基于希尔伯特变换的金融高频数据相关性检验

王天宇

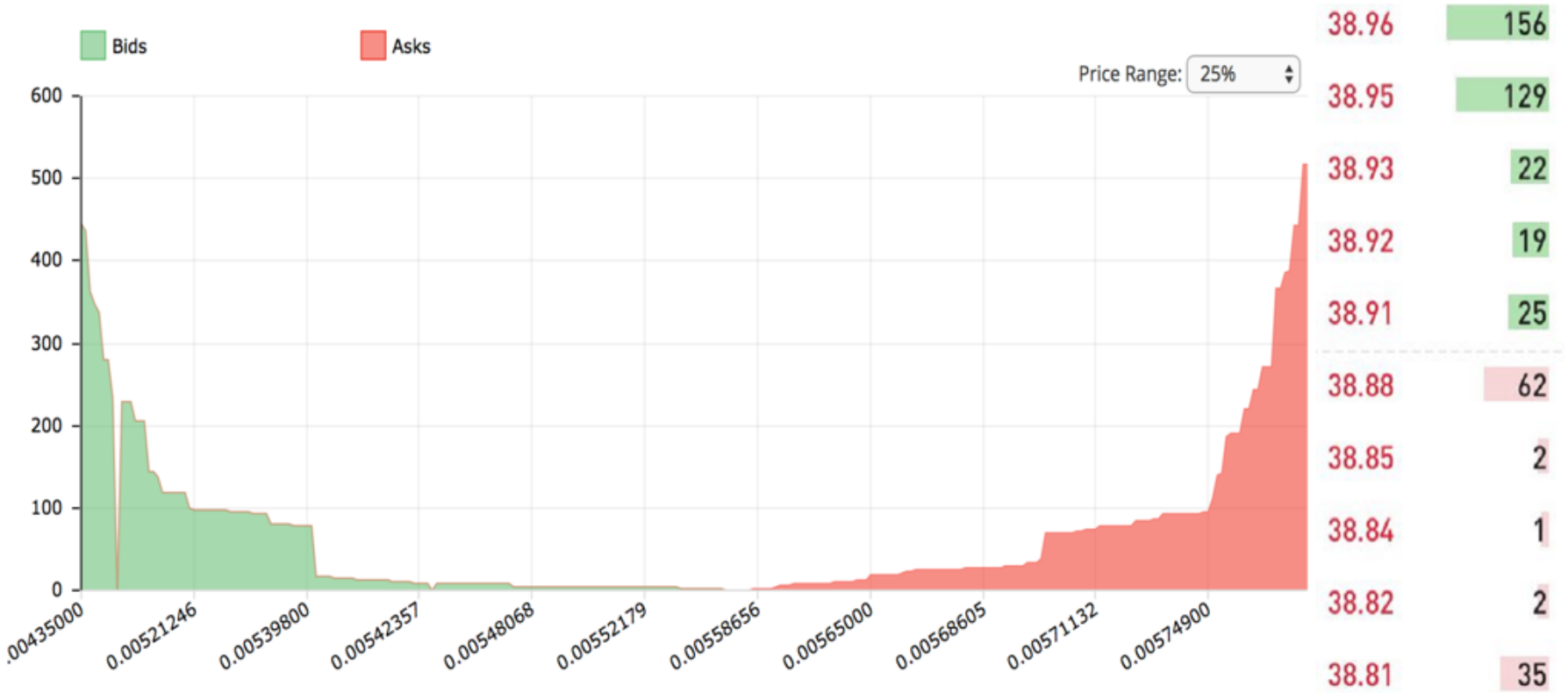
上海交通大学

2018 年 6 月 7 日

背景介绍

- 传统相关性检验方法：皮尔森相关系数
- 分析高频数据时存在问题：
  - 高频数据时间间隔不均等
  - 皮尔森相关系数无法展现数据间的联动关系(lead-lag)
- 文献：Wilinski M, Ikeda Y, Aoyama H. Complex correlation approach for high frequency financial data[J]. Journal of Statistical Mechanics: Theory and Experiment, 2018, 2018(2): 023405.
- 数据：上证50成分股2018年6月1日1分钟级交易数据
  - 数据来源：Tushare(免费、开源的Python财经数据接口包)
  - 可推广至tick级交易数据及其它投资标的，出于展示目的选用该数据

订单簿(order book)结构



高频数据结构

	TradingDay	Contract	Time	Update_millisec	av1	ap1	bv1	bp1	lastprice	tradevol	SystemTime
0	20170508	m1709	09:00:01	245	164	2851	1805	2850	2850	726732	2017-05-08 09:00:00.924
1	20170508	m1709	09:00:01	731	24	2851	1064	2850	2850	729428	2017-05-08 09:00:01.393
2	20170508	m1709	09:00:02	245	2327	2852	345	2851	2851	731082	2017-05-08 09:00:01.877
3	20170508	m1709	09:00:02	745	2111	2852	200	2851	2851	732614	2017-05-08 09:00:02.377
4	20170508	m1709	09:00:03	239	46	2851	312	2850	2850	735446	2017-05-08 09:00:02.877
5	20170508	m1709	09:00:03	742	44	2850	183	2849	2850	736842	2017-05-08 09:00:03.399
6	20170508	m1709	09:00:04	244	714	2850	23	2849	2850	737780	2017-05-08 09:00:03.890
7	20170508	m1709	09:00:04	722	88	2849	784	2848	2849	738124	2017-05-08 09:00:04.372
8	20170508	m1709	09:00:05	243	743	2849	645	2848	2849	738680	2017-05-08 09:00:04.876
9	20170508	m1709	09:00:05	744	419	2849	356	2848	2849	739824	2017-05-08 09:00:05.369
10	20170508	m1709	09:00:06	245	368	2849	749	2847	2848	740438	2017-05-08 09:00:05.874
11	20170508	m1709	09:00:06	742	285	2849	768	2847	2849	740892	2017-05-08 09:00:06.371
12	20170508	m1709	09:00:07	243	64	2848	881	2847	2848	741484	2017-05-08 09:00:06.877
13	20170508	m1709	09:00:07	739	76	2849	16	2848	2849	742538	2017-05-08 09:00:07.376
14	20170508	m1709	09:00:08	241	150	2849	121	2848	2848	743252	2017-05-08 09:00:07.868
15	20170508	m1709	09:00:08	739	6	2849	134	2848	2849	744212	2017-05-08 09:00:08.370
16	20170508	m1709	09:00:09	232	38	2849	34	2848	2848	744546	2017-05-08 09:00:08.870

