参赛队号：1947

2021年（第七届）全国大学生统计建模大赛

参赛学校：	华北理工大学
论文题目：	中国省级行政单位新冠疫情数据的预测研究：基于函数型数据分析方法
参赛队员：	耿雪倩 常畅 薛晓玮
指导老师：	常文千 李雪

目　　录

图表目录

摘 要

新冠肺炎疫情作为一次突发性、大规模的公共卫生事件，不但威胁着全人类的健康和生命安全，也给世界各国的疫情防控工作带来了巨大的挑战。本文提取了基于中国大陆 30 个省级行政单位（去除湖北省），从 2020 年 1 月 28 日到 2021 年 1 月 27 日的整年度新冠肺炎疫情相关的均匀数据，运用函数型数据分析的方法，探索了数据和参数估计的平滑方法、一维和多维函数型相关性、函数型主成分分析和主成分预测、函数型数据对函数型数据的回归模型等。函数型数据分析提取了新冠肺炎病毒在湖北省集中爆发后，中国大陆其他 30 个省级行政单位新增确诊与新增死亡病例随时间变化的规律与趋势信息，从数据的统计测度中挖掘出了新冠肺炎疫情重点防控阶段数据的变化特征、相关性及其影响因素以及各省级地区新冠疫情的阶段性变化，并根据模型对未来疫情的走势情况进行预测。发现：新冠疫情在某地爆发后，周边地区的新增确诊人数和新增死亡人数呈现出规律的周期性性质，新冠疫情在各省表现出了在集中时间爆发且不同省级行政单位疫情发展具有不同模式。根据新冠疫情的上述特点提出建议，利用新冠疫情爆发的周期性制定防疫措施，根据周期性特点构建完备的疫情防控屏障，完善群防群控的工作机制，根据各个周边省份疫情的流行模式，制订相适应的防控策略。

关键词：新冠疫情；省级；函数型数据；防控

Abstract

As a sudden and large-scale public health event, the new crown pneumonia epidemic not only threatens the health and life safety of all mankind, but also brings great challenges to how countries around the world respond to the epidemic. This article extracts the evenly recorded data related to the new crown pneumonia epidemic for the entire year from January 28, 2020 to January 27, 2021 based on 30 provincial administrative units in Mainland China (excluding Hubei Province), using the method of functional data analysis, Explored the smoothing method of data and parameters, one-dimensional and multi-dimensional functional correlation, functional principal component analysis and principal component prediction, regression model of functional data to functional data, etc. Functional data analysis extracts the regularity and trend information of new confirmed and new death cases over time in 30 other provincial administrative units in mainland China after the concentrated outbreak of the new crown pneumonia virus in Hubei Province, and digs out from the statistical measurement of the data The change characteristics, correlation and influencing factors of the data in the key prevention and control phases of the new crown pneumonia epidemic, as well as the phase changes of the new crown epidemic in various provincial-level regions, are analyzed, and the future epidemic situation is predicted based on the model. It was found that after the outbreak of the new crown epidemic in a certain place, the number of new confirmed cases and the number of new deaths in the surrounding areas showed a regular cyclical nature. The new crown epidemic showed that the outbreak occurred at a concentrated time in various provinces and the development of the epidemic situation in different provincial administrative units. Make recommendations based on the above characteristics of the new crown epidemic, use the periodicity of the new crown epidemic to formulate epidemic prevention measures, build a complete epidemic prevention

and control barrier based on the periodic characteristics, and improve the working mechanism of group prevention and control. According to the epidemic pattern of each surrounding province, Formulate appropriate prevention and control strategies.

Keywords: COVID-19; provincial level; functional data; prevention and control

第1章 引言

1.1 课题研究背景

新型冠状病毒肺炎疫情（Corona Virus Disease 2019，COVID-19）是一起涉及全球范围的重大突发公共卫生事件，世界卫生组织将其命名为“2019 冠状病毒病”。2020 年 1 月，中国遭受了新冠肺炎的大规模袭击，此次新冠肺炎疫情是新中国成立以来在我国发生的传播速度最快、感染范围最广、防控难度最大的重大突发公共卫生事件。2020 年 2 月疫情在全球范围爆发，成为全球的重大公共卫生事件。疫情一直持续至今，对人民的生命安全及经济发展等方面带来了巨大损失。经过全国人民群众的艰苦努力，新冠疫情在我国得到有效控制，防控形势总体持续向好，在疫情防控常态化中生产生活秩序逐步恢复，社会经济得到平稳有效发展。当前，全国疫情防控阻击战已经取得了重大的战略成果，但我们仍须时刻保持高度警惕，持续抓好“外防输入、内防反弹”工作。与此同时，国际疫情防控形势依然严峻，全球范围内疫情仍在快速蔓延，多个国家出现爆发性增长，疫情防控呈现长期性、复杂性和不确定性。

当前我国疫情形势控制成果显著，但依然需要面对诸多困难。随着疫情的缓解，国内一些监管机构放松了警惕，人员流动逐渐增加，因节日、假期产生的人员聚集行为越来越多，这带来疫情再次反弹的风险。其次，一些国外政府疫情防控的重视与警惕程度不足，国际疫情快速蔓延带来的输入性风险增加，对境外疫情输入、偷渡人员入境等输入性风险的防控策略和政策举措仍需完善。另外，源自进口冷链食品的新冠病毒感染风险依然存在，对口岸环节开展的预防消毒工作和进口冷链食品新冠病毒的风险监测更需要常态化坚持。新冠疫情作为全球性突发的公共卫生危机，不坚持实施有效防控措施，疫情的恶化势必影响到每个国家，导致世界各国都再度陷入恐慌。

1.2 国内外的研究现状

1.2.1 关于新型冠状病毒的认识与致病机制

在新冠肺炎病毒的流行病学特征方面，2020 年 1 月，Chan et al.(2020)[1]第一次确定了新冠肺炎病毒（COVID-19）有“人传人”和“无症状感染者”的现象，并且列举了新冠肺炎病毒在同一个家庭中六名成员间互相传播的情况，同时对“无症状感染者”这一人群提出了需要特别重视的提醒，这为后来的隔离措施、全民核酸检测等防控政策提供了最原始的理论支撑。同月，中国疾控中心、山东第一医科大学的科研人员公布对新冠病毒未来进化、适应和传播开展的研究结果，Lu et al.(2020)[2]发现新冠肺炎病毒的最原始宿主大概率是蝙蝠，但是人类感染病毒是因为其他中间宿主还是直接源于蝙蝠尚不能确定。2020 年 3 月，Li et al.(2020)[3]对新冠肺炎病毒感染的流行性、传染性进行探索，发现有近 90%的确诊患者没有相关记录，表明了其对新冠肺炎病毒大范围迅速传播的忧虑，是探索新冠肺炎疫情未来可能走向的重要思考。2020 年 2 月，Daniel et al.(2020)[4]合作探索了 SARS-CoV-2 病毒传染性、流行性强的一些主要因素。Hoffmann et al.(2020)[5]的一些研究证明，防治 SARS-CoV-2 病毒感染的途径可利用来自 SARS-S 的抗体或疫苗，其研究结论对于 SARS-CoV-2 的传播性和致病机制的认识有关键性，同时对治疗性干预的靶标问题也进行了说明。上述研究为治疗药物及疫苗研发工作的开展提供了思路。

1.2.2 新冠肺炎的病理表现和相关的防控策略

2020 年 1 月，黄朝林等(2020)[6] 根据多例新冠肺炎病毒确诊者的临床观察数据，发现新冠肺炎病毒与 2002 年爆发的 SARS（严重急性呼吸系统综合征）有累似的临床表现。由于现行的核酸检测的精准度不能保证所有的新冠肺炎病毒携带者均能被检测出来，2020 年 2 月 26 日 Ai et al.(2020)[7]的研究探索了胸部 CT

扫描对发现新冠肺炎病例的巨大作用，通过对咽拭子和 CT 扫描的诊断对比，表明在胸部 CT 相较于 RT-PCR 方法有更高的敏感性。因此，影像学检查可以作为降低 FNR(False Negative Rate)的一种重要方法。2020 年 2 月，Jin et al.(2020)[8]总结了一线临床诊治经验，发布了全球首部结合了一线诊治经验的循证指南，其内容涵盖疾病流行病学、病因学、诊断学、中医及西医治疗、护理学、医院感染控制等方面，为广大医务人员和人民群众提供指导与参考。

1.2.3 新冠肺炎疫情的大流行给世界经济带来的巨大影响

据世界贸易组织（WTO）估算，2020年全年，全球实际GDP下降大约在4.8％到11.1％，全球整体贸易总额大约为2019年的68％～87％，其中农业贸易额大约为2019年的为87.3％～93.5％，降幅较大。目前，现有文献主要从疫情对产业链的负面效应和对进出口限制的维度出发分析新冠肺炎疫情的流行给世界经济带来的巨大影响。关于疫情对产业链的负面影响，Wang et al.(2020)[9]和Mahagan et al.(2021)[10]从供给方面、Lin et al.(2020)[11]和Chang et al.(2021)[12]从需求方面开展研究，阐述了新冠疫情期间各国所采取的管控措施对产业链产生的负面影响，以及产业链阶段性断裂的原因和结果。关于进出口限制对产业经济的影响，现有研究主要从限制措施的合规性、贸易产品和粮食安全等方面着手，边永民(2020)[13]和李先德等(2020)[14]认为各国因为新冠疫情防治而制定的贸易限制政策及其合规性值得商榷，在阻碍经济贸易的同时，全球范围的产业经济安全也受到极大威胁。

1.3 数据驱动的统计测度研究对国家防控疫情的重要作用

有关新冠疫情数据的统计测度研究有利于准确把握国内外疫情防控的阶段性变化，因时、因势调整工作着力点和应对举措，结合最新的国际、国内疫情蔓延与遏止形势，及时颁布行之有效的政策举措，同时兼顾疫情的防控、人民的生

产生活和国家的经济进步。但是直至目前为止，国内外学者的研究大多集中在新型冠状病毒的认识与致病机制、新冠肺炎的病理表现和相关的防控策略以及新冠肺炎疫情的流行给世界经济带来的影响，缺少对新冠病毒在集中地点（如湖北省）爆发后，同一国家其他同级单位（如中国的其他省级行政单位）针对新冠疫情统计数据特征的分析探讨和以省级单位视角的防控措施的全面研究。

本文通过函数型数据分析，利用统计学理论支撑国家对具体省级行政单位疫情防控的指导策略，通过数据趋势对防控策略进行优化，落实防控举措，增强针对性和有效性，并对未来疫情防控形势进行预测，具有极强的现实意义和必要性。数据来自国泰安 CSMAR 数据库专题研究系列中新冠疫情与经济研究专区，调取了我国省级单位的统计数据（2020-1-28 到 2021-1-27），主要包括总量数据指标和新增数据指标两类。总量指标有累计确诊病例、累计死亡病例、累计治愈病例、现有疑似病例、现有危重症病例等，这些数据的收集可以反映同一地区随着时间变化累计病例的总量变化趋势，从而反映新冠疫情发展总体情况以及严重程度。新增数据指标有新增确诊病例、新增死亡病例、新增治愈病例等，反映了我国各省级单位新冠疫情各种案例的单纯增长情况，体现疫情的扩散速度。

挖掘这些指标所萃取的信息有利于判断我国省级单位疫情的变化趋势，建立预警机制，帮助全民了解疫情动态情况。根据确诊病例和新增死亡病例等的数据变化，分析监测各个省级行政单位在特定时间内的疫情扩散方向、速度与程度，制订与完善抗击疫情政策，同时验证已实施的政策与方法的有效性。积极推进防疫工作，及时发现并制止局部潜在的疫情传播风险，对于国家疫情防控层面具有重要意义。

第2章 数据与统计方法的介绍

2.1 数据来源和分析方法

本文的分析主要基于中国大陆 30 个省级行政单位（去除湖北省），从 2020 年 1 月 28 日到 2021 年 1 月 27 日的整年度新冠疫情相关均匀数据。通过对下图 2.1 中，新增确诊与新增死亡病例人数随时间变化的规律与趋势进行统计分析，以期从数据的统计测度中挖掘出新冠疫情在某地（湖北省）集中爆发后，对国家和其他省级行政单位疫情防控工作有益的信息。经过对比，各省份之间存在一些差异，与境外输入等因素有关。

