

数据整合驱动的极大似然估计教学改革探索 ——以转录动力学建模为例

杨喜艳¹, 王子豪², 吴亚豪¹

1. 广东金融学院金融数学与统计学院, 广东 广州 510521

2. 中山大学数学学院, 广东 广州 510275

DOI:10.61369/ASDS.2025040012

摘 要 : 极大似然估计 (MLE) 是统计推断中的核心方法, 广泛应用于生命科学数据建模。随着现代生物技术的发展, 实验数据呈现多样化特征, 如何有效整合不同类型的数据以提高 MLE 准确性, 已成为统计建模与生命科学交叉研究中的重要问题。本文以转录动力学为例, 探讨如何通过整合 nascent RNA 表达数据与转录启动时间数据, 精确估计随机动力学模型中的关键参数, 以增强学生对数据驱动建模的理解, 为相关课程的教学改革提供思路。

关 键 词 : 数据整合; 极大似然估计; 转录动力学建模

Exploring of Teaching Reform of Maximum Likelihood Estimation Driven by Data Integration: A Case Study in Transcription Dynamics Modeling

Yang Xiyan¹, Wang Zihao², Wu Yahao¹

1. School of Financial Mathematics and Statistics, Guangdong University of Finance, Guangzhou, Guangdong 510521

2. School of Mathematics, Sun Yat-sen University, Guangzhou, Guangdong 510275

Abstract : Maximum likelihood estimation (MLE) is a fundamental method in statistical inference and is widely used in life science data modeling. With the advancement of modern biological technologies, experimental data increasingly exhibit diverse characteristics. Effectively integrating different types of data to improve the accuracy of MLE has become an important issue in the interdisciplinary research combining statistical modeling and life sciences. This paper uses transcription dynamics as a case study to explore how the integration of nascent RNA expression data and transcription initiation timing data can enable precise estimation of key parameters in stochastic dynamical models. The approach enhances students' understanding of data-driven modeling and provides new perspectives for teaching reform in related courses.

Keywords : data integration; maximum likelihood estimation; transcriptional dynamics modeling

引言

随着现代生物技术的发展, 尤其是在分子水平上的基因表达研究中, 实验数据的复杂性和多样性呈现出前所未有的挑战^[1,2]。在转录动力学研究中, 传统的基于稳态基因表达数据单一数据的建模方式已经无法全面捕捉真实的转录过程^[3]。随着单分子成像技术的发展, 从影像数据中能够获得表征动态转录过程的时间序列数据, 这为深入分析转录调控机制提供了新机遇。然而, 如何从时间序列数据中提取并有效整合不同类型的数据, 提升极大似然估计的准确性和模型的可识性, 仍然是当前统计建模中的重要难题。在传统的 MLE 教学中, 学生通常仅通过单一数据来推断模型参数, 缺乏对数据整合与模型复杂性关系的深入理解。因此, 本文提出将数据整合的理念引入统计推断课程的教学改革与实践中, 利用转录动力学作为真实科研案例, 让学生理解不同类型的数据在模型构建中的作用, 并掌握如何通过联合似然函数精确推断转录动力学模型中的关键参数。此举不仅帮助学生提高统计推断能力, 更通过案例教学引导学生理解科研创

基金项目:

2022年度广东金融学院校级教学质量与教学改革工程项目: 课程思政理念下概率论与数理统计的教学改革与实践 (项目编号: 23);

广东省教育科学规划课题 (高等教育专题): OBE 理念下“大学数学”问题驱动下的启发式教学模式实践研究 (项目编号: 2024GXJK590)。

作者简介:

杨喜艳 (1983—), 女, 湖南邵阳人, 副教授, 研究方向: 计算生物学与生物统计;

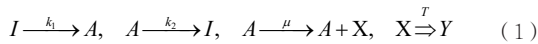
王子豪 (1993—), 男, 山西晋城人, 副教授, 研究方向: 计算生物学与生物信息学;

吴亚豪 (1978—), 男, 湖北黄冈人, 讲师, 研究方向: 动力系统。

新与学科交叉的重要性,从而实现课程思政与专业知识教学的有机结合。同时,结合启发式教学模式,通过引导式问题设计与学生自主探索,激发其对统计推断背后逻辑的深入思考。通过这种方式,学生不仅能够掌握数据整合下 MLE 的基本原理与方法,更能树立科学探索的责任感和使命感,培养其对多学科交叉学科问题的理解与解决能力。这种教学模式为大数据时代下的教学改革提供了新的思路,推动学生综合能力的提升。

