学校代码: 11658　　学　　号: 2013207010004
分 类 号: ****　　密　　级: 无

海南师范大学
Hainan Normal University

硕士学位论文

基于状态空间模型的房地产“情绪指数”研究

作 者 姓 名: 李彩雯
学 科 专 业: 应用数学
研 究 方 向: 金融数学
指 导 教 师: 蒋文江教授
申请学位类别: 理学硕士
提 交 日 期: 二〇一五年五月

A study based on State space model of real estate "Sentiment Index"

A Dissertation Submitted for the Degree of Master

Candidate: LI,Caiwen

Supervisor: Professor JIANG,Wenjiang

Hainan Normal University

Haikou, China

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得海南师范大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名:_____ 日 期:_____

学位论文著作权声明

本论文作者声明:

□ 本论文全部成果均为本人和指导教师合作研究取得，本人和指导教师都有权使用本成果学术内容（有第三方约定者除外）。

□ 本论文为指导教师指导下，本人独自完成。本人独自享有本论文的全部著作权。

学位论文作者签名:_____ 指导教师签名:_____

日 期:_____ 日 期:_____

学位论文版权使用授权书

本学位论文作者完全了解海南师范大学有关保留、使用学位论文的规定，即：海南师范大学有权保留并向国家有关部门或机构送交学位论文的复印件和电子文本，允许论文被查阅和借阅。本人授权海南师范大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其它复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名: _____ 指导教师签名:_____

日 期: _____ 日 期:_____

摘　要

近年来，房地产行业一直受到广泛关注，国内外多数学者采用房产基价求解法、指标体系测度法、统计检验测度法来评测房产的价格和估计房地产市场的泡沫，并预测房地产泡沫破灭在即，但多年过去了，这一切并没有发生。所以想到的问题是这种传统的分析方法在中国是否有效？有没有其他替代的方法可用？依据行为金融学理论可知市场参与者的心理因素以及价值共同决定了其价格的运动变化，这种心理因素对价格的走势有着极为重要的作用。所以本文主要想利用房地产市场相关行业的股市数据来综合评测出房地产参与者情绪指数这个潜变量，并通过评价房地产市场参与者的情绪指数来分析和预测房地产价格走势。这里我们假定房地产市场参与者的情绪指数是隐藏在房地产价格变化趋势背后的一种潜在变量。本文主要采用股市的数据，因为股市的数据获取方便、有质量保证，且能够即时更新。本文的核心思想是通过建立状态空间方程导出房地产市场参与者的情绪指数，首先联合使用EM 算法和Kalman 算法来估计模型中的未知参数，然后再用参数确定之后的模型，利用可观测到的股市相关行业的数据来分析情绪指数的变化规律和趋势，具体操作是：在给定参数初值后，应用Kalman 算法得到最初的潜在变量，再将其带回到状态空间模型，应用EM 算法更新参数，所以在每一次更新过程中应用EM算法和Kalman算法，这样不仅可以估计出方程的参数，还可以得到潜在变量，即房地产市场参与者的情绪指数。

关键词：行为金融学、房地产市场参与者的情绪指数、Kalman滤波、状态空间模型、EM算法

Abstract

In recent years, the real estate industry has got so much attention, most scholars use a lot of different ways to evaluate real estate prices and estimate the real estate market bubble.But these traditional prediction methods failed so many times .So it shows us that these kind of methods have huge flaws. Based on behavioral finance theory, we know that market participants' psychological factors and the price both can determine the price movement, and this kind of psychological factors has become an extremely important role on the price trend . So this paper will use the real estate related industries stock data to evaluate real estate participants' sentiment index comprehensively.Here we assume that participants' sentiment index in the real estate market is hidden ,it is a kind of latent variable hiding behind the real estate price trends. So we want to use a practical and data reliable way to estimate this potential variable. The core thought is establishing the state space equation ,giving parameter initial values, applying Kalman algorithm to get initial potential variable,then bringing them back state space model, applying the EM algorithm to update parameter.That means in each update process we need use EM algorithm and Kalman algorithm, So in this way ,we can estimate the parameters of the equations and we can gain latent variables at the same time.

Keywords：Behavioral finance theory,Kalman algorithm,State-space models,EM algorithm,Real estate “Sentiment Index”

目 录

第一章 引言 **1**

1.1 研究背景、方法及意义 1

1.2 文献综述及研究评价 2

1.2.1 文献综述 2

1.2.2 研究评价 3

1.3 应用概念及算法简介 4

1.3.1 行为金融学 4

1.3.2 状态空间模型思想介绍 4

1.3.3 Kalman滤波算法 5

1.3.4 EM算法 6

第二章 模型的建立及算法介绍 **9**

2.1 状态空间模型建立 9

2.2 本文算法流程介绍 9

2.2.1 算法开始前的初步计算 9

2.2.2 状态空间方程参数的期望表示 10

2.2.3 状态空间方程的参数表示 11

2.2.4 整体算法思路解释 12

2.2.5 算法流程图及详解 13

第三章 实证演练 **15**

3.1 前期准备 15

3.1.1 数据的选取及合理性 15

3.1.2 数据的预处理及合理性 15

3.1.3 软件包的选取及合理性 15

3.2 模型选择 16

3.2.1 不同方式定义模型结果对比 16

3.2.2 四大相关板块状态空间模型参数估计情况 16

3.2.3 房地产“情绪指数”总体评价 19

3.2.4 建立房地产“情绪指数”回归方程 23

第四章 总结与展望 **25**

4.1 研究总结 25

4.2 研究展望 25

4.2.1 预测应用 25

4.2.2 多变量合一应用 26

4.2.3 ARMA模型应用 26

参考文献 **27**

攻读硕士学位期间发表论文清单 **29**

致　　谢 **31**

第一章　引言

§1.1　研究背景、方法及意义

房地产业一直以来是人们关注的热点，近年来广为流传的泡沫破灭说一直没有得到实现，房价的持续上涨造成房产投资者的恐慌和迷茫，目前多数研究学者纠结于研究房产基价求解和指标式房产价值评价等问题，不过很显然房地产业的变化趋势并不是完全由表面的实体价值决定的，一定存在一种潜在的变量推动着其进行变化。同时多次对泡沫破灭预测的失败表明这一套评估体系并不能较为准确的预测房价未来走势。由行为金融学的理论可知，市场参与者的心理因素以及价格共同决定了其价格的运动变化，这种心理因素对价格的走势有着极为重要的作用。然而，这个重要的行为金融学变量又是不可直接观测的，对这类不可直接观测和度量的行为金融变量的度量是当今金融学数据研究的一个热点，在当代概率论统计的理论体系之中，处理这类潜变量的一个较为成熟的模型是所谓的状态空间模型，其基本框架是建立状态空间模型，通过可观测的变量获取不可观测变量的信息，那么这种不可观测的变量在本文又是指什么呢？这种不可观测的变量正是一种潜在的信息，并在一定程度上决定着房地产价格的未来变化趋势，而这种变化趋势是由潜在房地产市场背后的市场参与者情绪指数来决定的，所谓的房地产市场参与者情绪指数是潜在房地产相关可观测信息背后的一种决定性因素，包含了推动房地产背后变化的全部信息，当然如果想切实可行的得到这些决定房产变化的全部潜在信息，就需要得到决定房地产变化的全部外在观测信息。但是如何获得这些外在观测信息，如何获取这些可靠的数据又是一个值得思考的问题。

纵观国内外研究，可以发现，研究房地产的方法虽多且广，但在数据使用方面，多面临着数据采取难，可信度低，无时效性等现实问题，所以得到的研究成果多被质疑。那么选取一种可靠可取的公共数据资源是最佳的选择，由此我们想到采用股票数据，采用股票市场里的房地产相关的板块数据来综合的评价房地产潜在的“情绪指数”是可行的，一旦状态空间模型确定，我们可以有效的通过观测变量，获得潜在的不可观测变量。

但如何才能够确定状态空间模型呢？这就涉及到参数估计问题，所以我们首先想到估计存在潜在变量的参数估计方法EM算法，那么如何有效估计参数的同时，还能够通过Kalman方法得到不可观测变量；这就是本文采用的方法的核心之处，我们的核心思想是建立状态空间模型后，给定一组参数初值，然后确定初始状态的状态空间模型，应用Kalman滤波方法得到最初的不可观测变量，然后应用EM算法更新参数，直到极大似然函数收敛，在每一步更新过程中应用Kalman滤波，这样最终我们可以同时得到最优的参数估计和不可观测变量。

状态空间模型应用的Kalman算法不仅可以获得不可观测变量，而且还可以应用Kalman-Rauch平滑方法对可观测变量的未来趋势进行预测。这就意味着当投资者面临投资问题时，可以采用本文的方法，利用公共资源，即已有的房地产相关板块数据，直接获取过去到未来的房地产“情绪指数”，并且能够对这些板块的未来趋势进行预测，从而更直观的判断此时是否有投资的价值。所以房地产“情绪指数”就像一个风向标，可以牵动着投资者们的心，如果能够对其进行实时有效的评价，那么对房地产市场来说的意义非凡。

