**第1页**

伦敦大学学院统计科学系

理学硕士项目指南

统计2018–2019



**第2页**

**理学硕士项目安排2019**

学生应该计划在考试后短暂休息，然后再开始学习
项目。 在六月期间，所有监管人员可能会不时地离开 –
九月，参加会议或度假。 因此，学生应该看到他们的主管
***一旦***考试结束，就可以做出相互方便的安排
他们的项目。

2.在项目工作期间，学生和主管应安排定期会面（约一次
一周，尽可能）并且应该就完成项目工作达成合适的时间表
并生成一个书面帐户。

3.鼓励学生使用 **LaTeX**作为书面报告。 这是一个功能强大的程序
制作技术文件，非常值得学习 – 请参阅 STAT0034 Moodle
页 。

4.学生应将项目报告的***草稿***版本提交给主管以征求意见
**截至2019年8月9** 日 星期五 。

5.最后报告应在 **30** 日**星期五下午4点** 之前输入部门办公室
**2019年8月** 。 还应在Moodle的指定区域提交电子版
**2019年8月30** 日**星期五下午4点之前** 的硕士课程课程页面 。

6.硬拷贝和电子版必须相同。 逾期提交将导致严重
“迟到”处罚（请参阅 UCL学术规则 第 3.11 节） ）和报告
迟交超过五个工作日将收到零标记。 项目介绍
将 **开始** 在 **9月2日2019年周** 举行 （具体日期待定）。

7.学生应安排在 **19** 日 的一周可用 – **23** 日 **2019年8月** ，在他们的情况下，
主管需要与他们联系，询问他们的报告。

8.项目报告的长度取决于项目的主题，可能会有很大差异
项目之间。 长度在8,000到15,000字之间（不包括目录，
参考列表，以及任何表格，图形，计算机程序，计算机输出和附录）
普遍接受。 典型的项目报告长度在10,000到12,000字之间。 决赛
项目报告的版本应在首页上说明其字数，而绝对值
允许的最大值为16,500字。 超过16,500字的项目报告将产生10
商标中的百分点扣除，受条款规定的约束 学术的3.12
条例 ）。 通常要求完成和演示的工作量很大
足够的，并且材料以一种可以理解的方式呈现给同学们
相当的背景（所以8,000个单词可能只是一个非常理论的适当长度
或密集呈现的报告）。 另一方面，报告不应过于重复或含有
不必要或不相关的细节，甚至可能导致低于15,000字的标记。

9.有关审查员的指导，请参阅“研究生手册”第25–27页
将在一个项目中寻找。 这些指南也可以在第48–50页找到
文献。

我| 页

**第3页**

**理学硕士项目清单2019（理学硕士统计）**

| 项目名称 | SUBMITTING SUPERVISOR | 额外监事 | 计划适宜性 **硕士统计** | **硕士数据科学** | **硕士医统计** |
|---|---|---|---|---|---|
| **一类残疾收入保险模型** | Abourashchi，Niloufar | N / A | 是 | 没有 | 没有 |
| **多重递减模型** | Abourashchi，Niloufar | N / A | 是 | 没有 | 没有 |
| **随机试验中停止规则的研究** | Ambler，Gareth | N / A | 是 | 没有 | 是 |
| **探索引入HCM的成本效益** | Baio，Gianluca | 奥马尔 | 是 | 没有 | 是 |
| **临床实践中的风险 – SCD风险模型** | | Rumana | | | |
| **以成本效益的方式解决缺少随机数据的问题** | Baio，Gianluca | 戈麦斯， | 是 | 没有 | 是 |
| **分析** | | 曼努埃尔 | | | |
| **开发一种用于试验的NMA的交互式R / SHINY工具** | Baio，Gianluca | 贝拉尔迪， | 是 | 没有 | 是 |

