江霄骐

Shaw-J1029 15722633987@163.com

教育背景

浙江大学 **杭州,中国 2021.09-2025.07**

主修：物理学士

辅修：金融

- GPA：4.31 / 5.0 (88.1/100)
- 荣誉：浙江大学一等奖学金；其他荣誉（优秀学生等）
- 主要课程：高等代数、数学分析、计量经济学、概率论与数理统计、金融工程等。下学期将学习量化投资、数据挖掘与可视化、数据结构等。自学机器学习、深度学习(CS229和CS230)。

实习&经历

国盛证券|有色金属团队分析师助理 **（实习）上海，中国 2024.02-2024.07**

工作时间：10:00-18:00，官网链接：https://www.gszq.com/home

- 使用Python（GPT帮助）爬取智利铝出口数据，统一口径，消除手动下载的遗漏和格式不确定性
- 使用Excel进行数据清洗，如在大量的海关进出口数据中通过“ISNUMBER”“VLOOKUP”等函数进行数据清洗
- 通过Excel分析铝价市场波动和上游铝价加加工费的下游定价机制，估计目标公司（南山铝业）前景
- 结合全球汽车销量及预期、商用飞机龙头企业销量及预期，评估目标公司（南山铝业）的未来利润增长
- 协助撰写目标公司的深度报告，个人完成“公司概要”“主营业务数据解读”“财务分析”“行业纵览”等模块，并相应注入了个人的见解，约5000字
- 利用Excel的VBA撰写常用的宏模块，报表从wind和ifind中自动刷新每周金属价格数据
- 在策略会期间，查看目标公司报表，撰写提问提纲（如公司的销售经营情况、公司上下游生产项目推进情况），询问目标公司关于业务的情况并总结会议纪要，每个公司有所不同，从500-2500字不等

安永企业咨询有限公司|高级顾问助理 **（经历）北京,中国 2023.07-2023.09**

工作时间：10:00-18:00 官网链接：https://www.ey.com/zh_cn

涉及多个项目

【国投泰康】

- 提取会议记录，获取关键信息以记录项目的方向和进度，从而识别实际操作中的变化
- 根据现有信息检索、调整和修改技术关键词。补充和完善现状报告，添加数据人才选拔模版，从内部人才培养和外部人才引入两方面构建蓝图
- 将重要的访谈纪要重组后形成Powerpoint汇报材料，如将数据人才选拔分点概括并形成图表和文字描述
- 用Excel构建数据质量管理细则事项和流程，并用Powerpoint可视化成流程图，添加进中期汇报材料
- 为挖掘甲方重点关注内容，利用短期内的多份会议纪要和记录总共约10000字，合并加入txt文件后，使用Python的jieba模块进行中文分词，利用NLP的相关模块ngram进行高频词检索（权衡后未使用NLTK，因为ngram可以连续提取多个单词形成中文词组，而如“数据报告”等多词词组也可能在此之列），并使用word cloud模块，丢弃常用连词如“的”“有无”等后形成词云的分布和可视化图片，加入调研评估与分析报告
- 参与该项目的规划收尾工作，核验材料完整度和信息一致性

【北京银行】

*项目处于招标阶段，最高招标价格为80万元。**招标目标**为：挖掘去时点积累的全行支付流水数据，建立资产负债部、资金运营中心、信息技术部无边界敏捷团队，大数据人工智能研究团队研究理论，建立模型并部署应用。实现对日内支付规律挖掘，算法分析人工预报历史错误，主动提示，降低操作风险，形成有数据、有理论的研究报告；实现对未来时点的头寸预测，为制定各个业务板块流动性限额提供决策支持同时与市场交易结合，占据前瞻性性先机提高资金收益，降低负债成本，形成可落地、可持续应用的预测模型，并形成设计文档*

- 检索AI头寸预测模型。查阅各种AI模型在头寸预测中的历史使用情况。结合自身的AI知识储备，如LSTM、贝叶斯模型等，有针对性地补充检索信息，并阅读一些投研报告，添加Wavenet、Lasso回归等模型信息，以建立投标文件

- 提供搜索手稿的详细概述，协助填补遗漏部分。纠正错误部分
- 协助撰写技术标书中的LSTM部分原理，引用模型架构图片，增加标书技术可信度、竞争力，提高中标成功率，展现AI+银行头寸的交叉理解

浙商银行|投资银行部助理 （经历）苏州，中国 2022.07-2022.08

工作时间：8:30-18:00 官网链接：https://www.czbank.com/cn/index.shtml

- 了解NAFMII在债券承销领域的内部约束要求。整理和整合手稿材料
- 参与撰写募集说明书，多方核对主要信息并补充募集信息；利用外部征信网站了解通过发行债券筹集资金的过程
- 通过检查信用评级、簿记等，了解债券投资的估值原则（实际估值中，并依赖折现现金流模型，而是综合市场供需、信用评级、投资者情绪给出的综合报价），从DM-Lite中获取相关信息进行学习
- 参与银行债权投资白名单申报表的准备工作，以验证内部信息。学习使用DM-Lite和Wind

