定量研究

专题报告

证券研究报告

兴业证券 INDUSTRIAL SECURITIES

分析师：
徐寅
xuyinsh@xyzq.com.cn
S0190514070004

西学东渐--海外文献推荐系列之一百二十四

2021 年 7 月 15 日

报告关键点

本文重点对市场状态预测进行研究。市场常常会呈现出不同的状态，但是对市场状态的识别与划分并没有统一的明确标准；同时在固定的划分方式下，对未来市场状态的预测也是一项重要的议题。本文基于 ICC 聚类算法，对市场状态进行识别和预测。首先，本文基于 ICC 聚类算法，将 100 只股票的收益率序列作为输入变量，对历史上的市场状态划分为两类，分别定为牛市状态和熊市状态。从股票在不同状态下的平均收益来看，该方法具有较好的聚类效果。之后，本文基于 ICC 聚类算法，对未来的市场状态进行预测。结果表明，本文方法对于未来市场状态的预测具有较高的准确率。

相关报告

《西学东渐--海外文献推荐系列之一百二十三》2021-7-1
《西学东渐--海外文献推荐系列之一百二十二》2021-6-17
《西学东渐--海外文献推荐系列之一百二十一》2021-6-3

投资要点

- 西学东渐，是指从明朝末年到近代，西方学术思想向中国传播的历史过程。西学东渐不仅推动了中国在科学技术和思想文化方面的发展，也有力地促进了社会与政治的大变革。在今天，西学东渐仍有其重要的现实意义。作为 A 股市场上以量化投资为研究方向的卖方金融工程团队，在平日的工作中，常常深感海外相关领域的研究水平之高、内容之新。而这也促使我们通过大量的材料阅读，去粗取精，将认为最有价值的海外文献呈现在您的面前！
- 本文重点对市场状态预测进行研究。市场常常会呈现出不同的状态，比如牛市状态和熊市状态，但是对市场状态的识别与划分并没有统一的明确标准；同时在固定的划分方式下，对未来市场状态的预测也是一项重要的议题。本文基于 ICC 聚类算法，对市场状态进行识别和预测。首先，本文基于 ICC 聚类算法（Inverse Covariance Clustering，逆协方差聚类），将 100 只股票的收益率序列作为输入变量，对历史上的市场状态划分为两类，分别定为牛市状态和熊市状态。从股票在不同状态下的平均收益来看，该方法具有较好的聚类效果。之后，本文基于 ICC 聚类算法，对未来的市场状态进行预测，具体来讲，输入变量为过去 24 日的股票收益率，输出结果为下一个交易日的市场状态（牛市状态或熊市状态）。结果表明，本文方法对于未来市场状态的预测具有较高的准确率。本文的方法实际上是利用收益率时间序列数据进行了技术面的市场状态预测，没有借助宏观变量、情绪变量等，方法较为纯粹，为技术面择时的投资者提供了新的思路。

风险提示：文献中的结果均由相应作者通过历史数据统计、建模和测算完成，在政策、市场环境发生变化时模型存在失效的风险。

目录

报告正文

基于 ICC 聚类算法的市场状态预测模型

文献来源：

Pier Procacci, Tomaso Aste，Forecasting Market States. Quantitative Finance, 2019, 19(9): 1491-1498. https://doi.org/10.1080/14697688.2019.1622313

推荐原因：

市场常常会呈现出不同的状态，比如牛市状态和熊市状态，但是对市场状态的识别与划分并没有统一的明确标准；同时在固定的划分方式下，对未来市场状态的预测也是一项重要的议题。本文基于 ICC 聚类算法，对市场状态进行识别和预测。首先，本文基于 ICC 聚类算法（Inverse Covariance Clustering，逆协方差聚类），将 100 只股票的收益率序列作为输入变量，对历史上的市场状态划分为两类，分别定为牛市状态和熊市状态。从股票在不同状态下的平均收益来看，该方法具有较好的聚类效果。之后，本文基于 ICC 聚类算法，对未来的市场状态进行预测，具体来讲，输入变量为过去 24 日的股票收益率，输出结果为下一个交易日的市场状态（牛市状态或熊市状态）。结果表明，本文方法对于未来市场状态的预测具有较高的准确率。本文的方法实际上是利用收益率时间序列数据进行了技术面的市场状态预测，没有借助宏观变量、情绪变量等，方法较为纯粹，为技术面择时的投资者提供了新的思路。

我们的思考：

本文方法的本质是用无监督学习的方式对市场状态进行划分与预测；之所以采取无监督学习的方式是因为牛市、熊市状态的定义本身较为模糊，常见的采用指数收益定义市场状态的方式没有具体考虑每一只股票的收益情况以及不同股票之间的相关性，具有一定的不足。本文的方法具有较强的可拓展性，市场状态其实也可以聚类为三种或更多种。另外本文的方法在计算上非常高效，可以应用于大量资产。本文对市场择时的投资者具有较强的借鉴意义。

1、引言

市场并不总是呈现出同一种状态，而是存在着价格更有可能上涨的“牛市”时期和价格更有可能下跌的“熊市”时期。这些不同的市场状态通常被是由一系列宏观经济变量、市场变量和情绪变量所决定的。本文提出了一种新的方法来定义、分析和预测市场状态。

众多学术文献曾提出过各种时间序列模型，以试图捕捉不同的市场状态。其中较受欢迎的方法是 TAR 模型（Tong，1978）和马尔可夫转换模型（Hamilton，1989）：TAR 模型试图在时间序列过程中估计“结构性断点”；马尔可夫转换模型

