学 校 代 码 10459

学号或申请号

密 级

郑州大学

硕士学位论文

基于聚类分析和主成分分析的股票市场研究

作 者 姓 名：郑 扬

导 师 姓 名：贾军国 张郑阳

学 科 门 类：理 学

专 业 名 称：应用数学（金融工程）

培 养 院 系：数学与统计学院

完 成 时 间：2017 年 5 月

A thesis(dissertation) submitted toZhengzhou Universityfor the degree of Master（doctor）

Research To Our Country Stock Market Basing OnClustering AnalysisAnd Principal Component Analysis

By Yang Zheng

Supervisor: Prof. Junguo Jia ;Zhengyang Zhang

Applied Mathematics（Financial Engineering）

SchoolofMathematicsandStatistics

May ,2017

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律责任由本人承担。

学位论文作者：郑扬　　日期：　　年　　月　　日

学位论文使用授权声明

本人在导师指导下完成的论文及相关的职务作品，知识产权归属郑州大学。根据郑州大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权郑州大学可以将本学位论文的全部或部分编入有关数据库进行检索，可以采用影印、缩印或者其他复制手段保存论文和汇编本学位论文。本人离校后发表、使用学位论文或与该学位论文直接相关的学术论文或成果时，第一署名单位仍然为郑州大学。保密论文在解密后应遵守此规定。

学位论文作者：郑扬

日期：　　年　　月　　日

摘要

中国股市从无到有，发展至今日渐完善，现已颇具规模。如何在不同板块众多种类的股票中挑选出具有投资价值的股票已经成为一个亟待解决的问题。面对股票的大量数据信息，财务数据、基本面数据、价格数据，如何从这些大数据高纬度的信息中找到有用的数据是一个值得研究的问题。我们已经步入了大数据的时代数据挖掘等与之相关名词越来越多的出现在我们的视野中。在数据挖掘的众多方法中，聚类分析的应用极其广泛，它的基本思想是根据数据之间的相似性度量来把数据分类，一般相似矩阵构造的好坏会决定聚类效果的好坏。这从直观上也很容易理解，如果把一个衡量数据之间关系的度量刻画的越精确，结果当然就越能反映数据之间的关系。我们会发现对数据进行分类之后仍然有很多问题需要解决，因为股票分类完成后还需要筛选出真正具有投资价值的股票。

针对上述问题，本文选取了中国股市不同板块间的 58 只股票，首先运用聚类分析方法对所选样本的收益率数据进行分类，把具有同质性（收益率密切联系）的股票归为一类，然后对每一类股票按各个财务指标用主成分分析的方法进行排序，这样我们便能筛选出不同类别之间具有投资潜力的一些股票。本文中运用的聚类分析方法有 K-means 和谱分析的方法，并将这两种聚类算法进行了对比。而且由于不同类别间的股票具有不同的收益特性，从不同的类别间进行选择相当于降低了投资的风险，也就是所谓的不把鸡蛋放在同一个篮子里，这对于投资者而言意义深远。

关键词: 股票市场，聚类分析，K-means，谱聚类，主成分分析

Abstract

China's stock market comes into being from scratch and now has been a considerable size.Stock market gradually improved and the intrinsic value of the return is the future direction of the stock market development.Therefore,blue chip stocks higher investment value will grow increasingly sought after by investors.Speculation in the past that high prices a serious deviation from the value of the phenomenon will be gradually corrected.How to choose the different sections in many types of stocks has investment value of the stock has become a serious problem.Faced with large amounts of data information, financial data, fundamental data and the stock price data, how to find useful data is a problem worthy of study data from these big high latitudes of the information.We have entered the era of big data,and data mining and other terms associated with the emergence of more and more come in our field of vision.Cluster analysis as data mining is an important method of application is extremely broad.The basic idea of clustering analysis is based on the similarity between metrics to the data classification, and general similarity matrix structure will determine the quality of the clustering effect,which is easy to intuitively understand.Because if a measure of the relationship between the data measurement more accurate depiction of the results of course, the more it can reflect the relationship between the data We will discover the following data classification still exist many problems to be solved, because when the stock classification is completed,we need to screen out the real investment value of the stock.

In response to these problems，this paper selects the Chinese stock market between different sections of 58 stocks.Firstly,we use cluster analysis method to classify the selected sample according to the yields of stocks.Next,the stocks which have homogeneity (whose yield have close contact) is classified as a class.Then,We sort stocks for each individual financial indicators used by the principal component analysis method. So that we will be able to filter out some stocks which have

investment potential between different categories.This paper uses the K-means clustering methods and spectral analysis clustering methods,and compares the differences between the two clustering methods. Besides,since the stocks between different categories have different earnings characteristics, which has reduced the risk of investment when choosing stocks from different categories.The so-called do not put your eggs in one basket.This has far-reaching significance for investors

Key words:

Stock market， Cluster analysis， K-means， Spectral clustering， Principal component analysis

目录

图和附表清单

符号说明

样本：$x_i=\left(x_{i1},x_{i2},...x_{ip}\right)^T$　　i=1, 2. . . n

样本 x_i 和样本 x_j 之间的距离：$d\left(x_i,x_j\right)$

$c_{\alpha\beta}$：变量 x_α,x_β 之间的相似系数

S：样本 $x_1,x_2,...x_n$ 协方差矩阵

R：样本 $x_1,x_2,...x_n$ 相关系数矩阵。

L:　Laplace 矩阵

W：相似矩阵

G：谱图

Cut 值：$cut(A,B)=\sum_{i\in A,j\in B}w_{ij}$

因子矩阵：

$$X=\begin{pmatrix}x_{11},x_{12},\cdots,x_{1p}\\x_{21},x_{22},\cdots,x_{2p}\\\cdots\\x_{n1},x_{n2},\cdots,x_{np}\end{pmatrix}$$

λ：主成分的方差

F_1,F_2,F_3 ...:综合主成分

1 引言

随着我国经济高速发展，人们的金融意识日渐加强。作为经济市场的一个重要组成部分—股票市场，正逐渐走向成熟和规范，越来越多的投资者将目光转移到股票。历史证明股票不仅在过去为投资者提供了长期可观的经济收益，也会在将来为投资者提供良好的投资机遇。

随着市场化经济的不断深入，我国的经济建设也步入了高速发展阶段。金融市场上日渐丰富的理财产品更是层出不穷，随着理财意识的不断增强，越来越多的人开始关注理财产品，关注投资。作为金融市场的主力军，股票市场的稳定发展，不仅有利于经济的发展，更对社会的稳定起到了重要的作用。

然而我们都知道“股市有风险，入市须谨慎”。股票市场的变幻无常，想要从股市中赚钱，仅凭借运气是玩玩不可的。投资过股票的人都知道，股票价格涨跌无常，股市更是变化莫测，投资者想要在股票市场中获取良好的收益，就必须对股票进行深入和细致的研究，唯有如此才能成为一个成功的投资者。一般而言，对股票的研究不仅包括对上市公司历史业绩、财务状况、发展前景等以基本面为主的研究分析，还包括技术分析。

投资者要想获得满意的投资回报，必须考虑投资哪些行业以及行业中的具有投资潜力的股票。另外，中国股市上很多投资者是散户，不具备专业化的分析能力和思维方式。一般对股票市场的分析分为基本面分析和技术分析，很多投资者会忽略基本面分析的问题，但事实上基本面才是最能反映股票投资潜力的最有用的信息。股票的基本面分析包括宏观分析、中观分析和微观分析三大部分。宏观分析是指对国家经济、政治、文化的分析，微观分析是指对公司的分析，中观分析是指行业和地区的分析。股票的基本面分析从微观层面上主要包括四个方面：一是该上市公司处于什么行业，该行业是受到政策鼓励、扶持还是受到政策限制的；二是该上市公司经营状况如何，营业收入、盈利水平等财务指标，同比数据是上升还是下降；三是分析该上市公司有无发展潜力；四是分析股票价格波动，是否有一些机构的主力资金经常关注并且入驻其中。所以，最重要的事情是对各家公司的分析、评价，挑选真正的具有内在价值的股票做为投资标的。对公司进行财务情况评价都是通过多指标分析的，所以第一

件事情是对多指标进行降维处理，找到具有决定性的重要指标作为分析指标，对我们至关重要。

股市通常按照不同的角度将众多股票划分为不同的版块，每个版块的股票具有相似的特性。面对众多股票繁杂的财务数据，如何选出具有投资价值的股票，这是一个值得研究的问题。然而关于股票的数据有很多，衡量一只股票也不能仅仅只用某一个或某几个指标，也就是说每个股票有很多的特征属性。在数学中这就是一个高纬度的大样本数据的分类处理问题，首先基本的想法是对数据进行降维或者找到能处理高维数据的方法。数据降维的方法也有很多，比如说主成分分析以及基于主成分分析的各种推广。其次是分类问题，如何对大样本数据进行分类，就是一些基本的聚类方法的运用，比如谱聚类，稀疏矩阵聚类等，这些都归于多元统计分析的范畴。

近年来，“大数据”和“数据挖掘”越来越多的出现在人们的视野中[1]。大数据，顾名思义就是数据量很大，维度可能也很高的数据。我们常常说进入了人一个大数据的时代，大数据为我们提供了很多研究依据，我们不用担心统计分析没有意义，因为众所周知，统计分析都是建立在大样本数据之上的。什么是数据挖掘？相信很多人对此并不是十分了解，数据挖掘是一种通过数理模式来分析企业存储的大量资料，用以找出消费者喜好和行为的方法。这当然是比较片面的说法，从专业的角度来说，数据挖掘是从大量的数据中搜索出隐藏于其中的有着特殊关系的信息的过程。一般情况下可以分为三个步骤：数据准备，找寻规律、规律表示。

数据挖掘的任务主要包括聚类分析、演变分析、关联分析等，并借助计算机的力量，运用机器学习、模式识别等方法来达到目标。本文中我们便利用聚类分析的方法对我们所选取的样本数据进行分析处理[2]。

1.1 聚类方法研究意义及发展现状

聚类分析是研究分类问题的多元数据分析方法，依据数据本身的分布特征或事物的某些属性，把事物划分成不同的类，是类间的相似性尽可能的小，每一类之间的相似性尽可能的大。聚类方法分为动态聚类和静态聚类，动态聚类适用于海量数据之间的聚类分析，但一般情况下要指定聚类的数目，这对整个动态聚类的结果意义重大，所以经常需要做的工作是凭借尝试不同的聚类数目

进行聚类，经过比较筛选得到最优的聚类质量。

聚类分析是一种重要的数据分析手段，对数据对象进行区分和分类。通过数据的聚类分析，能有效地发现隐含在数据中的分布特性，进而为进一步充分并且有效地利用数据奠定了良好基础。随着信息技术的迅猛发展，对数据聚类所面临的不仅仅是数据量越来越大的问题，更重要的还有数据的高维度问题。维度在数据中的难点是非常常见的现象，尤其是在高纬度的环境中，Bellman 最早提出维度灾难这一说法。泛指数据过高的维数引发在数据处理中的一系列问题。本文以此为出发点，运用高维数据聚类方法对数据进行处理[15][16]。

对所考察的对象进行聚类分析帮助我们把股票划分成不同的类型，而这些划分是根据收益率进行的。所以实际上我们是把具有不同收益率性质的股票区分了出来，这对我们做投资而言意义重大。我们都知道，投资的时候要做到“不要把鸡蛋放在同一个篮子中”，聚类分析相当于为我们提供了一个有力的的理论支撑。投资于不同收益率性质的股票可以帮助我们实现投资组合的分散化，大大降低投资组合的风险，所以对股票进行聚类分析对于投资组合而言意义重大。

国内研究数据挖掘相对较晚，然而近几年发展迅猛，越来越多的学者投身到数据挖掘相关领域中，司文武和钱涛很巧妙的将谱聚类引入到半监督中，谱聚类是聚类分析方法中的一种。它是利用少部分标签的数据来辅助大量未标签的数据，以此来进行非监督的学习，并且提高了聚类方法的性能。利用被标记过的数据的信息，点与点之间形成距离矩阵，并对之进行调整。谱聚类也是在此基础上进行的。大量实践表明，该方法获得的聚类性能要优于一般的方法。

谱聚类建立在谱图理论的基础上，其本质是把聚类转化成图论中的最优划分问题，有非常广泛的应用前景。本文作为一篇集理论研究与实证分析于一体的综合性文章，旨在向读者引入这种比较好的思想方法，对广大投资者在进行股票投资过程中选择绩优股或龙头股提供好的理论依据。

目前对谱聚类的研究还处于起步阶段，没有完整的理论体系用以描述谱聚类的性能和局限性。因此国内对这方面的研究非常少。我们相信未来的研究一定会朝这个方向发展。

1.2 主成分分析方法的研究背景

在用统计分析方法研究多变量问题时，变量个数太多会增加课题的复杂性。

人们自然希望减少变量个数但同时得到较多信息。大多数情况下，变量之间是存在相关性的，具有相关性的变量，可以通俗地理解为变量之间所表达的一定的重叠信息。主成分分析是对原来提出的所有变量，将重复的变量(变量关系联系紧密)中删去多余的变量，建立尽可能少的新的综合性的变量，使得这些新变量之间是两两不相关的，并且这些新变量在反映信息时要尽可能的保持原有的信息。

