大数据背景下精准科研信息服务

2020 年版

作者：王敏杰

时间：2020-05-21

版本：0.1

ElegantLaTeX Program

Victory won't come to us unless we go to it. — M. Moore

目录

前言

根据基本科学指标数据库（Essential Science Indicators，简称 ESI）发布的最新统计数据显示：

1、我国师范类院校有 ESI 学科的 25 所，北京师范大学进入 ESI 学科数量最多。从入选的学科来看，其中化学学科的频次最高。

2、我校工程学近十年累积被引频次 2781，距离 ESI 前 1% 学科阈值线 2843, 接近度 97.82%，有望入选 ESI 学科，但竞争依然激烈。

3、根据贝叶斯数学模型分析，我校工程学科 2020 年有约 80% 的概率进入 ESI 前百分之一学科。

进度表

- 文献调研（3 月底完成）
- 数据获取（4 月中旬完成）
- 数学分析和模型评估（5 月初完成）
- 报告初稿（5 月中旬完成）
- 研讨会（待定）
- 正式稿发布（待定）

关于本文档

本报告使用 R 和 Stan 语言完成，数据和代码存放在 GitHub 仓库`https://github.com/perlatex/ElegantBookdown4IS`，欢迎批评指正。

感谢

I am very grateful to Ben Bales from the Stan Development Team for his patience in guiding Stan code. 感谢彭凤老师在图书购买上提供的帮助，感谢研究生李晨阳协助完成数据收集和整理工作。感谢科睿唯安 (原汤森路透) 公司赵宇先生提供了非常专业地技术解释。

作者简介

王敏杰，四川师范大学研究生公选课《数据科学中的 R 语言》和《社会科学中的统计学》授课老师，毕业于西南交通大学量子物理专业，爱好数据科学，喜欢用 R 和 Stan 统计编程，联系方式 38552109@qq.com

第 1 章 师范院校

本章横向比较了我国师范类高校 (Top30) 近十年的发展情况，然后统计了各学科进入前百分之一 ESI 学科的情况。

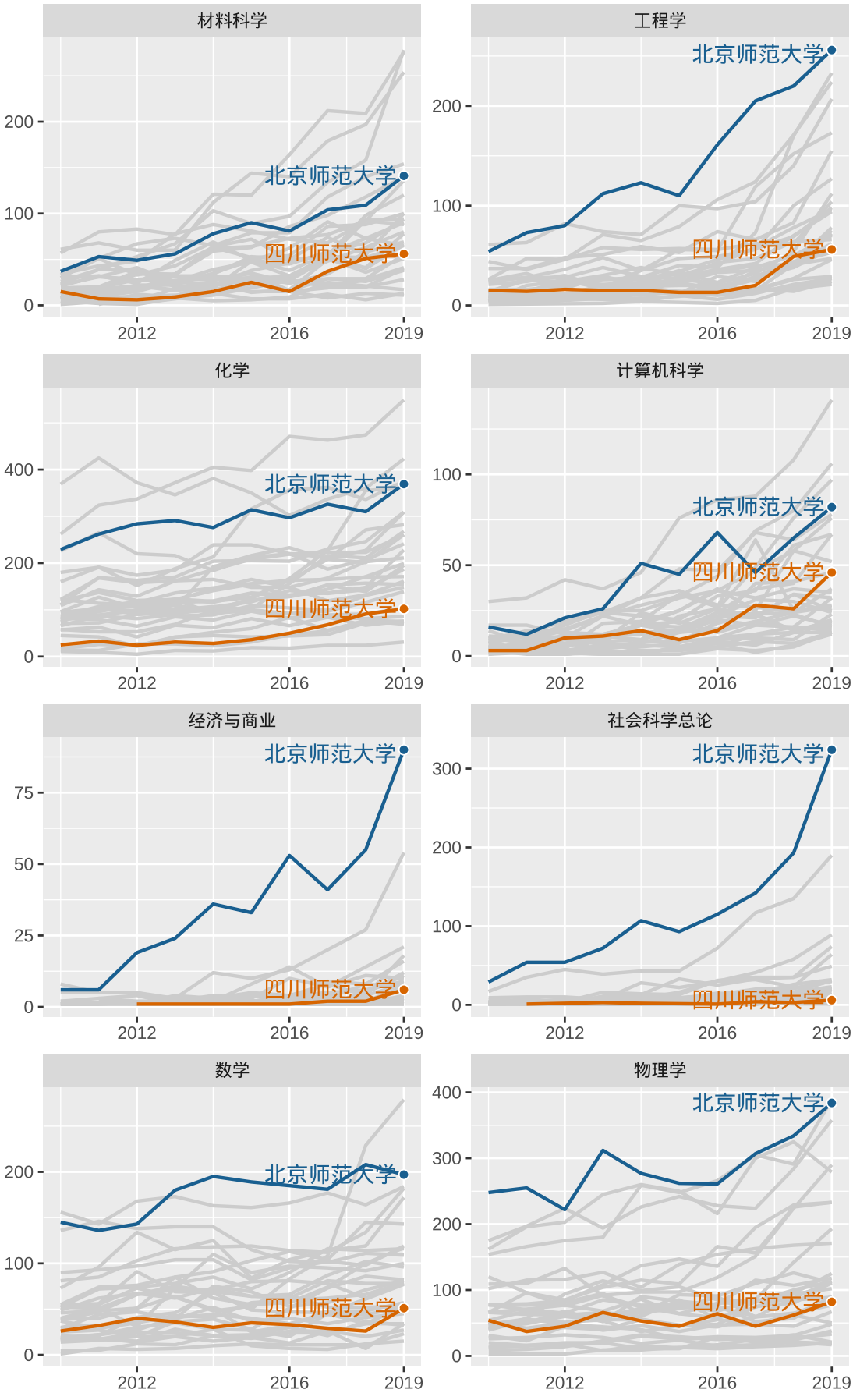
1.1 学校列表

表 1.1: 我国师范类学校 (top30) 列表

school	univ
北京师范大学	Beijing Normal University
华东师范大学	East China Normal University
华中师范大学	Central China Normal University
南京师范大学	Nanjing Normal University
湖南师范大学	Hunan Normal University
东北师范大学	Northeast Normal University
华南师范大学	South China Normal University
陕西师范大学	Shaanxi Normal University
首都师范大学	Capital Normal University
浙江师范大学	Zhejiang Normal University
山东师范大学	Shandong Normal University
天津师范大学	Tianjin Normal University
福建师范大学	Fujian Normal University
河南师范大学	Henan Normal University
江西师范大学	Jiangxi Normal University
上海师范大学	Shanghai Normal University
安徽师范大学	Anhui Normal University
西北师范大学	Northwest Normal University
广西师范大学	Guangxi Normal University
杭州师范大学	Hangzhou Normal University
云南师范大学	Yunnan Normal University
哈尔滨师范大学	Harbin Normal University
河北师范大学	Hebei Normal University
江苏师范大学	Jiangsu Normal University
四川师范大学	Sichuan Normal University
辽宁师范大学	Liaoning Normal University
重庆师范大学	Chongqing Normal University
曲阜师范大学	Qufu Normal University
贵州师范大学	Guizhou Normal University
海南师范大学	Hainan Normal University