通常根据马尔可夫链进行建模，利用建模得到的状态变量来反映市场状态的变化。然而，TAR 模型在应用过程中存在一定问题：当经济时间序列发生结构性中断时，TAR 模型不能确定地建立起来；而且对重大经济事件的先验知识可能导致推断偏差（Campbell 等，1997）。而马尔可夫转换模型则面临着维数灾难的问题：特别是对于更复杂的动态变化（Hamilton，1989），则需要依赖变分推断方法或 MCMC 方法（Tsay，2005；Kim 和 Nelson，1999）。这意味着在多元变量的情况下，尤其是从相关性结构中提取切换信息时，将难以进行相应的估计。

其他方法侧重于将观察结果聚类成组：根据某些比较标准来发现“相似的”数据对象。时间序列聚类的研究主要分为两类：子序列聚类和点聚类。子序列聚类涉及数据滑动窗口的聚类，通常以发现重复模式为目标，例如动态时间弯曲方法（Liao，2005）、层次方法（Neville-Manning 和 Witten，1997）或模式发现方法（Ren 等，2017）。相反，在点聚类方法中，在各个时间点 t 上的多元观测值都被分类到相应的聚类中。最常见的方法是通过距离度量来实现的（Grabarnik 和 Särkkä，2001；Focardi 和 Fabozzi，2004；Zolhavarih 等，2014；Hendricks 等，2016；Hallac 等，2016）。

在多元的情况下，市场的不同状态不仅反映在收益和损失上，也反映在价格的相对动态上。事实上，牛市和熊市存在的相关性结构变化表明这些市场状态存在结构性差异。为方便起见，业内最常见的方法是假设相关性结构是平稳的（Black 和 Litterman，1992；Duffie 和 Pan，1997）。然而，股票之间的相关性不是恒定不变的，而是会随时间变化而变化，（Lin 等，1994 ；Ang 和 Bekaert，2002；Musmeci 等，2016），其在市场剧烈波动期间大幅增加，尤其是在市场整体下行时（Ang 和 Chen，2002；Cizeau 等，2010；Schmitt 等，2013）。事实上，相关文献提出了一些方法来对时变相关性进行建模和预测。例如，Bollerslev (1990) 的广义自回归条件异方差模型 (GARCH 模型) 或 Engle (2002) 的动态条件相关模型 (DCC 模型)。然而，由于维数灾难的问题，这些模型通常无法应用于多个资产：随变量数量增加，参数数量会呈超线性增加（Danielsson，2011）。其他方法根据滚动窗口计算时变相关矩阵，然后聚焦于时变相关矩阵的变化，例如 RiskMetrics 等估计量（Longerstaey 和 Spencer，1996；Lee 和 Stevenson，2003）。然而，这些方法只使用小部分数据，这会导致估计量有较大差异；并且在高维情况下，可能会导致不确定的估计（Laloux 等，1999）。

Hallac 等人（2017）提出了一种名为 TICC（Toeplitz Inverse Covariance Clustering，即 Toeplitz 逆协方差聚类）的聚类算法，该算法最初是针对电动汽车提出的，通过与参考稀疏精度矩阵（逆协方差矩阵）相关的似然测度构建状态分类。然而，他们的方法并没有孤立地考虑每一个观测对象，而是将较短的子序列进行了聚类，使得在子序列上构造的协方差矩阵能够提供跨时间偏相关的信息。在这种情况下，通过对每个状态的精度矩阵添加 Toeplitz 约束，将跨时间偏相关系数约束为常数，从而实现协方差的平稳性。尽管作者考虑的数据结构与金融中的嘈杂数据有所不同，但是从金融学的角度来看，该方法仍有许多吸引人的特点。

本文在 Hallac 等人（2017）的基础上提出了一种类似的基于协方差的聚类方法。然而，我们只考虑单一的观测值，并没有在精度矩阵上强制使用 Toeplitz 结

构。因此，我们称这种方法为 **ICC（Inverse Covariance Clustering，逆协方差聚类）**。与 Hallac 等人（2017）类似，我们对市场状态之间的频繁切换进行惩罚，以此来加强时间上的一致性。与 Hallac 等人方法的另一个区别是，我们没有直接最大化似然度，而是根据马氏距离将状态分类（De Maess- chalck 等，2000）。我们在金融时间序列的背景下实证了这种方法，并对稀疏性和时间一致性所起的作用进行了详细的分析，同时评估了聚类的重要性。最后，我们证明了聚类分类可以用于样本外的一步预测。

我们的方法简化并阐明了“市场状态”的定义，通过一个稀疏精度矩阵和一个期望值向量来识别每个状态，其中期望值向量与一组多元观测数据有关。下文中，用J_k表示市场状态“k”的精度矩阵，该矩阵代表了系统变量之间的偏相关结构。在多元正态情况下，当且仅当J_k对应的元素等于零时，两个节点条件独立。稀疏精确矩阵提供了一个易于解释的、直观的市场状态结构，其中，该市场状态的依赖关系直接在一个稀疏网络中相互关联。此外，稀疏性将参数的数量从n^2阶（变量的数量为 n）减少到 n 阶，以防止过拟合（Lauritzen，1996），并过滤掉了噪声相关性（Barfuss 等，2016；Musmeci 等，2017）。