§1.2 文献综述及研究评价

§1.2.1 文献综述

房地产“情绪指数”的定义引自市场情绪指数，市场情绪是指整体市场所有市场参与人士观点的综合展现，情绪指数是用来衡量社会情绪波动状况的综合指数。所以我们这里房地产的“情绪情绪”是指房地产市场参与人士情绪波动状况的综合指数，这里我们假定它是一种潜在房地产背后的价值决定性因素。本文之后的内容将直接用房地产“情绪指数”一词来表示房地产市场参与者的情绪指数，此外由于国内外没有对专门的房地产“情绪指数”进行定义，所以本文的房地产“情绪指数”都会用引号括起来。

下面就简单介绍下房地产相关研究的现状，并对其存在的问题进行阐述，主要从相关的房地产价格指数评价和房地产泡沫这两个研究最广泛的方面进行文献分析和评价。

一、房地产价格指数：美国具有代表性的住房价格指数是OFHEO住宅价格指数，采用Fannie Mae 和Freddie Macin的统计数据为基础，应用Repeat-Sales模型进行计算。英国的住宅价格指数由HBOS（Halifax and Bank of Scotland）编制，采用全国住宅按揭贷款数据，主要包括每月大约15,000个住房购买数据。模型方面，其中Repeat-Sales模型和Hedonic模型是较为常用的价格指数研究模型，此外Gress在Hedonic模型中采用空间自相关价格模型，冷凯君对Hedonic模型进行半参数法估计，陈安明采用主成份分析法消除Hedonic模型共线性问题，其中国内对相关研究较为全面的是郭红领的中国房地产信心指数研究，这篇文章具体的阐述了房地产信心指数的评测方法和指数评价标准，并以深圳房地产为例展开了研究。

二、房地产泡沫评价：国外方面。Gala Garino and Lucio Sarno采用RALS(residual－augmented least squares) 估计的DF检验和RALS协整回归DF 统计检验、体制转换检验对英国住房市场泡沫进行了分析；Charles H immelberg通过房价收入比、房价租金比建立模型得到美国1980-2004年46 个城市房地产的房地产价格波动和基础价格, 说明美国在此期间没有明显的房地产泡沫。国内方面，赵安平等人利用基本房价建立状态空间模型，并将基本房价视作状态变量，从而计算出房价泡沫。钟晓兵研究了21 世纪以来我国房地产泡沫测度及生成机制，指出了泡沫测度的各项指标，包括房地产投资增长率/GDP增长率、房价收入比、房地产开发贷款额/金融机构贷款总额等指标来评价房地产泡沫。李维哲和曲波提出房地产泡沫的预警指标，分为生产类指标、交易状况类指标、消费状况类指标和金融类指标四大类。张朝阳研究了信贷扩张与房地产泡沫之间的相关性，并得到信贷扩张会伴随着房地产价格的持续快速上涨而上涨的结论。刘琳、黄英和刘洪玉借鉴国外从信用角度分析泡沫的方法分析房地产泡沫的形成机制, 建立适合我国房地产市场的房地产泡沫测度系数, 并进行实证分析。

房地产泡沫破灭评价：关海玲评价说房地产市场的泡沫受到金融运行状况、实体经济发展状况、全民的社会心态等多方面的影响。彭湘、卿斌评论说由于房地产市场资金来源枯竭、需求萎缩，房地产产业将面临崩溃的局面。易宪容也曾在2011年时评论说房地产泡沫破灭为时不远。王南指出了房地产泡沫破灭三个时间节点，并评论说如果房地产泡沫没有被行政的力量刺破，也没有被税收等经济的力量刺破，那么它将躲不过被自然的力量刺破。赵晓讲述了房地产泡沫破灭的两个场景，一是改革进程缓慢，房地产泡沫越吹

越大。二是改革迅速取得突破，房地产泡沫破灭提前到来。

§1.2.2 研究评价

通过参考目前对房地产市场的研究不难发现。对于宏观研究法，多数学者采用指标、指数研究法，应用统计经济模型对各项指标进行评价，并综合全面具体的来阐述房地产市场发展的现状。但是这种方法的弊端是数据采集，多存在不可靠性，如问卷调查得到的数据过于局限，无法应用此类数据进行进行房地产综合市场的评价；或是数据采集无时效性，多数得到的数据为陈旧数据，无法进行实时的房地产市场发展趋势进行评测；此外指标研究法，无法准确的找到指标与房地产市场的内在联系，只能通过相关系数来进行评价，这样就无法找到真正影响房地产变化趋势的变量。

对于房地产泡沫的评价，学者们从房地产宏观指标的理解，弱化到对微观概念泡沫的求解上，但是碍于对泡沫定义的限制，多数学者纠结于房地产基价的求解，一般采用指标法进行房产基价求解或是直接应用指标对房地产市场泡沫进行客观的评价；但房地产的基价是否由这些指标决定，这是未知的，只能说房地产存在潜在价值来决定着这种发展趋势。

纵观国内外关于房地产市场的各类研究，离不开经济统计模型法、指标评价法、测度评测法、综合分析法等几个基本方法。但是这些方法多存在片面性，无法全面具体的评价房地产市场的整体发展趋势，更是受数据采取的限制，很难对房地产未来发展进行准确的预测。

更主要是多数学者在过去的十年一直声称房地产泡沫即将破灭，但十年过去了，房地产产业还是在稳步发展，房地产价格还是稳中有升，说明之前的传统评价方法和理论分析法并不能合理的对房地产价格趋势进行预测和分析。

§1.3 应用概念及算法简介

§1.3.1 行为金融学

行为金融学主要是研究市场参与者心理、社会、认知等情感因素和市场价格对个人和机构的经济决策、回报和资源分配造成的影响。行为模型通常从心理学、神经科学和微观经济理论等来分析决策者的行为。它从微观个体行为以及产生这种行为的心理等动因来解释、研究和预测金融市场的发展。这一研究视角通过分析金融市场主体在市场行为中的偏差和反常，来寻求不同市场主体在不同环境下的经营理念及决策行为特征，力求建立一种能正确反映市场主体实际决策行为和市场运行状况的描述性模型。

行为金融学解释了为什么市场参与者会做出非理性与理性决策，为什么有时甚至会得到与假设相反的结果。这类错误直接影响到价格和收益，造成市场的低效率。同时行为金融学的研究也包括调查其他参与者如何利用套利等手段造成市场效率低下等问题。行为金融学主要还会分析过度反应导致的市场趋势和在极端情况下的泡沫和崩溃等信息，这种反应主要体现在投资者的注意力、过度自信、乐观、模仿和噪声交易等心理表现上。

此外技术分析师也认为行为金融学是技术分析的理论基础。多数投资者都是具有损失厌恶的心理，即不愿意出售股份或其他权益，这也能解释为什么房价很少慢慢地下降到市场清算期间的需求较低的水平。

§1.3.2 状态空间模型思想介绍

状态空间模型提供了一种有效的方法来估计不可观测变量,且广泛的应用于多变量时间序列。

状态空间模型的定义：设Y是包含k个变量的$k\times 1$维的可观测向量，这些变量与$m\times 1$维向量X有关，X 称为状态向量。

定义量测方程(Measurement Equation):

$$y_t = z_t x_t + b_t + \varepsilon_t, \quad t = 1, 2, ...T$$

式中T表示样本长度,z_t是$k\times m$矩阵，b_t是$k\times 1$向量，ε_t是$k\times 1$向量，其均值为0，协方差矩阵为H_t 的连续的不相关扰动项，即$E[\varepsilon_t]=0$,$var[\varepsilon_t]=H_t$

一般地，$\{x_t\}$是不可观测的，然而可表示成一阶马尔可夫(Markov)过程。

下面定义转移方程(Transition Equation):

$$x_t = A_t x_{t-1} + c_t + R_t \eta_t, \quad t = 1, 2, ...T$$

式中A_t是$m\times m$矩阵，c_t是$m\times 1$向量，R_t是$m\times g$矩阵，η_t是$g\times 1$向量，其均值为0，协方差矩阵为Q_t 的连续的不相关扰动项，即$E[\eta_t]=0$,$var[\eta_t]$

注：若使上述的状态空间模型成立，还需要满足下面两个假定：

(1)初始状态向量x_0的均值为0，协方差矩阵为P_0，即$E[\eta_t]=0$,$var[\eta_t]=P_0$

(2)在所有的时间区间上，扰动项ε_t和η_t是相互独立的，且它们和初始状态x_0 也不相关，即$E[\varepsilon_t\eta_s]=0, t, s=1,2,...T$且$E[\varepsilon_t x_0]=0, E[\eta_t x_0]=0$

量测方程中的矩阵z_t, b_t, H_t与转移方程中T_t, c_t, R_t, Q_t矩阵统称为系统矩阵。一般情况下，它们都被假定为非随机的。因此，尽管它们能随时间改变，但是都是可以预先确定的。对于任一时刻t，$\{y_t\}$能够被表示为当前的和过去的ε_t和η_t 及初始向量x_0 的线性组合，所以模型是线性的。

对于任何特殊的统计模型，x_t的定义是由结构确定的。它的元素一般包含具有实际解释意义的成分，例如趋势或季节要素。状态空间模型的目标是，所建立的状态向量x_t包含了系统在时刻t的所有有关信息，同时又使用尽可能少的元素。所以如果状态空间模型的状态向量具有最小维数，则称为最小实现(Minimal Realization)。对一个好的状态空间模型，最小实现是一个基本准则。