分钟数据结构

	trade_date	time	symbol	askprice1	askvolume1	bidprice1	bidvolume1	open	high	low	close	turnover	volume	vwap
0	20180601	93500	600000.SH	10.58	5300.0	10.56	22400.0	10.56	10.59	10.55	10.56	953867.0	90330.0	10.559803
1	20180601	93600	600000.SH	10.60	20000.0	10.59	3925.0	10.56	10.62	10.56	10.60	1880508.0	177515.0	10.593516
2	20180601	93700	600000.SH	10.62	7700.0	10.61	1350.0	10.60	10.64	10.59	10.61	1914401.0	180250.0	10.620810
3	20180601	93800	600000.SH	10.57	4100.0	10.56	33300.0	10.61	10.62	10.56	10.57	2122646.0	200575.0	10.582804
4	20180601	93900	600000.SH	10.59	5075.0	10.58	14400.0	10.56	10.59	10.56	10.58	895766.0	84743.0	10.570383
5	20180601	94000	600000.SH	10.59	25515.0	10.58	100.0	10.59	10.60	10.58	10.58	658361.0	62200.0	10.584582
6	20180601	94100	600000.SH	10.56	100.0	10.55	32500.0	10.59	10.59	10.53	10.56	2212913.0	209700.0	10.552756
7	20180601	94200	600000.SH	10.55	14400.0	10.54	6400.0	10.57	10.57	10.55	10.55	927640.0	87900.0	10.553356
8	20180601	94300	600000.SH	10.57	2600.0	10.56	8000.0	10.55	10.58	10.55	10.57	446735.0	42300.0	10.561111
9	20180601	94400	600000.SH	10.58	33900.0	10.56	45700.0	10.57	10.58	10.56	10.58	602482.0	57000.0	10.569860
10	20180601	94500	600000.SH	10.58	27200.0	10.57	18600.0	10.57	10.58	10.56	10.58	708090.0	66994.0	10.569454
11	20180601	94600	600000.SH	10.56	10700.0	10.55	64606.0	10.57	10.57	10.55	10.55	935765.0	88600.0	10.561682
12	20180601	94700	600000.SH	10.58	21400.0	10.57	2700.0	10.56	10.58	10.55	10.58	461592.0	43700.0	10.562746
13	20180601	94800	600000.SH	10.57	1400.0	10.56	8400.0	10.57	10.59	10.57	10.57	441050.0	41700.0	10.576739
14	20180601	94900	600000.SH	10.58	509.0	10.57	17200.0	10.57	10.59	10.57	10.58	363037.0	34300.0	10.584169
15	20180601	95000	600000.SH	10.59	9515.0	10.58	14000.0	10.59	10.59	10.57	10.58	817512.0	77249.0	10.582817
16	20180601	95100	600000.SH	10.60	178958.0	10.58	5200.0	10.59	10.60	10.58	10.60	622580.0	58800.0	10.588095

数据预处理

- 行业分类：申银万国一级行业分类[1]
- 计算中间价(Mid Price)：

$$P_i(t) = \frac{A_i(t) + B_i(t)}{2} \tag{1}$$

- 价格对数化以保证收益等指标的可加性

$$p_i(t) = log(P_i(t)) \tag{2}$$

[1] http://www.swsindex.com/idx0530.aspx

方法原理—傅里叶变换

- 时间轴放缩$[0, T] \rightarrow [0, 2\pi]$
- 计算傅里叶级数系数：

$$
\begin{aligned}
a_k(dp_i) &= \frac{1}{\pi}\int_0^{2\pi} cos(kt)dp_i(t) \\
b_k(dp_i) &= \frac{1}{\pi}\int_0^{2\pi} sin(kt)dp_i(t)
\end{aligned}
\tag{3}
$$

- 离散化：

$$
\begin{aligned}
a_k(dp_i) &= \frac{p_i(t_N) - p_i(t_1)}{\pi} - \frac{1}{\pi}\sum_{m=1}^{N-1} p_i(t_m)(cos(kt_{m+1}) - cos(kt_m)) \\
b_k(dp_i) &= -\frac{1}{\pi}\sum_{m=1}^{N-1} p_i(t_m)(sin(kt_{m+1}) - sin(kt_m))
\end{aligned}
\tag{4}
$$

方法原理—傅里叶变换

- 由Malliavin论文[2] 中结果得协方差矩阵计算公式：

$$a_0(\Sigma_{ij}) = \lim_{\tau \to 0} \frac{\pi\tau}{T} \sum_{k=1}^{T/2\tau} [a_k(dp_i)a_k(dp_j) + b_k(dp_i)b_k(dp_j)] \tag{5}$$

- 在实际计算中，取$\tau = 1min$，得协方差矩阵：

$$\hat{\sigma}_{ij}^2 = 2\pi a_0(\Sigma_{ij}) \tag{6}$$

及相关矩阵

$$\rho_{ij} = \frac{\hat{\sigma}_{ij}^2}{\hat{\sigma}_{ii}\hat{\sigma}_{jj}} \tag{7}$$

[2] Malliavin and M.E.Mancino, Finance and Stochastic 6, 49(2002)