在分类过程中，首先设定聚类的数量 K（在本文中我们设定 K=2），然后将多元观测对象随机分配到不同聚类中。从这 K 组数据中，我们计算样本均值μ_k和精度矩阵J_k，然后我们迭代地将样本重新分给最小的聚类。以下是方程 1:

$$M_{t,k} = d_{t.k}^2 + \gamma \, \mathbf{1}\{K_{t-1} \neq k\}$$

其中$X_t = [x_{t,1}, x_{t,2}, \dots, x_{t,n}]$是在时间 t（=1, …, T）的 n 个股票的多元观测值；μ_k是聚类 k 均值的向量；J_k是聚类 k 的（稀疏）精度矩阵；$d_{t,k}^2 = (X_t - \mu_k)^T J_k (X_t - \mu_k)$是聚类 k 中观测对象$X_t$相对于聚类质心$\mu_k$的马氏距离（Mahalanobis distance）的平方；γ 是惩罚系数，用于对状态切换进行惩罚；K_{t-1}是在 t-1 时刻观测对象所属的聚类。我们也考虑了使用最大似然或最小欧氏距离的聚类，但本文只呈现了马氏距离的方法，其结果相对更好。具体来说，欧氏距离在区分正收益和负收益方面非常有效，但不能很好地区分金融危机前和金融危机后的时期；最大似然法较好地确定了危机时期，但它在区分“牛市”和“熊市”市场状态时不那么清晰。需要注意的是，使用的马氏距离聚类产生了很高的似然度，但不是最大的。

通过 Viterbi 算法（Viterbi，1967；Bishop，2006），能够提高聚类的计算效率，该算法将一个 O(K^T)过程转换为 O(KT)过程。通过 TMFG-LoGo 网络过滤方法，从每个聚类的观测数据中有效地计算稀疏精度矩阵J_k（Massara 等，2015；Barfuss 等，2016）。相对于其他方法（例如 GLASSO 方法，Friedman 等，2008），TMFG-LoGo 方法被证明是更有效和更易实施的，特别是在只有少量数据可用的情况下（Barfuss 等，2016；Aste 和 Di Matteo，2017）。该过程通过一个内部专用的 Python 包实现，这是该方法第一次被应用于金融领域和市场状态分析。

本文测试的股票池为罗素 1000 指数成份股，时间区间为 1995 年 1 月 2 日至 2015 年 12 月 31 日，共涉及 2490 只美股。对于每个资产 i=1,…,n，我们计算了相应的每日对数回报$r_i(t) = \log\big(P_i(t)\big) - \log\big(P_i(t-1)\big)$，其中$P_i(t)$是股票 i 在 t 日的收盘价。

2、聚类

如上所述，本文的主要目标是在控制时间一致性的同时，有效地将有噪声的多元时间序列聚类为不同状态。在第一个实验中，我们考虑了 1995 年 1 月 2 日至 2015 年 12 月 31 日之间的整个数据集，并将其分为两种市场状态。为了探究算法中每个模块的作用，以及其与传统基准方法之间的区别，我们共研究了五个模型：

(a) ICC 模型——稀疏精度矩阵和时间一致性

(b) ICC 模型——全精度矩阵和时间一致性

(c) ICC 模型——稀疏精度矩阵

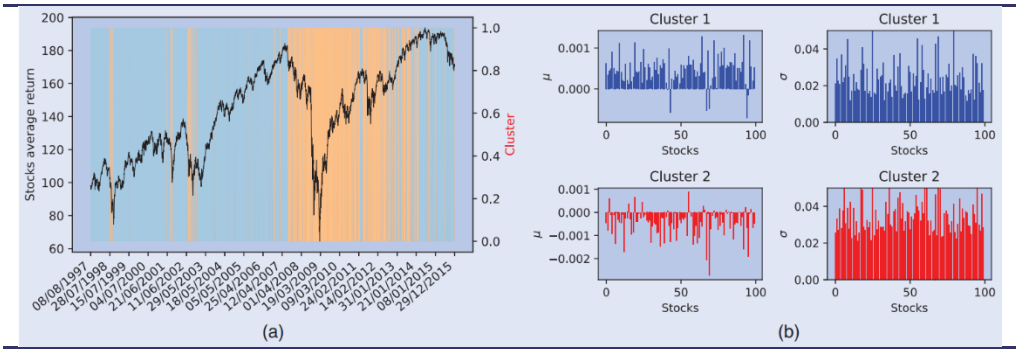
(d) ICC 模型——全精度矩阵

(e) 高斯混合模型——全协方差

模型(a)是本文提出的 ICC 方法。模型(b)使用了全精度矩阵J_k，而不是稀疏矩阵。模型(c)放宽了时间一致性限制，允许方程 1 中的 γ=0。模型(d)中 γ=0 且使用全精度矩阵。最后，模型(e)是一个传统的高斯混合模型（Bishop，2006），由于其与 ICC 方法具有相似性，因此作为基准方法。我们针对通过聚类得到的两个市场状态，通过对市场属性和时间一致性进行分析，比较了生成的两个聚类。**首先，我们从观察期内持续交易的股票中随机选择了 100 只股票作为研究对象。**上述随机选择是为了避免选择偏差。然后，我们利用随机重采样来评估考虑不同股票时的稳健性。

我们通过网格搜索优化了时间一致性参数：在本文的两个实验中，对于 ICC 稀疏模型(a)，γ=16；对于 ICC 全模型(b)，γ=14.7。实验得到的两个参考精度矩阵J_1和J_2有 344 个非零项（依赖网络边），其中 142 个是两种状态共有的，表现出一定的差异程度，但两种市场状态之间也存在显著的重叠。分配给每个聚类的样本数分别为：聚类 1 为 3295，聚类 2 为 1704。图表 1 用不同颜色的背景展示了两个聚类。我们可以观察到有较好的连续性。例如，聚类 1 中平均的连续天数是 25.3 天。我们还注意到，聚类 1（蓝色背景）倾向于与市场价格上涨的时期有关，而聚类 2（橙色背景）在金融危机和市场低迷时期出现得更多。可以发现该方法自动将“牛市”时期（正平均回报）分配给聚类 1，将“熊市”时期（负平均回报）分配给聚类 2。例如，我们可以在图 1(a)中观察到：2001-2002 年互联网泡沫危机期间的 52 个连续观测数据被归入熊市聚类 2，同时 2007-2008 年全球金融危机期间的 211 个连续观测数据也被归入熊市聚类 2。从图 1(b)我们可以观察到，牛市聚类 1 所有股票的平均回报为正数，而熊市聚类 2 的平均回报则为负数。此外，两种聚类的标准差也不同。