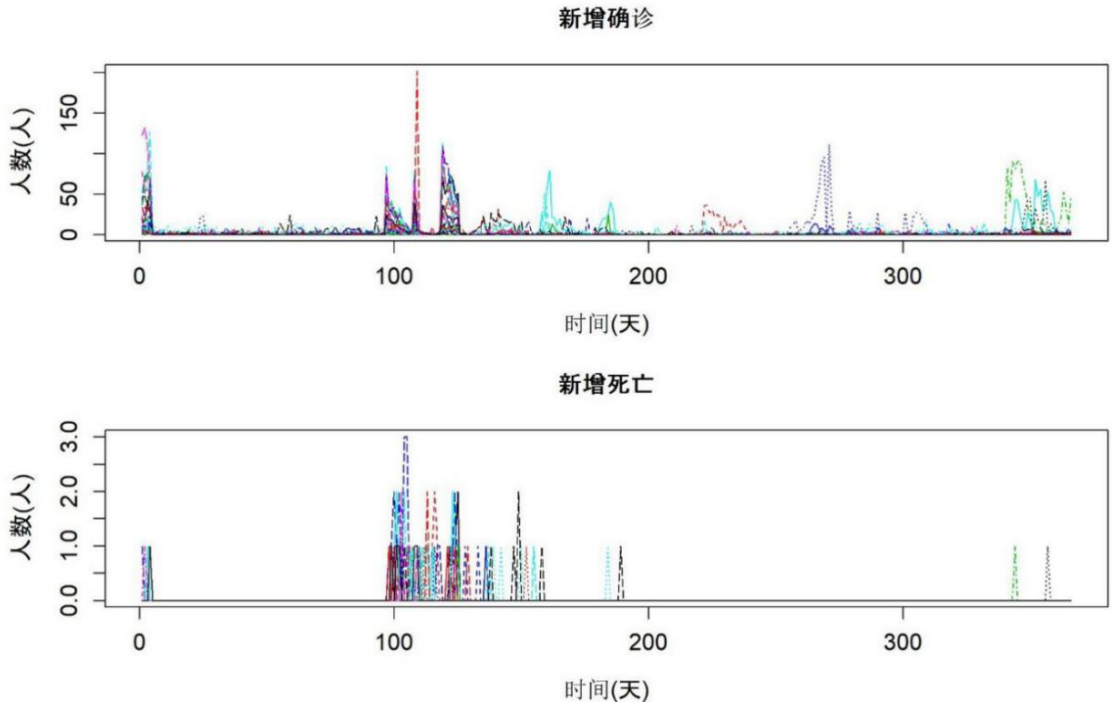

图 2.1 30 个省级单位新增确诊与新增死亡病例人数

采用的统计学方法是近年来越来越受到学界和业界重视的函数型数据分析，重点参考了王国长(2012)[15]基于非参数技术的典型相关分析、函数型数据降维方法和闫星宇(2020)[16]函数型线性回归的研究，主要工作包括对函数型指标在限制和非限制条件下的傅立叶基拟合和粗糙惩罚、一元函数型指标的相关性分析、一元函数型主成分分析及其粗糙惩罚方法、基于函数型主成分的预测方法、多元函

数型主成分分析及其粗糙惩罚方法、多元函数型指标的相关性分析以及函数型指标对函数型指标的回归分析。所有的数据处理、分析建模和可视化设置等均基于R语言（R studio，相关程序见附录），调用的程序包括："tidyr"，"fda"和"fields"等。

2.2 新冠疫情初期防治工作出现的问题

众所周知，2020年春节较往年更早，春运也随之提前，加上国内疫情开始的武汉是重要的交通枢纽，大量的人口由此流入、流出，直接导致了病毒在全国范围内传播、扩散。经过大约三个月的酝酿，从2020年4月底开始，除湖北省外的各省份新增确诊以及新增死亡人数激增。各个省份主要表现出的突出问题包括：

（1） 此次疫情有突发性和复杂性等特点，各省的医疗机构尚未进入战时状态，必要医疗资源均出现严重短缺，专业医护人员缺口较大，不能在短期内处理、安置集中大量出现的确诊患者、疑似患者和密切接触者；

（2） 一些省份在疫情初期重视程度不够，不能及早决策进行辖区内全面的封锁和核酸检测。同时，疫情爆发初期地方的核酸检测能力有限，导致新冠病毒的携带者以及阳性动物等在潜伏期内（根据目前的研究，最长的潜伏期为14天）不能被及时发现并进行医学观察，造成了大面积的传播；

（3） 初期对新冠肺炎的病因尚不了解，没有针对疫情的专用医疗方案，预防控制措施（例如密切接触者的判定标准和管理要求）尚不完善。

第3章 均匀一元函数型数据新增确诊的分析

3.1 非限制和正数限制条件下的傅里叶基拟合

3.1.1 非限制条件下的傅里叶基拟合

对于新冠疫情数据这一系列明显具有周期性的观测值，可以采用傅里叶基拟合的方式，为接下来的分析提供函数型数据。其中傅里叶基为增加频率不断增加的sine和cosine函数，其形式如下：

$$1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t) \ldots, \sin(m\omega t), \cos(m\omega t) \ldots$$

其中常数ω决定了第一组（$\sin(\omega t), \cos(\omega t)$)波动的周期，其他组的波动周期由$\frac{2\pi}{mw}$计算。$f(t)$为拟合后的函数型数据，$\phi_{(t)}$为傅里叶基向量，C为傅里叶基向量对应的参数。其中傅里叶基为增加频率不断增加的sine和cosine函数，其形式和参数估计的方法和最小二乘法相同，如公式（1）所示：

$$f(t) = \phi(t)^T C$$
$$\mathrm{SSE_C(C)} = \sum_{i=1}^{n} (y_i - f(t_i))^2 \quad (1)$$

3.1.2 正数限制条件下的傅里叶基拟合

本文的新增确诊人数在实际统计中为正数，但是如果不做正数的限制，其傅里叶基函数拟合结果在一些时间点可能会出现负数，这是违背现实的。因此，可以利用指数函数恒为正数的特点产生限制。$e^{W(t_i)}$作为拟合后的函数型数据，而不是上文中的$f(t)$，经过指数函数的严格限制，利用公式（2），新增确诊这一函数型数据的估计将不会出现与现实相违背的负值。

$$W(t) = \phi(t)^T C$$
$$\mathrm{SSE_C(C)} = \sum_{i=1}^{n} \left(y_i - e^{W(t_i)}\right)^2 \quad (2)$$

3.1.3 傅里叶基拟合的粗糙惩罚

傅里叶基直接拟合后的结果学习了太多不必要的数据特征，结果不平滑，存在着过拟合的现象，不再展示。下列公式（3）和（4）为分别对非限制和正数限制条件下的傅里叶基拟合进行粗糙惩罚的方式, L为Harmonic Acceleration微分算子，平滑系数预设为$\lambda = 10^4$，得到的粗糙惩罚结果如下图 3.1 所示，粗糙惩罚后的数据均呈现出较为平滑状态，且下部分正数限制（借助指数函数）下的粗糙惩罚估计未出现负值。粗糙惩罚后的结果更好的提取了 30 个省份新增确诊人数的趋势变化信息，提供了合理的函数型数据原材料。

Harmonic Acceleration 微分算子：$Lf = \omega^2 Df + D^3 f$

$$\mathrm{PENSSE_L(f)} = \sum_{i=1}^{31}(y_i - f(t_i))^2 + \lambda \int [Lf(t)]^2\, dt \qquad (3)$$

$$LW = \omega^2 DW + D^3 W$$

$$\mathrm{PENSSE_\lambda(W)} = \sum_{i=1}^{31}\left(y_i - e^{W(t_i)}\right)^2 + \lambda \int [LW(t)]^2\, dt \qquad (4)$$

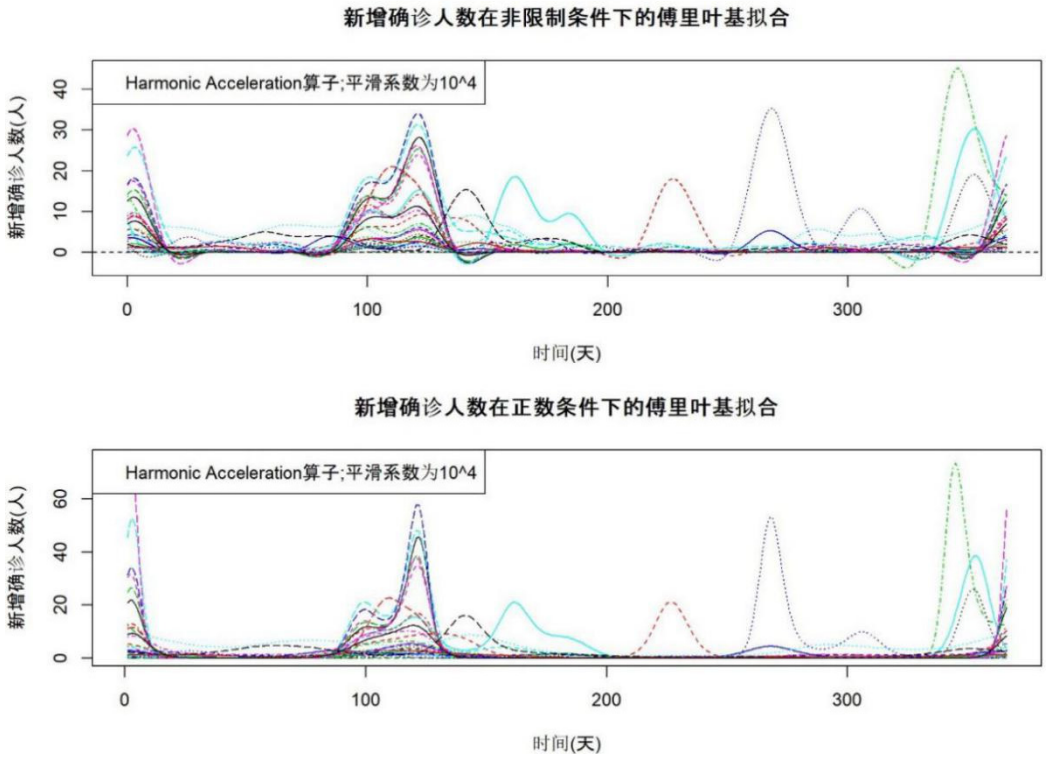

图 3.1 粗糙惩罚后的傅里叶基拟合

3.2 30 个省级行政单位新增确诊的相关性分析

$x_i(t)$，其中 i = 1 … 31，代表省级行政单位i在时间t的新增确诊人数，$\bar{x}(t)$为时间t所对应的新增确诊均值，$\sigma_{xx}(s,t)$为样本的函数型协方差，$\rho_{xx}(s,t)$为样本的函数型自相关系数。

$$\bar{x}(t) = \frac{1}{31}\sum_{i=1}^{31} x_i\,(t)$$

$$\sigma_{xx}(s,t) = \frac{1}{31}\sum_{i=1}^{31}\big(x_i(s) - \bar{x}(s)\big)\big(x_i(t) - x(t)\big)$$

$$\rho_{xx}(s,t) = \frac{\sum_{i=1}^{n}\big(x_i(s)-\bar{x}(s)\big)\big(x_i(t)-\bar{x}(t)\big)}{\sqrt{\sum_{i=1}^{n}\big(x_i(s)-\bar{x}(s)\big)^2}\sqrt{\sum_{i=1}^{n}\big(x_i(t)-\bar{x}(t)\big)^2}} \qquad (5)$$

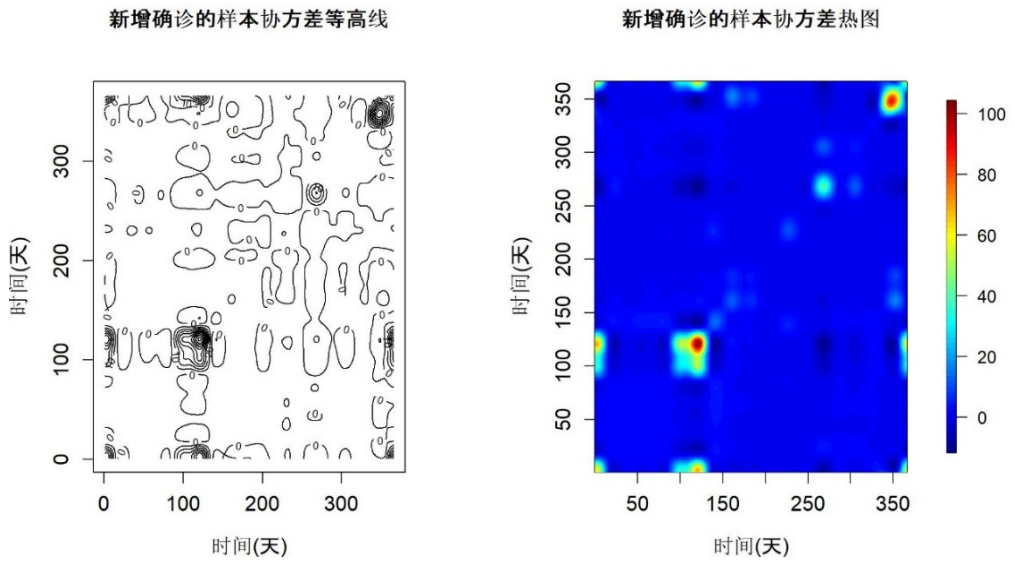

图 3.2 函数型样本协方差

函数型样本协方差如上图 3.2 所示，等高线图和热图均对不同时期的函数型协方差表现出不明显的差别，样本协方差相对变化不大。但是图 3.3 的函数型样本相关系数却表现出了明显的相关性：

