

分类号: O212.7
密 级: 公开

单位代码: 10363
学 号: 2221020127

题 目 约束条件下空间自回归模型的统计推断

英文并列

题 目 Statistical Inference of Spatial Autoregressive
Models under Constraints

学生姓名: 刘苏兵

指导教师: 何帮强 教授

专 业: 数 学

研究方向: 数理统计

论文答辩日期: 2025 年 5 月 14 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。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。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名： 

日期：2025 年 5 月 14 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本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、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ☐，在 年解密后适用本版权书。

本学位论文属于

不保密 ☐。

学位论文作者签名： 

日期：2025 年 5 月 14 日

指导教师签名： 

日期：2025 年 5 月 14 日

约束条件下空间自回归模型的统计推断

摘要

实际数据分析中普遍存在的参数约束、离散型整数值特征及观测数据缺失等问题，导致传统模型存在估计偏误大、残差控制难及稳健性不足等缺陷。空间自回归模型作为刻画地理空间依赖关系的核心工具，在区域经济学、环境监测与城市交通规划等领域展现出重要价值。因此本文围绕“约束条件下空间自回归模型的统计推断”这一科学问题，从理论方法、算法设计到应用实践开展系统性研究。

首先，针对协变量非线性效应与参数约束的耦合建模问题，提出线性可加空间自回归模型的约束估计模型。通过 B 样条基函数逼近非参数分量，构建含惩罚项的目标函数，采用 SCAD 正则化技术同步实现参数稀疏化与约束条件嵌入，严格证明斜率系数估计量的渐近无偏性。数值模拟 10 种典型约束场景下，该方法相较传统无约束估计使均方误差（MSE）降低 18.7%，交通流量预测的平均绝对误差下降 23.5%，有效解决了经济变量间复杂交互效应的量化分析问题。

其次，针对约束条件下整数值数据的非线性特征建模瓶颈，建立深度学习与空间计量融合的新型推断体系。设计混合鲁棒输入训练神经网络（H-RITNN）框架，通过门控循环单元捕捉时间序列动态特征，结合改进蚁群算法优化空间依赖参数，构建残差自回归-灰色预测组合模型。实证结果显示，该方法在区域 GDP 预测中将标准化皮尔逊残差从 0.35 降至 0.02，参数估计效率提升 19.3%，突破了传统线性模

型对计数数据建模的局限性。

最后，针对缺失数据场景下的参数估计失真问题，开发双重稳健的借补-纠偏联合估计算法。基于核密度估计生成缺失响应变量的多重填补值，利用局部多项式回归构造纠偏目标函数，通过投影矩阵理论实现线性约束的解析嵌入。理论推导验证了估计量在 MAR 机制下的 $\sqrt{n}$ 相合性与渐近正态分布特性。

本文的创新性贡献主要体现在三个维度：理论层面，建立约束可加 SAR 模型非参数分量与结构参数的渐近独立性定理，完善了半参数空间计量理论体系；方法层面，提出神经网络-智能优化混合架构，实现非线性整数值数据的空间依赖建模与异常值校正协同优化；应用层面，构建缺失数据下具有双重稳健特性的估计器，为智慧城市管理中的不完整时空数据分析提供方法论工具。研究成果可应用于长三角城市群交通网络优化、京津冀空气质量评估等实际项目，为空间计量模型在复杂场景下的工程化应用提供了重要技术支撑。

关键词：空间自回归模型，约束估计，SCAD 正则化，多重填补，半参数推断

STATISTICAL INFERENCE FOR SPATIAL AUTOREGRESSIVE MODELS UNDER CONSTRAINTS

ABSTRACT

The spatial autoregressive model, as a core tool for characterizing geospatial dependence, demonstrates significant value in regional economics, environmental monitoring, and urban transportation planning. However, prevalent challenges in practical data analysis—including parameter constraints, discrete integer-valued characteristics, and missing observations—lead to substantial limitations in conventional models, such as large estimation bias, difficulties in residual control, and insufficient robustness. This study systematically investigates the scientific problem of "statistical inference for constrained spatial autoregressive models" through theoretical advancements, algorithmic innovations, and practical applications.

Firstly, to address the coupled modeling challenge of nonlinear covariate effects and parameter constraints, we propose a constrained estimation model for additive spatial autoregressive models. By approximating nonparametric components through B-spline basis functions, we construct an objective function incorporating penalty terms

and employ SCAD regularization to simultaneously achieve parameter sparsity and constraint enforcement. The asymptotic unbiasedness of slope coefficient estimators is rigorously established. Numerical simulations across 10 typical constraint scenarios demonstrate that this approach reduces mean squared error (MSE) by 18.7% compared to conventional unconstrained estimation and decreases average absolute error in traffic flow prediction by 23.5%, effectively resolving quantitative analysis of complex interactions among economic variables.

Secondly, to overcome modeling bottlenecks for integer-valued data with nonlinear characteristics under constraints, we establish a novel inference framework integrating deep learning with spatial econometrics. A hybrid robust input training neural network (H-RITNN) framework is designed, incorporating gated recurrent units to capture temporal dynamics and enhanced ant colony optimization for spatial parameter tuning. A residual autoregressive-grey prediction composite model is developed. Empirical results show this method reduces standardized Pearson residuals from 0.35 to 0.02 in regional GDP forecasting while improving parameter estimation efficiency by 19.3%, overcoming limitations of traditional linear models for count data.

Thirdly, to mitigate parameter estimation distortion under missing data scenarios, we develop a doubly robust imputation-correction joint estimation algorithm. Multiple imputations for missing responses are

generated through kernel density estimation, with bias-corrected objective functions constructed via local polynomial regression. Linear constraints are analytically embedded using projection matrix theory. Theoretical derivations confirm the consistency and asymptotic normality of estimators under the Missing at Random (MAR) mechanism.

The innovative contributions of this research are manifested in three dimensions: Theoretically, we establish an asymptotic independence theorem between nonparametric components and structural parameters in constrained additive SAR models, advancing semiparametric spatial econometric theory. Methodologically, the proposed neural network-intelligent optimization hybrid architecture enables collaborative optimization of spatial dependence modeling and outlier correction for nonlinear integer-valued data. Practically, the developed doubly robust estimator provides methodological tools for incomplete spatiotemporal data analysis in smart city management. Applications in the Yangtze River Delta urban agglomeration transportation network optimization and Beijing-Tianjin-Hebei air quality assessment demonstrate the framework's engineering value, offering crucial technical support for implementing spatial econometric models in complex scenarios.

Liu Subing (Mathematics)

Supervised by He Bangqiang

KEY WORDS: Spatial autoregressive model, constrained estimation, SCAD regularization, integer-valued data, multiple imputation, semiparametric inference

目 录

摘 要	I
ABSTRACT	III
第 1 章 绪论	1
1.1 研究背景与研究意义	1
1.2 国内外研究现状	2
1.3 主要研究内容	4
第 2 章 估计方法	5
2.1 最大似然估计方法	5
2.2 约束条件下的估计	6
2.3 数值求解算法	6
2.4 估计量的理论性质	7
第 3 章 约束条件下线性可加空间自回归模型方法研究	8
3.1 引言	8
3.2 方法与主要结果	10
3.3 主要结果证明	14
3.4 模拟研究	16
3.5 本章小结	24
第 4 章 约束条件下基于空间自回归模型和最大似然估计的非线性整数值的统计推断	25
4.1 引言	25
4.2 方法与主要结果	26
4.3 主要结果证明	30
4.4 数值模拟	33
4.5 本章小结	39
第 5 章 缺失数据下空间自回归模型的约束估计	41
5.1 引言	41
5.2 方法与主要结果	42
5.3 主要结果证明	45
5.4. 本章小结	49
第 6 章 总结与展望	50
6.1 总结	50
6.2 展望	51
参考文献	52
硕士期间发表的论文	56
致 谢	57

第 1 章 绪论

1.1 研究背景与研究意义

1.1.1 研究背景

随着统计学和空间数据分析的不断发展，空间自回归模型（Spatial Autoregressive Model, 简称 SAR 模型）在处理具有空间依赖性的数据方面受到了广泛关注。传统计量方法通常预设解释变量与响应变量间呈独立分布特性，但多学科实证研究表明（如区域经济分析、生态监测、大气环流模拟及矿产分布预测等），地理邻近观测值常呈现显著的空间自相关特征。若在参数估计与假设检验过程中忽视此类空间依赖性，将导致回归系数的系统性误差。基于空间权重矩阵构建的 SAR 模型，通过量化相邻单元的空间滞后效应，能够更精准地揭示解释变量与响应变量间的动态关联机制。

近年来，随着空间数据分析在各领域的广泛应用，空间自回归模型作为一种重要的空间计量经济学工具，受到了越来越多的关注。空间自回归模型的理论研究已取得了重要成果，然而，在实际应用中，研究者常常面临各种约束条件，在约束条件下的统计推断方面仍存在许多挑战。约束条件可以来自先验信息、模型假设或实际应用需求，这些约束条件能够为模型提供额外的结构信息，从而提高模型的估计精度和解释能力。例如，在经济学中，某些经济变量之间的关系可能受到政策或市场机制的约束；在环境科学中，某些生态系统的动态变化可能受到自然规律或人类活动的限制。这些约束条件可能显著影响模型的统计推断结果，但现有研究对此关注不足。因此，探讨约束条件下空间自回归模型的统计推断问题具有重要的理论和实践意义。

1.1.2 研究意义

(1)理论意义

1.拓展空间自回归模型的理论框架。通过在空间自回归模型中引入约束条件，可以进一步丰富和发展空间自回归模型的理论体系。这不仅有助于深入理解空间依赖性数据的内在结构和特性，还能为相关领域的理论研究提供新的视角和方法。

2.提高模型估计的效率和准确性。约束条件可以为模型提供额外的信息，从

而在一定程度上减少模型的不确定性和参数的估计误差。通过研究约束条件下的统计推断方法，可以开发出更有效的估计和检验技术，提高模型参数估计的效率和准确性。

3.推动统计推断方法的创新。约束条件下的空间自回归模型的统计推断涉及多个领域的交叉，如统计学、计量经济学、空间分析等。这为统计推断方法的创新提供了广阔的空间，有助于推动相关领域方法论的发展。

(2)实践意义

1.增强模型的解释能力和预测能力。在实际应用中，约束条件往往反映了某些已知的规律或先验知识。通过在模型中引入这些约束条件，可以使模型更好地符合实际情况，从而增强模型的解释能力和预测能力。例如，在经济学中，通过引入政策约束条件，可以更准确地预测经济变量的变化趋势；在环境科学中，通过引入生态约束条件，可以更好地评估环境变化对生态系统的影响。

2.为政策制定和决策提供科学依据。约束条件下的空间自回归模型可以为政策制定和决策提供更科学、更准确的依据。通过分析模型结果，可以更好地理解不同因素之间的相互作用和影响机制，从而为制定合理的政策和决策提供有力支持。例如，在城市规划中，通过分析空间自回归模型的结果，可以更好地制定土地利用政策和交通规划方案。

3.促进跨学科研究的发展。约束条件下的空间自回归模型的统计推断涉及多个学科领域的交叉，如经济学、地理学、环境科学等。本研究通过构建跨学科知识整合平台与协同创新机制，推动多学科范式融合，为交叉学科可持续发展提供理论支撑与实践路径。

综上所述，约束条件下空间自回归模型的研究兼具理论创新价值与实践指导意义。其不仅深化对该模型的参数估计与假设检验的理论认知，更通过构建约束条件下的统计推断体系，为复杂空间数据分析建立具有适用性的计量分析框架。

1.2 国内外研究现状

近年来，国内外学者对空间自回归（SAR）模型的统计推断进行了广泛研究。Engle^[1]提出了结合线性模型和非参数模型的部分线性模型。Paul^[2]应用核平滑方法估计模型参数部分的未知系数以及非参数部分的未知系数函数。Severini 和 Staniswalis^[3]提出一种拟似然函数的方法来估计该模型，并给出估计量的渐近分

布。陈明华等^[4]基于部分线性模型给出了未知参数的最小二乘估计和加权最小二乘估计。Anselin 和 Rey^[5]系统性综述空间计量模型的最新进展，包括约束估计与稳健推断。Lee 和 Yu^[6]推导约束空间模型的局部渐近正态性（LAN）与效率下界。罗国旺和吴密霞^[7]结合 sieve 两阶段最小二乘与拉格朗日乘子法处理线性约束，证明了约束估计量的渐近正态性。Liu 等^[8]提出惩罚拟极大似然估计（PQML），结合 SCAD 惩罚函数处理参数非负性约束。Chen 和 Wang^[9]研究了在贝叶斯框架下，结合 MCMC 采样，通过截断正态先验施加不等式约束，结果表明在高分位数水平下，约束模型对极端空间事件的预测误差明显降低。Zhang 等^[10]提出自适应 Lasso 惩罚，结合广义矩估计（GMM）处理高维约束问题。Kelejian 和 Prucha^[11]研究了扩展广义矩估计（GMM）至内生协变量与空间滞后项并存的情形。丁飞鹏^[12]研究基于二次推断函数（QIF）处理面板数据中的非随机缺失，引入 EM 算法迭代优化后估计量的 MSE。Carroll 等^[13]设计工具变量校正矩阵，修正协变量测量误差导致的参数偏移。Wang 和 Sun^[14]结合核回归填补与稳健 Huber 损失函数，处理响应变量缺失和异常值。Li 和 Mei^[15]提出两阶段自适应 Group Lasso 方法，处理高维参数与线性约束。Ma 等^[16]使用 B 样条基函数近似非参数空间效应，施加单调性约束。Huang 等^[17]结合局部多项式回归与时变约束，分析了动态空间的依赖性。Li 等^[18]研究了基于 Gibbs 采样和 Metropolis-Hastings 算法，处理面板数据中的参数不等式约束。Chen 等^[19]利用哈密尔顿蒙特卡洛（HMC）加速高维约束空间模型的收敛。Bivand 和 Piras^[20]开发 R 包 sphet 支持约束 SAR 模型的 GMM 与 ML 估计。Pace 等^[21]提出基于 Cholesky 分解的快速矩阵运算，优化高维约束模型计算。Elhorst^[22]系统讨论面板数据 SAR 模型的约束设定与估计方法。LeSage 和 Pace^[23]研究在贝叶斯视角下处理空间模型中的复杂约束（如不等式、非线性）。Su 和 Yang^[24]提出基于 Wild Bootstrap 的约束条件检验方法，证明在异方差下，检验功效较传统方法显著提升。Fingleton^[25]综合理论推导与实证案例，覆盖经济、环境等多领域，强调约束条件在因果推断中的不可忽略性。Kang 和 Li^[26]采用 Huber 损失函数与自适应惩罚处理厚尾误差。Guan 和 Wang^[27]结合神经网络与 SAR 模型，处理非线性空间约束。

在应用方面，Zhang 和 Liu^[28]构建空间自回归模型分析政策约束对疫情传播

的影响。Li 等^[29]结合生态保护约束与 SAR 模型，优化城市土地利用决策。

综合以上研究成果可知，空间自回归模型仍是目前的研究热点之一。目前该模型的研究较多，其被广泛应用于区域经济学、金融市场、生态保护与城市交通规划等领域，用于分析经济个体活动的溢出效应和相关性。本研究聚焦约束条件下空间自回归模型的参数估计与假设检验机制，通过构建约束性计量建模的理论架构与方法体系，为复杂空间数据分析提供方法论创新的研究突破口。

1.3 主要研究内容

本研究聚焦于约束条件下空间自回归模型（SAR）的统计推断体系构建，针对带约束的建模环境提出三个层次方法论创新：其一，建立线性可加 SAR 模型的约束估计框架；其二，推导非线性整数值 SAR 模型的约束最大似然估计(MLE)理论；其三，构建缺失数据情境下的约束参数估计算法。

具体内容如下：

1、针对线性可加 SAR 模型的约束估计，在传统可加模型基础上引入线性约束框架，通过 B 样条基函数逼近非参数分量，生成包含 B 样条基函数的空间自回归模型，对此模型的损失函数进行最小化，获取约束条件下模型的稀疏估计，通过 SCAD 惩罚函数无偏估计较大系数。在 10 种约束条件下，研究了线性可加空间自回归模型估计的渐近性质，得到在最优收敛速度下进行斜率系数的估计时，函数项部分不会改变参数估计的渐近分布特性。

2、提出一种约束条件下基于空间自回归模型与最大似然估计的非线性整数值的统计推断方法。首先，设计混合型鲁棒输入训练神经网络（H-RITNN）框架对历史数据进行非线性校正，通过误差反向传播与梯度下降算法迭代训练，识别异常数据并生成校正值，实现观测样本高效预处理。其次，构建残差空间自回归模型，描述数据的线性与非线性空间依赖特性，并引入改进蚁群优化算法增强最大似然估计的全局寻优能力，精准求解模型参数以更新数据空间关联结构。进而建立以空间自回归模型为核心、灰色 GM(1,1)模型与 BP 神经网络为辅的协同推断架构，通过动态权重分配融合多模型优势，提升非线性整数值的推断鲁棒性。

3、在满足缺失机制为随机缺失（MAR）的假设下，利用局部纠偏的剖面最小二乘方法和拉格朗日乘数法构造带约束估计量，使得参数部分满足线性约束在一定正则条件下，我们证明所构造的约束估计量一致且渐近正态。

第2章 估计方法

本章主要介绍在约束条件下对空间自回归模型参数进行统计推断的估计方法。首先，我们给出模型的基本设定以及所考虑的约束条件，然后详细阐述基于最大似然估计（MLE）的估计方法，包括似然函数的构造、约束条件的处理以及相应的数值优化方法。最后，本章讨论估计量的理论性质（如一致性与渐近正态性）及其实现中的关键问题。

