分类号_____ 密　　级_____

U D C _____ 学校代码_____

硕士研究生学位论文

量化交易策略模型应用研究

学院（部、所）：统计与数学学院

专　　业：应用统计学

姓　　名：李　艺

导　　师：陈贻娟

论文起止时间：　2013 年 9 月—2014 年 3 月

学位论文原创性声明

声明：本人所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文版权使用授权书

本人完全了解云南财经大学有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文和论文电子版，允许学位论文被查阅或借阅；学校可以公布学位论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存、汇编、发表学位论文；授权学校将学位论文的全文或部分内容编入、提供有关数据库进行检索。

（保密的学位论文在解密后遵循此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘　要

随着电子化交易平台在全世界股票市场的逐渐成熟，量化交易成为各证券市场的新宠。在国外，量化交易被广泛应用于投资机构的实际操作中，近年来迅速兴起但饱受争议的高频交易就属于量化交易的新兴方式，因其具有程序化交易、巨大的交易量、极短的持仓时间、单笔收益率低而总收益高等特点，成为国内外投资研究领域关注的焦点。而我国证券市场起步较晚，电子化交易平台也在逐步完善，量化交易模式也处于发展阶段，加之我国 T+0 交易即将重启，这将对我国量化交易的推广起到重大的推动作用。因此，针对我国证券市场的量化交易模型研究具有非常重要的现实意义，是我国证券市场投资研究的一个重要方向。本文以量化交易策略模型为研究对象，选取了六只样本股票以及一只预测股票，涉及当日交易明细、量化交易明细，其中包括成交价格、成交量、买卖前三手数量、成交时间、量化交易成交量等股票市场数据，利用策略模型对预测股票的量化交易进行策略研究预测。

首先，文章依先前对于股票投资策略的研究，首先对股票价格波动与大单投资策略之间的关系进行研究分析，利用列联表检验及独立性检验，发现量化交易策略与价格波动之间极有可能存在着一定的非独立的关系。

其次，为深入探究这种非独立型关系，文章利用累积 probit 模型对样本数据进行拟合，并使用所拟合的模型对预测集股票进行策略分析，得出 1）即期变量较前期变量对于投资决策有更为显著的影响，2）量化交易策略偏向于与市场价格变化做同向交易，3）在高成交量的活跃市场氛围下，量化交易策略更倾向于进行买入交易，而在成交量比较低迷的情况下量化交易策略更倾向于进行卖出策略，4）买入挂单和卖出挂单对量化交易策略交替产生影响，本期高卖出挂单量趋向于使量化交易策略进行卖出交易而使得下一期量化交易策略趋向于进行买入交易，5）量化交易策略的实施对于股价走势的具有较为严格的要求，随后对其预测集进行预测并观察策略误判率。

再次，为了探究不同的模型方法对策略决定的影响，文章利用贝叶斯网络和支持向量机（SVM）对数据进行模型训练以及对预测集进行预测，并将其与累积

probit 模型的预测结果进行精度比较，得出三种模型对于交易策略的预测，各有优势，在预测精度方面 SVM 方法较好，而对于模型解释性，累积 probit 模型有较为明显的优势，而贝叶斯网络处于以上两种模型中间。

最后，根据以上分析发现策略的制定实施与价格具有非常密切的关系，文章利用多元时间序列模型，对股票价格进行拟合分析与预测，发现 VARX 模型和状态空间模型都能够对股价进行很好的预测。

文章结尾结合量化交易的特点，对我国证券市场面对新型交易方式在监管层面提出一些合理化建议。

关键词：量化交易；累积 probit 模型；贝叶斯网络；支持向量机；多元时间序列

Abstract

With the electronic trading platform in the worldwide stock market gradually mature, quantitative trading becomes prevalent in worldwidestock market. Quantitative trading is widely used in operation of investment institutions, abroad.HFT,rapid rise in recent years but controversial,belongs to emerging of quantitative trading way. Because of it's characters, such asprogramming trading, huge volume and yield of extreme position time, and low profitable in one trade but high profitable in total revenue, HTF has already become the focus in the field of investment researchdomestically and internationally. The mushrooming of securities market in China,however, isrelatively late. The electronic trading platformsimprove gradually, and quantitative trading platform is still in the development stage. Meanwhile, T + 0 transaction in China will restart shortly, will be to our country in this play a major role in the promotion of trade, which would dramatically affect the popularization of quantitative trading.Therefore, the quantitative trading model research on Chinese securities market has significant practical implications as well as provides further research directions for marketing investment research. The object of this paper is trading strategy based on quantitative model. The data is chosen from six sample stocks and one predictable stock, which involves in transaction details and quantitative transaction detailsthat including third-hand before clinch a deal the price, volume, sales amount, transaction time, quantitative trading volume, and other stock market data. The purpose of the paper is to understand the quantitative trading strategy and to make trading decision by using strategy model. In accordance with previous research on stock investment strategy, this article first analyzes the relationship between the fluctuation of stock price and the large investment strategy. The article finds some dependent relationship between quantitative trading strategy and fluctuation of price. Secondly, to further explore the relationship between the dependent model, the article uses the cumulative probit model to fit the sample data,

and use the fitted model to forecast stock strategy analysis.By doing this procedure, Ifind that 1)Spot variables have more significant affect on investment decision than prophase variables 2) quantitative trading strategy prefer to make synthetic deal with market price changes, 3) quantitative trading strategies prefer to purchase stock under active stock market with high volume and sell stock under stock market with low volume. 4) buying are cancelled and selling are cancelled affect quantitative trading strategy alternation, the current high volume sold deity tend to make quantitative trading strategy to sell trading makes the next phase of quantitative trading strategy, tends to buying, quantitative trading strategy is affected by the changes of buy and sell pending. 5) the implementation of quantitative trading strategy for the share price has strict requirements. I predict the predictable set and observe the rate of error. Then, to explore different ways to model the impact on the strategy decision the article uses Bayesian network and support vector machine (SVM) model to predict the prediction set and start the model training. By compared with cumulative probit model, I find that each of these three model has their own advantages. For example, SVM is better under prediction precisely criterion. Under explanation criterion, cumulative probit model has better performance. Bayesian network is located between two models. Finally, I find strong relationship between creating strategy and share price by analysis above. The paper tries to fit the stock price and make prediction by multivariable time series model, and find that VARX model and state space model have great prediction performance for stock price. Finally, combine with the characteristics of quantitative trading, Isupply several reasonable suggestions about monitoring the new trading way in Chinese securities market.

Keywords: Quantitative rading;Ordered probit model;Bayesian network;SVM; Multivariate time series

目 录

第一章 引言

第一节 研究背景

量化交易有多种不同的叫法，比如自动化交易（Automated Trading），算法交易（Algorithmic Trading），等等。其实到目前为止,行业内并没有对量化交易有一个明确而被大众都认可的定义，但只要是通过根据预先编制的计算机程序指令来完成这笔交易就应该属于量化交易的类别[1]。分类程序交易可以分为两个级别的决策和实施。量化交易的程序化是指以各种实时/历史数据为输入变量，将这些变量数据输入到事先设计好的算法计算得出交易决策的过程，整个决策过程包括：对于决策资产，在什么时间以怎样的价位进行什么样的（买/卖）操作以及买卖的数量等；而进行量化交易的决策执行则是计算机算法来优化交易订单执行的过程。

数量投资相对于传统投资方法的优越性主要来自两个方面：首先，现代投资组合理论强调通过多元化的投资组合来消除非系统性风险，以降低整体的投资风险。但事实上，由于一个人的视野和精力是有限的，基金管理人、科研人员都不可能进行大范围的选股和高频率验证并计算选股的好坏，投资策略的形成没有宽度和广度，那么它形成了一孔之见。通过人工选择的投资组合是很难实现最优分配的，而且并不能确保追求利润的同时进行风险管理和达到投资目标。而量化投资的角度较宽，且在计算机的帮助下能够高效、海量的对信息进行处理，更广泛的搜索和验证投资机会并准确的处理，以消除投资组合配置的限制。其二，行为金融学认为，投资者是不理性的。任何投资个体的判断和决策过程中会受到认知，情感，意志等心理因素不同程度影响。基金经理和投资研究分析师在一段时间内跟踪一只股票，因为总是关心股价表现与基本面的变化，可能会出现不同程度的情感依赖，即“和股票相恋”。如果有一个下降的趋势，也可能是从因为过度自信、抵制心理分析的出发点进行考虑，并最终导致非理性行为偏差,。而量化投资依靠计算机配置投资组合，克服了人性弱点，使投资决策更科学、更理性。

A 股市场的发展程度决定了当前市场上不可能存在完全量化的基金产品。量

化从一开始也不是作为定性的对立面而提出的方法，它是将定性分析中的技术分析策略用模型固化，替代过程中可以用电脑进行的部分并将其效用极大优化。应该架设怎样的平台、构建怎样的模型、输入怎样的因子，都是建立在定性分析上的总结。而为了预防小概率事件的发生，还应该为模型配备精良的开发团队，包括定性和定量分析专家，来跟踪观测模型的合理度、与市场趋势的匹配度以及实际投资表现。

量化投资的特点：

1、纪律性

所有的决策都是依据我们研究出来的模型做出的。我们有三个模型：一是大类资产配置模型、二是行业模型、三是股票模型。根据大类资产配置决定股票和债券投资比例；按照行业配置模型确定超配或低配的行业；依靠股票模型挑选股票。纪律性最主要的表现在于我们能够信任并依靠我们的模型，每一天的决定之前，首先先对所有模式进行运行，并根据模型的结果做出决策，而不是由自己的感觉。纪律性可以克服人性中的缺点，如贪婪、恐惧，还可以极大的克服侥幸心理和认知偏差，行为金融理论在这方面有很多的讨论。纪律性的另外一个好处是可跟踪。量化投资作为一个定性想法的合理应用，客观地在组合中去体现这样的组合思想。我们的每一个决策都是有理有据的，特别是有数据支持的。

2、系统性

具体表现为“ 三多 ”。首先是多层次，包括在资产配置、行业类别的选择、股票选择等方面，我们有三层的模型；其次是多角度，定量投资的核心投资思想包括宏观周期、市场结构、估值、成长、盈利质量、分析师盈利预测、市场情绪等多个角度；再者就是多数据，就是对海量数据的处理。

3、套利思想

定量投资主要的目标在于寻找错误定价，通过全面、系统性的扫描捕捉错误定价、错误估值带来的套利机会。、

4、概率取胜

这表现为两个方面，一是依靠股票组合来获得利润，而不是一个或几个股票。二是定量投资不断的从历史中挖掘有望在未来重复的历史规律并且加以利

用。

第二节 文献综述

量化交易策略源自于算法交易（Algorithmic Trading），许多学者在对量化交易策略进行研究时，往往在算法交易模型的基础上对量化交易策略模型进行调整和分析。对于量化交易策略的研究，主要集中在以下几种交易策略的研究。

作为量化交易的一种，高频交易成为现今证券市场特别是欧美等发达证券市场大机构争相研究使用的交易模式之一。Kearns、Kulasza 和 Nevmyvaka（2010）在《Empirical Limitations on High Frequency profitableity》[2]一文中基于 TAQ 数据库的股票数据信息，对高频交易的盈利能力进行了跟踪分析。文章中，作者对高频交易年盈利能力进行了计量。根据前文中我们提到的高频交易的特点，作者假设每笔高频交易都能够获利，在该假设前提下，高频交易的年盈利在 210 亿美元左右，远远高于其他交易策略。说明高频交易策略在获利能力上具有相当的优势。

T ·Hendershott 和 R ·Riordan（2009）在《Algorithmic trading and information》[3]一文中首次对算法交易进行了具体的模型计算。文章中选取了包括苹果、全美铝业在内的在纽交所和纳斯达克上市 120 只股票作为样本股票，对交易策略模型进行分析。该文章首次提出了具体的样本股票，并利用该 120 只股票数据对算法交易进行了具体的分析。该文章的贡献在于其为后来研究者进行算法交易的量化分析奠定了一定的基础。

Cvitanic 和 Kirilenko（2010）在《High Frequency Trader and Asset Prices》一文中首次利用理论模型来解释说明高频交易对市场的影响。文章中指出，高频交易对于价格发现机制具有显著作用，它可以凭借其交易特点，使众多的影响供求关系变动的因素全部集中于交易所的交易池内，并通过公开竞争叫价的方式，将各种因素转化为一个统一的成交价格。并且，随着时间的推移和其他各种条件的变化，市场成交价格也在不断地进行调整。