主成分分析（主分量分析）是一种设法将原来的变量重新组合成一组新的互相无关的几个综合性的变量，然后在综合变量中，根据需求选取较少却能尽可能的表达原有因子信息的变量。在这种方法也是数学上用来降维的一种方法。

主成分分析不仅应用于数学建模、数理分析，还在人口统计学、分子动力学模拟等发挥着重大的作用，是一种常用的多变量统计分析方法。

本文采用主成分分析公司的财务指标从而对股票进行投资价值的分析，具体解释如下：根据企业的经营状况指标，采用多指标的分析方法，对公司给出综合性评价。要想对公司做出正确地、合理地判断，当然需要尽可能多的选取指标，每个指标在不同程度上都会反映所研究问题的信息，有的甚至几十个、上百个指标。越多的指标反应的信息就越多，但这样会在计算产生很大的麻烦，不仅如此，还会对以后的结果分析产生很大的麻烦。因此，我们需要尽可能的减少指标个数，但有不能降低指标中反应的信息。在数学上，这就是高维数据的处理问题。在解决处理高维数据时，进行降维是十分必要而且重要的，因为当指标之间存在相关性时，可以说相关的指标之间反映的信息有一定的重叠。所以对数据进行降维处理是十分必要的。主要原因如下：

首先高维数据的多重共线性会导致解空间的不稳定，从而导致结果的不连贯；其次高维空间本身就可能具有稀疏性，比如一维正态分布有 68%的值会落在正负标准差之间，而在十维空间上时只有 0.02%的概率落在正负标准差之间；最后，过多的变量会妨碍规律的寻找。降维的目的主要有以下三个原因：可以减少变量的个数并确保变量之间是相互独立的，除此之外可以提供一个框架来对结果做出解释。

于是，在数学上，主成分分析就是这一问题的多元统计方法。它是将多个具有相关性的复杂指标，通过一定的方法，减少为少量且不相关的指标。主要目的有两个，一是降维，二是解释各个变量之间的内在关系。

2 研究方法

2.1 聚类分析

人类一般是通过对认识的对象进行分类的方式来认识世界。研究分类问题的一种重要的多元数据分析方法便是聚类分析，聚类分析有极其广泛的应用背景。经济学中，为了了解不同地区城镇居民的收入及消费情况，需要划分不同的类型去研究；在产品质量管理中，要根据各产品的某些重要指标将其划分成一等品、二等品、三等品等。

在经济、社会以及人口的研究中，存在着大量需要分类研究和构造分类模式的问题。历史的经验表明，人们主要靠经验和专业知识对所研究的问题进行定性分类，这会导致分类结果带有主观性，不能很好地揭示客观事物的内在本质差别。特别的，对于多因素和多指标分类问题，定性分类很难实现准确的分类。聚类分析便随之产生了。同样的，聚类分析还适合于多因素、多指标的分类问题，并可以实现较为准确的分类（陈琦[7]）。

聚类分析是多元统计方法中的一种。把研究的途径是把所研究的对象进行分类，使得同一类中对象之间的相关性比其他类的相关性更强。我们认为所研究的样本之间存在程度不同的相关性，有的相似性比较大，而有些相似性比较小。根据样本的多个观测指标，把能够度量样本相似程度的变量用统计量找出来，以此为基础，将某些相似度大的样本聚成一类，另一些相似度大的聚成一类。

2.1.1 聚类分析的基本思想

聚类分析的基本思想源于划分理论，该算法将聚类问题看做图论中的无向图多路划分问题。数据点可看做无向图 G（V，E）的顶点 V，$E=\{w_{ij}\}$ 是加权边的集合，可以度量基于某种相似度计算的两点间的距离。聚类数据的相似度量矩阵用 W 表示，将之看作图的邻接矩阵。于是，该邻接矩阵包含了聚类所需的所有信息。接下来我们定义一个划分准则，并优化这一准则，优化的标准是使得同一类的点具有较高的相似性，不同类的点与点之间具有较大的相异性。聚类分析一般寻求客观的分类方法，在进行聚类分析以前，对总体到底有几种

类型并不知道（究竟分几类较为合适需要从计算中探索调整）。

2.1.2 谱图相关知识

2.1.3 图分割

图分割（Graph Cut）是把一个连通图的某些边切断，这样一个连通图就变成一些联通的子图。被切断的边的权重之和称为 cut 值。直观地理解，具有较大权重的边不会被切断，这意味着相似的点被保存在同一个子图中，反之，不太相似的点(权重较小)则被分开。

用邻接矩阵的形式表示连通图连通图，记为 W。若两个点不相连，则权重值为零，否则设记为w_{ij}。设 A 和 B 分别为连通图的两个子集，$A \cap B = \varnothing$，则 A 和 B 之间的 cut 定义为：$cut(A,B) = \sum_{i \in A, j \in B} w_{ij}$

2.1.4 样本间的相似性度量——距离

样本：受审查客体的反映形象或其自身的一部分，是从总体中抽出的一部分个体。样本中所包含的个体数目称样本容量或含量，用符号 N 或 n 表示。样本越大从总体中提取的信息就越多，对总体的代表性就越好，因此一般情况都抽取大样本进行研究。

设有 n 个样本的多元观测数据：

$x_i = \left(x_{i1}, x_{i2}, ... x_{ip}\right)^T$　　i=1, 2. . . n

这时，每个样品可看成 p 元空间的一个点，n 个样本组成 p 元空间的 n 个点。

我们自然用各点之间的距离来衡量各样本之间的相似性程度（或靠近程度）。

设$d\left(x_i, x_j\right)$是样本x_i和样本x_j之间的距离，一般要求它满足以下条件：

$d\left(x_i, x_j\right) \geq 0$，且$d\left(x_i, x_j\right) = 0$当且仅当$x_i = x_j$；

$d\left(x_i, x_j\right) = d\left(x_j, x_i\right)$；

$d\left(x_i, x_j\right) \leq d\left(x_i, x_k\right) + d\left(x_k, x_j\right)$；

注：有些距离不满足上述最后一条，在聚类分析中广义上我们仍称它为距离。

常见的距离的分类有很多种，如欧式距离，明科夫斯基距离，切比雪夫距离等，我们在下面的分析中用明科夫斯基距离，明科夫斯基的距离按如下定义：

$d\left(x_i,x_j\right)=\left[\sum_{k=1}^{p}|x_{ik}-x_{jk}|^m\right]^{\overline{m}}$ m≥1

当 m=2 时为欧式距离，m=1 时为绝对距离。

2.1.5 变量间的相似性度量——相似系数

用相似系数来衡量变量之间的相似性程度。假设用 $c_{\alpha\beta}$ 表示变量 x_α,x_β 之间的相似系数，$c_{\alpha\beta}$ 满足：$|c_{\alpha\beta}|\leq 1$ 且 $c_{\alpha\alpha}=1$；

$c_{\alpha\beta}=\pm 1$ 当且仅当 $x_\alpha=cx_\beta(c\neq 0)$；

$c_{\alpha\beta}=c_{\beta\alpha}$.

$c_{\alpha\beta}$ 的绝对值越接近于 1，变量 x_α 和 x_β 的关联性就越大。

本文中我们运用相关系数的定义如下：

设样本 $x_1,x_2,...x_n$ 算得协方差矩阵 S 与相关系数矩阵 R。

设 $S=\left(s_{ij}\right)_{p\times p}$，$R=\left(r_{ij}\right)_{p\times p}$

则变量 x_α,x_β 的相关系数为 $r_{\alpha\beta}=\dfrac{s_{\alpha\beta}}{\sqrt{s_{\alpha\alpha}s_{\beta\beta}}}=\dfrac{\sum_{i=1}^{n}\left(x_{i\alpha}-\overline{x_\alpha}\right)\left(x_{i\beta}-\overline{x_\beta}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{i\alpha}-\overline{x_\alpha}\right)^2\sum_{i=1}^{n}\left(x_{i\beta}-\overline{x_\beta}\right)^2}}$

$r_{\alpha\beta}$ 是变量 x_α 的观测值 $\left(x_{1\alpha},x_{2\alpha},...x_{n\alpha}\right)^T$ 与 x_β 的观测值 $\left(x_{1\beta},x_{2\beta},...x_{n\beta}\right)^T$ 间的相关系数。

2.1.6 K-means 聚类算法

K-means 算法的基本思想：以空间中 k 个点为初始中心进行聚类，把靠近它们的对象分别归类。通过迭代，逐次更新各聚类中心的值，直至得到最优的聚类结果。假设把样本集分为成 c 个类别，算法描述如下：

(1)适当选择 c 个类的初始中心；

(2)在第 k 次迭代中，对任意一个样本，分别求出其到 c 个中心的距离，将该样本归到距离最短的中心所在的类；

(3)利用均值等方法更新该类的中心值；

(4)对于 c 个聚类中心，如果利用(2)(3)的迭代法更新后，值保持不变，则迭

代结束，否则继续迭代。

2.1.7 谱聚类法

谱聚类算法起源于谱图划分，是一个非常流行的高性能方法。下面先介绍一下关于图论的一些基础知识，这些知识很简单，但对于谱聚类算法起着重要的作用[3]。

定义 1.1：令 V=$V=\{v_1,v_2,...v_N\}$表示 N 个数据点构成的集合，E 代表 V 中任意两点连线构成的集合，则 G=(V，E)构成一个无向图。

定义 1.2 ：对 1.1 中集合 E 中每个元素赋予一个非负的数值w_{ij}，i，j=1，2...N

我们可以得到一个邻接非负矩阵$w=\left(w_{ij}\right)_{i,j=1,...N}$，该矩阵就是无向图的权重矩阵。如果$w_{ij}=0$说明点$v_i$和$v_j$之间没有连接，且该矩阵是对称的。

把数据点看作无向图 G(V，E)的顶点 V，E={ w_{ij} }（加权边的集合）表示基于某一相似性度量计算的两点间的相似度。W 表示待聚类数据点的相似度矩阵，它包含了聚类所需的所有信息。然后我们定义一个划分准则，使得同一类内的点具有较高的相似性，而不同类之间的点具有较高的相异性。

相似矩阵通常用 W 表示，有时也称为亲合矩阵。该矩阵的定义为: $\boldsymbol{w_{ij}} = \mathbf{e}^{\left(-\frac{d(s_i,s_j)}{2\sigma^2}\right)}$

其中σ为事先定义好的参数，需人为确定；s_i表示每个数据样本点，d(s_i,s_j)一般取为$\|s_i-s_j\|^2$，。

图中每个顶点的度是由相似矩阵的每行元素相加得到的，某顶点$v_i\in V$的度表示如下:

$$d_i=\sum_{j=1}^{n} w_{ij}$$

由所有对角元素d_i构成的对角矩阵即为度矩阵，通常用 D 表示。

拉普拉斯矩阵分为规范拉普拉斯矩阵和非规范拉普拉斯矩阵。

非规范拉普拉斯矩阵定义为：

L=D-W

谱聚类算法的一般框架:

输入：n 个数据点集合$\{x_1, x_2, ...x_n\}$

输出：数据点集的划分结果

算法：

Step 1：定义数据点间的相似性度量；

Step 2：根据该相似性度量，构造数据点集的相似度矩阵 W；

Step 3：计算 Laplace 矩阵 L=D-W；

Step 4：分别对规范或非规范的拉普拉斯矩阵：$Lf = \lambda Df$ 或 $Lf = \lambda f$ ，计算其特征值和特征向量；

Step 5：将数据点映射到由一个或多个特征向量确定的低维空间中；

Step 6：根据数据点在新的空间中的表示，划分数据点到两类或多类中。

2.1.8 谱聚类算法与 K-means 聚类算法的比较

K-means 算法优点：简单有效，不太受初始化条件的影响。但同时也有一些致命的缺点，比如不能处理不同尺寸和不同密度的簇、非球形簇等，并且 K-means 算法经常会使问题陷入局部最优的境地，不一定能得到全局最最优解。K-means 算法在进行非线性聚类时结果略显粗糙[4][5]。

谱聚类算法不局限于具体的模型，是一种基于相似矩阵建立在凸论基础上的的特征提取方法，与主成分分析有密切的关系。谱聚类算法目前应用广泛，因为对于谱聚类算法而言，解决线性问题不会受到初值以及局部收敛等问题的制约。进行谱聚类分析时不需要假设数据来自于 Gauss 分布，而且可以很好地应用于非线性聚类。如果相似矩阵具有优良的性质，谱聚类算法的效果会让人十分满意。但谱聚类算法本身也存在一定难度，因为构造一个好的相似矩阵并不是一件简单的事情。构造的相似矩阵不同，谱聚类的结果也会不太相同，也就是说谱聚类对于相似矩阵来说性能并不稳定，谱聚类算法同样不能保证聚类结果能正确反映数据的真正结构。