| | | | | | |
|---|---|---|---|---|---|
| 水平数据 | | 安德里亚 | | | |
| 隐马尔可夫模型及其在金融中的应用 | Beskos， | N / A | 是 | 是 | 没有 |
| 校准的计算方法 | 亚历山德罗 | | | | |
| 使用隐马尔可夫模型 Beskos 了解地震数据， | | N / A | 是 | 是 | 没有 |
| | 亚历山德罗 | | | | |
| BAYESIAN对变点模型的推断 | Beskos， | N / A | 是 | 是 | 没有 |
| | 亚历山德罗 | | | | |
| 交换率的多元随机波动率模型 | Beskos， | N / A | 是 | 是 | 没有 |
| 和他们使用STAN进行校准 | 亚历山德罗 | | | | |
| 对不均匀抽样的空间点过程的推断 | 钱德勒，理查德N / A. | | 是 | 没有 | 没有 |
| EFFORT | | | | | |
| 环境统计中的其他主题 | 钱德勒，理查德N / A. | | 是 | 是 | 没有 |
| 西北太平洋台风模拟 | 钱德勒，理查德N / A. | | 是 | 是 | 没有 |
| 有条件和部分区域开关的COPULAS： | 多诺夫，亚历克斯 | N / A | 是 | 没有 | 没有 |
| 基于混合分布的逼近 | | | | | |

第4页

| | | | | | |
|---|---|---|---|---|---|
| 利用极值理论和COPULAS评估市场 | 多诺夫，亚历克斯 | N / A | 是 | 没有 | 没有 |
| 风险 | | | | | |
| 用金融资产模拟金融资产的依赖结构 | 多诺夫，亚历克斯 | N / A | 是 | 没有 | 没有 |
| REGIME-SWITCHING VINE COPULA | | | | | |
| 因子聚合物在多元化模拟中的应用 | 多诺夫，亚历克斯 | N / A | 是 | 没有 | 没有 |
| 系统风险 | | | | | |
| 从样本中估计模式 | 汤姆，谢谢 | N / A | 是 | 是 | 没有 |
| R光谱校准程序 | 汤姆，谢谢 | N / A | 是 | 是 | 没有 |
| 关于高分析的一个或两个不同的项目 | 汤姆，谢谢 | N / A | 是 | 是 | 没有 |
| 尺寸光谱数据，标题将完全依赖 | | | | | |
| 什么是项目覆盖范围 | | | | | |
| 分析开放式问题的文件 | 吉约曼， | Bernardoni， | 是 | 是 | 没有 |
| | 亚瑟 | 米尔科 | | | |
| 使用GPUS的高斯过程仿真器 | 吉拉斯，塞尔 | N / A | 是 | 是 | 没有 |
| 使用分层结构的高斯过程仿真器 | 吉拉斯，塞尔 | N / A | 是 | 是 | 没有 |
| 具有相互作用的线性模型中的贝叶斯变量选择格里芬，吉姆 | | N / A | 是 | 是 | 没有 |
| 鲁棒推理的气体模型 | 格里芬，吉姆 | N / A | 是 | 是 | 没有 |
| 脉冲响应函数的局部投影 | 格里芬，吉姆 | N / A | 是 | 是 | 没有 |
| 多点模式数据依赖性的调查 | 霍纳，托马斯N / A. | | 是 | 是 | 没有 |
| 点模式数据中抑制的研究 | 霍纳，托马斯N / A. | | 是 | 是 | 没有 |
| 点模式数据的研究 | 霍纳，托马斯N / A. | | 是 | 是 | 没有 |
| 工作流程结构和统一的界面设计 | 基拉伊，弗兰兹 | N / A | 是 | 是 | 没有 |
| 用于统计建模和机器学习的工具箱 | | | | | |
| 结构的存在（INC。时间和层次） | | | | | |
| 广义误差的理论和实证研究 | 基拉伊，弗兰兹 | N / A | 是 | 是 | 没有 |
| 估算，定量模型验证和模型 | | | | | |
| 比较元学方法 | | | | | |
| 贝叶斯的高级马尔可夫链蒙特卡罗方法 | 利文斯通， | N / A | 是 | 是 | 没有 |
| 计算 | 塞缪尔 | | | | |
| BAYESIAN模型与遗传因素的推断 | 利文斯通， | N / A | 是 | 是 | 没有 |
| | 塞缪尔 | | | | |

