清華大學
Tsinghua University

曲线拟合应用

自动化系 赵虹

曲线拟合定义

- 已知数据点
  - $(x_i, y_i), i = 0, \dots, m$
- 线性无关函数组
  - $\Phi = span\{\varphi_1(x), \varphi_2(x), \dots, \varphi_n(x)\}$
- 拟合函数
  - $S^*(x) = \sum_j a_j^* \varphi_j(x)$
- 目标函数
  - $\min\|\delta\|_2^2 = \sum_{i=0}^m [S^*(x_i) - y_i]^2$

应用实例1：原子弹爆炸能量估计

- The Trinity Test (July 16, 1945)
  - Alamogordo Test Range, Jornada del Muerto desert, New Mexico

http://nuclearweaponarchive.org/Usa/Tests/Trinity.html

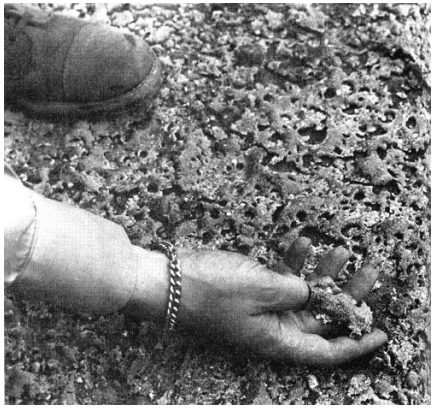
https://thatsmaths.com/2014/09/18/how-big-was-the-bomb/

The Trinity Test

http://nuclearweaponarchive.org/Usa/Tests/Trinty1.mpg

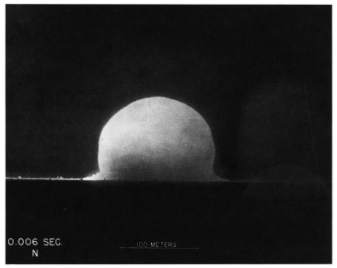
The Trinity Test

Oppenheimer and Groves inspecting the remains of the Trinity test tower, 9 September 1945.

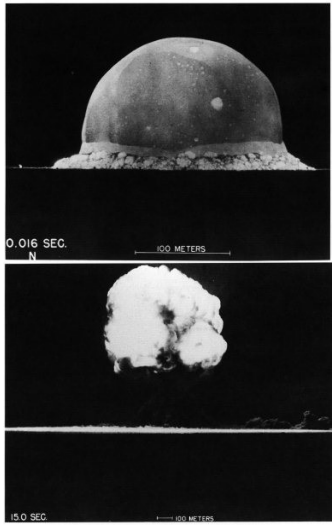


The heat of the Trinity explosion melted the sandy soil around the tower to form a glassy crust known as "trinitite".

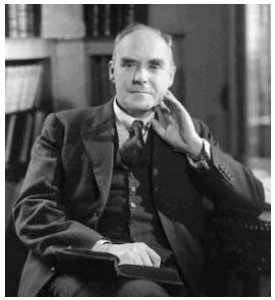
How Big Was The Blast?



In 1947, photographs of the Trinity Test were released by the US Army. They aroused great interest and appeared in newspapers and magazines all over the world.



The Man Who "Decoded" The Secret



Geoffrey Ingram Taylor (1886-1975) was one of the great scientists of the twentieth century. He was a meteorologist in his early career and became a major figure in fluid dynamics, his ideas having wide application in meteorology, oceanography, engineering and hydraulics.

Taylor was asked by the British government to study mechanical ways of measuring the bomb's yield (energy output).

Taylor was not directly involved in the bomb's development, and for security reasons worked independent of the US Manhattan project.

Dimension Analysis

- The energy would be released from a small volume, and would produce a very strong shock wave that would expand in approximately a spherical shape.
- Taylor recognized that the parameters in the problem are the energy E, the density ρ of air, the radius R(t) of the blast wave, and the time t since the blast.

$$R(t) = \left(\frac{t^2 E}{\rho}\right)^{\frac{1}{5}}$$

The Blast

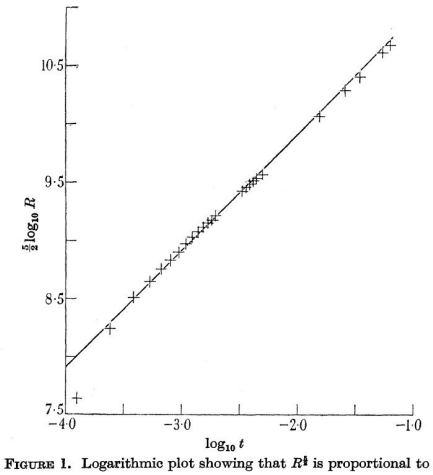


FIGURE 1. Logarithmic plot showing that $R^{\frac{5}{2}}$ is proportional to t.