2.1.9 谱聚类目前存在的问题

1、如何构造相似矩阵：谱聚类算法中相似矩阵的构造依赖于对不同数据样本之间定义的相似函数。在本文中，我们的算法上述提到的相似函数的公式，其中的参数σ是我们认为选取的。由于人为因素的影响，使得该方法的的确定性

受到一定的影响，并且使得方法的局限性显现了出来。对于这一点，我试图在本文之中进行改进，但囿于目前的参考文献并未有很大的改进。当前，绝大部分的文献中都没有通用的确定σ的方法。所以，对这一问题系统的研究将成为未来研究的一个热点[17]。

2、聚类的数目的确定：聚类数目多少对聚类结果的好坏有直接影响。目前已有的自动确定聚类数目的方法有：基于 Rcut 的谱聚类算法，基于距离的启发方法，以及非线性降维算法。令人遗憾的是，这些算法在一些特点非常明显的数据中并未有很好的结果。所以，聚类数目如何确定是一个关键性的难题。也未来研究的一个重要方向。在本文中，我们采用的聚类数目根据我们所选的股票板块确定。

3、如何选取拉普拉斯矩阵：从很多文献中，我们看到，聚类算法选取的拉普拉斯矩阵大体上有三种方式。遗憾的是，目前还没有确定每种方式的适用环境。有文献中提到，如果拉普拉斯矩阵对角线的元素相差不多时，三种方法得到的聚类结果基本相同，否则的话结果将差别很大。

4、如何将谱聚类算法推广到大规模的数据处理问题中：谱聚类涉及到求解特征值和特征向量，计算复杂度较高，进行大规模的计算难度较大。因此，研究一种更加高效、可扩展的算法将是一个非常有意义的课题[14]。

2. 2 主成分分析

2.2.1 主成分分析的主要思想

主成分分析是一种数理统计方法，又叫做主成分回归分析法。它最早是由 K.皮尔森对非随机变量研究的时候引入的，接着 H.霍尔林在 1933 年将主成分分析这个方法推广到随机变量。这时候，主成分分析的方法才正式的进入我们的视野，我们一般将经过线性变换后的少数变量叫做主成分，这些主成分都是由原来的较多个数的变量通过线性组合转变过来的，并且这些主成分之间是互不相关的。这些互不相关的变量还要尽可能的保留原来变量所要表达的信息，也就是说，主成分分析方法可以有效的降低维数，同时保持了数据对方差的贡献程度。这样，有利于我们抓住主要矛盾，提高分析效率。

主成分分析主要是用来解决数据处理时的难题。当面对大量数据的时候，

数据的高纬度、以及数据之间的相关性给数据的计算与分析带来了不小的麻烦。倘若盲目的减少变来变量个数，很有可能损失重要的信息，所以给出错误的结论。由此，如何在保证不降低数据包含信息的同时，降低数据的维度，是一个亟待解决，且非常有意义的课题。为此，人们不断地探索和研究，主成分分析的主要思想也源于此，解决了数据高纬度、高相关性的问题，在数据分析方面得到了广泛的应用。

2.2.2 主成分分析的数学模型

主成分分析是怎么达到降低维度的目的呢？它是将原因子，通过线性变化，转化为新的主成分，并通过贡献率挑选合适的主成分，从而达到降维的思想。通俗的讲：假定共有 n 个因子，以及选取了 p 个指标反应所选样本，X_{ij} 表示第 i 家公司的第 j 个指标，即：

$$X=\begin{pmatrix} X_{11},X_{12}\cdots X_{1p} \\ X_{21},X_{22}\cdots X_{2p} \\ \cdots \\ X_{n1},X_{n2}\cdots X_{np} \end{pmatrix}$$

通过主成分分析的计算，将原有因子重新排列组合得到新的 p 个主成分。用 Z 表示。具体结构如下所示：

$Z_1=k_{11}X_1+k_{12}X_2+k_{13}X_3+\cdots+k_1X_p;$

$Z_2=k_{21}X_1+k_{22}X_2+k_{23}X_3+\cdots+k_{2p}X_p;$

$\cdots$

$Z_p=k_{p1}X_1+k_{p2}X_2+k_{p3}X_3+\cdots+k_{pp}X_p;$

其中，各个主成分之间互不相关。

记 λ 表示每个主成分的方差，则 $\lambda_i/\sum\lambda_i$ 表示第 i 个主成分 Z_i 的方差的贡献率。那么，$\sum_{i=1}^{k}\left(\sum\left(\lambda_i/\sum\lambda_i\right)\right)$ 为前 K 个主成分的累计贡献率。

累计贡献率是我们选取主成分的主要依据。我们选取主成分，使得累计贡献率达到我们要求的置信水平。这样，主成分会在一定程度上，尽可能的保持原有因子所包含的信息。该置信水平由投资者自己设定，通常是在 80%-90%之

间。

2.2.3 主成分分析的计算步骤与方法

结合上述主成分分析的主要思想、原理、以及数学模型，下面我们给出主成分分析的计算步骤归纳，具体如下：

构建原始因子矩阵，该矩阵的各元素即为原因子。如：

$$X=\begin{pmatrix} x_{11},x_{12},\cdots,x_{1p} \\ x_{21},x_{22},\cdots,x_{2p} \\ \cdots \\ x_{n1},x_{n2},\cdots,x_{np} \end{pmatrix}$$

因子矩阵标准化的过程，这一过程主要是因为各个因子度量单位的不同、

因为不同的因子有不同的单位，有的是数值，有的又是百分比，单位不同，这样的指标的数量级有着非常大的差异。理论上，不同单位的数据是不能够做加减乘除计算的，因为这样的计算没有意义，计算的结果也不合理，为了解决单位不一致的问题，标准化处理，也就是保持每一个变量满足平均值是 0，方差是 1，其中变换的公式是：$\mathrm{y}_{ij}=\dfrac{x_{ij}-\overline{x}_j}{s_i}$

（3）计算出相关矩阵 C，其中 C 为$C=Y^TY$

（4）计算相关矩阵 C 的特征值和特征向量（统计性质），以及贡献率与累计贡献率。

（5）并按大小进行排序，剔除不显著因子。也就是确定主成分的个数。从较多的变量当中提取较少的主成分是我们的目的，一般来说，提取的主成分应当使得累计贡献率达到一定的值，即置信度。置信度因人而异，有的人会将它设置的很高，比如说 95%，也有的人设为 80%。置信度越高，也就要选出的主成分对原因子的解释程度越高，其选出的主成分个数也会较多，因此，设定置信度也非常的重要。在本文中，我们将置信度设为 85%，即，主成分至少要包含原因子 85%的信息。

（6）确定线性预测函数。根据选出的主成分和对应的特征向量，我们可以给出以给出预测函数。各个主成分前面的系数为它们方差占全部方差的比重。

（7）结果分析。对预测值进行排序，可以给出样本客观公正的评价。

2.2.4 主成分分析的优点

主成分分析最大的优点在于能够把一组初始变量，也可称为原因子，通过特定方法线性组合后，有新的一组因子互不相关的因子重新组合，这组新的变量，叫做主成分。我们当然是希望选取较少的主成分，还能够尽可能多的表达原因子包含的信息。最重要的是，我们选取这组主成分中，最有价值，包含信息最多的主成分，舍弃包含信息较少的主成分，从而达到了降低维度的目的。为了更好地表达因子所包含的信息，在数学中，我们引入方差这一概念来衡量。总方差表达的是原来所有因子所包含的信息。每个主成分也对应有自己的方差，各自方差占总方差的大小，称为贡献度。方差的大小代表了该主成分所包含的信息，方差越大，表明该主成分所包含的信息越高，反之，方差越小，也就表明该主成分所包含的信息越少。因此。当原因子重新线性组合后，得到新的主成分，按照方差的从大到小进行排列，同理，也就是按照主成分所包含的信息程度来排列。最大方差的主成分，我们称为第一主成分，依次，称为第二主成分……。计算各个主成分的贡献度，以及累计贡献度，依次来对主成分进行取舍。

通过主成分分析方法后，原因子的数据可以由较少维度的主成分来线性组合，降低维度，从而给我们的计算带来了很大的便利，大大提高运算效率。其次，根据累计贡献度的计算，我们的对主成分进行取舍的时候，是要满足我们给定的置信水平的，故主成分能够很好的表达原因子所要表达的信息，不会造成信息缺失。最后，原因子之间是具有相关性，换句话说也就是各因子所表达的信息有重叠，这给后面的分析带来了很大的麻烦。而主成分之间具有了很好地性质，各个主成分之间是互不相关的，这样很好地解决了这一难题，给分析个因素的印象提供了方便。

2.2.5 主成分分析和聚类分析的区别

聚类分析和主成分分析方法都是数学分析的方法。聚类分析是通过对称矩阵的来分析相关性的问题。它的优点是简单直观，结论明确。缺点是没有办法处理大样本的情况，在实践中会发现，虽然样本之间存在紧密关系，但事实上却没有任何关系，聚类分析方法本身无法识别这类错误。

主成分分析的方法是将一组变量通过线性变换，转换为一组互不相关的变

量，从而达到降维的目的。它的优点在于降低维数后的变量还能反应原有因子的大部分信息，从而更加客观地做出更加科学的评价。但主成分分析也有一定的缺点，当主成分的符号既有正也有负的时候，指标的实际含义就不太明朗了。

3 股票分类实证分析

3.1 前提及假设

对股票最简单直接的分析方法就是对收益率序列进行分析研究，本文中选取 4 个板块的股票进行分析。众所周知，板块是指有共同特征的股票群。股票市场中的板块的划分依据有很多角度，而在每一板块中有几十甚至上百种股票。如何从众多股票及众多公司的财务数据中全面、准确地分析出各板块中的绩优股和龙头股是一个值得研究的问题，本文试从聚类分析的角度对上述问题做一些探讨[6]。

每个行业板块的收益率序列应该具有相似的性质，而不同的板块之间收益率序列的性质可能会有很大的区别。通常我们对所研究的股票进行聚类之后会在每个类别之中用主成分分析的方法，对股票的财务数据进行分析给出每个类别之中相对而言具有投资价值的股票[7]。

一般情况下，采用平均收益率评价公司，根据收益率从大到小排序，收益率越大，表明公司效益越好。然而收益率是一个特别简单的统计量，因此，本文我们不采用这种方法。另一个常用的的方法是相关分析。由于收益率序列间的相关系数只能反应线性相关性，对于非线性关系而言基本是不可行的。我们认为可以采用聚类分析方法来探讨不同板块股票之间的关系。在进行分类时，我们仍采用收益率序列。当经济处于不同的发展阶段，板块之间的表现也会有很大不同，这对我们的分析造成了一定的影响。由于经济周期的划分同样存在很大的不确定性，所以我们暂时不考虑不同经济周期对我们的划分结果的影响(劳兰珺[13])。

3.2 数据处理

为了消除原数据量纲的不同影响，我们先对数据进行处理，采用正态标准数学变换。常用的变换方法有两种：标准差变换和极差变换。在这里我们采用标准差变换。经变换后的各指标的均值为 0，标准差为 1。

指标标准化：

本文中对股票数据进行分类时，由于采用的是收益率序列，所以不考虑量

纲的不同（收益率量纲相同），但对每个分类结果进行主成分分析时，首先要对数据进行标准化。

3.3 样本数据的分类

3.3.1 K-means 算法分类

首先对数据进行处理，因为不同行业和板块的股票价格可能相差很大，首先把股票价格转化为收益率，使所有的数据具有可比性，这样的出来的结果才具有参考价值[18]。

我们选取了 4 个股票板块的数据进行分析（股票的分类来自于申银万国分类软件），四个板块分别为商业银行板块，房地产板块，港口板块和食品饮料板块。选取的股票名称和代码如下所示：

房地产板块：万科，世纪星源，深振业，深物业，沙河股份，中冠，招商地产，深深房，中洲控股，中航地产，泛海控股，华侨城，金融街，绿景控股，珠江控股，荣安地产；

港口板块：深赤湾，盐田港，珠海港，北部湾港，厦门港务，日照港，上港集团，锦州港，重庆港九，营口港.，皖江物流，天津港，唐山港，连云港，宁波港，大连港；

饮料板块：泸州老窖古井贡酒，五粮液，顺鑫农业，洋河股份，伊力特，金种子酒，贵州茅台，老白干酒，沱牌舍得，山西汾酒；

银行板块：中信银行，中国银行，建设银行，光大银行，工商银行，交通银行，农业银行，北京银行，南京银行，招商银行，民生银行，华夏银行，浦发银行，宁波银行，平安银行。

根据 K-means 方法，我们的分类结果如下：

第一类：万科 A，招商地产，中洲控股，泛海控股，华侨城，金融街，绿景控股，深赤湾 A，上港集团，皖江物流。

第二类：世纪星源，深振业，深物业，沙河股份，中冠，深深房，中航地产，珠江控股，荣安地产，盐田港，珠海港，北部湾港，厦门港务，日照港，锦州港，重庆港九，营口港，天津港，唐山港，连云港，宁波港，大连港。