研究&实验

ΛCDM 宇宙中的大规模平面结构探测 （进行中）杭州，中国 2023.11-2024.08

Renyue Cen 教授指导，课题来自 P.J.E Peebles 教授

使用 Numpy 和 Cupy 在多个 GPU/DCU 上的并行计算加速计算。最终目标是确认ΛCDM 模拟中关于大规模平面结构在观测世界中的合理性

- 加载质量大于1e12太阳质量的Uchuu星系进行结构化盘探测（包括数据清理和从h5py到npy文件的转换等）
- 搭建GPU平台，建立数据化模型，全覆盖向量化计算
- 连接服务器并进行Cupy的GPU加速，使用npy文件作为结果进行迭代探测

挑战与解决

- 前期项目的最大问题是建模。在前两个项目经历中，我遇到了同样的问题，但相比第一次使用python，这次建模显得没有那样仓促。这个项目是我所做最大的项目，其中包括源数据获取、数据建模、远程平台搭建、GPU配置、可视化等多个模块。同样，在该项目之前，我对GPU加速是完全不熟悉的，因此该项目的运行依赖于远程平台的搭建。IDP聚类的项目数据量较小，并且没有高额的计算成本，因此在pc上就可以自助完成，仅利用pip就可以完成所有工作，但该项目的计算成本极高，模型很大，简单的pc平台和cpu计算已经无法在有意义的时间范围内完成，因此只能在试错中搭建和学习模型搭建GPU模型。
- 该项目是我的个人项目，因此没有讨论和交流的空间，而导师所提供的均为数理逻辑上的指导，无法在技术层面帮到我，因此简单至conda环境搭建、Numpy并行化的进一步学习，难至模型搭建、Cuda和Cupy配置，均是个人学习和摸索完成的。我的模型来源于P.J.E Peebles的*Flat Patterns in Cosmic Structure*，模型的总体搭建步骤并不复杂，但在并行计算优化上却有很大的难度：我从Uchuu模拟中获得的数据集是宇宙的snapshot暗物质晕坐标，在低红移处数据量很大；同时，由于一开始并未接触到远程平台，所以我使用的模型仍然是基于Numpy的模型范本，并计划在通过调试后迁移至Cupy和GPU进行计算；如果将模型全部置于并行化，内存很容易溢出，CPU无法正常计算。因此，我估计了计算空间成本，尝试将一部分计算分配到for循环中，在单次循环中仅使用200条样本数据进行计算（在计算中，由于仍然使用基于CPU的Numpy计算，我需要将原有数据降低维度来压缩调试的时间成本，这就要求我构建同样格式、但数据量小的多的调试数据样本；为做到这一点，我从原始数据中进行了随机采样，得到了原数据约千分之一的样本容量）。该模型在计算上与深度学习类似：当我将数据和计算向量化以后，发现模型具有非常大的矩阵运算要求，并且对数据格式非常敏感，因此我设计了每一个计算步骤，将所有的计算逻辑都清晰地呈现出来。在这个模块，还没有写使用文稿，或许在后期会补上一个README。
- 模型搭建花费的时间成本很长，从单步的for循环，一直到Numpy覆盖的全向量化，我经过了多次试错，并最终得到了模型的初版；然后，我自学了一部分OOP的内容，将我的模型进行了封装。整个过程大概经历了4-5个月。
- 在pc上完成Numpy的封装后，为对原始数据进行有效地计算，我需要将代码和模型迁移到GPU平台。该平台是紫金山天文台的Starburst，具有4块40G的A100，在一个月内进行多次计算已经足够。在这个阶段，我又遇到了一些学习壁垒：由于并不了解Linux的终端命令，我花大概1周时间初步了解了Linux终端的命令逻辑；在GPU平台上，使用Slurm作为作业提交系统，该程序管理模式我并不熟悉，因此摸索作业提交系统我也花了一些时间，大概1-2周；搞清楚这些问题以后，我发现Cupy的安装有一些问题：由于Starburst的Cuda版本较老为cuda11.8，我无法适配地安装Cupy，而对Cupy降级后，我又发现Cupy无法支持当前版本的Numpy，因此我遇到了经典的环境适配和版本适配问题。通过与平台管理人员的交流，我大概完成了该版本适配，并尝试在平台上调试成功。
- 在Numpy向Cupy的迁移过程当中，由于平台有超过一块显卡，我自主学习ProcessPoolExecutor，并在模块教程下完成了多卡协同的配置。这样，我完成了对显卡的配置学习，并能够在不主动调用Cuda的情况下使用Cupy进行并行计算。

