摘 要

配对交易是一种在投资者中十分流行的交易策略。投资的基本原则是低买高卖，但是人们往往很难对某个资产的内在价值做出精确的估计。配对交易把关注的焦点放在配对中的一项资产相对于另一项资产是否存在明显高估或低估的情形，从而规避了对内在价值的估计。参与配对的各个资产的价格从长期来看应该保持一种均衡关系，当相对价格严重偏离长期均衡时，交易者买入低估的资产，同时卖出高估的资产，以求在相对价格回归均值时平仓获利。

配对交易策略主要在两方面有所不同，一是如何选择配对，二是交易规则设置。本文旨在提出一种用于配对选择的新方法，并比较它与其他常用方法的优劣。通常，人们对股票池先按行业分类再在同行业内选择配对，这种方式由于限制搜索范围，可能错失优质配对，而且它被广泛采用，导致获利空间缩小。假使不囿于同行业内而在全部证券组成的超大空间内选择配对，那么待检验的候选配对的数量会非常之多，而且面临多重假设检验的不利影响。本文提出的方法在挖掘数据自身特征的基础上更有效地将全部证券预先划分为多个类簇，从而不必执行过度的搜索，同时，这种分类方法不是基于诸如行业这种广泛采用的显而易见的分类特征。

配对交易策略实施的前提是标的资产可以被做空。我国于 2010 年 3 月 30 日正式启动融资融券业务。本文以沪深 300 指数成分股作为股票池，利用主成分分析法提取股票价格时间序列的主要特征，然后基于数据特征使用 OPTICS 密度聚类算法将股票池中的全部股票划分为多个类簇，再在每一类簇中对候选配对依次执行协整检验、计算并检查 Hurst 指数、半衰期、均值回归频率等步骤，最终选出符合条件的配对。密度聚类算法可以深入挖掘数据点之间的深层次关系，而行业分类不仅带有分类编制者的主观性，而且未必能反映数据点的主要特征。OPTICS 算法不要求事先指定类别数量、假定类别形

态，且善于识别异常值，因此是本文课题下理想的聚类算法。联合使用协整检验、Hurst 指数、半衰期、均值回归频率则是为了保证相对价格有较强的均值回归特性。

本文分别采用不进行分类、按行业分类和 OPTICS 算法分类三种预处理方式结合固定阈值交易规则进行交易模拟。为降低偶然性，本文在 2017-2020 年、2018-2021 年、2019-2022 年三个时间段上展开实证。实证结果表明，与按行业分类相比，OPTICS 算法分类在投资回报率、夏普比率和交易有效性等方面都更胜一筹。与不进行分类相比，OPTICS 算法分类的交易有效性更高，在投资回报率和夏普比率方面两者则互有输赢。但是，OPTICS 算法分类方式下的计算效率远高于不进行分类。因此，综合来看，OPTICS 算法分类更优。本文的实证结果还表明，按行业分类的交易表现往往差于不进行分类。按行业分类虽然看似较为合理，但也许并不值得采纳。

关键词：配对交易；股票配对；配对选择；OPTICS 聚类；

ABSTRACT

Pairs trading is a widely used trading strategy among investors. The basic principle of investment is buying low and selling high, but it is often difficult to accurately estimate the intrinsic value of a single asset. Pairs trading focuses on whether one asset in the pair is significantly overvalued or undervalued compared to another asset, thus avoiding the estimation of intrinsic value. The spread of the prices of the assets participating in the pair should maintain an equilibrium in the long run. When the spread severely deviates from the long-term equilibrium, investors buy the undervalued asset and sell the overvalued asset at the same time, and close the positions to gain profits when the spread reverses to the mean.

Pairs trading strategies are different in two aspects. One is how to select a pair, and the other is the setting of trading rules. This paper aims to propose a method for pair selection and compare its advantages and disadvantages with other methods. Generally, people screen pairs in the same industry by testing whether there is a cointegration relationship between the historical prices of the assets. However, due to the limitation of search scope, this method will undoubtedly reduce the possibility of trading, and because it is widely used, the profit margin is relatively small. If one screens pairs in the super large space composed of all securities without being confined to industry category, the number of candidate pairs to be tested will be very large and the adverse impact of multiple hypothesis tests comes up. The method proposed in this paper can effectively divide all securities into various categories in advance, so that excessive search is not necessary. In the meanwhile, this classification is not based on an obvious feature which is widely used.

The prerequisite for pairs trading is that the underlying assets can be shorted. China officially launched securities margin trading on March 30th, 2010. This paper takes the constituent stocks of the Shanghai and Shenzhen 300 Index as the stock pool, uses the principal component analysis method to extract the main characteristics of the stock price time series, and then uses the OPTICS density

clustering algorithm to divide all the stocks in the stock pool into various clusters based on the data characteristics, and then performs the cointegration test on the candidate pairs in each cluster, and calculates and checks the Hurst index, half-life, mean-reversion frequency. Finally, the eligible pairs are acquired. Density clustering algorithm can deeply explore the deep relationship between data points, while industry classification not only has the subjectivity of classification compilers, but also may not reflect the main characteristic of data points. The OPTICS algorithm does not need to specify the number of categories or assume the shape of categories in advance, and is good at identifying outliers, so it is an ideal clustering algorithm for this topic. The joint use of cointegration test, Hurst index, half-life and mean-reversion frequency is to ensure that the spread does have strong mean- reversion characteristic.

In this paper, three pre-processing methods, namely, no classification, classification according to industry and classification based on OPTICS algorithm, are used respectively in combination with the fixed threshold trading rule for trading simulation. In order to reduce the occasionality, this paper conducts empirical research in three periods: 2017-2020, 2018-2021 and 2019-2022. The empirical results show that the OPTICS algorithm classification is superior to the industry classification in terms of ROI, Sharp ratio and trading effectiveness. Compared with no classification, OPTICS algorithm classification has higher trading efficiency, and in terms of ROI and Sharp ratio each has its own merit. However, the calculation efficiency of the OPTICS algorithm method is much higher. So in general, the OPTICS way is better. The empirical results also show that the performance of the industry classification is worse than no classification. Although classification by industry seems reasonable, it may not be worth adopting.

Keywords: Pairs Trading; Stock Pairs; Pairs Selection; OPTICS Clustering;

目 录

1.绪　论

1.1 研究背景

配对交易是统计套利的一个分支。统计套利可以被划分为因子模型和配对交易两大类。有时人们会将配对交易和统计套利这两个术语交替使用，其实，不是所有的统计套利都是配对交易，但所有的配对交易都是统计套利。

配对交易产生于 20 世纪 80 年代后期知名投资银行摩根士丹利旗下的分析师 Nunzio Tartaglia 领导的一支主要由数学家和物理学家组成的研究团队。全球最知名的几家对冲基金，比如 PDT Partners 和 D.E. Shaw 就是该团队的成员离开公司之后创建的。[1]作为统计套利的有效方法，配对交易一经出现，很快就成为一种在机构和个人投资者中十分流行的策略。但是，随着市场和技术迅速地演进和发展，套利的机会变得越来越少，收益率也越来越低。于是，在 2000 年代初，曾经广受欢迎的配对交易策略进入到一段冰河期。直到十年之后，研究者对这一领域的兴趣才又重新点燃，为该领域的研究引入各种先进的技术方法。这些先进的技术方法持续累积，推动配对交易策略向纵深发展，形成各种各样复杂程度不同、涵盖各类资产投资的框架体系，让其更接近于成为一种真正意义上的兼具通用性和稳健性的交易策略。

在综合众多文献的观点的基础上，配对交易可以被定义为：它是一种通过利用两个或多个有共同运动趋势的标的资产之间存在错误的相对价格来进行套利的方法。这些资产的价格从长期来看应该保持一种均衡关系，当相对价格严重偏离长期均衡时，它通过买入低估的资产，同时卖出高估的资产，以求在相对价格回归均值时平仓获利。

这一策略背后的思想可以溯源到投资的基本原则，即买入低估的资产并卖出高估的资产。然而，要确定某项资产是否的确被高估了或低估了，我们必

须先知道其内在价值。内在价值是价值投资的出发点，但却很难精确计算，往往只能近似估计。配对交易则试图利用价格相关性来解决这一难题。如果两项资产具有相同的特征和风险敞口，那么我们就有理由推断它们的价格走势也会相近。这样做的好处是，对资产内在价值的估计就不再是必须的，重点只需要放在判断配对中的一项资产相对于另一项资产来说，其价值是否存在被明显低估或高估的情形。我们只需要关注配对中的各项资产价格之间的关系，即价差，如果价差突然增大，那就可能是其中一只证券定价偏高，或者另一只证券定价偏低，或者两种情况兼而有之。在这种情况下，我们做空定价偏高的证券并做多定价偏低的证券以求在将来获利，因为我们预计，错误定价会在将来自发地修正，价差会逐渐收敛到长期均衡水平。这里必须补充说明的是，配对交易最简单的配对形式是两项资产，但它也可以扩展到一个包含 N 项资产的满足上述均值回归条件的资产组合。