（1）等高线图和热图在远离对角线的位置均出现了相关性高达 0.7-0.8 的区域，说明疫情的爆发前后相关，间隔 5-6 个月会有集中地新冠疫情爆发，具有周期性的特点。这与美国哈佛大学公共卫生学院的研究人员利用美国的 β 属冠状病毒的时间序列观测数据，模拟出的新冠病毒在温带地区的传播方式与特征的结果类似。可以

利用疫情的周期性构建完备的疫情防控屏障，在事前防范，超前部署，多重准备来消除疫情治理隐患。

（2） 右侧的热图，对角线附近的一些块状区域呈现出 0.5 以上的高相关性，说明新冠疫情会在集中的时间爆发，确诊病例出现后需要高度重视，普及医学观察的意义、法律依据和需要特别注意的问题等，对密切接触人群采取医学隔离检测，14 天集中观察等措施（重点观察有发热、咳嗽、气促等急性呼吸道感染的类似症状）。每日统计汇总，且实时向所在地的卫生健康部门咨询和报告。

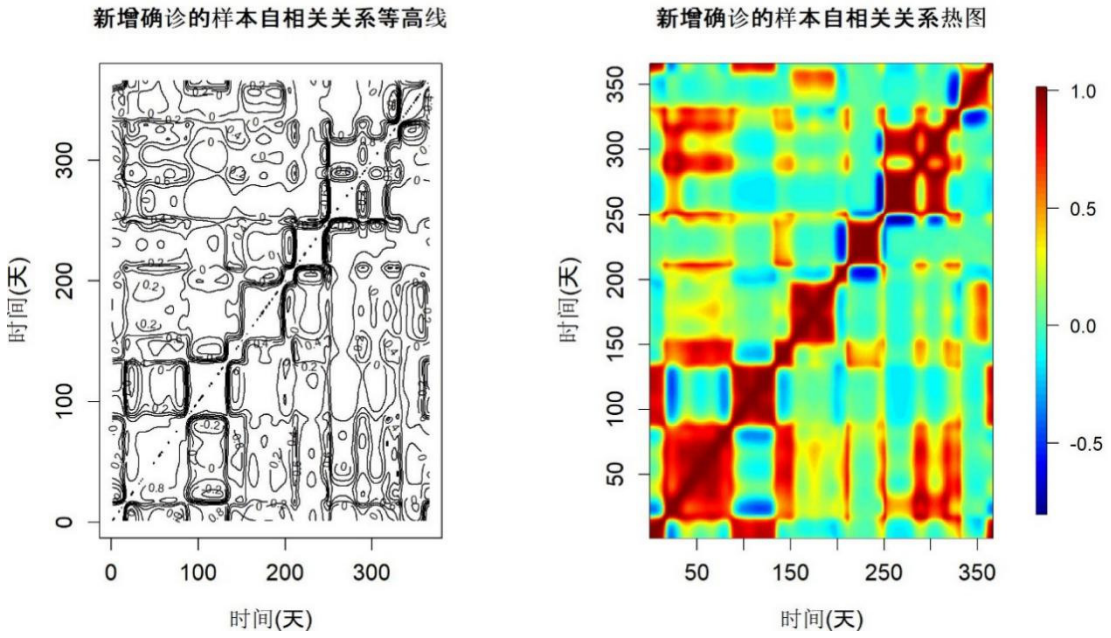

图 3.3 函数型样本自相关系数

3.3 一元函数型主成分分析

为了尽可能的保持原始数据最多的信息，防止限制条件对傅里叶拟合造成过多的信息丢失，因此在非限制条件下对 30 条新增确诊曲线的均匀数据做函数型主成分分解，目标是得到多条包含原始数据信息的函数型主成分曲线，提取出非限制条件下函数型原始数据的绝大部分信息，进行下文的主成分得分的方差分析、主成分平滑和和函数型主成分预测等内容。

3.3.1 函数型主成分的提取方法和提取效果

（1） 函数型主成分的提取方法，包括以下几个步骤

中国大陆 30 个省级行政单位所分别对应的 30 条新增确诊人数的曲线：$X_1^*(t), X_2^*(t), \ldots X_i^*(t) \ldots, X_{31}^*(t)$，其中$i = 1 \ldots 31$；

数据中心化：$X_i(t) = X_i^*(t) - \bar{X}(t)$；

均值曲线：$\bar{X}(t) = \frac{1}{31}\sum_{i=1}^{31} X_i(t)$；

第一函数型主成分曲线：$w_1(t)$，代表 30 条曲线强度第一的波动方向；

第一函数型主成分得分：$s_{i1} = \int w_1(t)X_i(t)\,dt$；

当时$\int w_1^2(t)dt = 1$最大化$\sum_{i=1}^{n} s_{i1}^2$；

第二函数型主成分曲线：$w_2(t)$，代表 30 条曲线第二强的波动方向，与第一函数型主成分曲线$w_1(t)$的积分正交；

第二函数型主成分得分：$s_{i2} = \int w_2(t)X_i(t)\,dt$；

当时$\int w_1^2(t)dt = 1$和$\int w_1(t)w_2(t)\,dt = 0$时最大化$\sum_{i=1}^{n} s_{i1}^2$；

同样的，可以得到其他积分后相互正交的函数型主成分曲线$w_3(t), \ldots, w_k(t), \ldots$。

（2） 函数型主成分的提取效果

最大化$\sum_{i=1}^{n} s_{i1}^2$ 得到的各个主成分得分的方差结果、利用相对比例计算累计方差被解释的份额。新增确诊人数的函数型主成分提取效果如下

图 3.4 所示，从其中左图可以直观的观测出前四个函数型主成分得分$s_{i1}, s_{i2}, s_{i3}, s_{i4}$可以解释 30 条新增确诊曲线的大部分波动信息，而之后的主成分得分包含的信息较少；由右图得到前四个主成分得分的累计方差占总方差的 90%左右，函数型主成分平滑和对新增确诊人数的函数型主成分预测可以提取包含绝大多数信息的前四个主成分$w_1(t), w_2(t), w_3(t), w_4(t)$。

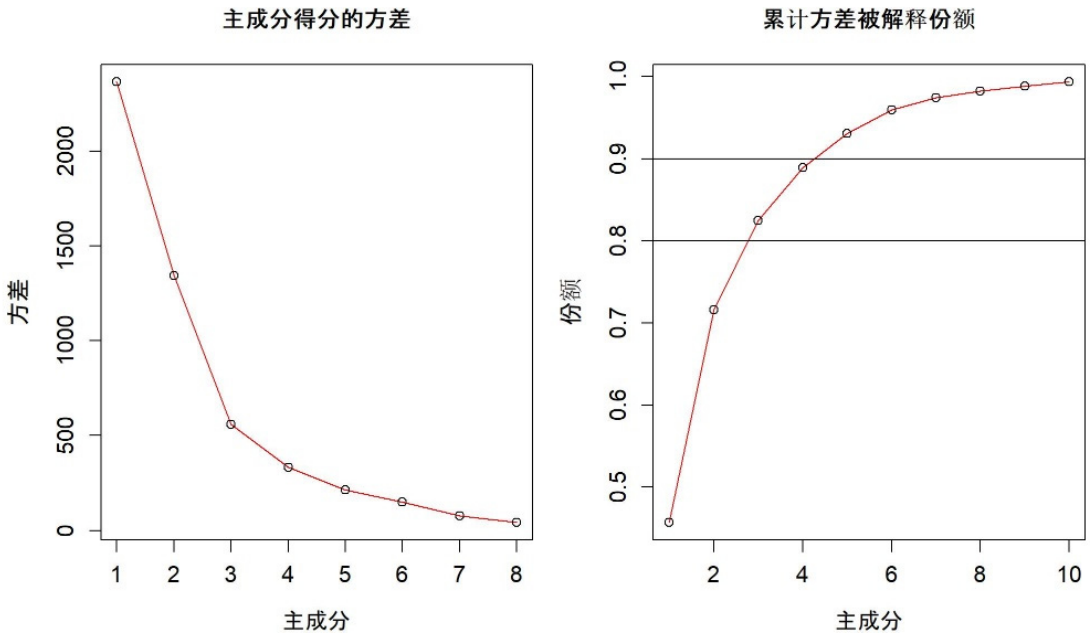

图 3.4 函数型样本相关系数

3.3.2 函数型主成分的计算技巧与结果展示

30 条新增确诊曲线： $X_1^*(t), X_2^*(t), \ldots X_i^*(t) \ldots, X_{31}^*(t)$，其中$i = 1 \ldots 31$分别代表中国大陆 30 个省级行政单位（如北京市、天津市和河北省等）的新增确诊人数数据，以北京市，河北省和天津市为例，如

图 3.5 的上部分所示。数据中心化的计算方式为：$X_i(t) = X_i^*(t) - \overline{X}(t)$，其中$\overline{X}(t)$为 30 个省级行政单位对应函数型变量的均值，也以北京市，河北省和天津市为例，如

图 3.5 的下部分所示。中心化的新增确诊数据可以更好地反映各个省级单位在其他省份的平均水平和平均波动幅度，以及该行政单位表现出的特异性信息。由原始的拟合数据和中心化后的数据综合分析出，天津市在这 366 天的中期有两个疫情集中爆发的时段（2020 年 4 月左右和 9 月左右），河北省在末期（2020 年 11 月左右）有一个疫情集中爆发时段，北京市的新增确诊虽然偶有出现，但是新冠疫情的爆发规模并没有过多地超过国内 30 个省级单位的新增确诊人数均值，对比来说疫情的控制相对有力、得当。

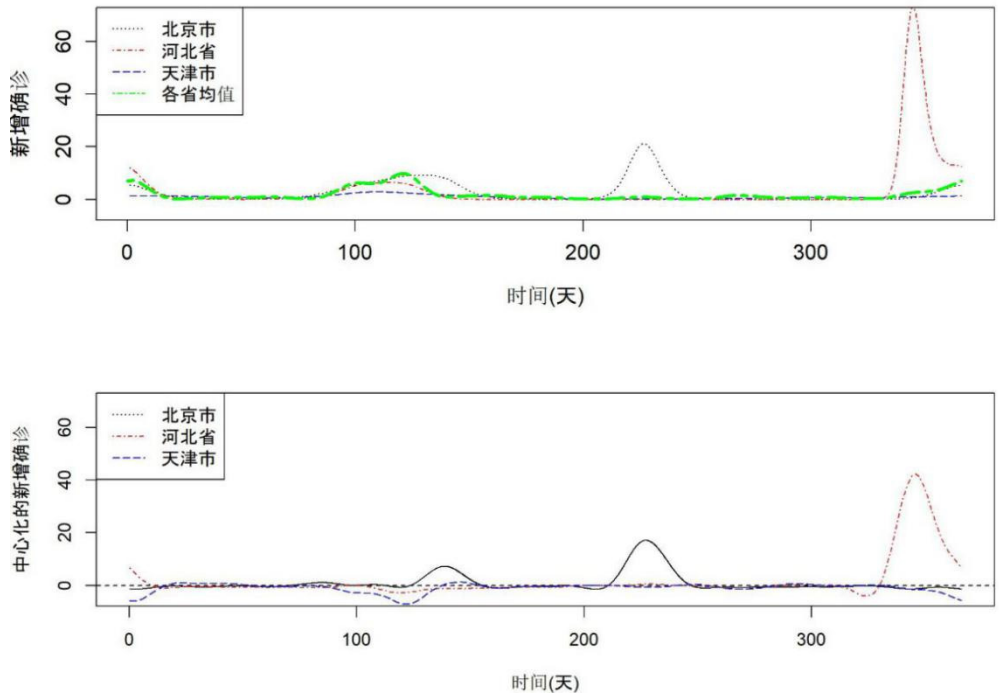

图 3.5 中心化的新增确诊数据

样本的函数型协方差：

$$\sigma(s,t) = n^{-1}x(t)x^T(t) = n^{-1}\phi^T(s)C^TC\phi(t);$$

函数型主成分向量：

$$w(t) = \phi^T(t)b，\text{定义：}\Phi = \int \phi(t)\phi^T(t)\,dt;$$

带入$\int \sigma(s,t)\,w(t)dt = \rho w(s)$可得：

$n^{-1}\phi^T(s)C^TC\int \phi(t)\phi^T(t)dtb = \rho\phi^T(s)b$，即为：$n^{-1}C^TCWb = \rho b$，且$b^TWb = 1$；

