.265 (28.97)	0.080
$k = 12$	0.297 (17.37)	0.018 (8.34)	-0.026 (-1.92)	0.015 (1.48)	0.243 (24.86)	



 6 量化投资与机器学习

Table 5. Cont.

Panel C. Results from Model (3)

	Intercept	LNTFP	LNTFR	CORR	Adj. R ²
$k = 1$	0.244 (13.76)	0.060 (8.59)	0.004 (2.35)	0.295 (26.86)	0.100
$k = 2$	0.223 (13.19)	0.066 (12.27)	0.002 (0.98)	0.305 (27.02)	0.104
$k = 3$	0.231 (10.95)	0.073 (8.40)	0.004 (2.06)	0.274 (26.36)	0.085
$k = 4$	0.236 (12.04)	0.073 (11.11)	0.004 (2.19)	0.260 (26.30)	0.079
$k = 5$	0.230 (11.56)	0.075 (10.25)	0.005 (2.47)	0.273 (25.09)	0.086
$k = 6$	0.226 (12.02)	0.085 (13.77)	0.007 (2.84)	0.246 (24.65)	0.073
$k = 7$	0.240 (11.11)	0.075 (9.07)	0.007 (3.23)	0.249 (23.84)	0.073
$k = 8$	0.230 (11.61)	0.074 (13.17)	0.007 (3.71)	0.269 (25.14)	0.083
$k = 9$	0.245 (12.34)	0.073 (11.23)	0.006 (3.13)	0.241 (24.06)	0.070
$k = 10$	0.241 (12.94)	0.079 (12.21)	0.005 (2.53)	0.234 (22.14)	0.068
$k = 11$	0.231 (13.03)	0.077 (12.57)	0.007 (3.63)	0.261 (28.55)	0.082
$k = 12$	0.242 (12.11)	0.076 (10.69)	0.004 (1.88)	0.238 (24.43)	

 量化投资与机器学习

Panel D. Results from Model (4)

	Intercept	LNTFP	LNTFR	ASV	ASV × LNTFP	ASV × LNTFR	CORR	Adj. R ²
$k = 1$	0.244 (13.87)	0.060 (8.75)	0.004 (2.38)	-0.003 (-0.12)	0.001 (0.06)	0.005 (0.61)	0.295 (26.84)	0.101
$k = 2$	0.224 (13.25)	0.066 (12.11)	0.002 (1.20)	-0.017 (-0.48)	0.006 (0.24)	0.009 (1.04)	0.305 (27.09)	0.105
$k = 3$	0.232 (10.98)	0.073 (8.33)	0.005 (2.66)	0.005 (0.13)	-0.019 (-0.73)	0.021 (2.26)	0.273 (26.27)	0.086
$k = 4$	0.237 (12.05)	0.073 (11.07)	0.004 (2.38)	-0.017 (-0.40)	0.000 (0.00)	0.009 (0.96)	0.259 (26.25)	0.082
$k = 5$	0.231 (11.60)	0.074 (10.04)	0.005 (2.54)	-0.003 (-0.09)	-0.012 (-0.42)	0.001 (0.14)	0.272 (25.05)	0.088
$k = 6$	0.225 (12.08)	0.085 (13.63)	0.007 (3.15)	-0.016 (-0.41)	-0.001 (-0.02)	0.020 (2.10)	0.246 (24.69)	0.075
$k = 7$	0.240 (11.17)	0.074 (9.05)	0.008 (3.41)	0.000 (0.02)	-0.009 (-0.36)	0.015 (1.54)	0.249 (23.96)	0.075
$k = 8$	0.229 (11.56)	0.075 (13.71)	0.007 (3.69)	-0.034 (-1.39)	0.021 (0.96)	0.002 (0.24)	0.268 (25.03)	0.086
$k = 9$	0.243 (12.26)	0.074 (11.36)	0.007 (3.35)	0.016 (0.63)	-0.023 (-1.01)	0.010 (1.08)	0.240 (23.98)	0.072
$k = 10$	0.239 (12.91)	0.079 (12.26)	0.006 (2.60)	0.011 (0.49)	-0.013 (-0.68)	0.019 (2.18)	0.234 (22.03)	0.069
$k = 11$	0.229 (13.04)	0.078 (12.74)	0.008 (3.82)	-0.007 (-0.23)	0.001 (0.05)	0.016 (1.91)	0.260 (28.46)	0.084
$k = 12$	0.241 (12.14)	0.077 (10.96)	0.005 (2.50)	-0.001 (-0.03)	-0.008 (-0.38)	0.013 (2.05)	0.238 (24.26)	0.070

总结

以上两部分可以知道：

- News Co-Occurrences的长期均值（LNTFP）与股票之间相关性的关联度更大
- News Co-Occurrences的短期变化（LNTFR）与投资者对股票的异常关注关联度更大
- News Co-Occurrences能够显著预测未来个股之间的相关性CORR，且长期的均值LNTFP比短期的变动LNTFR具有更强的预测性，且不随着预测间隔的增加出现衰减。

量化投资与机器学习微信公众号，是业内垂直于**量化投资**、**对冲基金**、**Fintech**、**人工智能**、**大数据**等领域的主流自媒体。公众号拥有来自**公募**、**私募**、**券商**、**期货**、**银行**、**保险**、**高校**等行业**20W+**关注者，连续2年被腾讯云+社区评选为“年度最佳作者”。

收录于合集 [#深度研读系列](#) 21

< 上一篇

Rebeco：A股低风险异象的实证研究

下一篇 >

供应链数据因子化研究：Customer Momentum

People who liked this content also liked

北大满哥与奥迪的罗生门