配对交易的应用场景十分广泛。由于股票和 ETF 市场能够提供大量的潜在配对，所以，在国外，配对交易策略为大多数股票投资者所青睐。Jacobs 和 Weber（2015）[2]对全球 34 个金融市场进行的实证研究表明，配对交易在新兴市场最为有利可图，这或许是因为新兴市场的市场有效性更低。然而，股票和 ETF 也有局限性。做空对于配对交易策略至关重要，在股票和 ETF 上实施配对交易策略时，投资者必须考虑卖空所需的抵押品、被做空证券的出借方是否容易找到，以及不同国家对于做空的限制政策等等问题。在其他大类资产领域，诸如大宗商品、外汇，也有很多支持配对交易盈利性的相关研究报告。期货，特别是大宗商品期货，是配对交易策略的重要战场。早在 30 多年前研究者就已经发现原材料商品价格存在共同运动的现象。期货的优势在于对做空没有限制。外汇市场上的统计套利虽然不甚流行，但也有相当一部分交易者使用各国货币展开配对交易。某些外汇交易的流动性低，导致直接交易无法进行，这时交易者需要借助于中介，即交易量大的货币（一般是美元），以便在目标货币（比如加元和澳元）之间进行对冲交易。和期货一样，外汇市场的优势是不限制做空。总的来说，在每一类资产领域，都存在着长期价格联动和短期价格失效的现象。只要市场表现违反一价定律，通过配对交易进行统计套利就大有可为。

1.2 研究目的

通常人们更注重寻找最优的交易规则，但是，选择合适的配对与交易规则设置一样重要。一个优秀的交易规则模型可能受累于一个不甚理想的配对而难以发挥出其应有的效果。因此，能否选出合适的配对在很大程度上决定配对交易策略的成功与否。

但是，随着获取交易数据变得越来越容易以及配对交易策略的广泛传播和应用，交易者们越来越难以挖掘出能够带来丰厚回报的配对。如果能够另辟蹊径找到其他人没有发现的有利可图的配对，那么交易的盈利水平就会大大增加。但是，这样的机会并不容易实现。一方面，如果交易者希望搜索尽可能多的交易可能性而不对搜索范围施加任何限制，那么他就不得不面临着从海量的候选配对中找出合适配对的窘境。假设市场中一共有 n 只可交易的证券，将每一只证券与其他所有的证券进行组合，得到的候选配对一共有$\frac{n\times(n-1)}{2}$个。这时，会出现两个问题。一是效率问题：随着 n 的增大，对全部候选配对进行均值回归特性检验的计算成本会急剧膨胀，这对计算机硬件和处理时间都提出了更高要求。二是多重假设检验问题：根据假设检验的定义，在α的显著性水平下，第一类错误出现的概率为α。那么在 m 次假设检验中，第一类错误至少出现一次的概率可以表示为$1-（1-\alpha）^{m}$。在实践中，α的取值一般为 1%、5%或者 10%。但无论α取其中哪一个值，当 m 足够大的时候，上式的值都会接近于 1。也就是说，选出错误配对的可能性约等于 100%。为了减轻多重假设检验的不良影响，研究者提出 Bonferroni 等多种校正方法，但是 Bonferroni 校正方法实际上过于保守。Harlacher（2016）[3]研究发现 Bonferroni 校正方法会导致十分保守的配对选择结果，以至于妨碍有真实协整关系的配对被发现。因此，他建议对全部证券进行合理的预先划分，然后在各个类簇中构建候选配对，目的是减少候选配对的数量，从而减少假设检验的数量。另一方面，如果交易者严格地限制搜索范围，即只在同一行业内寻找合适的证券配对，这会大大减少假设检验的数量，从而降低选出错误配对的可能性，而且属于同一行业的证券往往有着相同的特征和风险敞口，有理由认为它们的价格走势更有可能呈现共同运动的趋势。但是，这种简单的处理方式会带来一

个明显的缺点，即采用这种方式的人会很多，交易者很难找到与众不同的、尚未被广泛关注的配对，因此其获利空间就会比较小。

综上所述，我们需要提出一种更优的用于配对选择的方法：它能够有效地将全部证券预先划分为多个类簇，从而不必执行过度的搜索，同时，这种分类方法不是基于某种广泛采用的显而易见的特征（例如行业）。

1.3 研究方法和框架

本文运用了文献研究法、定性研究法和定量研究法。本文对国内外学者在配对交易领域的相关文献做了认真的梳理，对相关研究成果形成了全面和客观的认识。本文对配对交易理论和方法进行了深入探究，在此基础上搭建新的配对选择模型并从理论上阐述模型和模型所使用方法的合理性。本文使用数学和统计工具，在历史数据的基础上开展实证研究，佐证了所提出的模型的有效性。

由于沪深 300 指数成分股具有成交量大、流动性好、公司质地优良等优点，而且都是融资融券标的股票，可以满足做空需求，因此，本文选取沪深 300 指数成分股作为股票池。然后，分别采取不做预先划分、根据上市公司所属行业做预先划分、使用机器学习中的聚类算法 OPTICS 做预先划分三种不同的方式来构建股票类簇，并在每一类簇内对其中所有可能的股票配对依次进行协整检验、计算价差时间序列的 Hurst 指数、计算均值回归过程半衰期和均值回归频率等筛选步骤，选出符合全部筛选条件的股票配对。本文基于最终选出的股票对构建资产组合。构建资产组合有两种方式，一种方式是涵盖所有选出的股票对，另一种方式是仅包含特定数量的根据特定标准优选出来的股票对，这样做的原因是，在第一种方式下资产组合中的股票对数量可能会过多而且差异过大，不仅不适合真实情景下的操作，而且不利于横向比较。此外，本文将 2017-2022 年的股价数据划分为 2017-2020 年、2018-2021 年、2019-2022 年三个不同区间用于实证研究，目的是降低实证结果的偶然性。本文将在每一个区间上构建上述两种资产组合并按照固定阈值准则展开交易模拟，最后将三种分类方式下的交易表现进行比较以得出三种方式孰优孰劣的结论。