图表 1、时间序列聚类结果

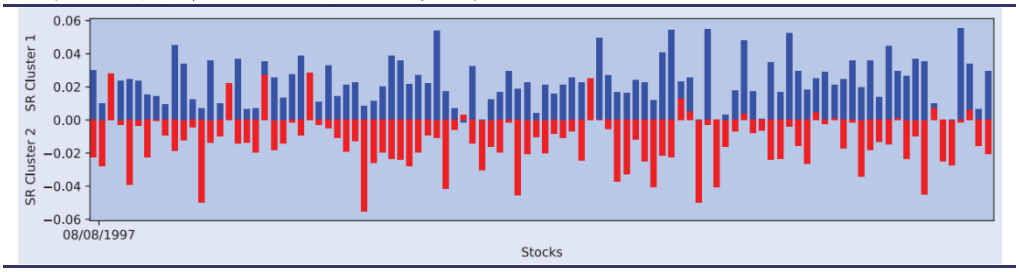


资料来源：Quantitative Finance，兴业证券经济与金融研究院整理

注：子图(a)显示了 100 只股票在各个 t 时刻的累积平均回报；在此图中，蓝色背景表示分为聚类 1 的时间点，而橙色背景表示分为聚类 2 的时间点。子图(b)显示了分别用两个聚类中的收益数据计算的 100 只股票的平均值和标准差。我们观察到，聚类 1 显示出正的平均回报（“牛市”状态）和较低的波动水平，而聚类 2 呈现负的平均回报（“熊市”状态）和较高的波动水平。(a)时间序列分割结果，(b)每个时间聚类的每个聚类的均值（左）和标准差（右）。

为了比较风险调整收益，我们计算了每个聚类中每只股票的夏普比率：牛市聚类的平均年化夏普比率为 1.2，第 5 百分位数和第 95 百分位数分别为 0.84 和 1.78，而熊市聚类的平均夏普比率为-0.96，第 5 百分位数和第 95 百分位数分别为-1.03 和-0.24。因此，这两个聚类有着非常不同的风险收益情况。图表 2 展示了这两个聚类中 100 只股票的夏普比率，其中蓝色柱代表根据聚类 1 中的对数回报计算的夏普比率，而红色柱代表根据聚类 2 中的对数回报计算的夏普比率。为了验证结果的稳健性和通用性，我们对另外一组随机的 100 只股票计算夏普比率，并将该过程重复了 100 次。对于所有重新抽样的股票组合，我们发现均可产生关于牛市和熊市的连续性聚类结果，其中至少 75%的股票在牛市状态下夏普比率大于零，且在熊市状态下夏普比率显著小于零。在 100 次重采样中，两个聚类的平均天数分别为 3451 和 1293。

图表 2、样本中 100 只股票的夏普比率（SR）



资料来源：Quantitative Finance，兴业证券经济与金融研究院整理

为了评估稀疏性和时间一致性的作用，我们对“备选”ICC 模型(b)-(d)和 GMM 模型(e)进行了相同的分析。

图表 3 展示了 100 次重采样中在两个聚类中具有正/负夏普比率的股票数量。在表格中，每一对数据代表的是在牛市中夏普比率为正（左）和在熊市中夏普比率为负（右）的股票数量。我们发现，在没有时间一致性约束的情况下，两种 ICC 模型(c 和 d)均能进行有效分类。然而，当考虑时间一致性约束时，ICC 全模型(b)

受到约束的影响较为显著，分类效果有所减弱，而 ICC 稀疏模型(a)提供了稳健的结果。就收益风险角度而言，GMM 的聚类效果最差。

图表 3、正/负的夏普比率（“牛市”，“熊市”）状态

	Median	5th percentile	95th percentile
GMM	(69,64)	(48,53)	(75,81)
ICC Full, $\gamma = 0$	(77,78)	(67,71)	(92,98)
ICC Sparse, $\gamma = 0$	(85,87)	(69,75)	(96,95)
ICC Full, $\gamma = 14.7$	(73,74)	(68,65)	(78,80)
ICC Sparse, $\gamma = 16$	(75,81)	(65,69)	(86,90)

资料来源：Quantitative Finance，兴业证券经济与金融研究院整理

从时间一致性的角度来说，图表 4 展示了 5 个模型的聚类所产生的切换次数和分段长度。当没有时间一致性约束时（即模型(c)、(d)），ICC 提供了较短的时间一致性结果。当被限制时间一致性时，ICC 全模型(b)在不同样本间呈现出不同的时间一致性，有些在整个周期内只有几次切换，而另一些则有几百次切换。ICC 稀疏模型(a)在整个周期中则有几百次切换，小于 GMM (e)中切换次数的 1/3。