Taylor's value for the energy was 17 kiloton TNT equivalent, close to the value later announced by President Truman. Taylor's result caused quite a stir and, according to his biographer, Batchelor (1996), Taylor was "mildly admonished by the US Army for publishing his deductions". But it was a mathematical tour-de-force. Considering the many sources of error, the estimate was, in his own words, "better than the nature of the measurements permits one to expect".

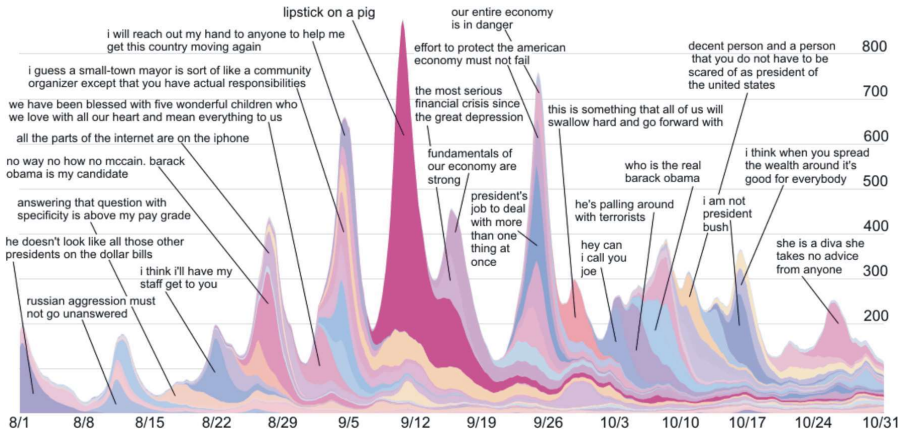
https://thatsmaths.com/2014/09/18/how-big-was-the-bomb/

应用实例2：信息传播

- 定义
  - 信息传播是个人、组织和团体通过符号和媒介交流信息，向其他个人或团体传递信息、观念、态度或情意，以期发生相应变化的活动[1]。
- 表示方法
  - 微观：个人接受并转发某一条信息的概率大小
  - 宏观：某一条信息浏览量（或点击量、搜索量、转发量等指标）随时间变化的函数

[1]百度百科词条：信息传播

Memetracker[2]



Top 50 quotes in the news and blogs over time, during the U.S. presidential election 2008.

[2] https://snap.stanford.edu/memetracker/

百度搜索指数

- 关键词“中美贸易战”在2019年5月1日—2019年7月8日的搜索指数随时间变化.
- 百度搜索指数：反应互联网用户对关键词的搜索及关注程度.[3]

[3]http://index.baidu.com/

信息传播模型

- 微观模型：预测个人转发信息的概率
  - 线性阈值(LT)模型
  - 独立级联(IC)模型
- 宏观模型：拟合信息转发量等指标的变化
  - 线性影响(LIM)模型
  - 传染病(SIR)模型

传染病(SIR)模型

- 传染病模型提出于上世纪20年代(Kermack W. O. and McKendrick A. G., 1927)
- 目的：用来刻画传染病的传播情况从而能够预测和控制其传播过程
- 将人群分为三大类
  - S（Susceptible 嫌疑者）：可能被传染的群体；
  - I（Infectious 传染者）：具有传染性的群体；
  - R（Recovered 恢复者）：恢复健康的群体，具有免疫。

SIR模型在信息传播领域的引入

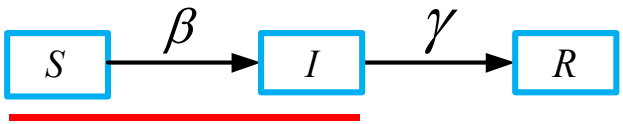
- 上世纪60年代被引入信息传播领域(Daley D. J. and Kendall D. G., 1964)
- 方法：类比传染病传染和信息传播过程的相似性