1.2 全景大数据

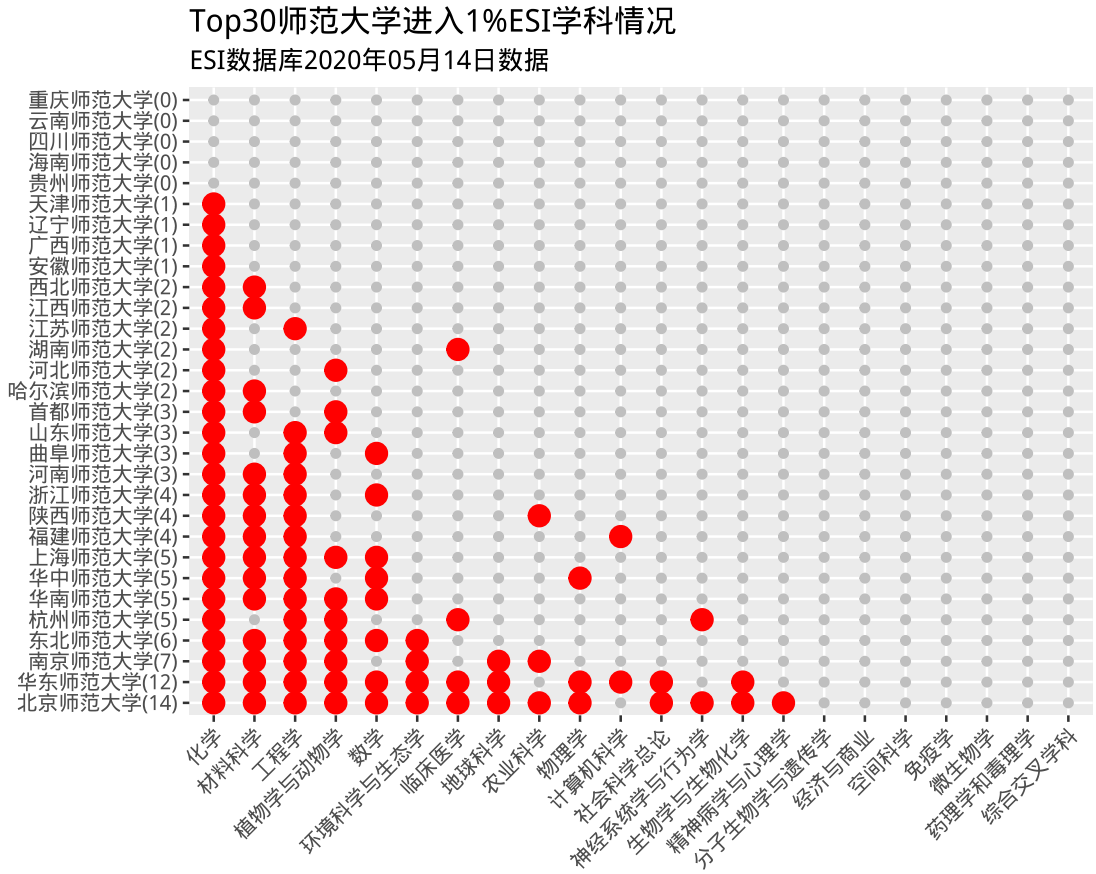
我国师范类高校科研论文的产出情况



这里我们高亮了**北京师范大学**和**四川师范大学**两所高校的发展曲线，灰色背景的其他 28 所高校的发展情况。可见，近几年我国师范类高校科研论文的产出整体上稳步提升，符合科学发展规律。但也明显看到，四川师范大学作为西部高校，与东部发达地区的院校还存在一定的距离。

1.3 师范类院校进入前百分之一 ESI 学科的数量

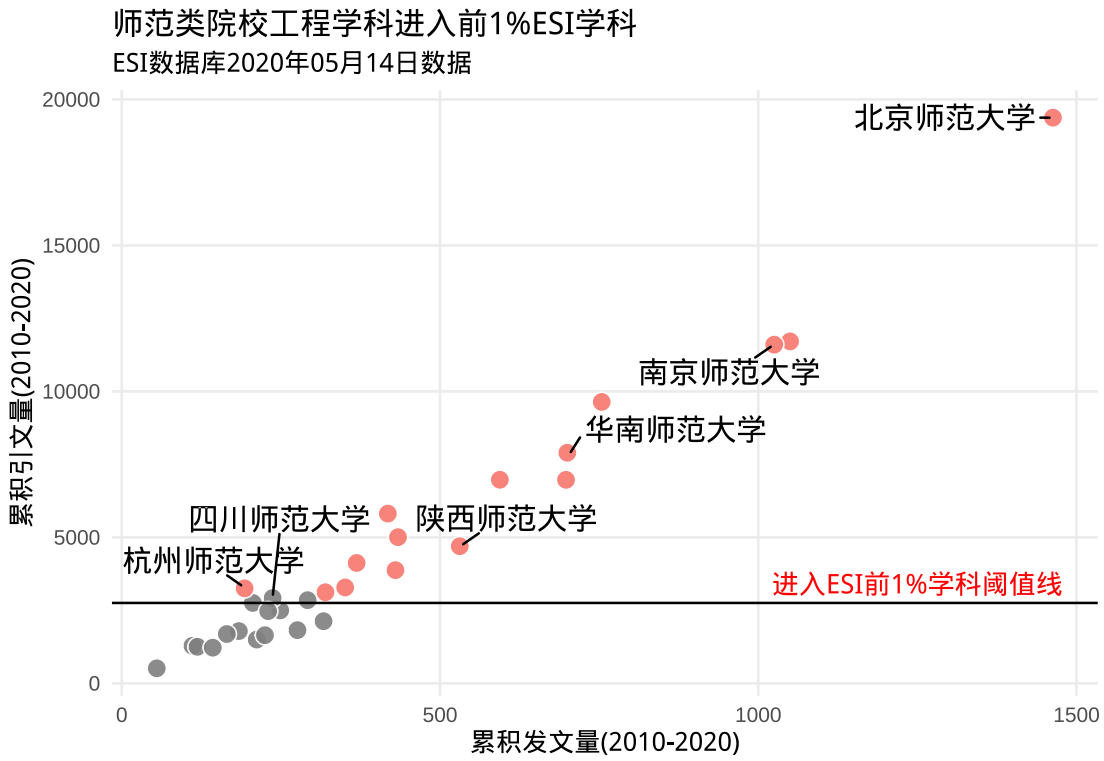
这里我们整理了师范类院校进入前百分之一 ESI 学科的数量。从学校来看，师范类院校有 ESI 学科的 25 所，其中最多的是北京师范大学 14 个学科，华东师范大学 12 个学科，南京师范大学 8 个学科。从入选的学科来看，化学学科、材料学科和工程学入选频次最高。