第三类：泸州老窖古井贡酒，五粮液，顺鑫农业，洋河股份，伊力特，金

种子酒，贵州茅台，老白干酒，沱牌舍得，山西汾酒。

第四类：中信银行，工商银行，交通银行，北京银行，南京银行，招商银行，建设银行，民生银行，光大银行，华夏银行，浦发银行，农业银行，宁波银行，平安银行，中国银行。

3.3.2 K-means 分类结果及分析

分类结果如图 1—图 4 所示，每个图代表按照 K-means 方法分类的其中一类，分类结果并不是与板块完全一致，只要是因为我们采用的方法研究的是有关股票的更本质的性质来对股票进行分类,结果如表 3.1,表 3.2 表 3.3，表 3.4 所示。

从表 3.1,表 3.2 表 3.3，表 3.4 中我们可以看到：用 K-means 方法的分类结果基本上与板块的划分结果一致。这说明不同板块之间股票的收益率具有的特征确实不同。关于这一点很容易理解，因为不同的板块或者说是行业，行业的利益驱动因素不同。行业的发展是内部因素与外部因素相互作用的结果，利益驱动是推动行业发展的根本动力。在大的经济框架下，不同板块受到的影响因素基本相同，但所受影响却有很大的差别。一般情况下我们会认为宏观经济对各个企业的影响方向一致，所以我们暂不考虑宏观经济的影响。

我们的分类结果中，港口板块和房地产板块有所交叉，深赤湾 A，上港集团，皖江物流属于港口板块，但根据我们的划分结果将其归到房地产板块。世纪星源，深振业，深物业，沙河股份，中冠，深深房，中航地产，珠江控股，荣安地产属于房地产板块，聚类结果显示它们和港口板块的利益性质比较接近。这或许是由于房地产和港口板块之间有相互影响，而房地产和食品饮料及银行板块的联系相对较弱造成的。

K-means 方法是一种很简单经典的方法，在数据的聚类中应用非常广泛。前文中提到，K-means 方法有一定的缺陷。我们在进行实证分析的研究过程中也发现了这种现象，根据这种方法对数据进行分类时，每次的分类结果并不太一致，这可能是由于数据本身结构的复杂性造成的，因为前文中我们提到 K-means 算法不适合解决非球形簇的数据。因为我们对所选取的样本数据本身的结构并没有总体的定位，所以并不能保证 K-means 算法的精确性。

表 3.1K-means 分类第一类

万科 A000002.SZ
招商地产 000024.SZ
中洲控股 000042.SZ
泛海控股 000046.SZ
华侨城 A000069.SZ
金融街 000402.SZ
绿景控股 000502.SZ
深赤湾 A000022.SZ
上港集团 600018.SH
皖江物流 600575.SH

表 3. 2K-means 分类第二类

世纪星源 000005.SZ	深深房 A000029.SZ
深振业 A000006.SZ	中航地产 000043.SZ
深物业 A000011.SZ	珠江控股 000505.SZ
沙河股份 000014.SZ	荣安地产 000517.SZ
中冠 A000018.SZ	大连港 601880.SH
盐田港 000088.SZ	重庆港九 600279.SH
珠海港 000507.SZ	营口港 600317.SH
北部湾港 000582.SZ	天津港 600717.SH
厦门港务 000905.SZ	唐山港 601000.SH
日照港 600017.SH	连云港 601008.SH
锦州港 600190.SH	宁波港 601018.SH

表 3.3 K-means 分类第三类

泸州老窖 000568.SZ
古井贡酒 000596.SZ
五粮液 000858.SZ
顺鑫农业 000860.SZ
洋河股份 002304.SZ
伊力特 600197.SH
金种子酒 600199.SH
贵州茅台 600519.SH
老白干酒 600559.SH
沱牌舍得 600702.SH
山西汾酒 600809.SH

表 3. 4 K-means 分类第四类

中信银行 601998.SH	南京银行 601009.SH
中国银行 601988.SH	招商银行 600036.SH
建设银行 601939.SH	民生银行 600016.SH
光大银行 601818.SH	华夏银行 600015.SH
工商银行 601398.SH	浦发银行 600000.SH
交通银行 601328.SH	宁波银行 002142.SZ
农业银行 601288.SH	平安银行 000001.SZ
北京银行 601169.SH	

针对上述问题，我们又提出了一种方法——谱聚类。谱聚类相对而言适用范围更加广泛，并且能很好地处理 K-means 算法不能解决的问题。同时，谱聚类算法的稳定性也比较好。在定义不同股票之间的相关性关系时，谱聚类定义的关系矩阵更加严格和准确。

3.3.3 谱分析分类结果及分析

第一类：世纪星源，深振业，深物业沙河股份，中冠，深深房，中航地产，珠江控股；

第二类：泛海控股，上港集团，荣安地产，锦州港，深赤湾，重庆港九，盐田港，营口港；

珠海港，天津港，北部湾港，唐山港，厦门港务，连云港，日照港，宁波港，大连港；

第三类：中信银行，中国银行，建设银行，民生银行，工商银行，浦发银

行，光大银行华夏银行，交通银行，招商银行，宁波银行，农业银行，平安银行，北京银行，南京银行，；

第四类：皖江物流，洋河股份，中洲控股，伊力特，绿景控股，金种子酒，泸州老窖，贵州茅台，古井贡酒，老白干酒，五粮液，沱牌舍得，顺鑫农业，山西汾酒；

分类结果如表 3.5，表 3.6，表 3.7，表 3.8 所示。根据谱聚类的分类结果，如果以板块划分为标准，可以看到谱聚类相比于 K-means 聚类更好的对股票进行了划分。虽然结果并不完全准确，但已经达到了很好的标准。其中，泛海控股和荣安地产和港口板块的大部分股票归属于同一个类别；房地产板块中的中洲控股和绿景控股以及港口板块中的皖江物流归属于食品饮料板块，大部分股票的分类符合板块的行业划分标准。

表 3. 5 谱分析第一类

世纪星源 000005.SZ
深振业 A000006.SZ
深物业 A000011.SZ
沙河股份 000014.SZ
中冠 A000018.SZ
深深房 A000029.SZ
中航地产 000043.SZ
珠江控股 000505.SZ

表 3. 6 谱分析第二类

泛海控股 000046.SZ	上港集团 600018.SH
荣安地产 000517.SZ	锦州港 600190.SH
深赤湾 A000022.SZ	重庆港九 600279.SH
盐田港 000088.SZ	营口港 600317.SH
珠海港 000507.SZ	天津港 600717.SH
北部湾港 000582.SZ	唐山港 601000.SH
厦门港务 000905.SZ	连云港 601008.SH
日照港 600017.SH	宁波港 601018.SH
大连港 601880.SH	

表 3. 7 谱分析第三类

中信银行 601998.SH	南京银行 601009.SH
中国银行 601988.SH	招商银行 600036.SH
建设银行 601939.SH	民生银行 600016.SH
光大银行 601818.SH	华夏银行 600015.SH
工商银行 601398.SH	浦发银行 600000.SH
交通银行 601328.SH	宁波银行 002142.SZ
农业银行 601288.SH	平安银行 000001.SZ
北京银行 601169.SH	

表 3. 8 谱分析第四类

皖江物流 600575.SH	洋河股份 002304.SZ
中洲控股 000042.SZ	伊力特 600197.SH
绿景控股 000502.SZ	金种子酒 600199.SH
泸州老窖 000568.SZ	贵州茅台 600519.SH
古井贡酒 000596.SZ	老白干酒 600559.SH
五粮液 000858.SZ	沱牌舍得 600702.SH
顺鑫农业 000860.SZ	山西汾酒 600809.SH

4 分类内部绩效评估

4.1. 第三类股票绩效评估

股票的划分已经完成，但是每个股票板块的上市公司也是越来越多，比如上面已经划分好的银行板块，共选有 15 家。下面我们主要采用主成分分析方法，对银行板块上市公司的财务状况和经营状况进行分析，并对其进行排序，给出一个客观公正的评价[10]。

随着我国全球经济一体化和经济市场全球化的逐步推进，上市公司面临着变幻莫测的市场。本文以银行板块为例阐述绩效分析的重要性。商业银行是金融业最为核心的一部分，也借此机会得到迅猛的发展。它是我国经济中枢组成部分，经过多年的发展，已经朝着高效稳健的体系方向进行发展，这在我国的整个国民经济的发展中起到了十分重要的积极作用。而且，商业银行的业绩不仅关系到银行自己的发展问题，还会对整个金融体系产生深远的影响。一直以来，我国商业银行一直重视速度，却忽视了最为重要的一部分，那就是银行的发展质量。在我国的银行的发展中，存在着诸多问题，例如，银行管理水平不够，缺乏创新，不良贷款多，风险也很高等。这些问题都会给银行的长期发展带来一定的负面影响。因此，对银行经营绩效进行评价分析便显得尤为重要。中国的商业银行，要想在市场竞争中处于不败之地，就必须提高竞争力和效率，来应对激烈的行业竞争中带来的挑战。其次，评价体系完善，也有利于银行的自我评价，并根据自我评价对银行内部的运行机制进行调整和改革。只有提升了效率和盈利能力，银行股东和投资者的投资回报才能得到有效的风险控制，增加银行在社会中的透明度。

本文主要采用主成分分析方法，根据银行公布的财务指标，进行分析排序，研究在同行中的发展地位，从而，完善商业银行发展体系，就是要求我们未雨绸缪，当意识到与同行业的银行发展水平存在差距时，要有压迫感和紧迫感，更加明确努力方向，加快有效发展和内部控制。只有这样，才能应对日趋激烈的国际竞争环境。

与国外的研究相比，我国银行的绩效评估起步较晚。我们使用主成分分析，将定量与定性相结合，对同业银行进行分析评价。从盈利能力，安全能力，经营能力，流动能力，等各个方面对行业进行全面分析。总的来说，就是通过银

行公布的财务指标来构建评估模型，并对其绩效进行分析。本文分别对 4 个板块的上市公司的经营情况进行评价分析。采用的分析方法是主成分分析，本文将对主成分分析的基本思想，原理，计算步骤，及其优点进行全面的阐述。另外，秉着定量与定性相结合的重要原则，构建商业间绩效评估体系，得到该行业的综合得分和相应的排名，给银行一个客观的参考。

银行间的绩效评估是一个系统的工作，本文的研究主要集中在整体的经营分析并不具体研究的规模效率和技术效率等等。本文以银行板块的商业银行作为样本，对它们的经营情况进行研究，基本思路是：根据商业银行经营的实际情况，选取合适的指标，运用主成分分析的方法，对商业银行的经营效果进行客观公正的评估，从而给银行提醒，督促其改善自身的经营模式，提高其工作效率[24]。

4.1.1 财务指标选取原则

评估商业银行的经营效果需要一套较好的指标。这些指标要能全面正确的反映商业银行的整体运营情况，使得各个银行间可以根据这些指标来做出比较。商业银行是一个庞大而复杂的金融机构，影响其经营情况的因素有很多，需要从各个层面[25][26]，全方位的选取多个指标来建立一个有效的评估体系，通过这个指标体系，可以分析评估商业银行的经营效果，为了保证我们能够全方位地，合理地分析商业银行的经营情况，应该遵循如下的原则[11][12]：

（1）全面性。选取的财务指标覆盖面要广，可以真实的反应银行经营管理安全性，流动性，盈利性等方面。构建指标体系就是要对银行的各方面经营给出一个综合的评价，并根据评价结果来衡量银行整体的经营管理能力，所以建立全面的指标体系非常重要，它要能反映银行的安全能力，发展能力，流动能力和经营能力。

（2）重点性。虽然前面讲到，指标的选取要全面，但也不是越多越好。指标选取的个数要适中，太多显得冗杂，太少则很难全面反映银行的真实情况。这就要求选取指标时要遵循重点性的原则，在进行评估的时候，要抓住主要矛盾。将全面性和重点性结合起来，这样才能是选取的指标既不遗漏重要的信息，又主次分明，并且突出评估的重点。

（3）层次性。如果说选取的有些指标相关性太高，都反映了同一个事实，

那么我们就应该将这些指标化为一类。因此，选取的指标要进行分类，做到层次分明。

（4）可操作性。选取的指标数据都应该是非常容易就能获取到的。这样做的好处就是便于实证分析。否则，评估工作将无从开展。本文所采用的指标源于各银行所公布的财务报告里，数据源于国泰安。

4.1.2 财务指标的选取

对商业银行进行评估，本身就是一个相当复杂的问题。财务指标的选取是否科学，直接影响了评估结果的客观性，正确性和可参考性。一套有效的财务指标不仅评价银行近期发生的经济活动，更重要的一个方面是评价银行的未来发展能力。评估的结果好坏反映了银行是否实现了健康的可持续发展。因此，我们构建指标体系的总思路就是要强调全面，可持续性[31]。