对象 \ 情景	传染病传染过程	信息传播过程
传播物	传染病	信息，谣言等
S状态	未感染过疾病	未知信息
I状态	染病具有传染性	已知并转发信息
R状态	恢复具有免疫	对信息失去兴趣

SIR模型

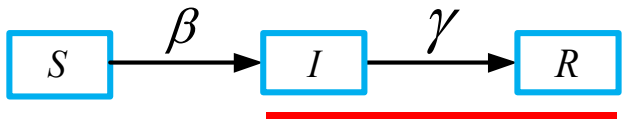
- 状态变量
  - $S(t)$，$I(t)$，$R(t)$: t 时刻处于S，I，R状态群体比例
- 前提假设
  - 在模型中所有人是一致的；
  - 所有用户互联（网络拓扑是完全网络）；
  - 网络的总人口始终不变（无生灭现象）

SIR模型



- 传播过程
  - 在Δt时间内每个S状态用户有$I(t)$概率与一个传播者接触
  - β: 上述接触中成功“传染”的概率
  - S状态比例的变化$\Delta S(t) = -\beta S(t)I(t)$

SIR模型



- 恢复过程
  - 在Δt时间内每个I状态用户有γ概率恢复
  - 恢复的用户状态将从I变化为R
  - I状态比例的变化$\Delta I(t) = \beta S(t)I(t) - \gamma I(t)$
  - R状态比例的变化$\Delta R(t) = \gamma I(t)$

SIR模型—动力学方程

- 令$\Delta t \to 0$ 可以得到下述动力学方程

$$\frac{dS(t)}{dt} = -\beta \cdot S(t) \cdot I(t)$$

$$\frac{dI(t)}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t)$$

$$\frac{dR(t)}{dt} = \gamma \cdot I(t)$$

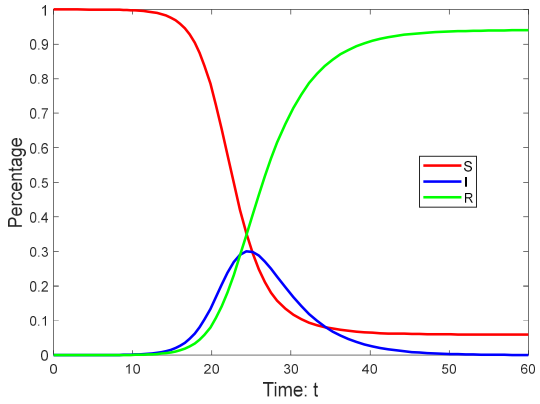
$$0 < S(t), I(t), R(t) < 1$$

$$0 < \beta, \gamma < 1$$

$$S(t) + I(t) + R(t) = 1$$

SIR模型—举例

- $\beta = 0.75,\ \gamma = 0.25$, 初态$[S_0, I_0, R_0] \approx [1,0,0]$



SIR模型拟合

- 实际Twitter观测数据（对于某一条具体信息）
  - 传播量序列 $X(t), (t = 1, \dots, m)$
  - 归一化 $x(t) = X(t)/N$, 其中的N为总人数
- 模型中与实际数据的对应：$I(t)$
- 拟合参数：β, γ
- 优化目标：

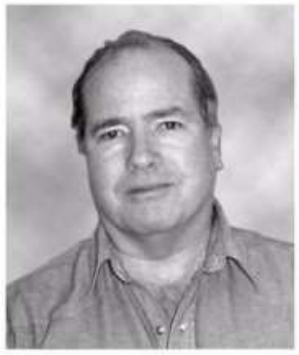
$$\min_{\beta,\gamma} e(\beta,\gamma) = \sum_{i=1}^{m} \|I(t) - x(t)\|_2^2$$

SIR模型拟合（续）

- 难点：模型无法给出闭式表达
- 求解优化目标的方案
  - 格点穷举法
    - 将参数β, γ在其取值范围之内按一定步长枚举
    - 复杂度高，精度受限
  - 智能优化方法
    - 模拟退火方法(SA)
    - 遗传算法(GA)
    - 粒子群算法(PSO)
    - …

粒子群(PSO)优化算法

- 模拟鸟群捕食行为设计的一种群智算法(1995)



