arXiv:2303.11830v1 [math.HO] 20 Mar 2023

为什么每位数据科学家都必须是柏拉图主义者

Collin Joseph Van Cuyk* Liang Hong*

2023年3月22日

摘要

数据科学家不是数学家，但他们在日常工作中大量运用数学。数学家可以研究无法通过五种感官感知的数学对象，而数据科学家则必须处理我们能够观察到的现实世界数据。这条微妙的界线表明，数据科学家对数学所持的哲学立场可能会对其工作产生重大影响。通过考察不同的数学哲学观点如何影响对数据科学中基本模型假设的诠释，我们得出结论：使用现代概率与统计工具的数据科学家必须是柏拉图主义者。

关键词与短语：数学哲学；数据科学；柏拉图主义；有限主义；形式主义。

1 引言

数学是每位数据科学家不可或缺的工具。尽管大多数数据科学家很少审视自己对数学持有的哲学立场，但各种哲学观点都对数据科学家的基本假设提出了挑战。在当前文献中，关于数学哲学有三种占主导地位的思想流派：（i）柏拉图主义，（ii）有限主义，以及（iii）形式主义。这三种流派的名称在文献中略有不同；例如，参见 Benacerraf 和 Putnam（1983年）、George 和 Vellenman（2002年）、Henle（1992年）、Kleene（1950年），以及 Kunen（1980年、2011年、2012年）。本文将采用 Kunen（2012年）第三章中的术语。具体而言，我们将柏拉图主义者定义为认为整个数学都是真实存在的人。显然，现代数学有很大一部分无法在我们的现实世界中观察到，例如许多良基集合。柏拉图主义者相信，这部分数学存在于一个我们无法凭借五种感官感知的世界中。其次，我们将有限主义者定义为只研究有限数学对象的人。有限主义者拒绝确定无穷的概念，尽管他们允许潜在无穷的概念，但会明确指出，这一过程永远无法

*数学科学系，德克萨斯大学达拉斯分校，美国得克萨斯州理查森市西坎贝尔路800号，邮编75080。

真正完成。简言之，有限主义者只相信有限数学。最后，我们将形式主义者定义为只承认有限主义者所接受的事物真实存在，但也承认无限数学具有智识上的趣味性的人。也就是说，他们只承认有限主义所接受的事物有意义，却把任何无限数学都视为一场智力游戏。尽管这三种哲学立场都不被认为是“正确”的立场，我们将表明，每一种立场都会对数据科学中的基本模型假设产生不可忽视的影响。具体而言，我们将得出结论：一位诚实的数据科学家必须是柏拉图主义者。本文所说的数据科学家，是指以运用现代概率论和统计学中的常用工具分析数据为职业的人。

2 数据科学中的基本模型假设

数据科学最根本的假设是，存在一个潜在且不可观测的概率律，它支配着所有可观测数据。数据科学家会根据所要解决的问题，对这一概率律作出假设。例子包括但不限于密度模型、线性模型或时间序列模型。为了说明这一点，只需考虑参数密度模型。大多数数据科学教科书对这类模型的描述如下：假设数据 X_1、...、X_n 独立同分布（iid），服从真实（但不可观测的）概率密度函数 $f^\star$。由于 $f^\star$ 不可观测，数据科学家便假定一个参数模型

$$\mathcal{M} = \{f_\theta \mid \theta \in \Theta\},$$

其中 Θ 称为参数空间，是有限维欧几里得空间的一个子集。例如，若考虑一个伽马模型，其形状参数为 $\alpha > 0$ 、率参数为 $\lambda > 0$ ，且 α 和 λ 均未知，那么

$$\mathcal{M} = \{f_{(\alpha,\lambda)} \mid (\alpha, \lambda) \in \mathbb{R}_+ \times \mathbb{R}_+\},$$

其中 $\mathbb{R}_+ = (0, \infty)$，并且

$$f_{(\alpha,\lambda)}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, x > 0,$$

且 Γ是伽马函数。

如果数据科学家是柏拉图主义者，那么这个伽马模型 $\mathcal{M}$ 就完全说得通：模型描述中使用的每个要素都有明确定义，而且 $\mathcal{M}$本质上是一个由不可数多个伽马分布构成的族。也就是说，柏拉图主义者对这个模型的理解，与数据科学教科书中的内容相同（例如 Lehmann 1983年和 Schervish 1995年）。

然而，对于一位有限主义者来说，上述模型描述不能按照教科书的意图来解读。更糟的是，这个模型对有限主义者而言毫无意义。首先，由于有限主义者拒绝接受任何确定无穷，伽马模型的参数空间，即 $\mathbb{R}_+ \times \mathbb{R}_+$，不能是不可数集。有限主义者只允许参数空间 Θ 为有限集。这显然不同于数据科学教科书中的描述；例如，参见 Lehmann（1983年）和 Schervish（1995年）。稍加思索便会发现，还有一个更

