学习和解释经验数据中的复杂分布

摘要

为了适应经验数据分布，然后以生成方式解释它们，这是了解各种学科数据的结构和 动态的常见研究范式。但是，以前的工作主要是尝试以个案的方式拟合或解释经验数据分 布。面对现实世界中复杂的数据分布，我们可以通过统一但简约的参数化模型来拟合和解 释它们吗？

在本文中，我们将复杂的经验数据视为由动态系统生成，该系统将均匀随机性数据作 为输入。通过对数据的生成动力学建模，我们展示了一个四参数动态模型以及推理和模拟 算法，它能够拟合并生成一系列分布，范围从高斯、指数、幂律、拉伸指数 （Weibull）），到具有多尺度复杂性的复杂变体。我们的模型可以通过统一的微分方程来 解释，而不是黑盒子，它可以捕捉潜在的动力学。我们的框架可以以原则的方式构建更强 大的模型。我们通过各种合成数据集验证我们的模型。然后我们应用我们的模型来自不同 学科的 16 个真实数据集。我们展示了通过最广泛使用来拟合这些数据集的系统偏差方法， 并显示我们的模型的优越性。简而言之，我们的模型可能提供一个框架，以便在经验数据 中拟合复杂的分布，更重要的是，了解它们的生成机制。

CCS概念

**计算数学→分布函数；**集体过程；

**网络→网络动态；**社交媒体网络；

关键词

复杂分布；重尾分布；生存分析；动态模型；解释性。

ACM 参考格式：

郑承熙，崔鹏，朱文武。2018.学习和解释经验数据中的复杂分布。在第 24 届 ACM SIGKDD 知识发现与数据挖掘国际会议（KDD '18）的会议录中。美国纽约 ACM，10 页。 https://doi.org/10.1145/ 3219819.3220073。

1 引言

通过参数模型拟合经验数据分布，然后以生成方式解释它们是理解数据基础的结构和 生成动力学的主要科学范例，其广泛用于各种领域，包括生物学[9]，物理学[1,5]，社会科学[12,15]，计算机科学[8,32]等。例如，通过调查网络的幂律度分布[2]，物理学家发现了网络演化随机网络中的动态。通过检查达尔文和爱因斯坦[18]或在线合作[30]的对应模式的响应时间分布，社会科学家试图揭示人类行为的决策动态。通过拟合高斯混合模型[19]，贝叶斯方法[14]，甚至深度生成模型[10]的数据分布，计算机科学家试图找到观察数据集的聚类结构和生成动力学。简而言之，这种科学范式适用于广泛的数据科学任务。

然而，以前的工作主要是试图以个案的方式拟合或解释复杂的经验数据。例如，高斯分布最广泛用于拟合窄尾数据分布。大量的文献试图通过幂律分布[5]，威布尔分布（或拉伸的指数分布）[11]来模拟重尾数据，等等。特定混合模型也用于拟合复杂的多尺度分布[23,30,32]。像 GAN 这样的深度生成网络在拟合一维参数分布方面表现出有限的功效[24]。因此，我们是否可以拥有一个统一的模型来拟合和解释现实世界中各种复杂的数据分布？回答这个问题至关重要。

在本文中，我们试图通过调查其生成动态来拟合经验数据中的复杂分布。我们的模型的直观显示如下：我们将具有复杂分布的经验数据视为从动态系统生成，其采用均匀随机性作为输入。我们不是直接以个案的方式对各种复杂的分布进行建模，而是尝试对其统一的、可能是简约的生成动态进行建模，从而生成所有这些复杂的分布。表 1 显示了一个例子：不是通过四参数动态模型拟合高斯，指数，幂律，拉伸指数，以及它们在多尺度方案中具有复杂性的复杂变体，我们可以捕获它们所有。我们的框架可以以原则的方式构建更复杂的动态模型，提供了有效的推理方法和模拟算法。此外，我们不是通过黑盒模型，而是通过统一的动态微分方程来解释这些复杂分布的生成动力学。至于实验，我们通过各种合成数据集分析模型的属性，并通过来自各个学科的 16 个经验数据集进一步验证它。我们的模型准确地拟合了所有这些复杂的经验数据（图 5）。我们的模型可能提供一个框架，以适应在现实世界中观察到的复杂分布，更重要的是，了解它们的生成机制。我们总结了我们的归纳如下：

* 统一能力：我们提出了一个通用模型来拟合经验数据中的各种复杂分布，以及推理和模拟算法。
* 简约：我们的模型只有四个参数来捕捉经验分布中的多尺度复杂性。
* 可解释性：我们的模型由统一的生成动力学方程解释。所有参数都有明确的物理意义。
* 实用性：我们的模型可以准确地拟合各种经验数据集，并且可以以原则的方式推广到更复杂的情况。

论文的大纲是：调查，模型，机制，实验，讨论和结论。复用性：软件和数据库是开放源码，[www.calvinzang.com](http://www.calvinzang.com)。