然而，对于任一特殊问题的状态空间模型的表示形式却不是惟一的，这一点很容易验证。考虑通过定义一个任意的非奇异矩阵B，得到新的状态向量$x_t^* = Bx_t$。用矩阵B 左乘转移方程，得到

$$x_t^* = T_t^* x_{t-1}^* + c_t^* + R_t^* \eta_t$$

式中$T_t^* = BT_tB^{-1}, c_t^* = Bc_t, R_t^* = BR_t$ 。相应的量测方程是

$$y_t = z_t^* x_t^* + b_t + \varepsilon_t$$

式中$z_t^t = z_tB^{-1}$。

系统矩阵Z_t, H_t, T_t, R_t, Q_t依赖于一个未知参数的集合，状态空间模型的一个主要的任务就是估计这些参数，这些参数被称为超参数(Hyperparameters)。超参数确定了模型的随机性质，而系统参数b_t或c_t等参数仅影响确定性的可观测变量和状态的期望值。

§1.3.3 Kalman滤波算法

Kalman滤波算法是最优化自回归数据处理算法，其核心思想是以最小均方误差为最佳估计准则，主要采用信号与噪声的线性状态空间模型，并运用递推的方法来解决线性优化滤波问题，即利用前一时刻地估计值和现时刻的观测值来更新对状态变量的估计，求出现时刻的估计值。Kalman滤波要比直接从全部的过去数据里进行每步滤波要高效，它适合于实时处理和计算机运算。

具体求解方法如下：

假设系统可用一个线性随机微分方程来描述：

$$x_k = Ax_{k-1} + Bu_k + w_k$$

再加上系统的测量值：

$$z_k = Hx_k + v_k$$

x_k是k时刻的系统状态；u_k是k时刻对系统的控制量；A和B是系统参数，对于多模型系统，他们为矩阵；z_k是k时刻的测量值；H是测量系统的参数，对于多测量系统，H为矩阵；w_k和v_k分别表示过程和测量的噪声，为高斯白噪声，协方差分别是Q,R。

第一步：利用系统的过程模型，来预测下一状态的系统。

假设现在的系统状态是k，根据系统的模型，可以基于系统的上一状态而预测出现在状态：

$$x_{k|k-1} = Ax_{k-1|k-1} + Bu_k \quad (1)$$

式(1)中，$x_{k|k-1}$是利用上一状态预测的结果，$x_{k-1|k-1}$是上一状态最优的结果,u_k为现在状态的控制量，如果没有控制量，它可以为0。

$$P_{k|k-1} = AP_{k-1|k-1}A' + Q \quad (2)$$

P表示协方差,其中式(2)中，$P_{k|k-1}$是$x_{k|k-1}$ 对应的协方差，$P_{k-1|k-1}$是$x_{k-1|k-1}$对应的协方差，A’表示A的转置矩阵，Q 是系统过程的协方差，式子1，2是对系统的预测。

第二步：得到现在状态(k)的最优化估算值$x_{k|k}$

$$x_{k|k} = x_{k|k-1} + Kg_k(Z_k - Hx_{k|k-1}) \quad (3)$$

其中Kg为卡尔曼增益(Kalman Gain)：

$$Kg_k = \frac{P_{k|k-1}H'}{HP_{k|k-1}H' + R} \quad (4)$$

$x_{k|k}$即为k状态下最优的估算值,为了要令卡尔曼滤波器不断的运行下去直到系统过程结束；
我们还要更新k状态下$x_{k|k}$的协方差：

$$P_{k|k} = (I - Kg_kH)P_{k|k-1} \quad (5)$$

其中I 为1的矩阵，对于单模型单测量，$I = 1$。当系统进入k+1状态时，$P_{k|k}$就是式子(2)的$P_{k-1|k-1}$。这样，算法就可以自回归的运算下去。

§1.3.4 EM算法

EM算法是Dempster Laind Rubin于1977年提出的，它是一种迭代算法，用于含有隐变量的概率参数模型的最大似然估计或极大后验概率估计，其目标是找出有隐性变量的概率模型的最大可能性解，它分为两个过程E-step 和M-step，E-step通过最初假设上一步得出的模型参数得到后验概率（期望），M-step 重新算出模型的参数，重复这个过程直到目标函数值收敛。EM算法广泛地应用于处理缺损数据，截尾数据，带有噪声等所谓的不完全数据中，在机器学习、高斯混合模型、HMM(隐性马尔科夫链)以及计算机视觉的数据聚类等研究方面也得到广泛应用。

方法简介：

假定集合$Z = (X, Y)$由观测数据X和未观测数据Y组成，X 和$Z = (X, Y)$分别称为不完整数据和完整数据。假设Z的联合概率密度被参数化地定义为$P(X, Y|\Phi)$，其中Φ表示要被估

计的参数。Φ的最大似然估计是求不完整数据的对数似然函数$L(X;\Phi)$的最大值而得到的：

$$L(\Phi;X)=\log p(X|\Phi)=\int \log p(X,Y|\Phi)dY$$

EM算法包括两个步骤：由E步和M步组成，它是通过迭代地最大化完整数据的对数似然函数$Lc(X;\Phi)$的期望来最大化不完整数据的对数似然函数，其中：

$$Lc(X;\Phi)=\log p(X,Y|\Phi)$$

假设在算法第t次迭代后Φ获得的估计记为$\Phi(t)$，则在$t+1$次迭代时

E-步：计算完整数据的对数似然函数的期望，记为：

$$Q(\Phi|\Phi(t))=E[Lc(\Phi;Z)|X;\Phi(t)]$$

M-步：通过最大化$Q(\Phi|\Phi(t))$来获得新的Φ。

通过重复上述两步，EM算法逐步更新模型的参数，使参数和训练样本的似然概率逐渐增大，最终找到一个极大点。直观地理解EM算法，它可被看作为一个逐次逼近算法：事先并不知道模型的参数，可以随机的选择一组参数或者事先粗略地给定某个初始参数Φ0，确定出对应于这组参数的初始状态，计算每个训练样本的可能结果的概率，在当前的状态下再由样本对参数更新，重新估计参数Φ，并在新的参数下重新确定模型的状态，这样，通过多次的迭代，循环直至某个收敛条件满足为止，就可以使得模型的参数逐渐逼近真实的参数值。

第二章　模型的建立及算法介绍

§2.1　状态空间模型建立

我们建立如下状态空间方程:

$$x_{t+1} = Ax_t + B + w_t, \quad w_t \sim N(0, Q)$$

$$y_t = Cx_t + v_t, \quad v_t \sim N(0, R)$$

模型解释：方程（1）是更新方程，方程（2）是观测方程，X,Y为两个随机变量，其中y_t是可观测的变量，且无缺失值，x_t是不可观测的变量（即隐藏变量），A、B、C是状态空间方程的系数，且待估计。w_t, v_t 是相互独立的高斯白噪声，且方差分别为Q、R。即除$\{y_t\}$可观测外，方程的其他信息未知。所以我们的目标是得到隐藏变量X的同时，估计方程的参数$\Phi = \{A, B, C, Q, R, \pi_1, V_1\}$

§2.2　本文算法流程介绍

§2.2.1　算法开始前的初步计算

根据状态方程和观测方程中的噪声为高斯白噪声，则有状态变量和观测变量的条件概率密度如下：

$$P(y_t|x_t) = \exp\{-\frac{1}{2}[y_t - Cx_t]'R^{-1}[y_t - Cx_t]\}(2\pi|R|)^{-1/2} \tag{2.1}$$

$$P(x_t|x_{t-1}) = \exp\{-\frac{1}{2}[x_t - Ax_{t-1} - B]'Q^{-1}[x_t - Ax_{t-1} - B]\}(2\pi|Q|)^{-1/2} \tag{2.2}$$

$$P(x_1) = \exp\{-\frac{1}{2}[x_1 - \pi_1]'V^{-1}[x_1 - \pi_1]'\}(2\pi|V_1|)^{-1/2} \tag{2.3}$$

则由马尔科夫性，可将$\{x_t\}, \{y_t\}$的联合概率密度表示为：

$$P(\{x_t\}, \{y_t\}) = P(x_1)\prod_{t=2}^{T} P(x_t|x_{t-1})\prod_{t=2}^{T} P(y_t|x_t) \tag{2.4}$$

展开并对其取log得到如下式子：

$$\begin{aligned}\log P(\{x_t\}, \{y_t\}) = &-\sum_{t=1}^{T}(\frac{1}{2}[y - Cx_t]^T R^{-1}[y - Cx_t]) - \frac{T}{2}\log|R| \\ &-\sum_{t=2}^{T}(\frac{1}{2}[x_t - Ax_{t-1} + B]^T Q^{-1}[x_t - Ax_{t-1} + B]) - \frac{T-1}{2}\log|Q| \\ &-\frac{1}{2}[x_1 - \pi_1]^T V_1^{-1}[x_1 - \pi_1] - \frac{1}{2}log|V_1| - T\log 2\pi\end{aligned} \tag{2.5}$$