Hull（2009）在《Options, futures and other derivatives》[4]一文中提出 Delta 中性策略作为量化交易的典型策略。Delta 是指该衍生证券的价格变化对其标的资产价格变化的比率，Delta 中性是指 Delta 为零的状态。如果在此基础上进行高频对冲，则 Delta 中性策略就成为了一种量化交易策略模式。该策略模式必须

在标的资产价格下降时将其卖出，在其价格上涨时将其买入，因此它在本质上是一种追涨杀跌交易方法。从国外交易经验看，做市交易是量化交易策略的主流。目前我国已经在利率掉期、国债等部分市场引入做市商交易制度，如果在更多市场实施该制度，量化交易还会有进一步发展。

王俊杰（2013）在《量化交易在中国股市的应用》[5]一文中指出，国内股票市场表现低迷,传统投资策略业绩平平。而获得更加稳定的收益越来越受到广大投资者的青睐。在此发展契机下，量化投资这一以追求绝对收益为目标投资策略得到广泛关注，并得到快速发展。王俊杰的文章以量化交易策略为主要的研究内容，通过对已有文献的研究成果的深刻总结，建立一个从投资时间的选择时到对股票的选择的实际可行的量化交易系统。在对时间的选择模型方面，文章系统分析了行业指数存在的行业轮动特征和持续性，并在时间序列的基础上，构建量化投资时机策略。该策略在回测期表现优于行业指数和动量策略。

镇磊（2010）在《基于高频数据处理方法对 A 股算法交易优化决策的量化分析研究》[6]一文中，首先介绍了算法交易的兴起和发展现状，以及当前算法交易的常用算法和主流设计思想，并通过对 A 股市场的发展分析，对在 A 股市场进行算法交易所需要注意的问题进行分析总结。文章在高频数据处理方法的基础上进行再研究，并提出了一种适合 A 股市场交易规则的交易算法，这种交易算法对无交互效应和有交互效应两种情况下交易策略的设计进行分别考虑。无交互效应的模型首先通过 ACD 模型建模得到交易持续期序列，选择交易时间点，然后分别对每个交易期间的成交量分布和每个期间的价格变化进行预测，最后根据价格的变化对交易量进行调整。而对于交互效应，通过事件分析法引入交互因子，并对其影响进行修正建模分析，得到相应的交易策略。在算法的具体实施过程中，文章对传统方面进行了一定的创新。在预测交易时间时，提出了一种带非对称效应的扩展 ACD 模型以解释外在因素对未来交易量持续期的影响;在预测交易量分布，考虑不同时间段的记忆周期差异，提出了一种基于自相关的分时 VWAP 算法

第三节 国内外发展现状

2000 年，总部位于纽约的第一个完全电子化的美国期权交易所——国际证券交易所（ISE）建立了。截止到 2010 年，美国提供完全电子化或者电子化与交

易大厅相结合的期权交易平台的交易所数量已经达到了七个。而根据总部设在波士顿的 Aite 集团估计，电子交易量已经从 2001 年的 25%迅速增至 2008 年的 85%，而随着信息技术的不断发展和广泛应用，电子交易量现已接近 100%。成熟的证券市场以及先进的计算机技术，为发达国家的资本市场技术的创新提供了有力的支撑，而随着算法交易的不断完善，交易速度的不断加快，量化交易在发达国家的资本市场得到广泛的应用，并成为机构投资者重要的操作手段。

根据信息显示，由于量化交易能在根本上实现全自动电脑化下单，从而排除人的情感因素达到不断优化交易模型同时能够适度地控制风险，因此被各大投资机构广泛使用。在美国证券期货市场，量化交易虽然仅出现 40 余年，但其占市场总交易量的比例已经超过了 60%。根据华联期货介绍，早在 2012 年上半年，量化交易量占国内股指期货交易量的比例已达 20%左右。在美国市场上，部分量化交易的基金也取得了较高的回报。例如基金巨头文艺复兴为例，2013 年 10 月份，文艺复兴基金在当月的回报率高达 8.65%，大幅超过股指 4.6%的回报率，该基金目前的规模约为 87 亿美元。此前莫尼塔统计报告显示，美国量化交易的基金近年来的复合平均回报约为 10.8%，高于市场的整体平均收益。国信证券监事会主席何诚颖认为，量化投资是金融中的“核武器”，在美国等地区呈现蓬勃发展的状态，并带来了巨大的规模与收益。

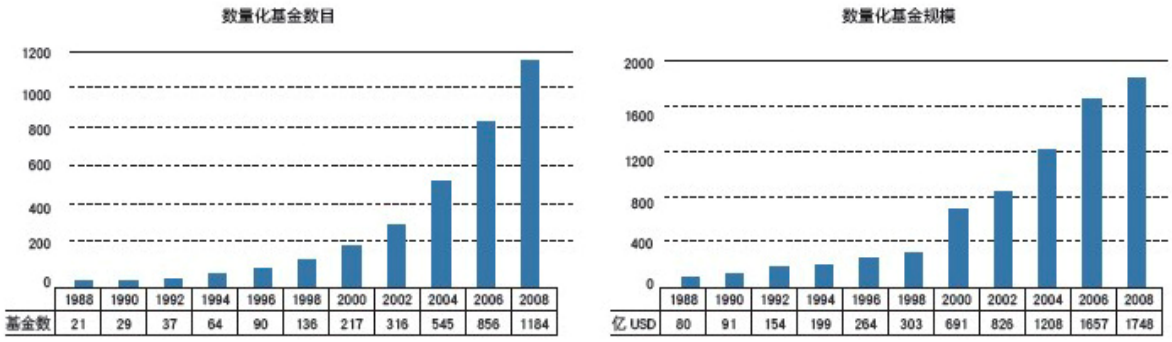

图 1.1 美国量化基金发展情况

从国际市场来看，量化交易真正的用武之地在于股票市场。“量化交易主要是赚个人投资者的钱，期货标的毕竟有限，在 A 股遍地散户的情况下，量化交易应该会取得非常好的回报，这是 A 股量化交易的红利，所以得趁早。”一位准备进军 A 股的私募基金经理说，国外的量化交易起步较早，国内的量化交易基本都是由海外教育背景研究员在做，目前国内已经开始有许多的量化交易机构

逐渐介入了沪深 300 股指期货的套利交易，所以现在沪深 300 的每日股价波动越来越趋于规则，套利的空间随之也越来越小，真正的机会则在分散的股票市场。

国内外机构被我国股票市场可观的收益所吸引。国内的主力投资机构中，政策限制较少的部分券商和保险已经先行试水股票市场。中国量化投资学会理事长、量化投资经理丁鹏认为，目前加总券商和险资自营的量化套利资金，以及公私募机构量化基金的规模，国内量化投资资金的体量已经达到 1000 亿元，其主要来源仍是券商和保险的自营盘。“ 最近一个多月经常接到很多香港那边打过来的电话，咨询我们的交易软件能不能实现全自动的量化交易问题。” 一家美国量化交易软件的中国代理商称，据他了解，由于有些国外的软件已经能够实现国内的数据接入，有国内的量化交易团队买下了该交易软件的代理权，在赚取了丰厚的利润后已经不对外出售。

第二章 量化交易策略的列联表分析

第一节 列联表分析方法介绍

（一）定义

列联表是观测数据按两个或更多属性（定性变量）分类时所列出的频数表，假定将 n 个个体根据两个属性 A 属性和 B 属性进行分类。属性 A 有 r 类：$A_1,\cdots,A_r$，属性 B 有 m 类：$B_1,\cdots,B_m$。n 个个体中既属于 A_i 属性又属于 B_j 属性的有 n_{ij} 个，其概率 $p_{ij}=\frac{n_{ij}}{N}$，则可构成以下二维 $r\times m$ 列联表：

表 2.1 二维 $r\times m$ 列联表

		列属性 B				合计
		B_1	B_2	$\cdots$	B_m	
行属性 A	A_1	p_{11}	p_{12}	$\cdots$	p_{1m}	p_{1+}
	A_2	p_{21}	p_{22}	$\cdots$	p_{2m}	p_{2+}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	A_r	p_{r1}	p_{2r}	$\cdots$	p_{rm}	p_{r+}
合计		p_{+1}	p_{+2}	$\cdots$	p_{+m}	1

其中：$p_{i+}=\sum_j p_{ij}, i=1,\cdots,r$ ；$p_{+j}=\sum_i p_{ij}, j=1,\cdots,m$ ；$1=\sum_i p_{i+}=\sum_j p_{+j}$。

（二）二维列联表独立性检验

$$\begin{cases} H_0: \text{分类数据独立} \\ H1: \text{分类数据不独立} \end{cases}$$

二维列联表的独立性检验实际上是分类数据的检验问题，是一种无方向检验问题。一个 $r\times m$ 的列联表实际上有 $r\cdot m$ 个类。当其中的两个属性相互独立时，个体在每一类中的概率由 p_{+1}，p_{+2}，…；p_{+m} 和 p_{1+}，p_{2+}，…；p_{r+} 完全确定。在独立性成立时，由似然函数：

$$\prod_{i=1}^{r}\prod_{j=1}^{m}p_{ij}^{n_{ij}}=\prod_{i=1}^{r}\prod_{j=1}^{m}(p_{i+}p_{+j})^{n_{ij}} \qquad (2.1.1)$$

得到参数 p_{i+} 和 p_{+j} 的最大似然估计分别为 $\hat{p}_{i+}=n_{i+}/n$ 和 $\hat{p}_{+j}=n_{+j}/n$，从而期望频数 np_{ij} 的估计为：

$$np_{ij} = n_{i+}n_{+j}/n$$

从而构造列联表独立性检验问题的 $Pearson\chi^2$ 统计量，

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{m}\frac{(n_{ij}-n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{m}\frac{(n_{ij}-n_{i+}n_{+j}/n)^2}{n_{i+}n_{+j}/n} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{m}\frac{n_{ij}^{\ 2}}{n_{i+}n_{+j}/n} - n\end{aligned} \tag{2.1.2}$$

其自由度为类别个数 $r\times m$ 减去 1 再减去参数个数 $(r+m-2)$，得其自由度为 $(r-1)(m-1)$。水平为 α 的 χ^2 检验的拒绝域为 $\chi^2 \geq \chi^2_{1-\alpha}((r-1)(m-1))$，即在 $\chi^2 \geq \chi^2_{1-\alpha}((r-1)(m-1))$ 时，认为独立性不存在，反之认为独立性成立。其对应的 p 值为自由度为 $(r-1)(m-1)$ 的 χ^2 大于等于 χ^2 统计量的的概率。

同样，我们可以利用似然比统计量来对列联表的独立性进行检验，其似然比检验统计量为：

$$-2\ln\Lambda = -2\sum_{i=1}^{r}\sum_{j=1}^{m}n_{ij}\ln(\frac{\hat{p}_{ij}}{n_{ij}/n}) = -2-2\sum_{i=1}^{r}\sum_{j=1}^{m}n_{ij}\ln(\frac{n_{i+}n_{+j}}{n\cdot n_{ij}}) \tag{2.1.3}$$

在原假设为真时，该似然比检验统计量渐近于 χ^2 分布，即似然比检验统计量的极限分布仍是 $\chi^2((r-1)(m-1))$。

（三）二维列联表相合性检验

二维列联表的相合性检验实际上是一种列联表有顺序的检验，用来检验有顺序的两种属性 A 属性和 B 属性的相关性问题。假设 A 属性从 $A_1,\cdots,$到A_r 以及 B 属性从 $B_1,\cdots,$到B_m 都有一个由小到大的顺序关系，而相合性检验的目的在于验证是否存在如下假设：即 A 属性值比较大的个体，它的 B 属性也比较大（或比较小）。而这种相合性存在两种情况，即正相合和负相合，而相合性检验我们通常使用 Kendall Tau-b 系数 τ 进行检验。令

$$\begin{cases} T_A=\sum_{i=1}^{r}\binom{n_{i+}}{2}=\sum_{i=1}^{r}\frac{n_{i+}(n_{i+}-1)}{2} \\ T_B=\sum_{i=1}^{r}\binom{n_{+j}}{2}=\sum_{i=1}^{r}\frac{n_{+j}(n_{+j}-1)}{2} \\ T_{AB}=\sum_{i=1}^{r}\sum_{j=1}^{m}\binom{n_{ij}}{2}=\sum_{i=1}^{r}\sum_{j=1}^{m}\frac{n_{ij}(n_{ij}-1)}{2} \end{cases} \quad (2.1.4)$$

其中$\binom{n_{i+}}{2}$，$\binom{n_{+j}}{2}$，$\binom{n_{ij}}{2}$为第i的第n_{i+}个数，第j的第n_{+j}个数，第i行第j列的n_{ij}个数中任取两个数的组合，故$T_A \geq T_{AB}$、$T_B \geq T_{AB}$，且

$$\tau=\frac{z}{\sqrt{[n(n-1)/2-T_A][n(n-1)/2-T_B]}} \quad (2.1.5)$$

其中$z=G-H$，$G=\sum_{i<k}\sum_{j<t}n_{ij}n_{kt}$，$H=\sum_{i<k}\sum_{j>t}n_{ij}n_{kt}$。