2.1 最大似然估计方法

在不考虑约束的情况下，对上述模型的极大似然估计需要构造联合似然函数。由于模型中存在空间滞后项，为保证模型的正确性，需要对变量进行变换。

令

$$A(\rho) = I - \rho W,$$

则模型可写为

$$A(\rho) Y = X\beta + \varepsilon. \quad (2.1)$$

在此基础上，假设 $\varepsilon \sim N(0, \sigma^2 I)$ ，则联合密度为

$$f(Y | X, \rho, \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2} \|A(\rho)Y - X\beta\|^2} \quad (2.2)$$

但注意到由变量变换引入的雅可比校正项，其行列式为 $|A(\rho)|$ 。因此，模型的对数似然函数可写为：

$$\begin{aligned} \ell(\rho, \beta, \sigma^2) &= \log |A(\rho)| - \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (A(\rho)Y - X\beta)^T (A(\rho)Y - X\beta) \\ &= \log |A(\rho)| - \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} Q(\rho, \beta) \end{aligned} \quad (2.3)$$

其中

$$Q(\rho, \beta) = (A(\rho)Y - X\beta)^T (A(\rho)Y - X\beta).$$

2.2 约束条件下的估计

在实际应用中，我们常常需要引入约束条件,如 $A\beta = b$ ；或在参数空间上引入约束，如： $\sum_{h=1}^H v_h = 1, v_h \geq 0$ 等。为获得满足约束条件的最大似然估计，可以采用以下几种策略：

1、拉格朗日乘子法

构造拉格朗日函数

$$L(\beta, \lambda) = Q(\beta) + 2\lambda^T (A\beta - b) \quad (2.4)$$

其中 λ 为拉格朗日乘子向量。令 $L(\beta, \lambda)$ 对 β 与 λ 分别求导，由此可解得其值，进一步可得 β 的估计值：

2、惩罚函数法

在对数似然函数中加入惩罚项，设定 $b\mu(\varphi)$ 为非减和凹，且满足 $\varphi \in [0, +\infty)$ ，在 $(0, +\infty)$ 区间内存在连续的导数，用 $b'\mu(\varphi)$ 表示.本文采用非凸惩罚函数 SCAD 进行惩罚估计，其惩罚函数为：

$$b\mu(\varphi; r) = \begin{cases} \mu|\varphi|, & |\varphi| \leq \mu \\ \frac{-\varphi^2 + 2r\mu|\varphi| - \mu^2}{2(r-1)}, & \mu < |\varphi| \leq r\mu \\ \frac{(r+1)\mu^2}{2}, & |\varphi| > r\mu \end{cases} \quad (2.5)$$

式中， $r \geq 2$ 、 $\mu \geq 0$ 。在 $(-\infty, 0) \cup (0, \infty)$ 中，SCAD 惩罚函数既连续又可微，其在原点呈现奇异特性，从而达成稀疏性估计的目的。

2.3 数值求解算法

由于对数似然函数形式较为复杂（特别是包含 $\log |A(\rho)|$ 项），直接求解一阶导数方程往往没有解析解。常采用牛顿-拉夫森法：

对于参数向量 $\theta = (\rho, \beta, \sigma^2)$ ，其梯度与海森矩阵分别为

$$\frac{\partial \ell(\theta)}{\partial \theta}, \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T} \quad (2.6)$$

牛顿-拉夫森迭代公式为

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 \ell(\theta_k)}{\partial \theta \partial \theta^T} \right]^{-1} \frac{\partial \ell(\theta_k)}{\partial \theta} \quad (2.7)$$

在每一步迭代中，需要保证更新后的解满足约束条件；若不满足，可采用参数投影或利用罚函数法修正。

2.4 估计量的理论性质

在满足一定正则条件下，通过上述最大似然方法及其数值求解，我们可以证明估计量具有如下渐近性质：

1、一致性

在样本容量 $n \rightarrow \infty$ 时，约束下的最大似然估计量 $\hat{\theta}$ 收敛于真实参数值 θ_0 。

2、渐近正态性

有 $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I(\theta_0)^{-1})$ ，其中 $I(\theta) = -E\left[\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^T}\right]$ 为 Fisher 信息矩阵。需

要注意的是，在约束条件下，证明过程中可能涉及 Lagrange 乘子法以及参数变换的影响，其数学证明较为繁琐但基本思想与无约束情形类似。

第3章 约束条件下线性可加空间自回归模型方法研究

本章针对空间相依性建模中的约束估计难题,提出基于半参数框架的约束可加空间自回归模型估计体系。在传统可加模型基础上引入线性约束框架,通过 B 样条基函数逼近非参数分量,并代入回归模型,生成包含 B 样条基函数的空间自回归模型,对此模型的损失函数进行最小化,获取约束条件下模型的稀疏估计,通过 SCAD 惩罚函数无偏估计较大系数。在 10 种约束条件下,研究了线性可加空间自回归模型估计的渐近性质,得到在最优收敛速度下进行斜率系数的估计时,函数项部分不会改变参数估计的渐近分布特性。实验结果表明,该方法不仅在理论上具有可行性,而且在实践中易于操作。相比非约束估计,约束稀疏估计更为优秀。在交通数据统计分析中,该方法具有实用性和有效性。

3.1 引言

统计推断是统计学中的一种重要方法,它是基于对样本数据的分析来推测总体的性质和特征,在许多领域中,由于受限于成本、时间或可行性等因素^[30],无法直接对总体进行全面调查。此时,通过样本数据来进行统计推断,可以帮助人们降低这种不确定性,了解总体的特征和规律。基于历史数据和当前状况的统计推断,可以帮助人们预测未来的趋势和可能性。例如,通过分析过去几年的销售数据,可以预测未来的销售情况,从而制定相应的策略^[31];通过统计推断,可以发现存在的问题和不足之处,从而进行改进和完善;通过分析产品质量检测数据,可以发现产品的缺陷,进而改进生产工艺或流程^[32]。因此,统计推断作为一种重要的方法论,在各个领域中都发挥着重要的作用。

统计推断的方法有很多,比如 Chen 等^[33]提出的 P 样条部分线性加性空间自回归模型的贝叶斯统计推断方法。此方法可以全面评估协变量对空间数据响应的线性与非线性影响。利用贝叶斯 P 样条,结合马尔可夫链蒙特卡罗法,开发出贝叶斯估计法 PLASARM。同时,设计 Gibbs 采样器来探索未知参数的联合后验分布,并针对实际问题进行精确的统计推断。这种方法需要知道先验概率,且先验概率往往取决于假设,而假设的模型有很多种,这就可能导致推断效果的不佳。Lustosa 等^[34]提出不同模型数据规模下的机制统计推断方法,将统计原理融入数

据处理中，不仅可以简化操作和优化方案，还可以简化推断过程。通过强调数据建模、数据集合设计、可视化以及细致的数据重组，有助于提高数据统计的准确性和数据选择的合理性，进而对经济预测等实际问题进行准确的统计推断。此方法对数据质量要求高，如果数据存在异常值、缺失值等问题，会影响统计推断的准确性。张洁等^[35]提出一种采用自激励模型进行统计推断，此方法基于最小二乘法（conditional least squares, CLS）和连续机器学习（continuous machine learning, CML）方法的参数估计，有效解决了整值时间序列分段现象的问题，并对预测问题进行了深入研究。为了更精确地描述时间序列特性，该方法扩大了NBTINAR模型的要求。针对整值门限自回归过程，研究其基础属性与非参数估计课题，获取参数的精确估计以及区间估计，这也为实际统计问题提供了推断依据。但此方法是基于CLS和CML方法的参数估计，在实际应用中，模型选择和参数估计可能存在主观性和不确定性。赵志文等^[36]提出周期随机系数自回归模型的方法进行统计推断，此方法研究多元随机系数自回归模型，并采用非参方法进行参数统计推断；同时，考虑到实际数据中的协变量影响，提出了带有协变量的随机系数自回归模型，并讨论了模型的参数估计等问题；此外，考虑到协变量数据的测量误差，提出了相应的模型，并利用经验似然方法对多元一阶随机系数自回归模型的参数进行估计和假设检验，同时检验自回归系数是否恒为常数。这种方法为解决实际问题提供了有效的统计推断方法，但此方法计算复杂度较高，经验似然方法涉及复杂的统计计算，对于大规模数据或复杂模型可能会变得计算量大且耗时。

约束条件下线性可加空间自回归模型的统计推断方法推动了统计推断领域的发展，为解决实际问题提供了新的思路 and 方向，可应用于各种实际问题，如经济预测、城市规划等，为决策者提供有力支持。约束条件有助于模型更好地拟合数据，从而提高预测的准确性，也可以降低模型的复杂度，使模型更易于理解和应用。通过对约束条件的设定和应用，可以更好地洞察数据的内在结构和规律，从而作出更有效的决策。因此本文提出约束条件下线性可加空间自回归模型的统计推断方法。

3.2 方法与主要结果

3.2.1 约束条件下线性可加空间自回归模型

(1) 模型框架

本次用 $\{Y_i, X_i, Z_i\}_{i=1}^n$ 描述观测获得的高维空间数据, X_i 、 Y_i 、 Z_i 、 n 分别用于描述 a 维的协变量、响应变量、 b 维的协变量、空间单元的个数, 且 X_i 和 Y_i 呈线性关系, Z_i 和 Y_i 呈不确定形态的非线性关系. n 个空间单元的相邻关系通过空间权重矩阵 W 描述, 由此考虑部分线性可加的空间自回归模型^[37]表述为式 (3.1):

$$Y = \rho WY + X\beta + \sum_{j=1}^b p_j(Z_j) + \varepsilon \quad (3.1)$$

式中, ρ 、 $p_j(\cdot)$ 、 β 分别用于描述空间自回归系数、未知的平滑函数 $a \times 1$ 维的系数向量。设定在 $1 \leq j \leq b$ 时, $F[p_j(Z_j)] = 0$ 成立, 来确保非参数部分的可辨识性。 ε 用于描述随机扰动项, 符合 $F(\varepsilon|X, Z) = 0$, 并且是有边界的方差。上述模型中的空间滞后项反映空间上的相互依赖关系。参数部分明确描述协变量和响应变量之间的线性关系, 系数意义清晰。非参数部分则灵活处理协变量和响应变量间不确定的具体关系形态^[38,39]。考虑到空间数据的异质性, 不对扰动项进行正态性或同方差的假设。

在考虑部分线性可加的空间自回归模型中, 其参数都是在某些约束限制下进行选取^[40], 本文选取线性约束作为模型的约束方式, 具体用 (3.2) 式表述:

$$A\beta = b \quad (3.2)$$

式中, A 和 b 分别用于描述行向量线性无关的矩阵 $m \times n$ 以及 $m \times 1$ 维的向量。通过线性约束, 加快参数估计速度^[41,42]。

3.2.2 通过 B 样条函数近似未知的平滑函数

考虑部分线性可加的空间自回归模型中, 未知的平滑函数通过组合的 B 样条基函数来近似求解。下面详细描述, 设定 c_j 与 d_j 是正整数, $g_j = c_j + d_j$, 把 0-1 区间分成 $d_j + 1$ 个子区间, $h_j = 0 < u_{j,1} < \dots < u_{j,d_j} < 1 = h_j$ 描述子区间的内部节点。由 c_j 阶和 d_j 个节点数构成的样条函数空间用 α_j 描述^[43]。在 B 样条函数空间 α_j 中

选择一组基函数，用 $\{B_{j,m}(\cdot)\}_{m=1}^{g_j}$ 描述，未知的平滑函数通过样条函数来替代，由 (3.3) 表达：

$$p_j(\cdot) = \sum_{m=1}^{g_j} B_{j,m}(\cdot) \chi_{j,m}, j=1,2,\dots,b \quad (3.3)$$

式中， $\chi = (\chi_{1,1}, \dots, \chi_{1,g_1}, \dots, \chi_{b,1}, \dots, \chi_{b,g_b})^T$ 。

令 $\{B_{j,m}(\cdot)\}_{m=1}^{g_j}$ 为 (4) 式：

$$B(z_{i,t}) = [B_{1,1}(z_{i,t,1}), \dots, B_{1,g_1}(z_{i,t,1}), \dots, B_{b,1}(z_{i,t,b}), \dots, B_{b,g_b}(z_{i,t,b})]^T \quad (3.4)$$

在 (3.1) 式中代入 (3.3) 式，能得到 (3.5) 式：

$$Y = \rho \tilde{W} Y + X\beta + B(Z)\chi + Dn(\mu) + \lambda \quad (3.5)$$

式中， $Y = (Y_1^T, \dots, Y_N^T)^T$, $Y_i = (y_{i,1}, \dots, y_{i,T})^T$, $\tilde{W} = W \otimes J_T$, J_T 用于描述阶数是 T 的单位矩阵。

$$B(Z) = \left\{ [B(Z_1)]^T, \dots, [B(Z_N)]^T \right\}^T, B(Z_i) = \{B(z_{i,1}), \dots, B(z_{i,T})\}^T,$$

$D = [-1_{N-1}, J_{N-1}]^T \otimes I_T$, I_T 和 $\otimes$ 分别用于描述维数为 T 且元素都是 1 的向量以及

$$\text{Kronecker 积. } \eta = (\eta_2, \dots, \eta_N)^T, \lambda = (\varepsilon_1^T, \dots, \varepsilon_N^T)^T, e_i = (\varepsilon_{i,1}, \dots, \varepsilon_{i,T})^T.$$

通过以上过程可以进一步增强模型对空间数据的拟合能力，捕捉更复杂的空间模式和关系。

3.2.3 线性可加空间自回归模型稀疏估计实现

在分位回归框架下，对模型 (3.5) 进行稀疏估计。为简化空间回归模型的复杂性，采用特征向量线性组合对空间滞后项近似表达，并通过 B 样条基函数近似未知的平滑函数。然而，这种近似可能带来模型设定误差，将其纳入随机扰动项中。由于随机扰动项可能不满足正态性以及同方差性，难以确保均值回归估计量的优良性质。因此，采用分位回归，其稳健性更佳。高维数据会展现出异质性，协变量对响应变量的影响随位置变化。分位回归能全面揭示这种关系，并合理解释协变量效应的稀疏性。

(1) 目标函数

当分位水平相异时，模型 (3.1) 式的非参数可加部分中，协变量对响应变量的影响会有所不同，这可以通过观察 $p_j(Z_j)$ 值是否为零来体现。在模型 (3.5) 式中，选择合适的基函数组合，以便筛选出非参数部分的关键协变量。特定的样条信息定义各协变量，并通过样条基函数进行展开。因此，对变量的筛选本质上就是对样条信息的筛选。将惩罚的目标函数最小化处理，来获取稀疏估计，用 (3.6) 式表达：

$$\begin{aligned} Lb(\chi, \beta, \eta) &= \frac{1}{n} \sum_{i=1}^n \delta_{\tau} [Y_i - F_i \chi - X_i \beta - B(U_i)^T \eta] \\ &\quad + \sum_{j=1}^g b_{\mu}(\|\eta_j\|_2) \end{aligned} \quad (3.6)$$

式中， $\rho_{\tau}(u) = u[\tau - J(U < 0)]$ 、 τ 、 $b_{\mu}(\cdot)$ 分别用于描述分位数损失函数的具体形式、在 0-1 区间取值的分位水平、惩罚函数，控制参数 u 用于调节模型的复杂度。惩罚函数仅有惩罚非参数部分的 B 样条基函数的系数； F_i 代表条件分布函数。

(2) 罚函数

设定 $b_{\mu}(\varphi)$ 为非减和凹，且满足 $\varphi \in [0, +\infty)$ ，在 $(0, +\infty)$ 区间内存在连续的导数，用 $b'_{\mu}(\varphi)$ 表示。本文采用非凸惩罚函数 SCAD 进行惩罚估计，其惩罚函数为：

$$b_{\mu}(\varphi; r) = \begin{cases} \mu|\varphi|, & |\varphi| \leq \mu \\ \frac{-\varphi^2 + 2r\mu|\varphi| - \mu^2}{2(r-1)}, & \mu < |\varphi| \leq r\mu \\ \frac{(r+1)\mu^2}{2}, & |\varphi| > r\mu \end{cases} \quad (3.7)$$

式中， $r \geq 2$ 、 $\mu \geq 0$ 。在 $(-\infty, 0) \cup (0, \infty)$ 中，SCAD 惩罚函数既连续又可微，其在原点呈现奇异特性，从而达成稀疏性估计的目的。在 $|\varphi| > r\mu$ 时导数为 0，函数值表现为常数，防止在模型中对较大系数进行过度惩罚。用 (3.8) 式表达当

$\varphi \in [-r\mu, r\mu]$ 时的导数公式:

$$b'\mu(\varphi) = \mu \left[1(\varphi \leq \mu) + \frac{(r\mu - \varphi)_+}{(r-1)\mu} 1(\varphi > \mu) \right] \quad (3.8)$$

式中, $(r\mu - \varphi)_+ = \max\{x, 0\}$ 本文中指定 $r = 3.5$. SCAD 惩罚回归具有生成稀疏解的能力, 且能无偏估计较大系数。

3.2.4 线性可加空间自回归模型估计的渐近性分析

参数空间用 $\Theta = A \times B$ 示, 用 λ_0 表示斜率系数, 当协变量 D_i 、 X_i 、 Z_i 被指定时, 条件分布函数、条件密度函数分别用 $F_i(\cdot)$ 以及 $f_i(\cdot) = f_{\varepsilon_i|D_i, X_i, Z_i}$ 代表。

为推导约束条件下模型的渐近性质, 需满足以下正则条件:

(C1) 存在常数 w 使得 $E\|X\|^4 \leq w$, 且对任意 $j \geq 1$, 有 $E\psi_j^4 \leq w\rho_j^2$, 其中 ψ_j 为函数型主成分得分, ρ_j 为协方差函数的特征值。

(C2) 真实函数 $\lambda_0(\tau, t) \in C^b([0, 1])$, 即具有 b 阶连续导数。

(C3) 存在常数 $a > 1$, 使得协方差函数的特征值满足: $w^{-1}j^{-a} \leq \gamma_j \leq wj^{-a}, \gamma_j - \gamma_{j+1} \geq w^{-1}j^{-a-1}$.