本文研究框架如图 1-1 所示。

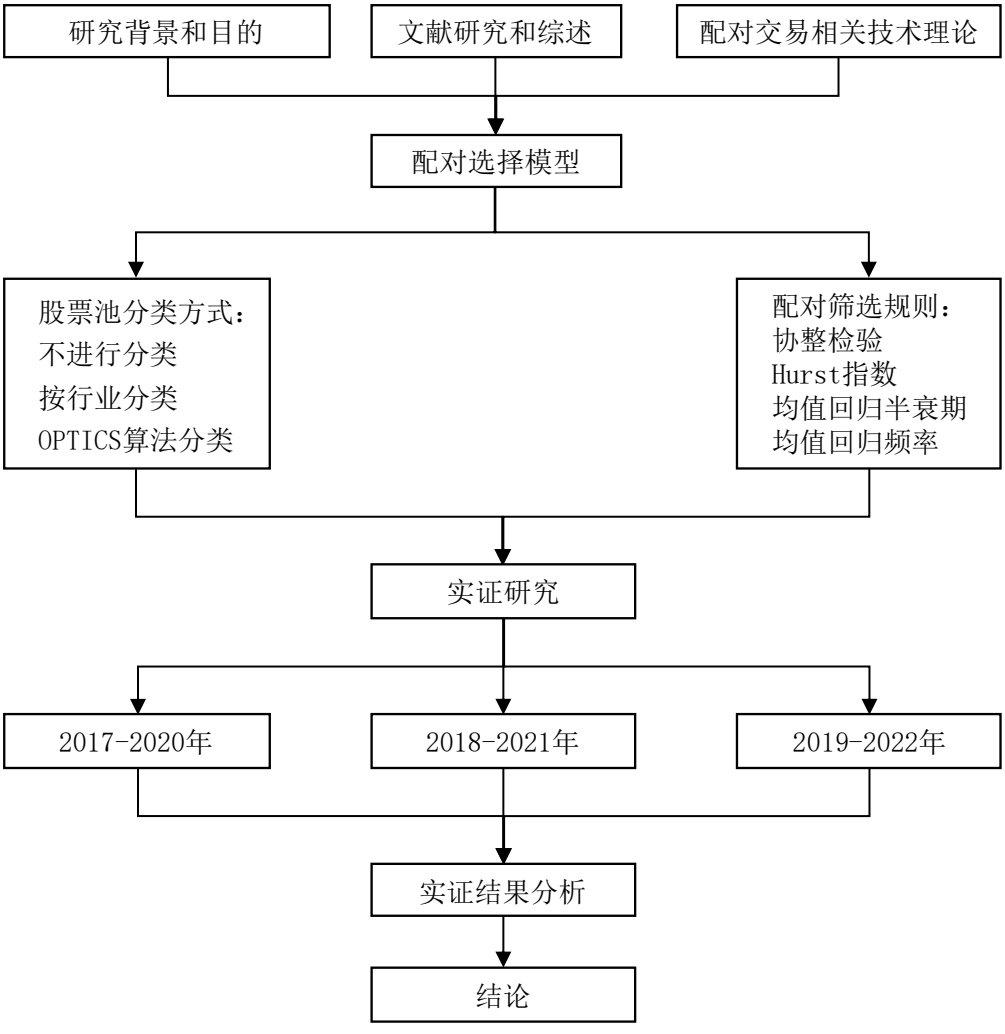

图 1-1 研究框架

本文的章节划分和主要内容如下：

第一章是绪论，主要阐述本文的研究背景、研究目的、研究方法和框架、创新之处。

第二章是文献综述，主要介绍国内外学者在配对交易领域的相关研究成果。本文将根据配对交易的两个主要阶段，即配对选择阶段和执行交易阶段，分开展开论述。

第三章介绍配对交易策略的相关理论基础，包括聚类算法、协整理论、Hurst 指数、半衰期，本文将重点阐释选择 OPTICS 算法的原因。

第四章介绍本文模型的框架结构，包括交易标的和数据集的概况、配对

筛选过程、交易规则、交易成本、收益与风险度量指标的计算等内容。

第五章展示模型运行的各项数据结果，并详细地比较了三种分类方式下的交易表现。

第六章是结论和未来展望，总结陈述了本文的研究结果，并提出了文章的不足之处和未来改进的方向。

1.4 本文创新

（1）以大数据的思维和方法研究配对交易。本文综合无监督学习分类算法、协整检验、Hurst 指数、均值回归半衰期、均值回归频率等技术，提出了一个较为完整的在配对交易中用于配对选择的模型框架。该模型基于大数据的思维和方法，普适性较强，能够处理规模庞大的股票池，而不仅仅局限于在一个较小的范围内对其中的少量股票进行数量极为有限的配对和简单的协整检验或者相关性检验，实证结果表明，本文提出的模型的性能和效果较好。

（2）在配对选择中引入 OPTICS 分类算法。本文利用 OPTICS 算法对股票池进行分类，相比于不进行分类，OPTICS 算法分类大大降低了配对选择的计算量并减轻了多重假设检验带来的不良影响；相比于按行业分类，OPTICS 算法分类能够更深入地挖掘数据集内数据点之间的深层次关系。实证结果表明，OPTICS 算法分类在交易的投资回报率、夏普比率、交易有效性方面均有较为明显的优势。

（3）本文的实证研究不局限于个别的股票配对，也不局限于个别的形成期和测试期。本文不仅构建了两种资产组合，且每一种资产组合都包含很多个股票配对，而且划分了三组不同的形成期和测试期，这样做降低了实验结果的偶然性，增强了文章论证的说服力。

2.文献综述

一般而言，要构建一个配对交易策略，交易者需要解决两个核心问题：一是，如何选择配对，配对的价差时间序列应当具有较强的均值回归特性；二是，设定交易规则，交易触发机制应当与选择配对的方法和价差时间序列的行为特征相匹配。所以，配对交易策略主要在两个方面有所不同:一是选择配对的方法，二是交易规则背后的思维逻辑。

本文考虑到所研究的问题，将按照配对选择阶段和交易阶段的划分，分别展开文献综述，然后对基于国内市场的配对交易实证研究展开综述，最后对以上文献进行评述。

2.1 配对选择阶段

配对选择阶段遵循两个步骤：（1）构建资产池及在此基础上的候选配对；（2）基于统计检验方法从候选配对中筛选出符合条件的配对。在步骤（1），投资者选择自己感兴趣的大类资产，比如股票、ETF、期货、外汇等等，并从资产池中搜寻可能的配对。在文献中，这一阶段的处理有两种方法。一是对所有证券的全部可能的组合进行地毯式搜索，比如 Krauss 等（2017）[5]和 Caldeira 等（2013）[6]的研究工作；二是按行业分类对全部证券先进行分组，然后再在组内形成配对，比如 Dunis 等（2010）[7]和 Do 等（2010）[8]的研究工作。第一种方法有助于挖掘出一些鲜为人知的合格配对，但计算量也很大；第二种方法有助于降低发现虚假的同向波动关系的可能性，但也限制了搜索空间。在步骤（2），投资者需要定义一套用于筛选的方法和标准，常用的方法包括最小距离法、相关性方法和协整检验方法。

2.1.1 最小距离方法

Gatev 等（2006）[9]率先提出用最小距离准则选择配对。这篇论文是配对交易研究领域的奠基之作，被后来者大量引用，影响深远。作者从 CRSP 数据库中选取了从 1962 年到 2002 年之间所有在流通的美股股票的每日价格数据，然后计算每只股票的累计回报率，并以跨度为 12 个月的形成期的第一个交易日为基准进行缩放处理，再根据缩放后的累计回报率计算形成期内各个股票之间的距离平方和，按照距离平方和大小对股票对进行排序。作者建议选择距离平方和最小的股票对用于配对交易。作者的实证研究表明，利用选出的股票对构建一个自融资组合，然后应用一种简单的基于固定阈值的交易规则就可以获得 11%的超额年化收益率。

但是，根据 Krauss（2015）[4]的分析，以距离平方和作为筛选标准并不是最优的。为了证明这一点，假设用证券x和y进行配对，$p_{x,t}$ 和 $p_{y,t}$ 是证券x和y的根据基准日缩放后的价格时间序列，那么价差为 $p_{y,t} - p_{x,t}$，价差的方差可以表示为

$$Var(p_{y,t} - p_{x,t}) = \frac{1}{T}\sum_{t=1}^{T}(p_{y,t} - p_{x,t})^2 - \left[\frac{1}{T}\sum_{t=1}^{T}(p_{y,t} - p_{x,t})\right]^2 \quad (2-1)$$

在式（2-1）中，$\sum_{t=1}^{T}(p_{y,t} - p_{x,t})^2$正好是距离平方和的表达式。最小距离准则要求（2-1）式的第一项尽可能小，这意味着价差的方差也会比较小。从逻辑上来说，这与提高配对交易的盈利性相抵触。因为只有当价差上下波动出现了明显偏离长期均值的情形时，交易者才有可能利用这种异常值进行套利。所以，一个好的配对应该兼具较大的价差方差和较强的均值回归特性。而最小距离准则没有考虑到这一点。

2.1.2 相关性方法

Chen 等（2017）[10]研究了利用证券收益率的相关性来筛选配对的方式。作者希望构建一种比 Gatev 等（2006）[9]的方法更具鲁棒性的配对选择方法。为了便于比较，他们使用了与 Gatev 等（2006）[9]相同的数据集。在样本内，资产 a 和 b 收益率之差的方差的计算公式如下：

$$Var(r_a - r_b) = Var(r_a) + Var(r_b) - 2\rho_{r_a,r_b}\sqrt{Var(r_a)}\sqrt{Var(r_b)} \quad (2-2)$$

其中，Var 表示方差，ρ_{r_a,r_b}表示资产收益率r_a和r_b的相关系数。从（2-2）式不难推出，即使 a 和 b 各自的资产收益率的方差都比较大，但只要二者之间的相关系数足够大，则资产收益率之差的方差就会比较小。

Chen 等（2017）[10]的研究结果表明，使用皮尔逊相关系数作为配对选择方法能带来更好的交易表现，平均月度收益率达到 1.7%，这一数字接近于 Gatev 等（2006）[9]使用最小距离法取得的收益率的两倍。但是，相关性方法无法从理论上解释均值回归过程，不能保证资产价格之间有长期均衡关系的存在，所以相关性方法也不是最优的配对选择方法。