τ的值在-1 和 1 之间，其越接近于 1，越倾向于认为正相合；越接近于-1 越倾向于负相合。

第二节　利用列联表检验进行案例分析

（一）数据介绍

本文所使用数据来自于 2013 年 12 月 27 日（星期五）上海证券交易指数，取自新浪财经[1]，涉及如下样本股票指数数据：

表 2.2 所使用样本股票信息

股票代码	股票名称	股票代码	股票名称

[1] 新浪财经 http://finance.sina.com.cn/

600016	民生银行	600228	昌九生化
600036	招商银行	600837	海通证券
600999	招商证券	601989	中国重工
601166	兴业银行		

其中，600016、600228、600036、600837、600999、601989 作为训练集股票样本，601166 作为预测集股票样本，所用数据涉及当日交易明细、量化交易明细，其中包括成交价格、成交量、买卖前三手数量、成交时间、量化交易成交量等股票市场数据。

（二）样本股票的选取

我国每日量化交易量呈现量大但分布不均的特点，有的股票当日大单成交量在几百笔之多，而有的股票当日大单成交量仅为几笔。对国内进行量化交易量化交易模型研究，为了便于研究，故选取当日内量化交易数量较多的几只股票，最终选取民生银行等七只股票作为研究对象。

（三）数据的处理

文章获取的原始数据为各股每日的交易明细，但原始数据中时间分布并不均匀，即时间间隔并不是严格的等距分布，所以首先将各变量按照时间加权，每 10 秒为一个单位时间，将各变量处理为每 10 秒数据：

$$X_{10}=\frac{1}{10}\sum_{i=1}^{n}X_i\cdot T_i$$

其中：X_{10}为各变量的单位时间数据，X_i为 10 秒内该变量出现的第i个数据，T_i为第i个数据所持续时间，我们以整 10 秒数作为每十秒的开始，处理示例如下：

时间	9:59:57	10:00:02	10:00:05	10:00:06	10:00:09
价格	9.87	9.89	9.88	9.90	9.88

则经时间加权的价格为：

$$price=\frac{1}{10}(9.87\times2+9.89\times3+9.88\times1+9.90\times3+9.88\times1)=9.887$$

在对量化交易的确认中，若单位时间内同时出现买卖相反的两笔量化交易数据，则对相反的两笔交易进行比较，选取交易额较大的一组交易方向作为该单位时间内的量化交易方向。为研究滞后变量对决策的影响，我们选取 strategyT6 为决策时期，其余变量的 T6 为同期数据，T5---T1 为各变量的滞后 1 期至滞后 5 期数据。各变量及其含义如下表：

表 2.3 各变量含义说明

符号	变量	符号	变量
strategy	交易决策	price	成交价格（元）
turnover	成交量（手）	buy	前三买入需求量和（手）
sell	前三卖出需求量和（手）	priInc	价格增长率
turnInc	成交量增长率	buyInc	买入需求量增长率
sellInc	卖出需求量增长率		

其中：strategy=-1 代表大单卖出交易；strategy=0 表示大单未进行交易；strategy=1 表示大单买入交易；$priInc=(price_t-price_{t-1})/price_{t-1}$；$turnInc=(turnover_t-turnover_{t-1})/turnover_{t-1}$；$buyInc=(buy_t-buy_{t-1})/buy_{t-1}$；$sellInc=(sell_t-sell_{t-1})/sell_{t-1}$。

（四）高频交易策略与价格波动之间的关系

为了探究大单量化交易策略与价格波动之间的关系，本文首先对大单量化交易策略与价格波动之间的独立性进行二维列联表检验。首先我们需要对价格波动进行定义：假设价格增长在 0.0005 为价格正向剧烈波动，记为 priIncT6A=1，价格增长在-0.0005 为反向剧烈波动记为 priIncT6A=-1，价格增长介于-0.0005 与 0.0005 之间为价格正常浮动，或价格未产生剧烈波动，记为 priIncT6A=0。对样本集数据进行二维列联表检验可做一下原假设与备择假设：

$$\begin{cases} H_0: \text{大单交易决策与价格波动之间独立} \\ H_1: \text{大单交易决策与价格波动之间不独立} \end{cases}$$

通过列联表检验可得以下检验分析结果：

表 2.4 列联表频数分布表

		priIncT6A			合计
		-1	0	1	
strategyT6	-1	358	674	200	1232
	0	1005	3755	912	5672
	1	250	976	474	1700
合计		1613	5405	1586	8604

由表 2.4 我们可以看出，量化交易策略与价格波动之间极有可能存在着一定的非独立的关系，为了确定是否存在这种非独立的关系，我们对列联表进行 χ^2 检验和似然比检验，其检验结果如表 2.5：

表 2.5 列联表卡方检验与似然比检验结果

统计量	自由度	值	概率
卡方	4	221.6647	<.0001
似然比卡方检验	4	203.0483	<.0001
Mantel-Haenszel 卡方	1	136.0598	<.0001
Phi 系数		0.1605	
列联系数		0.1585	
Cramer V		0.1135	

由表 2.5 检验结果可得：

$$\chi^2 = 221.6647, \quad p = P(\chi^2(4) \geq 221.6647) < 0.0001$$

$$-2\ln\Lambda = 203.0483, \quad p = P(\chi^2(4) \geq 203.0483) < 0.0001$$

由以上结果可以看出, 该组数据对应的 χ^2 统计量值为 221.6647, 所对应的 p 值为小于 0.0001；其对应的似然比卡方统计量为 203.0483，所对应的 p 值也小于 0.0001，故可得出结论：该组数据检验结果拒绝原假设，说明量化交易决策与价格波动之间不独立，即价格波动与决策制定之间具有一定的影响。

为了进一步探究这种影响之间的关系，我们对列联表进行二维列联表相合性检验，其检验结果如表 2.6：

表 2.6 二维列联表相合性检验结果

统计量	值	渐近标准误差
Gamma	0.2162	0.0193
Kendall's Tau-b	0.1173	0.0106
Stuart's Tau-c	0.0916	0.0084
Somers' D C|R	0.1208	0.011
Somers' D R|C	0.1139	0.0103
Pearson 相关	0.1258	0.0114
Spearman 相关	0.1261	0.0114

由检验结果可得：相合性检验的 kendall 统计量值为 τ=0.1173，其渐进标准误差为 0.0106，用统计量值加减两倍标准误差作为统计量的 95%的误差，可得统计量的置信区间为（0.0961,0.1385），其值为正，故可以判断量化交易策略与价格波动之间存在着正的关联，但是这种正的关联性并不是很强。也就说明，价格正向波动越大，越趋向于进行买入交易，价格反向波动越大，趋向于进行卖出交易。但是该方法并不能很明确的对量化交易策略进行判断，仅仅说明量化交易与价格波动之间存在一定的关联。为进一步探究量化交易策略的决定因素以及如何对预测集股票的量化交易决策进行决定，我们将使用累积 probit 模型、支持向量机以及贝叶斯神经网络对样本股票进行模型训练以及对预测集股票的量化交易策略进行决策判定，来探究量化交易策略的决定因素以及决策方法。

第三章 量化交易策略的累积 Probit 模型研究

第一节 累积 Probit 方法介绍

（一）方法论

假设有 N 个观察对象（标号为 $N=1,2,\cdots,N$），对此 N 个观察对象的某一特征 D_i 进行观察，而 D_i 的取值越高，说明该观察对象的 D_i 特征状况越好，而某观察对象的 D_i 的好坏或者数值的高低取决于该观察对象既有的多种因素。假设特征指数 D_i 是 K 个因素（决定变量）的一个线性函数，这 K 个因素的取值对于个人 i 来说，为 X_{ik}， $k=1,2,\cdots,K$。这意味着特征指数可以表示成：

$$D_i=\sum_{k=1}^{K}\beta_k X_{ik}+\varepsilon_i=C_i+\varepsilon_i \qquad (3.1.1)$$

其中，与第k个变量（$k=1,2,\cdots,K$）相关的系数为β_k，$C_i=\sum_{k=1}^{K}\beta_k X_{ik}$。如果$\beta_k>0$，对于某个特定的观察对象第$k$个因素取值的增加会导致他的特征状况变好，反之$\beta_k<0$则会使得特征状况变得糟糕。但是在实际当，往往会出现D_i与其导致因素之间的关系并不精确，所以方程还包含了一个误差项ε_i。

但是在实际情况中，D_i的取值所表示的观察对象所代表的特征指数却很难观测，例如一个人对一件事物的喜欢程度具体是多少。但是我们能够观察到的是一个观察对象对于D_i的感受程度，例如一个人对于某一事物的喜欢程度可以分为“厌恶”、“不喜欢”、“喜欢”、“非常喜欢”四个程度。而变量Y_i可以以下列方式与D_i进行关联：

$$\begin{cases} Yi=1, \quad \text{如果} D_i\le\delta_1 \\ Yi=2, \quad \text{如果} \delta_1\le D_i\le\delta_2 \\ \cdots\cdots \quad \cdots\cdots \\ Yi=M, \quad \text{如果} D_i\ge\delta_{M-1} \end{cases}, \quad m=1,2,\cdots,M,(M>2) \qquad (3.1.2)$$

方程 4.2.2 中的$\delta 1,\delta 2,\cdots,\delta_{M-1}\ge 0$，是跟方程 3.1.1 中的$\beta_k$一起有待通过样本估计的未知参数（$\delta 1<\delta 2<\cdots<\delta_{M-1}$）。一个观察对象对于该观测特征的取值$Y_i$取决于其特征指数$D_i$是否跨过某个临界值，则$Y_i$取值为$1,2,\cdots,M$的概率分别为：

$$\begin{cases} P(Y_i=1)=P(C_i+\varepsilon_i\le\delta_1)=P(\varepsilon_i\le\delta_1-C_i) \\ P(Y_i=2)=P(\delta_1\le C_i+\varepsilon_i\le\delta_2)=P(\delta_1\text{-}C_i\le\varepsilon_i\le\delta_2-C_i) \\ \cdots\cdots \\ P(Y_i=M)=P(C_i+\varepsilon_i\ge\delta_{M-1})=P(\varepsilon_i\ge\delta_{M-1}-C_i) \end{cases} \qquad (3.1.3)$$

N个观察中的每一次观察都被视为多项分布中的一次单一的抽取，在这种情况下，这个多项分布就会出现M种不同的结果。那么观察到这整个样本的状况的概率，就是个体观察值的概率的乘积，假设$N_1,N_2,\cdots,N_M$为M种不同结果每种结果观察到的次数，且有$N_1+N_2+\cdots+N_M=N$，则：

$$\begin{aligned} L &= [P(Yi=1)]^{N_1}[P(Yi=2)]^{N_2}\cdots[P(Yi=M)]^{N_M} \\ &= [F(\delta_1 - C_i)]^{N_1}[F(\delta_2 - C_i) - F(\delta_1 - C_i)]^{N_2}\cdots[1 - F(\delta_{M-1} - C_i)]^{N_M} \end{aligned} \quad (3.1.4)$$

其中 $F(x) = P(\varepsilon_i < x)$ 是误差的累积概率分布。若 $F(x)$ 已知，我们可以通过最大似然估计，即通过观察样本发生的概率最大化来得到 $\beta_k, \delta_1, \delta_2, \cdots, \delta_{M-1}$ 的值。而在实际应用中，我们很难得到 ε_i 的具体分布，所以我们常常假设误差项 ε_i 服从于某项已知的概率分布。

在实际中，我们常用的累积分布有两种：累积 Logistic 回归和累积 probit 回归, 它们的差别在于 :累积 logistic 回归是假定 ε_i 具逻辑分布的特点, 而累积 probit 回归则假定 ε_i 服从正态分布的结果。而逻辑分布除了在尾部比正态分布大很多之外，其余与正态分布非常相似。而在实际问题研究中，常常使用正态分布来对 ε_i 的分布进行假设。

