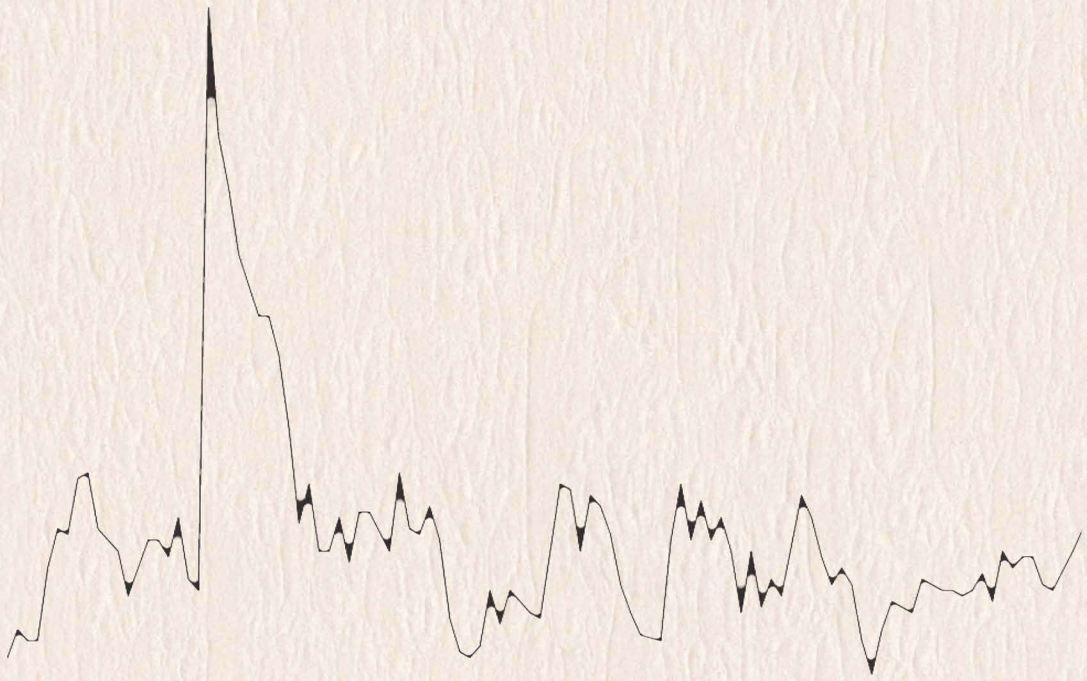
# 两类时间序列模型的异常值检测研究

尚华　著



中国书籍出版社
China Book Press

## 作者简介

尚华，1981年生于河南省辉县市。2003年于河南大学数学系应用数学专业本科毕业，2006年于重庆大学数理学院基础数学专业硕士研究生毕业，2016于首都经济贸易大学统计学院统计学专业博士研究生毕业。2006—2013在重庆邮电大学数理学院任教。现为华北科技学院理学院教师。主要研究方向为：时间序列、数据挖掘等。

ISBN 978-7-5068-6391-9
9 787506 863919 >
定价：25.00元

# 两类时间序列模型的异常值检测研究

尚华 著


中国书籍出版社
China Book Press

图书在版编目（CIP）数据

两类时间序列模型的异常值检测研究 / 尚华著. — 北京：中国书籍出版社，2017.8

ISBN 978-7-5068-6391-9

Ⅰ. ①两… Ⅱ. ①尚… Ⅲ. ①时间序列分析—异常值—研究 Ⅳ. ①O211.61

中国版本图书馆CIP数据核字（2017）第200833号

两类时间序列模型的异常值检测研究

尚华 著

特约编辑 马立平
责任编辑 李雯璐
责任印制 孙马飞 马 芝
封面设计 东方美迪
出版发行 中国书籍出版社
地 址 北京市丰台区三路居路97号（邮编：100073）
电 话 （010）52257143（总编室） （010）52257140（发行部）
电子邮箱 eo@chinabp.com.cn
经 销 全国新华书店
印 刷 北京睿和名扬印刷有限公司
开 本 787毫米×1092毫米 1/32
字 数 140千字
印 张 7
版 次 2017年8月第1版 2017年8月第1次印刷
书 号 ISBN 978-7-5068-6391-9
定 价 25.00元

# 前 言

时间序列模型自出现以来，就广泛应用于商业、经济及科学观测等各个领域。在有的场景下，比如信用卡诈骗、医学 MRI 或 PET 扫描图分析等，我们更感兴趣的是时间序列数据中的异常值，以便发现可能存在的诈骗、病变等。整值时间序列和多元时间序列是时间序列分析的重要组成部分，关于这两类时间序列模型异常值检测的文献相对较少，编著一本内容相对新颖且具有一定的理论意义和应用价值的有关这两类时间序列模型异常值检测的研究专著，是作者近年的心愿。今日本书付梓，终可聊以慰藉，但因水平及能力所限，殷切希望各位读者批评指正。

本书为华北科技学院理学院教师尚华近年来的一些研究及学习心得，希可为从事时间序列异常值检测的研究和应

用人员提供参考。本书不图在阐述前人的理论和方法方面求多求全，而力求内容能够新颖和切合实用。

本书分为四部分。第一部分包含第一章和第二章，是绪论和时间序列异常值检验的基础理论部分，为读者提供了必要的理论基础。第二部分包含第三章，主要是受加性和新息异常值污染的 INAR(1) 模型参数的 CLS 估计，并给出估计的理论推导和证明。第三部分包含第四章和第五章，主要是基于贝叶斯方法对 INAR(1) 模型和 VAR 模型分别进行异常值检验研究，并给出相应的模拟实验和实例分析，这部分内容是本书的重点和难点部分。第四部分包含第六章，是本书的总结与展望。