2.1.3 协整检验方法

协整检验方法通过检验两只证券的价格之间是否存在协整关系来筛选配对。如果证券 Y 和 X 之间存在协整关系，那么它们的线性组合 $(p_y - \beta p_x)$ 是一个平稳时间序列，其中，p_y和p_x分别表示 Y 和 X 的价格时间序列，β表示协整系数。需要指出的是，协整和相关性二者之间没有必然的关系。两个时间序列之间存在协整关系，但可能相关性较低，相关性较高的时间序列之间也可能不存在协整关系。[11]

在该领域，Vidyamurthy（2004）[12]的工作是被引用最多的。Vidyamurthy 提出了一个将协整检验方法用于配对交易的理论框架。该框架包含三个关键步骤:(1)在基本面信息或者统计信息的基础上，选出存在协整关系的配对。(2)使用特定的方法检验配对的可交易性。(3)设计基于非参数方法的交易规则。在整个过程中，Vidyamurthy 并没有执行严格的协整检验，而是基于实用原则选择技术路线，但是利用协整检验方法选择配对的思想是其理论框架背后的指导原则。

Krauss（2015）[4]认为，对于配对选择，协整检验方法比最小距离法更严谨，因为协整检验能发现计量经济学意义上更有说服力的长期均衡关系。Huck 和 Afawubo（2015）[13]使用标准普尔 500 指数成分股，在不同参数设置以及考虑风险载荷和交易成本的条件下，对协整检验方法和最小距离法做了比较研究。研究结果表明，协整检验方法要显著优于最小距离法，这证明了协整检

验方法在验证不同证券价格之间是否存在稳健的长期均衡关系方面是一种更为有效的方法。

Chan（2013）[14]指出，利用交易型开放式指数基金（ETF）进行配对交易相比于单一股票配对有更大的优势，如果在样本内数据中发现 ETF 之间存在协整关系，那么这种协整关系大概率地会在样本外数据中继续保持。

胡伦超等（2016）[36]构建了一个协整检验和最小距离方法并用的两阶段配对交易模型，并在上证 50 指数成分股上展开实证研究。通过验证可行性和有效性后，他们发现所提出的两阶段配对交易策略相较于仅考虑协整关系的配对交易模型，不仅可以降低模型风险，而且能够取得更高的收益率。

2.1.4 其他方法

最小距离法、相关性方法和协整检验方法是文献中提到最多的三种配对选择方法，但除此之外，还有研究者提出其他一些方法。Ramos-Requena 等（2017）[15]通过实证研究发现，与协整检验和相关性等经典方法相比，使用 Hurst 指数取得了更好的实验结果，尤其是当投资组合包含的配对较多时。因此，他们建议利用 Hurst 指数对配对进行排序和筛选。

除了统计方法之外，还有根据基本面信息从理论上确立配对资产之间的长期均衡关系的做法。例如，Dunis 等（2006）[16]选择原油和汽油进行配对，因为汽油是从原油加工中提取的。类似的情形还发生在玉米和乙醇之间[17]，因为玉米是生产乙醇的主要原料。也有一些研究工作将行业分类和统计检验结合起来筛选配对。张翠（2013）[37]基于 A 股市场做了实证研究，分别按照行业和行业上下游对公司股票进行分类，在此分类的基础上搜索符合要求的股票配对，研究发现无论按行业分类还是行业上下游分类都有助于提高配对交易的收益率，并且按行业上下游分类的收益率比按行业分类的更高。鲍鹏飞（2017）[38]以沪深 300 指数成分股作为股票池，先按行业分类，再按照上游企业、下游企业进行分类，其研究表明在同行业中的上游企业或下游企业内部进行配对的交易结果要好于上下游企业之间交叉配对的交易结果。

2.2 配对交易阶段

交易规则方面的研究主要集中于基于阈值的交易机制。

2.2.1 基于阈值的交易机制

在 Gatev 等人的那篇经典文章中，交易规则的设置是基于价差的发散和收敛过程。[9]倘若价差偏离长期均值的幅度超过历史标准差的两倍，交易者应该进行开仓操作。具体而言，当偏离幅度为正的两倍标准差时，交易者卖出价差；当偏离幅度为负的两倍标准差时，交易者买入价差。开仓之后如果价差收敛回到了长期均值，那么交易者应该将所持有的仓位进行平仓操作。这一交易机制虽然简单有效，但它的隐患是，倘若价差没有按照预期收敛，则会导致交易损失。

Elliott 等（2005）[18]运用时间序列方法来研究配对交易策略。作者用高斯·马尔科夫链来描述价差的均值回归过程。他们建立了包含状态方程和测量方程的状态空间模型，并计算出开仓阈值。根据 Do 等（2006）[19]的说法，该方法有三大优点：（1）模型易于处理，卡尔曼滤波器和状态空间模型可用于估计参数。（2）可以发展出连续时间模型用于预测。（3）该方法本质上是基于均值回归，而均值回归是配对交易的关键。Do 等人也对该方法提出了批评。他们指出，首先，价差应该使用对数价格而非价格；其次，这种严格的模型只适用于存在收益率平价关系的证券，而在实践中，收益率平价是较为罕见的。

当基本面或经济上的原因引发结构性崩溃，使得原来有共同运动趋势的配对资产不再强相关时，使用传统的配对交易策略就容易招致失败。这种崩溃可能导致价差出现异常大的偏离，而且无法回复到长期均值。在这种情况下，继续押注价差收敛将会很危险。为了解决如何检验均值偏离是暂时的还是永久的这一问题，Bock 和 Mestel（2009）[20]将马尔可夫区制转移模型和统计套利的研究工作结合起来，给不同区制设定不同的开仓阈值和止损阈值，开发出一套行之有效的配对交易规则。

Bertram（2010）[21]假设证券价格服从 Ornstein-Uhlenbeck 过程，并推导出统计套利交易的解析公式。首先，他通过研究 first-passage time 问题（first-

passage time 是指从某个初始状态到首次触发阈值的时间），推导出了交易时间长度的均值和方差的表达式。然后，他推导出单位时间内收益的期望和方差的表达式。以单位时间内期望收益或者夏普比率最大化为目标，就可以求解出最优交易阈值。Zeng 和 Lee（2014）[22]改进了 Bertram（2010）[21]的研究工作，Bertram 只允许在交易过程中做多，为了同时允许做空，Zeng 和 Lee 推导出了存在上下边界的 O-U 过程下的 first-passage time 的期望值的表达式，然后，他们将最大化单位时间内期望收益问题简化为求解方程的问题。

Bogomolov（2013）[23]在 Renko 和 Kagi 图表的基础上提出了一种新的配对交易方法。这种方法利用了交易过程可变性的相关统计信息。它既不需要复杂的数学推导，也不需要对价差的性质做出严格假设。基于 Renko 和 Kagi 模型的交易策略通过测度交易过程可变性来定义交易过程从转折点处向一个方向移动多少距离才能使得反向操作变为有利可图。

胡伦超等（2016）[36]为了验证所提出的两阶段配对交易模型的稳定性，对模型做了参数的敏感性分析。他们发现交易时长、交易触发阈值等模型参数的设置对交易模型的收益率有一定的影响。文章指出，交易的时间跨度不应过短，否则无法确保价差在期末能及时回复到长期均值，并且交易阈值应该设置为较为适中的值，这样才能在整体上最大限度地获利。

欧阳红兵等（2015）[39]使用 AR(1)模型拟合价差时间序列，在此基础上再使用数值算法对交易持续时长和间隔时长以及交易次数进行估计，最后求解利润最大化目标下的最优阈值。作者对同时在内地和香港两地上市的公司股票进行配对交易测试，实证结果证明其提出的求解最优阈值的方法是有效的，另外，使用该方法的固定参数模型相比于时变参数模型有更优秀的交易表现。

胡文伟等（2017）[40]对传统的固定参数模型进行了改进，利用 Sarsa 强化学习算法和 ε-greedy 策略，将传统的凭借主观经验来设置固定参数的办法升级为自适应模式下参数动态优化的方法。作者选取中国债券市场上流通量最大的前五只债券进行实证研究，结果表明基于强化学习的动态参数优化方法能显著地提高投资收益率和索提诺比率，并降低最大回撤率，减少交易次数，提升交易的效率。