$\hat{\beta}_k$ 为系数 β_k 的估计值，$\hat{C}_i = \sum_{k=1}^{K} \hat{\beta}_k X_{ik}$ 。使用 $\hat{C}_i$ 以及临界值 $\delta_1, \delta_2, \cdots, \delta_{M-1}$ 的估计值 $\hat{\delta}_1, \hat{\delta}_2, \cdots, \hat{\delta}_{M-1}$，可以对样本中每个个体处于不同剥夺程度的概率进行估计。这些估计可分别用 $P_{i1}, P_{i2}, \cdots, P_{iM}$，计算如下：

$$\begin{cases} P(Y_i = 1) = P(\varepsilon_i \le \delta_1 - C_i) = F(\delta_1 - C_i) \\ P(Y_i = 2) = P(\delta_1 - C_i \le \varepsilon_i \le \delta_2 - C_i) = F(\delta_2 - C_i) - F(\delta_1 - C_i) \\ \cdots\cdots\cdots \\ P(Y_i = M) = P(\varepsilon_i \ge \delta_{M-1} - C_i) = 1 - F(\delta_{M-1} - C_i) \end{cases} \quad (3.1.5)$$

其中 $\sum_{j=1}^{M} \hat{p}_{ij} = 1$。

对于 SAS 等常用软件，累积回归并不明显包含截距项，即方程 3.1.1 中包含的 $\beta_k (k = 1, 2, \cdots, K)$ 全部都是斜率，而截距项包含在临界值中。

（二）累积分布函数

标准正态变量（SNV）X 的累积分布函数为：

$$P(X<x)=\Phi(x)=\int_{-\infty}^{x}\frac{1}{2\pi}\exp(\frac{-X^2}{2})dx \tag{3.1.6}$$

则：

$$\begin{cases} P(Y_i=1)=\Phi(\delta_1-C_i) \\ P(Y_i=2)=\Phi(\delta_2\text{-}C_i)-\Phi(\delta_1-C_i) \\ \cdots\cdots \\ P(Y_i=M)=1-\Phi(\delta_{M-1}-C_i) \end{cases} \tag{3.1.7}$$

则对 $\beta_k,\delta_1,\delta_2,\cdots,\delta_{M-1}$ 的估计通过使用正态分布函数 $\Phi(x)$ 取代 $F(x)$ 然后据似然函数获得。

（三）连续变量的边际效应

当方程中一个自变量为连续变量，而且当该变量值改变的时候，出现不同结果的概率的变化程度，称为连续变量的边际效应。当为正态分布的情况下，X_{ik}（第 i 个观测样本的第 k 个决定变量的值）的边际效应为：

$$\begin{cases} \dfrac{\partial P(Y_i=1)}{\partial X_{ik}}=\dfrac{d}{dC_i}[\Phi(\delta_1-C_i)]\dfrac{\partial C_i}{\partial X_{ik}}=-\Phi'(\delta_1-C_i)\beta_k \\ \dfrac{\partial P(Y_i=2)}{\partial X_{ik}}=\dfrac{d}{dC_i}[\Phi(\delta_2-C_i)-\Phi(\delta_1-C_i)]\dfrac{\partial C_i}{\partial X_{ik}}=[\Phi'(\delta_2-C_i)-\Phi'(\delta_1-C_i)]\beta_k \\ \cdots\cdots \\ \dfrac{\partial P(Y_i=M)}{\partial X_{ik}}=\dfrac{d}{dC_i}[1-\Phi(\delta_{M-1}-C_i)]\dfrac{\partial C_i}{\partial X_{ik}}=\Phi'(\delta_{M-1}-C_i)\beta_k \end{cases} \tag{3.1.8}$$

其中，$\Phi'(x)=\dfrac{d\Phi(x)}{dx}$。

（四）平行斜率假说

累积 logistic 和累积 probit 模型的一个关键假设是方程 4.2.1 的斜率系数 β_k 并

不随着所关注的特征Y_i的变化，即在 probit 模型的情况下：

$$\begin{cases} p_1 = \alpha_1 + \sum_{k=1}^{K} \beta_k X_{ik} \\ p_1 + p_2 = \alpha_2 + \sum_{k=1}^{K} \beta_k X_{ik} \\ \cdots \\ p_1 + p_2 + \cdots + p_{M-1} = \alpha_{M-1} + \sum_{k=1}^{K} \beta_k X_{ik} \\ p_1 + p_2 + \cdots + p_M = 1 \end{cases} \tag{3.1.9}$$

而为检验该假设，我们常常需对数据进行多类别 probit 模型估计，多类别 probit 模型允许斜率系数β_k在不同的结果$i = 1, 2, \cdots, M$之间不同。当累积 probit 模型估计K个系数时，多类别 probit 模型估计$K(M-1)$个系数，如果从累积 probit 模型中计算所得的概率为P_1，而从多类别模型中计算得概率为P_2，计算$2(P_2 - P_1)$，并与$\chi^2(K(M-1))$比较。当然，这并不是严格的似然比检验，因为累积 probit 模型并不嵌套在多类别 probit 模型之中[7]。因此这个检验只是提示性的：一个非常大的χ^2值会是引起担心的依据，一个中等大的则不是。

第二节 利用累积 probit 模型对量化交易策略进行判断

在战略决策制定过程中，累积 logit 模型和累积 probit 模型具有广泛的应用价值。对于决策分布服从正态分布或者逻辑分布时常会用到这两种方法。而在实际生活中，正态分布较逻辑分布更为常见，所以本文尝试用累积 probit 模型对数据进行拟合分析。为了便于分析，我们首先对数据进行标准化处理。

首先，我们对数据的数据及滞后 1—5 期数据进行拟合，我们首先需要对模型的等斜率假设以及模型β系数为 0 进行检验，其检验结果如下表：

表 3.1 β系数 0 假设检验与等斜率检验

	检验全局零假设: BETA=0			等斜率假设的评分检验		
检验	卡方	自由度	Pr > 卡方	卡方	自由度	Pr > 卡方
似然比	331.8204	48	<.0001	76.7080	48	0.00529

评分	328.7806	48	<.0001
Wald	310.9066	48	<.0001

由检验全局零假设即 β 系数 0 假设检验结果可得，其似然比 χ^2 检验值为 331.8204，其对应的 p 值小于 0.0001，所以拒绝原假设，即表明该模型的 β 系数并不全都为 0。而根据其等斜率假设的 χ^2 检验值为 76.7080，其对应的 p 值为 0.00529，根据 χ^2 检验只是提示性的：一个非常大的 χ^2 值会是引起担心的依据，一个中等大的则不是，我们可以认为该模型符合等斜率假设，可以对模型进行进一步的估计。

基于以上假设检验，我们对数据的决策时期以及其滞后 1 至 5 期的数据进行拟合，其拟合结果如表 3.2：

表 3.2 累积 probit 模型拟合结果

最大似然估计值分析						
参数		估计值	Pr > 卡方	参数	估计值	Pr > 卡方
Intercept	-1	-1.0832	<.0001	priceT3	-22.6404	0.262
Intercept	0	0.8734	<.0001	turnoverT3	-0.0133	0.3732
priceT6		84.9704	<.0001	buyT3	-0.0254	0.804
turnoverT6		-0.0537	0.0002	sellT3	-0.1562	0.1964
buyT6		0.1753	0.0173	priIncT3	0.0761	0.3168
sellT6		-0.3784	<.0001	turnIncT3	-0.00197	0.8732
priIncT6		-0.4581	<.0001	buyIncT3	0.0136	0.3119
turnIncT6		0.00634	0.6092	sellIncT3	-0.00366	0.7734
buyIncT6		-0.0053	0.6895	priceT2	24.9277	0.2116
sellIncT6		0.015	0.2389	turnoverT2	-0.0138	0.3532
priceT5		-35.1849	0.085	buyT2	0.0111	0.9134
turnoverT5		0.00372	0.8035	sellT2	0.1462	0.2263
buyT5		-0.3045	0.0032	priIncT2	-0.047	0.5212
sellT5		0.7019	<.0001	turnIncT2	0.000568	0.9633
priIncT5		-0.2586	0.0006	buyIncT2	-0.0196	0.1461
turnIncT5		-0.0122	0.3438	sellIncT2	-0.0255	0.0552
buyIncT5		0.00608	0.6474	priceT1	-6.6127	0.6785
sellIncT5		0.0176	0.1727	turnoverT1	-0.0234	0.1023
priceT4		-45.461	0.0445	buyT1	0.1028	0.1603

turnoverT4		-0.049	0.0412	sellT1	-0.2082	0.0664
buyT4		0.0601	0.5582	priIncT1	-0.0085	0.502
sellT4		-0.1725	0.1568	turnIncT1	-0.0003	0.9803
priIncT4		-0.0442	0.5607	buyIncT1	-0.0148	0.2189
turnIncT4		0.0154	0.2189	sellIncT1	0.0101	0.4005
buyIncT4		-0.0248	0.0648			
sellIncT4		0.00822	0.517			

由表 3.2 可以看出，量化交易策略同期中（T6 时期）对于交易策略有较大的影响，其中策略同期中（$\alpha = 0.05$）priceT6、turnoverT6、buyT6、sellT6 和 priIncT6 对于策略有显著的影响，而滞后一期中仅有 buyT5、sellT5 和 priIncT5 三个变量对量化交易策略产生显著影响，而滞后 2 期以上则不会对量化交易策略的实施产生显著影响。这也就说明，高频更接近与一种即期交易策略，其往往更注重即期交易数据对策略本身产生的影响，而对于交易期以前的交易数据其影响往往并不明显。

为了更好地对量化交易策略进行分析，本文将对量化交易策略产生非显著影响的变量剔除之后，对模型再次进行拟合，可以得到以下分析结果：

表 3.2 剔除非显著影响变量后累积 probit 模型拟合结果

最大似然估计值分析						
参数		自由度	估计值	标准误差	Wald 卡方	Pr > 卡方
Intercept	-1	1	-1.0751	0.0168	4076.015	<.0001
Intercept	0	1	0.8629	0.0156	3060.363	<.0001
priceT6		1	0.0046	0.0211	0.0475	0.0274
turnoverT6		1	-0.0833	0.0128	42.1226	<.0001
buyT6		1	0.2048	0.0654	9.8081	0.0017
sellT6		1	-0.4501	0.0824	29.8308	<.0001
priIncT6		1	-0.0745	0.0153	23.7524	<.0001
buyT5		1	-0.2005	0.0653	9.4251	0.0021
sellT5		1	0.4033	0.0824	23.9662	<.0001
priIncT5		1	-0.0239	0.012	3.943	0.0471

根据以上分析结果，可得累积 probit 模型：

$$\begin{cases} \delta_1 = -1.0715 + 0.0046 priceT6 - 0.0833 turnoverT6 + 0.2048 buyT6 - 0.4501 sellT6 \\ \quad -0.0745 priIncT6 - 0.2005 buyT5 + 0.4033 sellT5 - 0.0239 priIncT5 \\ \delta_2 = 0.08629 + 0.0046 priceT6 - 0.0833 turnoverT6 + 0.2048 buyT6 - 0.4501 sellT6 \\ \quad -0.0745 priIncT6 - 0.2005 buyT5 + 0.4033 sellT5 - 0.0239 priIncT5 \\ P(strategyT6 = -1) = \Phi(\delta_1) \\ P(strategyT6 = 0) = \Phi(\delta_2) - \Phi(\delta_1) \\ P(strategyT6 = 1) = 1 - P(strategyT6 = -1) - P(strategyT6 = 0) \end{cases}$$

将数据上式可得各决策的概率和累积概率，并依此对决策的选择进行判断。

由以上模型分析结果我们可得以下结论：

（1）量化交易策略偏向于与市场价格变化做同向交易，即当股票的市场价格上涨时，量化交易策略更倾向于进行买入交易，当股票价格下跌时，量化交易策略更倾向于进行卖出交易，即追涨杀跌；

（2）在高成交量的活跃市场氛围下，量化交易策略更倾向于进行买入交易，而在成交量比较低迷的情况下量化交易策略更倾向于进行卖出策略；

（3）买入挂单和卖出挂单对量化交易策略交替产生影响，本期高卖出挂单量趋向于使量化交易策略进行卖出交易而使得下一期量化交易策略趋向于进行买入交易，而高买入挂单量对量化交易策略有着相反的影响，即高买入挂单量趋向于使量化交易策略进行买入交易而使得下一期量化交易策略趋向于进行卖出交易；

（4）priIncT6 和 priIncT5 具有相同的符号，说明量化交易策略的实施对于股价走势的具有较为严格的要求，只有在股价走势相对稳定时更趋向于进行量化交易。