依据上面的原则，我们选取的财务指标有六个，分别是总资产报酬率、销售利润率、应收账款周转率、股东权益比率、流动比率、资本积累率。

1) X1=总资产报酬率，是指企业的利润总额与总资产的比值，它表示企业的总资产的获利能力，是评价一个企业盈利能力的重要指标。

总资产报酬率=利润总额/总资产；

2) X2=销售利润率，是指企业主营业务利润与主营业务收入的比值。它同样也反映了企业的盈利能力。

销售利润率=主营业务利润/主营业务收入；

3) X3=流动比率，是指企业流动资产与流动负债的比值。它是用来衡量银行的债务偿还能力，体现了银行的资金流动能力。

流动比率=流动资产/流动负债；

4) X4=总资产增长率。总资产可以反映一个银行的综合实力，银行总资产规模的增长速度在一定程度上体现了银行未来发展的能力，因此，总资产增长率也是评价银行成长发展能力的一个重要指标。

总资产增长率=（本年总资产-上年总资产）/上年总资产；-

5) X5=营业收入增长率。企业营业收入的变动情况也是衡量一个企业成长发展状况的一个重要指标。

营业收入增长率=（本年营业收入-上年营业收入）/上年营业收入；

6) X6=净利润增长率。净利润实质利润总额除去所得税的剩余部分。它体现的是一个企业的经营最终结果，如果净利润越多，表示企业的经营效果越好；反之，则越差。因此，净利润增长率也是衡量一个企业未来发展能力的重要指标。该指标越大，表明该企业正处于在同行中的竞争力强。

净利润增长率=（本年净利润-上年净利润）/上年净利润；

7) X7=资本积累率；是指本年股东收益与上年股东收益的差额与上年股东收益的比值。它反映了银行资本本积累能力，是评估企业未来发展的一个重要指标。积累的资本越多，其未来的发展能力也就越强。

资产积累率=（本年股东权益-上年股东权益）/上年股东权益；

4.1.3 商业银行评估的实证分析

根据上面对主成分分析的相关知识，结合选取的财务指标评价体系，本章就对上面选出的银行板块的 15 家银行进行实证分析。这 15 家银行分别是中信银行，中国银行，建设银行，光大银行，工商银行，交通银行，农业银行，北京银行，南京银行，招商银行，民生银行，华夏银行，浦发银行，宁波银行，平安银行。在对这 15 家银行进行评估的时候，采用主成分分析的方法，通过财务指标评价体系，对这 15 家银行进行实证分析，根据预测值进行排序，以便达到利用指标对银行的经营效果进行客观公正评估的目的[21]。

本文以中信银行，中国银行，建设银行，光大银行，工商银行，交通银行，农业银行，北京银行，南京银行，招商银行，民生银行，华夏银行，浦发银行，宁波银行，平安银行为样本，进行经营评估，通过分析得出的结论对我国银行的发展现状，做出客观的评价。

它们对应的代码如表所示：

表 4.1 证券名称及代码

000001.SZ	平安银行	601288.SH	农业银行
002142.SZ	宁波银行	601328.SH	交通银行
600000.SH	浦发银行	601398.SH	工商银行
600015.SH	华夏银行	601818.SH	光大银行
600016.SH	民生银行	601939.SH	建设银行
600036.SH	招商银行	601988.SH	中国银行
601009.SH	南京银行	601998.SH	中信银行
601169.SH	北京银行		

实证分析侧重银行的安全性，流动性，盈利性和发展性进行分析，指标共有六个，X1=总资产报酬率，X2=销售利润率，X3=流动比率，X4=总资产增长率；X5=营业收入增长率；X6=净利润增长率；X7=资本积累率；X8=股东收益权益。

本文采用这八个财务指标，采用主成分分析的方法，来对各个公司的业务能力进行综合分析。这八个指标就是主成分分析中的原因子，本文使用主成分分析的方法可以降维，用更少的，并且互不相关的指标重新对这 15 个公司进行评价排序，这就是本文的主要思想。本文对样本选取的 15 只股票进行评价，给投资者提供客观的参考依据。

首先，根据公司公布的财务报表提取我们想要的财务数据，并对每组的数据进行初步的处理，即标准化。因为各个指标之间的单位有可能是不一样的，若用原始的数据直接进行计算，得出的结果是不能够得到合理的解释的，尤其是当它们的数量级相差很大时，更是没有办法直接运算，否则，单位比较大的量会对结果起决定性作用，而单位较小的量可能根本无法发挥作用。为了解决这个问题，我们需要对数据进行初步的处理，将均值化为 0，方差化为 1。这样，标准化之后的矩阵才具有可比性。如表 4.1 所示：

表 4.1　证券财务数据（标准化）

代码	000001.SZ	002142.SZ	600000.SH	600015.SH	600016.SH
名称	平安银行	宁波银行	浦发银行	华夏银行	民生银行
总资产报酬率	-1.490	-0.648	-0.546	-0.931	0.916
销售利润率	-2.224	0.330	-0.464	-0.529	-0.190
流动比率	2.755	-1.016	-0.696	-0.547	1.553
总资产增长率	0.606	1.118	0.285	-0.658	0.618
营业收入增长率	1.405	0.535	0.229	-0.783	0.092
净利润增长率	1.120	1.200	-0.209	0.021	-0.346
资本积累率	0.114	0.719	0.739	-0.244	0.219
代码	600036.SH	601009.SH	601169.SH	601288.SH	601328.SH
名称	招商银行	南京银行	北京银行	农业银行	交通银行
总资产报酬率	0.967	-1.077	0.490	0.350	-0.617
销售利润率	-0.885	-0.761	1.894	0.202	0.383
流动比率	-0.055	-0.629	0.515	-0.359	-0.646
总资产增长率	-1.068	2.832	-0.428	-0.529	-0.098
营业收入增长率	0.544	2.451	0.094	-1.158	-0.818
净利润增长率	0.234	2.450	0.924	-0.912	-0.747
资本积累率	-0.798	2.865	-0.613	-0.009	-1.348
代码	601398.SH	601818.SH	601939.SH	601988.SH	601998.SH
名称	工商银行	光大银行	建设银行	中国银行	中信银行
总资产报酬率	1.355	-0.221	1.980	0.271	-0.799
销售利润率	1.106	0.003	1.063	0.716	-0.642
流动比率	-0.347	0.257	-0.460	-0.673	0.349
总资产增长率	-0.394	-0.194	-0.325	-0.904	-0.860
营业收入增长率	-0.695	0.378	-0.705	-1.227	-0.342
净利润增长率	-0.880	-0.631	-0.842	-0.743	-0.640
资本积累率	-0.537	0.264	-0.995	0.237	-0.612

根据前面选取的因子矩阵，我们可以得出各因子之间的相关性。相关系数如:4.2 所示：

表 4.2　相关系数矩阵

	X1	X2	X3	X4	X5	X6	X7
X1	1.000	0.622	-0.135	-0.408	-0.443	-0.493	-0.439
X2	0.622	1.000	-0.467	-0.280	-0.539	-0.338	-0.322
X3	-0.135	-0.467	1.000	0.036	0.323	0.143	-0.094
X4	-0.408	-0.280	0.036	1.000	0.781	0.753	0.824
X5	-0.443	-0.539	0.323	0.781	1.000	0.864	0.690
X6	-0.493	-0.338	0.143	0.753	0.864	1.000	0.671
X7	-0.439	-0.322	-0.094	0.824	0.690	0.671	1.000

在表中，我们可以看出各因子之间具有相关性。其中，总资产增长率与资本积累率的相关性最高，达到 82.4%，其次，总资产增长率与营业收入增长率的相关性也很高，达到 78%。另外，也有一些指标之间具有负相关的关系。从表中可以看出，总资产回报率除了与销售利润率之间具有负相关外，与其他都具有正相关的关系。

特征值、贡献率和累计贡献率

根据计算，特征值，贡献率和累计贡献率的结果如表 4.3 所示。

表 4.3 特征值、贡献率和累计贡献率

主成分	特征值	贡献率	累计贡献率
1	3.936	0.562	0.562
2	1.407	0.201	0.763
3	0.801	0.114	0.878
4	0.371	0.053	0.931
5	0.272	0.039	0.969
6	0.146	0.021	0.990
7	0.068	0.010	1.000

从表中可以看出，前三个主成分的累计贡献率可以达到 87.8%，超过 85%，已经可以很好的表达原因子所要传达的信息，很好的解释这 15 家公司的经营效果及未来发展能力。其中，第一主成分的贡献率达到 56.2%，第二主成分的贡献率是 20.1%，第三主成分的贡献率为 11.4%。由此可以得出前三个主成分的累计贡献率达到 87.8%，降低了指标个数，从而减少了计算量。

因子载荷矩阵。通过因子载荷矩阵可以分析各个主成分的经济含义。给新生成的主成分赋予新的含义是主成分分析的一个重要步骤。因子载荷矩阵如表

4.5 所示：

表 4.4　主成分因子载荷矩阵

主成分分析	第一主成分	第二主成分	第三主成分	第四主成分	第五主成分	第六主成分	第七主成分
总资产报酬率	-0.339	-0.206	0.689	-0.421	-0.376	-0.015	-0.220
销售利润率	-0.315	-0.524	0.312	0.462	0.456	-0.129	0.301
流动比率	0.131	0.680	0.523	0.032	0.467	-0.163	-0.047
总资产增长率	0.432	-0.295	0.164	-0.174	0.361	0.704	-0.209
营业收入增长率	0.465	0.014	0.294	0.053	-0.381	0.078	0.737
净利润增长率	0.442	-0.146	0.190	0.575	-0.298	-0.237	-0.522
资本积累率	0.413	-0.335	-0.067	-0.495	0.262	-0.632	0.011

主成分是原因子之间的线性组合，根据因子载荷矩阵，我们可以得到各个主成分的函数形式，以及预测函数的表达形式。

$$Z_1 = -0.339X_1 - 0.315X_2 + 0.131X_3 + 0.432X_4 + 0.465X_5 + 0.442X_6 + 0.413X_7$$

$$Z_2 = -0.206X_1 - 0.524X_2 + 0.68X_3 - 0.295X_4 + 0.014X_5 - 0.146X_6 - 0.335X_7$$

$$Z_3 = 0.689X_1 + 0.312X_2 + 0.523X_3 + 0.164X_4 + 0.294X_5 + 0.190X_6 - 0.067X_7$$

其中，预测函数：

$$Z = 3.936Z_1 + 1.407Z_2 + 0.801Z_3$$

通过预测函数，我们可以给出这 15 家银行的预测值，并根据预测值进行排名，排名结果就是我们对这 15 家银行的绩效评估。在这里，我们给出这 15 家银行的一个综合排名。如表 4.6 所示：

表 4.5　综合评价结果

排名	代码	名称	第一主成分	第二主成分	第三主成分	预测值
1	601009.SH	南京银行	5.151	-1.927	0.149	2.526
2	000001.SZ	平安银行	3.024	2.984	0.435	2.350
3	002142.SZ	宁波银行	1.542	-1.469	-0.354	0.531
4	600016.SH	民生银行	0.200	0.763	1.432	0.429
5	600000.SH	浦发银行	0.682	-0.416	-0.861	0.202
6	601818.SH	光大银行	0.030	0.284	-0.075	0.065
7	600036.SH	招商银行	-0.490	0.782	0.444	-0.067
8	601998.SH	中信银行	-0.547	1.286	-0.891	-0.151
9	600015.SH	华夏银行	-0.329	0.359	-1.411	-0.274
10	601169.SH	北京银行	-0.682	-0.545	1.372	-0.336
11	601328.SH	交通银行	-1.306	0.066	-0.951	-0.830
12	601288.SH	农业银行	-1.403	-0.146	-0.483	-0.874
13	601988.SH	中国银行	-1.597	-0.609	-0.607	-1.090
14	601398.SH	工商银行	-1.957	-0.680	0.698	-1.157
15	601939.SH	建设银行	-2.318	-0.734	1.102	-1.324

总体来看，在银行中，南京银行排名为第一位。由于本文选取的都是财务指标，因此，在财务方面，做的比较好的银行有南京银行、平安银行、宁波银行以及民生银行。

4.2 第一类股票主成分分析

第一类样本股票有世纪星源，深振业，深物业沙河股份，中冠，深深房，

中航地产，珠江控股；其中房地产类股票居多。

现对第一类股票运用主成分分析，进行绩效评估。

表 4.6　证券名称及代码

证券代码	证券简称
000005.SZ	世纪星源
000006.SZ	深振业 A
000011.SZ	深物业 A
000014.SZ	沙河股份
000018.SZ	中冠 A
000029.SZ	深深房 A
000043.SZ	中航地产
000505.SZ	珠江控股

财务指标仍然选择上述的七个，X1=总资产报酬率，X2=销售利润率，X3=流动比率，X4=总资产增长率；X5=营业收入增长率；X6=净利润增长率；X7=资本积累率。