分解上式为：

$$n^{-1}W\Phi^{1/2}C^TC\Phi^{1/2}\Phi^{1/2}b = \rho\Phi^{1/2}b,$$

定义$u = W^{1/2}b$，带入$n^{-1}\Phi^{1/2}C^TC\Phi^{1/2}u = \rho u$，同时满足$u^Tu = 1$。

u即为$n^{-1}\Phi^{1/2}C^TC\Phi^{1/2}$的特征根对应的特征向量。由此，从原始数据中求解出 4 个函数型主成分$w(t)$。如下图 3.6 所示，左图的粗红线展示的是 30 个省级单位的均值，右图展示的是利用上述方法计算的四个函数型主成分，其中小幅

度波动过多，需要进一步的平滑处理（也就是粗糙惩罚）。

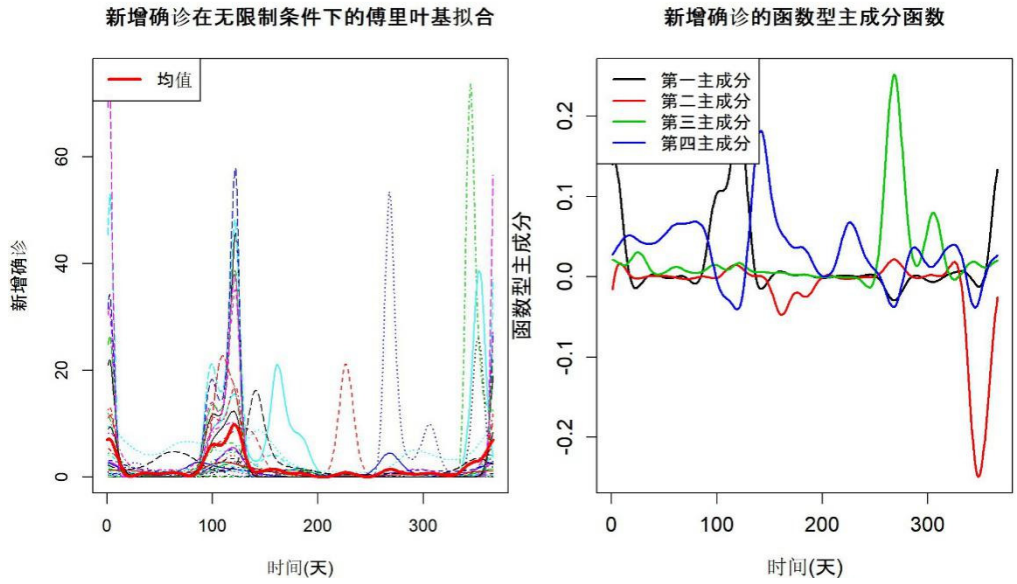

图 3.6 傅里叶基拟合和函数型主成分

3.3.3 函数型主成分的平滑方法和函数型主成分估计方法

（1） 函数型主成分的平滑

为了减少函数型主成分的高频震荡，更好的提取其趋势性信息并利用主成分构建函数型数据的预测模型，需要对提取的函数型主成分$w(t)$的噪声（粗糙惩罚）进行平滑处理，替换上文 3.3.1 中最大化$\sum_{i=1}^{n} s_{i1}^2$ 得到的$w_1(t), w_2(t), w_3(t), w_4(t)$结果。

新的平滑函数型主成分计算方法为最大化公式：$\frac{Var[\int w(t)x_i(t)dt]}{\int w(t)^2dt+\lambda\int[Lw(t)]^2dt}$，即为对 Harmonic Acceleration 微分算子下的$w(t)$进行粗糙惩罚，定义粗糙惩罚的尺度为$\int w(t)^2dt + \lambda\int[Lw(t)]^2\,dt$，得到平滑前与后的对比如下图 3.7 所示，通过上图未经过粗糙惩罚的函数型主成分和经过粗糙惩罚的函数型主成分对比可以发现，函数型主成分变得更为平滑，说明影响信息提取与识别的噪声被移除，可以利用更新的平滑$w_1(t), w_2(t), w_3(t), w_4(t)$对河北省新增确诊人数进行主成分估计、预测。

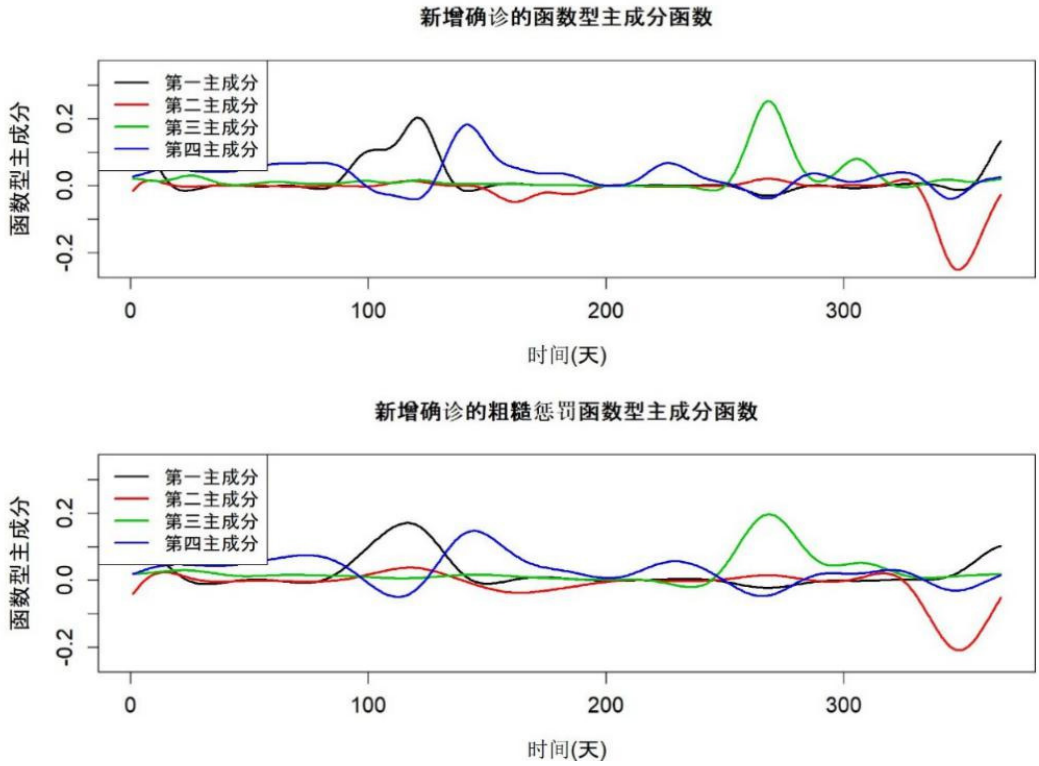

图 3.7 函数型主成分的平滑处理

（2） 函数型主成分估计方法（预测河北省新增确诊人数）

基于函数型主成分对河北省新增确诊人数进行预测，从 30 个省份新增确诊的函数型数据中剔除河北省的数据，并截取前 355 天的数据作为训练集，定义新的$X(t)$，截取 356 天之后的数据作为模型对未来预测值的对比基准，利用下述模型：

$$X(t) = \beta_0 + \beta_1 \mathrm{w}_1(\mathrm{t}) + \beta_2 \mathrm{w}_1(\mathrm{t}) + \beta_3 \mathrm{w}_1(\mathrm{t}) + \beta_4 \mathrm{w}_1(\mathrm{t}) + \epsilon(t)$$

其中$\epsilon(t)$为独立同分布的随机误差 （6）

分别得到 4 个函数型主成分$w_1(t), w_2(t), w_3(t), w_4(t)$对应的参数估计值$\beta_1, \beta_2, \beta_3, \beta_4$以及截距项$\beta_0$。利用上述结果，分别计算对前 355 天河北省新增确诊人数的估计和对 356 天到 366 天河北省新增确诊人数的估计，预测结果如下图 3.8 所示。容易看到，河北省新增确诊人数的观测值基本被上述两部分预测结果准确预测。基于函数型主成分的预测模型预测的河北省新增确诊人数的结果基本可信。

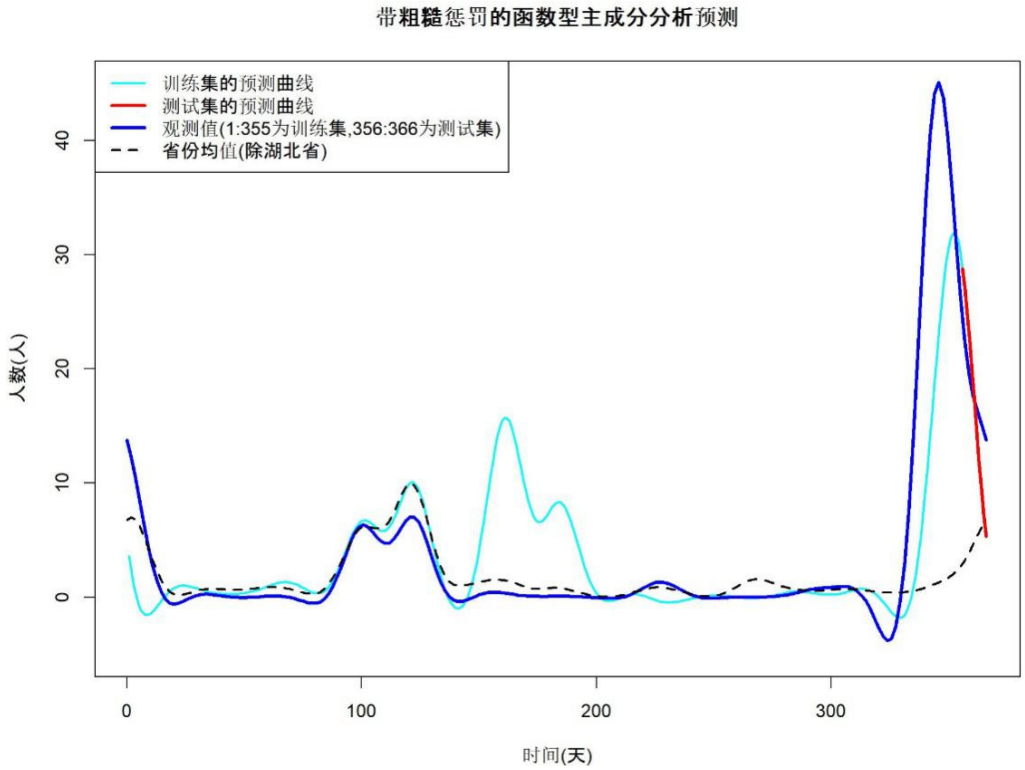

图 3.8 基于函数型主成分估计的预测

第4章 多元函数型数据分析

4.1 新增确诊与新增死亡的相位图

对新增确诊与新增死亡的数据继续用傅里叶基拟合和基于 Harmonic Acceleration 微分算子的平滑方法对粗糙曲线进行惩罚，平滑系数缩小为$\lambda = 10^1$。如下图 4.1 所示，下图 4.1，将原始数据与粗糙惩罚后的数据各自的相位图进行了对比。由引入粗糙惩罚后的相位图可以看出，30 个省份的均值相位图均接近椭圆形，新增确诊与新增死亡的数据随着时间的增减同步变化，反映出数据的周期性特征。此外，图中 31 个省级单位新增确诊和新增死亡的相互作用模式被清晰地分离开来，图中得到的 30 个相位图重叠率较低，由此可见各个省级单位对新增确诊与新增死亡的数据分别有其特异性的模式。