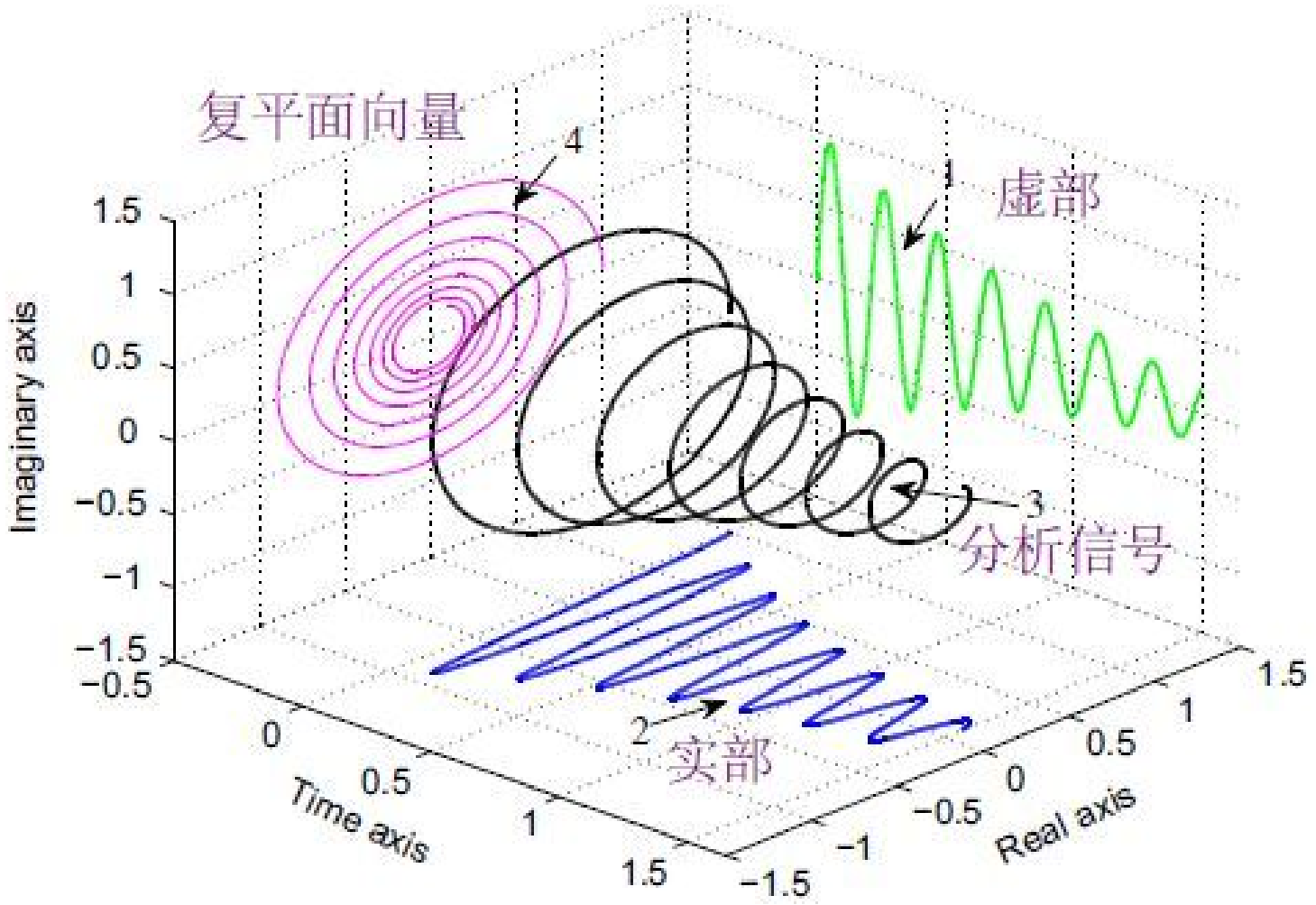
方法原理—希尔伯特变换的引入

- **傅里叶变换优点：**不要求时间序列间隔相等，不用重新采样导致不必要的数据损失
- **傅里叶变换缺点：**依然无法体现数据间的lead-lag 关系
- 引入CHPCA(Complex Hilbert Principle Component Analysis)与傅里叶变换相结合，其主要思想为希尔伯特变换：

$$H(Z,t) = p.v.\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{Z(s)}{t-s}ds \tag{8}$$

  - **实质：**Z与$\frac{1}{\pi t}$ 的卷积
  - **物理意义：**把信号的所有频率分量的相位推迟90度。

方法原理-希尔伯特变换


复平面向量
4
虚部
1
3
分析信号
2
实部
Imaginary axis
1.5
1
0.5
0
−0.5
−1
−1.5
−0.5
0
0.5
1
1.5
Time axis
1.5
1
0.5
0
−0.5
−1
−1.5
Real axis

方法原理-希尔伯特变换

- **应用方法：** $\hat{Z}(t) = Z(t) + iH(Z,t)$
- 希尔伯特变换效果：

$$\begin{aligned} H(sin(\cdot),x) &= -cosx \\ H(cos(\cdot),x) &= sinx \\ a_k(H(Z)) &= -b_k(Z) \\ b_k(H(Z)) &= a_k(Z) \end{aligned} \tag{9}$$

