基于函数型自适应聚类的股票收益波动模式比较*

王德青 何凌云 朱建平

内容提要: 股票收益波动具有典型的连续函数特征,将其纳入连续动态函数范畴分析,能够挖掘现有离散分析方法不能揭示的深层次信息。本文基于连续动态函数视角研究上证 50 指数样本股票收益波动的类别模式和时段特征: 首先由实际离散观测数据信息自行驱动,重构隐含在其中的本征收益波动函数; 进一步,利用函数型主成分正交分解收益函数波动的主趋势,在无核心信息损失的主成分降维基础上,引入自适应权重聚类分析客观划分股票收益函数波动的模式类别; 最后,利用函数型方差分析检验不同类别收益函数之间波动差异的显著性和稳健性,并基于波动函数周期性时段划分、图形展示和可视化剖析每一类别收益函数在不同时段波动的势能转化规律。研究发现: 上证综指股票收益波动的主导趋势可以分解为四个子模式,50 只股票存在五类显著的波动模式类别,并且五类波动模式的特征差异主要体现在本次研究区间的初始阶段。本文拓展了股票收益波动模式分类和差异因素分析的研究视角,能够为金融监管部门管理策略的制定和证券市场的投资组合配置提供实证支持。

关键词: 函数型主成分; 波动模式; 自适应聚类; 上证 50

DOI: 10. 19343/j. cnki. 11 - 1302/c. 2018. 09. 007

中图分类号: C812 **文献标识码:** A **文章编号:** 1002 - 4565(2018)09 - 0079 - 13

Comparing Returns Volatility Patterns of SSE50 Stock Shares via Functional Adaptive Clustering

Wang Deqing He Lingyun Zhu Jianping

Abstract: There is a typical character of continuous function existing in the volatility of stock shares. If analyzing its trajectory under continuous dynamic function domain, we can excavate more in-depth information, which cannot be revealed by the existing discrete analysis. The paper exploits the category patterns and time-varying characteristics of returns volatility of sample shares in Shanghai stock exchange 50 index (SSE50) based on the continuous dynamic functional perspective. Firstly, to construct the intrinsic volatility function driven by information implied in actual discrete observation. Further, to decompose the dominant mode of returns volatility function orthogonally using functional principal components. Thirdly, to perform adaptive weight clustering analysis on the fluctuation mode of stock returns based on the dimensionality reduction of principal components without losing core information. Lastly, to test the robustness and significance of

* 本文为国家自然科学基金"函数型数据的自适应分类预测方法及其在金融高频预测中的应用"(71701201)、教育部人文社会科学基金"金融市场的函数型数据挖掘方法与应用研究"(15YJCZH162)、2016 年度全国统计科学研究项目"函数型数据自适应聚类分析的方法与应用研究"(2016LY13)的阶段性研究成果,同时获江苏省自然科学基金青年项目"金融高频数据的函数型自适应分类预测方法研究"(BK20170268)、中国博士后科学基金项目"函数型数据挖掘理论、方法与应用"(2015M571839)、中央高校基本科研业务费专项基金"基于函数型数据分析的大数据挖掘方法及其经济管理应用"(2015WA01)的资助。

difference among different volatility patterns via functional analysis of variance, then display graphically and analyze visually groups' potential energy transformation rule in different intervals based on periodic divisions. The Empirical results show: 1) the dominant mode of returns volatility function of SSE50 can be divided into four sub-models. 2) 50 sample stock shares have five kinds of significant volatility patterns. 3) The characteristic differences of five volatility patterns are all reflected at the initial stage of this research interval. This paper expands the research perspective by classifying returns volatility modes of stock shares and analyzing difference factors, which can be the empirical support for the financial regulatory department to formulate management strategy and for the stock market to allocate portfolio.

Key words: Functional Principal Component; Volatility Patterns; Adaptive Clustering; SSE50

一、引言

伴随大数据时代信息采集与存储技术的进步,金融市场的复杂运行过程能够被高频率地实时刻画,数据展现形式不再是孤立的离散点,而是呈现出显著的连续函数特征。传统的金融波动建模方法多是基于规则的离散观测数据,忽略了收益函数的动态变化信息,即股票价格从时点 $t-1$ 到时点 t 连续的实时转化过程。从数据挖掘的视角来看,基于传统离散时间序列框架下的收益波动分析,难以准确确定生成实际观测数据的潜在随机过程。相反,如若将频率混杂、不等间隔的离散观测值纳入平滑的连续函数范畴,则能够测度函数曲线的变化速度及加速度,进而可以从交互的多个导函数视角挖掘静态数据潜在的动态信息(Ramsay 和 Ramsey,2002[1];严明义,2010[2])。分析思路的现实普适性和方法本身的相对优势,使得函数型数据分析(Functional Data Analysis,FDA)成为大数据时代金融数据挖掘的重要视角(Tsay,2016)[3]。核心意义上,FDA 的统计思想是将离散观测数据视作具有内在统一结构的函数整体,而非个体观测值按时间先后的顺序排列[4]。尽管实际收集的收益数据既有以年、月、日为刻度的稀疏数据,也有按小时、分钟或是逐笔记录的稠密数据,但在 FDA 框架下均可将其转化为平滑的连续可导函数,通过比较曲线斜率和曲率判断股价波动规律。

函数型数据分析是现代统计研究的前沿领域,并在诸多领域有着重要的应用。国外已有研究成果显示:相对传统离散数据分析,FDA 放松了数据采集的结构约束和分布假设,能够普遍适用于频率混杂、低信噪比、不等间隔观测的稀疏型数据,在股票收益的模式识别、趋势预测、策略定价等方面具有良好的应用前景(Hong,2013)[5]。目前,基于连续动态函数视角的股票收益波动研究成果主要有:Müller 等(2011)[6]将 FDA 引入标准普尔 500 指数日内高频波动的轨迹分析,其函数型回归建模和预测技术能够识别波动率的再现模式,有助于提升期货波动预测的准确率。针对现货电价波动剧烈导致马尔科夫状态转移模型的状态分类缺少理论基础的问题,Liebl(2013)[7]基于函数型主成分因子模型建立了电价现货与电力需求之间的日需求函数关系,并以欧洲能源交易价格的预测功效为评判标准例证了函数型建模相对现有能源价格预测方法的优势。Kokoszka 和 Reimherr(2013)[8]提出了函数型主成分分析的符号检验,以美国规模以上企业为分析对象,证实了日内价格曲线形状的不可预测性。针对每分钟电力需求的超短期预测,Shang(2013)[9]通过将季节性的单变量需求函数切片为函数型时间序列,提出了基于主成分降维和非参数 bootstrap 抽样的动态更新区间预测模型,其方法能够直接推广至标准普尔 500 指数的日内收益预测[10]。Kokoszka 和 Young(2013)[11]证明了收益率曲线建模时平稳性假定的重要性,并将传统的 KPSS 检验进行了函数型时间序列视角下的拓展。Kokoszka 等(2014)[12]以经典的 CAPM 模型为基础,提出了函数型动态因子预测模型,并将其用于评估标量和曲线因子对日内价格曲线形状的影响,研究发现日内石