冯玉茹（2019）[41]使用 GARCH 模型模拟价差的时变标准差序列，在此基础上确定开仓、平仓、止损等操作的触发信号，作者将 GARCH 模型和传

统的固定标准差模型分别应用于 A 股市场，实证结果显示，两种模型在样本内和样本外均有不错的收益，但 GARCH 模型比传统模型更加稳健。

于晓雨等（2020）[42]的研究旨在控制配对交易中的风险。有别于传统模型采用固定止损条件，作者提出在配对资产协整或部分协整的条件下采用遗传算法求解包含止损条件的最优交易阈值。作者将中证 500 和沪深 300 指数成分股按行业分类后选择配对，然后对所提出的模型展开实证研究。结果表明，与不设止损条件和设置 10%固定止损条件的最优阈值模型相比，带止损条件的最优阈值模型不仅实现了更高的投资收益率，而且在控制风险和损失方面也更为有效。

2.2.2 其他研究

Do 和 Faff（2012）[24]基于 1963 年到 2009 年之间的美股市场研究了交易成本对配对交易盈利性的影响。作者发现，在充分考虑了交易佣金、市场冲击和卖空费用的情况下，配对交易仍然是有利可图的，只是获利水平较从前来说有所降低。经他们精心挑选的配对构建的资产组合实现了每月大约 30 个基点的经风险调整后的资产回报率。在市值排名前 30%的股票上实施配对交易策略则能够实现平均每月 24 个基点的 alpha 超额收益。

Dunis 等（2006）[16]运用人工神经网络对原油和汽油的价差进行建模。作者训练神经网络以用于预测下一个交易日的价差相比当前交易日的变化率。其交易规则基于价差预测变化率和预先设定的阈值之间的比较结果。当前者大于后者时，开多仓或者继续持有多仓；当前者小于后者的相反数时，开空仓或者继续持有空仓；在除此以外的情形下，不持有任何仓位。该模型在不考虑交易成本的前提下取得了 15%的平均收益率，但是交易成本对收益的影响很大。后来，Dunis 等（2015）[17]在玉米和酒精的配对交易中调整了模型，在考虑交易成本的情况下取得了接近 20%的收益率。

2.3 国内市场的配对交易实证研究

丁秀玲和华仁海（2007）[43] 运用协整检验方法和格兰杰因果检验方法证

明大连商品交易所的豆粕期货和大豆期货的价格之间存在着长期均衡关系，他们在样本内模拟交易的结果表明无论是单独考虑多头套利交易或者空头套利交易，还是整体考虑所有套利交易，平均利润都大于零，但只有多头交易的平均利润在统计上显著，空头套利交易和所有套利交易的平均利润都不显著，而在样本外模拟交易的结果则显示套利交易的平均利润不显著。

王峥明（2010）[44]比较了基于收益预测和排序的配对交易策略和基于趋同性检验的配对交易策略在 A 股市场上的表现。作者选取了 A 股市场 2007 年到 2009 年数据进行实证研究后发现，二者都能取得高于市场整体表现的投资回报率。而在市场行情较差的时期它们的表现尤其好，在市场行情较好的时候则表现得一般。对二者进行比较后发现，基于趋同性检验的配对交易策略有更高的年平均收益率，但基于收益预测和排序的配对交易策略的收益波动性更小，夏普比率更高。

蔡燕等（2012）[45]将上证 180ETF 和沪深 300 股指期货进行配对，在经过协整检验后，分别建立基于 Elliot 框架和基于 O-U 随机过程的价差模型。通过比较模型预测结果和真实数据的相关性以及预测结果的标准差和真实数据的标准差后发现，该情形下 O-U 随机过程模型要优于 Elliot 框架模型。

赵胜民、闫红蕾（2015）[46]利用 A 股市场上融资融券标的股票和 ETF 的日频数据进行实证研究，按照融资融券标的分阶段扩容的时间和上市公司所属行业这两个维度进行分组，再基于转移模型使用俱乐部收敛检验和 log t 检验研究价差收敛性质和动态变化。研究发现按行业分类构建股票配对进行套利有着较大风险，融资融券双向交易机制未能充分发挥作用，但统计套利在 A 股市场上仍有一定的可行性，ETF 具有稳定的收敛关系，更适合作为配对标的。

2.4 文献述评

从国内外的相关文献中可以看到，配对交易领域的主要研究方法包括：距离方法运用非参数化的距离度量来寻找配对交易的机会；协整方法依赖于协整检验这一正式的统计方法来判断价差时间序列是否具有平稳性，相比于距离方法，协整方法从计量经济学理论上解释了均值回归过程，因此更有说

服力；时间序列方法无需考虑形成期，在既有配对的前提上，它致力于通过对价差的均值回复过程进行建模来确定最优交易规则；随机控制方法不需要预测价差的未来走势，也不需要设置形成期，它的目标是找到最优资产组合，确定最优仓位准则；还有相关文献较少的一些方法，例如机器学习、主成分分析、Hurst 指数、Copula 等等。

在配对选择的方法上，协整检验被广泛用于检验配对资产的价格之间是否存在长期均衡关系。另外，Hurst 指数在投资组合包含的配对较多时表现出了较好的筛选效果。一些研究提出在同行业或者行业上下游的限制性范围内选择配对。总的来说，针对配对选择策略的研究并不十分充分。本文综合无监督学习分类算法、协整检验、Hurst 指数、均值回归半衰期、均值回归频率等技术，提出了一个较为完整的在配对交易中用于配对选择的模型框架，是对该领域研究的有益补充。

3.配对交易相关技术理论

3.1 主成分分析

在搜索配对时，交易者希望找到那些有相同的系统风险敞口的证券。根据套利定价理论（Arbitrage Pricing Theory），这些证券从长期来看应该具有相同的期望回报。因此，任何偏离这一预期的价格都应当被视为错误定价，并为交易者套利提供契机。Jolliffe（2011）[25]指出，可以对资产回报率使用主成分分析法来提取证券背后的那些共同的风险因子。

PCA（Principle Component Analysis），即主成分分析法，是一种统计上常用的降维方法。它使用正交变换将一组可能线性相关的变量的观测值转换为一组线性无关的变量即主成分的线性表示。这种转换被定义为第一个主成分尽可能多地解释数据的变化，每一个后续的成分必须与前面的所有成分保持正交关系，同时在其和其之后的全部成分中能够最多地解释数据的变化。每一个成分可以被视为一个风险因子。

PCA 的运算方式如下：首先，根据证券 i 的价格时间序列 $P_i(t)$，计算它的收益率时间序列 $R_i(t)$，

$$R_i(t) = \frac{P_i(t) - P_i(t-1)}{P_i(t-1)} \qquad (3-1)$$

然后，对收益率时间序列进行标准化处理，即减去样本均值$\overline{R_i}$后再除以样本标准差σ_i，

$$r_i(t) = \frac{R_i(t) - \overline{R_i}}{\sigma_i} \qquad (3-2)$$

在标准化的收益率时间序列的基础上，计算相关系数矩阵ρ，公式为

$$\rho_{i,j} = \frac{1}{T-1}\sum_t r_{i,t} \times r_{j,t} \qquad (3-3)$$

使用收益率时间序列而非价格时间序列进行主成分分析的原因就在于收益率时间序列的相关系数矩阵对于估计价格的同向波动性来说更有信息量，而价格时间序列则可能由于时间趋势项表现出虚假的相关性。

下一步，提取特征向量和特征值以构造主成分。特征向量决定最大方差的方向，特征值决定对应特征向量的幅度。特征值和特征向量可以通过奇异值分解来求得。为此，将前面经过标准化处理后的 n 只证券的收益率时间序列拼合成一个矩阵，记为A，对A应用奇异值分解原理，得到

$$A = U\Sigma V^T \qquad (3-4)$$

其中，矩阵U和V都是正交矩阵，矩阵 Σ 对角线上的元素是 A^TA 的特征值的平方根，也被称为 A 的奇异值，沿着对角线按从大到小依次排列。

注意 $A^TA = V\Sigma^T\Sigma V^T$ ，而A^TA就是前面计算过的相关系数矩阵ρ，这样一来， V就能被确定下来。通过这种方法，就有可能找到矩阵A的特征向量。我们选择对应于最大方差的 k 个方向的 k 个特征向量，其中 k 代表特征数。特征向量越多，对数据的描摹就越好。