方法原理-希尔伯特变换

- 最终结果：

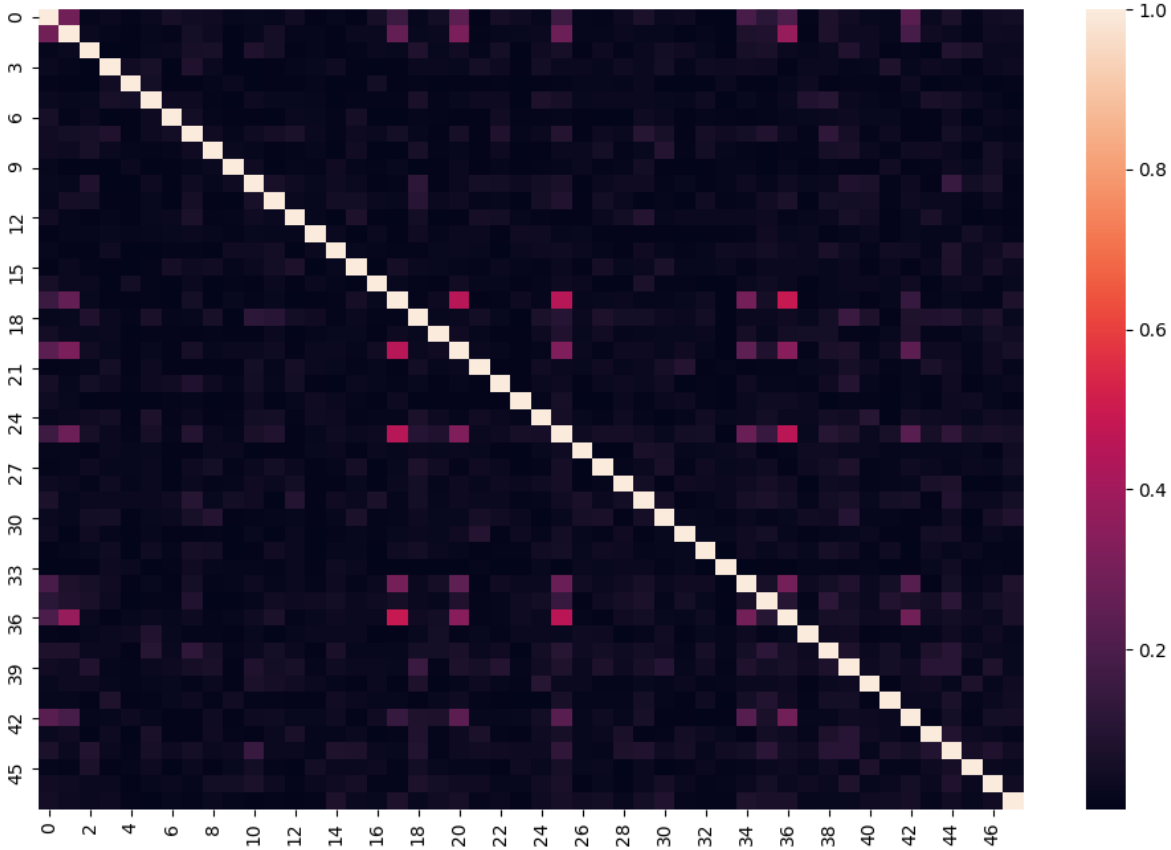
$$\begin{aligned} a_k(\hat{Z}) &= a_k(Z) + ia_k(H(Z)) = a_k(Z) - ib_k(Z) \\ b_k(\hat{Z}) &= b_k(Z) + ib_k(H(Z)) = b_k(Z) + ia_k(Z) \end{aligned} \tag{10}$$

- 协方差矩阵：

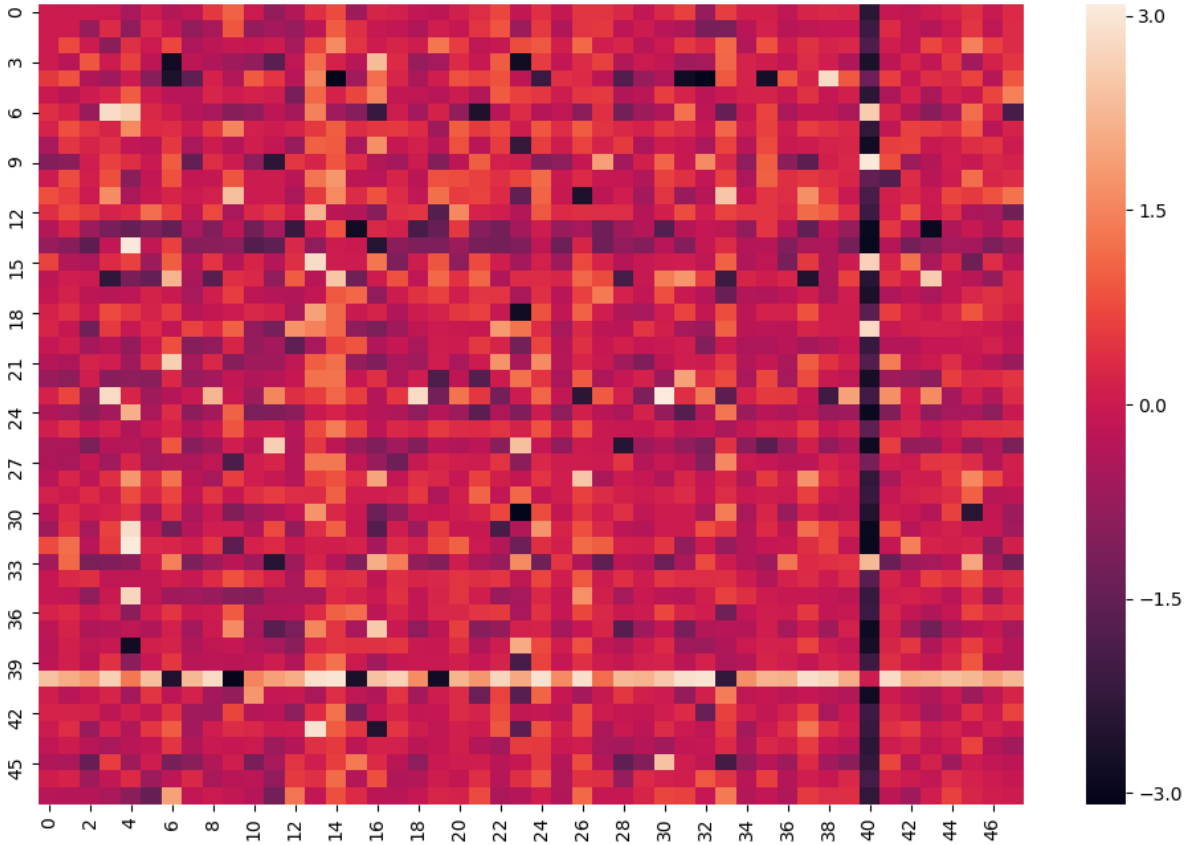
$$a_0(\Sigma_{ij}) = \frac{\pi\tau}{T} \sum_{k=1}^{T/2\tau} [a_k(\hat{dp_i})\overline{a_k(\hat{dp_j})} + b_k(\hat{dp_i})\overline{b_k(\hat{dp_j})}] \tag{11}$$