James Kennedy　　Russell Eberhart

粒子群(PSO)优化算法

- 鸟群的任务是找到最大的食物源（全局最优解）
  - 在整个搜寻的过程中，通过相互传递各自位置的信息，让其他的鸟知道食物源的位置。
  - 最终，整个鸟群都能聚集在食物源周围（即找到了最优解）

粒子群(PSO)算法的抽象描述

- 假设算法中有M个粒子，算法迭代n次
- 优化变量空间构成搜索空间（假设维度为k）
- 每个粒子i有三个参量
  - 当前在搜索空间中的位置: z_i；
  - 当前的移动速度: v_i；
  - 在当前位置上的适应度(fitness): $e_i(z_i)$ →(目标函数值)
- 最优位置记录
  - $pbest_i$：第i个粒子历史适应度最好的位置
  - $gbest$：所有粒子历史最好适应度对应位置

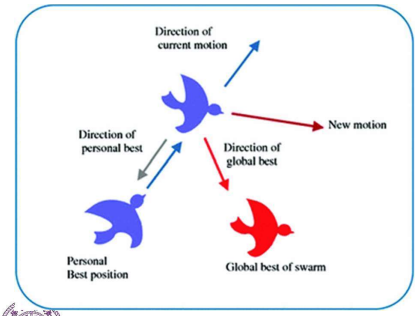
粒子群(PSO)算法的抽象描述（续）

- 在每轮迭代中，对每个粒子
  - 根据自身历史最优和全局历史更新最优更新速度

$$v_i^j \leftarrow v_i^j + 0.2 \times rnd \times \left(pbest_i^j - z_i^j\right) + 0.2 \times rnd \times \left(gbest^j - z_i^j\right)$$

Cognitive component

Social component



https://media.springernature.com/lw785/springer-static/image/art%3A10.1007%2Fs10462-013-9400-4/MediaObjects/10462_2013_9400_Fig2_HTML.gif

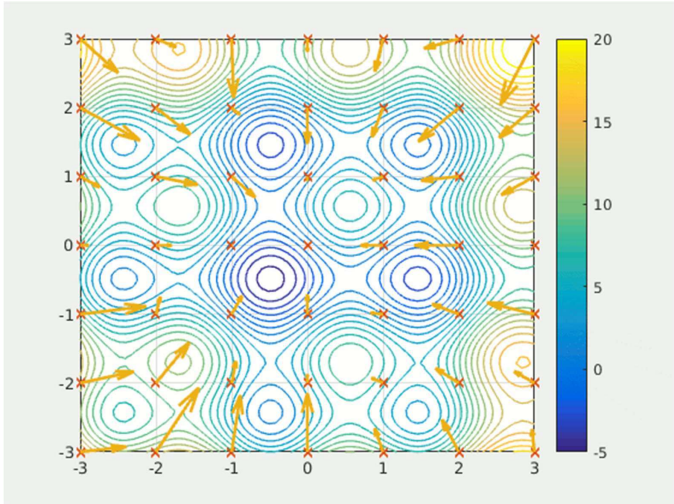
粒子群(PSO)算法的抽象描述（续）

- 在每轮迭代中，对每个粒子
  - 根据自身历史最优和全局历史更新最优更新速度

$$v_i^j \leftarrow v_i^j + 0.2 \times rnd \times \left(pbest_i^j - z_i^j\right) + 0.2 \times rnd \times \left(gbest^j - z_i^j\right)$$

  - 速度更新要每个维度分别进行
  - 每个维度的速度更新有一定的随机性
    (两个rnd是[0,1]间的随机数)
  - 利用更新的速度来更新位置: $z_i \leftarrow z_i + v_i$
  - 更新适应值$e_i(z_i), pest_i, gbest$
- 算法结束时返回当前$gbest$

粒子群算法示例

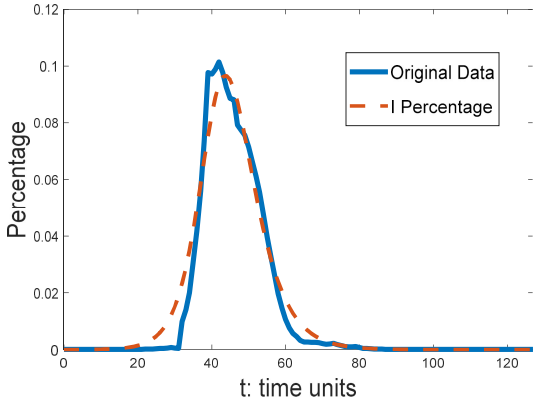


https://upload.wikimedia.org/wikipedia/commons/e/ec/ParticleSwarmArrowsAnimation.gif