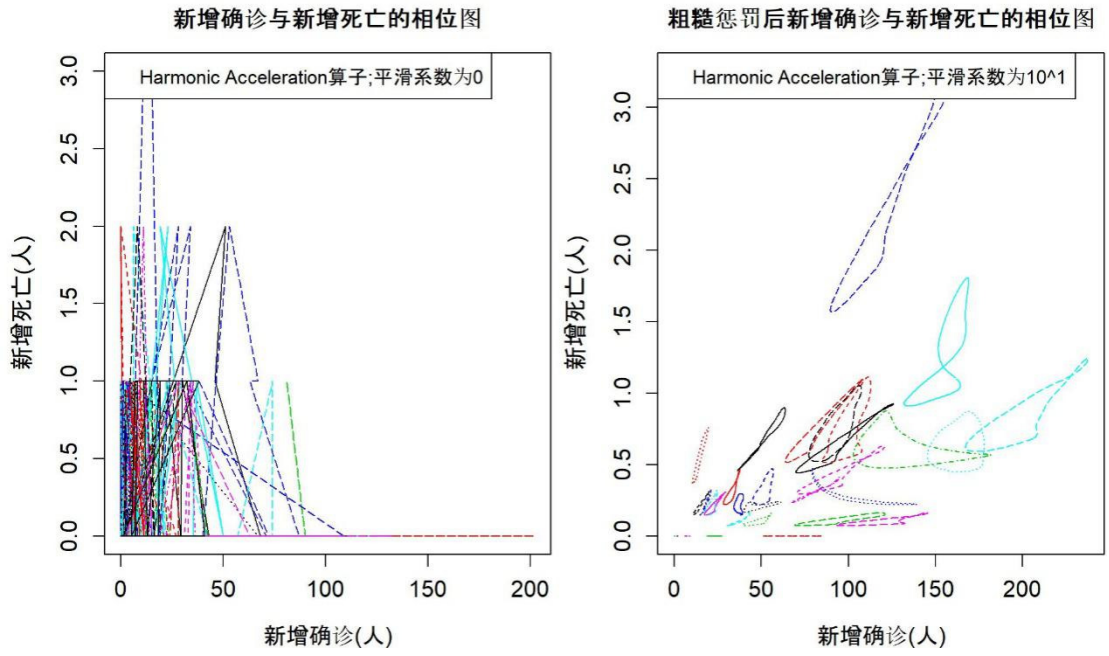

图 4.1 基于函数型主成分估计的预测

疫情初期正值人口流动性较大的春节，其余 30 个省级行政单位的最初病例多来源于湖北省内人口大规模向省外流动。经过时间的推移各省病例有外部输出逐步转变为以本土内传播为主。姚强、张柏杨等人通过 1 阶惩罚拟似然法拟合多

水平负二项回归对各省疫情发展趋势差异进行了分析，各省市置信区间的交叉情况也表明了各省疫情发展趋势间的差异性[17]。在疫情爆发初期，输入性病例基数是各省初期新增确诊增多的重要因素，由于各个省级行政单位启动一级响应的时间不同，对初期防止输入性病例的遏制效果影响了后续的疫情在本土的发展情况。此外在 1 月 23 日，全国各省均启动一级响应，本省疫情由输入性病例为主转变为本土传播为主。各省的疫情防控落实情况、对疫情趋势的总体研判、防疫积极性等均存在差异，导致各省新增确诊与新增死亡数据存在特异性。

4.2 30 个省级行政单位新增确诊和新增死亡的相关性分析

$x_i(t)$和$y_i(t)$，其中 $\mathrm{i} = 1 \ldots 31$，分别代表省级行政单位i在时间t的新增确诊和新增死亡人数，$\bar{x}(t)$和$\bar{y}(t)$为时间t所对应的新增确诊和新增死亡均值，下式(7)中$\rho_{xy}(s,t)$为样本的函数型相关系数，新增确诊和新增死亡的函数型自相关系数与式(5)有一样的定义。

$$\rho(s,t) = \frac{\sum_{i=1}^{n}(x_i(s)-\bar{x}(s))(y_i(t)-\bar{y}(t))}{\sqrt{\sum_{i=1}^{n}(x_i(s)-\bar{x}(s))^2}\sqrt{\sum_{i=1}^{n}(y_i(t)-\bar{y}(t))^2}} \qquad (7)$$

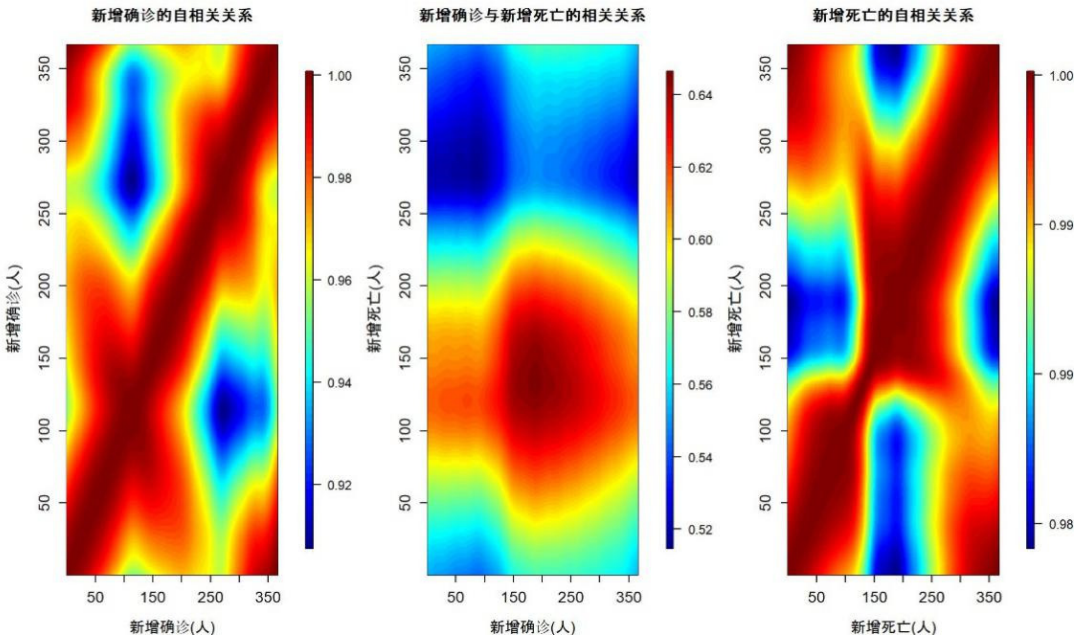

图 4.2 函数型相关系数热图

函数型样本自相关与相关的热图如上图 4.2 所示，新增确诊和新增死亡人数的自相关图位于两侧，其自相关性类似，对角线附近呈现高相关性，新增确诊人数和新增死亡人数均在集中的时间快速增长，突出了新冠疫情在集中时间爆发的特点。同时，疫情初期由于医疗资源受限，部分患者居家隔离，此时疫情确诊和死亡人数急剧上升，中间的两个函数型变量的相关系数在 4-7 月表现出了较强的相关性。伴随着全面核酸检测的开展，以及检测能力的提高，未被检测到的感染者被发现，因此新增确诊仍在增加。但由于我国调集优势医疗资源紧急设立发热门诊，大面积改建方舱医院，扩充新增确诊患者的收治量，使新增确诊的就医效率大大提高，死亡率得到控制；我国通过临床筛选出的“三药三方”疗效确切，医院实行严格三级分诊，就诊流程也得到了优化，降低了交叉感染的风险，这也在一定程度上控制住了新增死亡数，在周期末期新增确诊和新增死亡人数的相关关系已经实现大幅降低。

4.3 均值相位图与多维主成分分析后方差累计被解释份额

多维函数型主成分分析方法返回的新增确诊和新增死亡人数在 30 个省份的均值相位图接近椭圆形，可以反映出这一组函数型数据是存在周期性特征的，且随着时间的变化同步的增减，从而实现一个年度的循环。同时，通过计算新增确诊和新增死亡人数$(x(t), y(t))$在函数型主成分$(w_x(t), w_y(t))$的最大化方差$Var\left[\int w_x(t)x(t)dt + \int w_y(t)y(t)dt\right]$，得到函数型主成分得分对原始数据的波动信息的捕捉情况，前四个函数型主成分对$(x(t), y(t))$的解释力度高于 99.95%，基本的包含了数据的全部信息。

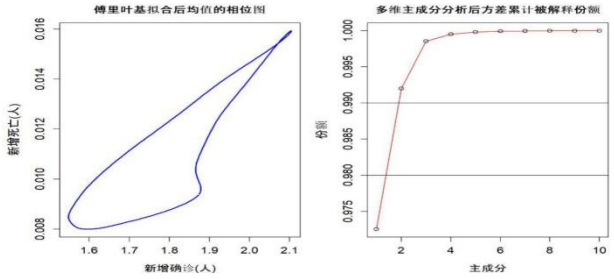

图 4.3 均值相位图与方差累计被解释份额

找到多组正交的函数型主成分，如下图 4.4：此时新增确诊的函数型主成分已经不能突出四个清楚的周期，并且新增死亡的函数型主成分两两类似，分别在 4-6 月份达到顶点和低点。对于新增确诊的主成分分析：第一主成分为黑色曲线表示，且在观察时间区间上较为稳定，代表的是 30 条曲线的等权重信息；第二主成分为红色，前半年较高后半年较低，代表了 30 条曲线的春、夏季和秋、冬季的对比；第三主成分曲线为绿色，代表春、秋季与夏、冬季的对比；第四主成分为蓝色，代表春、冬季与夏、秋季的对比。对于新增死亡的主成分分析：第一主成分和第二主成分的估计均为正；第三主成分和第四主成分的估计均为负，最大的差别出现在 4-7 月，此时新增确诊和新增死亡人数激增，带来强烈的曲线波动。

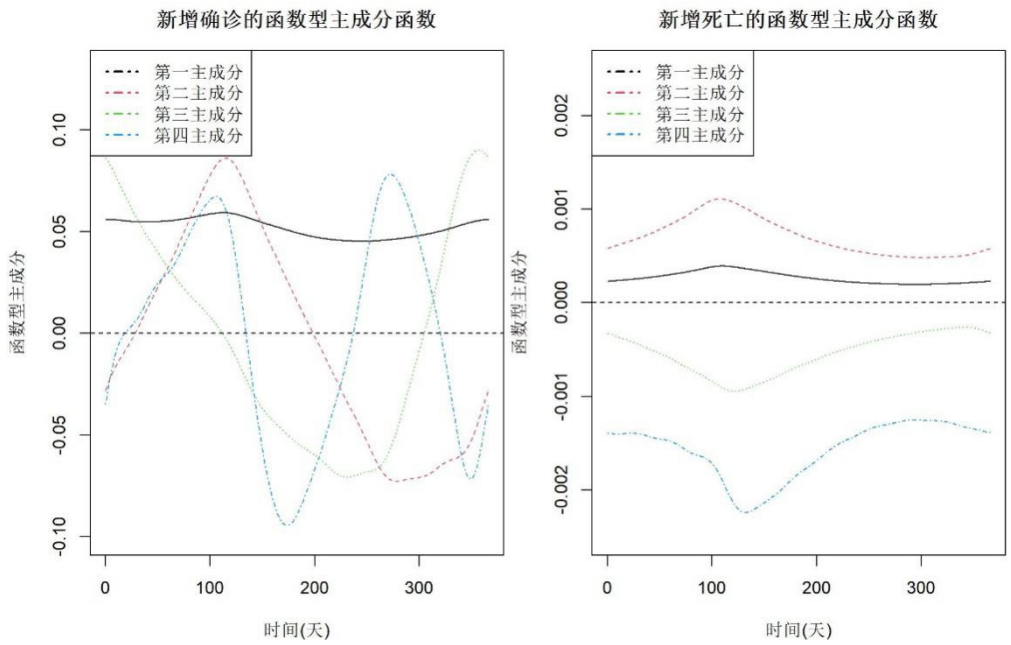

图 4.4 多维函数型主成分估计

4.4 函数型数据对函数型数据的回归分析

4.4.1 加入粗糙惩罚的傅里叶基拟合

基于 30 个省份的新增确诊和新增死亡人数，用公式(3)的方法进行带有粗糙惩罚的非限制条件下傅里叶基拟合（对函数型数据不预设限制条件会最大程度上保留原始数据的趋势性信息）。设置粗糙惩罚的平滑系数与上文中一维函数型粗糙惩罚形同，为$\lambda = 10^4$，得到的粗糙惩罚结果如下图 4.5 所示，上下两部分函数曲线均呈现平滑状态，即提取了 30 个省份新增确诊和新增死亡人数去除噪声的平滑波动信息，为下文的函数型对函数型数据回归提供了基于 30 个省份的新增确诊和新增死亡人数的自变量和因变量。对比这 366 天新冠疫情的周期数据发现，进入冬季后，不少省份新增确诊周期性的增加，但新增死亡人数却趋于稳定、变化较小。此时各个省级行政单位进入常态化防疫阶段，主要原因是：