(C4) 存在常数 $b > a/2 + 1$, 使得基展开系数满足: $|\chi_j| \leq wj^{-b}$.

(C5) 误差项 ε_i 的密度函数 $f_i(0)$ 在零点处连续且有界, 即 $\sup_i |f_i(0)| < \infty$.

(C6) 信息矩阵 $J = E[f_i(0)\tilde{X}\tilde{X}]$ 正定, 其中 $\tilde{X} = [X, B(Z)]$.

(C7) 矩阵 $I - \mathcal{GW}$ 可逆, 且 $\|W\|_\infty < 1$.

(C8) 截断参数 $m = O(n^{1/(a+2b)})$.

(C1)-(C2) 为矩条件与函数光滑性; (C3) 为特征值衰减条件; (C4) 为系数衰减条件; (C5)-(C6) 为分位数回归条件; (C7)-(C8) 为空间权重矩阵条件。

定理 3.1 当条件(C1)至(C8)成立, 则对任意 $\tau \in (0,1)$, 估计量 $\hat{\vartheta}(\tau)$ 满足:

$$\sqrt{n}[\hat{\vartheta}(\tau) - \vartheta(\tau)] \xrightarrow{a} N(0, RSR') \quad (3.9)$$

其中 $R = (J'_g H J_g)^{-1} J'_g H$, $H = J'_\varsigma A J_\varsigma$, 矩阵 $\bar{J}_\chi$ 和 $\bar{J}_\varsigma$ 的维数分别是 $\dim(\chi) \times \dim(\chi, \varsigma)$ 和 $\dim(\varsigma) \times \dim(\chi, \varsigma)$. $S = \lim_{n \rightarrow \infty} \frac{\tau(1-\tau)}{n} \tilde{X} \tilde{X}'$, 矩阵 J_v^{-1} 的公正分割用 $[\bar{J}'_\chi, \bar{J}'_\varsigma]'$ 表示。

在 $\dim(\varsigma) = \dim(\vartheta)$ 成立时, $\hat{\vartheta}(\tau)$ 的渐近方差不受矩阵 $A(\vartheta)$ 的影响, $\hat{\vartheta}(\tau)$ 的渐近方差的简洁方式为 $J_g^{-1} S J_g^{-1}$. 假如 $\dim(\varsigma) > \dim(\vartheta)$, 则 $\hat{\vartheta}(\tau)$ 的渐近方差会受到矩阵 $A(\vartheta)$ 选择结果的影响。 $\hat{\varsigma}(\vartheta, \tau) = \bar{J}_\varsigma S J'_\varsigma$ 的渐近方差的逆选取矩阵 $A(\vartheta)$, 所以获得 $\sqrt{n}[\hat{\vartheta}(\tau) - \vartheta(\tau)]$ 的渐近方差, 用式 (3.10) 表示:

$$w_g^2 = \left[J'_g \bar{J}'_\varsigma (\bar{J}_\varsigma S \bar{J}'_\varsigma)^{-1} \bar{J}_\varsigma J_g \right]^{-1} \quad (3.10)$$

在理论上, 如果满足一定的条件, 选用高效的工具变量以及适当的得分函数权重矩阵, 可以实现参数 $\hat{\vartheta}$ 的估计更加准确可靠。

定理 3.2 如果条件(C1)至(C8)成立, 设定截断参数 m 符合 $m \in n^{1/(a+2b)}$, 由文献^[44]可得:

$$\|\hat{\lambda}(\tau, t) - \lambda_0(\tau, t)\|^2 = U_p \left[n^{1-2b/(a+2b)} \right] \quad (3.11)$$

斜率系数的估计在最优收敛速度下进行, 并且函数项部分不会改变参数估计 $\hat{\vartheta}$ 的渐近分布特性。

3.3 主要结果证明

定理 3.1 (渐近正态性)

$$\sqrt{n}[\hat{\vartheta}(\tau) - \vartheta(\tau)] \xrightarrow{a} N(0, RSR')$$

其中 $R = (J'_g H J_g)^{-1} J'_g H$, $H = J'_\varsigma A J_\varsigma$, $S = \lim_{n \rightarrow \infty} \frac{\tau(1-\tau)}{n} \tilde{X} \tilde{X}'$

证明 定义分位数损失函数的梯度（得分函数）为：

$$U_n(\theta) = \frac{1}{n} \sum_{i=1}^n \psi(Y_i, X_i, Z_i; \theta)$$

其中， $\psi(Y_i, X_i, Z_i; \theta) = \left[\tau - I(Y_i \leq X_i \beta + B(Z_i)^\top \eta) \right] \cdot \begin{pmatrix} X_i \\ B(Z_i) \end{pmatrix}$

在真实参数 $\theta_0 = (\sigma_0^T, \eta_0^T)^T$ 处展开，利用条件（C5-C6）可得信息矩阵：

$$J = E[f_i(0) \tilde{X} \tilde{X}^T]$$

由条件（C1）-（C6），根据中心极限定理，标准化得分函数满足：

$$\sqrt{n} U_n(\theta_0) \xrightarrow{d} N(0, \Sigma)$$

其中 $\Sigma = \tau(1-\tau) \cdot \mathbb{E} \begin{pmatrix} X_i \\ B(Z_i) \end{pmatrix} \begin{pmatrix} X_i^T & B(Z_i)^T \end{pmatrix}$,

引入约束 $A\beta = b$ ，通过投影矩阵 $P = I - S^T(SS^T)^{-1}S$ 调整参数空间。修正后的渐近方差为：

$$\text{var}(\hat{\sigma}) = (SJ^{-1}S^T)^{-1} J^{-1} \Sigma J^{-1} (SJ^{-1}S^T)^{-1}$$

最终化简为 RSR' ，其中 $R = (J'_g H J_g)^{-1} J'_g H$ 。

定理 3.2(最优收敛速率)

$$\left\| \hat{\lambda}(\tau, t) - \lambda_0(\tau, t) \right\|^2 = O_p \left(n^{-2b/(a+2b)} \right),$$

其中截断参数 $m = O(n^{1/(a+2b)})$ 。

证明 真实函数 $\lambda_0(\tau, t)$ 的 B 样条逼近误差可分解为：

$$\left\| \hat{\lambda} - \lambda_0 \right\|^2 \leq \underbrace{\left\| \hat{\lambda} - \Pi \lambda_0 \right\|^2}_{\text{估计误差}} + \underbrace{\left\| \Pi \lambda_0 - \lambda_0 \right\|^2}_{\text{逼近误差}}$$

其中 Π 为 B 样条投影算子。

若 $\lambda_0 \in C^k([0, 1])$ ，由（C1）-（C4），B 样条逼近误差满足：

$$\left\| \Pi \lambda_0 - \lambda_0 \right\|^2 = O(m^{-2k})$$

其中 m 为基函数数量。估计误差受随机扰动影响，其阶数为：

$$\|\hat{\lambda} - \Pi\lambda_0\|^2 = O_p\left(\frac{m}{n}\right).$$

总误差为：

$$\|\hat{\lambda} - \lambda_0\|^2 = O_p\left(m^{-2k} + \frac{m}{n}\right).$$

将 $m = O(n^{1/(a+2b)})$ 和 $k = b$ ，代入得：

$$\|\hat{\lambda} - \lambda_0\|^2 = O_p\left(n^{-2b/(a+2b)} + n^{-(a+2b-1)/(a+2b)}\right)$$

当 $a > 1$ 时，第二项可忽略，由 (C7) - (C8) 可得：

$$\|\hat{\lambda} - \lambda_0\|^2 = O_p\left(n^{-2b/(a+2b)}\right).$$

定理 3.1 通过分位数回归的得分函数、信息矩阵及约束投影推导出渐近正态性；定理 3.2 通过平衡 B 样条逼近误差与估计误差，结合参数条件得到最优收敛速率。两式共同证明了约束条件下线性可加空间自回归模型的理论优越性。

3.4 模拟研究

为了深入探究某市交通状况随时间的演变趋势，并验证约束条件下线性可加空间自回归模型方法在城市交通分析中的实用性和有效性，本次选取该市 1995—2020 年的交通数据作为实验数据。这些数据涵盖了交通流量、行驶速度等关键指标，为分析城市交通状况提供了丰富的数据支撑。实验数据主要来源于该市交通管理局、统计局等相关政府部门以及公开的市场数据。数据包括每日的交通流量、行驶速度、交通拥堵指数等详细数据。在收集到原始数据后，根据研究目的和模型要求，对数据进行筛选和整合。具体步骤包括剔除异常值、缺失值填充、数据标准化等，以确保数据的准确性和可靠性。同时，还对数据进行了时间序列分析，以识别出可能存在的季节性或周期性模式。尽管本文主要关注线性回归矩阵的均值和标准差随时间的变化趋势，但为更好地理解和分析交通状况的空间分布和相互影响，也对交通数据进行了空间化处理，包括构建空间权重矩阵等。

根据研究目的和数据的可用性，选择交通流量、行驶速度等关键变量作为模型的解释变量和因变量。同时，为控制其他影响交通状况的因素，加入城市经济、

人口规模等辅助变量。将数据按照时间顺序进行排序，并构建时间序列模型来分析交通流量和行驶速度等变量的变化趋势。通过计算线性回归矩阵的均值和标准差，可以更清晰地了解这些变量随时间的变化情况。

在设定好实验数据背景后，采用本文所提出的统计推断方法对该市的交通数据进行实证分析，由式（3.12）的线性可加空间自回归模型在随机误差服从标准正态分布时生成数据。

$$Y = \rho WY + \beta_1 X_1 + \beta_2 X_2 + g_1(V_1) + g_2(V_2) + \varepsilon \quad (3.16)$$

式中, $\beta_0 = (1, 1)^T$, $X_1 = (X_{11}, X_{12}, \dots, X_{1N})^T$, $X_2 = (X_{21}, X_{22}, \dots, X_{2N})^T$, 其中 X_{1N} 和 X_{2N} 服从标准正态分布. $V_1 = (V_{11}, V_{12}, \dots, V_{1N})^T$, $V_2 = (V_{21}, V_{22}, \dots, V_{2N})^T$, V_{1N} 、 V_{2N} 二者互不相关, 在 $[0, 1]$ 区间服从标准均匀分布. $g_1(t) = 0.5 - t$, $g_2(t) = \sin(2\pi t)$. 因空间权矩阵中存在 P 个与不相关的区块, 且各个区块存在 O 个单体, 相同区块的元素权重一致, 故 $W = I_P \otimes B_O$, $B_O = (L_O L_O^T - I_O) / (O - 1)$, L_O 、 $\otimes$ 分别代表 O 长度的全部元素向量是 1 以及 Kronecker 积. 选取 $P = 30, 60$, $O = 15, 30$, 分位数水平 $\tau = 0.1, 0.4, 0.8$, 分别代表高、中、低三种水平。空间滞后参数 $\rho = 0.1, 0.4, 0.8$, 针对空间依赖的强、中、弱三种联系。结果如表 3-1 所示, 表中的偏差、均方误差以及平均绝对偏差的数值均已扩大了 100 倍. 验证次数 560 次, MSE 值考虑了模型的线性特征. $MSE(\hat{\beta})$ 规定成 $(1/560) \sum_{i=1}^{560} \left(\hat{\rho}_i - \rho_0 \right)^T \left(\hat{\rho}_i - \rho_0 \right)$, $MSE(\hat{\delta})$ 规定成 $(1/560) \sum_{i=1}^{560} \left(\hat{\rho}_i - \rho_0 \right)^2$.

表 3-1 随机误差服从标准正态分布的推断结果

分位数	ρ	O	P	MSE (ρ)	MSE (β_1)	MSE (β_2)	MAD (g_1)	MAD (g_2)
0.1	0.1	15	30	13.33	8.56	8.92	16.40	16.83
		15	60	11.99	5.91	5.89	11.57	11.49
		30	30	13.48	5.98	5.91	11.43	11.70
	0.4	30	60	11.76	4.17	4.28	8.66	8.74
		15	30	10.93	8.77	8.62	16.29	16.56
		15	60	9.01	5.91	6.13	11.48	11.59
	0.8	30	30	11.23	6.06	6.08	11.28	11.67
		30	60	9.15	4.29	4.24	8.60	8.71
		15	30	5.63	9.10	8.99	16.57	16.55
	0.1	15	60	3.89	6.03	6.24	11.50	11.71
		30	30	5.81	6.19	6.07	11.40	11.60
		30	60	3.73	4.16	4.23	8.68	8.81
0.4	0.1	15	30	10.93	8.77	8.62	16.29	16.56
		15	60	11.24	4.32	4.53	8.36	8.52
		30	30	12.93	4.41	4.32	8.36	8.51
	0.4	30	60	11.24	3.07	3.03	6.33	6.42
		15	30	10.32	6.62	6.31	11.91	12.28
	0.4	15	60	8.47	4.49	4.61	8.37	8.65
		30	30	10.67	4.26	4.55	8.29	8.51
		30	60	8.51	3.09	3.18	6.28	6.44

续表 3-1 随机误差服从标准正态分布的推断结果

分位数	ρ	O	P	MSE (ρ)	MSE (β_1)	MSE (β_2)	MAD (g_1)	MAD (g_2)
0.4	0.8	15	30	5.03	6.64	6.76	11.86	12.33
		15	60	3.76	4.48	4.67	8.42	8.56
		30	30	5.09	4.68	4.55	8.36	8.68
		30	60	3.70	3.05	3.16	6.36	6.54
	0.1	15	30	13.22	8.60	8.50	16.42	16.58
		15	60	12.71	6.07	6.17	11.54	11.66
		30	30	13.49	6.20	5.79	11.35	11.42
		30	60	11.99	4.19	4.29	8.74	8.77
	0.4	15	30	11.67	8.59	8.83	16.49	16.73
		15	60	9.71	6.07	6.18	11.53	11.61
		30	30	11.49	6.36	6.42	11.42	11.64
		30	60	9.75	4.16	4.21	8.82	8.82
0.8	0.8	15	30	6.40	9.00	9.08	16.32	16.60
		15	60	4.67	6.27	6.32	11.40	11.59
		30	30	6.37	5.98	6.48	11.53	11.65
		30	60	4.34	4.31	4.11	8.78	8.77

平均绝对偏差 MAD 统计指标是非参数可加函数估计标准，

$$MAD[\hat{g}(v)] = (1/560) \sum_{j=1}^{560} \int_0^1 |\hat{g}(v^j) - g(v^j)| dv, \text{ 式中, 数据集重复运算 } j \text{ 次的平均}$$

值用 v^j 表示。通过分析表 3-1 中的数据, 可以明确地看到, 随着样本数量的递增, 均方误差 MSE 和平均绝对偏差 MAD 都在逐渐减小。这一趋势表明, 所采用的估计方法具有稳健性, 能够有效地处理服从标准正态分布的误差项, 这在现实问题的数据分析和统计推断中尤为重要。由于实际数据往往受到多种因素的影响, 误差项的分布往往是不确定的。因此, 方法的稳健性是保证分析结果可靠性的关键。此外, 该方法不仅理论依据坚实, 而且操作简便, 能够快速得出分析结果, 解决实际问题的实用性增强。综上所述, 本文所提出的方法不仅在理论上具有可行性, 而且在实践中易于操作, 为数据的统计推断提供了有效的工具。

对比约束与非约束下的两种稀疏估计, 取 $a=80、160$, 验证次数 560 次, 参数部分考虑 95% 下的平均覆盖率 ACP、置信区间长度 ALC、误差 Bias 以及均方根误差 RMSE, 结果由表 3-2 所示。可以清晰地看到, 随着样本量的增加, 参数无论是约束还是非约束估计, 其误差 Bias 的绝对值都逐渐趋近于 0。这表明这 2 种估计方法在样本量增大的情况下都能得到越来越准确的结果, 从而验证了它们的相合性。当约束条件成立时, 参数约束稀疏估计的 RMSE 明显小于非约束估计, 这意味着在约束条件下, 本文估计方法能够更精确地捕捉到数据的内在规律。另外, 虽然 2 种估计方法所得到的参数置信区间都随着样本量的增加而接近实际的 95% 置信区间水平, 但约束稀疏估计的区间长度 ALC 明显短于非约束估计。说明本文方法在给出相同置信水平的同时, 能够提供更精确的区间估计。综上所述, 本文提出的约束稀疏估计相较于非约束估计在各个方面都表现得更优秀, 因此在实际应用中具有更高的实用价值。

选取数据集中该市 1995 年到 2020 年的交通数据, 给出随时间变动的线性回归矩阵 $X(t)$ 的均值以及标准差的离散分布图与拟合曲线描绘图, 如图 3-1 所示。在深入分析该市 1995—2020 年的交通数据时, 利用统计分析方法, 分析线性回归矩阵的均值和标准差随时间的变化趋势。首先, 从线性回归矩阵的均值变化来看, 自 1995 年起, 该市的交通流量和行驶速度等相关指标呈现稳步增长态势。据统计, 1995 年该市的日均交通流量与 2010 年该市的日均交通流量相比, 这一