第5页

| | | | | | |
|---|---|---|---|---|---|
| 抽样算法的收敛性 | 利文斯通， | N / A | 是 | 是 | 没有 |
| | 塞缪尔 | | | | |
| 预测天气变化对需求的影响 | 利文斯通， | N / A | 是 | 是 | 是 |
| 大伦敦的儿童救护服务 | 塞缪尔 | | | | |
| 脓毒症中器官功能障碍的模式 | Marra，Giampiero Palmer， | | 是 | 是 | 是 |

| | | | | | |
|---|---|---|---|---|---|
| 随着时间的推移，在观察研究中解决遗漏数据 | 马拉，Giampiero G | 爱德华兹； | 是 | 是 | 是 |
| 不同的混淆 | | 曼努埃尔 | | | |
| 重新使用COPULA模型在成本效益方面的应用分析 | 马拉，Giampiero | Gomes，曼努埃尔 | 是 | 是 | 是 |
| 预测足球博彩市场的运动 | Marra，Giampiero | N／A. | 是 | 是 | 没有 |
| 评估足球博彩市场的效率 | Marra，Giampiero | N／A. | 是 | 是 | 没有 |
| 学生提出的项目 | Marra，Giampiero | N／A. | 是 | 是 | 是 |
| 流式细胞术中细胞亚型的混合模型 | Manolopoulou，约安娜 | N／A | 是 | 是 | 是 |
| 用GAUSSIAN WELLS建模环形星系 | Manolopoulou，约安娜 | N／A | 是 | 是 | 没有 |
| 使用HAWKES过程建模计数过程 | Manolopoulou，约安娜 | N／A | 是 | 是 | 没有 |
| 贝叶斯多状态转换速率建模 | 欧文尼古拉斯 | 范登Hout，Ardo | 是 | 是 | 是 |
| 创建一个R包 | 诺思罗普，保罗 | N／A | 是 | 是 | 没有 |
| 设置较低的极值阈值 | 诺思罗普，保罗 | N／A | 是 | 没有 | 没有 |
| 利用贝叶斯极值模型预测开放集分类 | 诺思罗普，保罗 | N／A | 是 | 是 | 没有 |
| 调查随机选择中的置换测试的使用实验 | 奥基夫，艾丹 | N／A | 是 | 没有 | 是 |
| 具有非比例危险的时间模拟 | 奥基夫，艾丹 | N／A | 是 | 没有 | 是 |
| 比较临床表现和检测外翻 | Pavlou，Menelaos | Ambler，加雷思 | 是 | 没有 | 是 |
| 无线电广播比赛数据分析 | 扑克，Yvo | N／A | 是 | 是 | 没有 |



| | | | | | |
|---|---|---|---|---|---|
| 关于协方格矩阵的几何及其用法的几何统计推断 | 扑克，Yvo | N／A | 是 | 没有 | 没有 |
| 场效应晶体管的统计研究 | 扑克，Yvo | N／A | 是 | 没有 | 没有 |
| 中年中心极限定理 | 扑克，Yvo | Eltzner，本杰明 | 是 | 没有 | 没有 |
| 人体髋关节运动模型 | 扑克，Yvo | N／A | 是 | 没有 | 是 |
| 人口动态与SDES | 拉西亚斯，马蒂娜 | N／A | 是 | 没有 | 没有 |
| 随机微分方程及其应用 | 拉西亚斯，马蒂娜 | N／A | 是 | 没有 | 没有 |
| 条件和测试组成属性 | Sadeghi，Kayvan | N／A | 是 | 是 | 没有 |
| 用于指数随机图模型的模型多项式 | Sadeghi，Kayvan | N／A | 是 | 是 | 没有 |
| 离散决定点过程的限制性 | Sadeghi，Kayvan | N／A | 没有 | 是 | 没有 |
| 不确定性下的不确定性能力扩张竞争性电力市场 | Siddiqui，Afzal | N／A | 是 | 没有 | 没有 |
| 开关选项的评估 | Siddiqui，Afzal | N／A | 是 | 没有 | 没有 |
| 美国通货膨胀的福利成本 | 塞尔丘克，杰米尔 | N／A | 是 | 没有 | 没有 |
| 电力行业的风险管理 | Siddiqui，Afzal | N／A | 是 | 没有 | 没有 |
| 英国通货膨胀的福利成本 | 塞尔丘克，杰米尔 | N／A | 是 | 没有 | 没有 |
| 能源市场中的游戏理论投资分析 | Siddiqui，Afzal | N／A | 是 | 没有 | 没有 |
| 具有“弱协会”的因果网络 | 席尔瓦，里卡多 | N／A | 是 | 是 | 是 |
| 加强学习的因果关系 | 席尔瓦，里卡多 | N／A | 是 | 是 | 是 |
| 社会和空间网络中的因果效应模型其他相关数据 | 席尔瓦，里卡多 | N／A | 是 | 是 | 是 |
| 用SPOTIFY数据进行用户行为的大规模分析 | 席尔瓦，里卡多 | N／A | 是 | 是 | 没有 |
| 利用自然数据模拟时间序列数据的因果效应实验 | 席尔瓦，里卡多 | N／A | 是 | 是 | 是 |
| 学生LED项目 | 席尔瓦，里卡多 | N／A | 是 | 是 | 是 |
| DE-CLUSTERING EARTHQUAKE CATALOG DATA | Stavrianaki，卡捷琳娜 | N／A | 是 | 是 | 没有 |
| 地震的SHAZAM – 对区域的应用地震目录 | Stavrianaki，卡捷琳娜 | N／A | 是 | 是 | 没有 |