2 相关工作

我们主要回顾以下两个方面的相关工作：

**从经验数据中的简单分布到复杂分布**。窄尾分布，如指数分布和高斯分布，可以通过它们的均值和方差很好地捕获，它们具有充分研究的基础结构和动力学。相比之下，重尾分布，如幂律分布，拉伸指数分布，对数正态分布等，表现出更大的均匀无穷大方差，这意味着复杂的基础结构和数据动态。在重尾分布中，幂律分布是最着名的分布，因为它具有缩放特性[22]和生成机制[2]。关于幂律分布的广泛证据和讨论可以在[13,16]中找到。最近，越来越多的文献发现经验数据的分布比纯粹的幂律更复杂，从人类行为数据[26,32]，网络数据[3]到各种数据集，如图 5 所示。

**拟合复杂的分布**。可能使用先验或正则化因子的最大似然估计用于拟合窄尾分布[14]。另一方面，像 GAN 这样的深度生成网络通过拟合 1-D 参数分布得到验证，但表现出较大的偏差[24]。相比之下，对于复杂分布的拟合理论，比如偏斜或重尾分析[17]，尚未确定。以最典型的壳体-幂律分布为例，视觉检测和最小二乘拟合首先用于拟合幂律分布。后来，well-celebrated的工作[5]显示了最小二乘拟合方法的偏差，然后提出了参数方法（）根据最大似然原则来拟合幂律分布。PL方法已被广泛用于通过 大量科学论文拟合合理的幂律分布。然而，我们发现PL方法在检测现实世界数据中的幂律信号时表现出较大的偏差，如图5所示。失败的根源在于PL方法忽略了现实世界数据的复杂性[23,26,32]。如何通过统一模型拟合和解释经验数据集中的各种复杂分布在很大程 度上是未知的。

3 提出的方法

3.1 直观模型

我们的直观模型如下：我们将具有复杂分布的经验数据视为从（非线性）动态系统生 成，该系统将均匀随机性作为输入。我们试图捕捉它们统一的生成动力学，而不是对这个 动态系统的复杂输出（即各种数据分布）进行建模，简而言之，我们试图模拟产生复杂现 象的简单生成动力学。

我们的模型基于生存分析[32]，点过程[26]和动态系统[25,27,31]。数据可以通过危险率函数来建模，其描述了以为条件的随机变量的出现率，其中表示累积危险率。通过对危险率的建模，我们可以根据关系得到复杂的概率密度函数。我们进一步建立危险函数 2（x）与其相应的动 态系统之间的联系，以解释下一节中数据分布的生成机制。

3.2 模型

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

**图1：基本模型功能的图示。我们的模型生成一系列分布，包括幂律（PL），具有截止的PL，具有短尺度复杂度的PL，具有多尺度复杂度的PL，指数，拉伸指数（SE），具有短尺度复杂度的SE，SE具有多尺度复杂性等。**

在这里，我们提出简单但通用的基本模型，导致各种分布，如表 1 和图 1 所示。指定模型的危险率函数是：

(1)

*3.2.1 关键案例：基于幂律的分布。当时，基于幂律分布生成一系列分布。*

**引理3.1** 产生一系列分布，从指数分布，幂律分布，指数截止的幂律分布到复杂的多尺度分布。

**证明**：随机变量的概率密度函数为：

**指数分布**。当时，无论其他三个参数如何，都会产生指数分布，概率密度函数，如图1a灰度曲线所示。

**幂律分布**。当且时，。的另一个含义是最小值，称为，可以取：。

**具有截止的幂律分布**。当且，。

**复杂的多尺度分布**。当时，复数多尺度分布具有常数短尺度，幂律中尺度和指数长尺度。当时，。短尺度方程，它慢慢衰减到幂律中等范围。当，即，这是指数长尺度方程。当，这是幂律中等尺度方程。

*3.2.2一般情况：基于拉伸指数的分布。当基于拉伸指数分布生成一族分布时。*

**引理3.2**。生成家族分布，从指数分布，拉伸指数（Weibull）分布，拉伸指数分布，指数切割到复杂的多尺度分布。

**证明**：与上述理由类似，随机变量x的概率密度函数为：

**指数分布**。当。

**拉伸指数（Weibull）分布**。当且，会有

累积密度函数是。这是拉伸的指数分布。一些特殊情况：当时为指数分布，当近似正态分布。

**具有指数切割的拉伸指数分布**。当且，。

**复杂的多尺度分布**。当时，复数多尺度分布基于拉伸指数分布。当时，复数多尺度分布具有常数短尺度，拉伸指数具有中尺度和指数具有长尺度，当时，。在短尺度方案中，它缓慢地衰减到拉伸的指数中尺度范围。当，即， 。它是指数长尺度方程。当，，这是指数中等规模尺度。