由于上式中的似然函数含有不可观测变量，无法直接进行极大化。所以根据EM算法,首先

需要对完全数据的联合概率密度函数取条件期望，我们得到下式：

$$
\begin{aligned}
&E(\log P(\{x_t\},\{y_t\})|\Phi,\{y_n\}) = \Psi = \\
&-\frac{1}{2}\sum_{t=1}^{T}(E[Y_t^T R^{-1} Y_t] - E[Y_t^T R^{-1} C X_t] - E[(CX_t)^T R^{-1} Y_t] + E[(CX_t)^T R^{-1} C X_t]) - \frac{T}{2}\log|R| \\
&-\frac{1}{2}\sum_{t=1}^{T}(E[x_t^T Q^{-1} X_t] - E[X_t^T Q^{-1} A X_{t-1}] - E[(BX_{t-1})^T Q^{-1} X_t] - E[B^T Q^{-1} X_t] - E[X_t^T Q^{-1} B] \\
&+ E[(AX_{t-1})^T Q^{-1} A X_{t-1}] + E[B^T Q^{-1} A X_{t-1}] + E[A X_{t-1}^T Q^{-1} B] - B^T Q^{-1} B) - \frac{T}{2}\log|Q| \\
&-\frac{1}{2}(E[X_1^T V_1^T X_1] - E[\pi_1^T V_1^{-1} X_1] - E[X_1^T V^{-1} \pi_1] + \pi_1^T V_1^{-1} \pi_1) - \frac{1}{2}\log|V_1| - T\log\pi
\end{aligned}
\tag{2.6}
$$

这里由于上述条件期望较长，所以舍去每个条件期望的条件表示，但实际每个期望表示应理解为关于参数Φ和观测变量$\{y_n\}$的条件期望。

§2.2.2 状态空间方程参数的期望表示

我们的目标是极大化上式，然后对每个待估计参数求偏导，从而求得参数，并不断重复此过程进行参数更新，直至似然函数收敛。

这里我们每个参数都可以得到如下的显示解，所以我们逐一进行计算如下：

$$A_{j+1} = (\sum_{t=2}^{T} E[x_t^T x_{t-1}] - B^T E[x_{t-1}])(\sum_{t=2}^{T} E[x_t x_{t-1}^T])^{-1} \tag{2.7}$$

$$B_{j+1} = \frac{1}{T}\sum_{t=1}^{T}(E[x_t] - AE[x_{t-1}]) \tag{2.8}$$

$$C_{j+1} = (\sum_{t=1}^{T} E[y_t \hat{x}_t])(\sum_{t=1}^{T} E[x_t^T x_t])^{-1} \tag{2.9}$$

$$Q_{i,j+1} = \frac{1}{T}\sum_{t=2}^{T}(E[x_t^T x_t] - 2E[x_t^T x_{t-1}] - 2E[x_t]B + AE[x_{t-1}^T]A^T + 2BE[x_{t-1}^T]A^T + BB^T) \tag{2.10}$$

$$R_{j+1} = \frac{1}{T}\sum_{t=1}^{T}(E[y_t^T y_t] - 2E[y_t x_t]C^T + CE[x_t^T x_t]C^T) \tag{2.11}$$

但是可以看出，上式都是由关于X,Y的条件期望表示，所以下面的问题就是如何求解这些条件期望，或是如何表示出这些条件期望。

本文采用的方法是用Kalman滤波得到的$\hat{x}_t$，$\hat{P}_t$，$\hat{P}_{t,t-1}$等来表示这些条件期望。当然这些条件期望也可以直接进行表示，但是为了同时通过Kalman滤波得到不可观测变量，这种表示形式更易计算，且也是非常有效的。

下面的这些变量我们可以通过Kalman滤波算法得到，也就将上面各个参数中涉及含有X的条件期望表式出来了。

含有X的条件期望表示：

$$\hat{x}_t = E[x_t|\{y_t\},\Phi] \tag{2.12}$$

$$\hat{P}_t = E[x_t x_t^T | \{y_t\}, \Phi] \tag{2.13}$$

$$\hat{P}_{t,t-1} = E[x_t x_{t-1}^T | \{y_t\}, \Phi] \tag{2.14}$$

$$\hat{V}_t = var[x_t | y_t, \Phi] \tag{2.15}$$

$$\hat{V}_{t,t-1} = cov[x_t, x_{t-1} | \{y_t\}, \Phi] \tag{2.16}$$

但是可以从参数期望表达式可以看出，里面不仅含有形如$E[x_t|\Phi, \{y_n\}]$等关于X的条件期望，同时也含有形如$E[y_t|\Phi, \{y_n\}]$等关于Y的条件期望，所以这些我们可以直接进行求解，表示如下：

$$E[y_t y_t^T | \{y_t\}, \Phi] = y_t y_t^T \tag{2.17}$$

$$E[y_t y_{t-1}^T | \{y_t\}, \Phi] = y_t y_{t-1}^T \tag{2.18}$$

$$E[x_t y_t^T | \{y_t\}, \Phi] = x_t y_t^T \tag{2.19}$$

$$E[y_t x_{t-1}^T | \{y_t\}, \Phi] = y_t x_{t-1}^T \tag{2.20}$$

§2.2.3 状态空间方程的参数表示

根据上述的式子重新表示下参数的期望表示形式，去掉期望。则每次更新后的每个参数表示如下：

$$A_{j+1} = \sum_{t=2}^{T} (P_{t,t-1} - B^T x_{t-1}^T) \tag{2.21}$$

$$B_{j+1} = \frac{1}{T} \sum_{t=2}^{T} (\hat{x}_t - A\hat{x}_{t-1}) \tag{2.22}$$

$$C_{j+1} = (\sum_{t=1}^{T} y_t \hat{x}_t)(\sum_{t=1}^{T} P_t)^{-1} \tag{2.23}$$

$$Q_{j+1} = \frac{1}{T} \sum_{t=1}^{T} (P_t - 2P_{t,t-1}A^T - 2\hat{x}_t B^T + AP_{t-1}A^T + 2A\hat{x}_{t-1}B^T + BB^T) \tag{2.24}$$

$$R_{j+1} = \frac{1}{T} \sum_{t=1}^{T} (y_t^T y_t - 2y_t \hat{x}_t C^T + CP_t C^T) \tag{2.25}$$

$$\pi_1 = x_1, V_1 = P_1 - \pi_1^2 \tag{2.26}$$

现在我们得到了各个参数的表示形式，除了其中的含X 的信息我们不知道之外，其他的信息我们都已经知道了，所以下一步就是应用Kalman滤波估计不可观测的变量X。

具体的整体算法思路详见下面解释。

§2.2.4 整体算法思路解释

Kalman算法开始：

分析1：根据Kalman滤波算法，我们知道要开始整体的算法，状态空间方程的系统参数需要已知。但是我们设定的模型中，是需要估计系统参数的，所以我们想到给定参数Φ 的初值，这里我们采用蒙托卡罗投点法产生初值。

步骤1:给定系统参数初值Φ_0

分析2:在给定系统参数初值后，我们的Kalman滤波就可以开始运算了。将给定的参数初值代入方程后，通过滤波算法，可以得到初始的$\{x_{t0}\}$，这得到便是潜在的变量，即也是我们的想要得到的房地产“情绪指数”，但这步得到的只是初始状态下的潜在变量。最终我们需要的得到参数最优时的不可观测变量，那么如何得到最优的参数呢，就考虑到根据后面的EM算法求解。此外，这里Kalman算法我们采用了两步，包括Kalman滤波和Kalman-Rauch平滑；其实本文的观测变量$\{y_t\}$不存在缺失数据，暂且我们也不做预测，所以并不需要Kalman-Rauch平滑这步；但是考虑到投资者主要应用此算法进行预测，所以日后加上这步是不可避免的。

步骤2：初值给定后，开始Kalman算法（包括向前滤波和向后平滑两个过程）通过此步可以估计出不可观测变量$\{x_n\}$。

EM算法开始：

分析1：开始EM算法之前，我们的目标是求得最优的参数估计。所以，这里首先考虑求出完全数据的对数似然函数，即如上文所求；但考虑到对数似然函数中含有不可观测变量，无法直接进行求偏估参，所以我们想到对其取条件期望，然后极大化这个条件期望。

步骤1：求出$\{x_n\},\{y_n\}$的联合概率密度函数，并对其取对数似然函数，然后取条件期望。我们的目标是极大化这个条件似然函数。

分析2：这里我们采用的是显示求解法，因为本文所涉及的概率密度函数均为正态分布的，如上式可见，是比较容易得到显示解的。所以我们将对数似然函数具体表示出来后，分别对每一步取条件期望，因为条件期望具有可加性；得到具体条件似然函数表达式后，分别对每次待估计参数求偏导，得到各个参数的条件期望表示形式。