油期货形状与蓝筹股的股票价格密切相关。国内方面,岳敏和朱建平(2009)[13]运用函数型主成分分析(FPCA)捕捉了沪市股票收益率随时间波动的形式和方向。类似地,郭均鹏等(2012)[14]利用 FPCA 将 Shibor 市场中的期限利率的波动模式分解为长期平稳回归趋势和短期震荡扰动趋势,并基于 FPCA 得分进行聚类分析并对波动模式进行类别划分。Wang 等(2014)[15]基于 FPCA 探索了上证 50 股票月度收益率波动的主导模式,王德青等(2018)[16]则结合金融高频数据的统计特征,展望了大数据时代 FDA 在金融领域的研究方向。

综合来看,相较离散视角的波动建模,尽管 FDA 在数据处理、结构设计和模型生成等方面的优势,赋予了其卓越的识别、拟合与预测能力(Wang 等,2015)[17],但目前 FDA 在金融领域的研究还处于起步阶段,尚有不足。主要表现为: 多是直接套用其他领域的已有经验,鲜有结合金融背景的针对性机理分析[13]~[15],微观视角研究股票收益模式相似性的文献更少; 或是局限于 FPCA 的信息抽取对传统聚类和计量等模型进行拓展[6]~[12],并未基于客观的股价变化挖掘收益波动模式的类别差异及其动态演化规律; 此外,实证分析多是针对国外有效金融市场的宏观波动规律,总体指数信息掩盖了相关类股和个股收益的差异性特征,并且结论的有效性是否可以推广至国内有待考证。投资组合管理的关键是基于客观的业绩评估以正确判断股票的分类特征,而我国金融市场波动的特殊性使得单纯按行业板块的股票收益模式划分凸现诸多弊端[18]。因此,创新对股票收益数据的信息挖掘工具以提升对股市波动规律的认知,无论对金融监管部门宏观策略的制定,还是微观机构和个体的投资组合配置都具有重要意义。有鉴于此,本文以我国 A 股市场最具代表性的上证综指 50 只优质大盘样本股为分析对象,引入函数型建模思路以探索股票收益波动的主导趋势分解、类别模式差异及其动态演化特征,以期为研究我国股市的波动规律提供方法支持和实证参考。

二、研究思路与建模方法

FDA 首要工作是利用基函数展开的非参数平滑方法,由离散观测数据重构隐含在其中的连续本征函数。基于本文研究目的,首先使用函数型主成分分析对收益波动的主导趋势进行正交分解,剖析收益波动变异程度的时段特征。其次,利用函数型自适应聚类分析对股票收益波动的类别模式进行客观划分,并对不同类别模式之间存在显著差异的稳健性进行定量检验。最后,针对每一类别波动模式的特殊性进行可视化图形比较,探讨导致波动模式类别差异的核心因素。

(一)离散数据函数化

由离散数据重构本征函数主要有插值和平滑两种方法,二者核心差异在于是否存在扰动因素。基于实际问题的普遍性考虑,样本观测值 $y_{i1},\cdots,y_{iT_i}$ 多是本征函数 $\{f_i(t_j);t_j\in \mathrm{T},i=1,\cdots,n\}$ 带有噪音 $\varepsilon_i(t_j)$ 的离散实现,即 $f_i(t_j)=y_{it_j}+\varepsilon_i(t_j)$,本文主要讨论误差扰动下的粗糙惩罚平滑函数化方法。假定 $\Phi(t)=\{\phi_1(t),\cdots,\phi_L(t)\}$ 为 Hilbert 空间中的最优基函数(假定存在),则粗糙惩罚平滑的拟合残差平方和 $PENSSE_\kappa$ 为:

$$PENSSE_\kappa=\sum_{i=1}^{n}\{\sum_{j=1}^{T_i}[y_{ij}-f_i(t_j)]^2+\kappa\int[f_i''(t)]^2\mathrm{d}t\} \tag{1}$$

式(1)中的本征函数 $f_i(t)$ 为最小化惩罚残差平方和标准下由基函数线性逼近,即 $f_i(t)=\sum_{l=1}^{L}\beta_{il}\phi_l(t)$。平滑参数值 κ 用以权衡模型拟合优度与函数曲线平滑程度之间的比例关系,其最优值由最小化广义交叉验证 $GCV(\kappa)$ 标准确定,即:

$$GCV(\kappa)=\left(\frac{n}{n-df(\kappa)}\right)\left(\frac{PENSSE_\kappa}{n-df(\kappa)}\right) \tag{2}$$

式(2)中 $df(\kappa)$ 为投影矩阵 $S_{\Phi,\kappa}=\Phi(\Phi'\Phi+\kappa R)^{-1}\Phi'$ 的迹,即 $df(\kappa)=traceS_{\Phi,\kappa}$,其中 $R=$

$\int D^2\phi(s) \cdot D^2\phi'(s)\,ds$ 为基函数二阶导数外积积分构成的矩阵。基于上述符号表示，最小化式(1)的最优解为 $\hat{\beta} = (\Phi'\Phi + \kappa R)^{-1}\Phi' y$。关于惩罚平滑的函数化理论基础参阅 Kokoszka 等(2017)[19]，算法步骤详见 Ramsey 等(2009)[20]。

(二)函数型自适应聚类分析

传统离散数据的聚类分析往往难以直接应用于连续函数的分类问题[21][22]，基于降低算法计算成本和提升分类准确率的考虑，本文使用函数型数据的自适应赋权聚类分析以客观划分股票收益波动的模式类别，并对类别之间差异的显著性和聚类分析结果的稳健性进行 1000 次 Bootstrap 再抽样检验。

设 $V(s,t)$ 为 Hilbert 空间 $[0,T]^2$ 上的连续协方差函数，由 Mercer 引理存在正交的主成分函数序列 $\varphi_l(t)$ 及其对应的非负递减特征值 λ_l，使得:

$$\int_0^T V(s,t)\,\varphi_l(s)\,ds = \lambda_l\varphi_l(t) \qquad t \in [0,T], l \in N \tag{3}$$

$$\int_0^T \varphi_l(t)\,\varphi_m(t)\,dt = \delta_{lm} = \begin{cases} 1, & m = l \\ 0, & m \neq l \end{cases} \tag{4}$$

进一步，对于 $L^2(T)$ 上的二阶连续随机过程 $\{f(t), t \in [0,T]\}$，记 $\mu(t)$、$V(s,t)$ 分别为 $f(t)$ 的均值函数和协方差函数，则 $f(t)$ 的 Karhunen-Loève 展开式为:

$$f(t) = \mu(t) + \sum_{l=1}^{\infty}\zeta_l\varphi_l(t) \quad t \in [0,T] \tag{5}$$

式(5)中主成分得分 $\zeta_l(f)$ 是关于 $f(t)$ 的零均值随机变量，并且满足:

$$E[\zeta_l(f)\zeta_m(f)] = \lambda_l\delta_{lm} \qquad m,l \in N \tag{6}$$

由式(3)~(6)随机过程的 Karhunen-Loève 展开表达可知，$\zeta_l(f)$ 即去均值化函数 $f(t) - \mu(t)$ 在标准正交主成分基函数 $\varphi_l(t)$ 方向上的投影得分，并且 $\varphi_l(t)$ 由原始数据信息自适应确定。在式(5)的 Karhunen-Loève 展开下，不同函数之间的类别差异完全由其投影值 $\zeta_l(f)$ 之间的差异体现。其中 $\zeta_l(f)$ 的方差值为 λ_l，不妨假设其排列顺序满足 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$。为了体现不同 $\zeta_l(f)$ 分类效率的客观差异，令 $\beta_l = \lambda_l / \sum_{m\geqslant 1}\lambda_l$ 为 $\zeta_l(f)$ 的距离权重，定义函数 $f_i(t)$ 与 $f_j(t)$ 之间的自适应权重距离为:

$$d[f_i(t), f_j(t) \mid q] = [\sum_{l=1}^{\infty}(\beta_l \mid \zeta_l(f_i) - \zeta_l(f_j) \mid)^q]^{\frac{1}{q}} \tag{7}$$