用PSO算法拟合SIR模型

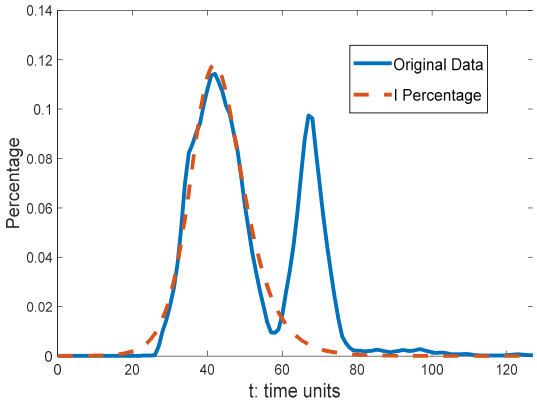
- 搜索空间$z = [\beta, \gamma]^T \in [0,1] \times [0,1]$
- 适应值fitness的计算
  - 根据$z_i = [\beta_i, \gamma_i]^T$, 利用SIR模型计算$I(t), t = 1, \dots, m$
  - 代入拟合误差中计算 $e_i(z_i) = \sum_{i=1}^{m} \|I(t) - x(t)\|_2^2$
- 其它参数的选取
  - 粒子数$M = 40$个
  - 迭代次数$n = 200$轮

拟合实例一



- 拟合效果良好(误差值0.0018)

拟合实例二



- SIR模型无法拟合传播情况复杂的曲线
- 误差值0.0516，误差是单峰情况误差的近30倍!

拟合原因较差的分析

- 不能够良好拟合的原因：模型的局限性
- SIR模型是刻画单一信息传播的模型
- 可以证明SIR模型中$I(t)$最多只能有一个峰值
- 传播曲线为多峰的信息
  - 相较传播曲线为单峰的信息更加复杂，包含多个子信息
  - 后续的传播峰值可能是由子信息的传播所造成的
  - 各个子信息之间可能会有相互影响(促进或抑制)

两信息传播相互影响实例

Chinese panic-buy salt over Japan nuclear threat

Beijing supermarkets run out of salt after false rumours circulate that iodised salt can help ward off radiation poisoning

▲ Customers flock to buy salt at a supermarket in Lanzhou, Gansu province. Photograph: China Daily/Reuters

Worried shoppers stripped stores of salt in Beijing, Shanghai and other parts of China on Thursday in the false belief it can guard against radiation exposure, even though any fallout from a crippled Japanese nuclear power plant is unlikely to reach the country.

日本核泄漏

↓ 极大促进

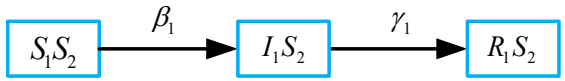
碘盐防辐射谣言

混合SIR模型—模型假设

- 假设一个信息由两个子信息E_1及E_2组成
- 每个用户一次只能传播一个子信息
- 两个子信息开始传播的时间可能不同
  (不妨设E_1先传播)
- 传播过一个子信息的群体对于另一个子信息的接受程度不同 → 子信息之间的相互影响

混合SIR模型

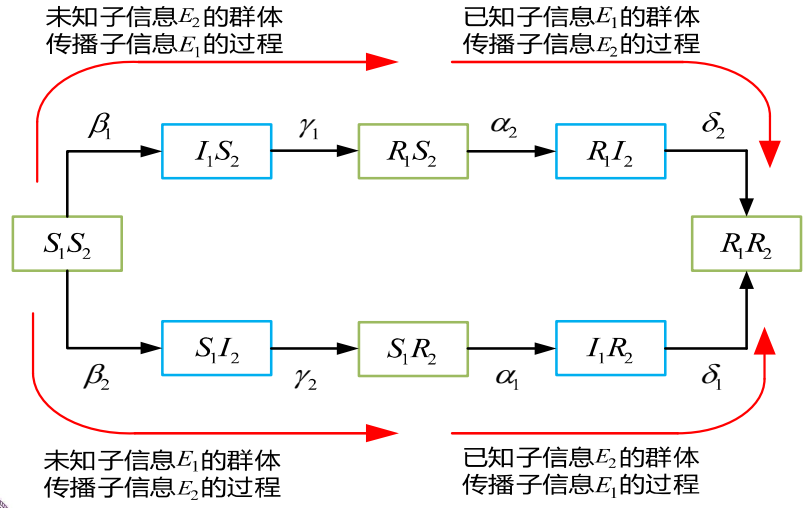
- **阶段1**：仅有子信息E_1传播时采用SIR模型刻画($[0,\tau]$)



