学校编码：10384　　分类号_____密级_____
学号：15420131152030　　UDC _____

厦门大学

硕 士 学 位 论 文

高维时变网络的分析方法研究

Research of High-Dimensional Time-Varying Network Analysis Method

张 声 威

指导教师姓名：方匡南 教授
专 业 名 称：统 计 学
论文提交日期：2015 年 12 月
论文答辩时间：2016 年 4 月
学位授予日期：2016 年 月

答辩委员会主席：_____
评 阅 人：_____

2016 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　月　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　月　日

摘要

网络分析尤其是相关网络分析方法一直以来应用于多领域的研究当中，在大数据浪潮不断革新数据分析方法的背景下，网络分析也面临着高维数据的处理问题。基于网络结构稀疏性的假定，许多相关研究提出关于高维网络数据的网络结构估计方法，这类方法主要对固定时期内的静态网络结构进行估计。然而，由于大量个体以及不确定影响因素的参与，网络个体间的关联结构也充斥着复杂性，结点间的关联结构也有可能随着时间的推移而发生动态的变化，仅凭借传统的静态网络分析并不能较好地刻画出网络结构的演化特征。

在本文的研究中，在详细论述相关网络理论、马尔科夫随机场、伊辛模型等基础理论框架的基础上，详细定义了动态的时变网络结构，针对高维二值型的网络数据提出了二值型时变网络结构估计问题的解决思路，将二值型时变网络的估计问题转化为权重调整的 L1 正则化下的稀疏逻辑回归，并应用投影梯度法进行最优化求解。

最后，文章还进行了模拟研究，并在实证研究部分呈现了时变网络分析在股票市场分析领域的应用，不但呈现出了个股关联结构随着时间推移呈现出的动态演化特征，并进一步利用时变网络对股票板块之间的联动效应进行分析。最后在建立特定板块内个股的时变网络的基础上，甄选出了收益率优于板块的绩优投资组合。

关键词：高维二值数据；时变网络；股票网络分析

Abstract

Related network analysis has been applied in many research fields. In the background where big data brings a lot innovation in data analysis method, the problem of modeling relational structure for high-dimensional data deserve further investigation in the field of network analysis. There is a rich literature in estimating the relational structure from high-dimensional data by assuming the sparsity of network structure, most of which focused on the time-invariant structure revealed through a certain time series of entity attributes. However, the network system operate with much complexity, the relational structure may be dynamic and change across different time points, which is invisible simply by including the network observations in traditional time-invariant network context.

In the study of this paper, theories of related network, Markov Random Field and Ising model is firstly expounded in section 2. In section 3, this paper defines the dynamic time-varying network in detail and introduce a reweighted l1-regulized logistic regression method to estimate the dynamic interactions between entities based on a series of binary expression values, where the projection gradient method is employed to get optimization solution.

Finally, the simulation studies in discussed in section 4. The empirical study in section 5 study the dynamic linkage between stocks by estimating the time-varying network of stock movements, the evolution characteristics of the time-varying stock network are also presented graphically.

Keywords: High-dimensional binary data; time-varying network; stock network analysis

目录

Contents

第一章　绪论

网络的发展一直伴随着人类乃至人类社会的演化进步，从我们身体中的大脑网络到多样的新陈代谢网络，从贯穿城市的水、电力网络再到连接地点的公共交通网络，从各种政治、经济、军事网络再到平日里的社交网络，从因特网这一由互相通信的计算机连接组成的全球网络再到万维网这一连接全球资料资源的统一空间，网络在我们所处世界的不同维度中主导着个体间的信息传播过程，有力地推动着社会的发展与时代的变迁。

1736 年，德国数学家欧拉在解决哥尼斯堡七桥问题时首次引用网络的观点对客观世界进行描述。数据是一组记录客观事物的可鉴别的符号。数据经过处理仍然是数据，只有经过解释，数据才有意义，才成为信息，在历经了数次信息化浪潮不断革新的今天，网络作为承载庞大数据流抑或信息流的载体，已然根植于我们生活的各个层面，我们也迫切地去不断关注身边的网络，人们愈加重视网络结构在信息传播过程中所起的重要作用。人们也更加积极地在社会活动中搭建起联通彼此的社会网络，摆脱了单向、不同步、分隔的信息传播机制，转而以协同作用的方式对信息进行生产与共享。