式(7)中 q 为通常的距离参数，在无核心损失的空间映射下，基于式(7)的函数型自适应聚类分析详见王德青等(2015)[23]。

(三)函数型方差分析

函数型方差分析用以比较两组及以上函数集合是否独立同分布生成于同一随机过程，其依据的比较对象为每一函数集合的均值。令 g 为潜在的类别数目，$f_{ij}(i = 1,\cdots,g; j = 1,\cdots,n_i)$ 为类别 i 中的第 j 个函数，则函数型数据方差分析的 F 统计量构造如式(8)所示:

$$F_n = \frac{\sum_{i=1}^{g} n_i \parallel \overline{f_{i\cdot}} - \overline{f_{\cdot\cdot}} \parallel^2 / (g-1)}{\sum_{i,j} \parallel f_{ij} - \overline{f_{i\cdot}} \parallel^2 / (n-g)} \tag{8}$$

其中 $f_{ij} = (f_{ij}(t_1),\cdots,f_{ij}(t_T))'$，$\|f\| = (\int f^2(t)\,dt)^{1/2}$ 为 L^2 范数；$\overline{f_{i\cdot}} = (\overline{f_{i\cdot}}(t_1),\cdots,\overline{f_{i\cdot}}(t_T))'$、$\overline{f_{\cdot\cdot}} = (\overline{f_{\cdot\cdot}}(t_1),\cdots,\overline{f_{\cdot\cdot}}(t_T))'$ 分别为类别和总体的函数均值，其计算公式为 $\overline{f_{i\cdot}}(t) = \sum_j^{n_i} f_{ij}(t)/n_i$，

$\overline{f_{..}}(t) = \sum_{i=1}^{g} n_i \overline{f_{i.}}(t)/n$; $n = \sum_{i=1}^{g} n_i$。基于上述符号表示,式(8)的等价形式为:

$$V_n = \sum_{i<j}^{g} n_i \| \overline{f_{i.}} - \overline{f_{j.}} \|^2 \tag{9}$$

给定类别均值函数相等的原假设 H_0: $\overline{f_{1.}} = \cdots = \overline{f_{g.}}$,分别计算显著性水平 α 下的临界值 $P_{H_0}\{F_n > F_{n,\alpha}\} = \alpha$、$P_{H_0}\{V_n > V_{n,\alpha}\} = \alpha$。通过比较 F_n 与 $F_{n,\alpha}$、V_n 与 $V_{n,\alpha}$ 大小即可做出是否拒绝 H_0 的判断,具体算法步骤详见 Cuevas 等(2004)[24]、Zhang 等(2014)[25]。

三、上证 50 指数股票收益波动的类别模式比较

(一) 收益波动函数的惩罚平滑

本文选取 2015 年 4 月至 2017 年 2 月上证综指样本股的月度收益数据为实证对象,数据来自万德(WIND)数据库。如何选择最优基函数是能够准确重构收益波动函数的关键,Ramsay 和 Ramsey(2002)[1] 给出了基函数类型选取的经验法则,基于收益波动的非周期性及重构收益函数的准确率考虑,本文选取节点与观测点数目相等的 B 样条基函数。同时,为了防止收益函数频繁波动导致的过拟合问题,对函数曲线的粗糙程度进行二阶惩罚。平滑参数 κ 的选取过程如图 1 所示。

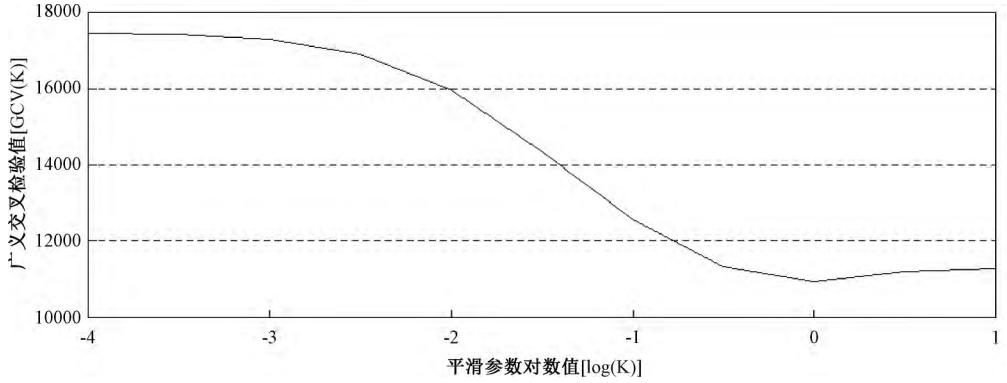


图 1　平滑参数 κ 选择的广义交叉验证

从图 1 中广义交叉验证函数 $GCV(\kappa)$ 先降后升的取值过程来看,当 $\log_{10}^{\kappa} = 0$ 时 $GCV(\kappa)$ 达到最小,即最优的惩罚平滑参数为 $\kappa = 1$。为了直观对比收益函数重构过程中粗糙惩罚的必要性,分别选取 $\kappa = 0$、$\kappa = 1$ 为平滑参数,绘制惩罚前后 50 只股票的收益函数及其均值的波动轨迹,如图 2 所示。

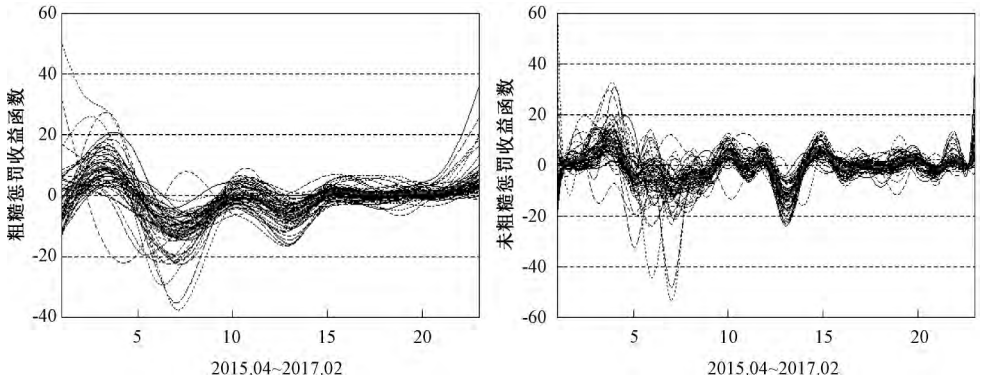


图 2　惩罚平滑(左侧)与未惩罚平滑(右侧)收益函数曲线

对比图 2 中的曲线波动轨迹发现,收益函数惩罚与否存在显著差异。具体来看,右侧未惩罚的

收益函数刻意拟合区间内的每一细微波动变化，导致函数曲线的峰谷数目较多且交替极为频繁，难以从大小不一的波动幅度中确定收益函数波动的核心规律。相对未曾惩罚收益函数频繁波动导致的高粗糙度，左侧惩罚后的收益函数更加平滑，通过抹平局部的细微变化凸显了收益波动的主体趋势。进一步从惩罚后的收益函数中可以看出，收益函数的核心波动信息集中于本次研究区间初始阶段，并且主体波动趋势可以大致划分为两次峰谷交替，能够与 2015—2017 年我国 A 股市场大涨和暴跌的股灾事件密切对应①。因此，本文后续使用 $\kappa = 1$ 为惩罚平滑参数，重构隐含在离散观测参数中的本征收益函数。函数化模型拟合的标准残差分布如图 3 所示，拟合残差的正态性概率分布检验如图 4 所示。