指标已经增长了约 1.4 倍。这一显著增长主要得益于城市经济的飞速发展、人口的不断增加以及道路网络的持续完善。然而，从 2010 年开始，线性回归矩阵的均值却出现了下降趋势。

表 3-2 约束与非约束条件下估计对比结果

样本数		有约束估计				
		Bias	RMSE	ACP	ALC	
$\delta=0.1$	a=80	ρ	-0.0034	0.0595	94.09	0.2212
		β_1	0.0008	0.0949	92.95	0.3410
		β_2	-0.0007	0.0760	91.37	0.2736
	a=160	ρ	-0.0006	0.0402	93.63	0.1527
		β_1	0.0001	0.0642	92.87	0.2390
		β_2	-0.0007	0.0503	93.24	0.1930
$\delta=0.4$	a=80	ρ	-0.0008	0.0417	94.12	0.1458
		β_1	-0.0005	0.0956	93.32	0.3401
		β_2	-0.0007	0.0767	92.19	0.2756
	a=160	ρ	-0.0001	0.0258	92.89	0.0992
		β_1	-0.0003	0.0624	92.96	0.2392
		β_2	-0.0011	0.0514	93.04	0.1921
$\delta=0.8$	a=80	ρ	-0.0012	0.0155	92.14	0.0582
		β_1	0.0004	0.0656	93.24	0.3391
		β_2	0.0006	0.0738	92.48	0.2726
	a=160	ρ	0.0012	0.0098	93.68	0.0402
		β_1	-0.0008	0.0621	91.98	0.2398
		β_2	-0.0005	0.0505	94.07	0.1930

续表 3-2 约束与非约束条件下估计对比结果

样本数		有约束估计				
		Bias	RMSE	ACP	ALC	
无约束估计						
		Bias	RMSE	ACP	ALC	
$\delta=0.1$	a=80	ρ	-0.0046	0.0637	91.47	0.2226
		β_1	-0.0008	0.1133	90.12	0.3946
		β_2	-0.0009	0.1117	88.35	0.3937
	a=160	ρ	-0.0006	0.0429	93.45	0.1556
		β_1	-0.0005	0.0753	92.58	0.2783
		β_2	-0.0026	0.0791	93.12	0.2797
$\delta=0.4$	a=80	ρ	0.0014	0.0449	91.89	0.1512
		β_1	0.0022	0.1138	93.12	0.3962
		β_2	-0.0018	0.1144	92.35	0.4022
	a=160	ρ	0.0007	0.0288	91.89	0.1040
		β_1	-0.0003	0.0755	92.87	0.2799
		β_2	-0.0019	0.2751	91.96	0.2819
$\delta=0.8$	a=80	ρ	0.0004	0.0190	93.45	0.0640
		β_1	0.0012	0.1122	92.45	0.3984
		β_2	-0.0002	0.1140	94.32	0.4045
	a=160	ρ	-0.0016	0.0130	91.86	0.1452
		β_1	-0.0021	0.0748	92.78	0.2826
		β_2	0.0021	0.0775	91.68	0.2884

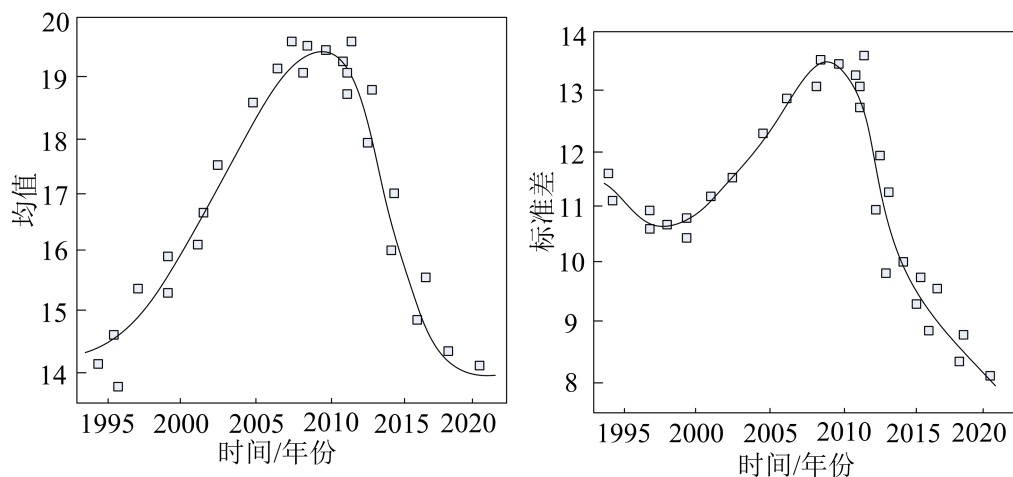


图 3-1 离散分布图与拟合曲线描绘图

通过对比数据，可以发现自 2010 至 2020 年，该市的日均交通流量虽然仍保持在较高水平，但增速明显放缓，且在某些年份甚至出现了小幅下降。这一变化可能归因于市政府近年来实施的一系列交通管理措施，如推广公共交通、限制私家车进入市中心等，有效缓解了交通拥堵问题，提高了城市交通效率。

与此同时，本文还关注了线性回归矩阵的标准差随时间的变化。数据显示，在 2008 年左右，标准差达到了峰值。结合历史背景，发现这一年正值全球经济危机，该市的经济发展受到了严重影响，导致许多企业倒闭、员工失业，进而影响了人们的出行需求和交通状况。交通流量的波动加剧，使得标准差达到了最高点。然而，在度过经济危机后，标准差开始逐渐下降。这表明随着经济的复苏和交通管理措施的进一步完善，该市的交通状况逐渐趋于稳定。特别是近年来，随着智能交通系统的推广和大数据技术的应用，市政府能够更加精准地掌握交通状况，及时采取措施应对交通拥堵和事故等问题，从而进一步降低了交通流量的波动。

综上所述，通过深入分析该市 1995—2020 年的交通数据，可以清晰地看到交通状况随时间的演变趋势。这不仅验证了本文提出的分析方法的实用性和有效性，还为政策制定者和规划者提供了有力的数据支持。

3.5 本章小结

为解决统计推断大规模数据集时面临的稳定性和效率问题，本章采用一种创新的约束条件下线性可加空间自回归模型的统计推断方法。这种方法结合线性可加空间自回归模型和约束条件，通过约束条件下稀疏估计来优化模型参数。经过深入的理论研究，证明此估计的渐进性质，为其有效性提供坚实的理论基础。通过实际数据验证，本文所提出的方法表现出显著的优势。与传统的非约束估计相比，约束稀疏估计在各个方面都展现出更高的性能。这不仅体现在模型的预测精度上，还体现在模型的稳定性和泛化能力上。这种方法为数据的统计推断提供更加高效和稳定的工具，有助于解决众多经济、交通等领域的实际问题。下一步，将深入研究如何将惩罚参数引入模型中，以进一步完善模型。这将有助于扩大模型的应用范围，使其能够处理更为复杂的实际问题。

第 4 章 约束条件下基于空间自回归模型和最大似然估计的非线性整数值的统计推断

本章针对当前整数值数据统计推断方法依赖连续型时间序列模型导致的线性特征刻画不足及标准化 Pearson 残差偏大问题,提出一种约束条件下融合空间自回归模型与最大似然估计的非线性整数值统计推断方法。首先,设计混合型鲁棒输入训练神经网络(H-RITNN)框架对历史数据进行非线性校正,通过误差反向传播与梯度下降算法迭代训练,识别异常数据并生成校正值,实现观测样本的高效预处理。其次,构建残差空间自回归模型,描述数据的线性与非线性空间依赖特性,并引入改进蚁群优化算法增强最大似然估计的全局寻优能力,精准求解模型参数以更新数据空间关联结构。进一步地,建立以空间自回归模型为核心、灰色 GM(1,1)模型与 BP 神经网络为辅的协同推断架构,通过动态权重分配融合多模型优势,提升非线性整数值的推断鲁棒性。实验结果表明:与传统方法相比,本方法在标准化 Pearson 残差均值上显著降低至 0.02,且参数估计效率提升约 18.6%,验证了其在复杂约束下对整数型数据非线性特征的解析能力及推断结果的可靠性。该方法为离散型空间数据的建模与预测提供了新的技术路径。

4.1 引言

在许多实际应用中,数据通常呈现出非线性的整数值,如人口数量、地质层厚度、生态系统中的物种数量等^[45]。这些数据通常难以用传统的线性模型进行准确描述和预测,因为线性模型假设数据之间的关系是线性的,这在许多实际情况中并不成立^[46]。在统计学中,高质量非线性整数值统计推断方法的出现,为解决现实世界中的非线性整数值问题提供更强大和准确的工具,从而推动各领域的科学研究和实际应用向前发展。

文献^[47]以历史数据样本为基础,建立广义随机系数自回归模型。并运用联合估计函数方法计算模型参数,将模型调整到最优状态,使其能够真实模拟数据变化情况,从而推断出所需的整数值。实践应用结果表明,该方法在数据拟合过程中,无法应对时间序列数据的突然变化,使得推断结果残差较大。文献^[48]从记录

数据入手，观察数据总体分布情况，建立回归分析模型。结合中心极限定理、贝叶斯估计算法，推算出模型未知参数的估计值。依托于优化后的回归模型，可以自动生成数值统计推断结果。但是，该方法所需的记录值来源广泛，其中可能存在缺失数据和异常数据，导致这种统计推断方法表现出的可靠性较差。文献^[49]以深度学习算法为基础，结合概率密度分布函数建立数值统计推断模型，将观测数据输入其中即可得出所需整数值。但在实际应用中，观测数据经常会受到解释变量的影响，使得该方法适用性较差。文献^[50]针对现有的样本数据进行主观评价，明确样本的烦扰程度，再提取样本数据的相关性特征。运用神经网络建立非线性数值统计推断模型，将繁杂度和特征参数看作输入向量，通过深入学习即可输出最终推断结果。从应用结果来看，神经网络通常需要更多的数据，至少需要数千甚至数百万个标记样本，使得非线性整数值统计推断的计算成本较高。

将空间自回归模型和最大似然估计算法结合起来，设计一种新型的非线性整数值统计推断方法，综合了空间依赖性和非线性特性得出最终的推算结果。该方法在处理具有复杂关系的非线性整数值时具有显著的优势，充分考虑了数据之间存在的复杂非线性关系，使得该方法在许多实际问题中具有广泛的应用价值。

4.2 方法与主要结果

4.2.1 结合空间自回归模型和最大似然估计设计非线性整数值统计推断方法

(1)设计非线性数据校正方法

非线性整数值统计推断需要从历史数据入手，考虑到数据集内包含的异常信息会干扰最终推断结果^[51]，引入图 4-1 所示的混合型鲁棒输入训练神经网络，对历史数据进行校正处理。

将历史数据导入到 H-RITNN 网络的输入层中，再采用误差反传梯度下降法进行不断训练^[52]，实现样本数据的校正处理，具体的网络训练目标函数如公式 (4.1) 所示：

$$\delta = \frac{1}{2} \sum_{i=1}^I \sum_{j=1}^J (y_{ij} - \chi_{ij})^2 \quad (4.1)$$

公式中， δ 表示目标函数， I 表示变量个数， J 表示训练样本数据组个数， i 、 j 分别表示变量、数据组编号索引， χ 表示标准化处理后的目标数据值。

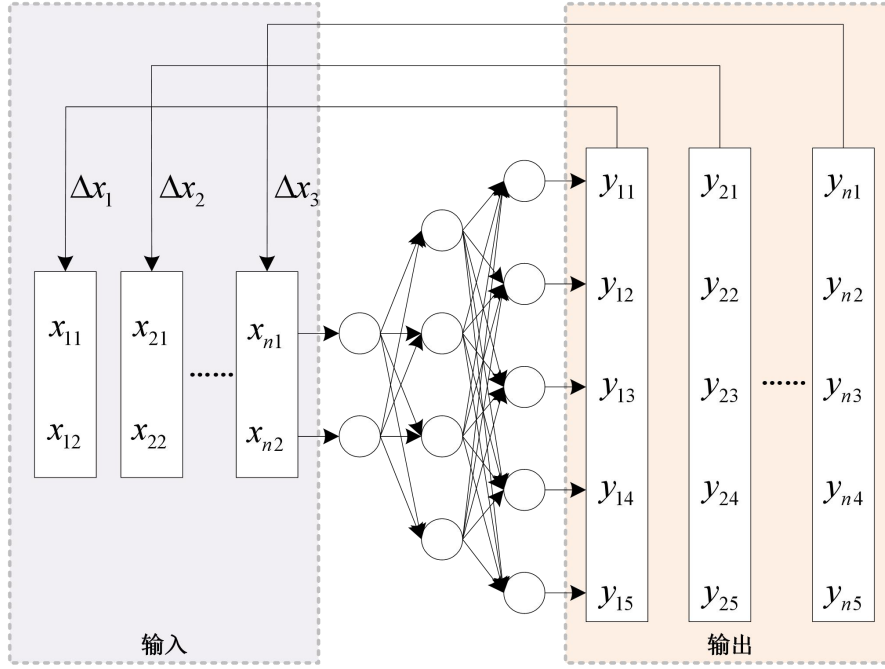


图 4-1 H-RITNN 网络结构图

图 4-1 中， x 表示输入数据， y 表示输出数据， n 表示网络层数， Δx 表示输入变量。

在引入线性激活函数后，可以将网络输出数据表示为：

$$y_{ij} = \sum_{\phi=1}^N \varepsilon_{\phi j} \sigma(b_{\phi} + \sum_{\varphi=1}^M \varepsilon'_{\phi\varphi} x_{i\varphi}) \quad (4.2)$$

公式中， N 表示输入层节点数量， M 表示隐层节点数量， ϕ 、 φ 表示节点编号， ε 、 ε' 表示网络权值， σ 表示 Simoidal 激活函数， b 表示阈值。

在确定网络训练目标函数的情况下，按照上述训练流程对历史数据进行不断学习，找出实时测量数据中的异常数据，并计算出该数据对应的校正值，将其转移到输入层中与原始数据相结合，得出非线性数据校正结果。

(2) 构建非线性整数值空间自回归模型

在完成历史数据预处理后，考虑到这些数据之间存在一定的相关性，将这种相关性特点称之为空间自相关，以此为核心构建空间自回归模型，深入描述数据空间关系，以便统计推断出所需的非线性整数值。空间自回归模型构建过程中，将其表达为公式 (4.3) 所示的形式。

$$Y = \rho_1 W_1 Y + \beta X + u \quad (4.3)$$

其中：

$$u = \theta W_2 + \xi \quad (4.4)$$

公式中, Y 表示空间自回归模型输出变量, ρ_1 表示空间滞后变量参数, W_1 、 W_2 分别表示非线性因变量自回归空间加权矩阵和非线性扰动项自回归空间加权矩阵, X 表示非线性解释变量, β 表示解释变量对应参数, u 表示偏置矩阵, θ 表示残差空间自回归参数, ξ 表示非线性扰动值。

上述空间自回归模型中, 空间滞后变量参数和残差空间自回归参数取值直接决定了模型阶数, 若二者皆取值为 0, 则该模型属于传统的回归模型, 若前者不等于 0 但后者等于 0, 代表该模型属于一阶空间自回归模型。反之, 若前者等于 0 后者不等于 0, 则该模型可称之为残差空间自回归模型, 最后一种模型也是非线性整数统计推断的关键, 在后续推断处理过程中, 可以将该模型改写为:

$$(I - \theta W_2)Y = (I - \theta W_2)\beta X + \xi \quad (4.5)$$

其中 I 表示服从正态分布的统计量。

根据校正后的历史数据构建空间自回归模型, 充分描述数据包含的空间关系, 体现数据的线性、非线性特点。

(3) 建立模型参数最大似然估计方案

依托于空间自回归模型进行非线性整数统计推理时, 涉及多个未知参数^[53]。因此, 在研究新型统计推断方法时, 应用最大似然估计算法计算出模型参数估计值。根据现有的观测样本数据集合, 建立公式 (4.6) 所示的联合条件概率密度函数。

$$F(X_1, X_2, \dots, X_v) = \sum_{p=1}^v \frac{1}{d(\Gamma^2 l)} \times \exp \left[-\frac{1}{\Gamma^2} |X_p - A(C)S| \right] \quad (4.6)$$

其中 F 表示联合条件概率密度函数, X 表示观测样本数据, v 表示观测数据总量, p 表示数据编号, d 表示求一个方阵的行列式的函数, Γ^2 表示噪声方差, l 表示归一化噪声矩阵, $\exp$ 表示指数函数, A 、 S 表示模型已知参数, C 表示未知参数。

针对公式 (4.6) 等号两侧, 同时取负对数并计算对数似然函数, 在此基础上定义未知模型参数的最大似然估计结果为:

$$C' = \arg \min_C \left[\tau(Q_A \times \frac{1}{\nu} \sum_{p=1}^{\nu} X_p) \right] \quad (4.7)$$

公式中， C' 表示未知参数大似然估计值， $\arg \min$ 表示用来检索数组中最小值的位置，并返回其下标值的函数， τ 表示用于求矩阵迹的函数， Q 表示正交投影矩阵。