在本书的研究和形成过程中，曾得到中国科学院数学与系统科学研究所陈敏研究员、首都经济贸易大学统计学院纪宏教授两位导师的悉心指导和帮助，是他们把我引入统计学领域；在本书的模拟实验方面曾得到首都经济贸易大学张贝贝老师的指导和帮助。我校单位的同事谭立云教授、刘海生教授、杨文光副教授等也给予作者不少帮助，在此一并向他们表示衷心的感谢！

本书受中央高校基本科研业务费（3142016023）、华北科技学院概率论与数理统计重点学科项目的资助，在此表示感谢！

尚华

2017 年 7 月

# 目 录

# 第一章　引言

## 1.1　研究意义和研究目的

### 1.1.1　研究意义

随着社会信息化的发展，经济、金融、生物、医学、教育、工业等都累积了大量的数据，而计算机的发展和应用则实现了这些海量数据的存储。现在面临的问题是，如何从这些海量的、无序的、看似杂乱无关的数据中提取有用的信息，以得到对过去工作的总结以及对未来的研究和工作的指导和建议。在这些保存的历史数据中，绝大部分都是根据时间顺序对历史事件的数值型记录的序列，称这些序列为时间序列（Time series）。时间序列数据在商业、经济及科学观测等各个社会领域中都广泛存在，比如商业零售行业中，在某一大型超市中，每天的销售额或每天的净利润；

气象预报研究中，某一地区每天的温度与湿度；生物医学中，某一症状病人在每个时刻的心跳变化等。

异常值是指那些不同于数据集中大部分的值。在静态数据中，若假定数据集本身来自于某一分布，异常值是指那些偏离此分布的值；若假定数据集中的观测值来源于某一模型，异常值是指那些偏离此模型的值。产生异常值的原因有很多，比如突然的天气原因、政策变化、书写错误等。异常值本身有时候还有很重要的意义，能提供很多有用的信息。比如：信用卡诈骗可能的表现为信用卡在不同地方几乎被同时使用，通过分析信用卡的使用情况，推测可能是信用卡诈骗。感应器在生活中经常被用于追踪各种环境周围的参数，它的突然变化可能是周围感兴趣的事件发生。在医学中，经常收集正常的 MRI 或 PET 扫描图，若有不同于这些的异常情形出现，则可能是癌症发生了。我们用卫星或远处遥感器收集大量地天气情况、气候变化等情况，以此根据可能的异常状况要预测突然的天气变化。

由于异常值的重要意义，异常值检测应运而生。通过大量的文献调查，发现大多数主流的时间序列异常值检测方法基本上都是针对 ARMA 或 ARIMA 模型的，即假定变

量为一元连续型的随机变量，但很多现实生活中的时间序列数据不是一元连续型的，可能为整值时间序列或多元时间序列数据。整值时间序列和多元时间序列是时间序列分析的重要组成部分，在交通、医学、金融等各个社会领域中都广泛存在。一阶整值自回归模型（First Order Integer-valued Autoregression，INAR(1)）模型和向量自回归（Vector Autoregression，VAR）模型是建模整值时间序列和多元时间序列最为成功的模型。

整值时间序列中最常见的是计数时间序列。计数时间序列是指在特定时间间隔中，某一事件发生的次数，它可能发生在生活的各个方面。例如：某种疾病每周的发病次数，某区域每月交通事故发生的次数，在股票市场每分钟的交易次数等。计数时间序列异常检测是比较少的。计数时间序列是正的、并且典型右偏的，这就需要一种特别的模型和程序。一般的异常值检测和剔除方法应用到计数时间序列异常中，经常会导致非整值出现，故不再适用。

INAR(1) 模型首先是由 McKenzie[1] (1985) 引入的，由于这个模型的简单性和解释的容易性，它在文献中被广泛研究，并应用到许多现实生活的实际问题，例如在统计过

程控制中的应用 Weiß[2] (2007)。因此，为解释某些数据集的离散性，激发我们来探索对离散时间序列的异常值检测。

在多元时间序列分析中，异常值的出现是不可预测的事件，可能会严重扭曲了一系列的分析。一个异常信号可能会是一个简单的确定性序列。但是这个信号是难以识别的，因为它嵌在一个数据集，数据集的动态结构既可能掩盖异常信号的观察也可能蔓延它的影响。因此，时间序列中的异常值检测问题比随机抽样中异常值检测要困难得多。异常的程度是由给定剩余观察值，其条件分布来判断，而不是边缘分布。所以异常值不一定是根据简单的图形检验其中最大或最小的数据，而是数据中与他们周围的观察值不一致的数据。

VAR 模型是多元时间序列分析中最成功的、最灵活的和易于使用的模型之一，它是一元时间序列自回归模型到动态多元时间序列的自然延伸。已经证明 VAR 模型对描述经济和金融时间序列的动态特性以及进行预测是非常有用的。不管是对一元时间序列模型还是以复杂理论为基础的联立方程组模型，它的预测精度都很高，除了数据描述和预测外，VAR 模型也可用结构推断和政策分析。