（1） 核酸检测能力大幅增强、范围逐步扩大，以往难以检测的症状较轻或潜在患者被锁定为新增确诊，并且这些新增确诊患者的治愈难度较小，死亡率较低；

（2） 根据对新冠疫情死亡病例的研究对比，对感染病毒后死亡率较高的人群有了有效的医治方案，并对其进行了更严格、成熟的密集接触者排查，大大降低了感染率；

（3） 医护人员的救治经验更加丰富，同时发热门诊、方舱医院的设立使得各省的医疗挤兑问题大大减轻，可以及时提供医疗服务，避免新增确诊由轻转重，由重转危。

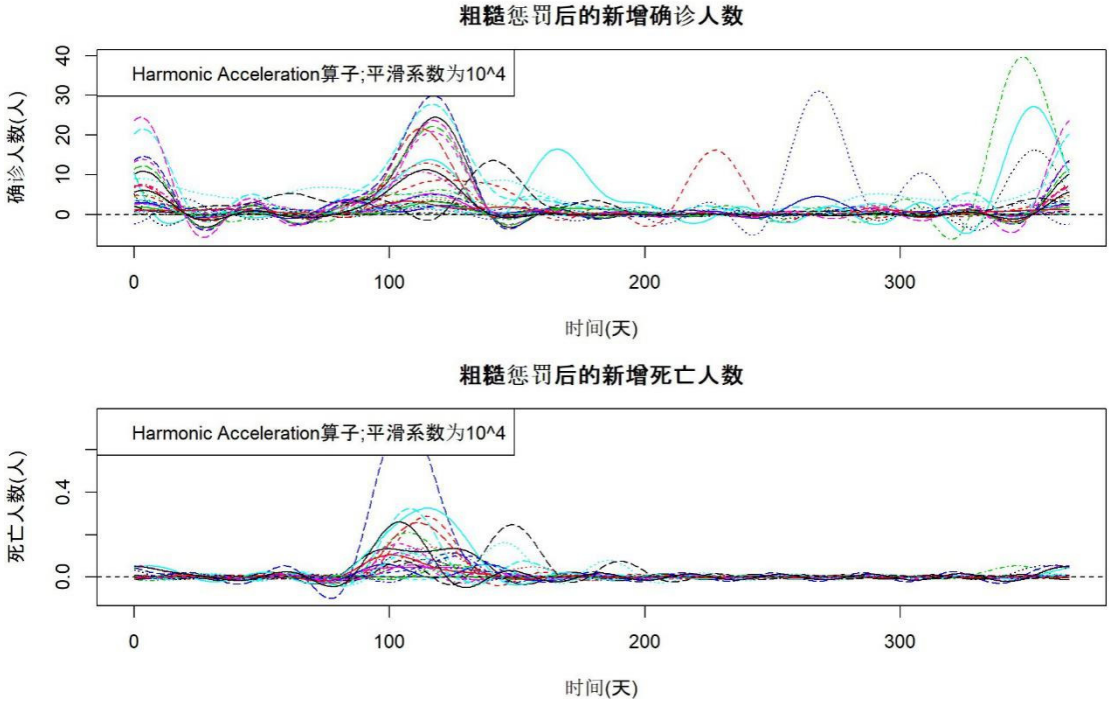

图 4.5 粗糙惩罚后的多维傅里叶基拟合

4.4.2 函数型对函数型数据的回归

（1） 模型建立

下式（8）中$(x_i(t), y_i(t))$代表省份i在时间t的新增确诊和新增死亡人数，$\epsilon_i(t)$为独立同分布的误差，且$i = 1 \ldots 31$，$\beta(s,t)$为时间s的新增确诊人数对时间t的新增死亡人数的作用。

$$y_i(t) = \int \beta(s,t) x_i(s) ds + \epsilon_i(t) \qquad (8)$$

（2） 参数估计方法

定义$\beta(s,t)=(\beta_0(t),\beta_1(s,t))^T$在$s$和$t$两个方向上的傅里叶基函数分别为$\phi(s)$和$\psi(t)$后，利用克罗内克张量积的计算法则分解累计积分平方误差。

$$\begin{aligned}SISE &= \sum\left[\int\left(y_i(t)-\int\beta(s,t)x_i(s)ds\right)^2dt\right]\\ &= \sum\left[\int\left(y_i(t)-\psi(t)B\int\phi(s)x_i(s)ds\right)^2dt\right]\\ &= \sum\left[\int\left(y_i(t)-\int\phi(s)x_i(s)ds\otimes\psi(t)vec(B)\right)^2dt\right]\end{aligned}$$

最小化累计积分平方误差$\boldsymbol{min}(SISE_\beta)$得到参数估计结果：

$$\begin{aligned}\widehat{\beta(s,t)} = &\left[\sum\left[\int\phi(s)x_i(s)ds\right]\left[\int\phi(s)x_i(s)ds\right]^T\right.\\ &\left.\otimes\int\psi(t)\psi(t)^Tdt\right]^{-1}\left[\sum\int\phi(s)x_i(s)ds\otimes\int\psi(t)y_i(t)dt\right]\end{aligned}$$

$\beta(s,t)$的参数估计结果如下图 4.6 所示，左侧为$\beta_0(t)$的估计结果，右侧为$\beta_1(s,t)$的估计结果。参考图 4.5 粗糙惩罚后的多维傅里叶基拟合，最明显的发现是：左侧的$\beta_0(t)$提取了新增死亡变化的趋势性信息，从 2020 年 1 月 28 日到 2021 年 1 月 27 日间， 30 各个省级行政单位的新增死亡人数呈现周期性的性质，并且在寒冷的冬、春季死亡率较高，温暖的夏、秋季死亡率较低；右侧的$\beta_1(s,t)$提取了去趋势后时间s的新增确诊对时间t的新增死亡的作用，值得注意的是，$\beta_1(s,t)$代表的信息补充了函数型数据新增确诊人数对新增死亡人数不能被$\beta_0(t)$提取的部分。

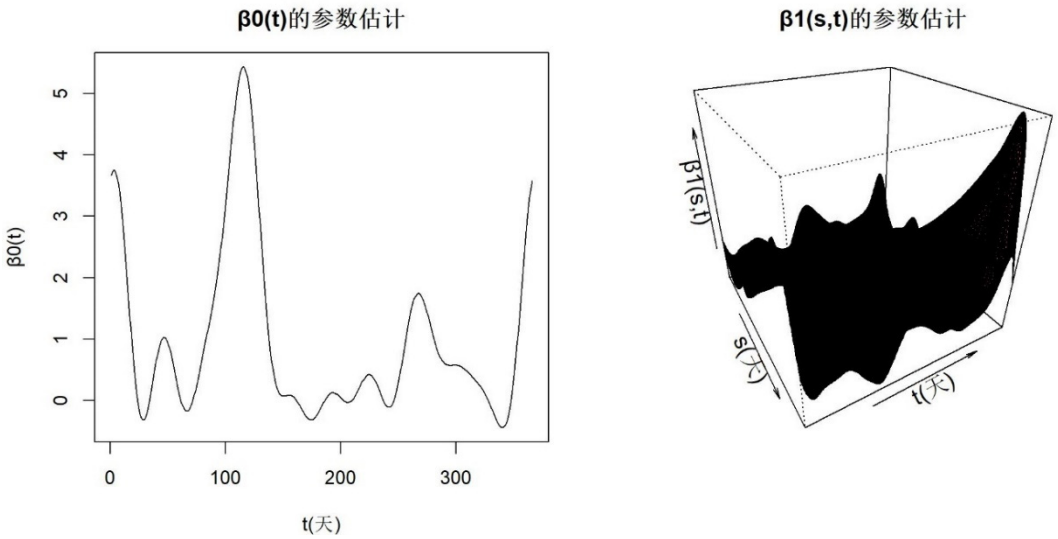

图 4.6 β的参数估计结果

在前 150 天，疫情尚未得到有效的控制，新增确诊对新增死亡的作用正向且较强，感染率和死亡率都在不断攀升。此时医疗挤兑问题严重，如果不能及时解决则很可能形成恶性循环，对疫情造成进一步的冲击。因此医疗资源在调配时应侧重遵循应急性原则，将资源向疫情重灾区倾斜，遏制疫情蔓延势头。在第 150 天到 330 天左右，新增确诊对新增死亡的作用较弱，全国本土疫情传播基本阻断，国内各省已经进入常态化防控阶段。境内疫情零星散发，但境外疫情快速扩散蔓延，境外输入病例可能会引起疫情反弹。但是从 330 天到 366 天，由于疫情的周期性因素，新增确诊和新增死亡人数再次攀升，因此医疗资源在调配时要考虑到疫情再次反弹的不确定性，更侧重动态调整原则，及时调动有限的资源随时服务于最需要的地区。

第5章 研究结论

5.1 利用新冠疫情爆发的周期性制定防疫措施

函数型样本的自相关关系、函数型主成分的提取以及新增死亡人数对新增确诊人数的函数型数据回归的参数估计均可揭示出新冠疫情相关的新增确诊人数和新增死亡人数呈现出规律的周期性性质：结合 30 个省级行政单位的综合情况，间隔 5-6 个月集中爆发的概率较高，且在寒冷的冬、春季发病率和死亡率较高，温暖的夏、秋季发病率和死亡率较低，由此带来的省级行政单位疫情防控的建议：

（1） 科学规范公众防控行为，完善群防群控的工作机制。新冠病毒在低紫外线、温度低的情况下存活的时间较长，因此多在寒冷季节加重，温暖季节平息。鼓励人民多参与体育活动，合理营养膳食，尤其注意换季影响，增强人体的免疫力，不让新冠病毒有机可乘；

（2） 利用疫情的周期性构建完备的疫情防控屏障。合理调配冬、春季的春运运量，减少走亲访友等聚集性活动和分散造成大规模人口流动的节日假期，推进有关疫情防控专业科普，提高公众的风险防控意识，降低相关的疫情传播风险；

（3） 加大快速监测技术的研发投入，构筑积极应对重大疫情和公共卫生的群众基础。促进我国科技工作者对病原基础研究、接种疫苗、抗病毒药物的研发，进一步做好疫情防控和检测产品的科技开发。海关等部门严控进口冷链食品和国外输入病例的入境风险，同时，居民群众注意安全防护，佩戴口罩、注意个人卫生、避免去海鲜市场、封闭环境等风险较高区域。

5.2 根据新冠疫情集中爆发且各省份之间各有模式的特点研究防控策略

上文傅里叶基拟合后的函数型数据、函数型相关系数热图和参数估计的相对大小都表现出了新冠疫情在集中时间爆发的特点；并且北京市，天津市和河北省新增确诊的变化特征以及粗糙惩罚后的省份数据各自的相位图揭示出了不同省

级行政单位疫情发展的不同模式。需要依据新冠疫情集中爆发且各省份之间各有模式的特点，因地制宜、因时制宜的研究防控策略。

（1）对密切接触者重点监控、疑似污染源采取必要消毒杀菌作业

对确诊患者的密切接触者在指定场所进行 14 天的集中医学观察（医院、方舱医院或改造的集中隔离点等）或居家医学观察并进行必要的核酸检测。期间，密切接触者按规定每天早、晚各进行一次体温测量，并实时监测身体健康状态。与此同时，精准的确定所有次级密切接触者，核酸检测后进行必要的管控和追踪记录。 为避免密切接触者再次扩大，根据确诊患者和密切接触者的活动、生活轨迹，对相应的场所采取有效的消杀作业，防止疫情集中爆发。

（2）根据不同省份之间地区差异，分情况实施防控措施

各省级单位医疗卫生资源、医护人员配备情况、对疫情反应时间、地理位置及人口流动性等因素不尽相同，造成了各省级单位对疫情来袭时不同的反映状态，因此需要根据不同省份的实际情况实施防控措施。

a. 对新冠疫情大规模爆发的地区。例如 2020 年 1-3 月份的湖北省、2020 年 11 月到 2021 年 1 月份的河北省等，需要集中全国的资源和力量进行抗“疫”援助，其主要采取的疫情防控措施有：实行“战时管制”的封闭隔离措施，仅开放疫情防控物资绿色通道，严格管控进出人员；根据党中央国务院的决策部署，紧急建成必须的隔离、救治医院，并将多处公共场所改建为方舱医院，全力接收新增确诊患者；对一线医护人员实行倾斜政策，鼓励医护人员参与一线防控；对进口冷链物品做到“采样全覆盖、样本全检测、包装全消杀、商品全追溯”。

b. 疫情蔓延规模还较小的地区，如：西藏、新疆等，采取相对温和的疫情防控措施：全力防范风险地区人员的无序输入和本地区的监管漏洞，以“外防输入、内防扩散”的策略应对；同样对进口冷链物品做到全覆盖的检测、消杀和追溯；在人口易聚集的场所，出入时进行体温测量和人员必要信息登记。

参考文献

[1]Chan J F W, Yuan S F, Kok K H,et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission:a study of a family cluster[J]. The Lancet, 2020, 395(10223): 514-523.