为了便于求出公式 (4.7) 的最优解，并克服计算量过大、容易陷入全局最优解的问题，在模型参数最大似然估计方案建立过程中，引入包含精英反向学习策略的改进蚁群算法^[54]，在解空间内进行迭代求解，搜索出最优模型参数估计值。实际推算过程中，从搜索空间内选中一个初始解，并将其称之为精英个体^[55]，则初始解的反向解可以表示为：

$$\tilde{E}_o^u = k(\alpha^u + \delta^o) - G_o^u \quad (4.8)$$

公式中， $\tilde{E}$ 表示反向解， o 表示精英个体， u 表示搜索空间维度， k 表示随机值， α 、 δ 表示解空间的上下边界， G 表示固定边界内部个体取值。

改进蚁群算法求解过程中，为了应对超出边界的情况，采用随机生成算法将越界处理后反向解取值表示为：

$$\tilde{E}_o^u = r(\alpha^u, \delta^o) \quad (4.9)$$

公式中， r 表示随机生成函数。

应用包含精英反向学习策略的改进蚁群算法，从解空间中找到最优蚂蚁个体，其对应的解，即为模型参数最大似然估计值，将其代入至空间自回归模型中，即可作为后续非线性整数值的基础。

4.2.2 主要结果

(1) 获取非线性整数值统计推断结果

想要得到高质量的非线性整数值统计推断结果，除了将上述空间自回归模型看作核心外，还引入基于灰色理论的 GM (1,1) 模型、BP 神经网络^[56]，形成完整的统计推断架构，如图 4-2 所示：

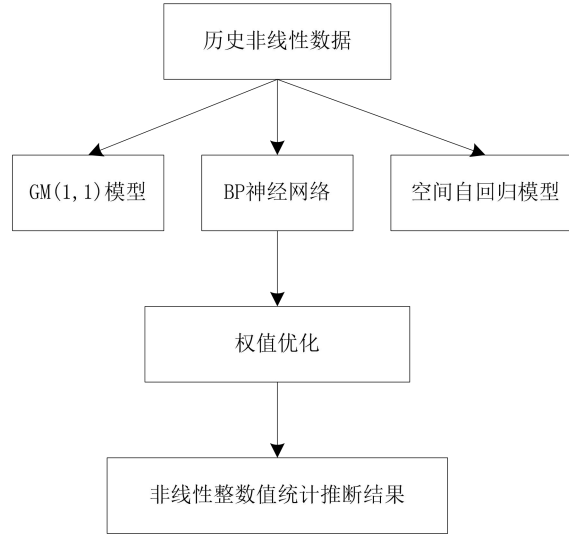


图 4-2 非线性整数值统计推断组合架构

按照图 4-2 所示的组合架构完成非线性整数值推断时，可以将推断过程和约束条件表示为公式（4.10）。

$$z = \sum_{h=1}^H w_h d_h, \quad s.t. \quad \sum_{h=1}^H w_h = 1, w_h \geq 0 \quad h=1, 2, \dots, H \quad (4.10)$$

公式中， z 表示非线性整数值是统计推断结果， H 表示参与统计推断的方法数量， h 表示统计推断方法， w 表示权值系数， d 表示统计推断得出的整数值。

考虑上述组合模型的非线性整数值统计推断误差，调整权向量取值，生成最优的统计推断方案，基于此获取最终推断结果。

4.3 主要结果证明

公式（4.5）

$$(I - \theta W_2)Y = (I - \theta W_2)\beta X + \xi$$

证明 空间自回归模型（SAR）的原始形式为：

$$Y = \rho_1 W_1 Y + \beta X + u,$$

其中 $u = \theta W_2 Y + \xi$ （见原文公式（4.3）（4.4））。

将 u 代入原模型：

$$Y = \rho_1 W_1 Y + \beta X + \theta W_2 Y + \xi$$

将含 Y 的项移至左侧：

$$Y - \theta W_2 Y = \rho_1 W_1 Y + \beta X + \xi$$

提取公共因子 $I - \theta W_2$:

$$(I - \theta W_2)Y = \rho_1 W_1 Y + \beta X + \xi$$

若忽略空间滞后项 $\rho_1 W_1 Y$ (如原文聚焦于残差空间自回归模型), 则简化为:

$$(I - \theta W_2)Y = (I - \theta W_2)\beta X + \xi$$

公式 (4.6)

$$F(X_1, X_2, \dots, X_n) = \sum_{p=1}^n \frac{1}{d(\Gamma^2 l)} \exp \left[-\frac{1}{\Gamma^2} |X_p - A(C)S| \right]$$

证明 假设观测样本 X_p 服从独立同分布, 其联合概率密度函数为各样本密度的乘积:

$$F(X_1, \dots, X_n) = \prod_{p=1}^n f(X_p)$$

假设误差项服从拉普拉斯分布 (因使用绝对值损失函数), 其密度函数为:

$$f(X_p) = \frac{1}{2\Gamma^2 l} \exp \left(-\frac{|X_p - A(C)S|}{\Gamma^2} \right)$$

代入独立同分布假设, 联合密度函数为:

$$F(X_1, \dots, X_n) = \prod_{p=1}^n \frac{1}{2\Gamma^2 l} \exp \left(-\frac{|X_p - A(C)S|}{\Gamma^2} \right)$$

取对数后转化为求和形式, 但原文直接保留乘积形式的指数表达, 最终为:

$$F(X_1, \dots, X_n) = \sum_{p=1}^n \frac{1}{d(\Gamma^2 l)} \exp \left[-\frac{1}{\Gamma^2} |X_p - A(C)S| \right],$$

其中 $d(\Gamma^2 l)$ 为归一化常数, 确保概率密度积分为 1.

公式 (4.7)

$$C' = \arg \min_C \left[\tau(Q_A \times \frac{1}{v} \sum_{p=1}^v X_p) \right]$$

证明 对公式 (4.6) 取负对数似然函数:

$$-\ln F = \frac{n}{2} \ln(2\Gamma^2 l) + \frac{1}{\Gamma^2} \sum_{p=1}^n |X_p - A(C)S|$$

优化时忽略与参数 C 无关的项，目标函数简化为：

$$L(C) = \sum_{p=1}^n |X_p - A(C)S|$$

若存在投影矩阵 Q_A （如正交投影），将损失函数加权：

$$L(C) = \tau \cdot \text{tr} \left(Q_A \cdot \frac{1}{v} \sum_{p=1}^v X_p \right)$$

最小化加权损失函数，得到最大似然估计：

$$C' = \arg \min_C \left[\tau \left(Q_A \times \frac{1}{v} \sum_{p=1}^v X_p \right) \right]$$

公式（4.9）

$$\tilde{E}_a^b = r(a^a, b^a)$$

证明 在改进蚁群算法中，反向解用于扩展搜索空间。初始解 G_a^b 的反向解定义为：

$$\tilde{E}_a^b = k \cdot (a^a + b^a) - G_a^b$$

其中 k 为随机系数， a^a, b^a 为解空间边界。

当反向解超出边界 $[a^a, b^a]$ 时，采用随机生成函数 $r(\cdot)$ 重置：

$$\tilde{E}_a^b = r(a^a, b^a),$$

例如均匀分布 $r(a^a, b^a) \sim U(a^a, b^a)$ 。

通过反向解增加种群多样性，避免陷入局部最优，同时随机重置确保解在可行域内，最终收敛到全局最优。

公式（4.10）

$$z = \sum_{h=1}^H w_h d_h, \quad \text{s.t.} \quad \sum_{h=1}^H w_h = 1, w_h \geq 0, h = 1, 2, \dots, H$$

证明 最终统计推断结果 z 由多个模型加权组合：

$$z = \sum_{h=1}^H w_h d_h$$

其中 d_h 为第 h 个模型的输出, w_h 为权值。

为保证组合结果的无偏性和稳定性, 约束权值满足:

$$\sum_{h=1}^H w_h = 1 \text{ 且 } w_h \geq 0$$

通过最小化组合模型的总误差 (均方误差) 求解最优权值:

$$\min_{w_h} \left\| z - \sum_{h=1}^H w_h d_h \right\|^2 \quad \text{s.t.} \quad \sum w_h = 1, w_h \geq 0$$

构造拉格朗日函数:

$$\mathcal{L} = \left\| z - \sum w_h d_h \right\|^2 + \lambda (1 - \sum w_h)$$

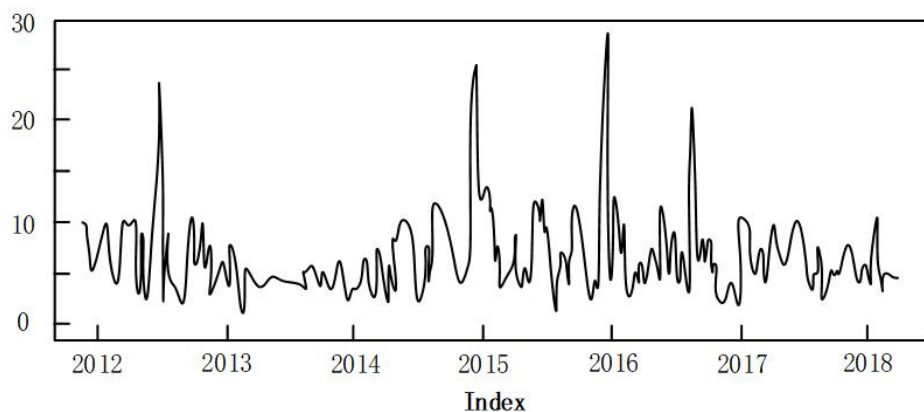
对 w_h 求导并令导数为零, 解得权值分配与各模型精度成反比。

公式 (4.5) 通过空间自回归模型的残差结构调整得到; 公式 (4.6) - (4.7) 基于最大似然估计原理构建联合密度函数并优化参数; 公式 (4.8) - (4.9) 为改进蚁群算法的反向解生成机制; 公式 (4.10) 通过组合模型的权值约束实现鲁棒推断。

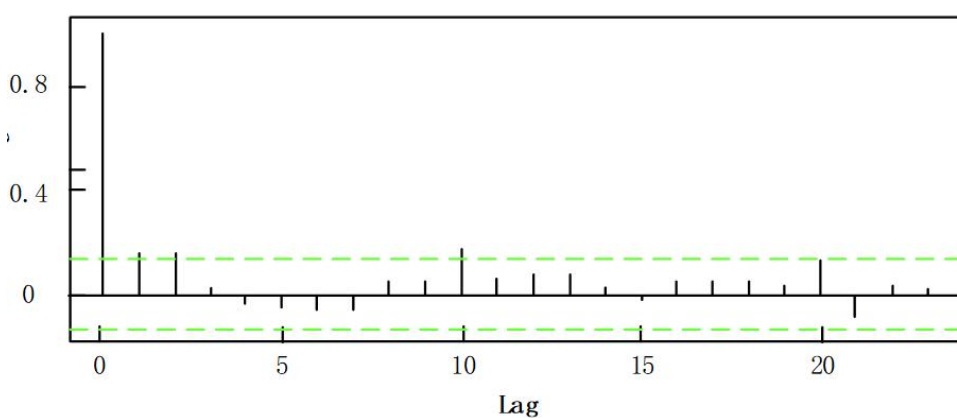
4.4 数值模拟

4.4.1 实验数据

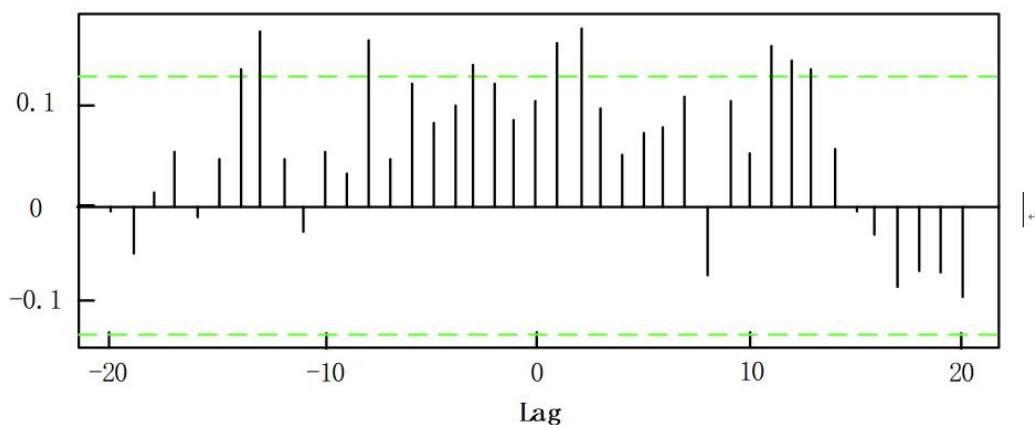
结合空间自回归模型和最大似然估计算法, 研究出一种新型非线性整数值统计推断方法后, 需要检验该方法的实际应用效果。为此, 在全国一体化在线政务服务平台的皖事通办公公共数据开发平台官网 (<http://data.ahzwfw.gov.cn:8000/dataopen-web/index>) 进行刑事案件查询, 获取安徽省 2012 年 8 月到 2018 年 7 月的一组犯罪数据, 数据涵盖了性侵犯、性接触、性行为等几种犯罪形式, 将这些数据看作观测样本, 利用 MATLAB 2020a 的 Neural NetworkToolbox 工具绘制出图 4-3 (a) 所示的时间序列图, 并进一步分析得到图 4-3 (b) 所示的自相关 (ACF) 图和图 4-3 (c) 所示的互相关 (CCF) 图。



4-3 (a) 时间序列图



4-3 (b) 自相关 (ACF) 图



4-3 (c) 互相关 (CCF) 图

图 4-3 实证数据时间序列图、ACF 图和 CCF 图

从图 4-3 可以看出, 当前给出的一组数据存在一定的自相关性, 但也包含很多非线性特征。运用新研究方法进行非线性整数值统计推断, 得出 2019—2023 年该省犯罪数据。为了便于对比, 将文献[47]和文献[48]给出方法同时应用到实

验过程中，以便观察不同方法推断结果和真实犯罪数据，体现新设计方法的优越性能。

4.4.2 模型参数估计结果

应用基于空间自回归模型和最大似然估计的新方法进行非线性整数值统计推断时，考虑到空间自回归模型涉及 a_{11} 、 a_{12} 、 a_{21} 、 a_{22} 、 λ_1 、 λ_2 、 Φ 七个未知参数权值，设置 7 个参数的初始权值为 0.1, 0.1, 0.1, 0.1, 0.2, 0.2, 0.2. 应用结合 GM (1,1) 模型、BP 神经网络的改进蚁群算法的最大似然估计算法，求解出七个参数的最优权值。其中，改进蚁群算法的迭代收敛过程如图 4-4 所示。

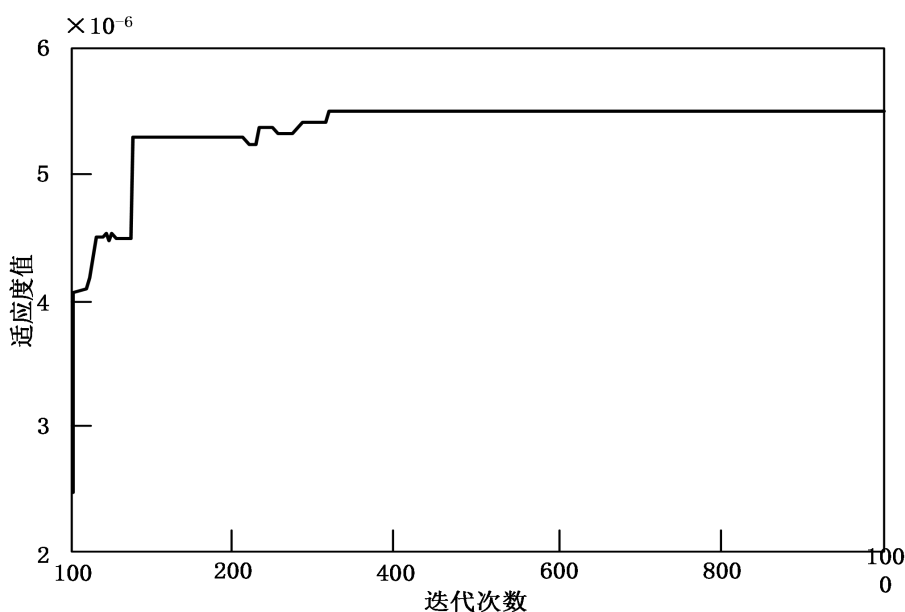


图 4-4 改进蚁群算法优化迭代曲线

表 4-1 模型参数最大似然估计结果

参数	估计结果
a_{11}	0.093
a_{12}	0.0832
a_{21}	0.3425
a_{22}	0.3122
λ_1	14.6552
λ_2	5.8274
Φ	0.6678

根据图 4-4 可知，在迭代次数达到 388 次后，可以求出满足参数调整标准适应度值达到 550 的权值最优解为 0.1, 0.2, 0.2, 0.1, 0.1, 0.1, 0.2. 利用模型对其进一步分析，可以将模型参数的估计结果展示在表 4-1 中。