1.4 师范院校各学科发展态势

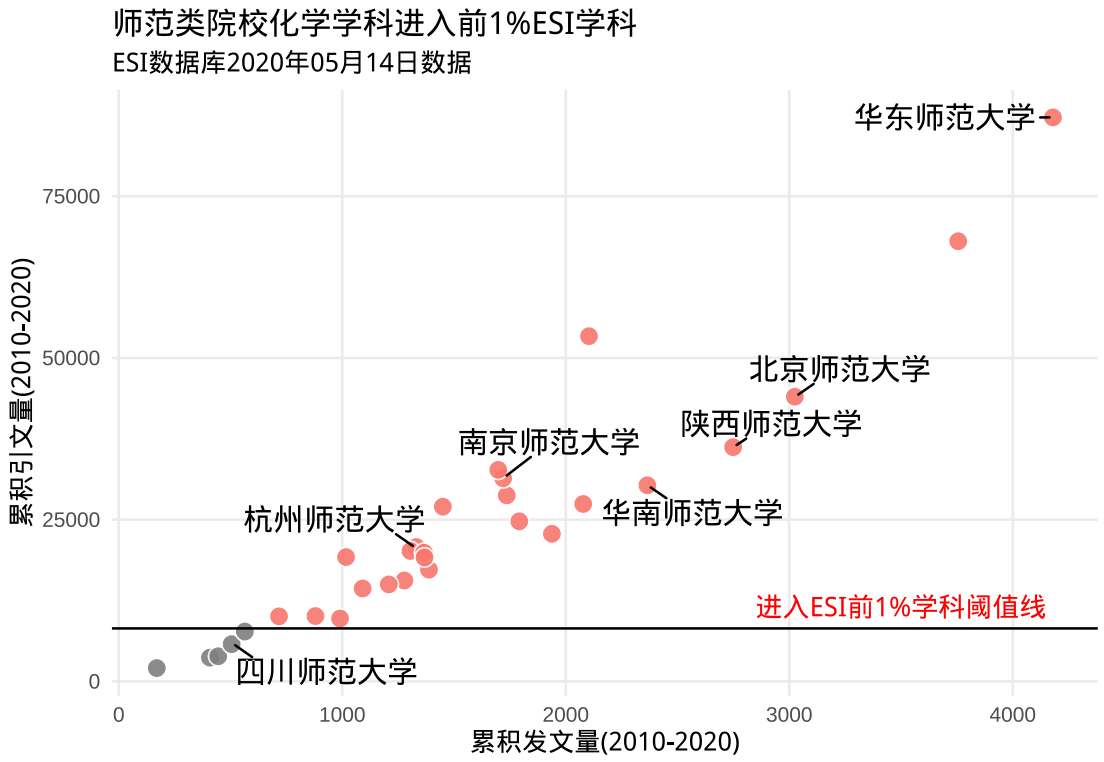
为跟踪学科发展态势，这里我们考察了各师范类高校在两个维度（累计产出和累计影响力）上的科研表现情况，图中红色标注表示该校已经进入前百分之一 ESI 学科，灰色表示还没有进入前百分之一 ESI 学科，由于 ESI 数据库比 SCI 数据库滞后两个月，因此图中阈值线附近的点，会有细微的偏差（可以理解为图中的阈值线会有细微的偏差）。