为了对模型的实际决策效果进行检验，我们预测集 601166 对模型的拟合效果进行检验。将预测集数据代入公式，并将预测结果与实际决策值进行比较，可得以下对比结果：

表 3.3 模型预测结果对比

预测概率和观测响应的关联			
一致部分所占百分比	63.6	Somers D	0.281
不一致部分所占百分比	35.5	Gamma	0.284

结值百分比	0.9	Tau-a	0.131

由以上对比结果,我们可以看出,该模型对预测集601166的误判率为35.5%,正确率为63.6%，有0.9%的交易决策无法进行判断，说明该模型拟合效果较好。

第四章 量化交易策略的支持向量机与贝叶斯网络方法研究

第一节 方法论——贝叶斯网络

（一）贝叶斯网络

贝叶斯网络在处理人工智能的不确定性问题问题上具有非常大的优势。贝叶斯网络是与概率统计相关联并将其应用于较为复杂领域从而对不确定性进行推理和数据分析的工具，是一种能够系统地描述随机变量之间关系的有效工具。贝叶斯网络的建立主要是进行事件的概率推理。用概率论处理不确定性的主要优点是保证推理结果的正确性。

一般来说，通过使用在有向无环图来表示随机变量节点的贝叶斯网络，这些随机变量可以是随机变量可以被观察到，或隐藏变量、未知参数。通过两个节点之间连接的箭头来表示这两个随机变量是是非条件独立或者具有因果关系的；若两个节点变量之间没有箭头相互连接则称该随机变量彼此间为条件独立。若两个节点之间通过一个单箭头将其连接在一起，表示其中一个节点是“父节点（因）”，我们以Y_i表示，另一个是“子节点（果）”，我们以X_i表示，N_i表示第i个节点，两节点就会产生一个条件概率值。

这里提到了有向无环图的概念，我们就不得不解释一下什么是有向无环图。有向无环图是指：如果一个有向图无法从某个顶点出发经过若干条边回到该点，

则这个图是一个有向无环图（DAG 图）。如图 4.1

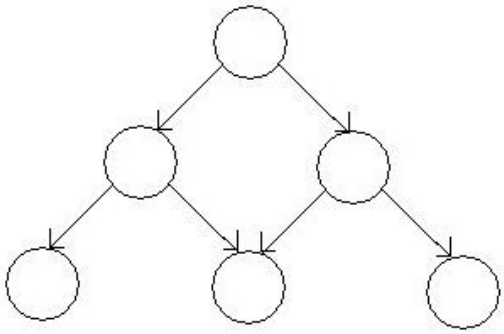

图 4.1 有向无环图

贝叶斯网络适用在节点的性质是属于离散型的情况下，当然对于非离散型变量，可以通过对变量取值范围进行分组划分，使其变为离散型随机变量，而后依照 $P(N_i \mid Y_i)$ 写出条件概率表, 其中 Y_i 表示每一行中各事件发生的概率, 而 X_i 表示每一列中各事件发生的概率，且每一行的概率必为 1。

（二）数学定义

提供了一种把联合概率分布分解为局部概率的方法[8]是贝叶斯网络的一个关键特征。令 $G=(I,E)$ 表示一个有向无环图，其中，I 为有向无环图中节点的集合，表示领域变量，E 为有向连接线段（即箭头）的集合，表示变量间的概率依赖关系，同时每个节点对应着一个条件概率分布表，表明该子节点与父节点之间存在的依赖关系。令 $X=(X_i)_{i\in I}$ 为其有向无环图中某一节点 i 所代表的随机变量，n_i 为其对应的取值，则有：

$$p(x)=\prod_{i=I} p(x_i \mid x_{pa(i)}) \qquad (4.1.1)$$

则称 X 为一个相对于一个有向无环图 G 的贝叶斯网络。$pa(i)$ 表示每个节点之“因”。

（三）贝叶斯网络的参数学习

贝叶斯网络的参数学习本质上是在其结构已知的情况下，并通过训练集来机器学习贝叶斯网络的每个节点上的概率分布表，而通过这种方法相比与早期通过专家的自身知识和经验来制定概率分布表，则具有非常高的适应性。而通过对数据的观测，可以将数据划分为完备数据集和不完备数据集。对于贝叶斯网络结构已知的不完备数据集，一般采用如 Gaussian 逼近、Monte-Carlo 方法，以及 EM（期望-极大化）算法求极大似然或极大后验来进行结构学习；而对于贝叶斯结构已知的完备数据集，通常使用最大似然估计来进行参数的训练与学习。（如表 4.1）

表 4.1 对于不同贝叶斯网络及数据集的参数估计方法

结构	观测值	方法
已知	完整	最大似然估计法 （MLE）
已知	部份	EM 算法 Monte-Carlo 方法 Gaussian 逼近
未知	完整	搜索整个模型空间
未知	部份	结构算法 EM 算法 Bound contraction

本文所用的数据集为完备数据集且贝叶斯网络结构已知，故本文在本节主要进行贝叶斯结构已知的完备数据集的参数学习的介绍。

对于贝叶斯结构已知的完备数据集，我们采取最大似然估计（MLE）来进行参数的学习和估计。在本文的上一节中，对累积 probit 模型进行参数估计时，同样使用的是最大似然估计法。

最大似然估计，也称为最大概似估计，是一种统计方法，它用来求一个样本集的相关概率密度函数的参数。这个方法最早是遗传学家以及统计学家罗纳德·费雪爵士在 1912 年至 1922 年间开始使用的。使用该方法对参数进行估计，需满足以下条件：

（1）数据集是完备数据集；

（2）各样本之间服从独立同分布。

此时我们可以用最大似然估计法（MLE）来求得参数。其对数最大似然函数为：

$$L = \frac{1}{N}\sum_{i=1}^{n}\sum_{i=1}^{s}\log(P(X_i \mid pa(X_i), D_i)) \qquad (4.1.2)$$

其中 $pa(X_i)$ 代表 X_i 的因变量，D_i 代表第 i 个观测值，N 代表观测数据的总数。由上式就可以借由观测值来估计出节点 U 的条件分配。如果当模型很复杂时，这时可能就要利用数值分析或其它最优化技巧来求出参数。

由统计学原理我们可知，利用 MLE 对参数进行估计有以下优势：

（1）一致性：随着观测个数即样本量的增加，参数最终收敛于其最佳可能值；

（2）渐进有效性：MLE 在于寻找使样本发生的可能性最大的参数 β 的估计值 $\hat{\beta}$，而所观测的样本数越多，$\hat{\beta}$ 越接近于其实际值；

（3）参数的不同分布形式不会对估计出的概率分布效果产生影响。

第二节　方法论——支持向量机

（一）支持向量机方法论

支持向量机属于一般化线性分类器，被认为是提克洛夫规范化（Tikhonov Regularization）方法的一个特例。这族分类器的特点是他们能够同时最小化经验误差与最大化几何边缘区，因此支持向量机也被称为最大边缘区分类器。支持向量机将向量从原始模式空间经过一个特定的函数的非线性变换映射到一个更高维的空间里，将一个低维非线性问题转化为某一个高维的线性问题，在这个空间里有一个最大间隔超平面，也称为最优分类超平面。在分开数据的超平面的两边建有两个互相平行的超平面，分隔超平面使两个平行超平面的距离最大化。假定平行超平面间的距离或差距越大，分类器的总误差越小。

（二）数学表示及求解

假设超平面的数学形式可以表示为：

$$W \cdot X - b = 0 \tag{4.2.1}$$

其中：X 是超平面上的点，W 是垂直于超平面的向量。

由于要求最大距离间隔，因此我们要知道支持向量以及与最佳超平面平行的并且离支持向量最近的超平面。我们可以看到这些平行超平面可以由方程族：

$$\begin{cases} W \cdot X - b = C \\ W \cdot X - b = -C \end{cases} \tag{4.2.2}$$

来表示，若该将W 和 X 视为二维向量，可以证明$C=1$，故可将平行超平面方程族写为：

$$\begin{cases} W \cdot X - b = 1 \\ W \cdot X - b = -1 \end{cases} \tag{4.2.3}$$

如果这些训练数据是线性可分的，那就可以找到这样两个超平面，在它们之间没有任何样本点并且这两个超平面之间的距离也最大。同样利用二维向量及两条平行线之间的而距离公式，我们很容易得出，两超平面的距离为$\frac{2}{|W|}$，则在二维超平面中，此超平面可表示为图 4.2。

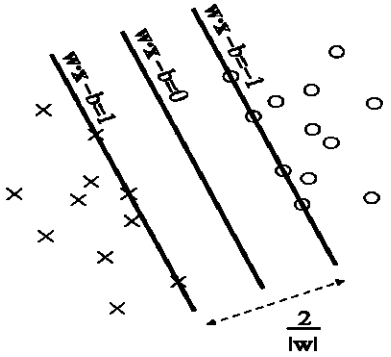

图 4.2 支持向量机的图形表示

由此，我们可以将超平面的求解转化为最小化$\|W\|^2$。为了将样本数据点全部划分到超平面间隔区域的外部，我们需将建立样本集约束条件：

$$\begin{cases} W \cdot x_i - b \geq 1 \\ W \cdot x_i - b \leq -1 \end{cases} \tag{4.2.4}$$

可以将 4.4.4 式写为：

$$y_i(\langle W, x_i \rangle - b) \geq 1, \quad i = 1, \cdots, n \tag{4.2.5}$$

至此，该超平面的求解问题转化成为分类间隔 $\frac{2}{\|W\|}$ 最大化即 $\|W\|^2$ 最小问题的求解。而使分类间隔最大实际上是对学习机推广能力的控制，这也是 SVM 的核心思想之一。统计学理论指出，在 N 维空间当中，假设样本分布在一个半径为 R 的超球范围内，类似于二维平面中的半径为 R 的圆形平面和三维空间中半径为 R 的球体，则满足条件 $\|W\|^2 \leq A$ 的正则超平面构成的指示函数集：

$$f(x, W, b) = \operatorname{sgn}(\langle W, x_i \rangle - b) \tag{4.2.6}$$

则满足 4.4.6 数据集的 VC 维满足一下的界：

$$h = \min([R^2 A^2], N) + 1 \tag{4.2.7}$$

因此，使 $\|W\|^2$ 最小就转化为使 VC 维的上界最小，从而实现结构风险最小化（Structure Risk Minimization，SRM）准则中对函数复杂性的选择。

则在线性可分的情况下，在 SRM 准则下的最优超平问题，就可以表示为如下的约束优化问题：

$$\begin{cases} \min \dfrac{1}{2}\|W\|^2 \\ s.t. \quad y_i(\langle W, x_i \rangle - b) \geq 1, \quad i = 1, \cdots, n \end{cases} \tag{4.2.8}$$

4.2.8 式的最优解可以通过求解拉格朗日函数的鞍点得到，定义如下拉格朗日函数：

$$L(W, b, \alpha) = \frac{1}{2}\|W\|^2 - \sum_{i=1}^{n} \alpha_i y_i(\langle W, x_i \rangle - b) \tag{4.2.9}$$

其中，$\alpha_i \geq 0$ 是个样本所对应的朗格朗日系数。

求解式 4.2.9 的最小值，则令该泛函对 W 和 b 求偏导，并令它们等于 0，就可以将上述求最优分类面的问题转化为较简单的对偶问题，其对偶问题由如下形式给出：

$$\begin{cases} Q(\alpha)=\sum_{k=1}^{n}\alpha_k-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle \\ s.t.\quad \sum_{i=1}^{n}\alpha_i y_i=0,\alpha_i\geq 0,\quad i=1,\cdots,n \end{cases} \tag{4.2.10}$$

这是一个不等式约束下的二次函数寻优问题，存在唯一解。以上最优化问题的最优解为:$\alpha^*=(\alpha_1^*,\cdots,\alpha_n^*)^T$，经过计算可得：

$$\begin{cases} W^*=\sum_{i=1}^{n}\alpha_i y_i x_i \\ b^*=-\frac{1}{2}\langle W^*,x_r+x_s\rangle \end{cases} \tag{4.2.11}$$

其中，x_r 和 x_s 是两类中任意的支持向量，从而可构造判别函数：

$$f(x)=\mathrm{sgn}(\langle W^*,x_i\rangle-b^*)=\mathrm{sgn}(\sum_{x\in SV}\alpha_i^* y_i\langle x\cdot x_i\rangle-b^*) \tag{4.2.12}$$

以上是在处理线性分类数据时的分类面的求解，而在处理非线性分类问题时，仅仅比线性情况多了一个非线性映射环节。现假定非线性映射为 $x\to\varphi(x)$，这时对偶形式的目标函数变为：

$$Q(\alpha)=\sum_{k=1}^{n}\alpha_k-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\langle\varphi(x_i),\varphi(x_j)\rangle \tag{4.2.13}$$

由于对偶形式中只出现两向量的内积运算，Vapnik 等人提出了满足 Mercer 条件的核函数 $K(x_i,x_j)$ 来代替内积运算，即 $K(x_i,x_j)=\langle\varphi(x_i),\varphi(x_j)\rangle$ 实现非线性软间隔分类。则其判别函数为：

$$f(x)=\mathrm{sgn}(\langle W^*,x_i\rangle-b^*)=\mathrm{sgn}(\sum_{x\in SV}\alpha_i^* y_i K(x_i,x_j)-b^*) \tag{4.2.14}$$