随着对网络研究的深入，人们在这一研究领域内倾注了无限的热情与想象，在过去数年中，与网络分析相关的概念不断推陈出新，丰富了网络分析的内涵并延伸着网络分析领域的外延，许多概念和度量方法被提出，用于描述网络的结构特性。复杂网络分析(complex network analysis)就一直是当前机器学习和数据挖掘研究的热点方向，人们越来越将注意力放在利用复杂网络对现实网络系统进行行为分析的研究上，各种网络分析模型的发展与应用囊括了不同领域，不仅仅是在计算机领域，数学、社会学、生物学甚至是经济金融领域的研究都越来越依赖于对复杂网络的建模与分析。

在许多自然、社会、信息科学的问题中，常常需要对一个复杂依赖网络(complex dependency network)中的大规模随机变量进行分析，复杂依赖网络又可称为相关网络或者是相依赖网络(related network)，这也是复杂网络分析研究领域的基本框架，例如针对染色体组中基因表达的网络分析，以及社交网络中不同个体的行为的网络分析等。本文的研究对象即为相关网络(related network)，在后文

中所提到的网络若没有特别说明即为相关网络，相关网络的具体概念将在后文进行阐述。

网络是由网络内的个体成员(结点)及成员的关联组成的，在具体应用中，我们常用图模型(graph model)来对网络系统进行描述，网络中的个体即对应图模型中的结点，网络中的个体关联即对应图模型中结点间的连线，根据个体关联的不同相依性又可以分为有向图模型和无向图模型。假定有图$G=(V,E)$，V表示网络中结点集合，E 是网络中结点之间的关联集合。 结点$u \in V$代表着网络中的一个个体，在基因网络中，u 作为一个基因存在，在股票网络中，u 是一只股票，在社交网络中，u 则为个人的标识。边$(u,v) \in E$ 则代表结点u 与结点v的关联，在不同类型的网络系统中有着不同的实际意义。

我们知道高维的复杂网络数据包含了大量的随机变量指标，当我们利用图模型去描述这些随机变量之间的网络关系时，就涉及到了概率图模型的概念。概率图模型是图论和概率论结合的产物，它作为一个囊括多变量分析且变量关系可视化的建模工具，主要包括两个大方向：无向图模型和有向图模型。无向图模型所蕴含的网络结构又称为马尔科夫网络(Markov networks)或者马尔科夫随机场(Markov random fields)，与之相关的应用有很多，典型的如基于马尔科夫随机场的图像处理与图像分割方法，也有和机器学习领域相结合进行模型参数估计的结构化学习方法。本文的研究对象相关网络属于无向图的范畴，因此本文将研究的重点放在无向图模型上。有向图又称贝叶斯网络(Bayes networks)，有向图的应用也很广，在此不进行深入展开。

在“大数据”概念不断被强调的时代背景下，很多网络数据存在着高维的特征，具体表现为样本数据小于甚至远远小于变量的维数，在这一情形下，传统的网络分析方法并不能很好地适应该种类型的数据，从金融市场、生物医学、政府职能部门、公众媒体乃至社交网络，大规模的数据资源已经成为业界与学界面临的重要机遇与挑战，也是复杂网络分析亟需考虑的关键问题。

经过近些年的研究积累，学界已经解决了稀疏的高维复杂网络数据的图结构估计问题，但是大部分的研究主要还是将视角放在反映固定时期内各网络节结点关联结构的静态网络分析上。静态网络分析仅提供了一个相对固定的视角，这与现实生活中各类网络内发生的真实情况并不相同，由于大量个体的参与，网络数

据随时间推移可能产生了巨大的、持续不断的变化，与其相对应的拓扑结构也因此随时间的推进而不断演化，对于一组有待分析的网络数据来说，它极有可能是高维的、动态的、充斥异质性元素和潜在噪声、不完整甚至是不可观测的，这些特质无疑给网络的理解和分析增加了复杂性，显然仅凭传统图论中的静态网络定义和模型并不能较好刻画出复杂网络的动态时空分布特质。