我们在图1中说明了上述理由。我们发现基于幂律分布的复杂分布（图1a）和拉伸指数分布（图1b）是由我们的简单危险函数生成的，可能在多尺度机制中具有复杂性。

3.3参数推理

我们的模型参数可以通过最大似然估计（MLE）框架来学习。观察一组数据的对数似然函数由下式给出：

根据的值，采用不同的形式。当时，对数似然函数是

然而当时，对数似然函数是：

关于的最大化等式5或6，受的约束导致估计的建模参数。但是，由于参数的物理意义明确，可以将先验知识应用于初始化。我们稍后会说明这一点。

该模型的另一个优点是所有参数都具有闭合形式的梯度。情况的梯度，即基于拉伸指数的模型，是：

约束条件：

当时，基于幂律的模型的梯度是：

约束条件：

我们可以通过许多基于梯度的优化算法来解决优化问题。例如，我们采用内点算法[4]，为了重现性，我们打开代码，参见第7节。

3.4 生成器

从累积分布函数生成随机数的最简单和最优雅的方法是逆变换方法[7]。首先我们从标准均匀分布生成一个随机数u。通过求解方程表示，是跟随分布

的数。我们将这个逆变换方法扩展到由的事实导致的危险率函数，其中。因此，，我们可以通过求解

得到所需的随机数，其中和从采样时没有区别。由于是单调递增函数，因此具有反函数，我们可以得到

即使没有闭合形式的反函数，我们也可以通过求解方程来得到数值，其中是从均匀分布生成的。

伪代码

4物理机制

在本节中，我们给出了模型的基本生成动力学，即Equ.1和各种分布，如表1所示。我们将复杂数据分布视为由（非线性）动态系统生成，采用均匀随机性作为输入：

4.1统一输入和增长

为了给出数据生成过程的动态视图，我们的第一步是通过连接点过程和生存分析来计算动态系统的输入。我们从标准均匀分布中采样个数的过程可以看作是一个随机点过程。给定泊松过程，然后是均匀分布在区间上。如果我们将标准化为，则遵循标准均匀分布。

我们将Equ.15中的替换为，导致代理的增长动态，其中的均匀到达时间：

例如，让，当时它产生幂律分布；并且当

时，拉伸指数分布为（详见第3节）。因此，，并且。我们可以得到它们的反函数，和。

通过在Equ.16中使用，我们获得了超过时间的增长曲线：

类似地，当时，通过将应用于等式16，，我们得到增长曲线：

4.2 生成动力学：基础模型

我们的第二步是通过连接生存分析和动态系统来逆向工程生成动力学。我们将Equ.17和Eq.18的导数推导到时间，我们得到幂律分布数据和拉伸指数分布数据的生成动力学如下：

我们发现方程19的线性优先附着产生幂律分布，而等于20的非线性优先附着产生拉伸指数分布，与随机网络中无标度观测的文献一致【2】。

4.3生成动力学：一般模型

在这里，我们给出了模型的生成动力学。当时，。

通过我们的建设，我们得到：

通过将上述方程的导数推导到时间，我们得到：

因此，从动态的角度来看，表现出复杂多尺度分布的复杂数据（由我们的模型Equ.1捕获）是从具有微分方程22的动态系统生成的，包括物理机制：非线性优先附着，增长系统，以及短期复杂度和长期复杂度。我们的动态包括Equ.19和20作为特殊情况。

因此，我们在随机网络场景中描述数据生成过程如下：

* 新节点在的泊松过程后进入网络，其中是最大观测时间，
* 并且，节点的程度，表示为，根据微分方程22随时间增长。

然后，该网络在时间的横截面度分布为，其中如Equ.1所示。

5.1综合数据分析

*5.1.1忽略完整性会导致系统偏差。*实际上，衡量现实世界数据量的分布要比纯幂律分布复杂得多。我们将在下一节中展示来自真实世界数据集的证据。在这里，我们研究了通过将着名的幂律拟合方法（表示为PL方法）[5]应用于复杂分布而引入的可能偏差。

**长期复杂性**。我们首先研究长期复杂性。参数作为建模长尺度复杂性的最简单形式。通过改变中的，我们得到了图3a中所示的一系列分布。当时的情况，如图3a中蓝色曲线的直线部分（幂律指数）所示，表明缺少长尺度复杂性。随着的增加，长尺度的复杂性将向短距离移动，直到两个部分重叠。实际上，长期制度的特征尺度是，短期制度的特征尺度是（参见第3节）。我们通过限制来避免这种重叠。我们生成104（一个相对较大的数据集，以获得合理的拟合结果，同时在基线方法的可扩展性限制内。我们将在稍后显示模型和基线的可扩展性。 ）采用每个特定的样本，并通过PL方法和我们的方法拟合和。图3e和绘制了（缩放指数）和作为的函数的平均估计值。我们发现PL方法估计的幂律缩放参数与虚线标记的真值之间的差异越来越大，增加，如图3e所示。相比之下，我们的模型很好地获得了真正的缩放值。对于图3i所示的短尺度的估计，PL方法严重高估了真实值，当时，真实值高达450倍。随着的增长，长期制度挤入短期制度，因此通过PL方法估计的降低到真实值。