- 目前，我已经完成了对红移=0.021，最低暗物质晕质量为10^12Msun、1.5*10^12Msun、1.55*10^12Msun的源数据的计算处理，并初步得到了一些结论：Peebles认为的平面结构，在LCDM模型中，确实是存在的，但稳健型仍需进一步计算得知，这是说，在宇宙的不同位置随机取点，并通过统计分布查看这些情况出现的概率，如果在地球上出现的观测结果在宇宙的任意一点都有稳健的概率能够观测到，那么就验证了LambdaCDM在真实宇宙中的该项观测模拟是具有解释力的。这是我将要做的工作。
- 当前，我已经处于项目的中后期，回过头来看这整个项目，确实应该有很强的关联逻辑：
  1、为初步建模验证和计算，应在CPU上搭建Numpy的向量化模型，并缩限原始数据大小以更时间经济地估计和调试模型效果
  2、为迁移模型，应做好远程平台的环境配置，包括conda环境搭建、Cuda和GPU配置、Slurm和Sbatch命令脚本撰写
  3、为有效计算，防止内存溢出，需要控制单次的计算量，通过多次调试完成。

感想

- 从该个人项目中，我学到多种实用技巧，包括numpy、cupy的数据转化和模型搭建、ssh远程平台的使用、GPU的配置等。一个支撑我坚持做下去该项目的理由是我的导师，即便我连续一两周没有什么项目进展，也会坚持给我发邮件询问项目开展情况，并提供一些有益的帮助；在8.12日，他还将与Peebles会面，并就我的项目进行一些探讨。我认为，在整个过程当中，持续的推动力是获得新知识的渴望以及项目进展的迫切想法。与此同时，在一个新项目开展的进程当中，我也学习到很有必要进行目标规划和实现方法梳理，这些东西能够帮助自己减少无谓的时间浪费，从而更高效地推动项目进展。

IDP 蛋白聚类研究——FUS 蛋白的 RGG 结构域探索　　杭州，中国 2023.03-2023.06

以机器学习为工具，使用 GMM 和 SC 方法对 FUS 蛋白的 RGG 结构域数据进行聚类，挖掘该蛋白在无定型结构下的一些显著特征，从而简化该 IDP 蛋白的结构空间，利于进一步的功能-结构探究。

- 量化IDP接触数据以获得接触矩阵
- 使用Python进行聚类、降维和数据可视化（ML）
- 使用Silhouette评估结果，GMM聚类的Silhouette值达到0.17，认为聚类有效
- 根据数据的凸性提出改进方案

挑战

- 第一个难点在于从头开始的代码学习。在此之前，我从未使用过python作为真实的建模工具，甚至对数据读取、数据评估等也没有了解，所以项目的前中期都在学习代码语言。学习的初始过程很艰难，因为没有系统的课程培训，而带教老师也无法抽出时间从头讲起。所以实际上我是通过不断的自我迭代和同学之间的互补交流完成补缺的。我会每周和同组同学一起学习python的用法以及聚类的基本思想和实现方式，这大概持续了5-6周。
- 第二个难点是接触矩阵，即聚类样本的获取。我们一开始得到的数据是来自大四学长的实验数据，这些数据比较混乱，无法作为输入。我和同组同学多次找到学长进行数据校验和核对，并最终确定了原始数据的格式是71*71的二阶张量。这也就进一步限定了IDP蛋白的长度，否则，该张量将会有其他的尺寸，如果仅用简单的机器学习方法，就无法进行有意义的聚类。在此期间，我们也意识到对原始数据的清洗和格式化是非常必要的，因为对于特定的算法，我们只接收格式化样本。
- 组内的三人有一位几乎没有时间投入该项目，因此我们将主要工作放在了我和另一位同学身上，将pre任务交给了该同学；但同时，我们更新我们的工作进度，他可以充分得知我们的进度：每隔两周，我们会约他出来讲解我们的解决思路。
- 正式聚类阶段的工作并不难，了解GMM和SC的工作原理后，对相应的聚类效果做评估，并各自得到聚类结果即可。由于GMM的假设中，需要数据集为完全的凸集，所以我们另外引入了谱聚类作为对照，检查聚类的效果。结果评估可以完全交给Silhouette，但由于数据处于较高维度，想要直观地查看聚类效果并不充分，所以我们使用PCA和t-SNE两种方法将71维的数据降维到2维以可视化。这两种方法我们也根据它们的实现方式分别进行了评估，最终我们一致认为t-SNE的可视化效果较好。
- 由于GMM和SC均需要通过手动选取k值来决定聚类，我们在选择k值上有疑惑。我们尝试使用了几个离散的k值进行聚类，并评估效果；我们和学长进行了进一步讨论，确定了无监督学习并不是要得到一些确定的标签，而是想查看和简化IDP的构型空间。因此，我们为查看该空间，权衡了轮廓系数（Silhouette）的评估，选取25作为k值，并以此认为RGG在时域中有较稳健的25个显著构型，这极大的降低了构型空间的参数，有利于对该蛋白的结构-功能特征进行进一步的探究。
- 在这个项目中，极大难点是编程语言，我通过自我学习和同学互助讨论克服了这一困难。这一经历也使我学会了一些有用的模块，如Numpy、Sklearn等，这在未来的数据建模和一些其他的课程中有很大的帮助。同时，我也通过多次交流、答疑等环节，降低了我们和实际工作之间的认知壁垒，使得工作有效地进行。