图表 4、时间一致性指标

	Median	5th percentile	95th percentile
		Number of switches	
GMM	785	540	874
ICC Full, $\gamma = 0$	1203	992	2176
ICC Sparse, $\gamma = 0$	1157	727	1421
ICC Full, $\gamma = 14.7$	204	54	306
ICC Sparse, $\gamma = 16$	208	120	298
		Segment length	
GMM	5.07	2.4	11.8
ICC Full, $\gamma = 0$	3.3	1.68	4.38
ICC Sparse, $\gamma = 0$	3.5	2.8	6.65
ICC Full, $\gamma = 14.7$	22.64	14.6	38.26
ICC Sparse, $\gamma = 16$	23.6	18	55.27

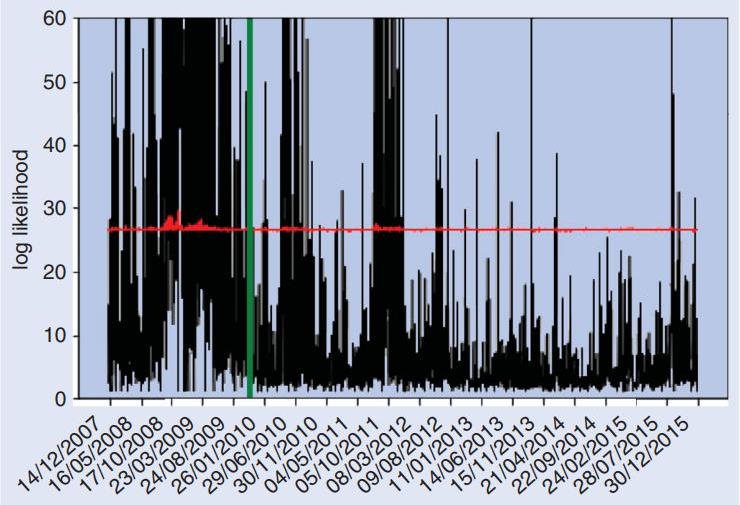
资料来源：Quantitative Finance，兴业证券经济与金融研究院整理

3、稀疏性的作用

在之前的研究中（Barfuss 等，2016），TMFG-LoGo 方法已被证明比其他过滤方法（包括 GLasso 和 Ridge）表现更好，其有效性更强，且具有固定稀疏度级别的优势，无需调整超参数（Massara 等，2015）。在本节中，我们将 TMFG-LoGo 方法与 Ridge l_2惩罚逆协方差方法（岭方法）在数据集上的表现进行比较，从统计显著性的方面评估 TMFG-LoGo 过滤过程。我们将广泛使用的 Ridge 惩罚作为经验逆协方差矩阵的稳健估计，并将其与 TMFG-LoGo 进行比较。结果表明，当应用于本文的数据集时，TMFG-LoGo 相比 Ridge 产生了更稳定的似然结果。我们使用 40%的数据（从 2007 年 12 月 31 日到 2015 年 12 月 31 日）作为测试集，并将测试集之前（2007 年 12 月 30 日之前）的 q 个观测数据作为训练集。通过在训练集内进行交叉验证，定义了 Ridge 惩罚参数。为了比较 TMFG-LoGo 和交叉验证

的 Ridge，我们使用两种协方差估计量计算了对数似然函数$L_{s,k} = 1/2(\log|J_k| - d_{s,k}^2 - plog(2\pi))$，并对它们进行了比较。图表 5 展示了使用 TMFG-Logo 和 Ridge 精度矩阵在 q=500 观测值上估计的训练集和测试集中的似然观测值。绿色竖线是训练集和测试集的分界线。随着时间的推移，TMFG-LoGo 的似然度更稳定，这表明该过程在过滤噪声方面是成功的。

图表 5、使用 TMFG(红线)和 Ridge(黑线)精度矩阵的对数似然值



资料来源：Quantitative Finance，兴业证券经济与金融研究院整理

图表 6 显示了在训练集和测试集中计算的概率的平均值、第 5 百分位数和第 95 百分位数的详细信息。如上所述，TMFG-LoGo 的似然度更稳定，TMFG-LoGo 的第 5 和第 95 百分位数只变化了几个百分点，而 Ridge 的变化超过一个数量级。当考虑不同的 q 值时，我们在 TMFG-LoGo 和 Ridge 中发现了类似的结果。值得注意的是，Ridge 的对数似然度在训练集和测试集上有很大的差异，这是过拟合的典型迹象。相反，TMFG 表现出很小的差异，表明 LoGo 过程充当了拓扑惩罚。

图表 6、TMFG 和 Ridge 对数似然度量在训练集(上)和测试集(下)的平均值、第 5 和第 95 百分位数

	Average	5th percentile	95th percentile
		Train set	
$\mathcal{L}_{\text{Ridge}}$	41.70	2.19	188.85
$\mathcal{L}_{\text{TMFG}}$	26.71	26.53	27.22
		Test set	
$\mathcal{L}_{\text{Ridge}}$	8.08	1.39	27.64
$\mathcal{L}_{\text{TMFG}}$	26.55	26.44	26.73

资料来源：Quantitative Finance，兴业证券经济与金融研究院整理

4、预测

在第二个实验中，基于先前的观察，我们使用本文的方法来预测市场的未来状态。为此，我们使用 65%数量的数据（从 1995 年 1 月 2 日至 2009 年 2 月 5 日）

作为训练集，提取了两个参考精度矩阵和平均值(J_1, μ_1)和(J_2, μ_2)。需要注意的是，这与我们在第一个实验中使用整个数据集不同。然后，我们预测了在给定时间 t 的情况下，接下来 t+h 时间的观测值属于状态 k 的概率。这是通过使用两个聚类的对数似然比（Neyman 和 Pearson，1933）执行逻辑回归来实现的，回归的时间区间为长度为Δ的滚动窗口：