首先，根据公司公布的财务报表提取我们想要的财务数据，并对每组的数据进行初步的处理，即标准化。计算结果如表 4.8 所示。

表 4.7　证券财务数据（标准化）

证券代码	000005.SZ	000006.SZ	000011.SZ	000014.SZ
证券简称	世纪星源	深振业 A	深物业 A	沙河股份
总资产报酬率	-1.07	0.28	-0.10	-0.50
销售利润率	-2.22	0.77	0.52	-0.11
流动比率	-1.60	0.97	0.27	1.38
总资产增长率	-0.38	-0.35	-0.36	-0.36
营业收入增长率	-0.36	-0.34	-0.36	-0.36
净利润增长率	-0.36	-0.35	-0.37	-0.37
资本积累率	0.10	0.10	0.08	0.10
证券代码	000018.SZ	000029.SZ	000043.SZ	000505.SZ
证券简称	中冠 A	深深房 A	中航地产	珠江控股
总资产报酬率	1.65	1.27	-0.69	-0.84
销售利润率	0.56	0.87	-0.06	-0.35
流动比率	-0.47	0.49	0.01	-1.04
总资产增长率	2.47	-0.35	-0.35	-0.32
营业收入增长率	2.47	-0.35	-0.36	-0.36
净利润增长率	2.47	-0.31	-0.34	-0.37
资本积累率	1.53	0.13	0.11	-2.14

根据前面选取的因子矩阵，我们可以得出各因子之间的相关性。相关系数如表 4.9 所示：

表 4.8　相关系数矩阵

相关系数	X1	X2	X3	X4	X5	X6	X7
X1	1.00	0.71	0.28	0.67	0.67	0.68	0.61
X2	0.71	1.00	0.69	0.24	0.23	0.23	0.23
X3	0.28	0.69	1.00	-0.19	-0.19	-0.19	0.24
X4	0.67	0.24	-0.19	1.00	1.00	1.00	0.61
X5	0.67	0.23	-0.19	1.00	1.00	1.00	0.62
X6	0.68	0.23	-0.19	1.00	1.00	1.00	0.62
X7	0.61	0.23	0.24	0.61	0.62	0.62	1.00

在表中，我们可以看出各因子之间具有相关性。其中，总资产增长率与销售利润率的相关性最高，达到 71%，其次，总资产增长率与营业收入增长率的相关性也很高，达到 68%。另外，也有一些指标之间具有负相关的关系。从表中可以看出，流动比率与总资产增长率、净利润增长率、营业收入增长率三个

指标的相关性都是负的。

根据进一步计算，特征值，贡献率和累计贡献率的结果如表 4.10 所示。

表 4.9　特征值、贡献率和累计贡献率

主成分	特征值	贡献率	累计贡献率
1	4.19	0.60	0.60
2	1.90	0.27	0.87
3	0.60	0.09	0.96
4	0.24	0.03	0.99
5	0.06	0.01	1.00
6	0.00	0.00	1.00
7	0.00	0.00	1.00

从表中可以看出，前两个主成分的累计贡献率可以达到 87%，超过 85%，已经可以很好的表达原因子所要传达的信息，很好的解释这类公司的经营效果及未来发展能力。其中，第一主成分的贡献率达到 60%，第二主成分的贡献率达到 27%，此时，前两个主成分的累计贡献率达到 87%，因此，我们选择前两个主成分，这样做有效降低了指标个数，从而减少了计算量。

因子载荷矩阵。通过因子载荷矩阵可以分析各个主成分的经济含义。给新的生成的主成分赋予新的含义是主成分分析的一个重要步骤。因子载荷矩阵如表 4.11 所示：

表 4.10　主成分因子载荷矩阵

	第一主成分	第二主成分	第三主成分	第四主成分	第五主成分	第六主成分	第七主成分
总资产报酬率	0.42	0.26	-0.18	0.68	-0.51	-0.01	0.02
销售利润率	0.23	0.58	-0.44	0.01	0.64	0.00	-0.01
流动比率	0.03	0.68	0.27	-0.53	-0.43	0.01	0.00
总资产增长率	0.46	-0.21	-0.11	-0.27	-0.03	-0.11	0.80
营业收入增长率	0.46	-0.21	-0.09	-0.26	-0.04	-0.64	-0.51
净利润增长率	0.46	-0.21	-0.09	-0.24	-0.04	0.76	-0.31
资本积累率	0.37	0.08	0.82	0.24	0.36	-0.01	0.01

主成分是原因子之间的线性组合，根据因子载荷矩阵，我们可以得到各个主成分的函数形式，以及预测函数的表达形式。

$$Z_1 = 0.42X_1 + 0.23X_2 + 0.03X_3 + 0.46X_4 + 0.46X_5 + 0.46X_6 + 0.37X_7$$

$$Z_2 = 0.26X_1 + 0.58X_2 + 0.68X_3 - 0.21X_4 - 0.21X_5 - 0.21X_6 + 0.08X_7$$

其中，预测函数：

$$Z = 0.6Z_1 + 0.27Z_2$$

通过预测函数，我们可以给出这类样本企业的预测值，并根据预测值进行排名，排名结果就是我们对这类样本的绩效评估。

根据预测函数给出第一类样本股票的排名，结果如表 4.12 所示。

表 4.11　证券综合评估

排名	代码	名称	第一主成分	第二主成分	预测值
1	000018.SZ	中冠 A	4.79	-0.99	2.601172
2	000029.SZ	深深房 A	0.31	1.39	0.563866
3	000006.SZ	深振业 A	-0.12	1.40	0.308002
4	000011.SZ	深物业 A	-0.39	0.69	-0.04464
5	000014.SZ	沙河股份	-0.66	0.98	-0.12951
6	000043.SZ	中航地产	-0.74	0.02	-0.43984
7	000505.SZ	珠江控股	-1.73	-1.08	-1.32789
8	000005.SZ	世纪星源	-1.47	-2.41	-1.53115

总体来看，在这类股票中，中冠 A 排名为第一位。由于本文选取的都是财务指标，因此，在财务方面，做的比较好的公司有中冠 A 、深深房 A、深振业 A 以及深物业 A。

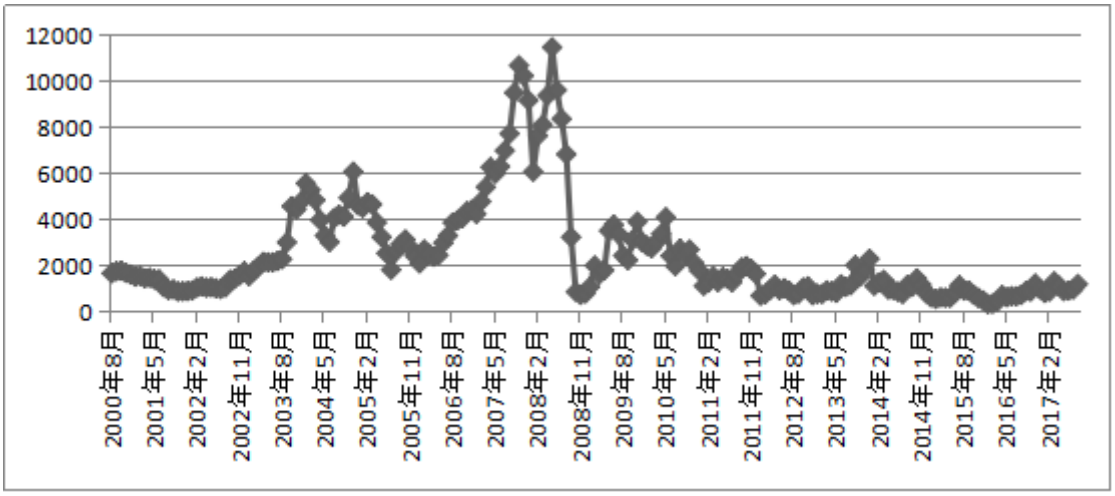
4.3 第二类股票主成分分析

第二类样本股票有泛海控股，上港集团，荣安地产，锦州港，深赤湾，重庆港九，盐田港，营口港；其中港口类股票居多。

港口业属于大型基础设施行业，现我国的港口布局已经逐步稳定，综合性大型枢纽港逐步完善。行业集中度较高，具有规模经济效益等特点。港口行业壁垒较高，不仅需要良好的地理位置、充裕雄厚的资金支持，更要符合国家的港口规划建设政策。纵观我国，港口布局基本完成，建设体系以主枢纽为骨干、区域性中型港口为辅助，小型港口作为补充。《全国沿海港口布局规划》将我国的沿海港口大致划分为五个港口群体，主要根据各个区域的经济发展特点、区域内港口的发展状况、各港口的运输关系、以及货物运输关系的合理性，分别为环渤海、东南沿海、珠江三角洲、长江三角洲、以及西南沿海。

2017 年上半年港口版块突然爆发，分析主要原因有两大刺激因素。

首先是波罗的海干散货运价指数（BDI）创新高。2017 年 3 月，BDI 指数达到 1333 点，创下 2014 年底以来的新高，相较于年内最低点 685，上涨 94.6%。其中，BDI 指数是根据几条主要航线的即期运费，利用加权平均的方法计算而成，反应的是即期市场的行情，故运费价格的波动是指数波动的主要影响因素。散装货物主要有钢材、农作物、煤、铁矿石等农业物资及工业原料等。BDI 指数之所以被称为国际贸易的晴雨表指数，是因为运费价格的走势与全球经济的景气荣枯、原材料行情高度相关。下图给出的是 2010 年至 2017 年 BDI 指数的走势图。



第二大刺激因素是“一带一路”概念的持续走高。一带一路战略是国家级顶层重点战略计划，对沿海国家的共同发展带来了新的机遇，打造新亚欧大陆桥、中蒙俄、中国-中亚-西亚、中国-中南的国际合作的新平台。港口作为水陆交通和物流的重要枢纽，直接受益“一带一路”的合作成果，充当重要的节点。2017 年 5 月 14 日至 15 日，由我国主办的一带一路国际合作高峰论坛，在北京隆重举行，29 位外国首脑、政府首脑等出席会议，对推动国际和地区合作具有重要的意义。

现对第二类股票运用主成分分析，进行绩效评估。

表 4.12　证券名称及代码

证券代码	证券简称
000046.SZ	泛海控股
000517.SZ	荣安地产
000022.SZ	深赤湾 A
000088.SZ	盐田港
000507.SZ	珠海港
000582.SZ	北部湾港
000905.SZ	厦门港务
600017.SH	日照港
601880.SH	大连港
600018.SH	上港集团
600190.SH	锦州港
600279.SH	重庆港九
600317.SH	营口港
600717.SH	天津港
601000.SH	唐山港
601008.SH	连云港
601018.SH	宁波港

财务指标仍然选择上述的七个，X1=总资产报酬率，X2=销售利润率，X3=流动比率，X4=总资产增长率；X5=营业收入增长率；X6=净利润增长率；X7=资本积累率。

首先，根据公司公布的财务报表提取我们想要的财务数据，并对每组的数据进行初步的处理，即标准化。计算结果如表所示。

表 4.13　证券财务数据（标准化）

证券代码	000046.SZ	000517.SZ	000022.SZ	000088.SZ	000507.SZ	000582.SZ
证券简称	泛海控股	荣安地产	深赤湾 A	盐田港	珠海港	北部湾港
总资产报酬率	-0.69	-1.30	1.81	1.26	-0.90	-0.13
销售利润率	0.10	-0.42	0.42	3.71	-0.54	-0.18
流动比率	0.72	1.35	-0.69	3.18	-0.34	-0.53
总资产增长率	2.44	1.01	-1.14	0.44	-0.56	1.28
营业收入增长率	2.88	-2.22	0.14	-0.39	0.18	-0.66
净利润增长率	2.50	-0.69	0.30	0.36	-2.94	-0.10
资本积累率	1.12	-0.69	-0.53	-0.28	-0.45	2.49
证券代码	000905.SZ	600017.SH	601880.SH	600018.SH	600190.SH	600279.SH
证券简称	厦门港务	日照港	大连港	上港集团	锦州港	重庆港九
总资产报酬率	0.26	-0.41	-0.73	1.29	-0.93	-0.87
销售利润率	-0.58	-0.26	-0.51	0.08	-0.46	-0.49
流动比率	0.10	-0.39	-0.09	-0.48	-0.83	-0.28
总资产增长率	-0.69	-0.49	-0.34	-0.41	-0.76	1.64
营业收入增长率	-0.06	-0.33	0.49	0.15	0.14	0.12
净利润增长率	0.22	-0.23	0.06	0.17	-0.08	-0.04
资本积累率	-0.22	-0.64	-0.64	-0.22	-0.59	1.79
证券代码	600317.SH	600717.SH	601000.SH	601008.SH	601018.SH	
证券简称	营口港	天津港	唐山港	连云港	宁波港	
总资产报酬率	-0.06	0.53	1.22	-1.23	0.89	
销售利润率	-0.09	-0.29	0.15	-0.58	-0.07	
流动比率	0.32	-0.18	-0.35	-0.97	-0.54	
总资产增长率	-0.52	-0.80	-0.26	-0.56	-0.27	
营业收入增长率	-0.11	-0.77	-0.14	-0.40	0.98	
净利润增长率	0.13	0.26	0.24	-0.28	0.12	
资本积累率	-0.55	-0.55	1.25	-0.76	-0.51	