第三节　利用支持向量机和贝叶斯网络进行案例分析

（一）利用贝叶斯网络和支持向量机对量化交易策略进行判断

贝叶斯网络和支持向量机都是数据挖掘中常用的技术手段，在股票预测中具有广泛的应用。本文将利用以上这两种数据挖掘方法对量化交易策略进行分析，并将其与上面所使用的累积 probit 模型进行对比。该节数据挖掘技术所用软件为 clementine12.0。

贝叶斯网络

贝叶斯网络又称信度网络，是 Bayes 方法的扩展，目前不确定知识表达和推理领域最有效的理论模型之一。从 1988 年由 Pearl 提出后，已经成为近几年来研究的热点。利用贝叶斯网络技术，有以下特点：

1、贝叶斯网络本身是一种非定性因果关联模型。贝叶斯网络与其他决策模型不同之处在于，其本身是一种将多元知识图解为可视化的概率知识表达与推理模型，更为详细地蕴含了网络各节点变量之间的因果关系及条件相关关系。

2、贝叶斯网络具有强大的非确定性问题处理能力。贝叶斯网络利用条件概率来表达各个变量之间的相关关系，能在有限的、不完整的、非确定的信息条件下进行学习和推理。

3、贝叶斯网络能够有效进行多源信息的表达与融合。贝叶斯网络可将故障诊断与维修决策相关的各种信息纳入到网络结构中，按节点的方式统一进行处理，能有效地按信息的相关关系进行融合。

以下将通过贝叶斯网络的方法对数据进行分析。

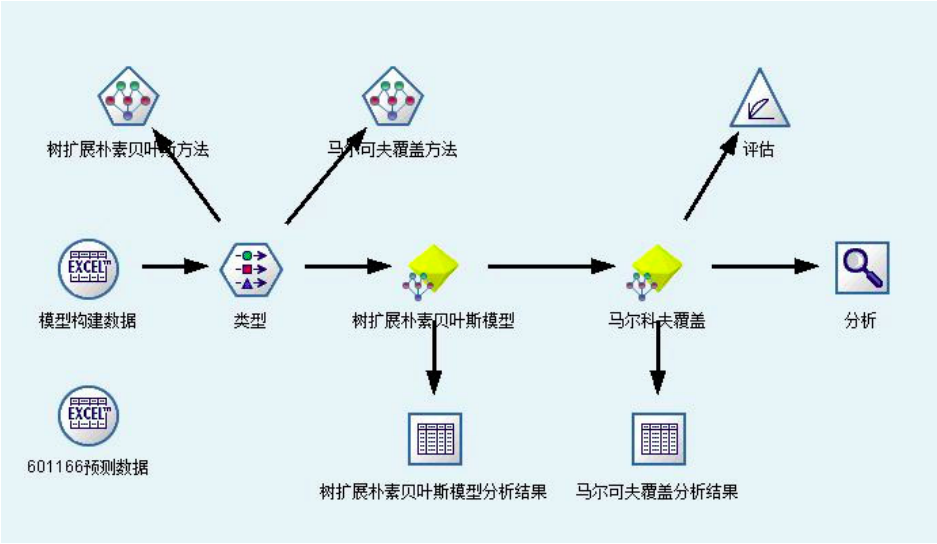

图 4.3 贝叶斯网络算法流程图

该方法首先利用模型构建数据，即训练集股票样本数据对模型进行训练，然后利用经训练集数据对预测数据进行分析比较。该方法中利用两种分析方法对模型进行构建，两种方法分别为：

树扩展朴素贝叶斯模型（TAN）：通过该方法可创建简单的贝叶斯网络模型，是一种基于标准朴素贝叶斯模型的改进模型。由于该模型允许每一个自变量除了依赖于目标变量之外，还依赖于其他自变量，因此该方法增加分类的准确度。

马尔可夫覆盖：该方法可以在包含目标变量的父项、子项及其子项的父项的数据集中选择节点集。马尔可夫覆盖基本上标识了需要预测目标变量的网络中的所有变量。用户认为这种构建网络的方法更为准确；但是，当处理大型数据集时，由于所包含的变量数较多，所以可能会消耗许多处理时间。

首先利用训练数据集对模型进行训练，其中图 5.1.1 中间部分树扩展朴素贝叶斯模型和马尔科夫覆盖模型分别由训练集通过树扩展朴素贝叶斯模型和马尔科夫覆盖方法训练而成，原始训练集的训练结果如下表 5.10：

表 4.2 训练集训练模型比较

方法	误判率
树扩展朴素贝叶斯模型	33.24%
马尔科夫覆盖	33.28%

由以上训练结果可以看出，对于训练集的预测，两种方法预测的误判率分

别为 33.24%和 33.28%，两种方法之间的差异并不明显，且对于数据有很好的预测判别，说明两种方法都能够进行量化交易策略的判别。

然后对预测集进行分析，将预测集链接入算法流程中，并利用已训练完成的模型对预测集数据进行分析比较，其两种方法分析结果的如表 4.3：

表 4.3 预测集分析比较结果

方法	误判率
树扩展朴素贝叶斯模型	29.43%
马尔科夫覆盖	29.36%

对于预测集的分析，我们可以看出通过树扩展朴素贝叶斯模型和马尔科夫覆盖这两种方法的预测结果的误判率均为 29.4%左右，说明这两种方法对于量化交易策略能够进行有效的分析，且分析准确率无明显差别。

支持向量机

支持向量机方法的基础理论为统计学习理论中的 VC 维理论和结构风险最小原理，其根据有限的样本信息在模型的复杂性（即对特定训练样本的学习精度）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳均衡点，以求获得最好的推广能力。

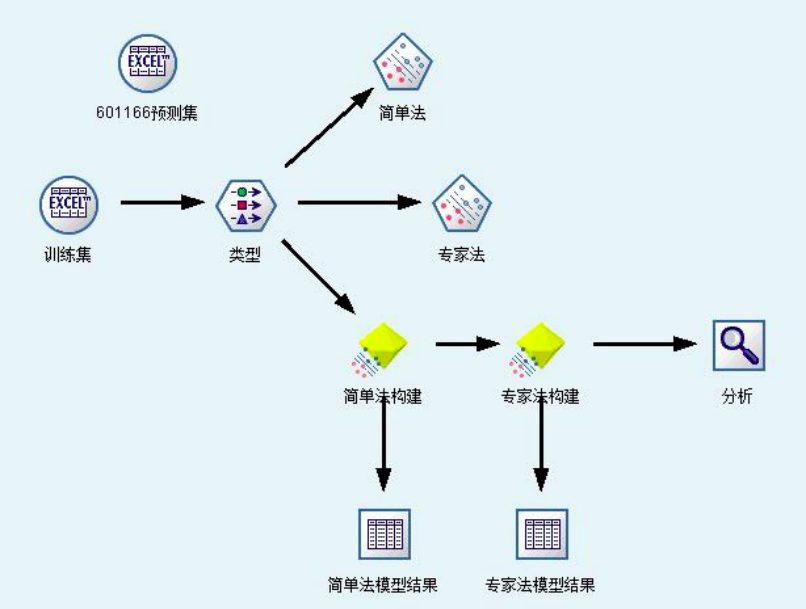

图 4.4 支持向量机算法流程图

通过训练集数据，对模型进行简单训练和专家训练，并对训练模型进行检验，其检验结果如表 4.4：

表 4.4 训练集训练模型比较

方法	误判率
SVM 简单法	17.1%
SVM 专家法	15.66%

由以上训练结果可以看出，对于训练集的预测，两种方法预测的误判率分别为 17.1%和 15.66%，两种方法之间的差异并不明显，且对于数据有很好的预测判别，说明两种方法都能够进行量化交易策略的判别。

然后对预测集进行分析，将预测集链接入算法流程中，并利用已训练完成的模型对预测集数据进行分析比较，其两种方法分析结果的如表 4.5：

表 4.5 预测集分析比较结果

方法	误判率
SVM 简单法	13.88%
SVM 专家法	13.81%

对于预测集的分析，我们可以看出通过简单法和专家法这两种方法的预测结果的误判率均为 13.8%左右，说明这两种方法对于量化交易策略能够进行有效的分析，且分析准确率无明显差别。

（二） 三种模型之间的比较

以上三种模型：累积 probit 模型、贝叶斯网络模型和支持向量机都能对量化交易策略的进行预测与决策，但是在模型预测过程中各有优势：

（1）在预测精度方面

通过以上三种方法对预测集的量化交易策略进行预测分析，我们可得以下模型检验比较结果：

表 4.6 三种模型预测效果比较

方法		误判率
累积 probit 模型		36.4%
贝叶斯网络模型	TAN 方法	29.43%
	马尔科夫覆盖	29.36%
支持向量机	简单法	13.88%
	专家法	13.81%

对比三种模型的误判率，我们可以看出：三种方法对于量化交易策略的预测具有明显差别：支持向量机方法较其他两种方法在预测中具有更高的准确率，更低的误判率，其误判率在 13.8%左右；而相比较而言，累积 probit 模型的误判率相对较高，为 36.4%。由此我们可以看出，从模型预测精度来看：

支持向量机＞贝叶斯网络＞累积probit模型

（2）在模型的可解释性方面

三种方法都是通过构建训练集模型来达到对预测集进行分析的目的，但是在对于模型的构建上具有明显的差异 :累积 probit 模型通过构建概率模型来对因变量进行预测，其建立的是具体的等式模型；贝叶斯网络模型是基于每个节点的概率分布表来进行分析判别，其利用最大似然估计来对模型的参数进行学习与估计；支持向量机则是利用从原始模式空间经过一个特定的函数的非线性变换映射到一个更高维的空间里，将一个低维非线性问题转化为某一个高维的线性问题，从而建立判别函数的方法对数据进行判别。通过其模型的构建方式，我们就可以看出，在对变量的解释性方面，累积 probit 模型对因变量具有更好更直观的解释性，我们可以通过其各自变量的参数对因变量进行解释；而支持向量机方法对于因变量的可解释性要比其他两种方式稍弱，其建立的仅仅是一种判别模型。所以在对变量的可解释性上：

累积probit模型＞贝叶斯网络＞支持向量机

综上对比三种量化交易策略分析方法，可以发现三种方法在策略分析上各有优势，无论是预测精度，还是影响因素的解释能力，在实际分析应用当中我们都经常会用到，所以在实际操作中，我们需将这三种方法加以结合，来对量

化交易策略进行综合分析比较。

第五章 量化交易策略的时间序列价格预测研究

第一节 方法论

（一）协整检验

如果时间序列的线性组合的单整阶数小于其成分的单整阶数，就称这些序列间存在协整，则那些线性组合系数称为协整向[9]。一个时间序列变量，仅在他们协整的时候，才具有自变量与因变量研究的价值。向量协整则意味着向量各分量之间存在长期关系。

协整检验的过程（Johansen 方法）

假设多元时间序列 $X_t=(x_{t1},\cdots\cdots x_{tk})^T, t=0,\pm1,\pm2,\cdots$，带有 D_t 的向量自回归模型 VAR（p）为：

$$X_t=\alpha+\sum_{i=1}^{p}\Phi(X_{t-j}+\gamma D_t+\mu_t),\qquad t=p+1,\cdots,n \tag{5.1.1}$$

假设有算子符号：

$$\Gamma_i=-(1-\Phi_1-\cdots-\Phi_i)\quad i=1,\cdots\cdots,p-1\ ;\ \Pi=-(1-\Phi_1-\cdots-\Phi_p) \tag{5.1.2}$$

则 VAR（p）可以重新表示成：

$$\Delta X_t=\Gamma_1\Delta X_{t-1}+\cdots+\Gamma_{p-1}\Delta X_{t-p-1}+\Pi X_{t-p}+\gamma D_t+\alpha+\mu_t \tag{5.1.3}$$

上式称为长期向量误差修正模型，其中矩阵 Π 包含了变量累积的长期影响，

它包含了序列之间的长期稳定的关系, 若它的秩为 0, 则该系统就不是协整的(假定所有在 X_t 中的变量至少为单整 $I(1)$ 的。若 Π 为满秩，则 X_t 的变量为平稳的。如果矩阵 Π 有秩 r，且 $0<r<k$，则可将 Π 分解为两个不同的（$k\times r$）矩阵 a 和 b，使 $\Pi=ab^T$，即 b 中包含了 r 个协整变量。b 的每一列为协整向量意味着 $b^TX_t\sim I(0)$。