步骤2：分别对条件似然函数中每个待估计参数求偏导，得到各个参数的条件期望表示形式。

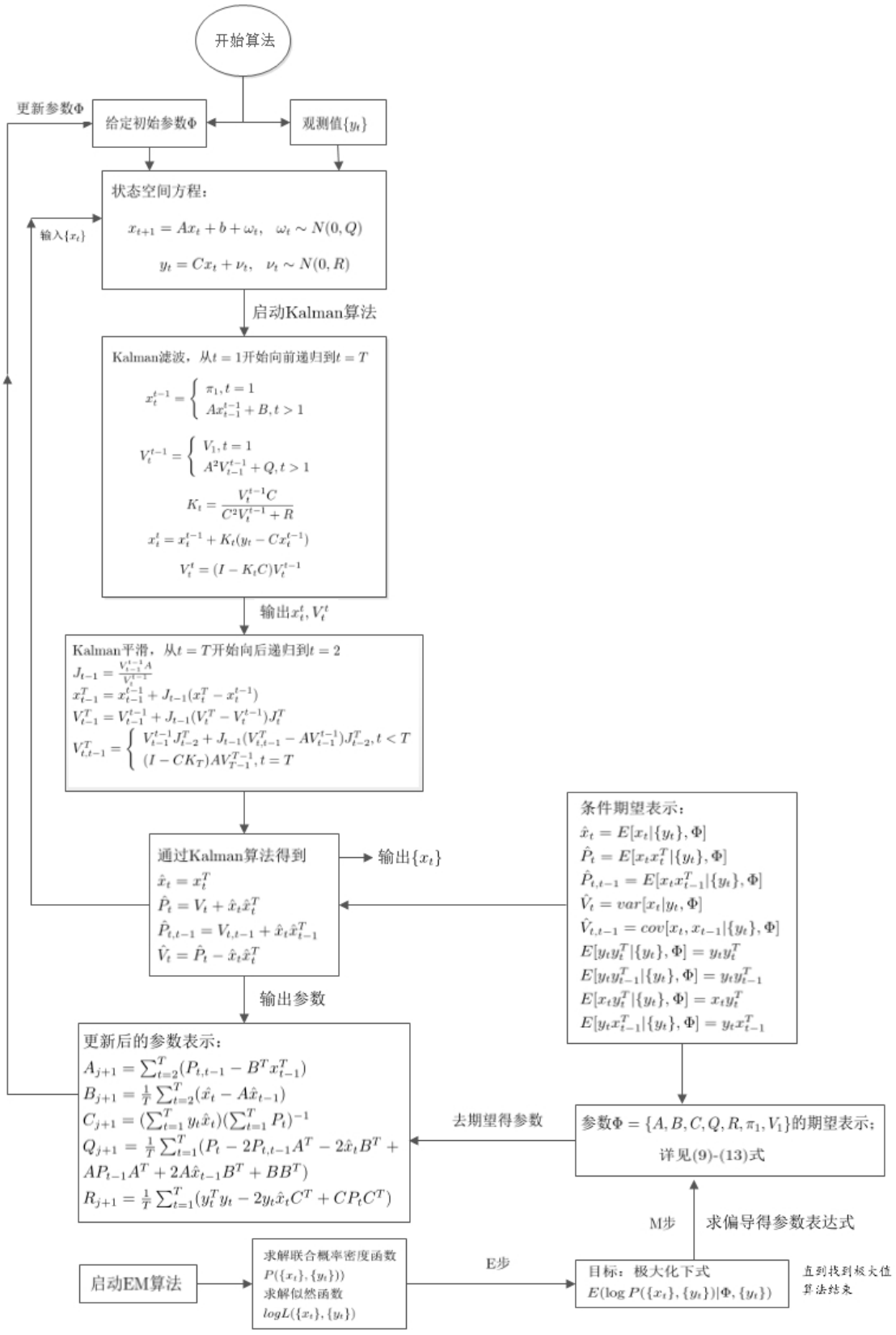
分析3：得到这些参数表达形式后，可以看到这些条件期望都是含$E[x_t x_t|\Phi,\{y_n\}]$,$E[x_t y_t|\Phi,\{y_n\}]$等式的条件期望形式。下面就考虑到根据A.M的结论，将各个参数的条件期望用Kalman滤波得到后的表示形式替换掉。

步骤3：去条件期望将参数由已经求得x_t和可观测的y_t等式表示，即每个参数中均含有观测变量和不可观测变量的信息。

分析4：表示出参数后，最终目标还是估计出这些参数。我们知道，要想得到最优参数估计，需要不断的更新参数，最终使得条件似然函数取得最大值。所以通过交替的使用Kalman算法和EM算法，不断的将这些参数Φ和不可观测变量$\{x_n\}$代入这个循环中，不断的更新不可观测变量$\{x_t\}$和所有待估计参数，最终得到使得条件似然函数收敛的最优值，此时得到的参数估计值和不可观测变量就是我们所求的。

步骤4:不断循环算法，取得使条件似然函数收敛的参数估计值和不可观测变量。

§2.2.5 算法流程图及详解



第三章　实证演练

§3.1　前期准备

§3.1.1　数据的选取及合理性

数据选取：本文选取来自上证指数中房地产板块、钢铁板块、水泥板块、建材板块2015年3月到2016年3月间的全部开盘价236个数据点作为初始数据。

合理性：通过前人研究及相关系数测定，我们知道钢铁产业、水泥产业、建材产业和房地产产业相关性相对最强，所以从这三大行业入手来评测隐藏在其背后的房地产“情绪指数”是合理的。

§3.1.2　数据的预处理及合理性

数据预处理：考虑到股市开盘价的具体数值有很大的波动性，所以在应用数据之前，先进行标准化处理。

合理性：虽然股票数据不存在消除量纲的问题，但是股市近一年波动较大，直接选用开盘价数据会带来较大的误差。经过标准化后，得到的数据部分大于0，部分小于0，平均值为0，这样就可以在一定程度上消除了较大跨度数据所带来的计算及画图问题，还可以有效避免了数据的自身变异问题。

§3.1.3　软件包的选取及合理性

本文选用R软件中的MARSS包来完成本文的算法应用和模型选取，MARSS包是较新的R软件包用来分析多变量时间序列，且MARSS包也选用的是状态空间模型，方便了本文的应用。

由于MARSS里包含众多模型，所以本文采用本文定义的模型进行定义；此外通过多次拟合和运算知道，通过改变模型定义的方式，会逐渐得到较优的结果。所以本文通过多种方式定义模型，并通过应用EM算法和Kalman算法来试图获得最优的参数和不可观测变量的估计。

§3.2 模型选择

§3.2.1 不同方式定义模型结果对比

根据MARSS软件包中模型，我们对本文的模型进行三种不同形式的定义

第一种：

不定义模型：选用MARSS包的自身定义功能，并未对任何参数形式进行定义。

第二种：

自定义模型：根据本文的模型将各个参数按照固定形式进行定义。

第三种：

Kemfit及参数定义模型：根据MARSS中自带的Kemfit模型，即固定部分参数，估计剩余参数；然后将估计得到的参数值代入模型，再次定义一个模型，估计出之前固定的参数值。

结果显示如下：

表 3.1: MARSS软件包三种模型输入方式拟合效果对比

	不定义模型	自定义模型	Kemfit模型	参数定义模型
迭代次数	> 500	155	129	23
极大似然函数	-15.47119	-13.78921	-14.17404	-12.51837
AIC	46.94238	39.57843	36.34807	31.03674
AICc	47.57674	39.94524	36.52123	31.14019

根据上表可以看出，选择第三种方式定义模型是最合理的，所以本文采用第三种方式定义模型，分别对股市的房地产板块、钢铁板块、水泥板块、建材板块进行了模型参数估计，并得到了下列各表，从表中不仅可以看到这些板块数据所构成状态空间模型中的参数值，还可以通过看水平误差看出估计参数结果的好坏。

§3.2.2 四大相关板块状态空间模型参数估计情况

下面四个表格分别为房地产板块状态空间模型参数估计表、钢铁板块状态空间模型参数估计表、水泥板块状态空间模型参数估计表、建材板块状态空间模型参数估计表。

表 3.2: 房地产板块参数估计值

	ML.Est	Std.Err	low.CI	up.CI
Step1	Kemfit模型			
R	0.00399	0.0319	4.62e-07	0.0158
B	-0.00081	0.0239	-4.76e-02	0.0460
Q	0.05846	0.0187	4.21e-02	0.0775
x0	-0.70471	0.2507	-1.20e+00	-0.2133
Step2	参数定义模型			
C	0.99944	0.0511	0.899	1.10

A	0.97096	0.0159	0.940	1.00
V0	0.00282	0.2748	0.236	0.35

从上表中可以看出房地产板块的状态空间模型参数R、B、Q、C、A的估计结果较好，水平误差都低于百分之五；但是$x0, V0$的估计误差在百分之二十五左右，这是可以理解的，因为初值本身就不够稳定，而且我们的目标就是估计R等参数。

表 3.3: 钢铁板块参数估计值

	ML.Est	Std.Err	low.CI	up.CI
Step1	Kemfit模型			
R	0.000573	0.0298	0.00118	0.00677
B	-0.005343	0.0106	-0.02616	0.01547
Q	0.026602	0.0111	0.01998	0.03417
x0	-0.026390	0.1643	-0.34848	0.29570
Step2	参数定义模型			
C	0.99997	0.0480	0.9060	1.094
A	0.99009	0.0107	0.9691	1.011
V0	0.00123	0.1780	0.0985	0.147

从上表中可以看出钢铁板块的状态空间模型参数R、B、Q、C、A的估计结果较好，水平误差均低于5%，但是同样$x0, V0$的估计误差较大。

表 3.4: 水泥板块参数估计值

	ML.Est	Std.Err	low.CI	up.CI
Step1	Kemfit模型			
R	0.000555	0.0317	0.00148	0.00733
B	-0.005496	0.0116	-0.02830	0.01731
Q	0.031413	0.0113	0.02408	0.03973
x0	0.238494	0.1792	-0.11266	0.58965
Step2	参数定义模型			
C	0.99994	0.0476	0.907	1.093
A	0.98642	0.0116	0.964	1.009
V0	0.00145	0.1943	0.117	0.175