根据前面选取的因子矩阵，我们可以得出各因子之间的相关性。相关系数如表所示：

表 4.14 相关系数矩阵

相关系数	X1	X2	X3	X4	X5	X6	X7
X1	1.00	0.52	0.11	-0.34	0.07	0.26	0.00
X2	0.52	1.00	0.78	0.12	0.00	0.23	-0.01
X3	0.11	0.78	1.00	0.38	-0.13	0.19	-0.06
X4	-0.34	0.12	0.38	1.00	0.27	0.40	0.69
X5	0.07	0.00	-0.13	0.27	1.00	0.52	0.20
X6	0.26	0.23	0.19	0.40	0.52	1.00	0.27
X7	0.00	-0.01	-0.06	0.69	0.20	0.27	1.00

在表中，我们可以看出各因子之间具有相关性。其中，销售利润率与流动比率的相关性最高，达到 78%，其次，总资产增长率与资本积累率的相关性也很高，达到 69%。另外，也有一些指标之间具有负相关的关系。从表中可以看出，总资产报酬率与资本积累率的相关性是负的，为-34%。

根据进一步计算，特征值，贡献率和累计贡献率的结果如表中所示。

表 4.15 特征值、贡献率和累计贡献率

主成分	特征值	贡献率	累计贡献率
1	2.38	0.34	0.34
2	1.90	0.27	0.61
3	1.33	0.19	0.80
4	0.80	0.11	0.92
5	0.43	0.06	0.98
6	0.11	0.02	0.99
7	0.05	0.01	1.00

从表中可以看出，前四个主成分的累计贡献率可以达到 92%，超过 85%，已经可以很好的表达原因子所要传达的信息，很好的解释这类公司的经营效果及未来发展能力。其中，第一主成分的贡献率达到 34%，第二主成分的贡献率达到 27%，第三主成分的贡献率达到 19%，第四主成分的贡献率达到 11%，此时，前两个主成分的累计贡献率达到 92%，因此，我们选择前四个主成分，这样做有效降低了指标个数，从而减少了计算量。

因子载荷矩阵。通过因子载荷矩阵可以分析各个主成分的经济含义。给新的生成的主成分赋予新的含义是主成分分析的一个重要步骤。因子载荷矩阵如表中所示：

表 4.16　主成分因子载荷矩阵

	第一主成分	第二主成分	第三主成分	第四主成分	第五主成分	第六主成分	第七主成分
总资产报酬率	0.17	0.44	0.51	0.49	0.00	-0.51	0.13
销售利润率	0.40	0.52	-0.08	0.04	-0.28	0.61	0.31
流动比率	0.40	0.39	-0.44	-0.26	-0.01	-0.36	-0.55
总资产增长率	0.47	-0.37	-0.37	0.00	-0.01	-0.37	0.61
营业收入增长率	0.29	-0.28	0.51	-0.43	-0.61	-0.08	-0.13
净利润增长率	0.47	-0.10	0.37	-0.23	0.73	0.21	-0.06
资本积累率	0.35	-0.41	-0.08	0.67	-0.13	0.22	-0.43

主成分是原因子之间的线性组合，根据因子载荷矩阵，我们可以得到各个主成分的函数形式，以及预测函数的表达形式。

$$Z_1 = 0.17X_1 + 0.4X_2 + 0.4X_3 + 0.47X_4 + 0.29X_5 + 0.47X_6 + 0.35X_7$$

$$Z_2 = 0.44X_1 + 0.52X_2 + 0.39X_3 - 0.37X_4 - 0.28X_5 - 0.1X_6 + 0.41X_7$$

$$Z_3 = 0.51X_1 - 0.08X_2 - 0.44X_3 - 0.37X_4 + 0.51X_5 + 0.37X_6 - 0.08X_7$$

$$Z_4 = 0.49X_1 + 0.04X_2 - 0.26X_3 - 0.0X_4 - 0.43X_5 - 0.23X_6 + 0.67X_7$$

其中，预测函数：$Z = 0.34Z_1 + 0.27Z_2 + 0.19Z_3 + 0.11Z_4$

通过预测函数，我们可以给出这类样本企业的预测值，并根据预测值进行排名，排名结果就是我们对这类样本的绩效评估，结果如表所示。

表 4.17　证券综合评估

排名	代码	名称	第一主成分	第二主成分	第三主成分	第四主成分	预测值
1	000088.SZ	盐田港	3.14	3.75	-1.26	-0.19	1.82
2	000022.SZ	深赤湾 A	-0.33	1.31	1.84	0.60	0.66
3	000046.SZ	泛海控股	3.76	-2.36	0.72	-1.58	0.59
4	601000.SH	唐山港	0.52	0.08	0.77	1.54	0.52
5	600018.SH	上港集团	-0.08	0.60	1.17	0.51	0.42
6	601018.SH	宁波港	-0.05	0.16	1.38	-0.22	0.26
7	600717.SH	天津港	-0.76	0.71	0.42	0.20	0.04
8	600317.SH	营口港	-0.33	0.48	0.07	-0.47	-0.02
9	000905.SZ	厦门港务	-0.46	0.19	0.46	-0.10	-0.03
10	000582.SZ	北部湾港	0.93	-1.64	-0.88	2.05	-0.06
11	600279.SH	重庆港九	0.95	-2.10	-0.99	0.79	-0.34
12	600017.SH	日照港	-0.99	0.09	-0.03	-0.35	-0.36
13	601880.SH	大连港	-0.59	-0.38	0.16	-1.01	-0.39
14	600190.SH	锦州港	-1.24	-0.48	0.30	-0.70	-0.57
15	000517.SZ	荣安地产	-0.60	0.34	-2.92	-0.36	-0.71
16	601008.SH	连云港	-1.61	-0.56	-0.18	-0.64	-0.81
17	000507.SZ	珠海港	-2.26	-0.18	-1.02	-0.08	-1.02

总体来看，在这类股票中，盐田港排名为第一位。由于本文选取的都是财务指标，因此，在财务方面，做的比较好的公司有盐田港、深赤湾 A、泛海控股以及唐山港。

4.4　第四类股票主成分分析

第四类样本股票有皖江物流，洋河股份，中洲控股，伊力特，绿景控股，金种子酒，泸州老窖，贵州茅台，古井贡酒，老白干酒，五粮液，沱牌舍得，顺鑫农业，山西汾酒；其中，酒类公司居多。

表 4.18　证券名称及代码

证券代码	证券简称
600575.SH	皖江物流
000042.SZ	中洲控股
000502.SZ	绿景控股
000568.SZ	泸州老窖
000596.SZ	古井贡酒

000858.SZ	五粮液
000860.SZ	顺鑫农业
002304.SZ	洋河股份
600197.SH	伊力特
600199.SH	金种子酒
600519.SH	贵州茅台
600559.SH	老白干酒
600702.SH	沱牌舍得
600809.SH	山西汾酒

财务指标仍然选择上述的七个，X1=总资产报酬率，X2=销售利润率，X3=流动比率，X4=总资产增长率；X5=营业收入增长率；X6=净利润增长率；X7=资本积累率。

首先，根据公司公布的财务报表提取我们想要的财务数据，并对每组的数据进行初步的处理，即标准化。计算结果如表所示。

表 4.19 证券财务数据（标准化）

证券代码	600575.SH	000042.SZ	000502.SZ	000568.SZ	000596.SZ
证券简称	皖江物流	中洲控股	绿景控股	泸州老窖	古井贡酒
总资产报酬率	-0.58	-0.78	-1.49	0.85	0.22
销售利润率	-0.40	0.04	-2.55	0.64	-0.02
流动比率	-0.46	-0.65	0.67	0.71	-0.52
总资产增长率	-0.89	3.14	-0.74	-0.56	0.16
营业收入增长率	-2.48	2.33	-0.29	0.11	0.36
净利润增长率	-1.71	0.42	2.97	-0.03	0.02
资本积累率	-0.23	0.75	-2.19	-0.91	0.88
证券代码	000858.SZ	000860.SZ	002304.SZ	600197.SH	600199.SH
证券简称	五粮液	顺鑫农业	洋河股份	伊力特	金种子酒
总资产报酬率	0.69	-0.72	1.49	0.85	-0.82
销售利润率	0.86	-0.39	0.95	0.37	-0.43
流动比率	2.83	-0.83	-0.47	0.49	-0.27
总资产增长率	-0.13	-0.42	-0.09	-0.20	-0.43
营业收入增长率	-0.02	0.12	0.28	-0.02	-0.57
净利润增长率	-0.14	-0.08	0.07	-0.34	-0.59
资本积累率	0.26	-0.14	0.92	0.01	-0.66

证券代码	600519.SH	600559.SH	600702.SH	600809.SH	
证券简称	贵州茅台	老白干酒	沱牌舍得	山西汾酒	
总资产报酬率	1.78	-0.56	-1.00	0.07	
销售利润率	1.83	-0.40	-0.51	0.01	
流动比率	0.53	-1.05	-0.65	-0.33	
总资产增长率	0.89	-0.48	-0.18	-0.06	
营业收入增长率	0.17	0.62	-0.63	0.03	
净利润增长率	-0.01	0.28	-0.64	-0.23	
资本积累率	1.94	0.11	-0.88	0.15	

根据前面选取的因子矩阵，我们可以得出各因子之间的相关性。相关系数如表所示：

表 4.20 相关系数矩阵

相关系数	X1	X2	X3	X4	X5	X6	X7
X1	1.00	0.88	0.38	0.08	0.12	-0.24	0.67
X2	0.88	1.00	0.25	0.31	0.21	-0.51	0.81
X3	0.38	0.25	1.00	-0.11	-0.08	0.17	-0.07
X4	0.08	0.31	-0.11	1.00	0.76	0.08	0.53
X5	0.12	0.21	-0.08	0.76	1.00	0.41	0.36
X6	-0.24	-0.51	0.17	0.08	0.41	1.00	-0.37
X7	0.67	0.81	-0.07	0.53	0.36	-0.37	1.00

在表中，我们可以看出各因子之间具有相关性。其中，总资产增长率与销售利润率的相关性最高，达到 88%，其次，总资产增长率与净利润增长率的相关性也很高，达到 76%。另外，也有一些指标之间具有负相关的关系。从表中可以看出，流动比率与总资产增长率的相关性都是负的。

根据进一步计算，特征值，贡献率和累计贡献率的结果如表中所示。

表 4.21 特征值、贡献率和累计贡献率

主成分	特征值	贡献率	累计贡献率
1	3.06	0.44	0.44
2	1.88	0.27	0.71
3	1.29	0.18	0.89
4	0.45	0.06	0.96
5	0.19	0.03	0.98
6	0.10	0.01	1.00
7	0.02	0.00	1.00

从表中可以看出，前两个主成分的累计贡献率可以达到 87%，超过 85%，已经可以很好的表达原因子所要传达的信息，很好的解释这类公司的经营效果及未来发展能力。其中，第一主成分的贡献率达到 60%，第二主成分的贡献率达到 27%，此时，前两个主成分的累计贡献率达到 87%，因此，我们选择前两个主成分，这样做有效降低了指标个数，从而减少了计算量。

因子载荷矩阵。通过因子载荷矩阵可以分析各个主成分的经济含义。给新的生成的主成分赋予新的含义是主成分分析的一个重要步骤。因子载荷矩阵如表中所示：

表 4.22　主成分因子载荷矩阵

	第一主成分	第二主成分	第三主成分	第四主成分	第五主成分	第六主成分	第七主成分
总资产报酬率	0.47	-0.24	0.28	-0.44	-0.08	-0.44	-0.50
销售利润率	0.54	-0.19	0.04	0.00	-0.27	-0.11	0.76
流动比率	0.09	-0.19	0.77	0.53	0.07	0.25	-0.09
总资产增长率	0.31	0.53	-0.12	0.52	0.25	-0.52	-0.07
营业收入增长率	0.24	0.62	0.10	-0.09	-0.60	0.38	-0.17
净利润增长率	-0.21	0.46	0.51	-0.47	0.35	-0.13	0.34
资本积累率	0.52	0.04	-0.19	-0.16	0.61	0.54	-0.06

主成分是原因子之间的线性组合，根据因子载荷矩阵，我们可以得到各个主成分的函数形式，以及预测函数的表达形式。

$$Z_1 = 0.47X_1 + 0.54X_2 + 0.09X_3 + 0.31X_4 + 0.24X_5 - 0.21X_6 + 0.52X_7$$
$$Z_2 = -0.24X_1 - 0.19X_2 - 0.19X_3 + 0.53X_4 + 0.62X_5 + 0.46X_6 + 0.04X_7$$
$$Z_3 = 0.28X_1 + 0.04X_2 + 0.77X_3 - 0.12X_4 + 0.10X_5 + 0.51X_6 - 0.19X_7$$

