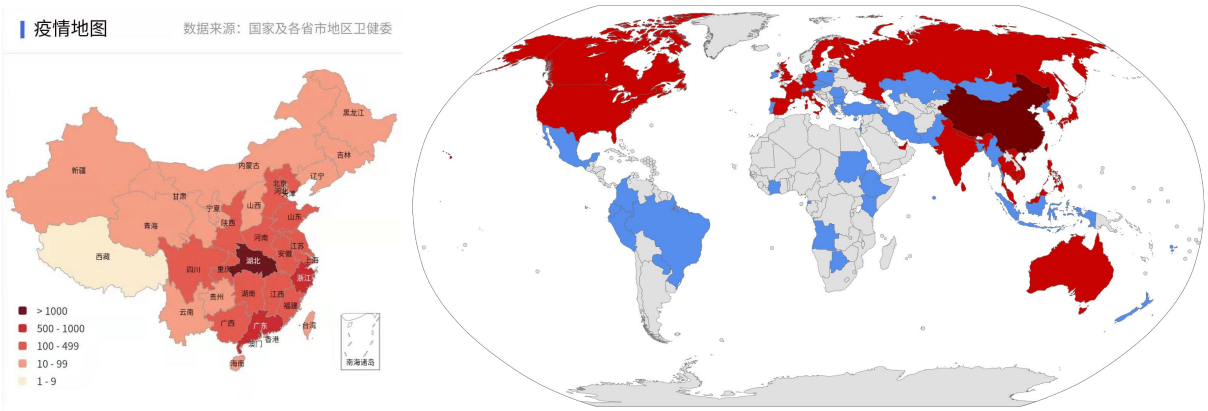
新型冠状病毒传染模型和预测分析

YanqiaoWu@FDU Math

引言:

2020年的春节在一场大疫情中度过。从2019年12月8日发现的第一起不明原因肺炎开始，这种类似SARS、被命名为2019-nCov的新型冠状病毒从疫源地武汉蔓延全国，并且海外多国也报告了确诊病例。截至2020年2月2日15时，全世界共有14568个确诊病例，其中中国有14423例确诊（含港澳台共32例）。中国大陆累计报告死亡病例304例。
全球及中国大陆的疫情情况如图:



新型冠状病毒疫情特点:

我们注意到，根据官方消息和已有的研究报告，这次疫情传播存在以下特点:
1. 病毒的潜伏期一般为3—7天，最长不超过14天。
2. 与非典等病毒不同，在潜伏期内有传染性。根据目前观察到的大量聚集性病例，病毒在潜伏期内的传染性很强。
3. 病毒爆发的时间点特殊，赶上春运时间。借助便捷的交通运输系统，疫情向疫源地以外蔓延的趋势明显。
4. 病毒的传染能力很强。根据目前各研究团队的计算，新型冠状病毒的基本传染系数R_0在2到5之间，传播能力比非典，埃博拉，H7N9强。
5. 病毒的检测困难，一是在症状上，发病时可能没有发热症状，二是当前的核酸测序检测不一定能够检查出来。
6. 感染病毒的发病症状不严重，大多数患者为轻症，病毒致死率低。
7. 中国政府采取了大力管控，湖北省多地实行封城，全国多省多地启动一级响应，境外，海外多国多地加强检查，这有利于遏制疫情的蔓延。

基本传染病学模型分析

考虑到疫情的以上特点，我们首先分析主流的传染病学模型在模拟这次疫情传播上的性能:

SIR模型:

SIR 模型中涉及三种人群：易感人群S（Susceptible），感染人群I（Infective）和移除人群R（Removed）
易感人群指未得病者，但缺乏免疫能力，与感染者接触后容易受到感染。感染人群指染上传染病的人。移除人群是因病愈（具有免疫力）或死亡而不再参与感染和被感染过程的人。

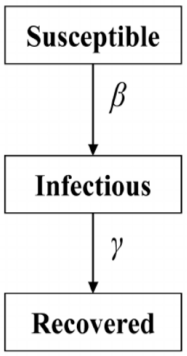
SIR模型的假设是:

1. 不考虑人口的出生、死亡、流动等种群动力因素。人口始终保持一个常数，即N(t)=K
2. 一个病人一旦与易感者接触就必然具有一定的传染力。假设 t 时刻单位时间内，一个病人能传染的易感者数目与此环

境内易感者总数S(t)成正比，比例系数为β，从而在t时刻单位时间内被所有病人传染的人数为β·S(t)·I(t)

3. t时刻，单位时间内从染病者中移出的人数与病人数量成正比，比例系数为γ，单位时间内移出者的数量为γi(t)。

模型过程和模型的微分方程表达式如下：



$$\frac{dS}{dt} = -\beta SI$$

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

SIR模型并不适用于刻画新型冠状病毒的传播，原因有以下几点：

1. 疫情的一个重要特点是有大量潜伏期的患者，且潜伏期内的患者传染能力较强。在SIR模型中，感染患者I既包括了潜伏期中的患者，也包括了已经被隔离的患者。这两类人群的传播能力（反应在β系数上）是不同的，但SIR并没有考虑到这一点。

2. SIR认为总人数N（N=S+I+R）是固定的。这个假设在较为封闭的区域或人口流动较小时成立，但显然不适用于春运这样大人口流动环境下的疫情传播模拟。

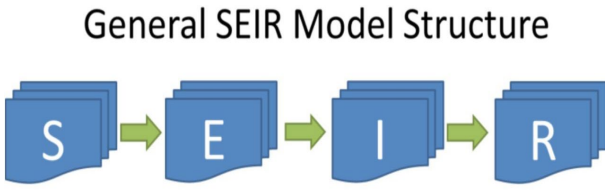
3. SIR无法反应对疫情的管控。针对疫情，政府出台各种防疫，治疫措施，大量市民提高了防范意识，对疫情的传播有巨大影响。

SEIR模型

SEIR模型是在SIR模型上基础上对潜伏人群进行了刻画。

SEIR将人群分为四类：易感人群S（Susceptible），潜伏人群E（Enfective），隔离（确诊）人群I（Infective）和移除人群R（Removed）

SEIR模型的微分表达式如下：



$$\frac{dS}{dt} = -\frac{\beta SI}{N}$$

$$\frac{dE}{dt} = \frac{\beta SI}{N} - \sigma E$$

$$\frac{dI}{dt} = \sigma E - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

3
SEIR模型虽然增加了潜伏期感染者，但是仍然不适用于新型冠状病毒疫情的建模，原因是：
1. 从微分方程表达式中，SEIR模型认为处在潜伏期的患者没有传染能力。这适用于SARS等在潜伏期内无传染力的病毒,但不适用于新型冠状病毒.
2. 仍然不能解决春运和管控防疫措施带来的变化.

模型建立

基于SEIR模型和新型冠状病毒的特点,我们建立如下的模型.

模型假设

1. 假设强力管控措施(封城,个人防护,隔离)和春运高峰期(节前3-4天)重合,即综合考虑春运和防疫管控的作用.可以依次将疫情分成两个阶段,阶段一是春运加防疫之前,阶段二是春运加防疫之后.在每个阶段内,假设总的人数N恒定.
2. 假设 t 时刻单位时间内，一个病人能传染的易感者数目与此环境内易感者总数S(t)成正比，比例系数为β，从而在t时刻单位时间内被所有病人传染的人数为β·S(t)·I(t)
3. 考虑潜伏期人群的病毒传播作用,设作用系数为σ.且由于系数β于每个人能够接触到的人数成正比,假设σ也与每个人能够接触到的人数成正比.因为用接触人数可以刻画σ和β的值,认为在每个阶段内σ和β的比值是定值.