[2]Lu R J, Zhao X, Li J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding[J]. The Lancet, 2020, 395(10224): 565-574.

[3]Li R Y, Sen P, Chen B, et al. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus(SARS-CoV2)[J]. Science, 2020, 368(6490): 489-493.

[4]Daniel W, Nianshuang W, S C K, et al.Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation[J]. Science, 2020, 367(6483): 1260-1263.

[5]Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor[J]. Cell, 2020, 181(2): 271-280.

[6]Huang C L, Wang Y M, Li X W, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan,China[J]. The Lancet, 2020, 395(10223): 497-506.

[7]Ai T, Yang Z L, Hou H Y, et al. Correlation of Chest CT and RT-PCR Testing for Coronavirus Disease 2019(COVID-19)in China: A Report of 1014 Cases[J]. Radiology, 2020, 296(2): E32-E40.

[8]Jin Y H, Cai L, Cheng Z S, et al. A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus(2019-nCoV) infected pneumonia(standard version)[J]. Military Medical Research, 2020, 7(1): 1-22.

[9]Wang Yubin, Wang Jingjing, Wang Xiaoyang. COVID-19,supply chain disruption and

China’s hog market: a dynamic analysis[J]. China Agricultural Economic Review, 2020, 12(3): 427-443.

[10]Mahajan K, Tomar S. COVID-19 and supply chain disruption: evidence from food markets in India[J]. Amer. J. Agr. Econ.,2021,103: 35-52.

[11]Lin B X, Zhang Y Y. Impact of the COVID-19 pandemic on agricultural exports[J].Journal of Integrative Agriculture, 2020, 19(12): 2937-2945.

[12]Chang H, Meyerhoefer C D. COVID‐19 and the Demand for Online Food Shopping Services: Empirical Evidence from Taiwan[J]. American Journal of Agricultural Economics, 2020, 103(2):448-456.

[13]边永民. 新型冠状病毒全球传播背景下限制国际贸易措施的合规性研究[J]. 国际贸易问题, 2020(7): 1-13.

[14]李先德, 孙致陆, 贾伟, 等. 新冠肺炎疫情对全球农产品市场与贸易的影响及对策建议[J]. 农业经济问题, 2020(8): 4-11.

[15]王国长. 函数数据回归与降维[D].东北师范大学,2012.

[16]闫星宇. 函数型线性回归的若干研究[D].华东师范大学,2020.

[17]姚强,张柏杨,李满娣,舒婷,吴晨瑶,杨燕玲,严可,蒋敏,朱彩蓉.非湖北省地区COVID-19 发病宏观影响因素及各省发病趋势差异初探[J].现代预防医学,2020,47(24)