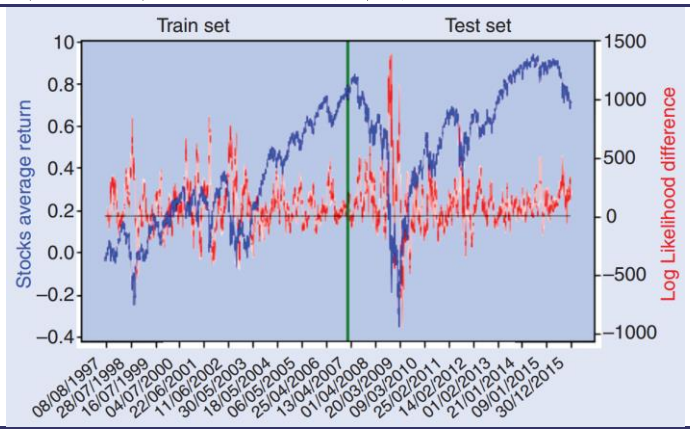
$$R_t = \sum_{s=t-\Delta+1}^{t} L_{s,1} - L_{s,2}$$

其中$L_{s,k} = 1/2(\log|J_k| - d_{s,k}^2 - plog(2\pi))$是观测对象$X_s$与聚类 k=1 或 2 相关联时的对数似然值。在本文的实验中，我们考虑Δ= 24天，因为这是第一次实验中 ICC (a)得到的分段的平均长度。图表 7 提供了每个聚类计算出的似然比的可视化表示。其中蓝线是股票平均回报，红线是对数似然度，绿色竖线表示训练集的结束和测试集的开始，黑色水平线表示回报为零。市场状态K_t相对于对数似然比R_t的逻辑回归可以写成如下方程：

$$P(K_{t+h} = 1,2 \mid R_t = x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

其中参数β_0和β_1是通过极大似然估计得到的（Bishop，2006）。我们通过训练集估计了所有参数（J_1, J_2, μ_1, μ_2, γ, β_0和 β_1），并通过交叉验证估计了训练集中的阈值为 0.54。然后，我们在给定对数似然比$R_t = x$的情况下，使用这些参数在测试集中预测第二天的状态。具体来说，如果$P(K_{t+1} = 1|R_t = x) > 0.54$，我们预测$\widehat{K}_{t+1} = 1$，否则$\widehat{K}_{t+1} = 2$。例如，对于 2010 年 3 月 30 日（测试集），我们预测了一个牛市状态，概率为$P(K_{30-Sep} = 1|R_{29-Mar}) = 0.77$，其中$R_{29-Mar}$使用了 2010 年 3 月 6 日至 2010 年 3 月 29 日（Δ= 24天，全部在测试集内）的观测数据，参数J_k, μ_k, γ, β_0和 β_1是在训练集上（2009 年 4 月 31 日之前）估计得到的数据。

图表 7、训练集和测试集的对数似然比和平均回报



资料来源：Quantitative Finance，兴业证券经济与金融研究院整理

为了评估以上方法的效果，我们将测试集上的预测结果与在第一次实验的整个时间段内的分类结果进行了比较（见图表 1）。我们使用了三个指标（Hastie 等，2008）来评估我们分类方法的效果：真正例率 TPR（True Positive Rate，正确归为

聚类 1 的样本数除以归为聚类 1 的样本总数),真负例率 TNR(True Negative Rate,正确归为聚类 2 的样本数除以归为聚类 2 的样本总数)和准确率 ACC(Accuracy,预测正确的样本数除以样本总数)。为了检验该方法的稳健性,我们对 100 只股票进行了随机重采样,并针对新的数据集进行了分类实验。我们重复这个过程 100 次,并计算了三个评价指标 TPR、TNR 和 ACC,结果展示在图表 8 中。在重采样中获得了良好的 ACC 水平,只有第 5 百分位略低于 50%。TPR 在第 5 百分位时也高于 50%,而 TNR 仅呈现良好的中位数结果,较低的第 5 百分位数值表明很难正确预测。这表明我们的方法有将样本过度分配到聚类 1 (牛市状态) 的倾向。尽管如此,通过使用 Aste 和 Di Matteo (2017) 提出的超几何分布,我们验证了这些 TNR 在 0.01 水平上具有统计显著性,表明对熊市状态具有显著预测能力。需要强调的是,目前的预测没有进行进一步的优化,模型表现可以通过多种方式加以改进。但这超出了本文的初衷:本文更关注于方法的简单性,而不是一味追求效果。

图表 8、使用 ICC 对数似然比作为自变量的样本外性能指标

	Median	5th percentile	95th percentile
TPR	0.68	0.51	0.93
TNR	0.52	0.39	0.78
ACC	0.54	0.47	0.69

资料来源:Quantitative Finance,兴业证券经济与金融研究院整理

为了将上文的结果与基准方法进行对比,我们将时间 t – 1 时收益为正的股票的占比作为自变量,来估计方程 3 中的逻辑回归,从而将本文的 ICC 对数似然与关于相关性结构的信息简化版本进行比较。采用相同的估计方式,通过交叉验证得到阈值为 0.61,结果见图表 9。这种简化信息方法可以提供接近 50%的中位数准确率,但与图表 8 中显示的 ICC 对数似然比的结果相比,该模型总体性能较差。

图表 9、使用收益为正的股票的比例作为自变量的样本外绩效指标

	Median	5th percentile	95th percentile
TPR	0.71	0.67	1
TNR	0.24	0.0	0.77
ACC	0.47	0.38	0.62