为达到简化目的，令 $Z_{0t}=\Delta X_t, Z_{1t}=(\Delta X_{t-1}^T,\cdots,\Delta X_{t-p-1}^T,D_t,1)^T, Z_{pt}=X_{t-p}$，令 Γ 含参数 $(\Gamma_1,\Gamma_2,\cdots,\Gamma_{p-1},\gamma,\alpha)$，模型就变为：

$$Z_{0t}=\Gamma Z_{1t}+ab^TZ_{pt}+\mu_t \tag{5.1.4}$$

令：

$$H_{ij}=\frac{1}{n}\sum_{t=1}^{n}Z_{it}Z_{jt}\qquad (i,j=0,1,\cdots,p) \tag{5.1.5}$$

记：

$$F_{ij}=H_{ij}-H_{i1}H_{11}^{-1}H_{ij}\qquad (i,j=0,1,\cdots,p) \tag{5.1.6}$$

设 $\hat{\lambda}_1,\hat{\lambda}_2,\cdots,\hat{\lambda}_k(\hat{\lambda}_1>\hat{\lambda}_2>\cdots>\hat{\lambda}_k>0)$ 为方程

$$\left|\lambda F_{kk}-F_{k0}F_{00}^{-1}F_{0k}\right|=0 \tag{5.1.7}$$

的特征值。

则，特征值检验的原假设和备择假设为：

$$H_0:Tr(\Pi)=\gamma \Leftrightarrow H_1:Tr(\Pi)=\gamma+1 \tag{5.1.8}$$

检验统计量为：

$$\lambda_{\max}(\gamma,\gamma+1)=-n\ln(1-\hat{\lambda}_{\gamma+1}) \tag{5.1.9}$$

迹检验:

当增加比γ个更多的λ_i，$Tr(\Pi)$是否会增加，$\lambda_i=0$等价于相应的迹统计量为0。这就导致了随γ递增的约束模型的似然比检验，要做的检验为：

$$H_0:Tr(\Pi)\leq\gamma \Leftrightarrow H_1:Tr(\Pi)>\gamma \tag{5.1.10}$$

检验统计量为：

$$\lambda_{Tr}(\gamma)=-n\sum_{i=\gamma+1}^{k}\ln(1-\hat{\lambda}_{\gamma+1}) \tag{5.1.11}$$

综上，Johansen方法包含以下步骤：

（1）确保时间序列向量的单整水平至少为$I(1)$；

（2）选择滞后阶数；

（3）选择系统的确定的成分；

（4）确定协整向量数目。

（二）VARX模型拟合

VAR模型（Vector Autoregressive Model）最早由Sims在1980年提出。该模型利用多方程联立的形式，其并不以经济理论为基础，在模型中的每个方程，内生变量对模型的全部内生变量的滞后值进行回归拟合，从而估计全部内生变量的动态关系[10]。可将VARX模型写为如下形式：

$$A(L)Y_t=B(L)e_t+C(L)X_t \tag{5.1.12}$$

其中，Y_t是k维输出变量，即因变量，表明有k个时间序列作为因变量，X_t为m维的输入变量，即自变量，由m个时间序列组成，e_t为k维不可观测的随机扰动过程，A、B、C为适当维的滞后算子L的矩阵，VARX模型为VARMAX模型的特例，其$B(L)$=1。

（三）状态空间模型拟合

空间状态模型为另一种多元时间序列模型，是VARX之外的另一种选择。

状态空间模型是一种动态模型，自变量为隐含着的时间。状态空间模型在经济时间序列分析中得到广泛的应用。其中应用最为普遍的状态空间模型是由 Akaike 提出并由 Mehra 进一步演绎发展的典型相关（canonical correlation）方法。由 Aoki 等人提出的估计向量值状态空间模型的新方法能得到内部平衡的状态空间模型，利用该种方法仅仅去掉系统矩阵中的相应元素就可以获得任意低阶近似模型而无需重新进行估计，并且只要原来的模型是稳定的，则由其得到的低阶近似模型也一定是稳定的。

构建状态空间模型有如下步骤：

（1）对相关的时间序列进行季节调整，并将季节要素序列外推；

（2）对季节调整后的时间序列进行单位根检验，确定单整阶数，然后在 ARIMA 过程中选择最接近的模型；

（3）求出 ARIMA 模型的系数；

（4）用 ARIMA 模型的系数准确表示正规状态空间模型，检验状态空间模型的可控制性；

（5）利用 Kalman 滤波公式估计状态向量，并对时间序列进行预测。

（6）把外推的季节要素与相应的预测值合并，就得到经济时间序列的预测结果。

一个独立于时间的线性状态空间的创新形式为：

$$Z_t = FZ_{t-1} + GX_t + Ke_{t-1} \qquad (5.1.13)$$

$$Y_t = HZ_t + e_t \qquad (5.1.14)$$

其中，Z_t 为不可观测的 n 维状态向量，F 为状态转移矩阵，G 为自变量，H 为因变量矩阵，K 为 Kalman 收益。第一个方程通常我们称之为状态转移方程，第二个称之为测量方程。这里 Z_t 是隐变量，而 Y_t 和 X_t 是可观测变量。

第二节　利用时间序列模型进行案例分析

（一）价格预测模型

在利用累积 probit 模型对量化交易策略我们得出结论,股票价格对于量化交易策略的实施具有非常显著的影响，而且股票价格一直是人们进行交易的首要参考，所以证券价格分析成为重要分析目标。对于价格的分析，往往使用单变量时间序列分析，这样的分析方法在于可以将所以的影响因素，无论是已知还是未知因素都包含在自回归当中，但是却无法具体分析其中一些重要因素的影响。本文将尝试使用多元时间序列模型来对股票价格进行预测。

（1）数据介绍

在这一节中，所使用的数据为兴业银行（601166）的日历史成交数据，数据所涉及的时间为 2013 年 1 月 7 日——2013 年 12 月 27 日,对于数据中因休市或停牌造成的无交易记录，本文采取移动加权平均的方法，利用无交易记录日所在周内数据进行加权平均，来补足当日数据。所以该节中所使用数据包括 245 个交易日数据，涉及收盘价（单位：元)、开盘价（单位：元)、最高价（单位：元)、最低价（单位：元)、成交量（单位：十万手)、成交额（单位：亿元）等六个变量。

（2）单位根检验

单位根检验是指检验序列中是否存在单位根，因为存在单位根就是非平稳时间序列了。单位根就是指单位根过程，可以证明，序列中存在单位根过程就不平稳，则回归分析中存在伪回归。所以首先对数据进行单位根检验，检验结果如下：

表 5.1 单位根检验 t 统计量

变量	t 统计量检验值	P 值
收盘价	-2.213	0.0264
开盘价	-2.135	0.0338
最高价	-2.145	0.0330
最低价	-2.093	0.0374
成交量	-2.117	0.0353
成交额	-3.375	0.000867

ADF 显著性水平为 0.05 的临界值为-3.43，而这六个检验统计量分别为

-2.213，-2.135，-2.145、-2.093、-2.117 和-3.375，则拒绝原假设，该数列的 6 个变量为非平稳的。

（3）协整检验

协整意味着向量分量之间存在长期关系，而我们第一步将先探求收盘价与开盘价、最高价、最低价之间是否协整，即是否存在这种长期关系。本文将使用 Johanse 方法中的迹检验方法进行协整检验，其检验结果如表 5.2：

表 5.2 ohanse 迹检验检验结果

H_0	H_a	检验统计量	0.1 临界值	0.05 临界值	0.01 临界值	结论
$\gamma \leq 5$	$\gamma > 5$	1.45	7.52	9.24	12.97	不显著
$\gamma \leq 4$	$\gamma > 4$	54.94	17.85	19.96	24.06	0.01 显著
$\gamma \leq 3$	$\gamma > 3$	114.77	32.00	34.91	41.07	0.01 显著
$\gamma \leq 2$	$\gamma > 2$	189.61	49.65	53.12	60.16	0.01 显著
$\gamma \leq 1$	$\gamma > 1$	301.66	71.86	76.07	84.45	0.01 显著
$\gamma = 0$	$\gamma > 0$	418.67	97.18	102.14	111.01	0.05 显著

通过 Johanse 迹检验检验结果我们可以看出，该模型可能存在 4 个协整向量。而通过 Johanse 特征根检验我们可以得出相同的结论。

（4）VARX 模型

利用 R 软件的（dse）软件包，带有输入变量的向量自回归移动平均模型为：

$$A(L)Y_t = B(L)e_t + C(L)X_t \qquad (5.2.1)$$

则根据拟合结果可写出如下 VARX 模型：

$$\begin{aligned} &\text{收盘价}_t + 0.053\times\text{收盘价}_{t-1} + 0.0031\times\text{收盘价}_{t-2} = \\ &\quad = -0.4248\times\text{开盘价}_t + 0.07943\times\text{开盘价}_{t-1} + 0.7431\times\text{最高价}_t + 0.1107\times\text{最高价}_{t-1} \\ &\quad + 0.6976\times\text{最低价}_t - 0.1556\times\text{最低价}_{t-1} + 0.0049\times\text{成交量}_t - 0.0076\times\text{成交量}_{t-1} \\ &\quad - 0.0027\times\text{成交额}_t + 0.0037\times\text{成交额}_{t-1} \end{aligned}$$

而图 5.3 给出了 VARX 模型拟合数据所得残差的 acf 图和 pacf 图：

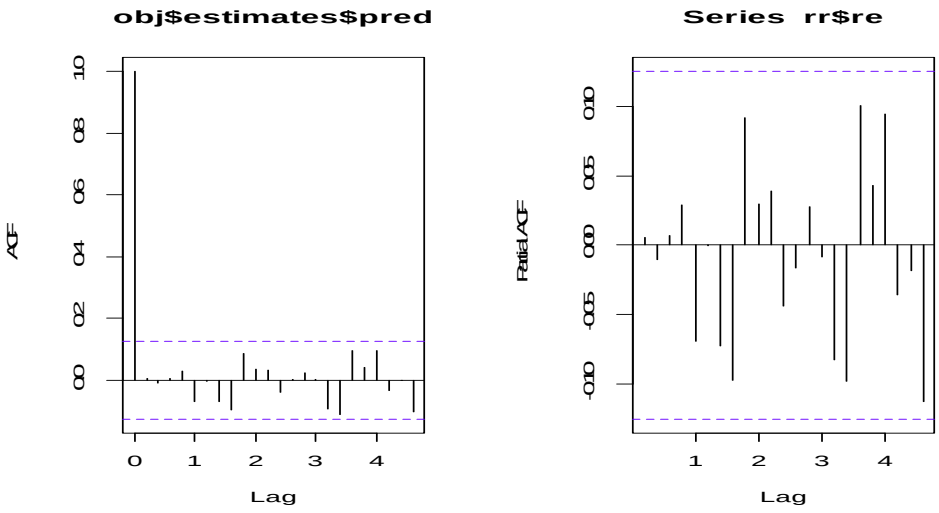

图 5.1 VARX 模型拟合数据所得残差的 acf 图和 pacf 图

（5）状态空间模型

空间状态模型为另一种多元时间序列模型，是 VARX 之外的另一种选择。一个独立于时间的线性状态空间的创新形式为：

$$Z_t = FZ_{t-1} + GX_t + Ke_{t-1} \tag{5.2.2}$$

$$Y_t = HZ_t + e_t \tag{5.2.3}$$

利用 R 软件可得状态空间模型拟合为：

$$F = \begin{bmatrix} 0 & 0 & -0.0110 \\ 1 & 0 & 0.0305 \\ 0 & 1 & 0.0418 \end{bmatrix}, \quad H = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

$$G = \begin{bmatrix} -0.0705 & 0.0787 & 0.0299 & -0.0093 & 0.0032 \\ 0.1002 & 0.0893 & -0.1649 & -0.0010 & 0.0004 \\ -0.4342 & 0.7575 & 0.6935 & 0.0075 & -0.0050 \end{bmatrix}, \quad K = \begin{bmatrix} -0.0110 \\ -0.0305 \\ -0.0418 \end{bmatrix}$$

其中，Z_t 为不可观测的 n 维状态向量，F 为状态转移矩阵，G 为自变量，H 为因变量矩阵，K 为 Kalman 收益。第一个方程通常我们称之为状态转移方程，第二个称之为测量方程。这里 Z_t 是隐变量，而 Y_t 和 X_t 是可观测变量。

（6）两种模型的比较

图 5.4 为 VARX 模型、状态空间模型以及原始数据的拟合效果图，其中绿色为 VARX 拟合模型，红色为状态空间模型，黑色为原始数据，可以发现两种模型拟合效果都非常好，故可使用这两种模型对数据进行分析和预测。