从上表中可以看出水泥板块的状态空间模型参数R、B、Q、C、A的估计结果也非常好，水平误差均低于5%，同样$x0, V0$的估计误差较大。

表 3.5: 建材板块参数估计值

	ML.Est	Std.Err	low.CI	up.CI
Step1	Kemfit模型			
R	0.000579	0.0321	0.00151	0.00757
B	-0.004700	0.0120	-0.02813	0.01873
Q	0.033363	0.0115	0.02565	0.04209
x0	0.040169	0.1844	-0.32116	0.40149
Step2	参数定义模型			
C	0.99994	0.0477	0.907	1.093
A	0.98559	0.0120	0.962	1.009
V0	0.00154	0.1991	0.123	0.184

从上表中可以看出建材板块的状态空间模型参数R、B、Q、C、A的估计结果也很好，水平误差均低于5%。

但是可以从上述四表中可以看出一定规律，其中A,C的值多接近1，这就意味着状态空间模型中两个变量前后时刻变化波动较小。同时可以看出每个参数的值都相对较小，间接的可以分析出，不可观测变量和可观测变量两者变化趋势相似。

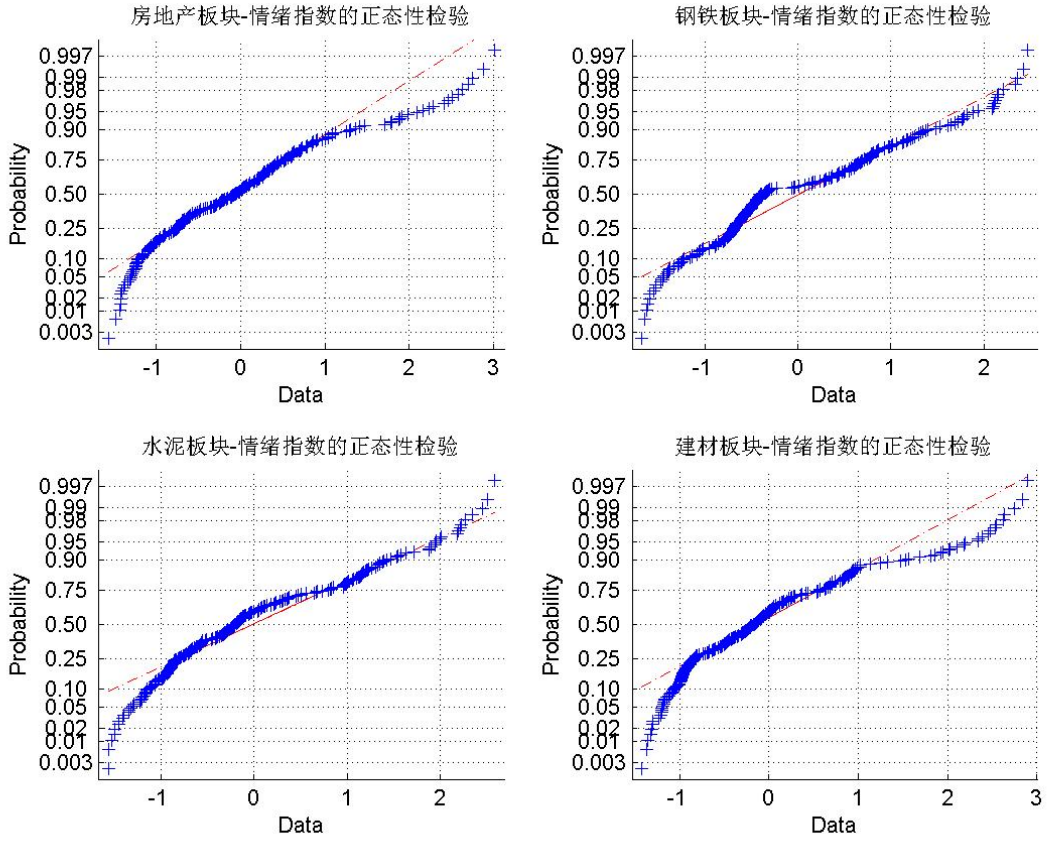
下表是上述四大板块参数估计的综合评价表，表中给出了迭代次数和极大似然估计值，其中AIC和AICc是评价模型好坏的标准，数值越小，说明模型越好。

表 3.6: 房地产、钢铁、水泥、建材板块参数估计评价表

	迭代次数	极大似然估计值	AIC	AICc
房地产	129	-14.17404	36.34807	36.52123
	23	-12.51837	31.03674	31.14019
钢铁	>500	88.3002	-168.6004	-168.4272
	23	88.70818	-171.4164	-171.3129
水泥	>500	69.57029	-131.1406	-130.9674
	23	70.23411	-134.4682	-134.3648
建材	>500	62.53379	-117.0676	-116.8944
	23	63.23841	-120.4768	-120.3734

§3.2.3 房地产“情绪指数”总体评价

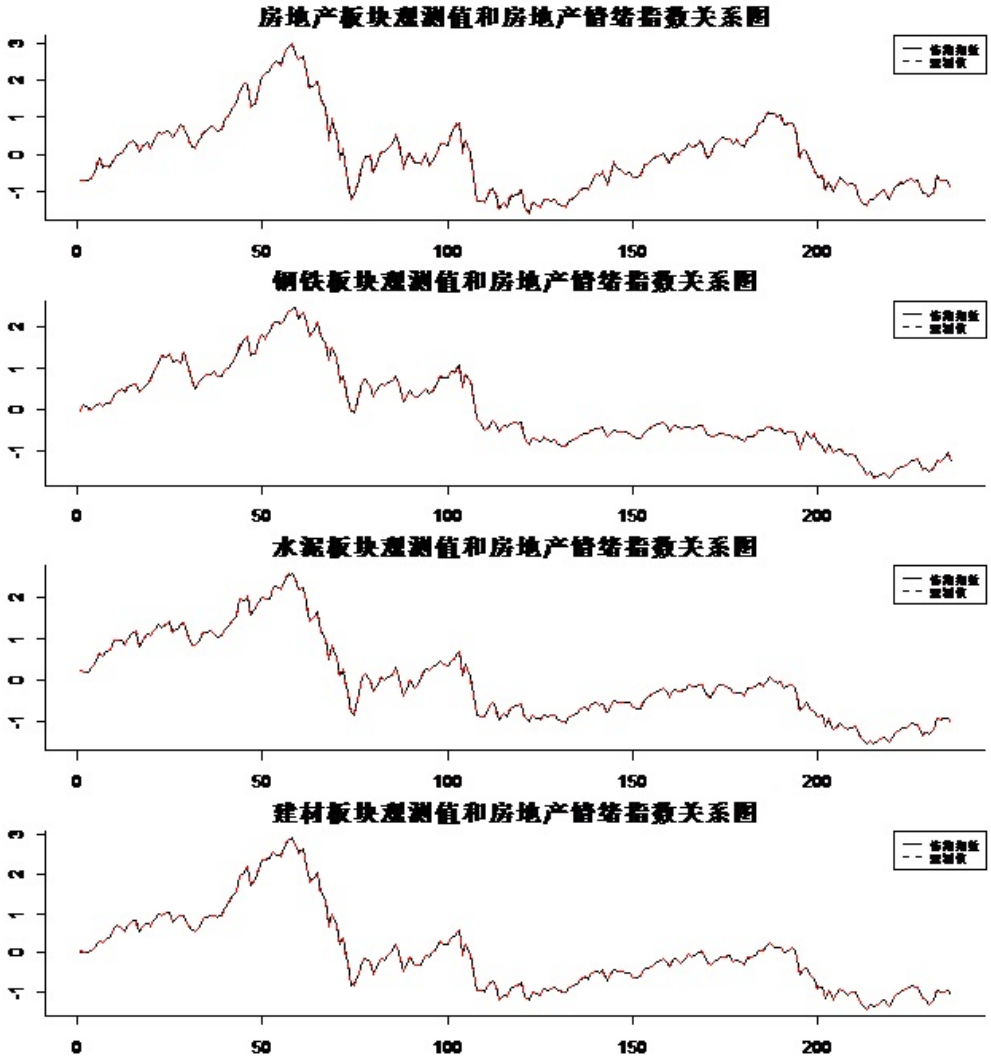
下面我们考查下这四大板块数据评测出的房地产“情绪指数”变量是否服从正态分布，我们画如下四图：



从上图可以看出，这四大板块评测出的房地产“情绪指数”并不严格地服从正态分布，所以我们需要进一步研究房地产“情绪指数”分布的性质。

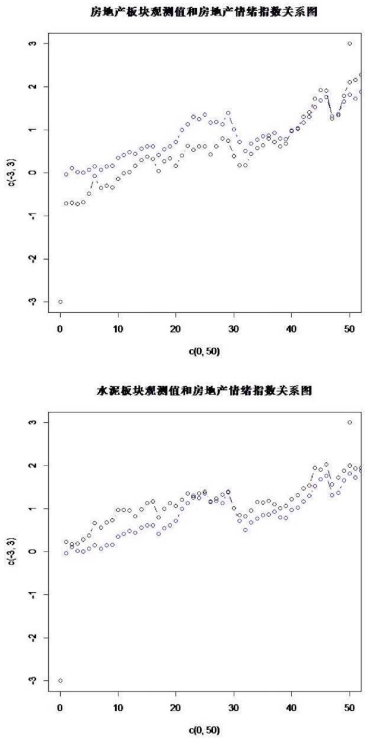
下面考虑想进一步直观的看出观测变量和不可观测变量的关系，所以我们画出下四图：

分别为房地产观测数据和不可观测数据观测图、钢铁观测数据和不可观测数据观测图、水泥观测数据和不可观测数据观测图、建材观测数据和不可观测数据观测图。



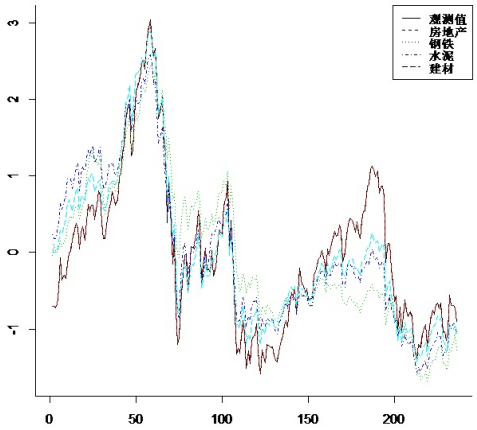
从上四表可以清晰的看出不可观测的情绪指数线基本与可观测的数据线重合，说明隐藏在数据里的信息基本可以由数据解释。房地产“情绪指数”波动情况基本与各相关板块股市指数波动形态一致，也就说在以后的应用当中，我们可以通过直接分析观测数据，就可以在一定程度上解释房地产的“情绪指数”。