数据和符号说明
数据:

数据	数据内容	数据来源	备注说明
全国每日疫情数据	确诊,疑似, 当日医学观察,累计医学观察人数	国家卫健委	早期数据缺失
湖北省每日疫情数据	确诊,疑似, 当日医学观察,累计医学观察人数	湖北省卫健委, 人民日报等	无疑似病例数据,部分数据缺失
广东省每日疫情数据	确诊,疑似, 当日医学观察,累计医学观察人数	广东省卫健委, 人民日报等	无疑似病例数据, 无累计医学观察数据,部分数据缺失
重庆市每日疫情数据	确诊,疑似, 当日医学观察,累计医学观察人数	重庆市卫健委, 人民日报等	无疑似病例数据,部分数据缺失
陕西省每日疫情数据	确诊,疑似, 当日医学观察,累计医学观察人数	陕西省卫健委, 人民日报等	无累计医学观察数据, 部分数据缺失
甘肃省每日疫情数据	确诊,疑似, 当日医学观察,累计医学观察人数	甘肃省卫健委, 人民日报等	无疑似病例数据,部分数据缺失

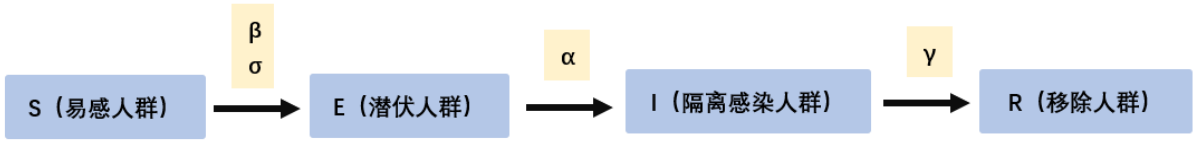
模型的建立
根据模型假设,建立以下模型:

1. 以SEIR为主要模型框架.
2. 分成两个阶段,第一个阶段是2020-01-25之前,第二个阶段为2020-01-25及之后,两个阶段内的模型使用SEIR框架.
3. 引入潜伏期人群的传染系数σ.σ正比于密切接触者,且σ=r·β,r为确定值.
4. 第一个阶段使用从2020-01-15开始的数据进行拟合,将第一个阶段预测出来的部分数据用于第二个阶段模型的建立.

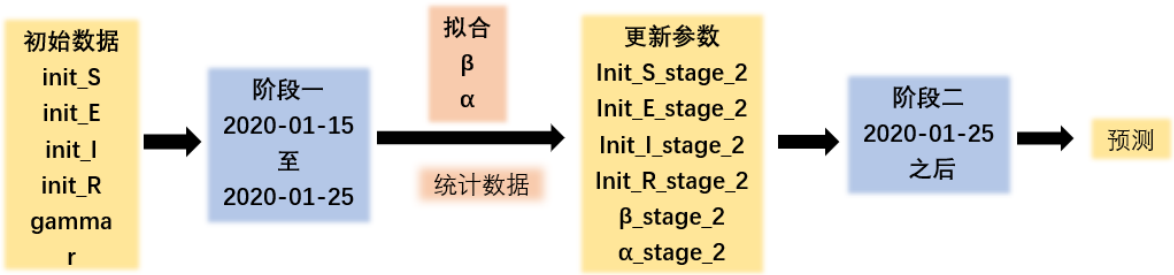
模型示意图:

- 单阶段模型(SEIR_modified)示意图:

4



- 双阶段模型示意图：



模型的微分方程表达式:

$$\frac{dS}{dt} = -\beta \cdot \frac{I \cdot S}{N} - \sigma \cdot \frac{I \cdot E}{N}$$
$$\frac{dE}{dt} = \beta \cdot \frac{I \cdot S}{N} + \sigma \cdot \frac{I \cdot E}{N} - \alpha \cdot E$$
$$\frac{dI}{dt} = \alpha \cdot E - \gamma \cdot I$$
$$\frac{dR}{dt} = \gamma \cdot I$$
$$\sigma = r \cdot \beta$$

参数的确定与估计方法

- 阶段1参数估计:

参数	参数意义	取值	估计方法	优化方法	取值范围	说明
init_S_stage_1	初始易感人群	100000	人工估计	-	50000-1000000	参考现有研究, 结合密切接触者估计
init_E_stage_1	初始潜伏人群	300	人工估计	-	>50	根据ICL研究报告估计
init_I_stage_1	初始感染人群	44	官方数据	-	-	-
init_R_stage_1	初始恢复人群	2	官方数据	-	-	-
beta_stage_1	发病人群传染系数	待估计	拟合估计	hyperopt算法	(0,0.4)	根据R0和密切接触者估计范围
r_stage_1	sigma beta比值	5	计算估计	-	-	计算方法见后
alpha_stage_1	潜伏人群被隔离速率	待估计	拟合估计	hyperopt算法	$(\frac{1}{14}, \frac{1}{3})$	取值范围计算见后
gamma_stage_1	治愈速率	0.0286	官方数据计算	-	-	-

- 阶段2参数估计:

5

参数	参数意义	取值	估计方法	优化方法	取值范围	说明
init_S_stage_2	初始易感人群	100000+ΔN	人工估计	-	50000-1000000	ΔN为新增易感人群,计算方法见后
init_E_stage_2	初始潜伏人群	取stage_1预测数据	根据stage_1计算	-	-	-
init_I_stage_2	初始感染人群	1975	官方数据	-	-	-
init_R_stage_2	初始恢复人群	49	官方数据	-	-	-
beta_stage_2	发病患者传染系数	$\frac{beta\_stage\_1}{k\_change\_ratio}$	根据stage_1结果确定	-	(0,0.4)	计算见后
r_stage_1	sigma beta比值	5/k_change_ratio	计算估计	-	-	计算方法见后
alpha_stage_1	潜伏-隔离速率	根据stage_1计算	-	-	$(\frac{1}{14},\frac{1}{3})$	取值范围算法见后
gamma_stage_1	治愈速率	0.0286	官方数据计算	-	-	-

参数计算方法:

k_change_ratio的计算

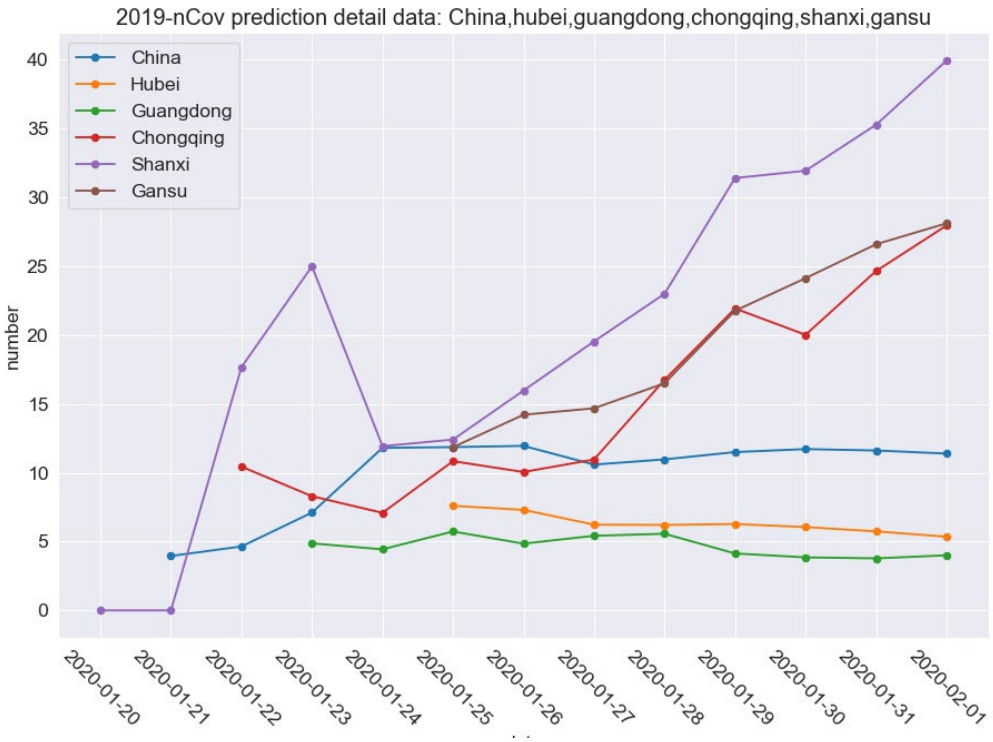
- k值:计一个人的密切接触者人数为k,beta与k成正比,sigma与k成正比.且:

$$\frac{\text{sigma}}{\text{beta}} = \text{k}$$

- 计算k,即需要估计每个人的密切接触者人数.
  卫健委公布了每日医学观察者的人数,由于医学观察者是确诊患者追踪道德密切接触者,所以可以用医学观察者人数和总确诊人数来估计密切接触者人数.
- 运用分层抽样的方法计算:
  1. 将所有省份按确诊人数进行新型冠状病毒感染情况等级的划分,一级为最严重,六级为最轻微.
     结果划分如下图:

2020.2.2数据统计												
update at 2020.2.2 15:00												确诊人数总和
一级区域:	湖北	疫源										9074
二级区域	浙江	广东	湖南	河南	>400							2221
三级区域	安徽	江西	重庆	江苏	四川	山东	200-400					1627
四级区域	北京	上海	福建	陕西	广西	河北	云南	黑龙江	90-200			1049
五级区域	辽宁	海南	山西	天津	甘肃	贵州	宁夏	内蒙古	吉林	新疆	20-90	408
六级区域	青海	西藏	第六级人数太少，不抽样									12
从上到下、从左至右依次减少												

我们按对角线抽取每一个分层的中位数省份,由于六级区域确诊人数太少,不予抽取.
抽取的省份是: 湖北,广东,重庆,陕西,甘肃
根据各省的数据,可以计算出每日累计医学观察者和累计确诊的比例,结果如图:



分析:根据疫情4-7天的潜伏期（丁香医生），24日以前完成的春运(20日至24日集中)接触效果在2020-01-25 -- 2020-02-01显现，从图像上看，多地确实显现出了春运带来的接触者影响，k值有所上升。
值得注意的时，2020-01-24及以后，全国的数据k值平稳，可以作为平均接触人数的参考

取全国的平均值可得11.491363226848797,计k=11.5 根据春运和封城后一个短周期(4天,从2020-01-28开始)由分层抽样的加权平均值可得9.22061813550857,计k_new=9.22
估算r值:对于发病被迅速隔离的人,其密切接触者会立即减小,设他能接触的只有两到三个人(可解释为亲属和义务人员),则r可取为5.

ΔN的估计
由于春运,易感人群会明显增加
基于一个简单的估计：
根据信息：武汉市人口1100W，封城前有500W人离开，截止2月1日24时武汉市有4109人感染，设两种人群（离开和非离开）的比例相同，则500W人中 有 4109·500/600 人感染，不妨设湖北流出感染者是其两倍,共约为7000人
根据平均k值，一个人密切接触者11.5人，则易感人群增加：7000×11.5约8W人
ΔN=80000

模型的拟合

拟合方法:

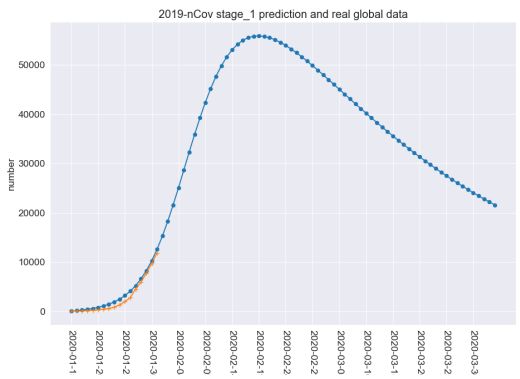
- hyperopt进行拟合,设置测试次数max_eval为1000次
- 数据点采样: 为了增加对比的数据点,加强模型的稳健性,假设每日日内的增长是线性的,每日均匀采样增加4个数据点.
- 损失函数: MSE