其中，预测函数：

$$Z = 0.44Z_1 + 0.27Z_2 + 0.18Z_3$$

通过预测函数，我们可以给出这类样本企业的预测值，并根据预测值进行排名，排名结果就是我们对这类样本的绩效评估。

根据预测函数给出第四类样本股票的排名，结果如表所示。

表 4.23　证券综合评估

排名	代码	名称	第一主成分	第二主成分	第三主成分	预测值
1	000042.SZ	中洲控股	1.44	3.62	-0.80	1.45

2	600519.SH	贵州茅台	3.22	-0.24	0.51	1.44
3	000858.SZ	五粮液	1.17	-1.01	2.30	0.66
4	002304.SZ	洋河股份	1.68	-0.26	-0.01	0.66
5	000596.SZ	古井贡酒	0.63	0.40	-0.48	0.30
6	600197.SH	伊力特	0.66	-0.64	0.48	0.20
7	000568.SZ	泸州老窖	0.19	-0.74	1.05	0.08
8	600809.SH	山西汾酒	0.13	-0.07	-0.37	-0.03
9	600559.SH	老白干酒	-0.58	0.67	-0.74	-0.21
10	000860.SZ	顺鑫农业	-0.78	0.21	-0.81	-0.43
11	600199.SH	金种子酒	-1.13	-0.54	-0.64	-0.76
12	600702.SH	沱牌舍得	-1.34	-0.34	-1.00	-0.86
13	000502.SZ	绿景控股	-4.11	1.43	2.00	-1.05
14	600575.SH	皖江物流	-1.17	-2.48	-1.50	-1.45

总体来看，在这类股票中，中洲控股排名为第一位，预测值为 1.45，第二名是贵州茅台，预测值为 1.44。由于本文选取的都是财务指标，因此，在财务方面，做的比较好的公司有中洲控股、贵州茅台、五粮液以及洋河股份。

贵州茅台今年以来，股价累计上涨达到 70%，在 2017 年的每一个月，都不断刷新股价记录。1 月曾两次创下历史新高，2 月三次创下历史新高，3 月则有十次创下历史新高。

5 展望与不足之处

近年来，随着互联网、云计算、三网融合与通信技术的迅猛发展，数据的大量快速增长已经成为很多行业共同面临的严峻挑战。这个日渐发展完善的信息社会已经进入了所谓的“大数据”时代。大数据改变着我们的生活方式和工作方式、企业的运作模式以及科学研究模式的改变。网络大数据给学术界带来了严重挑战和机遇。网络数据科学和系统科学等相关领域交叉的新兴学科方向正逐步成为学术界研究的热点问题。《Nature》和《Science》等许多杂志也都纷纷出版对大数据的研究成果。其中，在 2008 年，《Nature》出版的“Big Data”从各个方面分析了数据洪流带来的挑战，包括医药生物、互联网技术等。然后在 2011 年，《Science》设置专刊 “Dealing with Data”，用来讨论数据处理，以及海量数据带来的机遇。并在文章中指出，通过对数据的有效组织与分析，会发现更多有趣的现象，这些结果将对人们的社会生活巨大的变化。然而，现在面临的最大的问题是网络大数据的类型的复杂性、数据结构的复杂性以及内在模式的复杂性，这些问题造成了大数据存储、分析和挖掘环节的各种困难。

谱聚类算法是聚类分析中一个崭新和前沿的研究课题，由于方法本身不需要对数据的结构做出假设，并且该方法有识别非凸性数据分布的能力，适用于很多实际问题的解决。所以谱聚类在短短的几年之中就引起了国际学术界的广泛关注。谱聚类算法这一研究将大大丰富聚类算法的研究内容，并给聚类问题的求解提供了新的解决思路，具有非常大的应用潜力和科研价值。

主成分分析主要是用来解决数据处理时的难题。当面对大量数据的时候，数据的高纬度、以及数据之间的相关性给数据的计算与分析带来了不小的麻烦。倘若盲目的减少变来变量个数，很有可能损失重要的信息，所以给出错误的结论。通过主成分分析方法后，原因子的数据可以由较少维度的主成分来线性组合，降低维度，从而给我们的计算带来了很大的便利，大大提高运算效率。

为了更好地研究公司的内在价值，我们采用聚类分析和主成分分析分析两种方法。其中，我们选取的数据和财务指标都能很好的反映公司的盈利能力以及成长潜力。通过对上市公司股票的全面研究，能有效地缩小投资范围，确定投资价值。本文中以 4 个板块为例，对所用的研究方法做出详细的描述和分析。这四大板块分别是商业银行板块，房地产板块，港口板块和食品饮料板块。根

据谱聚类的分类结果，如果以板块划分为标准，可以看到谱聚类相比于 K-means 聚类更好的对股票进行了划分。虽然结果并不完全准确，但已经达到了很好的标准。其中，泛海控股和荣安地产和港口板块的大部分股票归属于同一个类别；房地产板块中的中洲控股和绿景控股以及港口板块中的皖江物流归属于食品饮料板块，大部分股票的分类符合板块的行业划分标准。其次，在对每种板块进行选股的时候，选取六个财务指标，采用主成分分析的方法。对股票池进行分析，给出股票的排名，一共投资者进行决策。

在本文中，我们通过谱聚类算法和主成分分析的一个实证应用，希望引起学者们的广泛关注，使读者对该领域有一定的认识。当然，由于作者的学术水平有限，对谱聚类算法存在的问题不能提出改进的方法。但我们相信随着对“大数据”时代的到来，数据的处理问题会成为一个非常热门的问题，将这种新的方法应用到各个研究领域意义深远。

最后，由于本人水平有限，本文存在不足之处。

1. 实证部分选取的板块过少，时间较短，精确度有待提高。
2. 对投资者在进行具体应用时，还可以收集更多的有关上市公司的资料和有关方面的信息，这样对股票的选择将会有一个更加细致和全面的研究，由此得到的投资建议将会具有更好的参考价值。

参考文献

[1] 丁世飞.靳奉祥.赵相伟.《现代数据分析与信息模式识别》[M].科学出版社.2013,60~67.
[2] 范金城.梅长林.《数据分析.》[M],科学出版社.2001,193~202.
[3] 李翔. 谱聚类算法分析及其在高维情形下的应用[D]. [硕士学位论文]. 中国科学技术大学, 2014.
[4] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述[J]. 计算机科学, 2008(07):14-18.
[5] Wiley. Anderson T W . An introduction to multivariate statistical analysis[J] ,1984.
[6] 郝瑞, 张悦. 基于因子分析和聚类分析的股票分析方法——以沪深 300 指数成分股为例[J]. 时代金融, 2014(26):135-137.
[7] 陈琦. 聚类分析和判别分析在股票投资中的应用[J]. 中国市场, 2011(26):69-72.
[8] 李庆东. 聚类分析在股票分析中的应用[J]. 辽宁石油化工大学学报, 2005(03):94-96..
[9] 王元卓, 靳小龙, 程学旗. 网络大数据:现状与展望[J]. 计算机学报, 2013(06):1125-1138.
[10] 江冬明. 主成分分析在证券市场个股评析中的作用[J]. 数理统计与管理,2001,(3).
[11] 潘琰.程小可. 上市公司经营业绩的主成分评价方法[J]. 会计研究, 2000,(1).
[12] 董逢谷. 《上市公司综合评价》[M]. 上海财经大学出版社, 2002.
[13] 劳兰珺,邵玉敏.中国股票市场行业收益率序列动态聚类分析[J].财经研究, 2004(11):75-82.
[14] 关昕,周积林.基于改进谱聚类的图像分割算法[J]. 计算机工程与应用, 2014(21):184-188.
[15] 牛科,张小琴,贾郭军.基于距离度量学习的集成谱聚类[J]. 计算机工程, 2015(01):207-210.
[16] 邓秀勤. 聚类分析在股票市场板块分析中的应用[J]. 数理统计与管理, 1999(05):1-4.
[17]王勇. 基于流形学习的分类与聚类方法及其应用研究[D]. [博士学位论文]. 国防科学技术大学, 2011.
[18] 陈国华, 廖小莲, 夏君. 证券投资分析的聚类分析方法, 2011[C].
[19] 黄玮强, 庄新田, 姚爽. 中国股票关联网络拓扑性质与聚类结构分析[J]. 管理科学, 2008(03):94-103.
[20] 孙玉侠. 数据挖掘中的谱聚类算法研究[D]. [硕士学位论文]. 中国海洋大学, 2010.
[21] 侯文. 对应用主成分法进行综合评价的探讨[J]. 数理统计与管理, 2006(02):211-214.
[22] 任若恩,王惠文.多元统计数据分析[M].北京:国防工业出版社, 1997.
[23] 黄宁. 关于主成分分析应用的思考[J]. 数理统计与管理, 1999(05):44-46.
[24]吴燕萍. 上市公司综合评价分析[J]. 统计与决策, 1998(04):20-22.
[25] 丁欢新. 商业银行竞争力评价的现状及评价指标构建[J]. 商业经济与管理, 2003(11):56-59.
[26] 赵昌昌, 曹学勤, 刘生元. 中外银行竞争力实证分析[J]. 当代经济科学, 2003(04):85-88.
[27]王益.中国银行业竞争力分析(上)[J].管理现代化,2002(5): 4～13.
[28] 鲁志勇,于良春.中国银行竞争力分析与实证研究[J].改革, 2002(3): 61～67.

[29] 张守凤,乐菲菲,李丽华.层次分析法在商业银行竞争力评价中的应用[J].统计与决策, 2003(2): 61～62.
[30] 姜波.我国商业银行竞争力比较模型及实证分析[J].商业研究, 2003(3): 9～11.
[31] 张守凤,柳兴国,柳华真.商业银行竞争力的模糊多属性评价方法[J].济南大学学报:自然科学版, 2003, 7(2): 122～124.
[32] 柯冰.钱省三.聚类分析和因子分析在股票研究中的应用.上海理工大学学报,2002(4):371~374.
[33] 冯伟.孙德山. 聚类分析在金融投资分析中的应用.辽宁师范大学学报, 2008(3)43 一 5.
[34] 李庆东, 李颖. 证券投资分析方法新探索——聚类分析方法应用[J]. 现代情报, 2005(11):225-227.
[35] 施大洋.杨朝军. 证券投资风格研究.经济师, 2005(9):130~210
[36] 狄明明.孙德山.聚类分析和支持向量机在股票研究中的应用.计算机技术与发展,2008(7):18~246.
[37] 司文武, 钱沄涛. 一种基于谱聚类的半监督聚类方法[J]. 计算机应用, 2005, 25(06):1347-1349.

附录

个人简历、在学期间发表的学术论文与研究成果

郑扬（1984 年 1 月），男，河南郑州人，2008 年郑州大学西亚斯国际学院毕业，获得管理学学士，2012 年郑州大学在职研究生，主要研究方向：金融工程。

已发表论文：

在《郑州师范教育》学刊 2015 年第 4 卷第 6 期第 55 页发表论文《基于主成分分析的房地产行业研究》。

致谢

本学位论文是在我的导师郑州大学数学与统计学院贾军国教授、中国农业银行河南省分行国际业务部总经理张郑阳两位老师的指导下完成的。两位老师治学严谨，对工作更是精益求精，良好的工作作风深深感染了我，也激励这我在未来的道路上勇敢前行。从选题、理论方法的研究、数据的下载与处理，然后到初稿的完成，修改直至论文的最终完成，每个细节都令人印象深刻，至今难忘，这一切都离开两位老师的殷切指导和宝贵的意见。张老师从论文选取的背景、贾老师从论文选取的数学方法分别给以我学术上的重点指导，在此谨向我的老师致以诚挚的谢意和崇高的敬意。

三年的研究生生活不仅带给我丰富的学术知识，也更加端正了我做人做事的态度。特别是作为一名在职研究生，这段经历是我人生中非常难忘和美好的时光。在职人员修习研究生学位并不是一件轻松的事情，虽然工作很劳累，但每次回到校园就感觉精神充沛，激情满满。尤其是郑大美丽的校园，浓厚的学习氛围，令人向往、流连忘返。这里简直是一座世外桃源，能够在这里学习是一件非常幸福的事情。

另外，在工作中，我发现专业知识的欠缺，所以非常珍惜这种难得的学习机会。通过这段时间的学习，我不仅提高了查阅文献、阅读文献的能力，还提高了独立思考的能力。从发现具有价值的问题，搜索各种国内外的文献、方法，结合众多学者的观点，在加上独自的思考能力，才能取得完美的结果。非常的感谢这三年研究生生活中我的老师和同学们，是他们让我的研究生生活变得多姿多彩。最后，我还要感谢这群一起学习的同窗们，是这份奇妙的缘分让我们相聚在一起，共度三年的美好时光，在期间，有欢乐、有泪水，有你们的支持和帮助，让我克服一切，直至完成本文。

如今，我的心情无法平静，有太多的无言帮助让我难以忘记，在此，请接收我真诚的敬意！最后我还要感谢生我养我的父母，是你们把我带到这个世界上，让我看到这么美丽的景色，这么丰富多彩的世界。在我的成长过程中，不断鼓励我，在我犯下错误时，不断鞭策我，在我遇到难关时，不断激励我，如今，我定会让你们为我骄傲！再次谢谢你们!