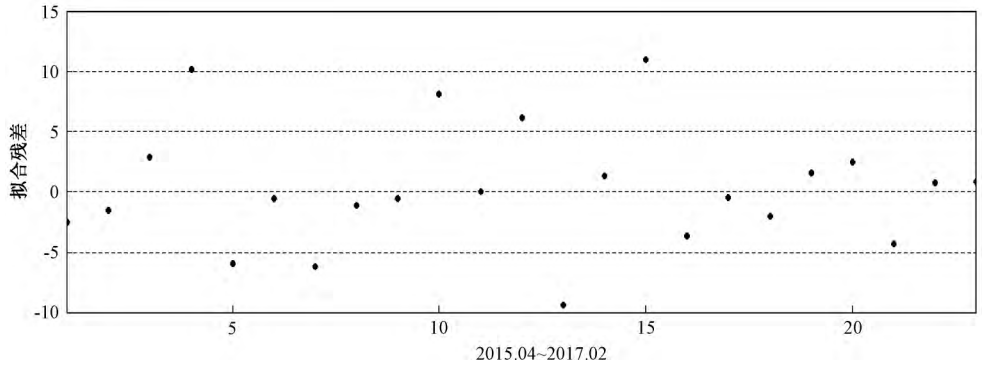


图 3 拟合残差散点分布

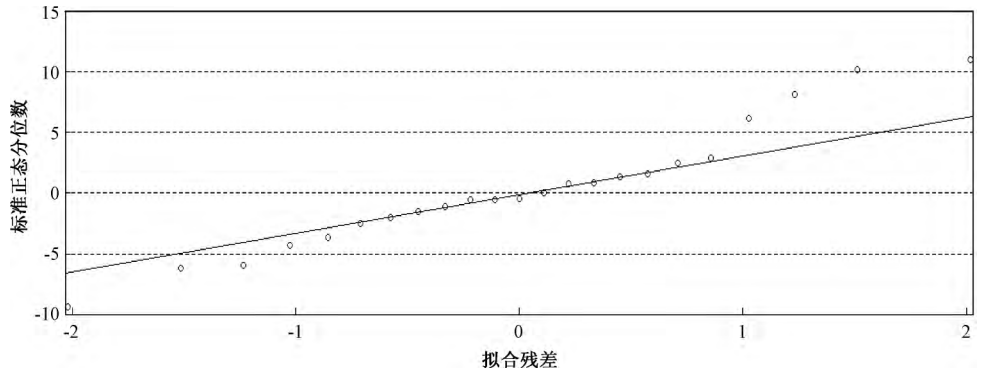


图 4 标准残差的正态概率分布检验

由图 3 的散点分布规律可以看出，拟合残差没有显现出任何系统性的结构特征，意味着函数化模型较好地拟合了离散数据，并且图 4 中正态概率检验再次印证了拟合残差分布的随机性。综上所述，基于本文设定的基函数类型、节点数目以及平滑参数值，能够准确地重构出离散观测数据中隐含的本征收益函数。

（二）基于波动趋势分解的类别模式划分

由于行业所属、股本结构、企业业绩、地域性等因素影响，不同股票收益波动的拓扑关系随不同

① 本文后续通过图形对比发现，粗糙惩罚的收益函数均值曲线与上海证券交易所发布的月度 K 线在快速上涨、断崖式下跌、政府救市、市场稳定等多个趋势转变的点位极为吻合。事实上，自 2015 年 5 月 A 股市场结束“二八分化”格局后迎来普涨行情，受金融、地产、煤炭、石油石化等权重股发力，2015 年 5 月 19 日、2015 年 5 月 22 日逾 200 只个股连续涨停。然而，2015 年 6 月始受美联储的加息预期导致的资产价格重新定位影响，全球股市集体大跌，A 股市场出现多次暴跌事件，特别是 2015 年 8 月 24 日大盘几近跌停。之后受政府救市政策影响，市场短期稳定。后期受救市反向操作行为影响，市场出现第二波幅度较小的下跌，之后政府和监管部门加大打击违规行为力度，并清理场外配资等行为，股市危机逐渐淡出。上述波动过程反映在惩罚收益函数曲线中则是前期明显的两个峰谷波动和后期逐渐平稳的轨迹。

时段动态变化,并且其形态特征具有典型的类别模式。为了从微观视角剖析收益函数波动的时段特征,利用函数型主成分对收益函数的主体波动趋势进行正交分解,方差最大化旋转后的四种子波动模式如图 5 所示。