1.4.1 工程学

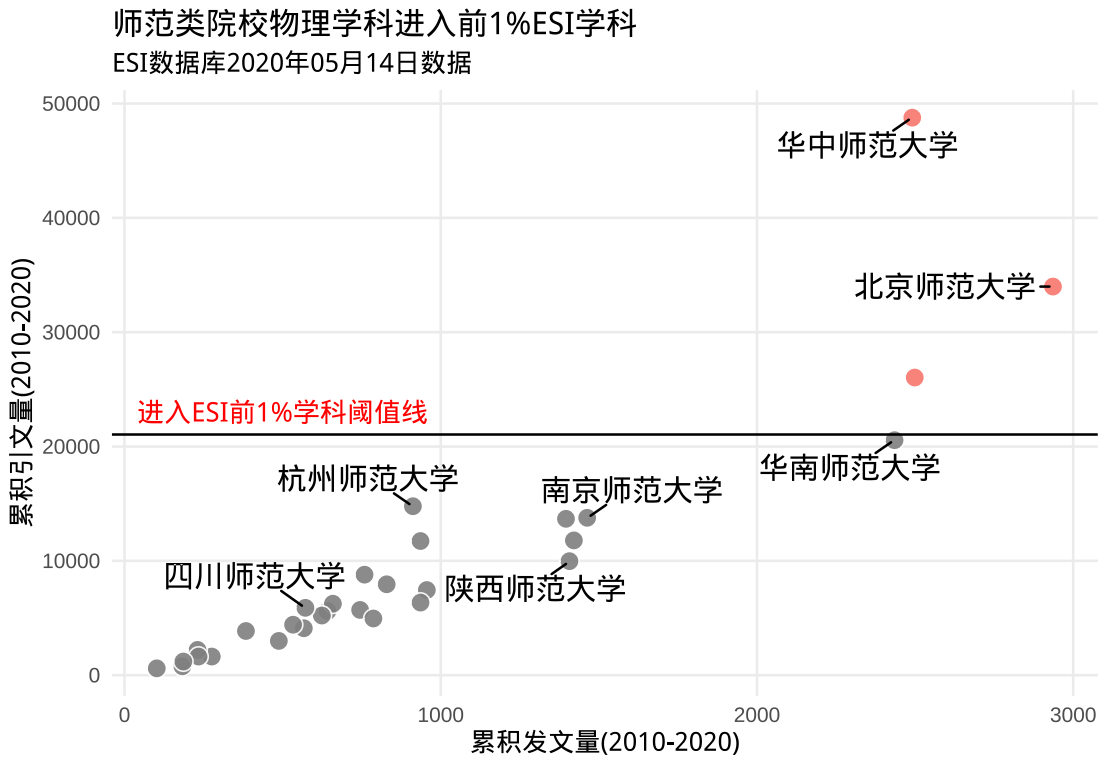


四川师范大学的科研产出超过杭州师范大学，但科研影响力差一点点，因此杭州师范大学率先进入了前百分之一 ESI 学科。

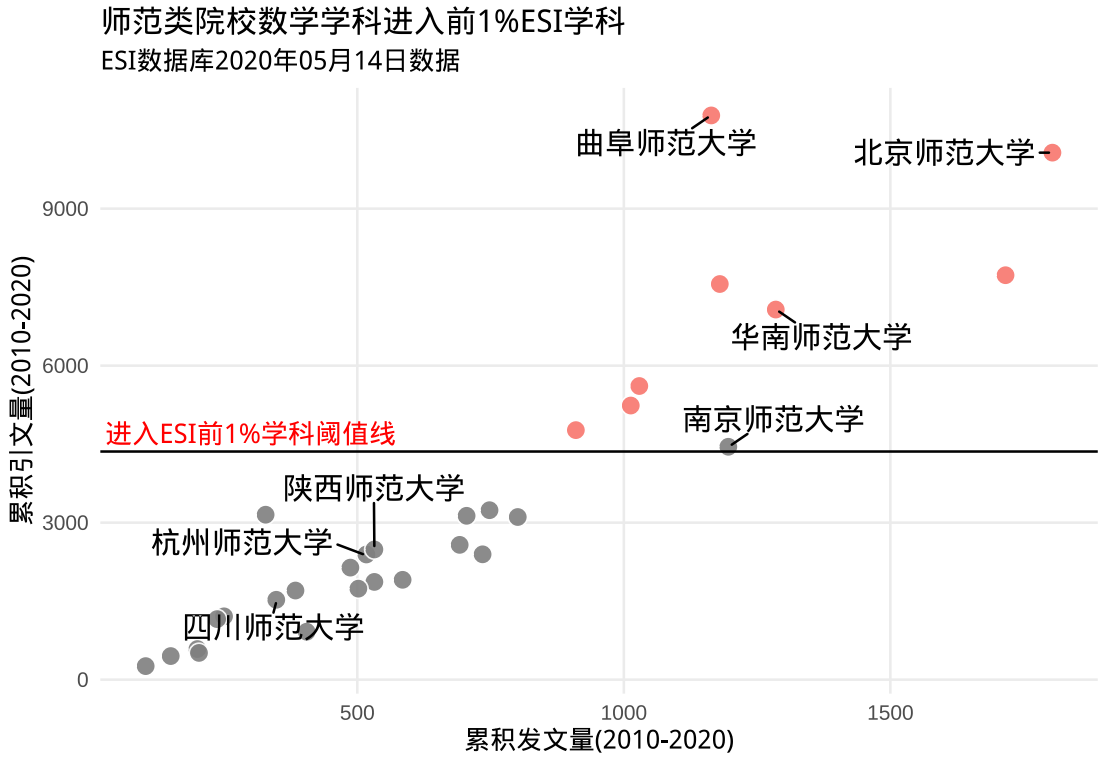
1.4.2 化学



1.4.3 物理学



1.4.4 数学



第 2 章 学科发展

2.1 潜力学科

当前四川师范大学在励精图治奋力耕耘，推动学科发展，其中**工程学**学科与进入 ESI 学科的阈值线最为接近，接近程度达到 97.8%。其他各学科的发展情况见表 2.1 。

表 2.1: 四川师范大学进入 22ESI 学科接近程度

学科	累积论文数	被引频次	阈值线	接近程度
工程学	237	2923	2755	106.10%
化学	505	5752	8188	70.25%
计算机科学	167	2473	3686	67.09%
材料科学	243	2811	6674	42.12%
数学	348	1526	4359	35.01%
物理学	572	5893	21050	28.00%
环境科学与生态学	94	536	4388	12.22%
植物学与动物学	54	336	2881	11.66%
精神病学与心理学	50	364	4077	8.93%
社会科学总论	25	107	1530	6.99%
经济与商业	15	241	4516	5.34%
农业科学	35	85	2361	3.60%
临床医学	16	99	3374	2.93%
神经系统学与行为学	14	150	6426	2.33%
微生物学	19	100	5492	1.82%
地球科学	22	104	6140	1.69%
药理学和毒理学	18	53	3453	1.53%
分子生物学与遗传学	40	215	14132	1.52%
生物学与生物化学	35	93	6316	1.47%
空间科学	2	152	40196	0.38%
综合交叉学科	1	2	2608	0.08%
免疫学	1	0	5149	0.00%

2.2 竞争对手

由表 2.1可以看出**工程学**是入选前百分之一 ESI 学科的潜力学科，但我们也要意识到，当前师范院校高校中，工程学进入 ESI 学科的有 14 所，未进入的 16 所，表 2.2列出了这未进入的 16 所高校的工程学科与阈值线的接近程度，可以看到，大学彼此之间竞争还很激烈。

表 2.2: 工程学科有可能进入 ESI 学科的师范大学

学校	累积论文数	被引频次	阈值线	接近程度
四川师范大学	237	2923	2755	106.10%
重庆师范大学	292	2850	2755	103.45%
广西师范大学	206	2755	2755	100.00%
西北师范大学	249	2500	2755	90.74%
云南师范大学	230	2474	2755	89.80%
湖南师范大学	317	2129	2755	77.28%
首都师范大学	276	1823	2755	66.17%
江西师范大学	184	1791	2755	65.01%
辽宁师范大学	165	1690	2755	61.34%
天津师范大学	225	1650	2755	59.89%
安徽师范大学	212	1501	2755	54.48%
贵州师范大学	111	1288	2755	46.75%
哈尔滨师范大学	119	1253	2755	45.48%
河北师范大学	143	1223	2755	44.39%
海南师范大学	55	515	2755	18.69%

第 3 章 学科预测

在前面一章，我们看到我校的潜力学科是工程学科，有望在 2020 年进入 ESI 的 1% 学科。本章的主要工作是，计算并预测川师工程学科 2020 年进入 ESI 前百分之一学科的概率，以及竞争对手的概率。