一、研究背景与方法

为了说明极大似然估计在生命科学中建模的应用,这里我们考虑分子水平上的基因转录。转录是一个复杂的过程,其中涉及到两个关键的阶段:启动和延伸,每一个阶段在基因表达的研究中起着非常重要的作用^[4,5]。数学建模为理解基因表达的统计特征如何从转录的分子机制中产生,提供了一种潜在的途径。目前已经发展出许多的随机基因表达的模型,其中应用最为广泛的是两状态电报模型,该模型用生化反应描述如下:



在方程(1)中, I 和 A 分别表示基因启动子非活性态和活性态, X 表示 nascent RNA, Y 表示 mature RNA。其中第一个反应描述了基因启动子以速率 k_1 从非活性态 I 切换到活性态 A , 第二个反应描述了基因启动子以速率 k_2 从活性态 A 切换到非活性态 I , 第三个反应表示在活性态时以速率 μ 合成 nascent RNA, 最后一个反应表示 nascent RNA 以时间 T 延伸变为 mature RNA。注意到(1)中每一个反应的发生都是随机的。

为了从时间序列数据中分析转录启动事件,首先需要确定启动时间分布。记 $P_1(t)$ 和 $P_2(t)$ 表示 t 时刻启动子驻留在非活性态和活性态的概率,并且记 $\mathbf{P}(t) = (P_1(t), P_2(t))^T$ 表示列向量。此外,定义启动矩阵 $\mathbf{A}_{ini} = (A_i)$, 其中 $A_{11} = -k_1$, $A_{12} = k_2$, $A_{21} = k_1$, $A_{22} = -(k_2 + \mu)$, 则我们可以写出如下的主方程

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{A}_{ini} \mathbf{P}(t), \quad (2)$$

其中初值 $\mathbf{P}(0) = (P_1(0), P_2(0))^T$ 。则方程(2)的解可表示为 $\mathbf{P}(t) = e^{\mathbf{A}_{ini}t} \mathbf{P}(0)$, 则可得启动时间分布的概率密度函数

$$f_{ini}(t) = \mathbf{u}_{ini} \mathbf{P}(t) = \mathbf{u}_{ini} e^{\mathbf{A}_{ini}t} \mathbf{P}(0), \quad (3)$$

其中 $\mathbf{u}_{ini} = (0, \mu)$ 。为了求得 $f_{ini}(t)$ 的显示表达, 对方程(3)两边取拉普拉斯变换

$$\tilde{f}_{ini}(s) = \mathbf{u}_{ini} (s\mathbf{I} - \mathbf{A}_{ini})^{-1} \mathbf{P}(0). \quad (4)$$

通过代入已知量, 可得模型(1)的拉普拉斯变换的显示表达

$$\tilde{f}_{ini}(s) = \frac{\mu s + k_1 \mu}{s^2 + (k_1 + k_2 + \mu)s + k_1 \mu}, \quad (5)$$

则特征方程 $s^2 + (k_1 + k_2 + \mu)s + k_1 \mu = 0$ 的根

$$\begin{aligned} -\lambda_1 &= \frac{-(k_1 + k_2 + \mu) - \sqrt{(k_1 + k_2 + \mu)^2 - 4k_1 \mu}}{2}, \\ -\lambda_2 &= \frac{-(k_1 + k_2 + \mu) + \sqrt{(k_1 + k_2 + \mu)^2 - 4k_1 \mu}}{2}. \end{aligned} \quad (6)$$

对方程(5)两边求逆拉普拉斯变换, 得启动时间分布

$$f_{ini}(t) = c_1 e^{-\lambda_1 t} + c_2 e^{-\lambda_2 t}, \quad (7)$$

其中系数

$$c_1 = \frac{\lambda_1(\mu - \lambda_2)}{\lambda_1 - \lambda_2}, \quad c_2 = \frac{\lambda_2(\lambda_1 - \mu)}{\lambda_1 - \lambda_2}. \quad (8)$$

接下来我们给出模型(1)中 nascent RNA 稳态分布的计数分布, 具有如下形式

$$P(M=m) = \begin{cases} C_{1,1,m} e^{-\lambda_1 T} + C_{2,1,m} e^{-\lambda_2 T} & m=0 \\ \sum_{k=0}^m \frac{T^{m-k} (C_{1,k,m} e^{-\lambda_1 T} + C_{2,k,m} e^{-\lambda_2 T})}{(m-k)!}, & m \geq 1 \end{cases} \quad (9)$$