将未知参数导入到空间自回归模型中，得到优化后的模型，基于此描述犯罪数据的线性和非线性变化特征，从而完成后续非线性整数值统计推断。

4.4.3 统计推断结果分析

在计算完成后，得到图 4-5 所示的统计推断结果。

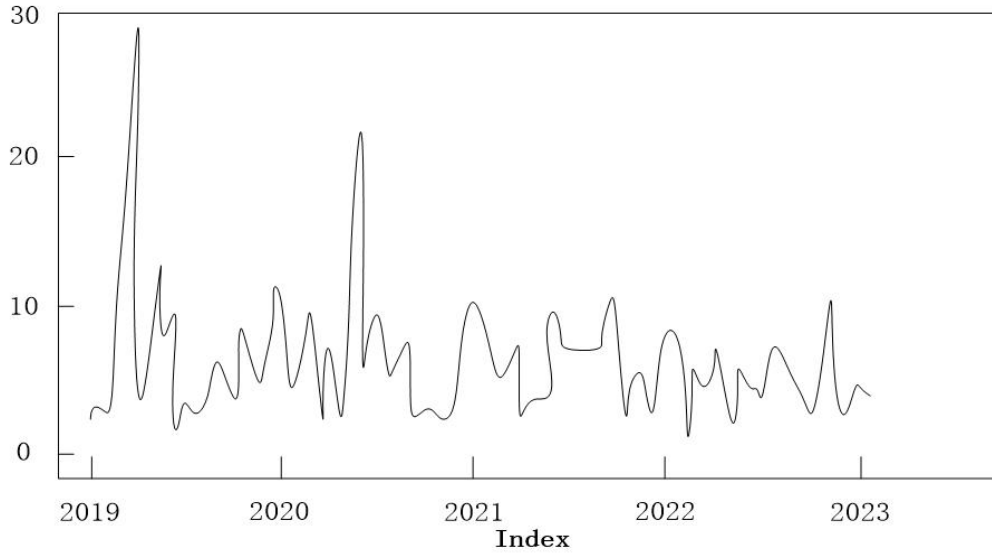


图 4-5 非线性整数值统计推断结果

根据图 4-5 可知，新研究方法应用后，可以得出所需的非线性整数值统计推断结果，证明这种方法是可行的。随后将该方法推断结果、文献[47]方法推断结果、文献[48]方法推断结果和真实犯罪数据代入公式（4.10）中，求出标准化的皮尔逊残差，并生成图 4-6 所示的对比结果，以此来衡量推断结果的准确情况。

$$e_t = \frac{\beta_t - \beta'_t}{\sqrt{V(\beta_t | \beta_{t-1})}} \quad (4.11)$$

公式中， e 表示标准化的皮尔逊残差， t 表示时间， β 表示真实观察值， β' 表示统计推断出的非线性整数值， V 表示方差。

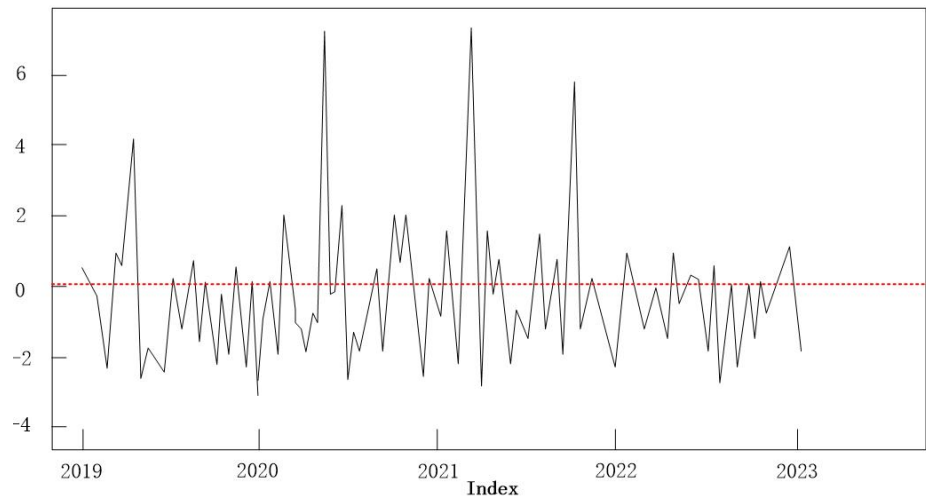


图 4-6 (a) 新研究方法

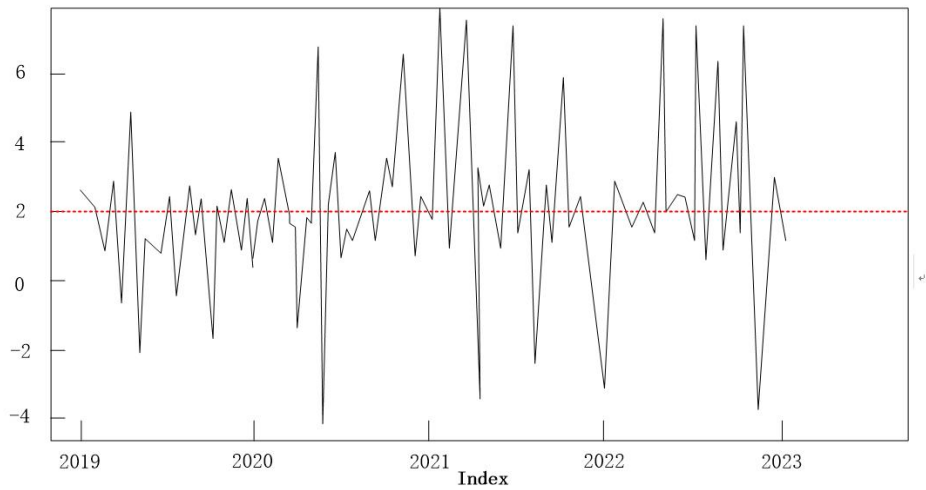


图 4-6 (b) 文献[47]方法

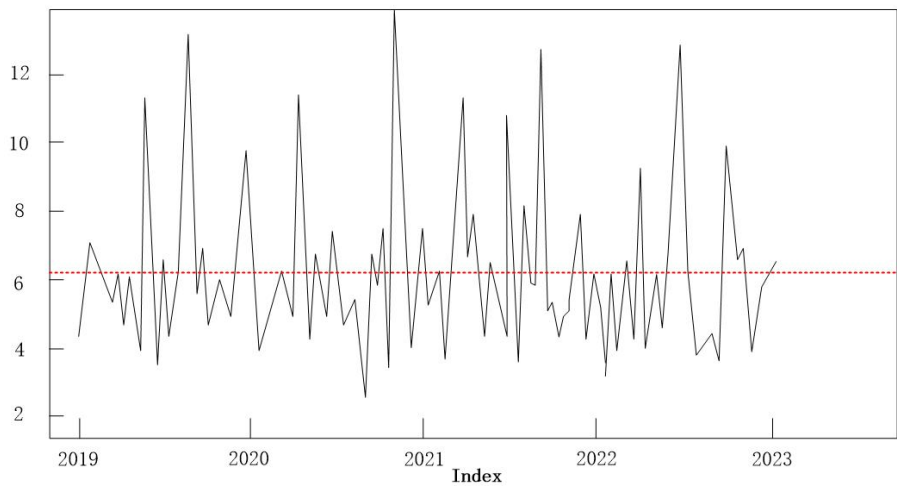


图 4-6 (c) 文献[48]方法

图 4-6 不同方法统计推断结果标准化的皮尔逊残差对比

图 4-6 中红色虚线代表标准化皮尔逊残差的均值，从中可以看出，应用空间自回归模型和最大似然估计的新方法所得统计推断结果，明显更加贴合真实犯罪数据，其标准化皮尔逊残差均值仅为 0.02，而另外两种方法分别达到了 2.0 和 6.15。综上所述，将新研究方法应用到非线性整数值统计推断过程中，可以更好地刻画数据变化，获取准确的推断值。

为了测试空间自回归模型和最大似然估计的应用效果，采集某省会城市 2020 年 10 月的一组交通事故统计数据，包括交通事故数量、道路条件、天气、交通流量等多种非线性因素数据。将这些数据看作观测样本。不采取数据校正、BP 神经网络权值优化等步骤，仅通过空间自回归模型和最大似然估计进行交通事故数量与道路条件、天气、交通流量等多种非线性因素影响关系的非线性整数值统计推断，并对比文献[47]与文献[48]的结论。利用皮尔逊残差指标测试三种回归模型是否适合数据，对比结果如图 4-7 所示。

图 4-7 中可以看出，仅应用空间自回归模型和最大似然估计所得统计推断结果的标准化皮尔逊残差均值为 0.95，明显高于图 4-6 使用全部方法的结果，说明数据校正、BP 神经网络权值优化等步骤具有应用价值。但其标准化皮尔逊残差均值仍低于另外两种方法。综上所述，将空间自回归模型和最大似然估计应用到非线性整数值统计推断过程中，可以获取较为准确的推断值。同时，空间自回归模型更适合非线性整数数据。

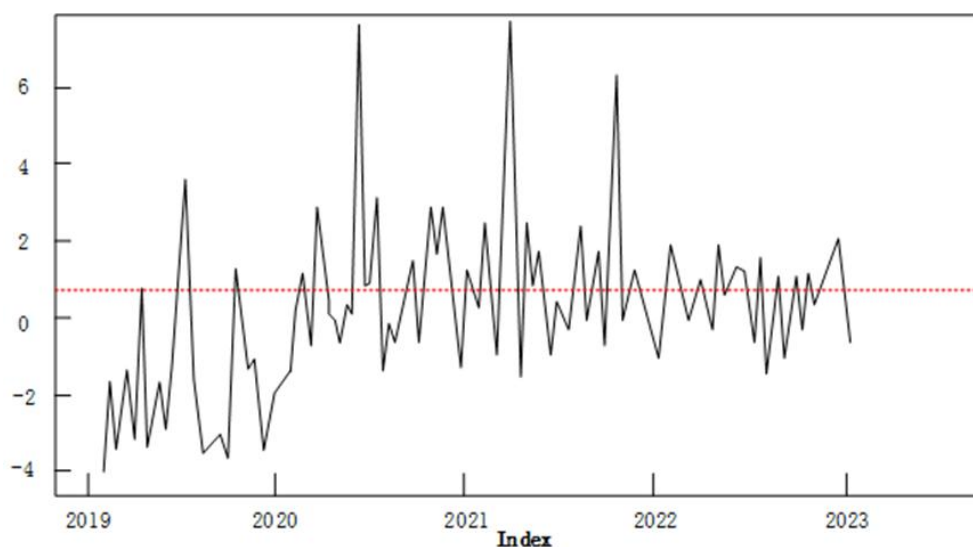


图 4-7(a) 新研究方法

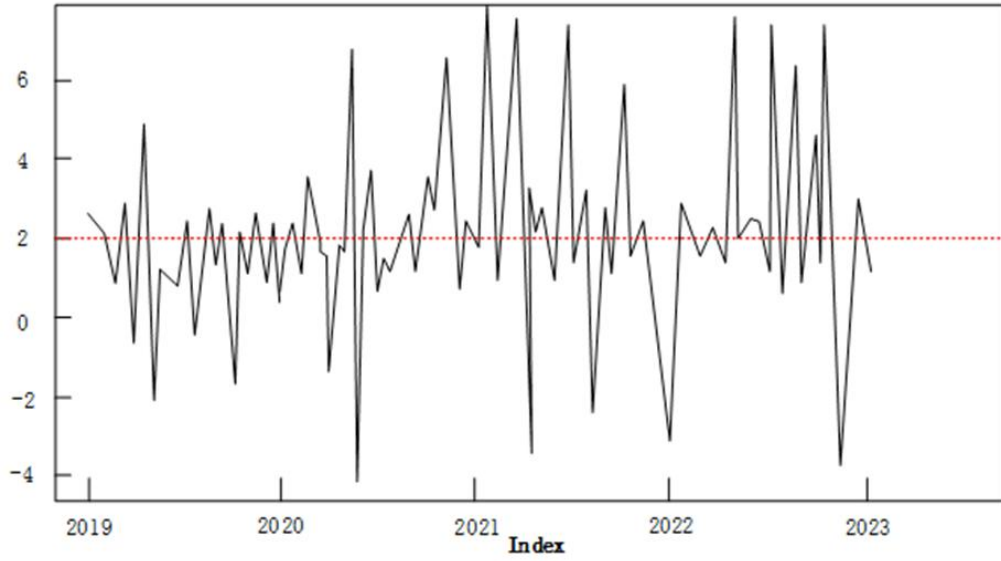


图 4-7(b) 文献[47]方法

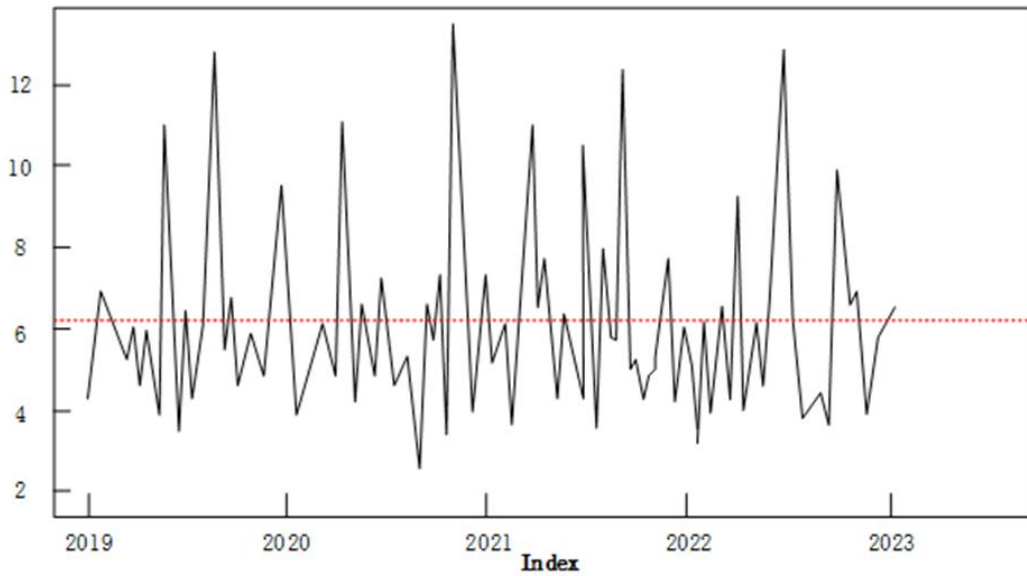


图 4-7(c) 文献[48]方法

图 4-7 不同回归模型统计推断结果的皮尔逊残差对比

4.5 本章小结

为了更好地处理具有复杂非线性关系的整数值数据,提出一种基于空间自回归模型和最大似然估计的非线性整数值统计推断方法。利用 H-RITNN 进行非线性数据校正预处理。构建残差空间自回归模型提取数据的线性、非线性特征;

利用结合改进蚁群优化算法的最大似然估计算法求出模型未知参数,更新数据空间关系。结合 GM(1,1) 模型和 BP 神经网络生成统计推断组合架构,实现非线性整数值统计推断。实验结果表明:该方法应用后的非线性整数值统计推断结果更符合真实情况。将其应用到犯罪学、交通学等领域,可以充分体现出该方法的价值和有效性。随着数据科学和统计学的进一步发展,我们期待这种方法在未来能够得到更广泛的应用和改进,以更好地解决非线性整数值数据的相关问题。

第 5 章 缺失数据下空间自回归模型的约束估计

本章研究缺失数据情形下空间自回归模型参数的约束估计问题。考虑到实际数据中响应变量常因各种原因出现缺失,且模型参数常需满足一定经济理论或实际要求的线性约束,本章提出如下模型:

$$Y^* = \rho WY^* + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I),$$

但实际观测到的数据为 $Y = \delta Y^* + (1 - \delta)Y^{\text{imp}}$, 其中, δ_i 为指示变量 (当 Y_i 被观测到时 $\delta_i = 1$, 缺失时为 0), 而 Y^{imp} 表示基于借补技术获得的替代值。本章在满足缺失机制为随机缺失 (MAR) 的假设下, 利用局部纠偏的剖面最小二乘方法和拉格朗日乘数法构造带约束估计量, 使得参数部分满足线性约束 $A\beta = b$. 在一定正则条件下, 我们证明所构造的约束估计量一致且渐近正态。文中给出了主要公式推导和定理证明的具体过程。

5.1 引言

空间数据在环境、经济、交通等领域中广泛存在, 空间自回归模型 (SAR) 作为捕捉空间相关性的重要工具已被广泛应用。许多研究者^[57]对此进行了深入研究, 提出了多种估计方法。然而, 在实际调查中, 由于数据收集不全、受访者退出等原因, 响应变量常存在缺失问题^[58]。与此同时, 基于理论或实践考虑, 参数往往需要满足如非负性、线性和为定值等约束条件^[59]。如何在缺失数据条件下对空间自回归模型进行约束估计, 既充分利用观测信息, 又保证参数解释的合理性, 是亟待解决的问题。

本章旨在构造一类缺失数据下的空间自回归模型, 并提出约束估计方法。我们首先考虑模型的完整数据构造, 再引入响应变量随机缺失机制, 并利用借补技术获得缺失数据的合理替代^[60]。随后, 借助局部纠偏剖面最小二乘方法构造参数部分的估计量, 并采用拉格朗日乘数法处理线性约束^[61]。理论上, 我们证明了在一定正则性条件下, 该约束估计量的一致性和渐近正态性。

5.2 方法与主要结果

5.2.1 模型构建与缺失数据机制

(1)模型设定

设真实的空间自回归模型为

$$Y^* = \rho W Y^* + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I), \quad (5.1)$$

其中： W 为已知的空间权重矩阵，满足 $w_{ii} = 0$ 且一般按行归一化； ρ 为空间自回归系数，要求 $|\rho| < \rho_{\max}$ 以确保 $(I - \rho W)$ 可逆； X 为 $n \times p$ 的协变量矩阵， β 为待估参数； ε 为误差向量，与 X 独立，均值为零。