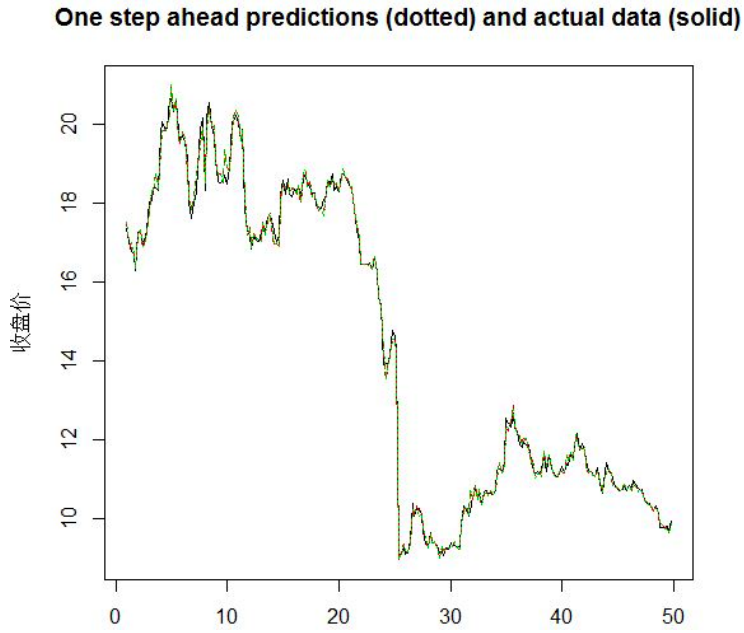

图 5.2 VARX 模型、状态空间模型拟合效果与原始数据的对比

而利用 VARX 模型对 2013 年 12 月 19 日——2013 年 12 月 27 日数据进行预测，其预测结果表 5.17。对比预测结果与实际数值我们发现该模型可以对实际数据有很好的预测效果。

图 5.3 VARX 模型预测值与原始数据的对比

日期	预测值	实际值
2013/12/19	10.18	10.1
2013/12/20	9.89	9.72
2013/12/23	9.85	9.73
2013/12/24	9.76	9.7
2013/12/25	9.79	9.73
2013/12/26	9.64	9.64
2013/12/27	9.90	9.68

第六章 主要结论及政策建议

第一节 主要结论

通过累计 probit 模型可以得出以下结论：

（1）量化交易策略偏向于与市场价格变化做同向交易，即当股票的市场价格上涨时，量化交易策略更倾向于进行买入交易，当股票价格下跌时，量化交易策略更倾向于进行卖出交易，即追涨杀跌；

（2）在高成交量的活跃市场氛围下，量化交易策略更倾向于进行买入交易，而在成交量比较低迷的情况下量化交易策略更倾向于进行卖出策略；

（3）买入挂单和卖出挂单对量化交易策略交替产生影响，本期高卖出挂单量趋向于使量化交易策略进行卖出交易而使得下一期量化交易策略趋向于进行买入交易，而高买入挂单量对量化交易策略有着相反的影响，即高买入挂单量趋向于使量化交易策略进行买入交易而使得下一期量化交易策略趋向于进行卖出交易；

（4）priIncT6 和 priIncT5 具有相同的符号，说明量化交易策略的实施对于股价走势的具有较为严格的要求，只有在股价走势相对稳定时更趋向于进行量化交易。

而通过对比累计 probit 模型、支持向量机以及贝叶斯网络三种方法，可以得出以下结论：

（1）在预测精度方面

通过以上三种方法对预测集的量化交易策略进行预测分析，我们可得以下模型检验比较结果：

表 5.14 三种模型预测效果比较

方法		误判率
累积 probit 模型		36.4%
贝叶斯网络模型	TAN 方法	29.43%
	马尔科夫覆盖	29.36%

支持向量机	简单法	13.88%
	专家法	13.81%

对比三种模型的误判率，我们可以看出：三种方法对于量化交易策略的预测具有明显差别：支持向量机方法较其他两种方法在预测中具有更高的准确率，更低的误判率，其误判率在 13.8%左右；而相比较而言，累积 probit 模型的误判率相对较高，为 36.4%。由此我们可以看出，从模型预测精度来看：

支持向量机＞贝叶斯网络＞累积probit模型

（2）在模型的可解释性方面

三种方法都是通过构建训练集模型来达到对预测集进行分析的目的，但是在对于模型的构建上具有明显的差异 :累积 probit 模型通过构建概率模型来对因变量进行预测，其建立的是具体的等式模型；贝叶斯网络模型是基于每个节点的概率分布表来进行分析判别，其利用最大似然估计来对模型的参数进行学习与估计；支持向量机则是利用从原始模式空间经过一个特定的函数的非线性变换映射到一个更高维的空间里，将一个低维非线性问题转化为某一个高维的线性问题，从而建立判别函数的方法对数据进行判别。通过其模型的构建方式，我们就可以看出，在对变量的解释性方面，累积 probit 模型对因变量具有更好更直观的解释性，我们可以通过其各自变量的参数对因变量进行解释；而支持向量机方法对于因变量的可解释性要比其他两种方式稍弱，其建立的仅仅是一种判别模型。所以在对变量的可解释性上：

累积probit模型＞贝叶斯网络＞支持向量机

综上对比三种量化交易策略分析方法，可以发现三种方法在策略分析上各有优势，无论是预测精度，还是影响因素的解释能力，在实际分析应用当中我们都经常会用到，所以在实际操作中，我们需将这三种方法加以结合，来对量化交易策略进行综合分析比较。

第二节　政策建议

量化交易作为新兴事物，其出现引发较多的争议以及导致市场剧烈波动产生种种问题也是在所难免。新的交易手段必定存在许多不完善甚至严重缺陷，但不能一味否定新生事物的生命力，而必须通过必要的监管措施引导其对市场发挥其应有的功能。美国和欧洲市场变得越来越有效，不能不说量化交易起到了一定的作用。随着中国金融市场的不断完善和发展，交易手段也不断推陈出新，交易方式和交易规则也在不断的进行，重新施行 T+0 的呼声也越来越高，刚刚出现的高频交易即将登陆中国股票市场。的确，在 T+0 交易中，一笔资金可以多次交易、反复买卖，在不增加市场资金存量的情况下，有效地提高市场的流通性、活跃程度和交易量，可以产生明显的资金放大效应。在市况较弱的情况下，“ T+0 ”交易制度一方面有利于减少投资者的投资风险，另一方面也将为投资者提供更多的短线交易机会，有助于投资者提高其盈利水平。在当前管理层已经实施“ 证券交易佣金浮动制 ”、投资者交易成本有所下降的情况下，也为实施“ T+0 ”回转交易提供了必要的技术准备。实施“ T+0 ”回转交易，还可为国家带来更多的印花税收入、为市场带来更多的短线机会、为券商带来更多的佣金收入，有利于“ 多赢 ”局面的形成，在一定程度上刺激当前交投清淡的弱市格局。

随着国内资本市场的不断成熟，以及国内资本市场逐步走向国际化，国内对于恢复 T+0 交易的呼声也越来越高，高频交易进入中国也只是时间问题。虽然现今为止量化交易尤其是高频交易的争论依然持续，但是量化交易的优势也是被大家所认可。对于量化交易存在的问题，我们可以通过以下监管方式予以消除：

（1）熔断制度的运用

熔断制度，就是在期货交易中，当价格波幅触及所规定的点数时，交易随之 停止一段时间，交易可以继续进行，但价幅不能超过规定点数之外的一种交易制度，在国内该制度被称为涨跌幅限制。设置“ 熔断 ”机制的目的是为了控制交易风险。“ 熔断 ”制度，是 1987 年世界股灾发生以后为了控制股票交易风险的一种交易制度，其后，“ 熔断 ”制度又被引入股指期货市场，它的设立为股指期货交易提供了一个减震器的作用。将熔断制度运用于量化交易特别是高频

交易监管当中，当出现非正常交易引起市场剧烈震荡时，执行该制度，可以暂停非正常波动股票的交易，交易所可以对非正常交易的原因进行核查，并及时纠正，起到稳定市场的作用。

（2）加强各量化交易尤其是高频交易商交易指令核查监管

回顾美股“5.6 闪崩”，东部时间 14:30 左右，某基金抛售价值超过 40 亿美元的 E-mini S&P 期货，而一部分买家为高频交易商。十分钟后，这些被买入的基金在算法的控制下，在各大高频交易商之间频繁交易，成为美股闪崩的导火索。归根结底我们可以发现，其根本原因为错误的执行了高频交易策略。加强各高频交易商交易质量的核查，其重点在于交易商自身的监督与核查。对于高频交易，交易商应该加强对算法的监管，以及对交易信号的正确识别，加强交易策略核查。

（3）明确量化交易特别是高频交易商的市场义务

有权利则必有义务。高频交易商拥有在市场上快速选择交易对手完成交易的权利，从公平公正的角度来讲，交易商则必须履行维护市场稳定、促进交易价格公允的义务。由文章一开始高频交易的特点我们可以看出，高频交易对加强市场的流动性、市场的价格发现等都发挥着举足轻重的作用，交易商同时在一定程度上还履行了做市商的部分功能，作为证券市场的监管部门必须正确认识到高频交易的影响范围，制定相应的监管规则，要求高频交易商履行维护证券市场有效性的义务，不得操纵市场、控制股市，杜绝非法交易，并需制定细致严厉的处罚措施。同时，高频交易商必须依法执行监管者对其设定的法律义务，遵守职业道德，维护市场信誉。

随着我国金融市场的不断创新和发展，量化交易发展所需的外部环境已逐步成熟。特别是 2010 年 4 月股指期货上市使中国股票市场具备了完善的套利机制，量化交易在中国将进入一个快速发展的阶段。由于国内外市场在交易机制等方面的不同，量化交易在国内市场的表现形式必然也有所不同。这就要求我们从中国的特殊市场条件出发，维护公平、透明、高效的市场秩序，并控制系统性风险，从而在监管中做到有的放矢、趋利避害，充分发挥虚拟经济对实体经济保驾护航的作用。

参考文献

[1]《量化交易轻松跑赢操盘手》金融电子化.2012 年 第 12 期 (2)

[2]Kearns,Kulasza,Nevmyvaka.Empirical Limitations on High Frequency profitableity [D]. Working Paper.2010.

[3]T·Hendershott,R·Riordan.Algorithmic trading and information.SSRN.2009.

[4]Hull,Options, futures and other derivatives[M].Pearson Prentice Hall.2009.

[5]王俊杰.量化交易在中国股市的应用[D].2013.

[6]镇磊.基于高频数据处理方法对 A 股算法交易优化决策的量化分析研究[D]. 中国科学技术大学 2010.

[7]乔治·H,邓特曼等著,吴晓刚主编.广义线性模型[M].上海,人民出版社,2011:340 页.

[8]宫秀军.贝叶斯学习理论及其研究[D].2002 年.

[9]吴喜之.复杂数据统计方法[M].北京,中国人民大学出版社,2011:196---197 页.

[10]李南成,马萍,徐舒.房地产价格的政策效应—基于 VARX 模型的研究[J].西南财经大学学报,2006.(3).51—54.

致　谢

时光蹁跹，岁月流逝，两年前踏入校园的情景恍如昨日。在毕业论文完成之际，心情却是颇为复杂，有欢乐，有感激，有留恋。回顾这三年的求学生涯，我收获的不仅是愈加丰厚的知识，更是人生的成长。成长的路上，身边的良师益友给予我极大的宽容与帮助，在此谨以最朴实的言语致以我最真诚的谢意。

感谢我的导师陈贻娟教授，能被陈老师收于门下，是我学习生涯中莫大的幸事。陈老师治学严谨、学识渊博、学风淳朴，从论文的选题、撰写到修改都给予我极为中肯的指导意见；陈老师宽容体贴、待人诚恳，对待学生如自己的孩子，在生活上教会我许多人生道理，当我步入困境时为我指点迷津，犹如醍醐灌顶。在此谨向陈老师致以诚挚的感谢和崇高的敬意。

感谢两年来一起前行的同学和舍友们，感谢一起参加各种比赛的同学伙伴，感谢云南财经大学统计与数学学院的全体老师，他们让我在一个良好的学术的环境中学习与成长，为我传道解惑，他们举办的学术讲座让我拓宽视野，引领我步入神圣的学术殿堂。

最后，感谢我的家人对我无微不至的关怀与支持，是他们让我衣食无忧，能让我静心的求学。养育之恩，无以为报。他们在物质上与精神上的无私支持，饱含着对我殷切的期望，同时也给予了我追求人生理想的坚实力量。

本人在读期间的研究成果

2013 年全国大学生统计建模大赛优秀奖；

云南省农业银行网点规划报告撰写。