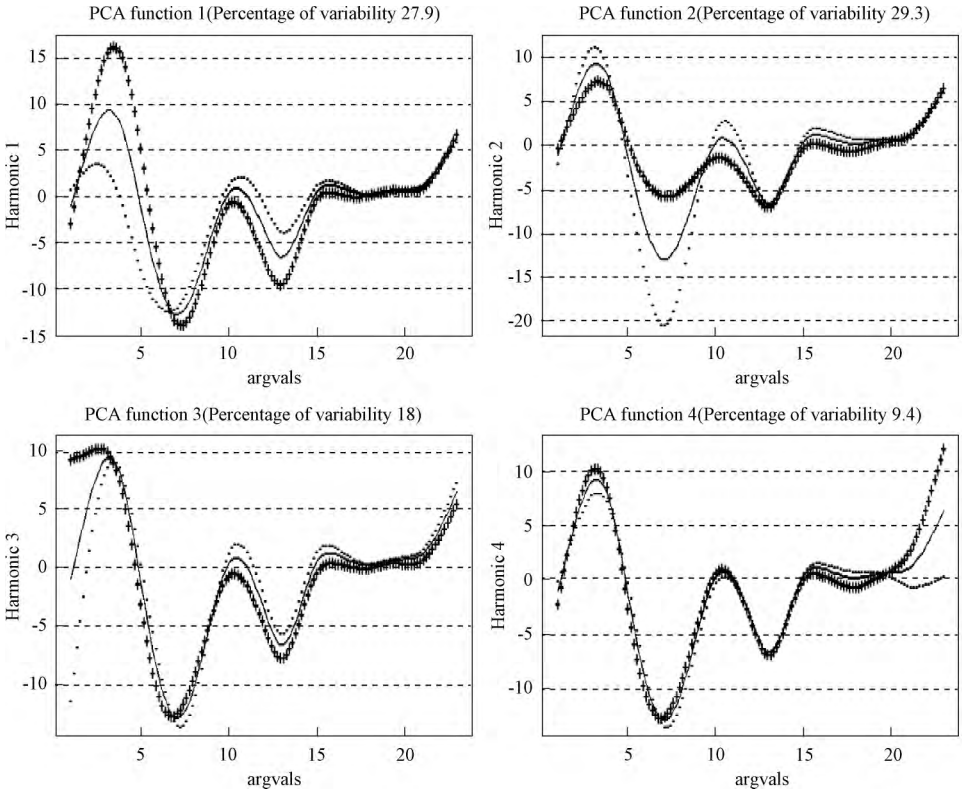


图 5　收益函数波动趋势的函数型主成分分解

由图 5 曲线波动的变异幅度及其时段特征可以看出,图中四个子波动模式能够对收益函数的波动规律进行针对性分解,并且四个主成分函数提取了原始收益函数 84.6% 的核心波动信息。具体来看,第一主成分函数主要体现收益波动的最高位阶段,第二主成分函数主要体现收益波动的最低谷阶段,第三、第四主成分函数则分别刻画收益函数初始和结尾阶段的波动特征。从信息提取贡献比来看,第一、第二主成分函数占比高达 57.2%,说明波峰与波谷是股票收益波动类别差异的核心信息时段。第三、第四主成分函数的信息含量较小,并且峰谷转换区间曲线重合度较大,说明收益函数在上述时段波动的模式类别差异并不显著。此外,虽然本次研究时段横跨两次峰谷交替,但波动的主要信息集中于第一峰谷。究其原因在于: 面对该时段内 A 股市场的第一次大涨和大跌,上市公司和投资个体在应对异常涨跌行情时存在策略和时滞差异[26],导致流动性危机并对金融体系的稳定产生冲击。在政府一系列救市政策出台后,尽管受反向操作行为影响,股市短时间出现第二次涨跌,但随着政府稳定市场政策和监管力度的加强,股市危机逐渐淡出,市场波动趋于正常。

与现有按股票所属行业板块的主观分类方法不同,本文基于股票收益率数据的函数型主成分分解结果进行自适应权重聚类分析,客观划分上证综指 50 只股票在本次研究区间内的波动模式类别。潜在类别数目的确定准则如图 6 所示,具体的类别划分结果见表 1,每一类别收益函数的波动轨迹如图 7 所示。

由图 6 的方差 – 类别数目曲线可知,当类别数目从 2 变化到 5 时,组内离差平方和下降趋势明显,但 5 类之后下降幅度显著减弱,并且分为 5 类时组间离差平方和达到最大。此外,使用 R 软件

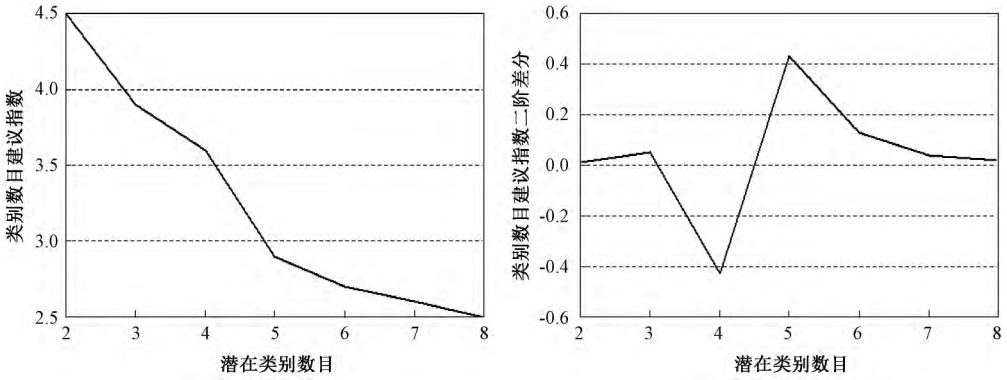


图 6 潜在类别数目确定准则

表 1 上证 50 指数样本股波动模式划分

股票代码	行业所属	类别	风格结构	股票代码	行业所属	类别	风格结构
600518	制造业	1	大起大落	600893	制造业	4	
600089	制造业	1		601186	建筑业	4	
600690	制造业	1		601390	建筑业	4	
601766	制造业	1		601668	建筑业	4	
600958	金融业	2	触底反弹	601800	建筑业	4	
600010	制造业	3	楔形整理	601989	制造业	4	
600030	金融业	3		600000	金融业	5	深 V 脉冲
600048	房地产业	3		600015	金融业	5	
600016	金融业	3		600104	制造业	5	
600109	金融业	3		600028	采矿业	5	
600036	金融业	3		600111	制造业	5	
600837	金融业	3		600519	制造业	5	
601166	金融业	3		600585	制造业	5	
600999	金融业	3		600583	采矿业	5	
601318	金融业	3		601088	采矿业	5	
601601	金融业	3		600887	制造业	5	
601688	金融业	3		601169	金融业	5	
601901	金融业	3		601288	金融业	5	
600018	交通运输	4	W 形筑底	601328	金融业	5	
600050	信息技术	4		601398	金融业	5	
600637	信息技术	4		601628	金融业	5	
600150	制造业	4		601818	金融业	5	
600256	综合类	4		601857	采矿业	5	
600406	信息技术	4		601988	金融业	5	
601006	交通运输	4		601998	金融业	5	

NbClust 函数计算的 23 个信息准则中 8 个支持分为 5 类。为了进一步验证 50 只股票分为 5 类的稳健性，利用函数型方差分析对初始分类结果进行类间差异显著性检验，1000 次 Bootstrap 再抽样的检验结果如图 8 所示。

从图 8 中收益函数均值曲线的轨迹可以看出，5 类收益函数的波动特征存在显著的模式差异，并且在 5% 显著性水平下的 Bootstrap 再抽样检验印证了类别之间差异显著的稳健性。结合图 7 中每一类别收益函数均值的波动特征，可以大致将 5 类股票收益函数的波动风格结构归纳为：大起大落、触底反弹、楔形整理、W 形筑底、深 V 脉冲。从表 1 的具体分类结果来看，第 1 类和第 3 类股票隶属同一行业或其子行业，说明股票之间的行业属性关联性较强，稳定的风格结构体现的是同质性投资信息。第 4 类股票主要由非金融业的实体经济行业构成，第 5 类股票则涵盖了所有行业的上