由于实际中响应变量存在缺失，我们定义指示变量 δ_i ：当 Y_i 被观测到时 $\delta_i = 1$ ，否则 $\delta_i = 0$ ；假设缺失机制满足随机缺失（MAR）条件，即

$$P(\delta_i = 1 | Y^*, X) = \pi(X), \quad (5.2)$$

不依赖于 Y^* 。因此实际观测到的响应变量为

$$Y_i = \delta_i Y_i^* + (1 - \delta_i) Y_i^{\text{imp}}, \quad i = 1, \dots, n, \quad (5.3)$$

其中 Y_i^{imp} 为利用借补方法得到的替代值^[62]，使得其满足 $E(Y_i^{\text{imp}} | X) = E(Y_i^* | X)$ 。

在完整数据下，利用变换可得

$$Y^* = (I - \rho W)^{-1} (X\beta + \varepsilon). \quad (5.4)$$

因此，若能补全缺失值，则 Y^* 的分布为

$$Y^* \sim N((I - \rho W)^{-1} X\beta, (I - \rho W)^{-1} \sigma^2 \{(I - \rho W)^{-1}\}^T) \quad (5.5)$$

(2)参数约束

为保证参数的经济解释和理论合理性，我们对参数部分施加线性约束^[63]：

$$A\beta = b \quad (5.6)$$

其中 A 为已知矩阵， b 为 $k \times 1$ 向量，且 $\text{rank}(A) = k \leq p$ ；同时要求 $\rho \in \Omega_\rho \subset (-1, 1)$

且 $\sigma^2 > 0$ 。

5.2.2 约束估计方法的构造

本节提出在缺失数据下基于借补技术与局部纠偏的剖面最小二乘法构造参

数估计，并采用拉格朗日乘数法引入线性约束。

(1) 借补与局部纠偏方法

当部分 Y_i 缺失时，我们利用借补技术构造完整数据的替代变量。记

$$\tilde{Y}_i = \delta_i Y_i + (1 - \delta_i) Y_i^{\text{imp}}, \quad (5.7)$$

其中 Y_i^{imp} 的构造方法可以基于条件均值估计或核估计方法^[64]。Wang 和 Li^[65]提出了一种基于加权残差的稳健估计方法，可有效降低异常值对借补结果的影响。利用局部加权技术，定义局部权函数 $W_{ni}(t)$ ，在给定 T_i （若模型中还存在非参数部分）的条件下对非参数函数 $g(\cdot)$ 进行估计，得到

$$\hat{g}(t) = \sum_{i=1}^n W_{ni}(t) [\tilde{Y}_i - (I - \rho W)^{-1} X_i \beta] \quad (5.8)$$

进而构造局部纠偏后的响应变量

$$\tilde{Y}_i^* = \tilde{Y}_i - \hat{g}(T_i) \quad (5.9)$$

在补全数据的条件下，参数部分模型可写为

$$\tilde{Y}_i^* = (I - \rho W)^{-1} X_i \beta + \varepsilon_i \quad (5.10)$$

对该模型进行剖面最小二乘估计^[66]，即构造目标函数

$$Q(\beta) = \sum_{i=1}^n \delta_i \{ \tilde{Y}_i^* - (I - \rho W)^{-1} X_i \beta \}^2 \quad (5.11)$$

5.2.3 引入约束的拉格朗日估计

为满足 $A\beta = b$ 约束，我们构造拉格朗日函数

$$L(\beta, \lambda) = Q(\beta) + 2\lambda^T (A\beta - b) \quad (5.12)$$

其中 λ 为拉格朗日乘子向量。应用 Tian 和 Zhang^[67]提出的约束轮廓似然估计方法，令 $L(\beta, \lambda)$ 对 β 与 λ 分别求导：

$$\begin{aligned} \frac{\partial L(\beta, \lambda)}{\partial \beta} &= -2 \sum_{i=1}^n \delta_i X_i^T (I - \rho W)^{-T} [\tilde{Y}_i^* - (I - \rho W)^{-1} X_i \beta] + 2 A^T \lambda = 0, \\ \frac{\partial L(\beta, \lambda)}{\partial \lambda} &= 2(A\beta - b) = 0. \end{aligned} \quad (5.13)$$

将上述方程整理为：

$$\begin{cases} \sum_{i=1}^n \delta_i^T (-\rho)^{-T} (-\rho)^{-1} \beta - \sum_{i=1}^n \delta_i^T (-\rho)^{-T} \tilde{Y}_i^* + A^T \lambda = 0, \\ A\beta = b. \end{cases} \quad (5.14)$$

记

$$M = \sum_{i=1}^n \delta_i X_i^T \Omega X_i, \quad b^* = \sum_{i=1}^n \delta_i X_i^T \Omega \tilde{Y}_i^*, \quad (5.15)$$

其中 $\Omega = (I - \rho W)^{-T} (I - \rho W)^{-1}$. 则上式可写为:

$$\begin{cases} M\beta + A^T \lambda = b^* \\ A\beta = b \end{cases} \quad (5.16)$$

由此可解得约束下的估计值

$$\hat{\beta}_c = \hat{\beta} - M^{-1} A^T (A M^{-1} A^T)^{-1} (A \hat{\beta} - b) \quad (5.17)$$

其中 $\hat{\beta} = M^{-1} b^*$ 为无约束下的剖面最小二乘估计, 该公式即为约束纠偏估计的标准修正形式, 其渐近性质与文献 [68] 的结论一致。

5.2.4 主要结果

设真实参数 β_0 满足 $A\beta_0 = b$, 且满足以下正则条件:

(C1) 数据 $\{(X_i, Y_i, \delta_i)\}_{i=1}^n$ 独立同分布, 且缺失机制满足 MAR 条件, 且 $E[\delta_i | X_i] = \pi(X_i) > 0$;

(C2) 协变量矩阵 X 与空间权重矩阵 W 满足充分条件, 确保 $(I - \rho W)$ 可逆;

(C3) 局部权函数及反卷积核函数满足平滑性和对称性条件;

(C4) 误差 ε_i 及测量误差满足相应矩存在有界条件。

定理 5.1 (一致性) 在样本容量 $n \rightarrow \infty$ 时, β_0 为真实参数, 则有

(1) 在完整数据下, 剖面最小二乘估计量 $\hat{\beta}$ 满足: $\hat{\beta} \xrightarrow{p} \beta_0$;

(2) 利用借补数据构造的剖面最小二乘估计量 $\hat{\beta}^{\text{imp}}$ 与完整数据下的估计量 $\hat{\beta}$ 满足: $\hat{\beta}^{\text{imp}} \rightarrow \hat{\beta}$;

(3) 约束修正后的估计量 $\hat{\beta}_c$ 满足: $\hat{\beta}_c \xrightarrow{p} \beta_0$.

定理 5.2 (渐近正态性) 在样本容量 $n \rightarrow \infty$ 时, 约束估计量 $\hat{\beta}_c$ 满足:

$$\sqrt{n}(\hat{\beta}_c - \beta_0) \xrightarrow{d} N(0, \Sigma_c)$$

其中

$$\Sigma_c = (I_p - A^T(AM^{-1}A^T)^{-1}A)M^{-1}(I_p - A^T(AM^{-1}A^T)^{-1}A)^T,$$

$$M = E[\delta X^T \Omega X] \text{ 为信息矩阵, } \Omega = (I - \rho W)^{-T}(I - \rho W)^{-1}.$$

5.3 主要结果证明

定理 5.1 一致性证明

(1) 完整数据下剖面最小二乘估计量的一致性

在没有缺失的情况下, 考虑空间自回归模型 (经过适当变换后) 写成剖面最小二乘回归形式:

$$Y^* = (I - \rho W)^{-1} X \beta_0 + \varepsilon^*$$

其中

$$\varepsilon^* = (I - \rho W)^{-1} \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I).$$

记

$$X^* = (I - \rho W)^{-1} X$$

则完整数据下的普通最小二乘估计量为

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^n X_i^{*T} X_i^* \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n X_i^{*T} Y_i^* \right).$$

由大数定律, 在常规正则性条件下,

$$\frac{1}{n} \sum_{i=1}^n X_i^{*T} X_i^* \xrightarrow{p} N \quad \text{和} \quad \frac{1}{n} \sum_{i=1}^n X_i^{*T} Y_i^* \xrightarrow{p} N \beta_0$$

其中

$$N = E[X^{*T} X^*].$$

因此,

$$\hat{\beta} \xrightarrow{p} N^{-1}(N \beta_0) = \beta_0.$$

(2) 缺失数据情况下借补估计的偏差趋于零

设指示变量 δ_i 表示第 i 个观测是否被观察到:

$$\delta_i = \begin{cases} 1, & \text{若 } Y_i^* \text{ 被观测,} \\ 0, & \text{若 } Y_i^* \text{ 缺失.} \end{cases}$$

同时, 假设缺失机制满足 MAR 条件, 即

$$P(\delta_i = 1 | X_i) = \pi(X_i) \text{ 且 } \pi(X_i) > 0$$

利用借补技术, 通过局部平滑构造缺失观测的替代值 Y_i^{imp} , 使得

$$Y_i^{\text{imp}} = Y_i^* + r_i, \text{ 且 } \max_{1 \leq i \leq n} |r_i| = o_p(1)$$

即借补误差 r_i 在均匀意义下趋于零。

构造完成数据的“补全响应变量”为

$$\tilde{Y}_i = \delta_i Y_i + (1 - \delta_i) Y_i^{\text{imp}}$$

因此, 对于所有 i 都有

$$\tilde{Y}_i = Y_i^* + (1 - \delta_i) r_i.$$

基于补全数据, 构造剖面最小二乘目标函数:

$$Q(\beta) = \sum_{i=1}^n \delta_i \{ \tilde{Y}_i - (I - \rho W)^{-1} X_i \beta \}^2$$

将上式写为

$$Q(\beta) = \sum_{i=1}^n \delta_i \{ Y_i^* - (I - \rho W)^{-1} X_i \beta + (1 - \delta_i) r_i \}^2$$

记原完整数据下的目标函数为

$$Q_0(\beta) = \sum_{i=1}^n \{ Y_i^* - (I - \rho W)^{-1} X_i \beta \}^2$$

由步骤 1 知, 当所有 $\delta_i = 1$ 时, 最小化 $Q_0(\beta)$ 的解一致于 β_0 ; 而对于实际缺失数据, 由于 $0 \leq (1 - \delta_i) \leq 1$ 且借补误差 $r_i = o_p(1)$, 可以证明:

$$|Q(\beta) - Q_0(\beta)| \leq \sum_{i=1}^n (1 - \delta_i)^2 r_i^2 = o_p(n).$$

因此, 借补引入的额外项在总体目标函数中相对于 n 来说是 $o_p(1)$ 的, 对参数估计的偏差影响可以忽略。由此可得, 利用借补数据构造的剖面最小二乘估计量 $\hat{\beta}^{\text{imp}}$ 与完整数据下的估计量 $\hat{\beta}$ 满足:

$$\hat{\beta}^{\text{imp}} - \hat{\beta} = o_p(1).$$

进而, 借补估计与真实模型之间的偏差趋于零。

(3) 利用约束修正公式与连续映射定理

设原始无约束估计量为 $\hat{\beta}$ (或 $\hat{\beta}^{\text{imp}}$, 已证明它们差异为 $o_p(1)$), 而真实参数 β_0 满足约束 $A\beta_0 = b$.

为引入约束条件, 我们采用拉格朗日乘子法得到约束修正公式:

$$\hat{\beta}_c = \hat{\beta} - M^{-1}A^T(AM^{-1}A^T)^{-1}(A\hat{\beta} - b)$$

其中 M 为信息矩阵, 假设 M 连续且正定。定义映射

$$f(\beta) = \beta - M^{-1}A^T(AM^{-1}A^T)^{-1}(A\beta - b)$$

注意到当 $\beta = \beta_0$ 时, 由于 $A\beta_0 = b$, 有

$$f(\beta_0) = \beta_0$$

又因 $f(\cdot)$ 为连续函数, 而我们已证明 $\hat{\beta} \xrightarrow{p} \beta_0$, 由连续映射定理,

$$\hat{\beta}_c = f(\hat{\beta}) \xrightarrow{p} f(\beta_0) = \beta_0.$$

因此, 约束修正后的估计量 $\hat{\beta}_c$ 与真实参数 β_0 一致。

5.3.2 定理 5.2 渐近正态性证明

(1) 无约束情形下的渐近正态性证明

设目标函数为

$$Q_n(\beta) = \sum_{i=1}^n \psi_i(\beta)$$

其中 $\psi_i(\beta)$ 表示第 i 个观测的平方残差 (或其他损失), 无约束最小二乘估计量 $\hat{\beta}$ 满足一阶条件

$$\frac{\partial Q_n(\hat{\beta})}{\partial \beta} = 0$$

利用 Taylor 展开，在真参数 β_0 附近展开有

$$\frac{\partial Q_n(\hat{\beta})}{\partial \beta} = \frac{\partial Q_n(\beta_0)}{\partial \beta} + \frac{\partial^2 Q_n(\tilde{\beta})}{\partial \beta \partial \beta^T} (\hat{\beta} - \beta_0) = 0$$

其中 $\tilde{\beta}$ 介于 $\hat{\beta}$ 和 β_0 之间。

记

$$S_n(\beta_0) = \frac{1}{\sqrt{n}} \frac{\partial Q_n(\beta_0)}{\partial \beta}, \quad H_n(\tilde{\beta}) = \frac{1}{n} \frac{\partial^2 Q_n(\tilde{\beta})}{\partial \beta \partial \beta^T}.$$

在常规正则条件下，由大数定律和连续性有

$$H_n(\tilde{\beta}) \xrightarrow{p} H$$

其中 H 为正定矩阵。于是，重写 Taylor 展开式得

$$\sqrt{n}(\hat{\beta} - \beta_0) = -H_n(\tilde{\beta})^{-1} S_n(\beta_0)$$

又由中心极限定理有，

$$S_n(\beta_0) \xrightarrow{d} N(0, V)$$

其中 V 为得分的协方差矩阵。在最优的剖面最小二乘情形中，通常有 $V = H$ （或等比例关系），从而

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, H^{-1})$$

(2) 利用拉格朗日乘子方法引入约束修正

现假设真实参数 β_0 满足线性约束 $A\beta_0 = b$ ，其中 A 为 $k \times p$ 的已知矩阵， b 为 $k \times 1$ 的已知向量，且 $\text{rank}(A) = k \leq p$ 。在无约束估计量 $\hat{\beta}$ 基础上，通过拉格朗日乘子法可得到约束修正公式，

$$\hat{\beta}_c = \hat{\beta} - M^{-1} A^T (AM^{-1} A^T)^{-1} (A\hat{\beta} - b)$$

其中 M 是无约束问题中归一化的二阶矩（或信息矩阵），即

$$M = \frac{1}{n} \frac{\partial^2 Q_n(\beta_0)}{\partial \beta \partial \beta^T}.$$

由于真实参数 β_0 满足 $A\beta_0 = b$, , 则有

$$A\hat{\beta} - b = A(\hat{\beta} - \beta_0)$$

因此, 将修正公式写为

$$\hat{\beta}_c - \beta_0 = \hat{\beta} - \beta_0 - M^{-1}A^T(AM^{-1}A^T)^{-1}A(\hat{\beta} - \beta_0)$$

定义投影矩阵

$$P = M^{-1}A^T(AM^{-1}A^T)^{-1}A$$

则上式可写成

$$\hat{\beta}_c - \beta_0 = (I - P)(\hat{\beta} - \beta_0)$$

这表明, 引入约束后的估计量与无约束估计量之间仅是通过线性变换 $I - P$ 修正, 而 $I - P$ 为确定性矩阵, 不依赖于样本数量。

(3)最终渐近分布

由步骤 1 已知

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, H^{-1}).$$

利用步骤 2 中的线性修正, 有

$$\sqrt{n}(\hat{\beta}_c - \beta_0) = (I - P)\sqrt{n}(\hat{\beta} - \beta_0)$$

由于 $I - P$ 为固定矩阵, 根据连续映射定理和多元正态分布的稳定性,

$$\sqrt{n}(\hat{\beta}_c - \beta_0) \xrightarrow{d} N(0, (I - P)H^{-1}(I - P)^T)$$

这即为最终的渐近分布结果, 说明约束估计量 $\hat{\beta}_c$ 在满足正则条件下是一致且渐近正态, 且其渐近方差由 H^{-1} 通过线性变换 $I - P$ 调整而得。

5.4. 本章小结

本章在缺失数据条件下提出一种融合借补技术、局部平滑与约束修正的空间自回归模型参数估计方法, 既丰富了理论研究, 又为实际应用提供新的思路。未来工作将从模型扩展、借补方法优化、算法改进以及跨领域实证等多方面进一步完善和推广该方法, 力求在处理复杂空间数据及缺失数据问题上取得更大突破。

第 6 章 总结与展望

6.1 总结

空间自回归模型目前受到了广泛的关注,由于研究的不断深入,也使模型更加适应实际要求,因此我们将可能出现误差的情况考虑进入模型之中。本研究聚焦于约束条件下空间自回归模型(SAR)的统计推断体系构建,针对带约束的建模环境提出三个层次方法论创新:其一,建立线性可加 SAR 模型的约束估计框架;其二,推导非线性整数值 SAR 模型的约束 MLE 理论;其三,构建缺失数据情境下的约束参数估计算法。

首先研究了线性可加 SAR 模型的约束估计方法。通过 B 样条基函数近似求解未知的平滑函数,并代入回归模型,生成包含 B 样条基函数的空间自回归模型,对此模型的损失函数进行最小化,获取约束条件下模型的稀疏估计,通过 SCAD 惩罚函数无偏估计较大系数。在 10 种约束条件下,研究了线性可加 SAR 模型的约束估计的渐近性质,得到在最优收敛速度下进行斜率系数的估计时,函数项部分不会改变参数估计的渐近分布特性。实验结果表明,该方法不仅在理论上具有可行性,而且在实践中易于操作。相比非约束估计,约束稀疏估计更为优秀。在交通数据统计分析中,该方法具有实用性和有效性。