严重的问题会令有限主义者提出异议。在现代概率论中，概率密度函数 f 是一个几乎处处非负的博雷尔可测函数，满足 $\int_{\infty}^{-\infty} f(t)dt = 1$，并且

$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

其中， F 是相应的累积分布函数。由于定积分是黎曼和的极限，而有限主义者无法接受这一概念，因此概率密度函数的概念对有限主义者来说毫无意义。因此，用伽马密度描述伽马模型，对有限主义者来说也毫无意义。同样，有限主义者也会否定概率密度函数与其相应分布函数之间的优美关系 $F'(x) = f(x)$。因此，数据科学家不可能是有限主义者。

那么形式主义者呢？数据科学家能成为形式主义者吗？形式主义者是一个“圆融”的人；他采取一种立场，以避免与柏拉图主义者或有限主义者发生冲突。尽管他的立场在数学领域无可指摘，但在数据科学领域，情况就不同了。形式主义者只相信有限数学是真实的。因此，上述伽马模型对他来说也毫无意义。他能否像对待无限数学那样，干脆把这个模型视为一种思想模型（在现实世界中或许毫无意义）？答案是否定的。数据科学家处理的是我们可以观察到的现实世界数据，因此，他的工作是提出一个“最佳”模型来描述这些现实世界数据。所以，诚实的数据科学家不能提出一个自己并不相信其存在的模型。

3 贝叶斯分析呢?

在上一节中，我们显然采用了频率学派方法。然而，如果数据科学家选择贝叶斯方法，也会面临同样的问题。为说明这一点，我们回顾一下：参数贝叶斯模型将数据 X_1、...、X_n 视为在给定参数 θ 的条件下独立同分布，并在 θ上设定先验分布 Π_0。也就是说，参数贝叶斯模型按层次结构指定如下：

$$\theta \sim \Pi_0 \text{ and } X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} f_\theta,$$

其中， f_θ 是由 $\theta \in \Theta$索引的 X空间上的概率密度函数。随后，通过贝叶斯公式将先验分布 Π_0 更新为后验分布 Π_n：

$$\Pi_n(A) = \frac{\int_A L(\theta)d\Pi_0(\theta)}{\int_\Theta L(\theta)d\Pi_0(\theta)}, \quad A \subset \Theta,$$

其中 $L(\theta) = \prod_{i=1}^{n} f(X_i)$ 是似然函数， A 是一个 Π_0-可测集。关于贝叶斯分析的详细论述，请参阅 Schervish（1995年）第 1 章和 Berger（2006年）第 4 章。由于积分和概率密度函数（因此也包括

似然函数）都涉及贝叶斯模型，因此，正如我们在上一节中所用的论证所示，数据科学家既不能是有限主义者，也不能是形式主义者。因此，数据科学家必须是柏拉图主义者。

4 结语

如果数据科学家希望自己的行为与其哲学立场一致，那么作为有限主义者，他们几乎无所作为。大多数（如果不是全部）统计推断的实际应用，都无法在绝对定义的论域或纯离散、纯有限模型中运行。我们无意评判哪种哲学更有道理；然而，就所有实际或应用目的而言，数据科学家无法像有限主义所要求的那样对自己施加如此严格的限制。形式主义者又如何呢？形式主义者认为，只有有限主义者所接受的数学才是真实的，因此也会面临其工作与哲学立场之间同样的矛盾和问题。形式主义者无法在道德上坦然地呈现自己认为不过是“一场游戏”的研究结果。因此，我们不得不 concl conclude，现代数据科学家若希望完成工作时确信其成果既真实又有意义，就必须持有柏拉图主义的哲学立场。

参考文献

Berger, J.O. (2006年).《统计决策理论与贝叶斯分析》。施普林格：纽约。Benacerraf, P. 和 Putnam, H. (1983年)。《数学哲学》，第二版。剑桥大学出版社：英国剑桥。George, A. 和 Velleman, D. J. (2002年)。《数学哲学》。布莱克韦尔出版社公司。Henle, J.M. (1991年).《快乐的形式主义者》。《数学情报》13(1), 12–18。Kleene, S.C. (1950年)。《元数学导论》。北荷兰出版社：阿姆斯特丹。Kunen, K. (1980年)。《集合论：独立性证明导论》。爱思唯尔：阿姆斯特丹。Kunen, K. (2011年)。《集合论》。学院出版社：英国伦敦。Kunen, K. (2012年)。《数学基础》，修订版。学院出版社：英国伦敦。Lehmann, E.L. (1983年)。《点估计理论》。威利：纽约。Schervish, M.J. (1995年)。《统计学理论》。施普林格：纽约。