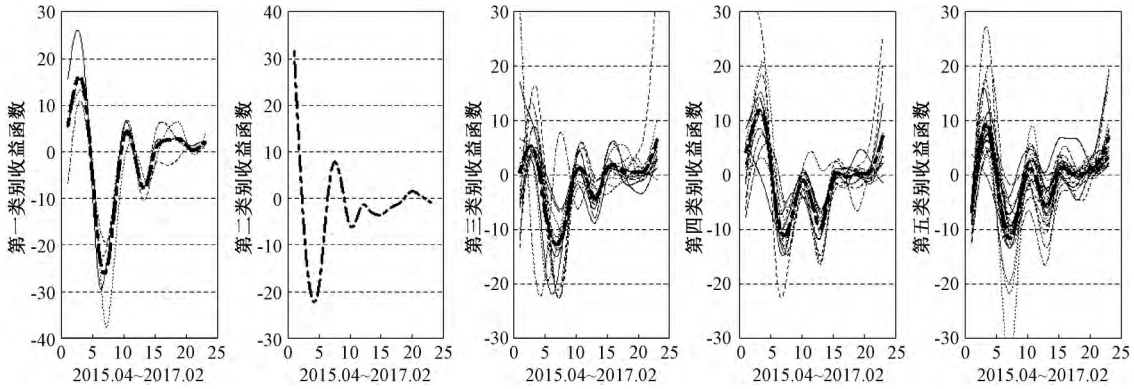


图 7　五类股票收益函数波动曲线

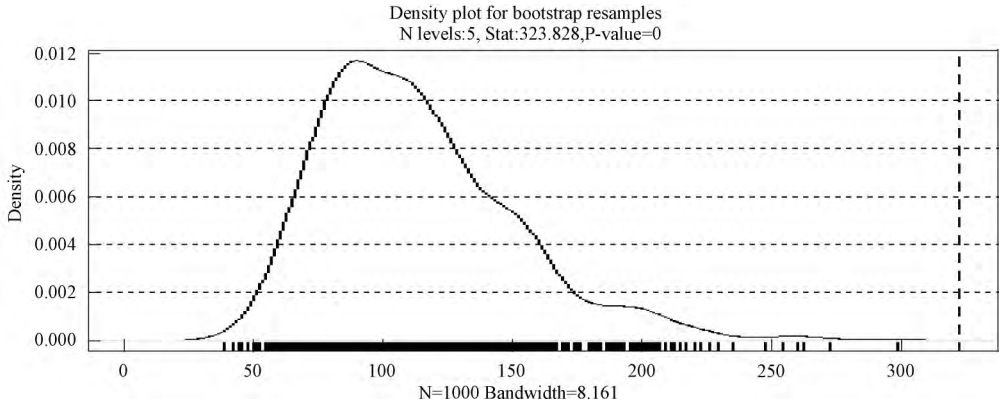


图 8　自适应聚类分析的函数型方差检验

市公司,差异较大的风格结构体现了差异性投资信息。特别地,第 2 类股票仅包含“东方证券”(600958)一只股票,原因在于该股票从 2015 年 4 月 36% 的收益断崖式跌至 2015 年 8 月的 -36%,并反弹至 2015 年 10 月的 23.7%,2016 年 1 月再次暴跌至 -24.8%,收益涨跌幅度之大远超过其他股票,说明我国股票市场实施的个股价格涨跌停板机制,目前还难以达到“理性判断”的制度效应,这也是 2016 年 1 月后我国股市进一步实施“熔断机制”的重要原因之一。结合每一类别股票的行业风格结构来看: 银行业和制造业的风格效应相对显著,价格联动性和收益协同性相对较高。此外,在行业分类的内部则存在按上市公司规模以及业绩等方面聚类的特征,但同一类内的股本结构和地区特征在本次研究区间内的关联性并不明显。

(三) 不同类别收益函数波动模式的图形分析

波动是股票市场的基本特征,多样复杂的影响因素引致股票收益的波动模式时刻处于动态变化之中,正确判断股票的类别特征需要深入挖掘并充分利用收益数据蕴含的分类信息。虽然图 7 的函数轨迹静态展示了不同类别股票收益的波动模式差别,但从中并不能获知波动模式在不同时段的动态变化信息。为了从整体上对比不同类别收益函数波动的时段特征,绘制每一类别收益函数在整个研究区间的波动曲面及其对应的变异程度等高线,如图 9 所示①。

由图 9 可以看出,尽管不同类别股票的收益函数均经历了两次以上峰谷转换,但波动的核心信息主要集中于初始阶段。原因在于 2014 年初以融资交易为代表的杠杆资金在短期内助推了股市

① 由于第 2 类股票仅包含“东方证券”(600958)一只股票,不存在类别内部不同股票之间的收益波动差异,因此图 9 仅展示另外 4 类股票在不同时段的收益波动信息。

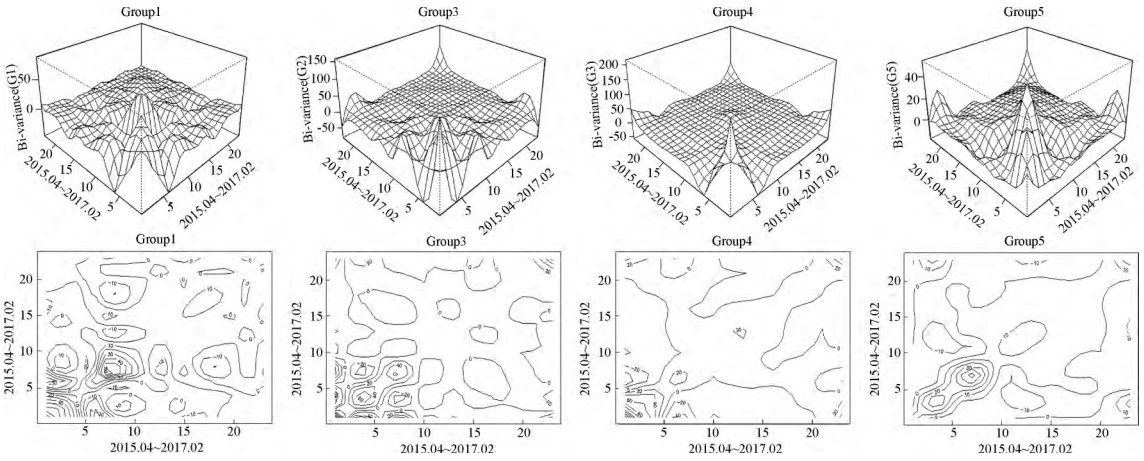
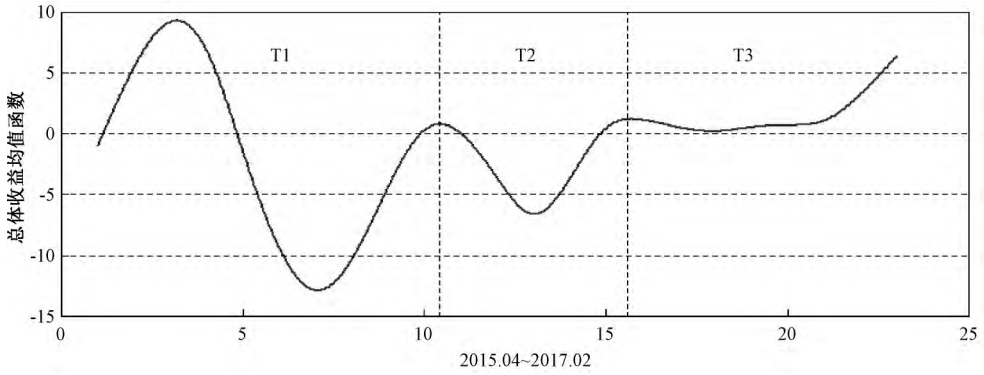


图 9　不同类别收益函数波动曲面及其信息等高线

的暴涨，但市场泡沫的杠杆比例前期并未使监管层意识到股市风险。在去杠杆阶段，卖空限制下上市公司为了躲避股价下跌选择临时停牌，引发流动性危机导致股价暴跌，高杠杆配资的市场结构则进一步加剧价格下跌。此外，在本轮股市出现的大幅波动中，程序化交易导致群体性行为，产生的信息不对称加剧了异质性投资者的市场恐慌情绪，使得个股的波动幅度、震荡时间跨度及政府的救市政策影响均差异较大。为了对比不同类别股票收益的时段波动差异，将整个研究区间按股灾前后进行周期性时段划分，如图 10 所示。在不同时段绘制总体和不同类别收益函数势能转换的相平面，如图 11 ~ 13 所示。