3.1 统计方法

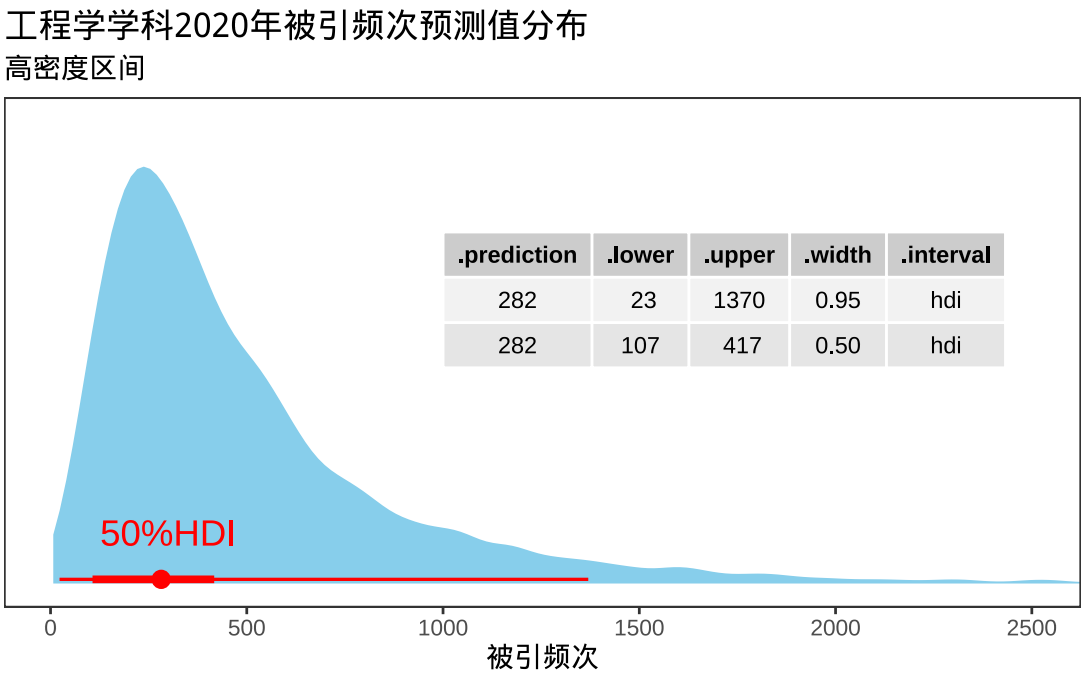
学科的发展与很多方面都有关系，因此建立一个完全正确的预测模型是不可能的。正如英国统计学家 George E. P. Box 所说 “All models are wrong, but some are useful.” 因此我们的模型是错误的，也可能没什么用，但我们依然坚持呈现出来，用图书馆人质朴的方式为我校的发展呐喊助威。

相关研究表明，科研论文被引频次服从负二项分布（具体可见附录），我们建立贝叶斯线性模型，并给定参数的先验概率：

$$
\begin{aligned}
y_i &\sim \text{NegBinomial}(\mu_i, \phi) \\
\log(\mu_i) &= \alpha + \gamma_{j[i]} + \beta x_i \\
\alpha &\sim \text{Normal}(0, 100) \\
\beta &\sim \text{Normal}(0, 10) \\
\gamma &\sim \text{Normal}(0, 2) \\
\phi &\sim \text{HalfCauchy}(0, 2.5)
\end{aligned}
$$

3.2 结果分析

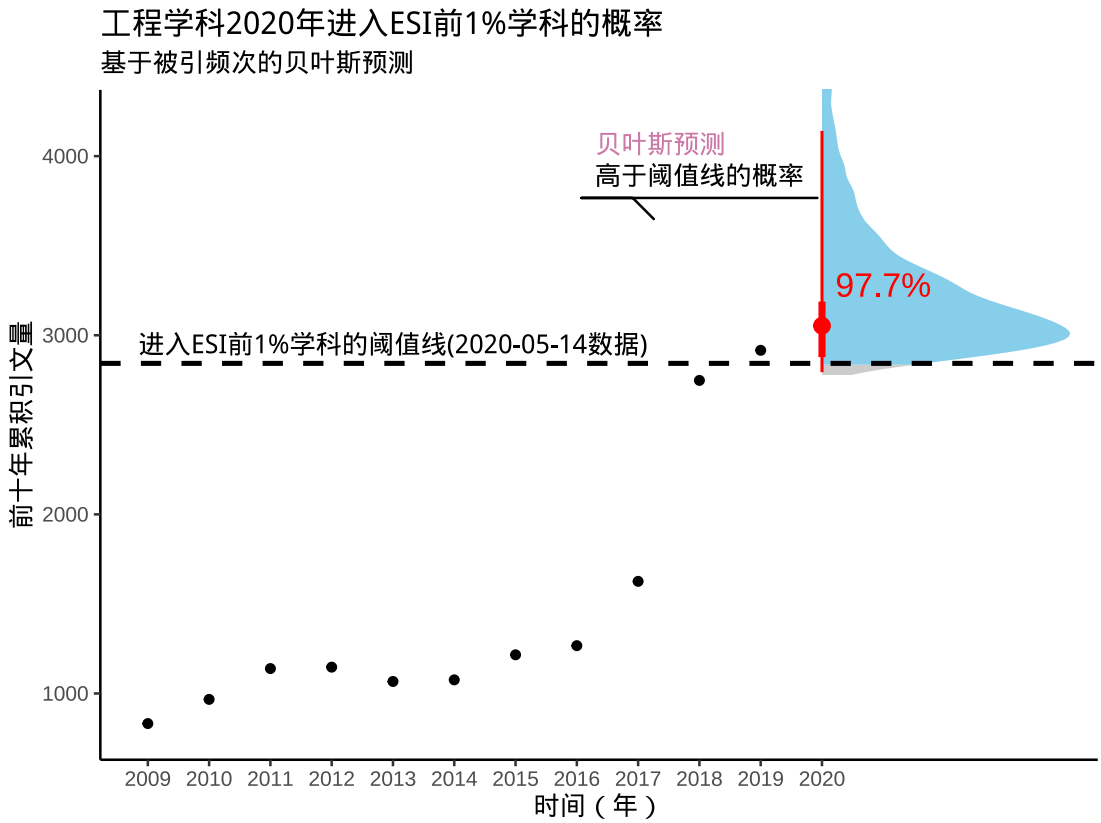
根据模型计算，我们预测了工程学科 2020 年的科研产出量的估计值 282，以及 50% 的可信赖区间 (107, 417)，模型评估见附录。



因此，四川师范大学近十年的累计科研影响力估计值以及分位数区间见下表3.1，在阈值线变化不大或者不变的前提下，2020 年进入 ESI 前百分之一学科的概率将为 79.0%

表 3.1: 四川师范大学工程学累计科研影响力估计值以及 2020 年进入 ESI 前百分之一学科的概率

year	pred_mean	quantile2.5	quantile97.5	prob_above_line
2020	3280	2845	4464	97.7%



3.3 竞争对手的概率

是否进入 ESI 前百分之一学科，取决于这个机构近十年累计被引频次，统计的周期是一个滚动的窗口，我们在预测 2020 年的情况，需要计算 2001 年-2020 年这个时间周期，如果 2010 年的被引频次很高，而 2020 年很低，那么十年为窗口的累计量就下滑，因此当前各学校的接近程度高不代表入选的概率也高。这里，我们采用相同的贝叶斯模型，计算竞争对手的工程学科入选概率.