住房价格、人均产值和地价的建模　　杭州，中国 2022.05-2022.06

使用 Stata 来探究、挖掘三者的线性关系，并利用计量经济学的经典假设检验对稳健性进行检测

- 获取三方的截面数据，并进行简单线性回归、t检验和F检验

- 分别进行共线性、异方差性和内生性检验，并转化模型至最优

挑战

- 该项目的主要前期问题是模型构想：在现实生活中，很少能直接找到少数几个满足简单线性关系的经济学指标：在经济学中，多因子被一般认可，并且共线性的限制会阻止我们选取高度相关的指标做回归；同时，即便有单指标依赖，许多依赖关系是非线性的。这就要求从统计指标中寻找直接相关性更弱的。选择地价和住房价格的线性关系假设是自然的，因为从中国开发商建设的商品住房定价方式来看，土地价格+建筑价格=房价是基本定价方式；选择人均产值与住房价格则是将抽象的需求关系转化过来，并浅显地假设该关系是线性的。
- 项目的中后期过程并不复杂，仅需分别对三者进行线性回归，并加入工具变量城市化率。值得一提的是，由于中国的数据范围较大，我所采用的是以省级行政区划为类别的2021年截面数据。该数据规模量较小，但用于呈现全国内的截面数据特征已经足够。尽管OLS模型的拟合效果较好，Adj R_squared 达到0.923，认为线性性较强，但White检验下chi2(5)的p值过小，无法拒绝异方差的问题；为修正该问题，我引入了二次项做二次回归分析，引入(x_2^2)，得到了非线性效应的显著结果，这说明我在之前所引入的“人均产值与住房价格线性相关”假设是错误的，而平方相关则是更好的拟合手段；然后，我同时引入了加权最小二乘法进行统计修正，解决了异方差问题。另外，为解决内生性问题，我引入了各省的城市化率作为工具变量并进行豪斯曼检验，F检验的p值为0.952，这表明工具变量选择正确，内生性问题得到有效解决。这样，就可以合理地认为加权最小二乘法的结果具有较高的解释力具有较好的稳健性。
- 在计量经济学的假设检验实际运转中，统计学的思路非常明晰，在进行回归分析的每一步，都应该敏锐地观察模型的问题；从此经历中，我也深刻认识了一些重要的统计学工具以及计量软件的用法。实际上，二次回归的扩展与我将来接触到统计学习中的SVM有一些相似，都是引入非线性来更全面地加固算法拟合度，这些知识的连贯性是较强的，这也使我在学习和实验过程中时不时回过头来查看自己是否能在不同的学科知识背景下获取某些联系，这的确有利于我加深印象和理解。

社会工作

浙江大学综合媒体中心工作人员 **杭州，中国 2021.09-2022.03**

- 规划和确定学校官方账号推文的形式和内容（按照先前分配的组别轮岗，每周更新；推文点击量均约300+）
- 参与文章诗歌撰写并以小组专题发表《韶光丨晖光半去，薄日将眠》 -浙江大学学生会，阅读量约1100
- 参与中期摄影并提供修改建议，在过程中对稿件的内容和主题进行及时修改并发表，如《ZJU丨125周年，听我轻轻说——521! 》 -浙江大学学生会，阅读量约2100；《新年倒计时丨水光潋滟，共祝良辰》 -浙江大学学生会，阅读量约1200；《韶光丨新桃换旧符》 -浙江大学学生会，阅读量约600
- 参与浙江大学学生会两会感想推文撰写和修改，对《时代的声音丨浙江大学学生会热议2022全国两会》文案完成审核与局部修改，参与格式排版和发表，阅读量约1200

浙江大学民族乐团二胡演奏员 **杭州，中国 2021.09-2023.09**

技能

语言

- CET6 617;
- TOFEL 104;
- GRE 327

计算机技能

- Python(熟练掌握Numpy和Cupy);
- Linux终端使用(主要是Sbatch和Slurm脚本递交);
- Mac和Windows终端使用;
- Excel (数据清理); Slides/PowerPoint(熟练掌握);
- Markdown(熟练掌握Typora，主要用于文档撰写)

运动；音乐

- 篮球；围棋；二胡（十级）