- **阶段2**：两个子信息共同传播时，用户的八种状态($[\tau,+\infty]$)

群体	含义	群体	含义
S_1S_2	尚未传播E_1以及E_2	I_1R_2	传播E_1且不再传播E_2
S_1I_2	传播E_2而尚未传播E_1	R_1S_2	尚未传播E_2且不再传播E_1
S_1R_2	尚未传播E_1且不再传播E_2	R_1I_2	传播E_2且不再传播E_1
I_1S_2	传播E_1而尚未传播E_2	R_1R_2	不再传播E_1和E_2

混合SIR模型—阶段2的状态转移



混合SIR模型—分析

- E_1的传播参数β_1和α_1
  - $\beta_1 \rightarrow$ 未传播过$E_2(S_1S_2)$, $\alpha_1 \rightarrow$ 传播过$E_2(S_1R_2)$
  - $\beta_1 < \alpha_1$: 传播过E_2更愿意传播$E_1 \rightarrow E_2$促进E_1传播
  - $\beta_1 > \alpha_1$: 传播过E_2不愿意传播$E_1 \rightarrow E_2$抑制E_1传播
  - 同理可分析β_2和α_2
- 恢复参数$\gamma_1, \gamma_2, \delta_1, \delta_2$
  - 实际情况：相较于传播过程，子信息对恢复过程的影响较小
  - 推测： $\gamma_1 \approx \delta_1$ 以及 $\gamma_2 \approx \delta_2$

混合SIR模型—动力学方程

$$\frac{dS_1S_2(t)}{dt} = -\beta_1 \cdot I_1(t) \cdot S_1S_2(t) - \beta_2 \cdot I_2(t) \cdot S_1S_2(t)$$

$$\frac{dI_1S_2(t)}{dt} = \beta_1 \cdot I_1(t) \cdot S_1S_2(t) - \gamma_1 \cdot I_1S_2(t) \qquad \frac{dS_1I_2(t)}{dt} = \beta_2 \cdot I_2(t) \cdot S_1S_2(t) - \gamma_2 \cdot S_1I_2(t)$$

$$\frac{dR_1S_2(t)}{dt} = -\alpha_2 \cdot I_2(t) \cdot R_1S_2(t) + \gamma_1 \cdot I_1S_2(t) \qquad \frac{dS_1R_2(t)}{dt} = -\alpha_1 \cdot I_1(t) \cdot S_1R_2(t) + \gamma_2 \cdot S_1I_2(t)$$

$$\frac{dR_1I_2(t)}{dt} = \alpha_2 \cdot I_2(t) \cdot R_1S_2(t) - \delta_2 \cdot R_1I_2(t) \qquad \frac{dI_1R_2(t)}{dt} = \alpha_1 \cdot I_1(t) \cdot S_1R_2(t) - \delta_1 \cdot I_1R_2(t)$$

$$\frac{dR_1R_2(t)}{dt} = \delta_2 \cdot R_1I_2(t) + \delta_1 \cdot I_1R_2(t)$$

$$I_1(t) = I_1S_2(t) + I_1R_2(t) \qquad I_2(t) = S_1I_2(t) + R_1I_2(t) \qquad I(t) = I_1(t) + I_2(t)$$