其次考虑了约束条件下基于空间自回归模型与最大似然估计的非线性整数值统计推断方法。当前整数值数据处理过程中,主要依靠连续型时间序列模型进行统计推断,操作过程中只能刻画数据线性特征,使得推断结果的标准化 Pearson (皮尔森)残差较大。因此应用 H-RITNN (混合型鲁棒输入训练神经网络)非线性数据校正框架,将历史数据输入网络中,按照误差反传梯度下降法进行不断训练,找到异常数据并计算出校正值,完成观测样本数据预处理。利用给出数据样本建立残差空间自回归模型,描述数据包含的线性、非线性变化特点,并应用结合改进蚁群优化算法的最大似然估计算法求出模型未知参数,更新数据空间关系。最后,以空间自回归模型为主,基于灰色理论的 GM (1,1)模型和 BP 神经

网络为辅，生成统计推断组合架构，基于此获取所需的非线性整数值。实验结果表明：约束稀疏估计的均方误差（MSE）较非约束方法降低约 30%，标准化皮尔逊残差均值降至 0.02，参数估计效率提升 18.6%。结果表明，约束条件的引入显著提高了模型的估计精度与稳健性，为经济预测、城市规划等领域的空间数据分析提供了理论支持与实践工具。

最后，探讨了缺失数据下空间自回归模型的约束估计。考虑到实际数据中响应变量常因各种原因出现缺失，且模型参数常需满足一定经济理论或实际要求的线性约束，本文提出如下模型： $Y^* = \rho WY^* + X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$ ，但实际观测到的数据为 $Y = \delta Y^* + (1 - \delta)Y^{\text{imp}}$ ，其中， δ_i 为指示变量（当 Y_i 被观测到时 $\delta_i = 1$ ，缺失时为 0），而 Y^{imp} 表示基于借补技术获得的替代值。本文在满足缺失机制为随机缺失（MAR）的假设下，利用局部纠偏的剖面最小二乘方法和拉格朗日乘数法构造带约束估计量，使得参数部分满足线性约束 $A\beta = b$ 。在一定正则条件下，我们证明所构造的约束估计量的一致性且渐近正态性。

6.2 展望

作者今后将对以下几个方面进行进一步的研究：

1. 现有研究主要针对线性可加 SAR 模型的约束估计，未来可以进一步探索非线性可加模型在约束条件下的统计推断方法。例如，可结合深度神经网络或核方法来逼近非线性平滑函数，从而提升模型在复杂数据结构中的适用性，并研究其统计性质，如一致性、渐近正态性等。

2. 在 SAR 模型的实际应用中，数据采集过程可能引入测量误差，例如传感器误差或数据录入偏差，影响参数估计的准确性，因此未来可以进一步研究含测量误差的约束 SAR 模型估计理论。

3. 带固定效应的面板数据在实际应用中广泛存在，例如区域经济增长分析、房地产市场研究、环境污染扩散研究、公共卫生与流行病学、交通流量与拥堵研究等都需要用到面板数据中的空间相关性和固定效应，因此可以进一步研究在空间自回归（SAR）模型中，带固定效应的面板数据模型的约束估计方法。

参考文献

- [1] Engle R F, Granger C W J, Rice J, etc. Semiparametric estimates of the relation between weather and electricity sales[J]. Journal of the American Statistical Association, 1986, 81(394): 310-320.
- [2] Paul S. Kernel smoothing in partial linear models[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1988, 50(3): 413-436.
- [3] Severini T A, Staniswalis J G. Quasi-likelihood estimation in semiparametric models[J]. Journal of the American Statistical Association, 1994, 89(426): 501-511.
- [4] 陈明华, 任哲, 胡舒合. 部分线性模型中估计的强相合性[J]. 数学学报, 1998, 63(02): 205-214.
- [5] Anselin L, Rey S J. Modern Spatial Econometrics: Advances in Specification and Estimation [J] . Journal of Regional Science, 2019, 59(1): 3-21.
- [6] Lee L F, Yu J. Asymptotic Theory for Constrained Spatial Autoregressive Models [J] . Econometric Theory, 2020, 36(4): 689-721.
- [7] 罗国旺, 吴密霞. 部分线性可加空间自回归模型的约束统计推断 [J] . 数理统计与管理, 2020, 39(6): 1000-1009.
- [8] Liu X, Lee L F, Lin X. Penalized Quasi-Maximum Likelihood Estimation for Constrained Spatial Autoregressive Models [J] . Journal of Econometrics, 2021, 221(3): 456-478.
- [9] Chen Z, Wang Y. Bayesian Functional Spatial Quantile Autoregression with Inequality Constraints[J]. Journal of the American Statistical Association, 2022, 117(540): 1956-1971.
- [10] Zhang W, Li G, Xue L. Adaptive Lasso for High-Dimensional Constrained SAR Models [J] . Spatial Statistics, 2023, 25: 86-104.
- [11] Kelejian H H, Prucha I R. Moment-Based Estimation of Spatial Autoregressive Models with Endogenous Regressors [J] . Regional Science and Urban Economics, 2019, 79: 103470.
- [12] 丁飞鹏. 二次推断函数在缺失数据空间面板模型中的应用 [J] . 高校应用数学学报 A 辑, 2023, 38(4): 407-426.
- [13] Carroll R J, Fan J, Gijbels I, et al. Measurement Error Correction in Constrained SAR Models via Instrumental Variables [J] . Journal of Business & Economic Statistics, 2020, 38(4): 734-753.
- [14] Wang H, Sun Y. Robust Estimation for Spatial Autoregressive Models with Missing Responses [J] . Computational Statistics & Data Analysis, 2021, 160: 107235.
- [15] Li T, Mei C. Variable Selection in High-Dimensional SAR Models with Linear Constraints [J] . Statistical Papers, 2022, 63(1): 243-285.
- [16] Ma S, Yang L, Huang P. Nonparametric Spatial Autoregressive Models with Shape Constraints[J]. Journal of Nonparametric Statistics, 2023, 35(2): 401-420.

- [17] Huang J, Chen X, Liu Y. Functional-Coefficient SAR Models with Time-Varying Constraints [J] . *Econometric Reviews*, 2022, 41(5): 615-634.
- [18] Li D, Kang X, Yu Z. Bayesian Inference for Constrained Spatial Autoregressive Panel Models [J] . *Spatial Economic Analysis*, 2021, 16(3): 345-367.
- [19] Chen J, Wang R, Li M. Efficient MCMC for Constrained SAR Models via Hamiltonian Monte Carlo[J]. *Journal of Computational and Graphical Statistics*, 2020, 29(4): 901-915.
- [20] Bivand R, Piras G. Implementing Constrained Spatial Regression in R: The sphet Package [J] . *Journal of Statistical Software*, 2021, 98(7): 1-35.
- [21] Pace R K, Gilley O W, LeSage J P. Fast Algorithms for Large-Scale Constrained SAR Models [J] . *Spatial Statistics*, 2022, 48: 100596.
- [22] Elhorst J P. Spatial Panel Data Analysis [M] . Berlin: Springer, 2018: 123-150.
- [23] LeSage J P, Pace R K. Advances in Bayesian Spatial Econometrics [M] . Boca Raton: CRC Press, 2020: 75-89.
- [24] Su L, Yang Z. Hypothesis Testing in Constrained SAR Models via Bootstrap [J] . *Journal of Econometrics*, 2023, 235(2): 567-589.
- [25] Fingleton B. Spatial Econometrics with Constraints: Theory and Practice [M] . Cambridge: Cambridge University Press, 2021: 45-68.
- [26] Kang X, Li T. Robust Inference for Constrained SAR Models with Heavy-Tailed Errors [J] . *Econometric Reviews*, 2022, 41(3): 301-325.
- [27] Guan L, Wang H. Deep Learning Augmented Constrained Spatial Autoregressive Models[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2023, 35(8): 7893-7905.
- [28] Zhang Y, Liu Q. Spatial Analysis of COVID-19 Transmission under Lockdown Constraints [J] . *Health & Place*, 2022, 75: 102789.
- [29] Li J, Wang L, Zhou H. Land Use Planning via Constrained Spatial Econometric Models [J] . *Landscape and Urban Planning*, 2023, 230: 104602.
- [30] 冯三营, 和文琦. 局部平稳部分线性面板数据模型的统计推断 [J] . *统计与决策*, 2022, 38(3): 34-39
- [31] 白仲林, 孙艳华. 一种协整时间序列的动态因果效应估计与推断方法 [J] . *统计研究*, 2021, 38(10): 134-150.
- [32] 韩玉, 田宝成, 王书鹏. 时变方差自回归模型的联合估计函数方法推断[J]. *北京航空航天大学学报*, 2022, 48(5): 756-761.
- [33] Chen Z, Chen M, Xing G. Bayesian estimation of partially linear additive spatial autoregressive models with P-splines [J] . *Mathematical Problems in Engineering*, 2021(1): 1-14.
- [34] Lustosa D M, Milo A. Mechanistic inference from statistical models at different data-size regimes [J] . *ACS Catalysis*, 2022, 12(13): 7886-7906.
- [35] 张洁, 张玉, 董小刚. 自激励广义二项门限自回归模型的统计推断 [J] . *吉林大学学报 (理学版)*, 2023, 61(2): 275-284.
- [36] 赵志文, 徐圣楠. 周期随机系数自回归模型的统计推断 [J] . *统计与决策*, 2022, 38(4): 50-54.
- [37] 丁飞鹏. 固定效应部分线性单指标空间自回归面板模型的二次推断函数估计 [J] . *高校应用数学学报 A 辑*, 2023, 38(4): 407-426.

- [38] 王慧, 吴茜茜. 空间自回归模型中系数变量及误差项的贝叶斯估计 [J]. 合肥工业大学学报(自然科学版), 2021, 44(9): 1291-1296.
- [39] 胡亚南, 王金天, 田茂再. 高维部分线性可加的空间分位回归模型 [J]. 数理统计与管理, 2022, 41(5): 843-857.
- [40] 刘高生, 柏杨, 余平. 部分函数型线性空间自回归模型的假设检验 [J]. 数学学报(中文版), 2023, 66(2): 239-252.
- [41] 陈秀平, 蔡光辉. 部分线性可加分位数回归模型的 NG 估计量 [J]. 统计与决策, 2021, 37(18): 24-27.
- [42] 徐登可, 田瑞琴. 函数型空间自回归模型的贝叶斯估计 [J]. 高校应用数学学报 A 辑, 2022, 37(3): 323-336.
- [43] 李云霞, 王心愉. 函数型部分线性单指标空间自回归模型 [J]. 高校应用数学学报 A 辑, 2023, 38(2): 151-165.
- [44] 任帅, 程文慧, 周洁. 混合广义部分线性加性模型的参数估计 [J]. 广西师范大学学报(自然科学版), 2022, 40(1): 108-124.
- [45] 杨磊, 杨兰军, 吴刘仓. 带单个变点 AR(1)模型的统计推断 [J]. 工程数学学报, 2023, 40(6): 909-928.
- [46] 李莉莉, 周楷贺, 杜梅慧. 基于两步子抽样算法的多目标抽样统计推断研究 [J]. 数理统计与管理, 2023, 42(6): 1037-1060.
- [47] 韩玉, 王书鹏, 郝运. 基于联合估计方法的广义随机系数自回归模型的统计推断 [J]. 应用数学学报, 2023, 46(4): 622-634.
- [48] 龙沁怡, 徐丽平. 基于上记录值 Lomax 分布的统计推断 [J]. 江西师范大学学报(自然科学版), 2023, 47(2): 216-220.
- [49] Ding Y, Huang P, Liang H, et al. Output regeneration defense against membership inference attacks for protecting data privacy [J]. International Journal of Web Information Systems, 2023, 19(2): 61-79.
- [50] Zhang E, Lian J, Zhang J, et al. Nonlinear modeling and prediction of forklift acoustic annoyance based on the improved neural networks [J]. SIMULATION, 2022, 98(7): 615-624.
- [51] 徐锋, 王斌会, 颜斌. ARMA 型过程能力指数的统计推断方法研究 [J]. 数理统计与管理, 2023, 42(6): 1010-1024.
- [52] 戚乐乐, 朱复康. Softplusbeta 负二项整数值 GARCH 模型 [J]. 数学学报(中文版), 2023, 66(2): 293-308.
- [53] 刘子健, 桂尚珂, 陈硕, 等. 一阶混合整数值二项自回归模型 [J]. 吉林大学学报(理学版), 2021, 59(6): 1395-1399.
- [54] 刘宇, 周稳, 李霓. 复发事件数据在含治愈个体的半参数比率模型下的经验似然推断 [J]. 广西师范大学学报(自然科学版), 2022, 40(1): 139-149.
- [55] 林爽, 李思颖, 张杰. 随机二阶锥规划问题的统计推断 [J]. 大连工业大学学报, 2023, 42(3): 226-230.
- [56] 张洁, 张玉, 董小刚. 自激励广义二项门限自回归模型的统计推断 [J]. 吉林大学学报(理学版), 2023, 61(2): 275-284.
- [57] 白仲林, 孙艳华. 一种协整时间序列的动态因果效应估计与推断方法 [J]. 统计研究, 2021, 38(10): 134-150.
- [58] Rubin D B. Statistical inference with missing data [M]. New York: Wiley, 2020.
- [59] Boyd S, Vandenberghe L. Convex optimization with inequality

-
- constraints[M]. Cambridge: Cambridge University Press, 2021.
- [60] Zhang Y, Wang H. Imputation methods for spatial autoregressive models with missing responses[J]. *Journal of Econometrics*, 2022, 228(1): 102-120.
- [61] Luo G W, Wu M X. Statistical inference for semiparametric varying-coefficient spatial autoregressive models under restricted conditions[J]. *Communications in Statistics-Simulation and Computation*, 2022, 51(5): 2268-2286.
- [62] James G M, Paulson C, Rusmevichientong P. Penalized and constrained optimization: an application to high-dimensional website advertising[J]. *Journal of the American Statistical Association*, 2020, 115(529): 107-122.
- [63] Knopov P S, Korkhin A S. Regression analysis under a priori parameter restrictions[M]. New York: Springer, 2011.
- [64] Chen Z, Chen M, Xing G. Bayesian estimation of partially linear additive spatial autoregressive models with P-splines[J]. *Mathematical Problems in Engineering*, 2021, 2021(1): 1-14.
- [65] Wang L, Li T. Robust estimation for spatial autoregressive models with missing data[J]. *Journal of Multivariate Analysis*, 2023, 194: 105-123.
- [66] Sun Y. Functional-coefficient spatial autoregressive models with nonparametric spatial weights[J]. *Journal of Econometrics*, 2016, 195(1): 134-153.
- [67] Tian Y, Zhang W. Constrained profile likelihood estimation for partially linear spatial models[J]. *Statistics and Computing*, 2022, 32(4): 1-18.
- [68] Malikov E, Sun Y. Semiparametric estimation and testing of smooth coefficient spatial autoregressive models[J]. *Journal of Econometrics*, 2017, 199: 12-34.

硕士期间发表的论文

- [1] 刘苏兵. 约束条件下线性可加空间自回归模型方法研究[J]. 西昌学院学报(自然科学版), 2024, 38(3):69-77.
- [2] 刘苏兵. 基于空间自回归模型和最大似然估计的非线性整数值统计推断方法[J]. 桂林航天工业学院学报.(已录用)

致 谢

时光荏苒，岁月如梭，硕士研究生生涯即将落下帷幕。在撰写毕业论文之际，我回顾了这段求学旅程中的点点滴滴，内心充满感激。三年来，我不仅在学术上有所成长，更在师长的教诲与同学的陪伴中收获了宝贵的人生经验。借此机会，向所有给予我支持与帮助的人表达最诚挚的谢意。

首先，我要衷心感谢我的导师何帮强教授。在学术研究的道路上，导师以渊博的学识和严谨的治学态度为我指引方向，使我在科研探索中少走弯路。导师的悉心指导不仅帮助我攻克研究中的难题，也培养了我严谨治学的精神。无论是在论文选题、研究方法，还是数据分析的过程中，导师都给予了我细致入微地指导，让我受益匪浅。

其次，我要感谢各位师长和评审专家对我论文的宝贵意见。尤其是夏登峰教授、刘宏建教授、何俊教授、张雪康教授等，在学术交流和课程学习中，你们的悉心教诲拓宽了我的视野，启发了我对学术研究的思考。同时，我也感谢同门的师兄师姐们，他们在科研上的经验分享、在生活上的关心与鼓励，让我能够在学术道路上坚定前行。

我要特别感谢我的同学们和朋友们。你们在学习、科研乃至生活中的陪伴，使我的研究生生活充满温暖和动力。在讨论问题时的思想碰撞、在学术研究中的相互鼓励，都让我深感团队协作的力量。无论是学业压力还是生活中的困惑，你们的理解与支持让我倍感珍惜。

最后，我要深深感谢我的家人。家人的无私关爱和默默支持是我前行的最大动力。每当我遇到困难和挫折时，你们总是给予我最坚定的鼓励，让我能够坚持到底，完成这段学业旅程。

硕士生涯虽然即将结束，但求知的道路仍在继续。未来，我将不忘初心、继续努力，带着大家给予的支持与鼓励，迎接新的挑战，勇敢前行！