最后，将原始矩阵 A 乘上特征向量得到一个大小为 n×k 的新矩阵。矩阵的维度被降低到选定的 k 个特征上。

对于如何决定 k 的数值，一般的做法是分析每个主成分所解释的总方差的比例，然后选取主成分以达到一定的解释比例。本文要在提取数据特征之后运用无监督学习算法，较高的特征维度会带来两个问题：1、找到无关特征的可能性会增加；2、维数灾难。维数灾难是由 Bellman（1957）[26]提出。当测量明显相似的数据点之间的距离时，维数灾难会导致这些数据点突然变得非常遥远，进而导致聚类过程变得无效。根据 Berkhin（2006）[27]的研究，当维数大于 15 时，维数灾难的影响十分严重。有鉴于此，k 的取值将根据经验进行选择，上限为 15。

3.2 聚类分析

聚类分析是一种根据研究对象的数据特征，将研究对象分组到不同类别的统计分析技术，属于无监督学习方法。

就本文研究的问题来说，理想的聚类算法应该具备以下特性：

（1）无须事先指定类别数量。因为缺少相关的信息，所以不应对类别数量进行预先假设，而应该由数据自己“说话”。

（2）无须将每一只证券都对应到一个类别。因为某些证券的价格时间序列可能表现得极为与众不同，它们应该被作为异常值处理。选择一个鲁棒性强的聚类算法将这些异常值剔除以免损害聚类效果，是很有必要的。

（3）分类必须严格。否则，会导致无效配对数目增多。

（4）无须预先假定类别形态。理由与（1）相同。

3.2.1 聚类算法概览

常用的聚类分析法可以被划分为三类：分区聚类（Partitioning Clustering）、层次聚类（Hierarchical Clustering）和密度聚类（Density-based Clustering）。

分区聚类，以 K-means 为代表性方法，其不适用于本文研究的原因有三：首先，这类算法不能较好地处理噪声数据和异常值。其次，它要求聚类形状为凸性。它假设数据围绕着中心呈现正态分布，这一假设过于苛刻。最后，它要求事先指定类别数量。

层次聚类算法基于动态建模思想，它需要交易者决定聚类何时终止，这可能会增加交易者不必要的偏见。最终交易者的选择可能十分接近于按照标准的搜索方法得到的结果，而这是我们从一开始就极力规避的。因此，层次聚类在本文的研究中也是不合适的。

与分区聚类、层次聚类相比，密度聚类有以下优点：首先，它对聚类形状没有要求，因此无须对数据分布做过于苛刻的假设。其次，它不必对数据集中的每一个点进行分组，因此它能够有效剔除异常值。最后，它不需要事先指定类别数量。因此，在本文的研究工作中，我们将采用密度聚类方法。

DBSCAN 算法是一种经典的密度聚类算法。在图 3-1 中，我们之所以能在每一个数据集中轻易发现哪些数据点可以归为一类，是因为每一类数据点都有一个典型的密度，这个密度要明显高于外面的点。 这就是 DBSCAN 算法背后的主要思想。

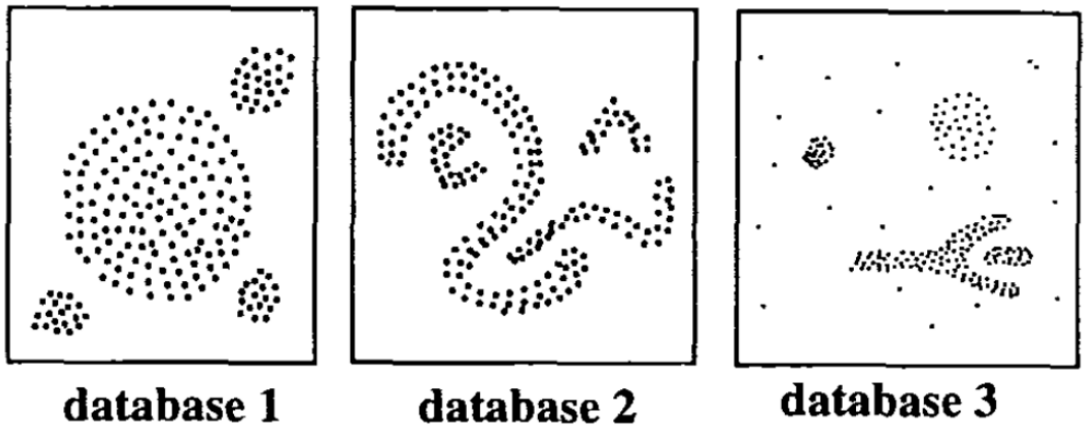

图 3-1 基于密度聚类的分类簇示例

本文不打算详细介绍 DBSCAN 算法的原理，只介绍它的优缺点。它的优点在于：（1）它可以处理任意形状的稠密数据集，而不像 K-means 那样局限于凸数据集；（2）它对异常数据点不敏感，能够识别出异常点；（3）结果无偏倚，相比较而言，K-means 之类的算法聚类结果受到初始值的影响会很大。它的缺点在于：（1）对密度差异较大的数据集的聚类效果不好；（2）需要投资者预先设定参数，即距离阈值ε和邻域样本数阈值$minPts$，参数设置对聚类效果的影响较大。因此，就本文的研究工作来说，虽然 DBSCAN 算法相较于分区聚类和层次聚类算法体现出了一些优势，但仍然不是最理想的选择。

3.2.2 OPTICS 算法

OPTICS，全称为 Ordering Points to Identify the Clustering Structure,也是由 DBSCAN 算法的研发团队开发出来的一种密度聚类算法。[28]与 DBSCAN 要求相似的数据密度不同，OPTICS 能够有效地作用于密度差异较大的数据集。所以，OPTICS 是 DBSCAN 的改进版。

图 3-2 形象地说明了 OPTICS 相比于 DBSCAN 的优势所在。

在这个例子中，DBSCAN 只能识别指定的距离参数 eps 以内的数据点为相邻点，于是，它就会无视图中右下角的类簇。如果我们放大 eps 的取值，虽然被遗漏的类簇能被识别了，但是也可能给其他类簇的识别带来不利影响，比如类簇 1 和类簇 2 被合并为一个类簇。所以在 DBSCAN 算法里，eps 的值很难预设。但是在 OPTICS 算法里，这个难题就不存在了。

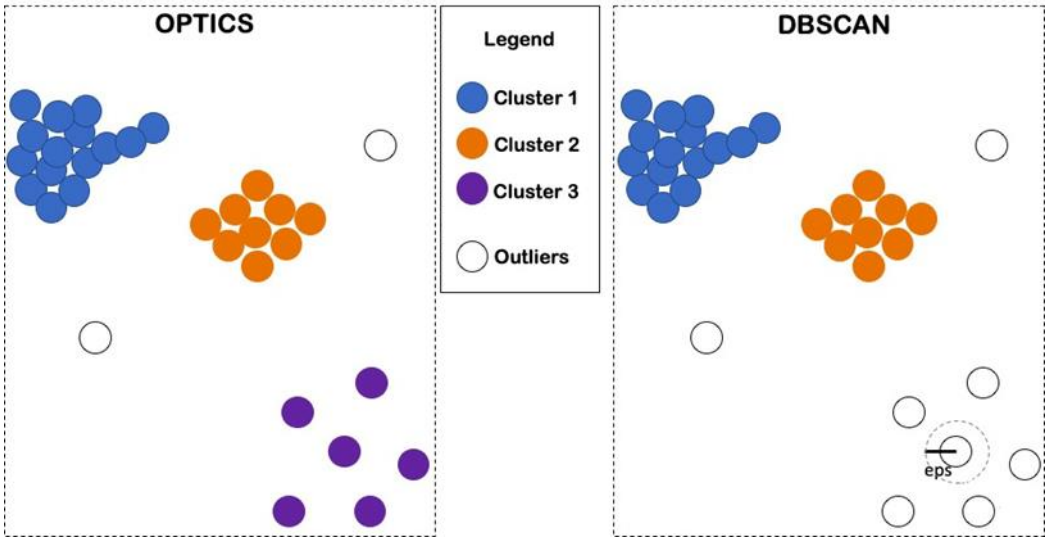

图 3-2 OPTICS 和 DBSCAN 的对比说明

为了充分考虑不同类簇所适用的不同距离标准，OPTICS 算法将样本点与周围的点进行逐一比较。图 3-3 对此做了说明。如果我们想知道 A 到 B 的距离是大是小，我们可以将 A 到 B 的距离与 A 到邻域内所有点的距离进行比较。从图中我们可以看出，A 到 B 的距离相较于到其他点的距离不算远，因此，A 和 B 可以被视为相邻点。