考虑下面这两个实际问题：

基因表达调控网络(gene regulatory networks)：基因表达调控网络一直是生物信息学研究的热点问题，生物系统的运行是一个复杂的动态过程，包括了一个基因的表达受其他基因的影响，而这个基因又会影响其他基因的表达，这种相互影响、相互制约关系构成了复杂的基因表达调控网络。基因调控网络通过分析基因表达数据，将孤立的基因水平，蛋白质水平的各种相互作用关系整合起来，通过构建合适的基因调控网络拓扑结构来模拟生物系统的调控机理。Luscombe(2004)等人指出，在不同的时间点，基因表达的调控网络具有整体依赖性，相较于稳定的结构，基因调控网络更有可能随着时间推移发生系统性的改变。对于高维的基因表达数据，有如下问题：给定一组包含 p 个基因表达水平，n 个时间点($n<<p$)，共计 $n*p$ 个观测的表达谱(microarray)数据，如何捕捉不同基因间调控机理及相应的时变动态特征?

股票市场分析：股票市场的交易过程也蕴含着大量信息，以一只股票的交易来说，从开盘价到收盘价，从成交量到换手率、从基本的收益率计算到复杂的技术指标分析，不同股票的多时点数据也就汇集成为一个庞大的数据集，足以支撑一个动态的股票网络分析系统，众多的信息也股票给投资者以多样的应用形式。例如，假设我们关注的某只股票的二级市场价格发生了明显的涨跌，我们也许会想要知道是否存在着其余股票和这只股票同时发生了异动？如果股票之间的这种关联性并没有在我们整个观测周期内持续，那么在哪些时间点这种关联关系能够“短暂”地被我们的分析体系所观测到？股票之间的联动关系强调了网络的拓扑结构，而联动关系随时间而发生的演化特征则强调了网络拓扑结构的动态性。同样，可以有和基因表达调控网络类似的问题：给定一组包含 p 个股票价格涨跌，n 个时间点(交易日)($n<<p$)，共计 $n*p$ 个观测的股票价格的网络数据，如何捕捉不同股票之间的价格联动机制及相应的时变动态特征？

上述的两个实际问题进一步说明了将研究视角由静态的网络结构上升至对动态的(时变)网络进行研究的巨大价值，动态的数据常常包含多种关系和多种连接，这就很容易把时变网络的动态特征与多维数据挖掘联系起来展开更加深入的研究。

在上述实际问题当中，每个基因和股票都可以视为网络中独立的个体(entity or node)，个体之间的潜在关系结构可能随着时间的推移而不断变化，我们要想深入了解某一网络中这种动态演化的机制，首先存在的问题就是不同时间点上的网络数据快照(snap shots)并不能直接呈现出个体结点之间的联动情况，对于一个生化复杂网络而言，现有的技术并不容许我们直接通过实验来定义考虑不同时点特异性的动态网络结构，我们可以直接获取的也许只有不同时间下的微阵列表达数据作为基因个体的时间序列测量指标，同样对于股票网络来说，我们可直接获取的是个股的交易信息，而不能获知股票间的联动情况。

本文的核心内容将从动态地进行网络分析的角度出发，研究如何从一组具有马尔科夫随机场理论内伊辛模型特性的二值网络时间序列数据出发，利用其中所蕴含的结点属性，建立一个随着时间的推移不断演化的动态的无向网络拓扑结构。结合相关的研究，本文将这种动态的网络定义为二值型时变网络(time-varying networks with binary data)。

基于这一基本思路，本文的研究内容主要由下列几个部分展开：

第一章引言介绍了本文的研究背景并简要介绍了网络分析这一研究主题，根据实际应用的需要明确了本文的研究方向即时变网络分析，明确了研究主线。

第二章基于现有的研究结果，论述了相关网络的主要理论框架，阐述了相关网络的估计以及高维问题下的处理方法，介绍了马尔科夫随机场与伊辛模型并强调了应用二值型网络数据进行网络分析的重要性。

第三章则在第二章所奠定的理论基础上，针对二值型时变网络，给出了规范化的详细定义，从理论上明确了研究问题。将二值型时变网络的估计转化为最优化求解问题，利用 L1 惩罚下的权重调整 Logistic 回归进行求解，并提出相应的最优化求解算法和参数调整方式。

第四章针对二值型时变网络的估计问题进行模拟研究，定义了估计结果优劣的衡量指标，按照二值型时变网络的理论框架生成随机网络数据集，图形化地呈