其中矩阵内元素均可表达为$\rho_{kl} = s_{kl}e^{-i\theta_{kl}}$，其中$s_{kl}$ 为幅度，可转化为相关性。$e^{-i\theta_{kl}}$为相位，可转化为lead-lag关系

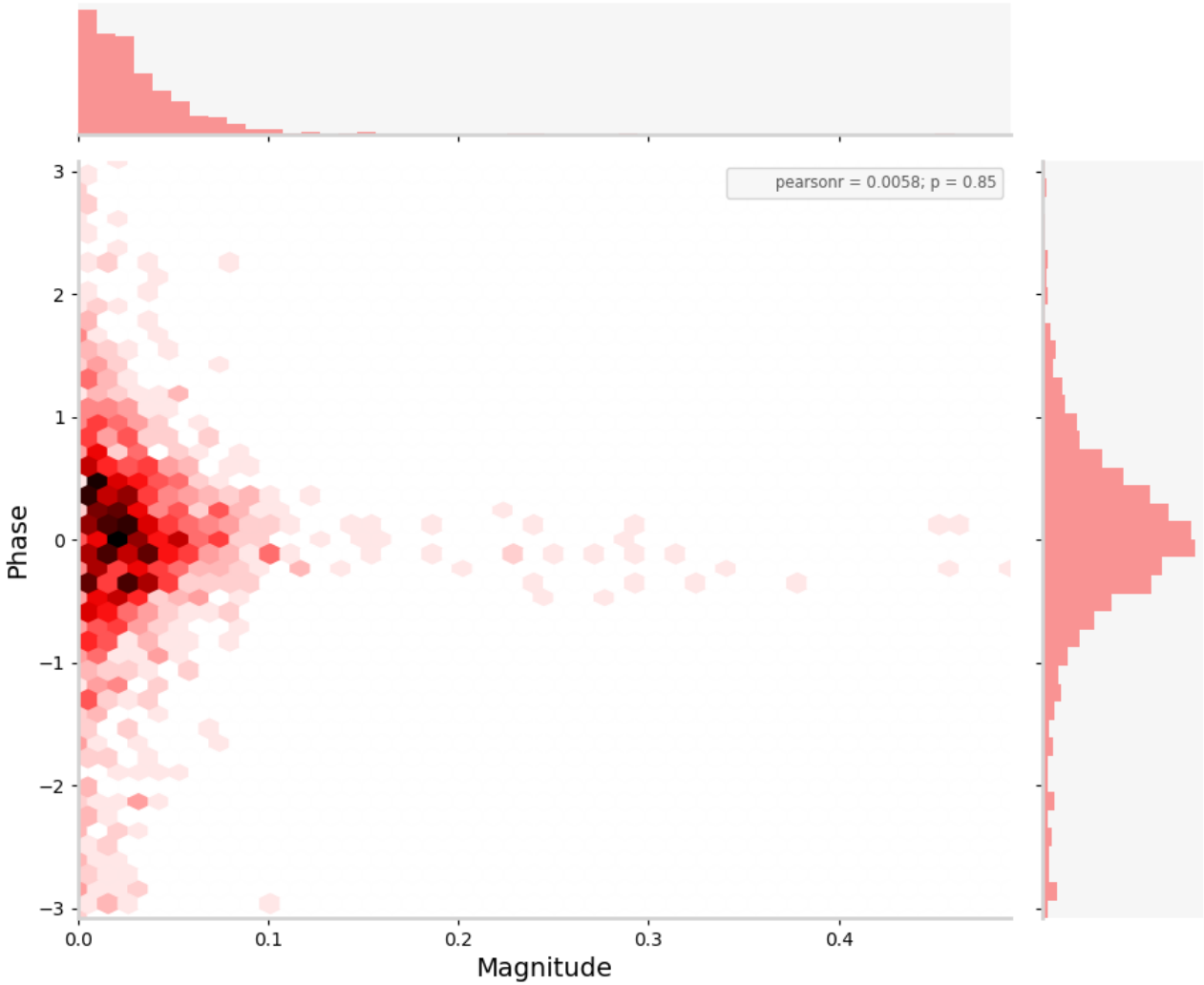