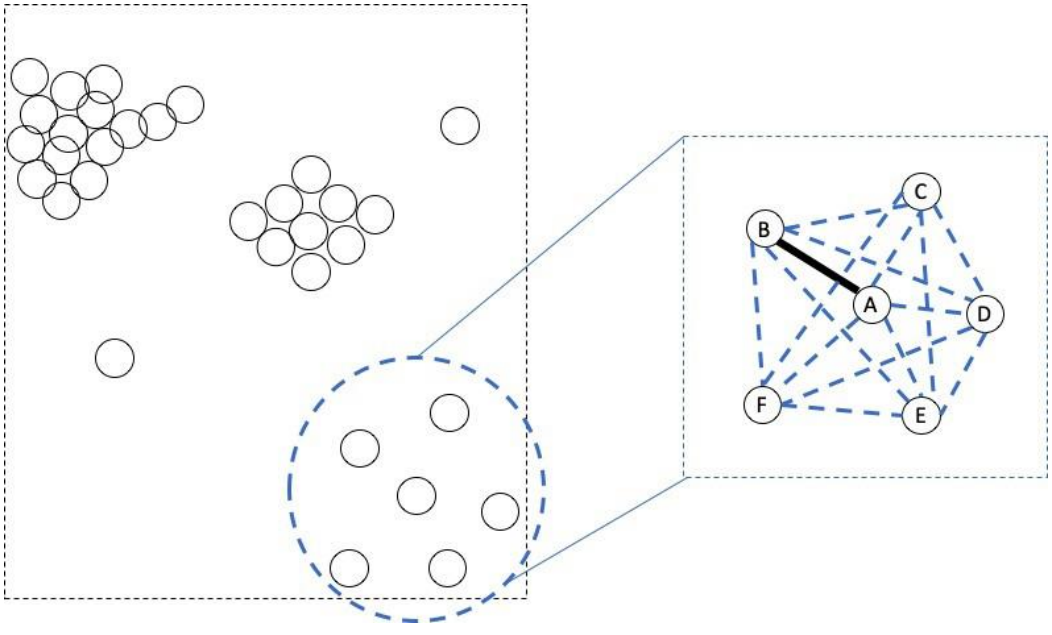

图 3-3 OPTICS 算法判断相邻点的方法

OPTICS 算法的核心思想是通过识别密度连接点来提取数据集的聚类结构。为构建数据集的密度表示，该算法会生成一张称之为“可达性图”的有序列

表。列表中每一个点都对应一个可达距离，可达距离衡量了数据集中其他点到达该点的难易程度。可达距离近似的数据点最有可能属于同一类簇。

OPTICS 算法有两个重要概念：

（1）核心距离：它是某个给定点被归类为核心点所要求的最小半径值。如果给定点不是核心点，则无需定义它的核心距离。

（2）可达距离：它是相对于另一个数据点定义的。点 p 和 q 之间的可达距离是 p 的核心距离和 p 与 q 之间的欧几里得距离(或基于其他度量尺度测算的距离)两者中间的较大值。这里要求 p 是核心点，意味着可达距离不可能小于核心距离。

图 3-4 形象地说明了核心距离和可达距离这两个概念。设定最少相邻点个数$minPts$为 5，半径ε为 6mm。在左图中，核心距离用ε'表示，它是让 p 成为核心点的最小距离。在半径为$\varepsilon'(\varepsilon' < \varepsilon)$的圆形区域内，已然包含 5 个点（等于$minPts$的值）。右图描述了可达距离。在$p$是一个核心点的前提下，$q_1$和$p$之间的可达距离大于这两个点之间的实际距离，$q_2$和$p$之间的可达距离等于这两点之间的实际距离。

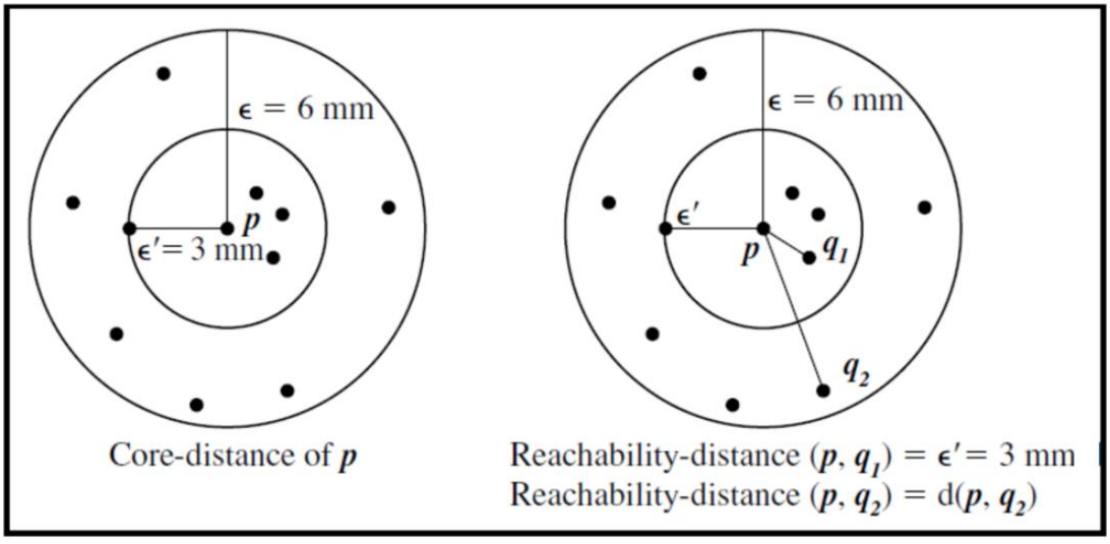

图 3-4 核心距离和可达距离

OPTICS 算法通过计算，返回数据点和它们对应的可达距离。返回的数据点按序排列，排序规则要求空间上最近的数据点要相邻。基于这些信息，就可以构造可达性图，沿 x 轴是按顺序排列的数据点，沿 y 轴是可达距离，如图 3-5 下方所示。图 3-5 左上方表示原始数据集，右上方绘制出了 OPTICS 算法

的生成树。不同的颜色表示不同的类簇。同属于一个类簇的数据点之间的距离更近，它们与相邻点之间的可达距离也较小，因此，在可达性图上，就会呈现出下陷的区域，密度越大的类簇，下陷的程度也越大。另外需要补充说明的是，黄色区域的数据点被视为噪声点，在其可达性图上找不到下陷的区域。