(9)的具体计算过程, 可参考文献^[6]。

二、基于数据整合的极大似然推断

(一) 模拟数据产生

为了探讨数据整合的方法在极大似然估计中的有效性, 我们使用模拟数据验证。假设模型(1)中生化反应的参数从一个大的参数空间产生

$$(k_1, k_2, \mu, T) \in [U(0.01, 10), U(0.01, 10), U(0.01, 50), U(1, 500)] = \Theta, \quad (10)$$

其中 U 表示均匀分布的随机变量。一旦参数被选中, 我们利用随机模拟算法产生一条长 10^6 s 的时间序列^[7], 通过记录相邻两次连续启动事件发生的时间间隔来收集启动时间的数据; 以及通过均匀采样收集 nascent RNA 的数据。

(二) 算法步骤

记 $P(\tau, m; \theta)$ 表示启动时间数据和 nascent RNA 数据的联合分布, 其中 $\theta = (k_1, k_2, \mu, T)$ 表示要估计的参数向量。为了方便, 我们限制最优参数的搜寻区域与(10)中相同。接下来, 我们定义联合似然函数 $L(\theta) = \prod_{i,j} P(\tau_i, m_j; \theta)$, 其中 τ_i 表示第 i 个启动时间数据, m_j 表示第 j 个 nascent RNA 计数的数据。我们采用无梯度优化算法寻找极大似然估计的参数^[8]

$$\theta^* = \arg \min_{\theta \in \Theta} \left(-\sum_{i,j} \ln P(\tau_i, m_j; \theta) \right) \quad (11)$$

这里负对数似然的最小化等价于似然的最大化。假设模型(2)中转录启动事件和延伸事件的相互独立, 相应地, 联合分布可假设为两个边缘分布的乘积, 即 $P(\tau, m; \theta) = f_{ini}(\tau; \theta) P(m; \theta)$ 。则优化问题变为

$$\theta^* = \arg \min_{\theta \in \Theta} \left(-\left(\sum_i \ln f_{ini}(\tau_i; \theta) + \sum_j \ln P(m_j; \theta) \right) \right), \quad (12)$$

其中 $f_{ini}(\tau; \theta)$ 和 $P(m; \theta)$ 分别由方程和给出。

为了比较, 我们也给出仅采用 nascent RNA 数据推断方法,

此时优化问题变为

$$\theta^* = \arg \min_{\theta \in \Theta} \left(-\sum_j \ln P(m_j; \theta) \right), \quad (13)$$

为了方便, 我们把(12)和(13)中的方法分别称为整合数据推

断方法和单一数据推断方法。在每一次推断中，当似然值收敛到某个最大值时，优化算法终止。在本文中，我们约定当迭代次数超过5000次时，算法将被停止运行。

(三) 推断结果

下面比较两种方法的推断效果。根据(10)中参数 k_1, k_2, μ, T 的取值空间，我们选取两组代表性的参数：第一组参数中，模型(1)的 nascent RNA 的稳态分布呈单峰分布；第二组参数中，模型(1)的 nascent RNA 的稳态分布呈双峰分布。图1展示了两种方法的推断结果。从图1A与C中可以看到，在这两组参数下，两种推断方法均能很好地拟合模拟数据的 nascent RNA 的稳态分布。然而，两种推断方法中启动时间的分布的推断有着明显的差异。从图1B与D中可以看到，数据整合的推断方法中的启动时间分布能够很好地拟合模拟数据的结果，而单一数据的推断方法中的启动时间分布极大地偏离了模拟数据的结果。此外，表1列出了模型的真实参数值，以及两种方法估计得到的参数值。可以看到，仅在采用数据整合方法时，推断出的参数与真实值高度一致；而用单一数据方法时，参数的估计值与真实值能够达到数量级的差异。总的来说，图1和表1说明，基于单一的 nascent RNA 计数数据推断时，会造成模型参数的不可识别，而用整合数据推断时，显著提升了模型参数的可识别性和推断的准确性。