资料来源:Quantitative Finance,兴业证券经济与金融研究院整理

5、结论

在本文中,我们提出了一种新的方法来定义、识别、分类和预测市场状态。本文的方法具有准确性、直观性和较强的预测能力,并且能够处理高维数据。我们针对多元非平稳的金融数据集进行了两个实验,结果表明该方法在识别和预测数据集结构方面具有较强的有效性。实验使用了两个聚类和 100 个变量,我们也证实了类似的结果可以适用于更多或更少的变量;当使用三个或更多的聚类时,

也会出现有趣的分类结果，本文选择两组只是为了简化问题。本文为使用多元分析预测股市回报的方法提供了新思路，也极大地简化了对“牛市”和“熊市”的解释。当然，在现实中，不止有两种市场状态，而且牛市和熊市的定义常常是模糊的。本文没有试图优化结果，而是更关注于简单性和可解释性，因此能够提供一个较大的开放性领域来完善本文的方法。我们还采用了几个可以在未来研究中修改的备选方法。例如，马氏距离是一个有效的方法；欧几里德距离和似然实验也可以产生不错的效果。此外，TMFG 网络、其他信息过滤网络、其他稀疏化方法也是可以采用的方法。时间一致性也可以使用隐式马尔可夫方法来实行。通过逻辑回归来预测市场状态只是众多回归方法中简单的一个，本文选择它的原因是，它可能可以更好地利用我们状态结构的信息内容。所有这些方法的选择都是出于简单和直观的动机，因为本文方法的主要成就之一是计算效率，允许我们将该方法应用于高维数据集。进一步的研究将包括新的信息来源（如新闻、经济指标、情绪等）。

参考文献

[1] Ang, A.和Bekaert, G., International asset allocation with timevarying correlations. Rev. Financial Stud., 2002, 15, 1137–1187.

[2] Ang, A.和Chen, J., Asymmetric correlations of equity portfolios. J. Financ. Econ., 2002, 63, 443–494.

[3] Aste, T.和Di Matteo, T., Causality network retrieval from short time series. ArXiv preprint arXiv:1706.01954, 2017.

[4] Barfuss, W., Massara, G.P., di Matteo, T.和Aste, T., Parsimonious modeling with information filtering networks. Phys. Rev. E, 2016, 94, 062306.

[5] Bishop, C.M., Pattern Recognition和Machine Learning, 2006 (Springer-Verlag: New York).

[6]Black, F.和Litterman, R., Global portfolio optimization. Financ. Anal. J., 1992, 48, 28–43.

[7] Bollerslev, T., Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. Rev. Econ. Stat., 1990, 4, 498–505.

[8] Campbell, J.Y., Lo, A.W.和MacKinlay, A.C., The Econometrics of Financial Markets, 1997 (Princeton University Press: Princeton, NJ).

[9] Cizeau, P., Potters, M.和Bouchaud, J.P., Correlation structure of extreme stock returns. Quant. Finance, 2010, 1, 217–222.

[10] Danielsson, J., Financial Risk Forecasting: The Theory和Practice of Forecasting Market Risk with Implementation in R和Matlab, 2011 (Wiley-Blackwell: Hoboken).

[11] De Maesschalck, R., Jouan-Rimbaud, D.和Massart, D.L., The Mahalanobis distance. Chemometr. Intell. Lab. Syst., 2000, 50, 1–18.

[12] Duffie, D.和Pan, J., An overview of value at risk. J. Derivatives, 1997, 4, 7–49.

[13]Engle, R., Dynamic conditional correlation. J. Bus. Econ. Stat., 2002, 20, 339–350.

[14] Focardi, S.M.和Fabozzi, F.J., A methodology for index tracking based on time-series clustering. Quant. Finance, 2004, 4, 417–425.

[15] Friedman, J., Hastie, T.和Tibshirani, R.,稀疏模型inverse covariance estimation with the graphical lasso. Biostatistics, 2008, 9, 432–441.

[16] Grabarnik, P.和Särkkä, A., Interacting neighbour point processes: Some models for clustering. J. Stat. Comput. Simul., 2001, 68, 103–125.

[17] Hallac, D., Nystrup, P.和Boyd, S., Greedy Gaussian segmentation of multivariate time series. ArXiv e-prints, 2016.