但是我们想要更加清晰的看出潜在变量和实际观测变量之间的差异，即更加清晰的看出从房地产相关产业的股市数据中得到的房地产“情绪指数”。我们画下图展示从0-50个数据点房地产板块观测值、水泥板块观测值和房地产“情绪指数”之间的差异。



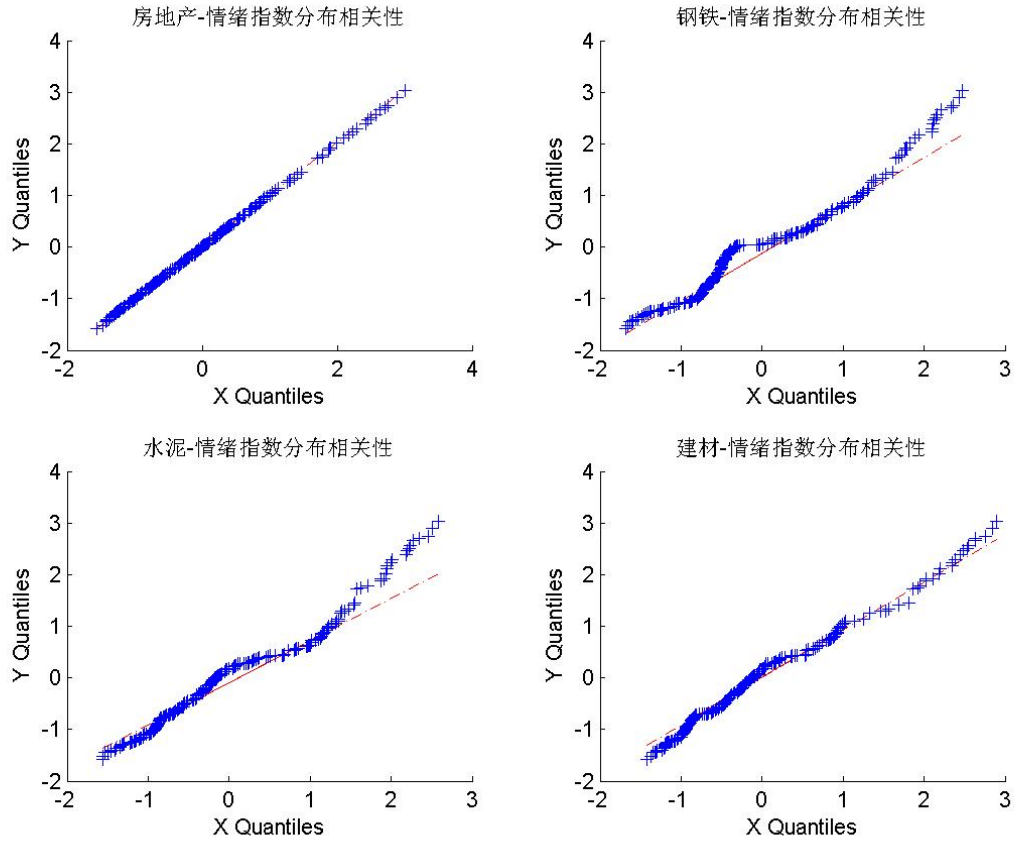
从上两图可以看出相关板块股市观测值与房地产“情绪指数”存在一定差异，如房地产板块观测值在40 个观测点之前一直波动与“情绪指数”一致，其数值一直小于“情绪指数”；但在40-50个观测点之间，“情绪指数”变化减弱，两者值趋于重合状态。但总体来看相关板块的股市观测值与房地产“情绪指数”变化趋势基本一致。

通过上面几个图我们能够看出我们所要评价的房地产“情绪指数”在相关板块中与其相应的观测变量之间的关系，但为了更好且直观的看出这些由相关板块观测变量得到的房地产“情绪指数”与房地产观测变量之间的关系，我们画出下图：



从上图可以看出，房地产板块的观测值与相关板块测得的房地产“情绪指数”变化趋势相对一致，都出现了三个峰值，且在第50个数据点的时候达到最大值。分析得出如下结论隐藏在其他板块中的房地产“情绪指数”与房地产自身的变化趋势较为一致，在一定程度上通过观测房地产板块的指数可以直接评价房地产“情绪指数”。

但是考虑到也许这些趋势是由于股市自身的趋势所引起，所以我们分别将房地产板块观测值与其他四大相关板块测得的房地产“情绪指数”画QQ图，考查他们是否来自同一个分布。见下图：



从上图可知，房地产、建材观测变量和房地产“情绪指数”变量可能来自同一分布，但是由于数据点不是很多，无法直接定论。至于钢铁、水泥板块的观测变量和房地产“情绪指数”变量之间可能存在一定关系，但是从观测数据来看，并不来自同一分布。

所以从这些得到图表当中很难分析出房地产“情绪指数”的性质，也很难通过房地产相关板块的观测变量与其进行分析对比。那么如何利用我们估计出的房地产“情绪指数”来对未来的房地产市场价格变化趋势进行预测，下面我们就考虑建立回归方程，考查下房地产观测变量和相关四大板块测得的房地产“情绪指数”之间的关系。

§3.2.4 建立房地产“情绪指数”回归方程

下面首先将房地产观测变量和相关四大板块测得的房地产“情绪指数”建立回归方程。

设如下

$$F = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \tag{3.1}$$

通过matlab进行拟合得到下面结果

表 3.7: 拟合参数值

	参数值	置信区间
β_1	1.0110	(0.9989,1.0232)
β_2	0.0004	(-0.0101,0.0109)
β_3	-0.0167	(-0.0449,0.0115)
β_4	0.0083	(-0.0141,0.0307)

其中$r^2 = 1, F = 1.9094 \times 10^5, p = 0$，$r^2$越接近1说明回归方程显著，所以此回归方程显著，$p < 0.05$说明回归模型成立。

所以我们可以得到如下的回归方程：

$$F_{sentiment} = 1.0110 X_{fdc} + 0.0004 X_{gt} - 0.0167 X_{sn} + 0.0083 X_{jc} \tag{3.2}$$

这样，当我们通过四大相关板块测得每个对应的房地产“情绪指数”后，我们可以整体的来评价房地产“情绪指数”。

第四章　总结与展望

§4.1　研究总结

本文从房地产“情绪指数”定义介绍开始，到算法流程，再到实证分析，再到应用评价，较为全面且具体的阐述了，一种评测房地产“情绪指数”的方法。

首先，从房地产“情绪指数”定义上讲，本文是根据市场情绪指数角度进行定义，目前国内外也并没有专门的文献对这一概念进行定义。但是本文主要是想借助这个“情绪指数”的定义给决定房地产价格潜在变化趋势的房地产市场参与者情绪指数一个抽象化的概念，这种“情绪指数”可以理解为一种具有波动性的变化趋势，也就是潜在房地产价格背后的一种决定性的因素。不过这个定义正确与否有待考证，也请多方学者给出建议。

从本文算法来评价，本文采用的是EM算法来估计状态空间方程的参数，但是核心算法和亮点是应用了Kalman滤波算法来估计不可观测变量，两个方法在更新过程中交替使用，能够有效的同时估计出状态空间模型的参数和不可观测变量，这种方法非常实用且便捷，也就是说对任何一种可以观测的变量，获得一定数据后，我们都可以评测出其潜在的变量，同时还可以建立相应的状态空间模型，并估计出方程的参数值。所以这种算法在未来的应用中会具有很大的前景，当然本文给出的算法虽然流程比较详尽，但是根据应用的不同，可能需要不同程度的进行改进。比如，在应用Kalman 算法时，如果需要作预测或是部分观测数据缺失，需要考虑加入Kalman-Rauch平滑进行向前滤波这步，这样才可以有效的估计出不可观测变量。

从实证分析的角度分析，本文主要是将上述算法应用到我们关心的房地产市场中，来通过评测房地产“情绪指数”来间接的评价房地产市场和预测房地产价格未来的变化趋势，但是考虑到主要是想验证算法的可行性和评价功能的有效性，本文并没有进行预测房地产价格变化趋势的相关分析，所以进一步的预测就需要市场参与者根据自己的需要来应用本文的方法来进行预测和分析。