$$S_1S_2(t) + R_1S_2(t) + S_1R_2(t) + R_1R_2(t) + I_1(t) + I_2(t) = 1$$

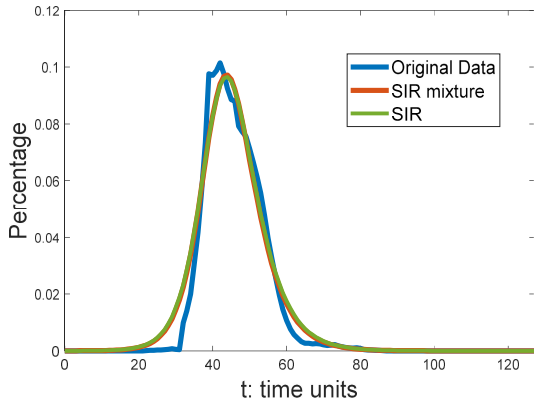
混合SIR模型拟合

- 实际Twitter观测数据（对于某一条具体信息）
  - 传播量序列 $X(t), (t = 1, \dots, m)$
  - 归一化 $x(t) = X(t)/N$, 其中的N为总人数
- 实际数据中无信息具体内容，仅有整体传播数据
  - 模型中与实际数据的对应：
    $I(t) = I_1S_2(t) + I_1R_2(t) + S_1I_2(t) + R_1I_2(t)$
- 拟合参数：$\theta = [\beta_1, \alpha_1, \beta_2, \alpha_2, \gamma_1, \delta_1, \gamma_2, \delta_2, \tau]^T$
- 优化目标： $\min_{\theta} \sum_{i=1}^{m} \|I(t) - x(t)\|_2^2$

用PSO算法拟合混合SIR模型

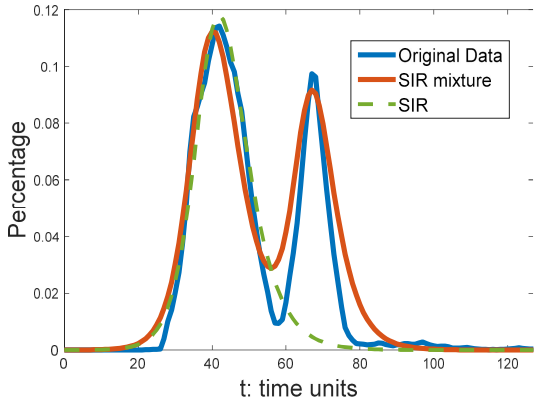
- 搜索空间z
  - $\beta_1, \alpha_1, \beta_2, \alpha_2, \gamma_1, \delta_1, \gamma_2, \delta_2$ 的范围均为$[0,1]$
  - τ 的范围：$[0, m]$
- 适应值fitness的计算
  - 根据$z_i = \theta_i$, 利用混合SIR模型计算
    $I(t) = I_1S_2(t) + I_1R_2(t) + S_1I_2(t) + R_1I_2(t)$, $t = 1, \dots, m$
  - 代入拟合误差中计算 $e_i(z_i) = \sum_{i=1}^{m} \|I(t) - x(t)\|_2^2$
- 其它参数的选取
  - 粒子数$M = 40$个 迭代次数$n = 200$轮

混合SIR模型拟合实例一



- 对单峰信息拟合良好(误差为0.0018)
- 学习得到的参数中$\tau \approx m$，说明了整个传播过程只有一个信息传播

混合SIR模型拟合实例二



- 对双峰事件的拟合效果优于SIR模型（误差为0.0041，相比SIR模型提升92%）

对学习参数的分析

- 传播参数结果分析
  - $\beta_2 \approx 0.54,\ \alpha_2 \approx 1,\ \beta_1 \approx 0.63, \alpha_1 \approx 0.52$
  - $\beta_2 < \alpha_2$: E_1促进E_2传播（有了先验知识）
  - $\beta_1 > \alpha_1$: E_2抑制E_1传播（用户总是更关注新的信息）
- 恢复参数结果分析
  - $\gamma_1 \approx \delta_1 \approx 0.35, \gamma_2 \approx \delta_2 \approx 0.57$
  - 用户对信息失去兴趣受信息间相互作用影响较小
- 时间间隔结果分析
  - $\tau \approx 25$：恰为实际数据中两个峰值间隔

总结

- 模型决定拟合效果的好坏（上限）
- 模型的可解释性可以从学习得到的参数上体现
- 拟合的方法实际是解决优化问题
- 采取有效的优化方法能够大幅提升拟合的效率和准确性

参考文献

- Kermack W. O. and McKendrick A. G., A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A*, 1927, 115(772): 700-721.
- Daley D. J. and Kendall D. G., Epidemics and rumours. *Nature*, 1964, 204(4963): 1118-1118.
- Eberhart R. and Kennedy J., A new optimizer using particle swarm theory. *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, IEEE*, 1995: 39-43.

思考

- 思考：四种误差在这两个实例中的体现?