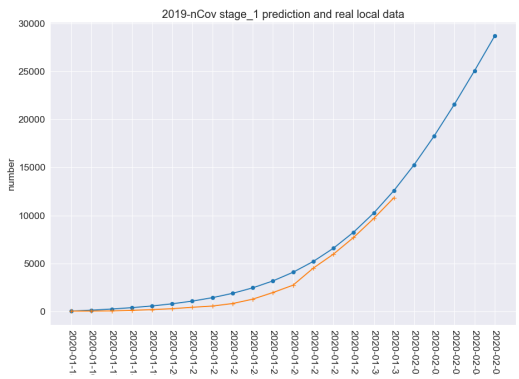
拟合结果

1. 参数结果:

参数	数值	参数	数值
alpha_stage_1	0.15	init_S_stage_1	100000
beta_stage_1	0.08	init_I_stage_1	44
gamma_stage_1	0.03	init_E_stage_1	300
r_stage_1	5	init_R_stage_1	2

1. 拟合结果(预测了自1月15日起80天)





模型的预测

参数更新

根据模型计算更新后的系数:

根据模型计算更新后的系数: $beta\_stage\_2 = \frac{beta\_stage\_1}{k\_change\_ratio} = 0.63$

$alpha\_stage\_2 = alpha\_stage\_1 \cdot k\_change\_ratio = 0.195$

$r\_stage\_2 = \frac{r\_stage\_1}{k\_change\_ratio} = 3.84$

根据stage_1的预测结果

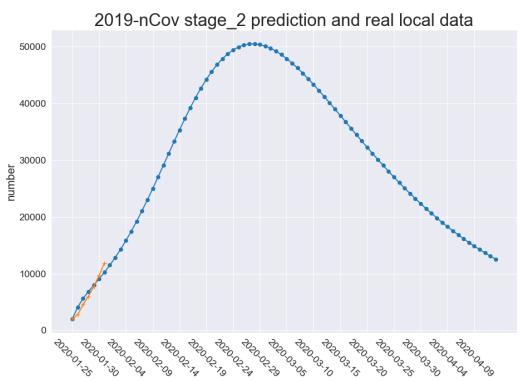
init_S=93836

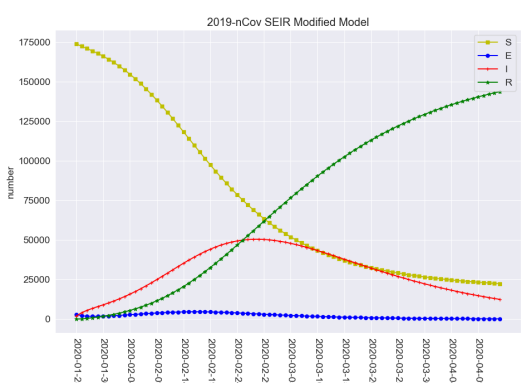
init_S_stage_2=93836+80000

init_E_stage_2=init_E=3003

参数	数值	参数	数值
alpha_stage_2	0.195	init_S_stage_2	173836
beta_stage_2	0.063	init_I_stage_2	1975
gamma_stage_2	0.03	init_E_stage_2	3003
r_stage_2	3.84	init_R_stage_2	49

预测结果





预测数据显示:
峰值为50492人,到达峰值的日期为2020-02-28.根据不同起点的数据进行模型训练,峰值大都稳定在2020年2月27日至2020年3月10日之间.

模型的评估:

- 模型的优点

  1. 较之SIR模型,我的模型区分了潜伏期患者和已经发病被隔离的患者,更加精确.
  2. 较之SEIR模型，我的模型考虑了新型冠状病毒在潜伏期的传染性.
  3. 我尝试使用有限的数据刻画了防疫管控措施和春运带来的影响.
  4. 在一定范围内,模型稳健性较好.
- 模型的缺点:
  1.运用SEIR的框架，始终无法很好地刻画动态问题，对长期目标的预测缺乏灵活性.
  2.在对参数敏感性的测试中,部分参数过于敏感.比如治愈率γ极大地影响了到达峰值的日期.

结论

预期新型冠状病毒感染峰值出现在2月底到3月上旬,峰值大约在4到5万人.随着进一步的管控和节后返工潮的到来，疫情的发展仍有诸多不确定性。