复杂网络中统计规律的检验

修格致

IRSGIS
Peking University

2019 年 7 月 26 日

Contents

统计检验

- 什么是统计检验?
- 为什么要做统计检验?
- 统计检验的对象是什么?
- 怎么算完成了统计检验?

统计检验的常见模型

- 假设检验
- 回归分析

复杂网络中的统计规律

- 有什么?
  - 小世界：三角形的比例 [Ref: Nat. Phy. Multiscale Navigation].
  - 无标度：幂律的度分布 $P(k) \sim k^{-\gamma}$.
- 有什么问题?
  - 数据相关性：规模效应、局部特征.
  - 统计工具缺陷：幂律数据 v.s. 线性误差.
  - ……

文献列表

- Testing Statistical Laws in Complex Systems
- Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups
- Polya filter

Testing Statistical Laws in Complex Systems

- 我们该如何正确地检验复杂系统中出现的统计规律?

摘要翻译

复杂系统中有着形形色色的统计规律，比如统计单词出现频率的 Zipf 定律，地震里氏等级的 Gutenberg-Richter 律，和复杂网络中可能会出现的无标度分布律等。检验这些定律需要更高的统计观点。本 letter 中，我们讨论了一个复杂系统所生成数据的一个常见现象：这些统计规律是如何受到观测的相关性影响的。我们首先证明了标准的最大似然估计如何得到“第二类错误”(错误的否定了结论)。然后我们提出了一个保守的方法来检验这些规律，并证明了这种方法找到的参数有更小的拒绝率和更大的置信区间。

Trends

- 1999: BA 模型被提出. 随后几年无标度分布在各种网络中被发现.
- 最近五年：很多研究证明，这其中的大部分并不是幂律分布.
  - 原因：
    1. 数据集变得更大
       (是否从侧面证明：即使数据不是幂律分布的，采样也是幂律分布的?) .
    2. 统计方法的提升：从最小二乘拟合变成了最大似然估计.

建立正规的统计模型

我们检验一个数据集服从何种分布的时候，通常依赖于以下两个假设：

1. 观测数据服从某分布 $p(x, \vec{\alpha})$, 比如说幂律分布

$$p(x; \alpha) = Cx^{-\alpha}$$

2. 经验观测 $x_1, x_2, \ldots, x_N$ 是独立的, 在这里独立指关于 i 和之前的变量 x_{i-1} 这两个意义.

其中第一条讲得是统计定律，后一条是关于统计检验的. 比如说对数似然函数的加和依赖于这条独立性. Why?

问题?

- 复杂网络中的数据相关性与与时间的关系?
- 独立性?

规模效应是否与统计检验要求的独立性有着本质的矛盾? 对假设 2 的背离可能会导致服从假设 1 的分布被否定。

Thank You!

My personal website: `gxiu.github.io`