	univ_cn	year	pred_mean	Q2.5	Q97.5	prob_above_line
1	广西师范大学	2020	2694.	2012	5038.	25.6%
2	华中师范大学	2020	3160.	2543	4686.	68.0%
3	四川师范大学	2020	3168.	2696.	4269.	79.3%
4	重庆师范大学	2020	2957.	2638	3667.	60.5%

我们是以阈值线不变或者变化很小为前提，进行的预测，事实上，阈值线每两个月就会调整一次，尽管我们进入 ESI 学科概率比较大，但也不能掉以轻心。如果需要了解其他学科的预测信息或者对预测模型有不同见解的，非常欢迎与本文作者交流探讨。

附录 A 统计口径

A.1 数据来源

基本科学指标数据库（Essential Science Indicators，简称 ESI）是衡量科学研究绩效、跟踪科学发展趋势的基本分析评价工具，它是基于 Clarivate Analytics 公司（原汤森路透知识产权与科技事业部）Web of Science（SCIE/SSCI）所收录的全球 11000 多种学术期刊的 1000 多万条文献记录而建立的计量分析数据库。目前，ESI 已成为当今世界范围内普遍用以评价高校、学术机构、国家/地区国际学术水平及影响力的重要评价指标工具之一，其数据库以学科分门别类 (共分 22 个学科)，采集面覆盖全球几万乃至十几万家不同研究单位的学科。

A.2 学科分类以及各学科进入 ESI 的阈值

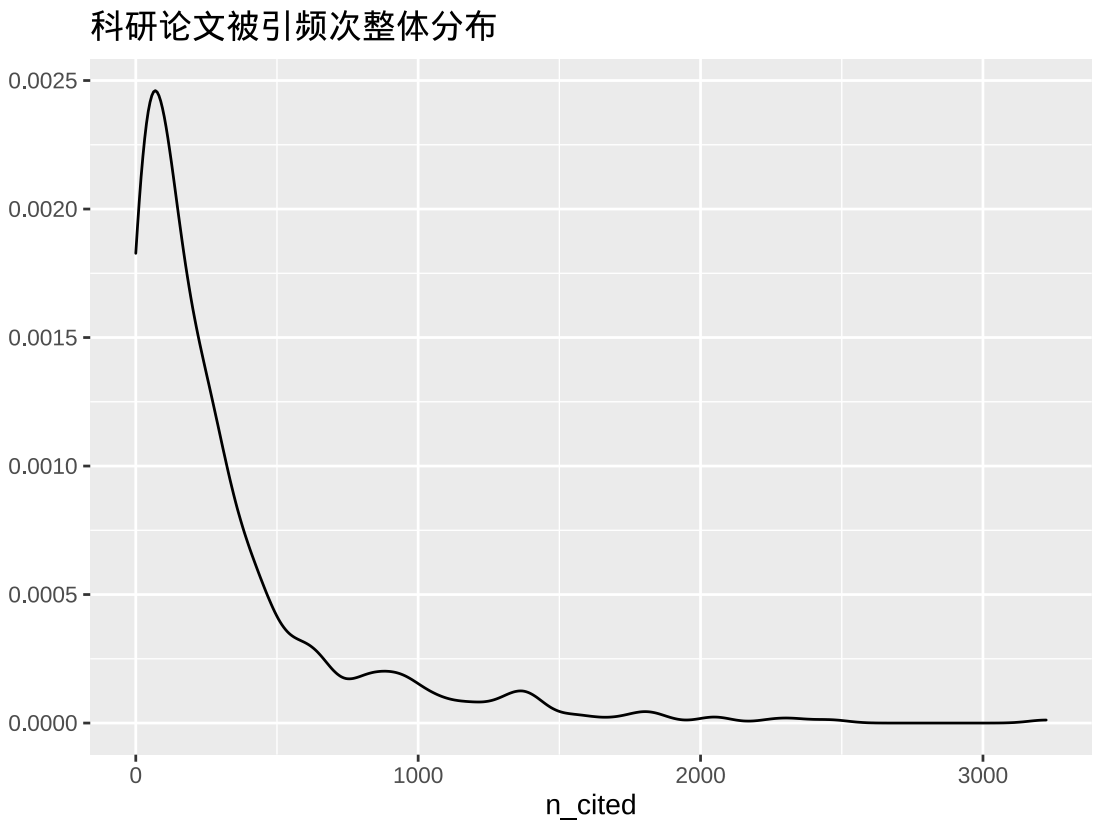
ESI 学科分类一种较为宽泛的学科分类模式。ESI 学科分类模式基于期刊分类，由自然科学与社会科学的 22 个学科构成。艺术与人文期刊没有被包含。每一本期刊只被划分至 22 个 ESI 学科中的一个，没有重叠的学科设置使得分析变得更为简单。被归类为跨学科学科 (Multidisciplinary field) 的 Science、Nature 与 PNAS 期刊，会被按照各篇文章的参考文献 (reference) 与引用文献 (citation)，重新为每篇文章单独分类，但每篇文章仍只会被分类到一个学科。