注：图中竖直的两条虚线是对波动周期的划分，分别对应 $x = 11$、$x = 16$。

图 10　波动函数均值的周期性时段划分

以总体函数均值为基准，对比 5 类收益函数的势能转换轨迹发现，除了纵向振幅和横向相位差异外，第 1、3、4、5 类收益函数的势能转换规律基本与总体函数保持一致，但第 2 类收益函数势能转换具有特殊性。具体来看，T1 时段尽管 A 股市场大盘收益正增长，但以第 1、3、4、5 类股票为代表的绝大多数股票的收益增长加速度为负，说明市场虽然繁荣但增长缺乏后劲，泡沫化危机开始形成。此时政策层面应提示杠杆牛市风险，强化信息披露以防止投资者的乐观情绪被过度放大。另外，在 T1 时段内第 1、2 类股票收益的振幅和相位相对较大，说明两类上市公司短时间内针对股市行情进行大幅调整的效率相对较高，这与两类公司的股本结构密切相关。T2 时段内，5 类股票收益均规则地遍历了一个完整周期，并且绝大多数股票的势能转化规律与市场整体保持一致，说明尽管政府稳定市场的政策影响存在滞后，但对上市公司和投资者均有较好的导向作用。因此，加强对金融创新风险源的认知以消除监管的时滞和低效，是金融市场发展不可漠视的重要因素。T3 时段

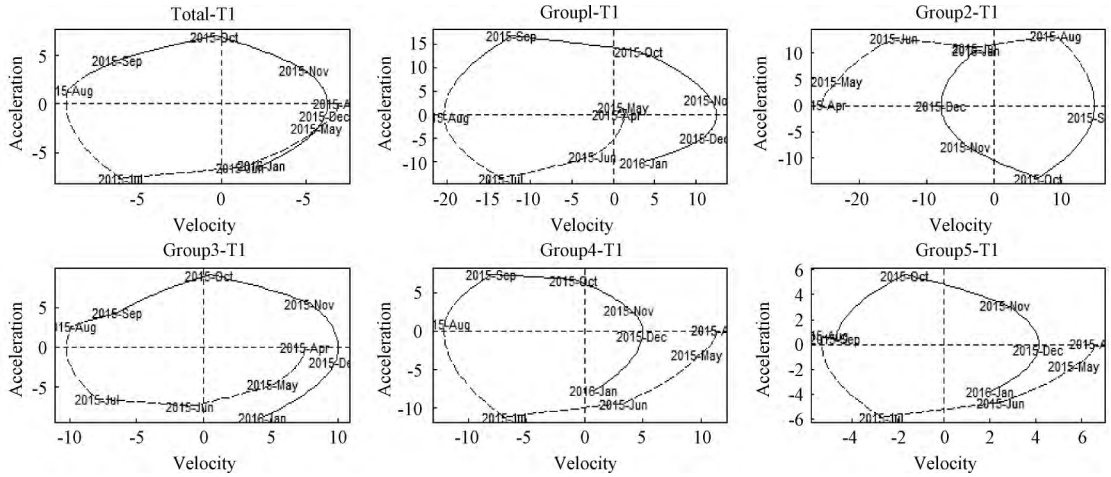


图 11　T1 时段波动函数平均势能转换

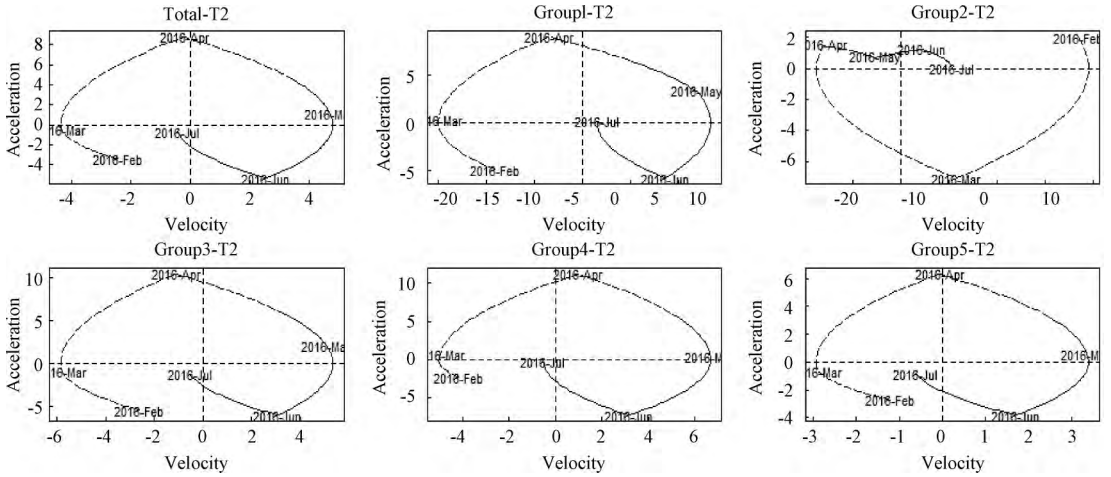


图 12　T2 时段波动函数平均势能转换

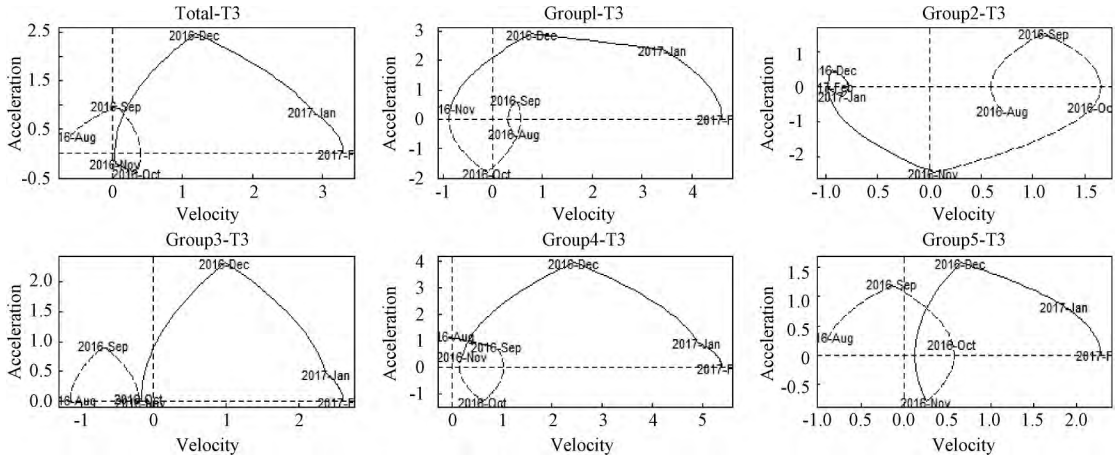


图 13　T3 时段波动函数平均势能转换

内，尽管 5 类股票收益轨迹差异较大，但振幅和相位都显著变小，说明政府一系列稳定市场的措施使得股市回归于正常的波动范围内，各类股票能够依据市场信号按其特有模式理性调整。截止 2017 年 2 月，除第 2 类股票收益呈下降趋势外，其余股票收益的增长速度为正，这与图 7 中 5 类股

票收益曲线的延展趋势以及后续股票的实际收益保持一致。

四、结论与展望

秉承金融大数据分析思想，本文将函数型建模思路引入股票收益的波动模式识别。以上证综指样本股的收益数据为实证对象，通过信息驱动自适应选择最优粗糙惩罚平滑参数，重构了隐含在离散收益数据中的本征收益函数。基于函数型主成分降维，将收益波动的主导趋势进行了无核心信息损失的分解；通过函数型自适应聚类分析将股票收益的波动模式进行客观分类，并可视化动态剖析了不同类别股票收益波动的模式差异。本文研究结论如下：