上证50成分股间相关性



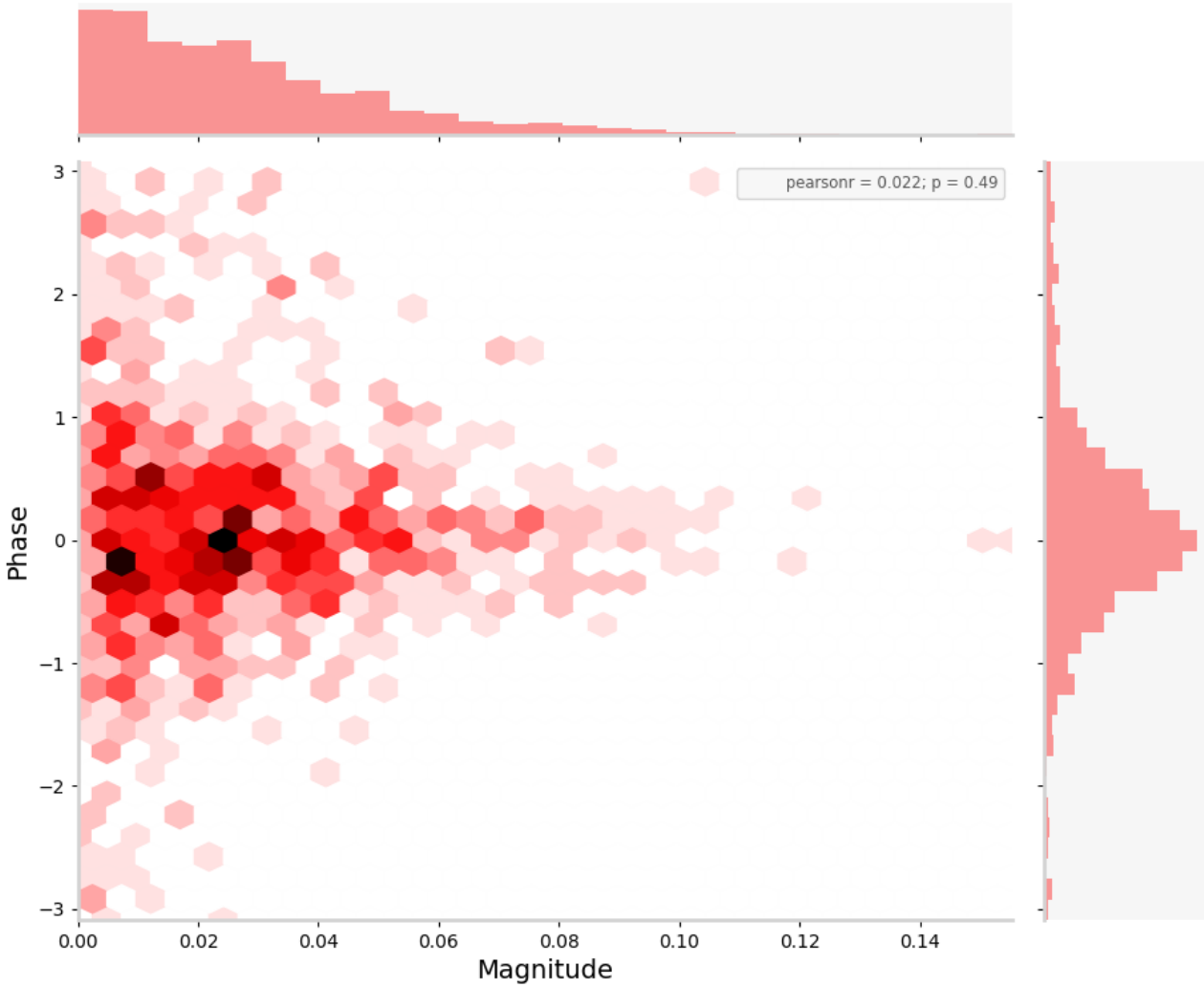
上证50成分股间相位分布



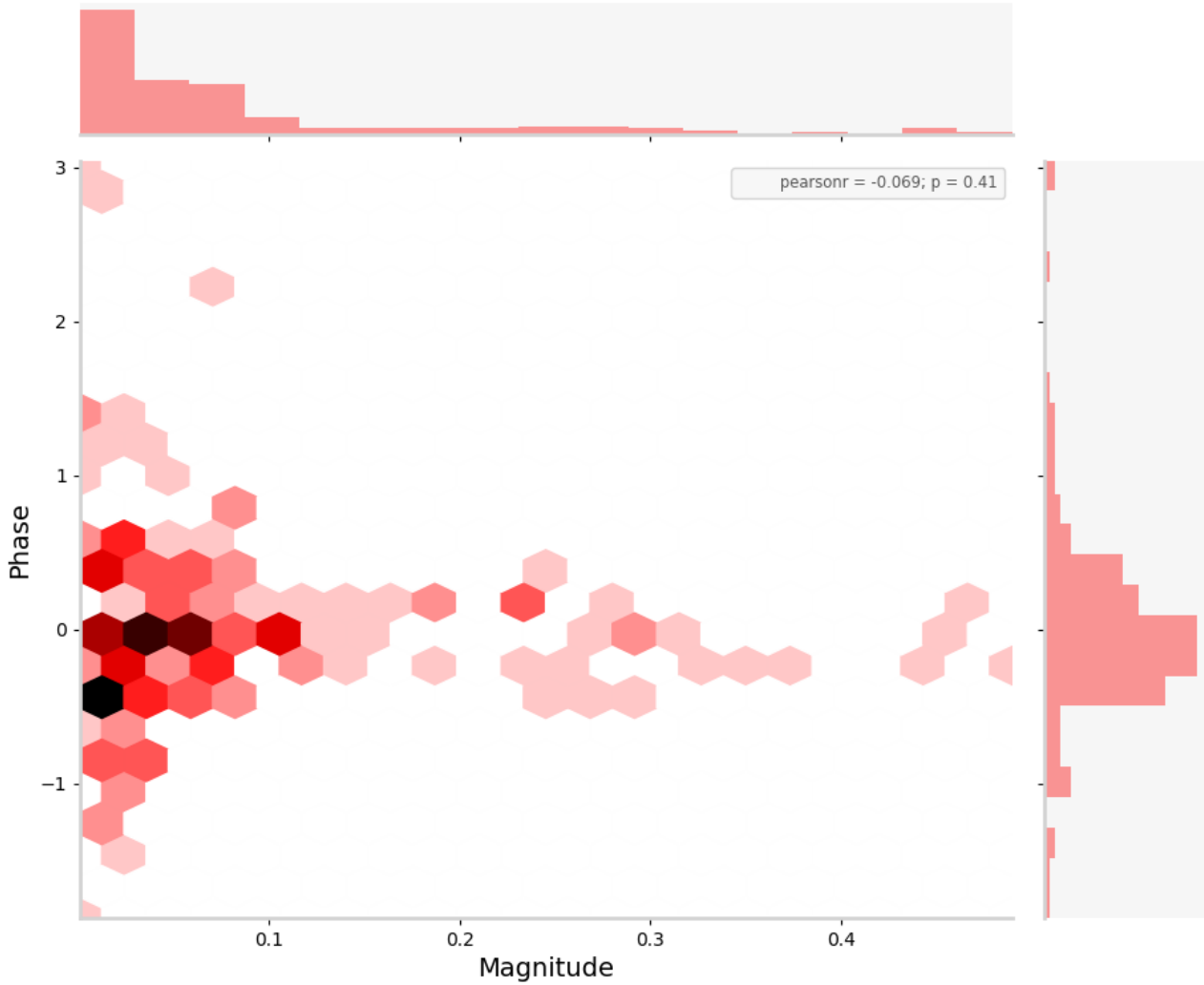
上证50成分股间相关性-相位分布



上证50成分股间相关性-相位分布(不同行业)