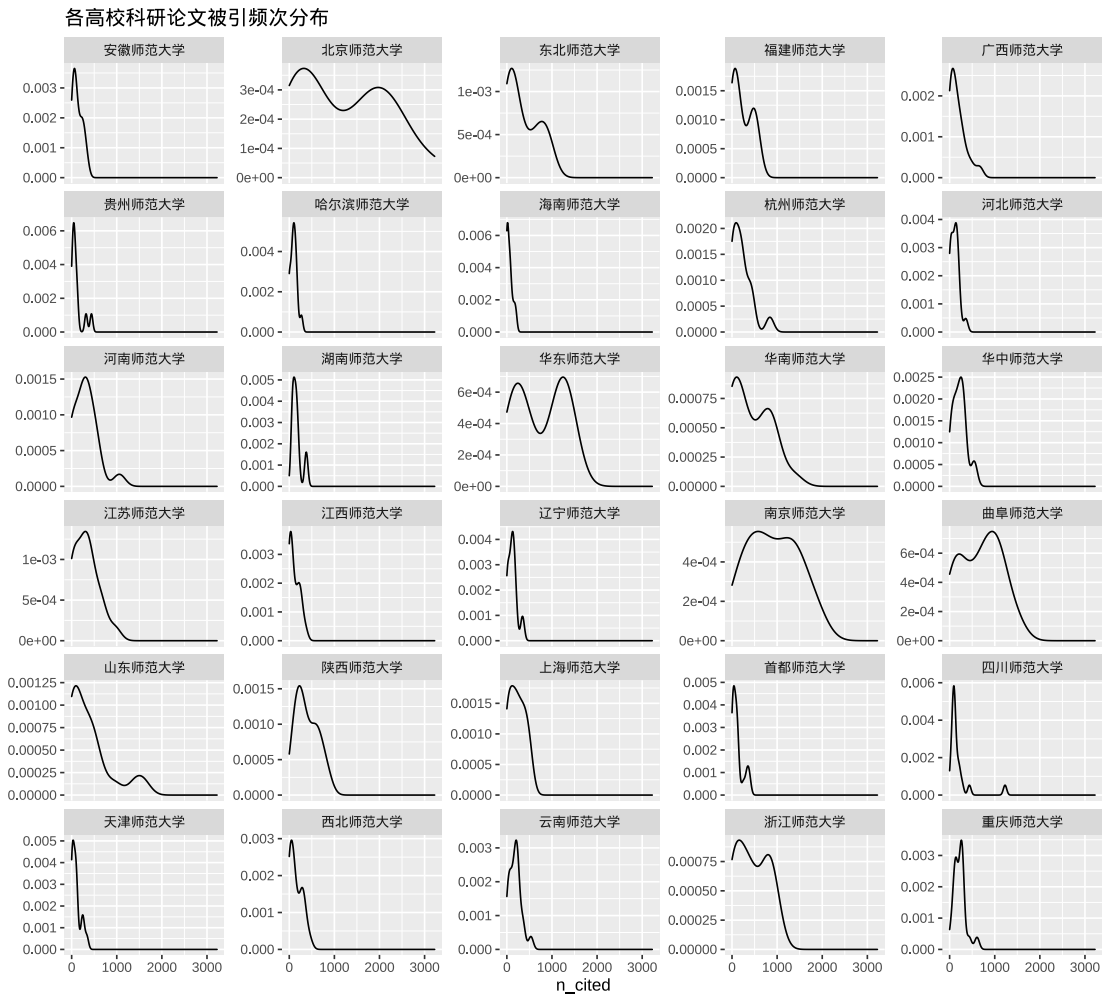
表 A.1: ESI 学科分类以及各学科进入 ESI 的阈值 (2020 年 5 月 14 日数据)

学科类型	学科	阈值 20200514
工学（3）	计算机科学 (Computer Science)	3686
	工程科学 (Engineering)	2755
	材料科学 (Materials Sciences)	6674
生命科学（4）	生物与生化 (Biology & Biochemistry)	6316
	环境 / 生态学 (Environment/Ecology)	4388
	微生物学 (Microbiology)	5492
	分子生物与遗传学 (Molecular Biology & Genetics)	14132
社会科学（2）	一般社会科学 (Social Sciences, General)	1530
	经济与商学 (Economics & Business)	4516
理学（5）	化学 (Chemistry)	8188
	地球科学 (Geosciences)	6140
	数学 (Mathematics)	4359
	物理学 (Physics)	21050
	空间科学 (Space Science)	40196
农学（2）	农业科学 (Agricultural Sciences)	2361
	植物与动物科学 (Plant & Animal Science)	2881
医学（5）	临床医学 (Clinical Medicine)	3374
	免疫学 (Immunology)	5149
	神经科学与行为 (Neuroscience & Behavior)	6426
	药理学与毒物学 (Pharmacology & Toxicology)	3453
	精神病学 / 心理学 (Psychology/Psychiatry)	4077
其他（1）	多学科 (Multidisciplinary)	2608