附录

```
################################################################################
#####1.初步的数据处理
# 读取数据,有台湾+30 个省级,港澳体量较小未加入
# 新增和死亡从 2020 年 1 月 28 日开始全国统计，2021 年 1 月 27 日结束,2020 年为闰年 366 天
d1=read.csv("C:/Users/chang/Desktop/2021 全国统计建模/数据/新增和死亡.csv")
dim(d1)
# 查看是否有缺失,把变量日期改为时间格式
str(d1)
which(is.na(d1))
# 查看数据基本情况
head(d1,30)
length(d1)
dim(d1)
names(d1)
library(tidyr)
#设置 sep，把省份这一变量
d1=d1[-which(d1$省级单位=="湖北省"),]
d2=d1[,c(1,2,4)]
d3=d1[,c(1,2,3)]
nd=spread(d2,省级单位,新增死亡)
dim(nd)
ni=spread(d3, 省级单位, 新增确诊)
dim(ni)
# 得到两个数据集 nd 和 ni，一个为新增死亡，另一个为新增确诊，
# 时间跨度为 2020-01-28 到 2020-01-27
ni$日期
nd$日期
# 统计的长度 366
nrow(ni)
nrow(nd)
dayin=ni[,-1]
```

```
daydeath=nd[,-1]
#缺失值 2020-3-25 西藏自治区、青海为 NA,实际调查为 0
which(!complete.cases(dayin))
dayin[143,27]=0
dayin[143,19]=0
which(!complete.cases(daydeath))
daydeath[143,27]=0
daydeath[143,19]=0
dayin[143,]
daydeath[143,]
data=list(data.matrix(dayin),data.matrix(daydeath),colnames(dayin))
str(data)
names(data)=c("新增确诊",'新增死亡','省级单位')
#############################################################################
#####2.函数型数据限制条件下的傅里叶基拟合、非限制条件下的傅里叶基拟合、相关分析
# install.packages("fda")
library(fda)
daybasis366 = create.fourier.basis(c(0,366),366)
# harmonic acceleration differential  算子
# define the harmonic acceleration differential operator
# L = D^3 f(t) + 0 * D^2 f(t) + w^2 * D^f(t) + 0 * f(t)
# It will not penalize a+bt, cos(wt) and sin(wt)
# w = 2pi/period of the process
# c(0,(2*pi/365)^2,0) is the vector of coefficients to
# the lower derivatives
harmLfd = vec2Lfd(c(0,(2*pi/366)^2,0), c(0,366))
c19fdPar = fdPar(daybasis366,harmLfd,1e4)
c19fd = smooth.basis(1:366,data$新增确诊,c19fdPar)
# Tell 3 information
# 1:fbasis - Type of basis functions
# 2: harmLfd - Define the roughness penalty
# 3: lambda - value of the smoothing parameter
```

```
jpeg("1.jpg",height=1300,width=2000,res=210)
par(mfrow=c(2,1),mar = c(4, 2, 4, 0.2))
plot(c19fd$fd,xlab='时间(天)',ylab='新增确诊人数(人)',bty="l")
legend("topleft",c("Harmonic Acceleration 算子;平滑系数为 10^4"))
title(main="新增确诊在非限制条件下的傅里叶基拟合")
# smoothing when considering the positive contraint;
posc19fd= smooth.pos(1:366,data$新增确诊,c19fdPar)
posc19fd = eval.posfd(1:366,posc19fd$Wfdobj)
matplot(posc19fd,xlab='时间(天)',ylab='新增确诊人数(人)',type="l")
legend("topleft",c("Harmonic Acceleration 算子;平滑系数为 10^4"))
title(main="新增确诊在正数条件下的傅里叶基拟合")
dev.off()
#相关分析
#协方差
c19var = var.fd(c19fd$fd)
tvvals = eval.bifd(1:366,1:366,c19var)
jpeg("2.jpg",height=1200,width=2000,res=210)
par(mfrow=c(1,2),mar = c(4, 4, 6, 5))
contour(1:366,1:366,tvvals,xlab=' 时 间 ( 天 )',ylab=' 时 间
(天)',cex.lab=1.2,cex.axis=1.2,main="新增确诊的样本协方差等高线")
# install.packages("fields")
library(fields)
image.plot(1:366,1:366,tvvals,xlab=' 时 间 ( 天 )',ylab=' 时 间
(天)',cex.lab=1.2,cex.axis=1.2,main="新增确诊的样本协方差热图")
dev.off()
# 协相关系数
c19cor = cor.fd(1:366,c19fd$fd)
jpeg("3.jpg",height=1200,width=2000,res=210)
par(mfrow=c(1,2),mar = c(4, 4, 4, 5))
contour(1:366,1:366,c19cor,xlab=' 时 间 ( 天 )',ylab=' 时 间
(天)',cex.lab=1.2,cex.axis=1.2,main="新增确诊的样本自相关关系等高线")
image.plot(1:366,1:366,c19cor,xlab=' 时 间 ( 天 )',ylab=' 时 间
(天)',cex.lab=1.2,cex.axis=1.2,main="新增确诊的样本自相关关系热图")
dev.off()
```

```
#############################################################################
#####3.函数型主成分分析
# 在分限制条件下对 30 条新增确诊曲线做 4 函数型主成分分解
c19pca = pca.fd(c19fd$fd,nharm=4)
names(c19pca)
c19pca$varprop
#temppca$values are the eigenvalues
jpeg("4.jpg",height=1200,width=2000,res=210)
par(mfrow=c(1,2),mar = c(4, 5, 4, 0.2))
plot(c19pca$values[1:8],xlab='主成分',ylab='方差',type='l',
     cex.lab=1.2,cex.axis=1.2,cex=2,col=2,main="主成分得分的方差")
points(c19pca$values[1:8])
# 新增确诊曲线被函数型主成分解释的百分比
# It shows that the top 3 FPCs explains more than 99% of total variations
plot(cumsum(c19pca$values[1:10])/sum(c19pca$values),xlab='主成分',
     ylab='份额',col=2,cex.lab=1.2,
     cex.axis=1.2,cex=2,type='l',main="累计方差被解释份额")
points(cumsum(c19pca$values[1:10])/sum(c19pca$values))
abline(h=0.9)
abline(h=0.8)
dev.off()
# 均值曲线
jpeg("5.jpg",height=1200,width=2000,res=210)
par(mfrow=c(1,2),mar = c(4, 4, 4, 0.2))
matplot(posc19fd,xlab='时间(天)',ylab='新增确诊',type="l")
title(main="新增确诊在无限制条件下的傅里叶基拟合")
legend("topleft",c("均值"),lwd=3,col=2)
lines(c19pca$meanfd,lwd=3,col=2)
# 函数型主成分
harmfd = c19pca$harmonics
harmvals = eval.fd(1:366,harmfd)
dim(harmvals) # The top 4 FPCs
# 画出第二条主成分曲线
# par(mfrow=c(1,1),mar = c(8, 8, 4, 2))
```

```
# plot(1:366,harmvals[,1],xlab=' 天 ',ylab=' 第 一 函 数 型 主 成 分 ',lwd=4,lty=1,cex.lab=2,cex.axis=2,type='l')
# plot all 4 FPCs
matplot(1:366,harmvals,xlab='时间(天)',ylab='函数型主成分',
        lwd=2,lty=1,cex.lab=1.2,cex.axis=1.2,type='l')
legend('topleft',c(' 第 一 主 成 分 ',' 第 二 主 成 分 ',' 第 三 主 成 分 ',' 第 四 主 成 分 '),col=1:4,lty=1,lwd=2)
title('新增确诊的函数型主成分函数')
dev.off()
# plot the first FPC scores vs. the second FPC scores
quartz()
par(mfrow=c(1,2),mar = c(4, 4, 4, 2))
plot(1:366,harmvals[,1],xlab=' 天 ',ylab=' 第 一 函 数 型 主 成 分 ',lwd=4,lty=1,cex.lab=2,cex.axis=2,type='l')
plot(c19pca$scores[,1:2],xlab='第一主成分得分',ylab='第二主成分得分',col=4,
     cex.lab=1.5,cex.axis=1.5,cex=1)
text(c19pca$scores[,1],c19pca$scores[,2],labels=data$省级单位,cex=0.5)

jpeg("6.jpg",height=1500,width=2000,res=210)
par(mfrow=c(2,1),mar = c(4, 4, 4, 2))
plot(1:366,posc19fd[,2],xlab=' 时 间 ( 天 )',ylab=' 新 增 确 诊 ',cex.lab=1.2,cex.axis=1.2,col="black",lwd=1,lty=3,ylim=c(-5,70),type="l")
lines(1:366,posc19fd[,9],cex.lab=1.2,cex.axis=1.2,col="red",lwd=1,lty=4)
lines(1:366,posc19fd[,26],cex.lab=1.2,cex.axis=1.2,col="blue",lwd=1,lty=5)
lines(c19pca$meanfd,cex.lab=1.2,cex.axis=1.2,col="green",lwd=3,lty=6)
legend("topleft", c(" 北 京 市 ", " 河 北 省"," 天 津 市 "," 各 省 均 值 "),col = c("black","red","blue","green"),lty=c(3,4,5,6),lwd=1)
daily$place[c(17,24,35)]
# 移除均值
plot(c19fd$fd[2]-c19pca$meanfd,xlab=' 时 间 ( 天 )',ylab=' 中 心 化 的 新 增 确 诊 ',cex.lab=1.2,cex.axis=1.2,col="black",lwd=1,lty=3,ylim=c(-5,70),type="l")
lines(c19fd$fd[9]-c19pca$meanfd,cex.lab=1.2,cex.axis=1.2,col="red",lwd=1,lty=4)
lines(c19fd$fd[26]-c19pca$meanfd,cex.lab=1.2,cex.axis=1.2,col="blue",lwd=1,lty=5)
legend("topleft", c(" 北 京 市 ", " 河 北 省 "," 天 津 市 "),col =
```

```
c("black","red","blue"),lty=c(3,4,5),lwd=1)
dev.off()
################################################################################
#####4.函数型主成分的平滑方法
# 对函数型主成分的粗糙性进行惩罚
# 采用上文平滑系数为 10^{4}的傅里叶基函数拟合结果
# 用 the harmonic acceleration differential 算子定义粗糙惩罚,平滑系数为= 10^5
c19fdPar1 = fdPar(daybasis366,harmLfd,1e5)
pc19pca = pca.fd(c19fd$fd,nharm=4,harmfdPar=c19fdPar1)
# 得到主成分
pc19fd = pc19pca$harmonics
pharmvals = eval.fd(1:366,pc19fd)
jpeg("7.jpg",height=1500,width=2000,res=210)
par(mfrow=c(2,1),mar = c(4, 5, 4, 0.2))
matplot(1:366,harmvals,xlab='时间(天)',ylab='函数型主成分',
        lwd=2,lty=1,cex.lab=1.2,cex.axis=1.2,type='l',ylim=c(-0.25,0.35))
legend('topleft',c('第一主成分','第二主成分','第三主成分','第四主成分'),col=1:4,lty=1,lwd=2)
title('新增确诊的函数型主成分函数')
matplot(1:366,pharmvals,xlab='时间(天)',ylab='函数型主成分',
        lwd=2,lty=1,cex.lab=1.2,cex.axis=1.2,type='l',ylim=c(-0.25,0.35))
legend('topleft',c('第一主成分','第二主成分','第三主成分','第四主成分'),col=1:4,lty=1,lwd=2)
title('新增确诊的粗糙惩罚函数型主成分函数')
dev.off()
################################################################################
#####5.函数型数据主成分估计
# Now let's do a bit of reconstruction for 河北
# 去除河北省数据新增确诊人数的主成分分析
c19fdPar2 = fdPar(daybasis366,harmLfd,1e1)
raw_9ppca = pca.fd(c19fd$fd[-9],nharm=4,harmfdPar=c19fdPar2)
# 得到主成分的函数
harms = raw_9ppca$harmonics
# Get the mean curve
```

```
meanfd = raw_9ppca$meanfd
# 河北 data
hbdat =eval.fd(1:366,c19fd$fd[9])
plot(hbdat,type='l')
# evaluate FPC in the days [1:132]
c19_9pca_vals = eval.fd(1:355,harms)
# evaluate the mean curve in the days [1:132]
mc19_9vals = eval.fd(1:355,meanfd)
# 移除均值曲线
hbdat_train = hbdat[1:355]-mc19_9vals
# 得到曲线的主成分得分
coef = lm(hbdat_train~c19_9pca_vals-1)$coef
coef
#预测
Rfd = coef[1]*harms[1]+coef[2]*harms[2]+
   coef[3]*harms[3]+coef[4]*harms[4]+meanfd
Rvals = eval.fd(1:366,Rfd)
quartz()
jpeg("8.jpg",height=1500,width=2000,res=210)
par(mfrow=c(1,1),mar = c(4, 4, 4, 2))
plot(1:366,Rvals,type='l',lwd=2,col=5,ylim=c(-5,45),xlab='时间(天)',ylab="人数(人)",main='带粗糙惩罚的函数型主成分预测',cex.lab=1,cex.axis=1)
lines(c19fd$fd[9],col=4,lwd=3)
lines(356:366,Rvals[356:366],col=2,lwd=3)
lines(meanfd,lty=2,col=1,lwd=2)
legend("topleft", c("训练集的预测曲线", "测试集的预测曲线","观测值(1:355 为训练集 ,356:366 为测试集)"," 省份均值(除湖北省)"),col = c(5,2,4,1),lty=c(1,1,1,2),lwd=c(2,3,3,2))
dev.off()
#############################################################################
#####6.多元函数型数据主成分分析
# Multivariate FPCA
jpeg("9.jpg",height=1300,width=2000,res=210)
par(mfrow=c(2,1),mar = c(4, 5, 4, 0.2))
```

```
matplot(data$ 新 增 确 诊 ,type='l',cex.lab=1.2,cex.axis=1.2,ylab=' 人 数 ( 人 )',xlab=" 时 间 (天)",main="新增确诊")
matplot(data$ 新 增 死 亡 ,type='l',cex.lab=1.2,cex.axis=1.2,ylab=' 人 数 ( 人 )',xlab=" 时 间 (天)",main="新增死亡")
dev.off()
# Setting up a general object
#  定义 harmonic acceleration differential 算子
harmaccelLfd <- vec2Lfd(c(0, (2*pi)^2, 0),c(0, 366))
basis21 <- create.fourier.basis(c(0,366),21)
# Smooth the data
str(data1)
data1=array(0,dim=c(366,30,2))
data1[,,1]=data$新增确诊
data1[,,2]=data$新增死亡
rownames(data1)=1:366
#命名
colnames(data1)=data$省级单位
column.names=data$省级单位
row.names=1:366
matrix.names=c("新增确诊","新增死亡")

# Take these vectors as input to the array.
?smooth.basisPar
data1=array(data1,dim=c(366,30,2),dimnames                                                    =
list(row.names,column.names,matrix.names))
mulfd <- smooth.basisPar(1:366,data1,basis21, Lfdobj=harmaccelLfd, lambda=1e1)$fd
rownames(mulfd$coefs)=1:366
names(mulfd$fdnames) = c("时间", "省份", "新增")
mulfd$fdnames[[3]] = c("确诊", "死亡")
#  画出平滑曲线
mulvals = eval.fd(1:366,mulfd)
jpeg("10.jpg",height=1200,width=2000,res=210)
par(mfrow=c(1,2),mar = c(4, 4.5, 3, 0.1))
matplot(data$新增确诊,data$新增死亡,type='l',cex.lab=1.2,cex.axis=1.2,xlab='新增确诊
```

```
(人)',ylab='新增死亡(人)',main="新增确诊与新增死亡的相位图")
legend("topleft",c("Harmonic Acceleration 算子;平滑系数为 0"))
matplot(37.5*mulvals[,,1],37.5*mulvals[,,2],type='l',cex.lab=1.2,cex.axis=1.2,xlab='新增确诊(人)',ylab='新增死亡(人)',main="粗糙惩罚后新增确诊与新增死亡的相位图")
legend("topleft",c("Harmonic Acceleration 算子;平滑系数为 10^1"))
dev.off()
mul.cor = cor.fd(1:366,mulfd)
library("fields")

jpeg("11.jpg",height=1200,width=2000,res=210)
par(mfrow=c(1,3),mar = c(4, 5, 4, 5))
#    contour(tfine,tfine,gait.cor[,,1,3],xlab='day',ylab='day',cex.lab=1.5,cex.axis=1.5)
image.plot(1:366,1:366,mul.cor[,,1,1],mul.cor,
                xlab='新增确诊(人)',ylab='新增确诊(人)',cex.lab=1.2,cex.axis=1.2,main="新增确诊的自相关关系")
image.plot(1:366,1:366,mul.cor[,,1,2],mul.cor,
                xlab='新增确诊(人)',ylab='新增死亡(人)',cex.lab=1.2,cex.axis=1.2,main="新增确诊与新增死亡的相关关系")
image.plot(1:366,1:366,mul.cor[,,1,3],mul.cor,
                xlab='新增死亡(人)',ylab='新增死亡(人)',cex.lab=1.2,cex.axis=1.2,main="新增死亡的自相关关系")
dev.off()
# Now a principle components analysis
mul.pca = pca.fd(mulfd,nharm=4)
data$新增死亡
quartz()
par(mfrow=c(2,1),mar = c(8, 8, 4, 2))
plot(mul.pca$meanfd,cex.lab=1.5,cex.axis=1.5)
plot(mul.pca$meanfd,cex.lab=1.5,cex.axis=1.5,ylim=c(0,0.03))

# Mean cycle
meanvals = eval.fd(1:366,mul.pca$meanfd)
jpeg("12.jpg",height=1300,width=2000,res=210)
par(mfrow=c(1,2),mar = c(4, 5, 3, 0.2))
```

```
plot(meanvals[,,1],meanvals[,,2],xlab='新增确诊(人)',ylab="新增死亡(人)",main="傅里叶基拟合后均值的相位图",
     cex.lab=1.2,cex.axis=1.2,type='l',lwd=2,col=4)
# 多维主成分分析后方差累计被解释份额
plot(cumsum(mul.pca$values[1:10])/sum(mul.pca$values),xlab='主成分',
     ylab='份额',col=2,cex.lab=1.2,
     cex.axis=1.2,cex=1.2,type='l',main="多维主成分分析后方差累计被解释份额")
points(cumsum(mul.pca$values[1:10])/sum(mul.pca$values))
abline(h=0.98)
abline(h=0.99)
dev.off()
# 画出 FPCs
jpeg("13.jpg",height=1300,width=2000,res=210)
par(mfrow=c(1,2),mar = c(4, 4, 3, 0.2))
plot(mul.pca$harmonics[,1,],lty=c(2,2,2,2),lwd=2,cex.lab=1.2,cex.axis=1.2,ylim=c(-0.1,0.13),xlab='时间(天)',ylab='函数型主成分',type='l')
legend('topleft',c('第一主成分','第二主成分','第三主成分','第四主成分'),col=1:4,lty=4,lwd=2)
title('新增确诊的函数型主成分函数')
plot(mul.pca$harmonics[,2,],lty=c(2,2,2,2),lwd=2,cex.lab=1.2,cex.axis=1.2,ylim=c(-0.0025,0.0025),xlab='时间(天)',ylab='函数型主成分',type="l")
legend('topleft',c('第一主成分','第二主成分','第三主成分','第四主成分'),col=1:4,lty=4,lwd=2)
title('新增死亡的函数型主成分函数')
dev.off()
#############################################################################
#######7.函数型数据对函数型数据的回归分析
Beta0Par = fdPar(basis21, 2, 1e-5)
Beta1fd  = bifd(matrix(0,21,21), basis21, basis21)
Beta1Par = bifdPar(Beta1fd, 2, 2, 1e3, 1e3)
BetaList = list(Beta0Par, Beta1Par)
#  定义自变量和因变量
harmLfd = vec2Lfd(c(0,(2*pi/366)^2,0), c(0,366))
c19fdPar = fdPar(basis21,harmLfd,1e-14)
```

```
infd = smooth.basis(1:366,data$新增确诊,c19fdPar)

#jpeg("14.jpg",height=1300,width=2000,res=210)
par(mfrow=c(2,1),mar = c(4, 4, 3, 0.2))
plot(infd,ylab='确诊人数(人)',xlab="时间(天)",cex.lab=1.2,cex.axis=1.2)
title("粗糙惩罚后的新增确诊人数")
legend("topleft",c("Harmonic Acceleration 算子;平滑系数为 10^4"))
deathfd = smooth.basis(1:366,data$新增死亡,c19fdPar)
plot(deathfd,ylab=' 死 亡 人 数 ( 人 )',xlab=" 时 间 ( 天 )",main=" 新 增 死 亡 ",cex.lab=1.2,cex.axis=1.2)
title("粗糙惩罚后的新增死亡人数")
legend("topleft",c("Harmonic Acceleration 算子;平滑系数为 10^4"))
dev.off()
#   函数型对函数型回归
mod = linmod(deathfd$fd, infd$fd, BetaList)
beta1mat = eval.bifd(1:366, 1:366, mod$beta1estbifd)
quartz()
persp(1:366, 1:366, beta1mat,axes=T,
      xlab="新增确诊(天)", ylab="新增死亡(天)",zlab="参数估计",
      cex.lab=1.5,cex.axis=1.5,theta=60,phi=30,col=3)
```

致谢

本论文在常文千老师、李雪老师的悉心指导下完成，从文章主题选取到具体的写作过程，论文初稿的修改直到最后定稿，常老师和李老师均热情而又及时的给予了高度专业的指点，在此向常文千老师、李雪老师表示深深的感谢和崇高的敬意！同时，在论文写作过程中，我们还参考了统计学、医学相关的书籍、论文，一并向王国长、闫星宇等所有有关的作者表示谢意。这里也要特别感谢西蒙菲沙大学的曹际国老师，曹老师在 Github 网站分享的关于函数型数据分析的 R 语言程序给了我们很大的启发。

在文章的写作期间，我们的身边始终围绕着家人、老师、朋友，他们的关爱和无私帮忙使得我们度过艰难，顺利地比赛要求的所有任务。回首这近 3 个月的论文写作过程，往事犹如昨日，他们的关心和帮忙是我们不能忘却的完美回忆。在此，向写作期间所有帮助过我们的人表示最衷心地感谢。

最后，我们要向百忙之中抽时间对本文进行审阅、评议的各位专家老师表示诚挚谢意。