上证50成分股间相关性-相位分布(同行业)

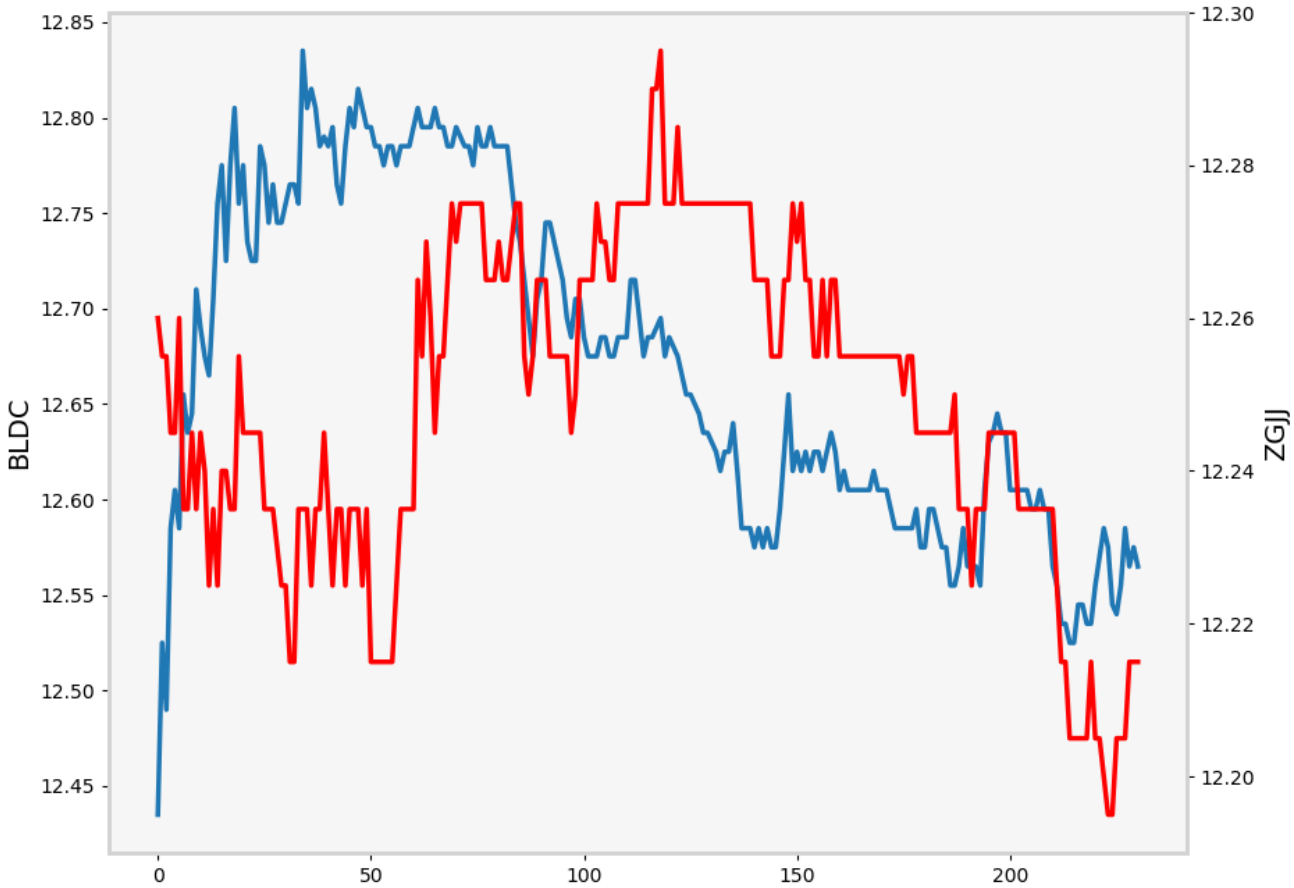


代表性股票对

高相关度股票对		高相位差股票对	
东方证券	中信证券	中国银行	江苏银行
东方证券	华泰证券	中国银行	中国建筑
中信证券	华泰证券	工商银行	山东黄金
		保利地产	中国交建

高相关度股票走势对比图(东方证券-中信证券)

高相位差股票走势对比图(保利地产-中国交建)