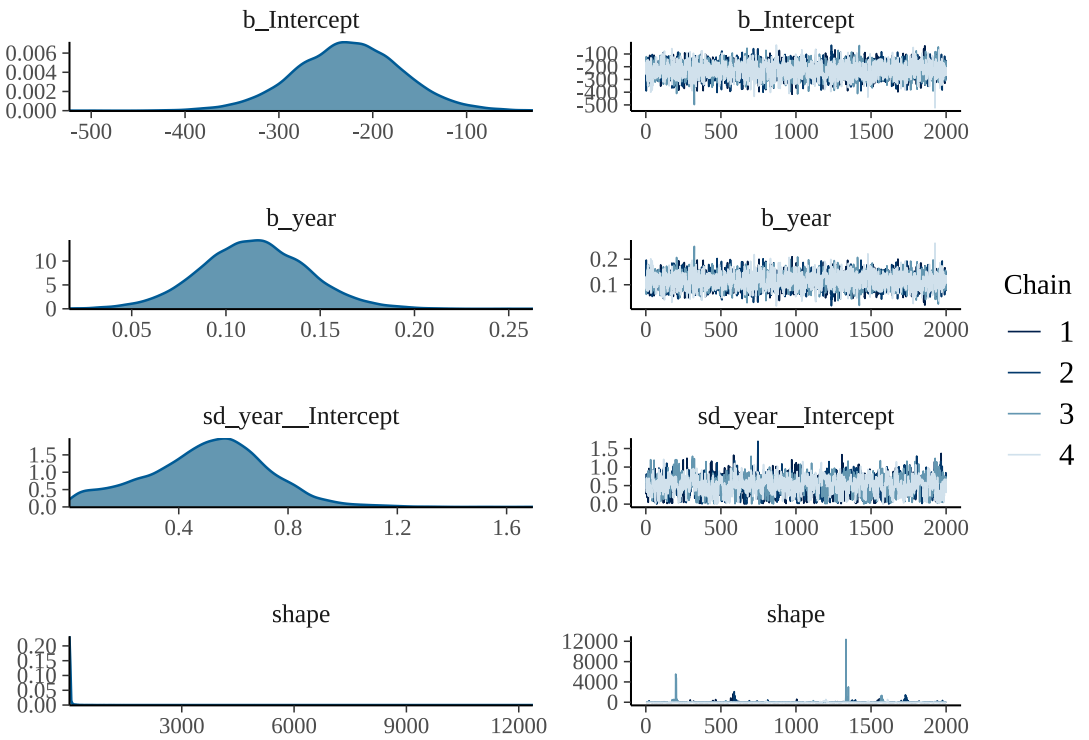
A.3 贝叶斯模型参数

A.3.1 被引频次为什么是负二项分布

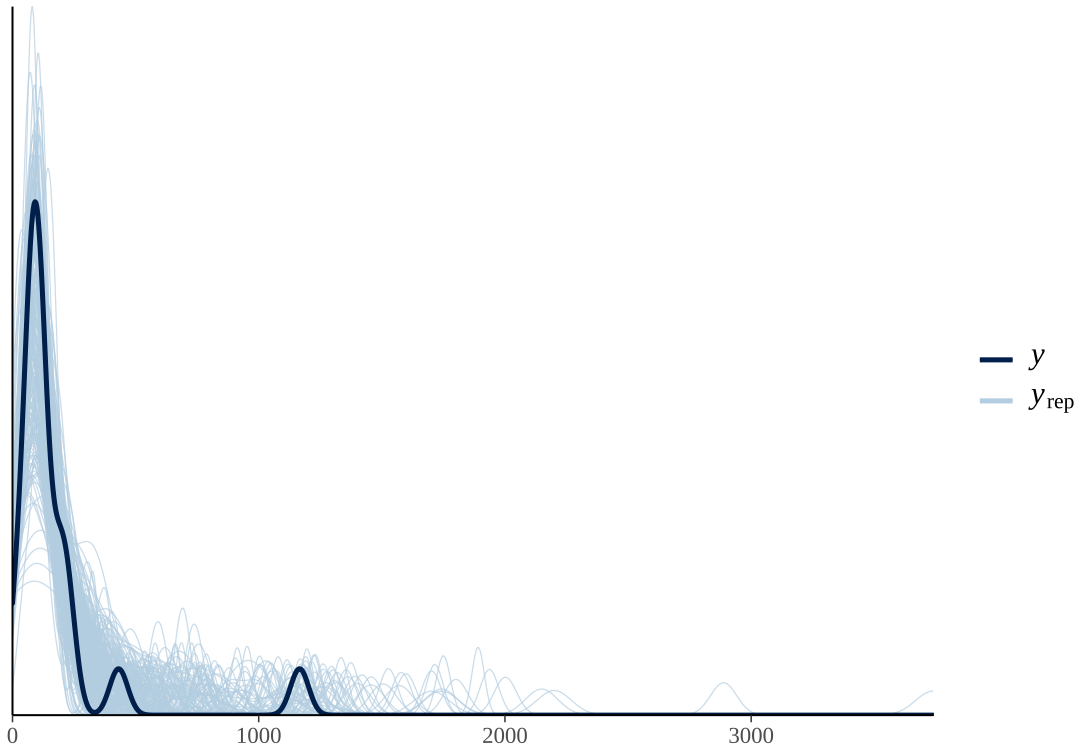



各高校科研论文被引频次分布
安徽师范大学
北京师范大学
东北师范大学
福建师范大学
广西师范大学
贵州师范大学
哈尔滨师范大学
海南师范大学
杭州师范大学
河北师范大学
河南师范大学
湖南师范大学
华东师范大学
华南师范大学
华中师范大学
江苏师范大学
江西师范大学
辽宁师范大学
南京师范大学
曲阜师范大学
山东师范大学
陕西师范大学
上海师范大学
首都师范大学
四川师范大学
天津师范大学
西北师范大学
云南师范大学
浙江师范大学
重庆师范大学
n_cited

A.3.2 后验概率分布



A.3.3 后验概率检验



A.3.4 ESI 数据完全不透明

5 月更新的 ESI 数据库收录论文的时间范围是 2010 年——2020 年 2 月底（十年零两个月）；

- 我们只能检索到年
  - 比如，7 月份发布时，ESI 数据库收录论文的时间范围是 2010 年 2 月——2020 年 4 月
  - 我们能检索的 2010 年——2020 年 7 月
- 在 ESI 检索到的数据，他们还要再筛查一次。
  - ESI 工程学和计算机这两个学科的引用有不少是来自会议论文的，但是 ESI 不统计来自会议论文的引用，所以实际表现没有您检索出的结果那么高。

表 A.2: 师范高校工程学科

univ_cn	cum_paper	cum_cited	web_of_science	cites	top_papers	is_enter	Threshold0514
安徽师范大学	212	1501	NA	NA	NA	NA	2755
北京师范大学	1463	19377	1408	17751	29	TRUE	2755
首都师范大学	276	1823	NA	NA	NA	NA	2755
华中师范大学	320	3121	312	2765	9	TRUE	2755
重庆师范大学	292	2850	NA	NA	NA	NA	2755
华东师范大学	1050	11714	1014	10104	21	TRUE	2755
福建师范大学	369	4124	362	3798	11	TRUE	2755
广西师范大学	206	2755	NA	NA	NA	NA	2755
贵州师范大学	111	1288	NA	NA	NA	NA	2755
海南师范大学	55	515	NA	NA	NA	NA	2755
杭州师范大学	193	3255	190	2939	6	TRUE	2755
哈尔滨师范大学	119	1253	NA	NA	NA	NA	2755
河北师范大学	143	1223	NA	NA	NA	NA	2755
河南师范大学	430	3878	418	3633	10	TRUE	2755
湖南师范大学	317	2129	NA	NA	NA	NA	2755
江苏师范大学	434	5008	428	4419	10	TRUE	2755
江西师范大学	184	1791	NA	NA	NA	NA	2755
辽宁师范大学	165	1690	NA	NA	NA	NA	2755
南京师范大学	1025	11602	998	10055	25	TRUE	2755
东北师范大学	418	5815	405	5201	7	TRUE	2755
西北师范大学	249	2500	NA	NA	NA	NA	2755
曲阜师范大学	754	9640	739	8403	28	TRUE	2755
陕西师范大学	531	4698	521	4195	10	TRUE	2755
山东师范大学	698	6972	677	6418	23	TRUE	2755
上海师范大学	351	3285	343	3008	5	TRUE	2755
四川师范大学	237	2923	NA	NA	NA	NA	2755
华南师范大学	700	7900	683	7241	16	TRUE	2755
天津师范大学	225	1650	NA	NA	NA	NA	2755
云南师范大学	230	2474	NA	NA	NA	NA	2755
浙江师范大学	594	6974	584	6285	18	TRUE	2755

参考文献

[1] R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[2] Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4

[3] Wickham, H., Grolemund, G. (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media. ISBN: 1491910399

[4] Xie Y, Allaire J, Grolemund G (2018). R Markdown: The Definitive Guide. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 9781138359338

[5] Gelman, A., Lee, D., & Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. Journal of Educational and Behavioral Statistics, 40(5), 530–543.

[6] Bürkner P (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1–28.

[7] Kay M (2020). tidybayes: Tidy Data and Geoms for Bayesian Models. doi: 10.5281/zenodo.1308151, R package version 2.0.3

[8] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin (2014). Bayesian Data Analysis. 3rd ed. Chapman and Hall/CRC.

[9] Kruschke, J. K. (2014). Doing Bayesian data analysis : a tutorial with R and BUGS. Burlington, MA: Academic Press.

[10] McElreath, R. (2015). *Statistical rethinking: A Bayesian course with examples in R and Stan.* Chapman & Hall/CRC Press.

[11] Kruschke, J. K. & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychonomic Bulletin & Review, 25:178-206.