[18] Hallac, D., Vare, S., Boyd, S.P.和Leskovec, J., Toeplitz inverse covariance-based clustering of multivariate time series data. CoRR, abs/1706.03161, 2017.
[19] Hamilton, J.D., A new approach to the economic analysis of nonstationary time series和the business cycle. Econometrica, 1989, 57, 357–384.
[20] Hastie, T., Tibshirani, R.和Friedman, J., The Elements of Statistical Learning, 2008 (Springer: New York).
[21] Hendricks, D., Gebbie, T.和Wilcox, D., Detecting intraday financial market states using temporal clustering. Quant. Finance, 2016, 16, 1657–1678.
[22] Kim, C.J.和Nelson, C., State-Space Models with Regime Switching: Classical和Gibbs-Sampling Approaches with Applications, 1999 (The MIT Press: Cambridge, MA).
[23] Laloux, L., Cizeau, P., Bouchaud, J.P.和Potters, M., Noise dressing of financial correlation matrices. Phys. Rev. Lett., 1999, 83, 1467–1470.
[24] Lauritzen, S.L., Graphical Models, 1996 (Oxford University Press: Oxford).
[25] Lee, S.和Stevenson, S., Time weighted portfolio optimisation. J. Prop. Investment Finance, 2003, 21, 233–249.
[26] Liao, W.T., Clustering of time series data – a survey. Pattern Recogn., 2005, 38, 1857–1874.
[27] Lin, W.L., Engle, R.和Ito, T., Do bulls和bears move across borders? International transmission of stock returns和volatility. Rev. Financ. Stud., 1994, 7, 507–38.
[28] Longerstaey, J.和Spencer, M., RiskMetrics Technical Document, J.P. Morgan/Reuters, 1996.
[29] Massara, G.P., di Matteo, T.和Aste, T., Network filtering for big data: Triangulated maximally filtered graph. CoRR, abs/1505.02445, 2015.
[30] Musmeci, N., Aste, T.和Di Matteo, T., What does past correlation structure tell us about the future? An answer from network filtering. ArXiv e-prints, 2016.
[31] Musmeci, N., Nicosia, V., Aste, T., Di Matteo, T.和Latora, V., The multiplex dependency structure of financial markets. Complexity, 2017, 2017, 13.
[32] Nevill-Manning, C.G.和Witten, I.H., Identifying hierarchical structure in sequences: A linear-time algorithm. J. Artif. Intell. Res., 1997, 7, 67–82.
[33] Neyman, J.和Pearson, E.S., IX. On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. London A: Math. Phys. Eng. Sci., 1933, 231, 289–337.
[34] Ren, L., Wei, Y., Cui, J.和Du, Y., A sliding window-based multistage clustering和probabilistic forecasting approach for large multivariate time series data. J. Stat. Comput. Simul., 2017, 87, 2494–2508.
[35] Schmitt, T.A., Chetalova, D., Schäfer, R.和Guhr, T., Nonstationarity in financial time series: Generic features和tail behavior. EPL (Europhys. Lett.), 2013, 103, 58003.
[36] Sharpe, W.F., Mutual fund performance. J. Bus., 1966, 39, 119–138.
[37] Sharpe, W.F., The Sharpe ratio. J. Portf. Manag., 1994, 21, 49–58.
[38] Tong, H., On a threshold model. In Pattern Recognition和Signal Processing, edited by C. H. Chen, 1978 (Sijthoff和Noordhoff: Amsterdam).
[39] Tsay, R., Analysis of Financial Time Series, 2nd ed., 2005 (Wiley: Hoboken, NJ).
[40] Viterbi, A., Error bounds for convolutional codes和an asymptotically optimum decoding algorithm. IEEE Trans. Inf. Theory, 1967, 13, 260–269.
[41] Zolhavarieh, S., Aghabozorgi, S.和Wah Teh, Y., A review of subsequent time series clustering. Sci. World J., 2014, 2014, 312521.

风险提示：文献中的结果均由相应作者通过历史数据统计、建模和测算完成，在政策、市场环境发生变化时模型存在失效的风险。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

投资评级说明

投资建议的评级标准	类别	评级	说明
报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后的12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅。其中：A股市场以上证综指或深圳成指为基准，香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	相对同期相关证券市场代表性指数涨幅大于15%
		审慎增持	相对同期相关证券市场代表性指数涨幅在5%～15%之间
		中性	相对同期相关证券市场代表性指数涨幅在-5%～5%之间
		减持	相对同期相关证券市场代表性指数涨幅小于-5%
		无评级	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级
	行业评级	推荐	相对表现优于同期相关证券市场代表性指数
		中性	相对表现与同期相关证券市场代表性指数持平
		回避	相对表现弱于同期相关证券市场代表性指数

信息披露

本公司在知晓的范围内履行信息披露义务。客户可登录 www.xyzq.com.cn 内幕交易防控栏内查询静默期安排和关联公司持股情况。

使用本研究报告的风险提示及法律声明

兴业证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

，本公司不会因接收人收到本报告而视其为客户。本报告中的信息、意见等均仅供客户参考，不构成所述证券买卖的出价或征价邀请或要约。该等信息、意见并未考虑到获取本报告人员的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本报告所载资料的来源被认为是可靠的，但本公司不保证其准确性或完整性，也不保证所包含的信息和建议不会发生任何变更。本公司并不对使用本报告所包含的材料产生的任何直接或间接损失或与此相关的其他任何损失承担任何责任。

本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌，过往表现不应作为日后的表现依据；在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告；本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

除非另行说明，本报告中所引用的关于业绩的数据代表过往表现。过往的业绩表现亦不应作为日后回报的预示。我们不承诺也不保证，任何所预示的回报会得以实现。分析中所做的回报预测可能是基于相应的假设。任何假设的变化可能会显著地影响所预测的回报。

本公司的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

本报告并非针对或意图发送予或为任何就发送、发布、可得到或使用此报告而使兴业证券股份有限公司及其关联子公司等违反当地的法律或法规或可致使兴业证券股份有限公司受制于相关法律或法规的任何地区、国家或其他管辖区域的公民或居民，包括但不限于美国及美国公民（1934 年美国《证券交易所》第 15a-6 条例定义为本「主要美国机构投资者」除外）。

特别声明

在法律许可的情况下，兴业证券股份有限公司可能会利差本报告中提及公司所发行的证券头寸并进行交易，也可能为这些公司提供或争取提供投资银行业务服务。因此，投资者应当考虑到兴业证券股份有限公司及/或其相关人员可能存在影响本报告观点客观性的潜在利益冲突。投资者请勿将本报告视为投资或其他决定的唯一信赖依据。

兴业证券研究

上海	北京	深圳
地址：上海浦东新区长柳路36号兴业证券大厦15层	地址：北京西城区锦什坊街35号北楼601-605	地址：深圳市福田区皇岗路5001号深业上城T2座52楼
邮编：200135	邮编：100033	邮编：518035
邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn	邮箱：research@xyzq.com.cn