方法原理-复相关矩阵的特征根分解

- 由于ρ为Hermite矩阵，其可被表示为：

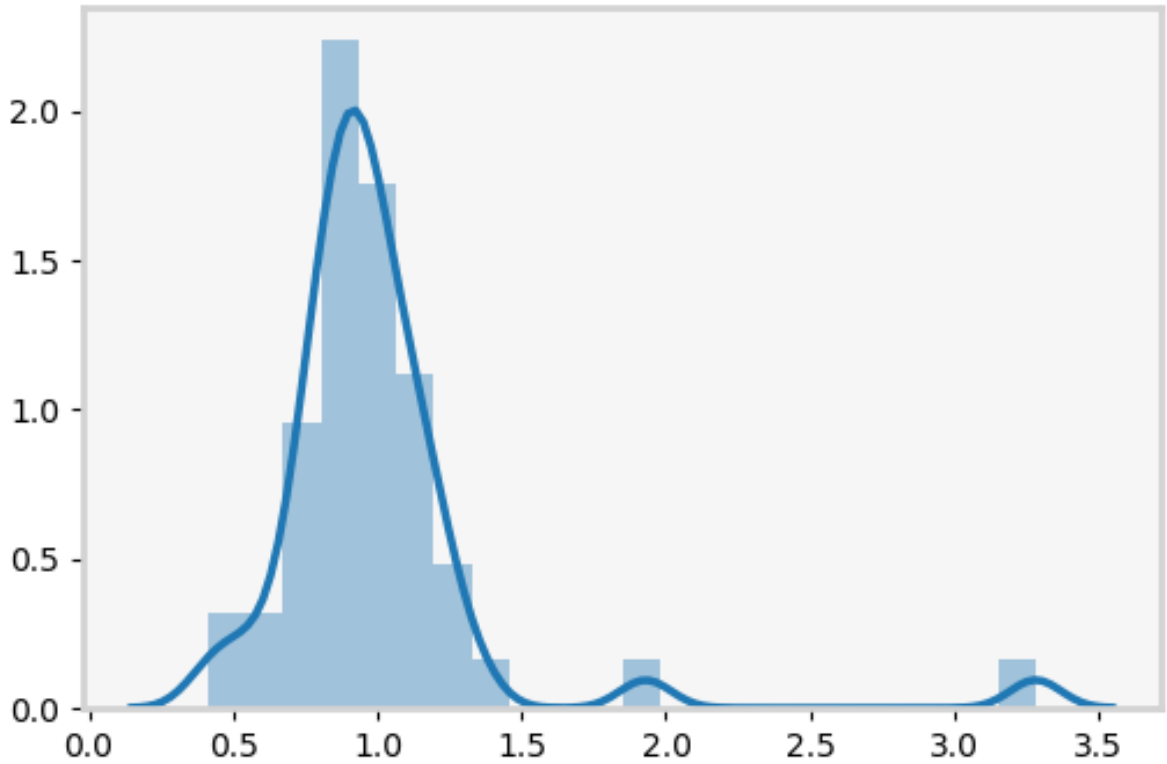
$$\rho = \sum_{i=1}^{N} \lambda_i V^{(i)} V^{(i)\dagger} \tag{12}$$

- 其中λ_i为特征根，$V^{(i)}$为对应特征向量。此时复主成分可被表示为：

$$CP_i(t) = \sum_{j=1}^{N} dp_j(t) V_j^{(i)} \tag{13}$$

  - dp_j：对数收益率
  - $V_j^{(i)}$：事实上为第j个时间序列与第i个复主成分的相关系数

特征根分布



方法原理-特征根及特征向量含义

- 最大特征根：市场模式(Market mode)
- 特征向量组成元素：对应股票与该复主成分的关系。
  - **实部**：与主成分的相关性
  - **虚部**：与主成分的领导/滞后关系
- 在日经225中随特征根大小变化体现出了可区分的差异，在上证50中效果一般。

特征向量元素分布

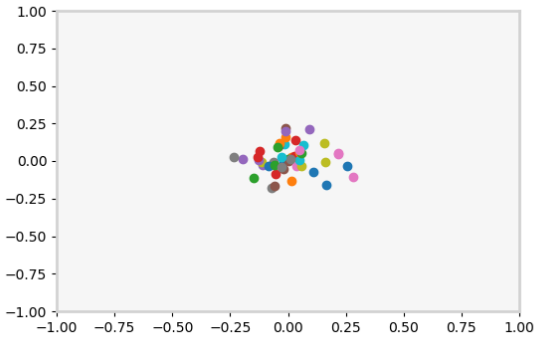


图: 最大特征根对应向量

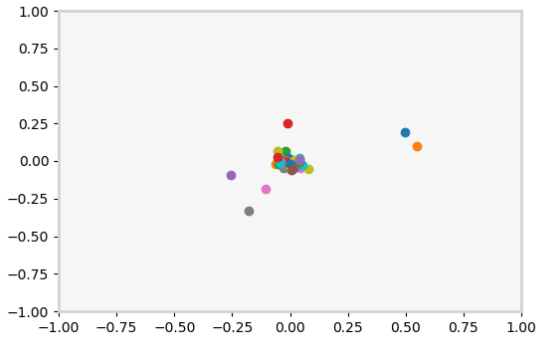


图: 第20大特征根对应向量

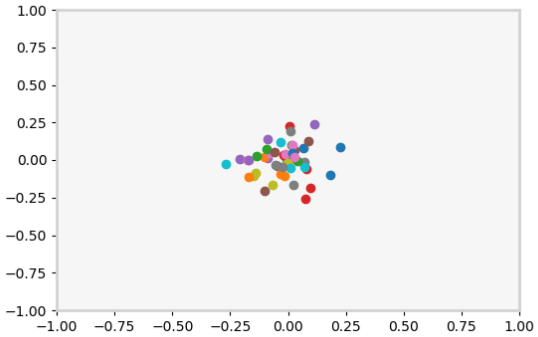


图: 第36大特征根对应向量

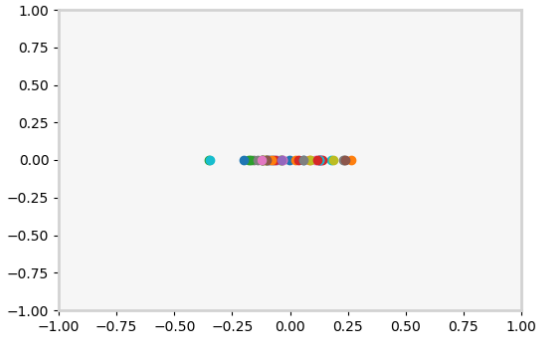


图: 最小特征根对应向量

后续研究方向

- 推广至期货市场/其他指数成分股
- 推广至tick级交易数据
- 研究特征根分解理论在中国市场失效的经济学原因