第一，上证综指样本股票收益的主导波动趋势可以大致分解为四个子模式，四个子模式分别刻画波动趋势的波峰、波谷、起始和结尾阶段，并且波峰和波谷的信息含量相对较大，即收益函数波动模式的类别差异主要集中于峰谷阶段。

第二，上证综指 5 0 只股票收益的波动模式存在显著的类别差异，可以总结为大起大落、触底反弹、楔形整理、W 形筑底、深 V 脉冲等五类。行业层面上，银行业和制造业股票内部风格结构稳定，股票收益联动性较强；相对股本结构和地区归属等因素，行业分类内部存在明显的按公司规模以及业绩方面聚类的子结构。

第三，收益波动模式的动态变化和势能转换规律具有显著的时段特征，并且类别之间波动模式的差异主要体现在本次研究区间的初始峰谷阶段，其原因是不同股票应对 A 股市场的大涨大跌行情存在策略和时滞差异。鉴于市场稳定政策的正向导向作用，政府应着力提升对风险的敏感度，消除监管的滞后性。

需要说明的是，本文研究时段是 2015 年 4 月至 2017 年 2 月，相对上证综指的成立时间来说是较短的样本区间，并且这一区间内 A 股市场经历了多次大涨（2015 年 5 月 19 日至 22 日）和暴跌（2015 年 6 月 19 日和 22 日、2015 年 7 月 4 日和 27 日）的极端事件，使得股票的分类特征和类别内部的风格结构稳定性有待进一步提升。作为后续研究，可以结合突变结构理论预测趋势的转变时机，确定做顶和做底的点位，在不同趋势阶段考察股票收益的波动模式。

参考文献

[1] Ramsay J O, Ramsey J B. Functional Data Analysis of the Dynamics of the Monthly Index of Nondurable Goods Production [J]. Journal of Econometrics, 2002, 107(1/2): 327 - 344.

[2] 严明义. 网上拍卖竞买者出价水平的动态演变模式研究 [J]. 统计研究, 2010, 27(3): 59 - 65.

[3] Tsay R S. Some Methods for Analyzing Big Dependent Data [J]. Journal of Business & Economic Statistics, 2016, 34(4): 1 - 47.

[4] 严明义. 函数性数据的统计分析：思想、方法和应用 [J]. 统计研究, 2007, 24(2): 87 - 94.

[5] Hong M. Potential Applications of Function Data Analysis in High-frequency Financial Research [J]. Journal of Business & Financial Affairs, 2013(2): 23 - 25.

[6] Müller H G, Sen R, Stadtmüller U. Functional Data Analysis for Volatility [J]. Journal of Econometrics, 2011, 165(2): 233 - 245.

[7] Liebl D. Modeling and Forecasting Electricity Spot Prices: A Functional Data Perspective [J]. Annals of Applied Statistics, 2013, 7(3): 1562 - 1592.

[8] Kokoszka P, Reimherr M. Predictability of Shapes of Intraday Price Curves [J]. Econometrics Journal, 2013, 16(3): 285 - 308.

[9] Shang H L. Functional Time Series Approach for Forecasting very Short-term Electricity Demand [J]. Journal of Applied Statistics, 2013, 40(1): 152 - 168.

[10] Shang H L. Forecasting Intraday S&P500 Index Returns: A Functional Time Series Approach [J]. Journal of Forecasting, 2017, 36(7): 741 - 755.

[11] Kokoszka P, Young G. KPSS Test for Functional time Series [J]. Statistics A Journal of Theoretical & Applied Statistics, 2016, 50(5): 957 - 973.

[12] Kokoszka P, Miao H, Zhang X. Functional Dynamic Factor Model for Intraday Price Curves [J]. Journal of Financial Econometrics, 2015, 13(2): 456 - 477.
[13] 岳敏,朱建平. 基于函数型主成分的中国股市波动研究 [J]. 统计与信息论坛, 2009, 24(3): 52 - 56.
[14] 郭均鹏, 孙钦堂, 李汶华. Shibor 市场中各期限利率波动模式分析—基于 FPCA 方法 [J]. 系统工程, 2012(12): 88 - 92.
[15] Wang Z, Sun Y, Li P. Functional Principal Components Analysis of Shanghai Stock Exchange 50 Index [J]. Discrete Dynamics in Nature and Society, 2014(3): 1 - 7.
[16] 王德青,等. 函数型数据聚类分析研究综述与展望 [J]. 数理统计与管理, 2018, 37(1): 51 - 63.
[17] Wang J L, Chiou J M, Mueller H G. Review of Functional Data Analysis [J]. Statistics, 2015(1): 1 - 41.
[18] 吴晓求. 大国金融中的中国资本市场 [J]. 金融论坛, 2015(5): 28 - 35.
[19] Kokoszka P, Reimherr M. Introduction to Functional Data Analysis [M]. London: Chapman and Hall/CRC Press, 2017.
[20] Ramsay J, Hooker G, Graves S. Functional Data Analysis with R and MATLAB [M]. NewYork: Springer Press, 2009.
[21] Chiou J M, Li P L. Functional Clustering and Identifying Substructures of Longitudinal Data [J]. Journal of the Royal Statistical Society, 2007, 69(4): 679 - 699.
[22] Jacques J, Preda C. Functional Data Clustering: A Survey [J]. Advances in Data Analysis & Classification, 2014, 8(3): 231 - 255.
[23] 王德青, 刘晓葳,朱建平. 基于自适应迭代更新的函数型数据聚类方法研究 [J]. 统计研究,2015,32(4): 91 - 96.
[24] Cuevas A, Febrero M, Fraiman R. An Anova Test for Functional Data [J]. Computational Statistics & Data Analysis, 2004, 47(1): 111 - 122.
[25] Zhang J T, Liang X. One-Way Anova for Functional Data via Globalizing the Point-wise F -test [J]. Scandinavian Journal of Statistics, 2014, 41(1): 51 - 71.
[26] 吴晓求. 股市危机: 结构缺陷与规制改革 [J]. 财贸经济, 2016, 37(1): 22 - 32.

作者简介

王德青,男,2014 年毕业于厦门大学,获经济统计学博士学位,现为中国矿业大学管理学院副教授、金融统计学硕士生导师、管理科学与工程博士后,厦门大学数据挖掘研究中心研究人员。研究方向为函数型数据挖掘、金融大数据分析。

何凌云,女,2008 年毕业于中国矿业大学,获管理学博士学位,北卡罗来纳大学夏洛特分校和南京大学商学院高级访问学者,现为中国矿业大学管理学院教授,江苏省能源经济与管理研究基地与国际能源政策研究中心副主任,应用经济学一级学科负责人,金融学统计学硕士生导师。研究方向为能源与环境金融。

朱建平,男,2003 年毕业于南开大学,获理学博士学位,现为厦门大学管理学院教授、博士生导师,厦门大学数据挖掘研究中心主任,中国统计学会副会长,教育部高等学校统计学类专业教学指导委员会秘书长,中国统计教育学会常务理事。研究方向为数理统计、数据挖掘。

(责任编辑: 郭明英)