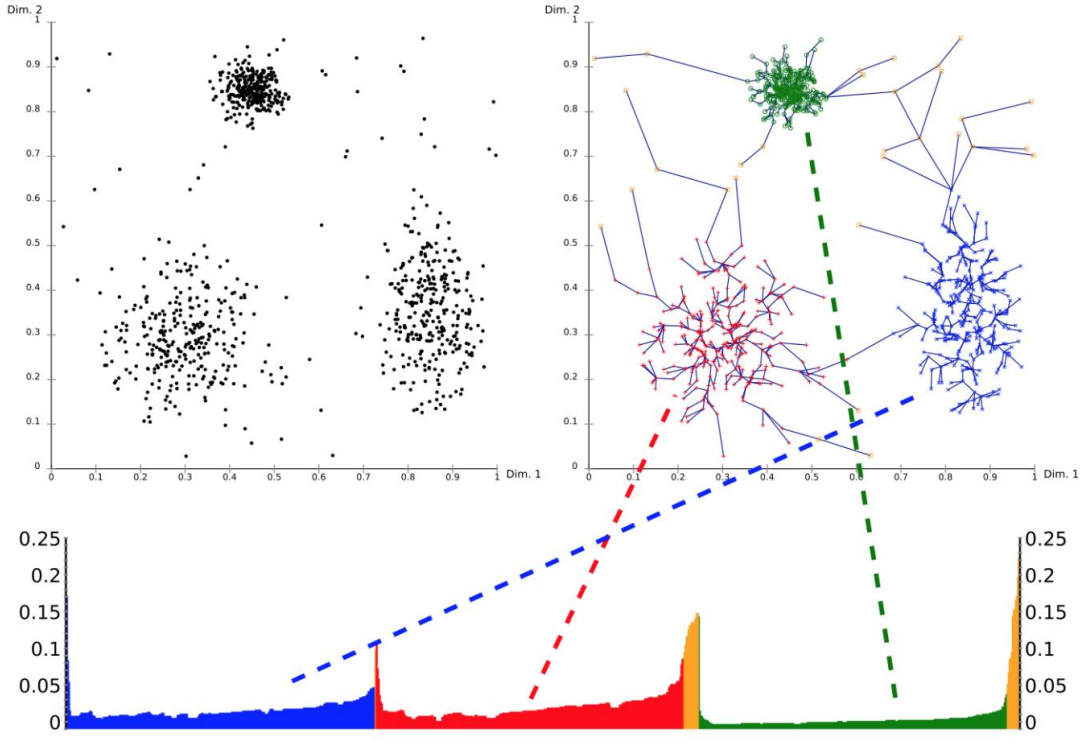

图 3-5 OPTICS 聚类提取

在了解完可达性图之后，我们现在介绍聚类提取流程。主要有两种方法。一种方法是检查可达性图，并在 y 轴上固定地设置一个合适的ε值，然后在数据集中的每个类簇上应用ε值，这样做的结果与参数（ε和$minPts$）设置相同的 DBSCAN 算法的结果一样。另一种做法是由一个自动程序为每一个类簇选择最合适的ε值，它通过定义可达性图上的每一个类簇边界内的最小陡度来实现。Ankerst 等（1999）[28]详细阐述了具体的实现过程。就本文的研究工作来说，它的关键之处在于成功地解决了为每一个类簇选择ε这一难题。它使整个聚类过程接近于无参数化，投资者只需要设置$minPts$这一个参数。

3.3 均值回归和平稳时间序列

Chan（2013）[14]指出，除了基于给定参数的交易回测，对时间序列进行统计检验也十分重要，因为统计检验囊括了时间序列的全部信息。而且，如果某个时间序列过程通过了统计检验，我们会对交易策略的盈利能力更有信心。

如果一个时间序列过程的联合概率分布不随时间或空间的改变而改变，就被定义为强平稳过程。对交易者来说至关重要的是，时间序列过程的均值和方差不会随着时间或空间的变化而变化。仅满足均值和方差不变性的时间序列过程被定义为弱平稳过程。对于金融时间序列分析来说，满足弱平稳性即可。

根据 Alexander 等人的研究[11]，价格、利率和收益等时间序列数据可以被认为是一阶单整序列，即经过一次差分后就变成平稳时间序列，而收益率可以被认为是平稳时间序列。但收益率无法被交易，只有价格是可交易的，而价格一般是一阶单整序列，即非平稳时间序列。因此，在只有一项资产的情况下，交易者无法利用平稳时间序列进行交易。当存在两项资产 x 和 y 的情况下，如果 $x(t)$ 和 $y(t)$ 之间存在协整关系，即存在协整系数 β 使得 $z(t) = x(t) - \beta \times y(t)$ 是平稳时间序列，交易者就能利用 $z(t)$ 进行交易。

平稳时间序列与统计套利有关的最重要的特征是均值回归特性。$z(t)$ 是一个平稳时间序列意味着：一方面，$z(t)$ 的均值不随着时间推移而改变。这表明，两项资产之间的协整关系是一种长期关系。另一方面，$z(t)$ 的均值不变性会让两项资产的价格始终保持一种“捆绑”关系，即在协整关系持续有效的前提下，其中一项资产的价格相对于另一项资产的价格不会显得过高或者过低，如果在某一时刻 $z(t)$ 的值偏离长期均值的幅度很大，那么它接下来将大概率向长期均值方向运动，表现出均值回归特性。

当两项资产存在协整关系时，那么导致它们各自的价格时间序列呈现非平稳的因素应该是相同的因子，或者用金融术语来说就是，这两项资产有相似的风险敞口，因此它们的价格才会同向波动。例如，身处同一行业的不同公司的股票、WTI 原油和布伦特原油等等，就可能存在着协整关系。

下面介绍几种用于检验时间序列的平稳性、协整关系和均值回归特性的

相关技术。

3.3.1 Augmented Dickey-Fuller 检验

Augmented Dickey-Fuller(ADF)检验是 Dickey 和 Fuller 于 1979 年提出的一种假设检验方法[29]，旨在检验某个时间序列过程是否存在单位根。单位根这一名称来源于自回归多项式在单位圆上是否有根值。存在单位根即表明该时间序列是非平稳过程。

对时间序列过程可以分别建立如下三种类型的模型：

$$\Delta y_t = \gamma y_{t-1} + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_k \Delta y_{t-k} + \epsilon_t \quad (3-5)$$

$$\Delta y_t = \alpha + \gamma y_{t-1} + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_k \Delta y_{t-k} + \epsilon_t \quad (3-6)$$

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_k \Delta y_{t-k} + \epsilon_t \quad (3-7)$$

其中$\Delta y_t = y_t - y_{t-1}$，$\alpha$是截距项，$\beta$是时间趋势项的系数，$k$是自回归过程的滞后阶数。ADF 检验的原假设为$\gamma = 0$，备择假设为$\gamma < 0$。检验统计量$t = \hat{\gamma}/\widehat{\sigma_\gamma}$,其中，$\hat{\gamma}$为$\gamma$的 OLS 估计量，$\widehat{\sigma_\gamma}$为$\gamma$的标准误。Dickey 和 Fuller 经研究发现上述统计量在$\gamma = 0$这一原假设成立的条件下虽然不服从 t 分布，但其极限分布存在，被称之为 Dickey-Fuller 分布。Dickey 和 Fuller 编制了 ADF 检验临界值表，后来 Mackinnon 对临界值表加以扩充，形成了目前广泛使用的 Mackinnon 临界值表。如果上述 t 统计量小于显著性水平对应的临界值，则在该显著性水平上拒绝原假设。

3.3.2 协整检验

协整检验由两位计量经济学家 Engle 和 Granger 于 1987 年首次提出[30]。在介绍协整这一概念之前，必须先阐明单整的概念。如果一个时间序列经过 d-1 次差分后仍不平稳，但经过 d 次差分后就变为平稳，则称该时间序列是 d 阶单整的，记为 I(d)。如果多个 d 阶单整的时间序列变量的某个线性组合的单整阶数小于 d，则称这些时间序列变量之间存在协整关系。在配对交易这一研究背景下，如果一组非平稳的 I(1)变量的某个线性组合是一个平稳的 I(0)变量，则它们之间存在协整关系。具体而言，考虑两个一阶单整的时间序列变量y_t和

x_t，协整关系意味着存在常数α和β，使得

$$y_t - \beta x_t = \alpha + \mu_t \qquad (3-8)$$

其中，μ_t是平稳时间序列。协整关系吸引我们的地方在于，它提供了一种方法来人工合成一个可用于交易的平稳时间序列。而在未经处理的原始金融数据当中，要找到一个平稳时间序列是一项极其困难艰巨的任务。

协整检验可以确定一组时间序列变量之间存在的稳定的、长期的关系。最常用的协整检验方法是 Engle-Granger 两步法（简称 E-G 两步法）和 Johansen 检验。对于两个证券价格时间序列，一般采用 E-G 两步法，其具体步骤如下：

1、使用 ADF 检验法，检验时间序列 y_t 和 x_t 是否存在单位根。如果原时间序列存在单位根，但是一阶差分后的时间序列不存在单位根，则继续执行步骤 2。

2、使用普通最小二乘法对 y_t 和 x_t 进行线性回归，得到残差序列 e_t。

3、使用 ADF 检验法，检验残差序列 e_t 是否存在单位根。如果拒绝原假设，即残差序列 e_t 不存在单位根，则时间序列 y_t 和 x_t 之间存在协整关系。

Armstrong（2001）[31]指出，E-G 两步法的一个主要问题是，自变量和因变量的选择可能会导致不同的结论。Johansen 检验方法，专门用于在有两个以上证券的情况下检验协整向量是否存在。本文仅在两只股票之间配对，Johansen 检验方法超出了本文的研究范围。

3.3.3 Hurst 指数

Hurst 指数也可以用来判定时间序列的平稳性。它的逻辑是：以几何布朗运动为参照，平稳时间序列的发散速度理应慢于几何布朗运动。Hurst 指数是一个标量值，帮助研究者识别某个时间序列过程是呈现均值回归或随机游走态势，还是存在长期趋势。[32]

Hurst 指数有很多种计算方法。本文采用如下方法对 Hurst 指数进行估计，该方法的思想是用对数价格序列的方差来估计时间序列的发散速度。对于某个任意的时间滞后项τ，对数价格序列的方差表示为：

$$Var(\tau) = < |\ln(t+\tau) - \ln(t)|^2 > \qquad (3-9)$$

一般而言，在几何布朗运动情形下，$Var(\tau)$和τ成正比：