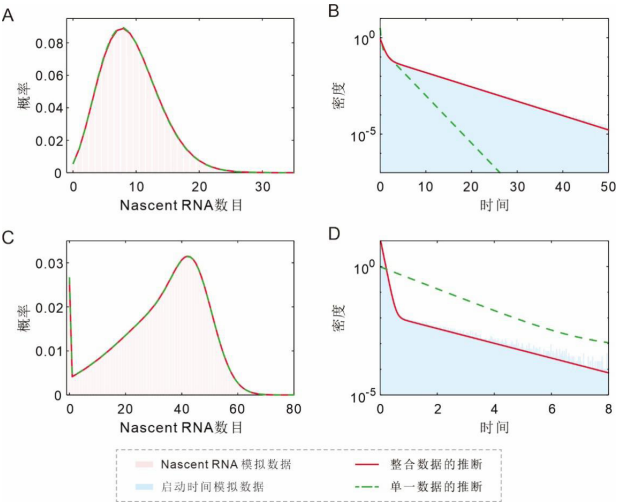


图1 两种推断方法在 nascent RNA 稳态分布和启动时间分布推断中的对比结果

表1 模型真实参数与两种推断方法估计值的对比

图 A-B	参数	k_1	k_2	μ	T
	模型使用参数	0.315	0.626	0.918	29.5
	模型估计参数				
	整合数据推断	0.31390	0.62072	0.91762	29.532
	单一数据推断	1.06876	2.29408	3.17266	9.0585

图 C-D	参数	k_1	k_2	μ	T
	模型使用参数	0.673	0.247	12.5	3.65
	模型估计参数				
	整合数据推断	0.67565	0.24905	12.4557	3.67277
	单一数据推断	0.052418	0.019282	0.97211	47.4379

三、结论与展望

本研究以转录动力学建模为例，展示了在极大似然估计中引入数据整合的优势。通过将 nascent RNA 表达数据与转录启动时间数据联合建模，显著提高了参数估计的准确性。这一结果不仅在研究上验证了数据整合的有效性，也在教学上为极大似然估计方法的教学改革提供了新的视角。传统极大似然估计教学中，往往侧重于基于单一类型数据进行参数估计，忽视了现实中数据的多样性与复杂性。本文中图 A 与图 B 所示结果清楚表明，仅依赖 nascent RNA 数据会导致严重的偏差。相比之下，整合 nascent RNA 数据和启动时间数据显著改善了模型参数的可识别性。因此，在教学中应引导学生理解数据多样性对建模结果的重要影响。

通过对科学研究问题的探讨，学生能够理解如何从不同类型的数据中构建联合似然函数，以及如何整合不同类型的数据进行推断，这种教学方式将比单纯公式推导更贴近科学研究实际，能够提高学生的数据驱动建模能力的培养。在教学过程中，结合科研实例引导学生思考生命科学背后的社会价值，如基因调控研究对于疾病理解、治疗策略设计的意义，有助于增强学生的科学责任感和社会使命感。这也为实现“课程思政”提供了自然融入的切入口，让学生在掌握专业知识的同时树立正确的科学观与价值观。该教学改革实践亦体现了 OBE 理念中“以学为中心”的启发式教学思路，通过具体问题引导学生主动思考与探索，在理解知识的同时提升其建模能力与综合素养。

参考文献

[1] Pichon X, Lagha M, Mueller F, Bertrand E. A growing toolbox to image gene expression in single cells: Sensitive approaches for demanding challenges. *Mol Cell*. 2018;71(3):468 - 480.
[2] Acosta J N, Falcone G J, Rajpurkar P, et al. Multimodal biomedical AI. *Nature Medicine*, 2022, 28(9): 1773-1784.
[3] Nicoll A G, Szavits-Nossan J, Evans M R, et al. Transient power-law behaviour following induction distinguishes between competing models of stochastic gene expression[J]. *Nature Communications*, 2025, 16(1): 2833.
[4] Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*. 2018;19(10):621 - 637.
[5] Aoi Y, Shilatifard A. Transcriptional elongation control in developmental gene expression, aging, and disease. *Mol Cell*. 2023;83(22):3972 - 3999.
[6] Yang X, Wang, Z, Shi C, Zhou T, Zhang J. Deciphering HIV-1 transcription initiation and elongation from single-molecule imaging data. *Research*. 2025; 8: 0645.
[7] Fu X, Zhou X, Gu D, Cao Z, Grima R. DelaySSAToolkit.jl: Stochastic simulation of reaction systems with time delays in Julia. *Bioinformatics*. 2022;38(17):4243 - 4245.
[8] Feldt R. BlackBoxOptim.jl. GitHub. 2019. [accessed 20 Dec 2023] <https://github.com/robertfeldt/BlackBoxOptim.jl>