总体来说，本文能够较好的利用状态空间模型，应用其良好的性质，估计出的房地产“情绪指数”。虽然可以看出，我们得到的房地产“情绪指数”信息基本与观测信息基本吻合，所以如果要粗略的对房地产市场就行评价，仅需要就相关板块的观测值进行评价或预测即可。所以之后的研究可能是考虑建立一个更为合理的模型，能够更由针对性的评测出潜在变量。

§4.2　研究展望

§4.2.1　预测应用

本文的出发点是为了能够评测出房地产“情绪指数”从而方便投资者自行进行预测，并能够做出合理的投资策略。所以下一步，我们研究的重点将放在如何预测房地产价格未来的变化趋势，也就是未来的“情绪指数”。

采用的预测方式可以有以下两种:

第一种：通过一定的统计方法预测出房地产相关板块观测变量未来一段时间内的走

势后，应用本文算法，对潜在变量进行估计，这样我们就可以同时得到跟观测变量等时间段的不可观测变量的预测值。

第二种：将房地产相关板块观测变量的样本待预测的部分设为缺失值，应用Kalman-Rauch方法，估计出不可观测变量，这样我们得到的不可观测变量就会含有预测的那个部分，然后我们再对预测的那部分结果进行分析。

§4.2.2 多变量合一应用

本文的应用考虑到各相关板块数据均来自股票市场，可能存在一定相关性，所以都是分别对各个板块进行滤波，最终得到四项房地产“情绪指数”。这样我们就需要分别研究这四项房地产“情绪指数”，或是将其进行回归分析，这就增加了一定了分析难度。

所以下面我们要考虑的是，可否通过一定方法消除或是克服观测变量之间相关性的影响，将多个观测变量合为一个多维变量，代入状态空间模型，得到一个综合的房地产“情绪指数”。

$$\begin{cases} Y_{fdc} = CX + v \\ Y_{gt} = CX + v \\ Y_{sn} = CX + v \\ Y_{jc} = CX + v. \end{cases} \Rightarrow Y = CX + v$$

§4.2.3 ARMA模型应用

考虑到本文得到的潜在变量与观测变量来自同一个分布，这样得到不可观测变量研究意义减弱，所以也许可以应用其他模型来进一步研究不可观测变量。

见如下ARMA观测方程：

$$Y_t^1 = C^1 X_t + Z^1 Y_{t-1} + U_1 + \varepsilon_t^1, \quad \varepsilon_t^1 \sim N(0, R^1)$$

$$Y_t^2 = C^2 X_t + Z^2 Y_{t-1} + U_2 + \varepsilon_t^2, \quad \varepsilon_t^2 \sim N(0, R^2)$$

$$Y_t^3 = C^3 X_t + Z^3 Y_{t-1} + U_3 + \varepsilon_t^3, \quad \varepsilon_t^3 \sim N(0, R^3)$$

.....

考虑建立上式的观测方程，因为我们选用的观测变量来自股市，如果这个时间序列是非平稳的，那这个变量本身一定具有自相关性，应表示成自相关函数，同时这个可观测变量又受着隐藏在其中的不可观测变量X控制，所以建立ARMA过程方程是可行的。

则我们建立的状态空间模型将可以表示如下:

$$x_{t+1} = Ax_t + B + w_t, \quad w_t \sim N(0, Q)$$

$$y_t = Cx_t + Zy_{t-1} + u + v_t, \quad v_t \sim N(0, R)$$

日后的研究将主要围绕上述三个研究方向进行，将对房地产“情绪指数”给与更加全面具体切合实际的评价。同时也会将此方法应用在更多更广泛的领域。

参考文献

[1] 关海玲,当前房地产泡沫的科学评估与破灭防控[D],社会科学家,2015(6),73-77

[2] 赵晓,房地产泡沫破灭的两种场景[D],2013(36),5-5

[3] 彭湘,卿斌,房地产泡沫破灭当在何时,经济视野[D],2013(15),396-397

[4] 王南,房地产泡沫破灭三个时间节点,经营管理者[D],2011(8),106-106

[5] 易宪容,房地产泡沫破灭为时不远,中国经济信息[D],2011(18),15-15

[6] 郭红领,中国房地产信心指数研究[D], 哈尔滨工业大学博士论文,2007

[7] 张舒涵,基于状态空间模型对我国房地产泡沫的研究[J],区域金融研究,2015,507(2),78-81

[8] 杨刚,王洪卫,金融文持对上海房地产市场发展的影响研究－－基于状态空问模型的实证检验[J],现代管理科学,2012(3),3-11

[9] 周文剑,基于状态空间模型的浙江房地产价格泡沫的实证检验[J],现代经济信息,2013(7),270-270

[10] 赵安平,范衍铭,基于卡尔曼滤波方法的房价泡沫测算－－以北京市场为例[J], 财贸研究,2011(1),59-65

[11] 庞培俊,李江,基于行为金融学的资产价格波动分析[J],中国集体经济,2008(21),99-100

[12] 王朝宗,资产价格波动原因探讨－－基于行为金融学理论的分析,总裁,2009(10),74-75

[13] Elizabeth Eli Holmes,Derivation of an EM algorithm for constrained and unconstrained multivariate autoregressive state-space (MARSS) models,R,2014,MARSSpackage

[14] Zhong Lusheng,Robust maximum-likelihood parameter estimation of stochastic state-space systems based on EM algorithm,PROGRESS IN NATURAL SCIENCE,2007,17(9),1095-1103

[15] Ryan Dong Chen, Christopher Gan, Baiding Hu,David A. Cohen,An Empirical Analysis of House Price Bubble: A Case Study of Beijing Housing Market,Research in Applied Economics,2013,1,77-97

[16] Goodman, A. C.,Thibodeau, T. G. (2008). Where are the Speculative Bubbles in US Housing Market? Journal of Housing Economics,2007.12(17), 117-137.

[17] ZHANG Xiao-xia,WU Chong-feng,The Bubbles of China Stock Market Based on Return Decomposition and Cumulative Return,JOURNAL OF DONGHUA UNIVERSITY(ENGLISH EDITION),2006,23(4),111-115

[18] Hou, Y. (2009). House Price Bubbles in Beijing and Shanghai? [J]A Multi-Indicator Analysis.International Journal of Housing Markets and Analysis, 3(1), 17-37.

[19] Zhong Lusheng,Robust maximum-likelihood state-space systems parameter estimation of stochastic based on EM algorithm[J],National Laboratory of ladustrial Control Technology,2007

[20] XIAO Qin and TAN Gee Kwang,Signal Extraction with Kalman Filter: A Study of the Hong Kong Property Price Bubbles[J],Randolph Economic Growth Centre Division of Economics,2006

[21] C.K. Chui • G. Chen,Kalman Filtering with Real-Time Applications[D]

[22] Ghahramani, Z. and G. E. Hinton(1996) Parameter estimation for linear dynamical systems.Technical report CRG-TR-96-2.

[23] Holmes, E. E. and W. F. Fagan. (2002) Validating population viability analysis for corrupted data sets. Ecology 83: 2379-2386.

[24] McLachlan, G. M. (1996) The EM algorithm and extensions.[J] Wiley, USA.

[25] Maybeck, P. S. (1979) Stochastic models, estimation and control. Volume 1.[D] Academic Press, New York, USA.

[26] Shumway, R. H. and Stoffer, D. S. (1982) An approach to time series smoothing and forecasting using the EM algorithm.[D] Journal of Time Series Analysis 3(4):253-264

[27] R. E. KALMAN,A New Approach to Linear Filtering and Prediction Problems ,Research Institute for Advanced Study, Baltimore, Md.Transactions of the ASME – Journal of Basic Engineering, 82 (Series D): 35-45. Copyright ? 1960 by ASME

攻读硕士学位期间取得的研究成果

一、已发表（包括已接受待发表）的论文，以及已投稿、或已成文打算投稿、或拟成文投稿的论文情况（只填写与学位论文内容相关的部分）：

序号	作者（全体作者，按顺序排列）	题　　目	发表或投稿刊物名称、级别	发表的卷期、年月、页码	相当于学位论文的哪一部分（章、节）	被索引收录情况
1、	蒋文江、李彩雯、刘鹏懿	度量股票市场情绪指数的新方法–基于状态空间模型	《海南师范大学学报（自然科学版）》、省级	2016年29卷第3期	第二章	
*	*	*	*	*	*	*

注：在”发表的卷期、年月、页码”栏：
1如果论文已发表，请填写发表的卷期、年月、页码；
2如果论文已被接受，填写将要发表的卷期、年月；
3以上都不是，请据实填写”已投稿”，”拟投稿”。
不够请另加页。

二、与学位内容相关的其它成果（包括专利、著作、获奖项目等）
1、*****。
2、*****。
3、*****。

致　　谢

在此，首先我要感谢我的导师蒋文江教授三年来以来对我无私的教诲。在整个论文撰写过程中，从论文选题，构思到最后定稿的各个环节，蒋老师都细心的指导我进行修改，使我得以最终完成这篇毕业论文。在学习中，蒋老师严谨的治学态度、丰富渊博的知识、敏锐的学术思维、精益求精的工作态度以及侮人不倦的师者风范是我终生学习的楷模。同时也要感谢三年来给予我帮助、关怀的数学与统计学院的老师们和领导们，感谢同窗三年给予我帮助的同学们。最后，我要向百忙之中抽时间对本文进行审阅、评议和参与本人论文答辩的各位老师表示感谢。