**第7页**

| | | | | | |
|---|---|---|---|---|---|
| **用于模拟认知功能的双向离散分布** | 范登豪特，ARDO | N／A | 是 | 是 | 是 |
| **参数时间依赖的多状态生存模型** | 范登豪特，ARDO | 尼古拉斯，欧文 | 是 | 是 | 是 |
| **存活数据的广义时间依赖逻辑模型** Van Den Hout， | | N／A | 是 | 是 | 是 |
| | ARDO | | | | |
| **多次测试中的错误发现控制** | 王腾尧 | N／A | 是 | 是 | 没有 |
| **矩阵完成** | 王腾尧 | N／A | 是 | 是 | 没有 |
| **高维主成分分析** | 王腾尧 | N／A | 是 | 是 | 没有 |
| **分类GASTRO的聚类拉曼数据 – 肠癌** | 薛景浩 | 托马斯，杰兰特 | 是 | 是 | 是 |
| **生物医学中国家科学数据的优化聚合** | 薛景浩 | 琼斯，马丁是的 | | 是 | 是 |
| **图像分析** | | | | | |
| **半监督机器学习分类拉曼图像** | 薛景浩 | 托马斯，杰兰特 | 是 | 是 | 是 |
| **卵巢癌** | | | | | |
| **伪狂犬病与多普勒癌的分类** | 薛景浩 | 托马斯，杰兰特 | 是 | 是 | 是 |
| **从拉曼图像** | | | | | |
| **结肠癌KAPPA统计学的META分析** | 薛景浩 | 托马斯，杰兰特 | 是 | 是 | 是 |
| **评定** | | | | | |
| **基于表征的分类** | 朱瑞 | 薛景浩是的 | | 是 | 没有 |

**第8页**

**理学硕士项目描述2019年**

**标题：** 一类残疾收入保险模型

**主管：** Niloufar Abourashchi博士

**适应性：** 理学硕士（财务背景优先）

**描述：**

这些项目的目的是说明马尔可夫随机过程的数学如何
通过多个状态模型的框架，在精密模拟科学中使用不同的
各种保险。 本主题将向您介绍一类残疾收入保险模型。
一个这样的例子是残疾保险模型或健康疾病模型，它提供和
如果她/他病得太重而无法工作，则承保残疾保险以取代保单持有人的收入。 对于
更多信息请联系导师。

**标题：** 多个减量模型

**主管：** Niloufar Abourashchi博士

**适应性：** 理学硕士（财务背景优先）

**描述：**

这些项目的目的是说明马尔可夫随机过程的数学如何
通过多个状态模型的框架，在精密模拟科学中使用不同的

各种保险。 选择此主题的学生将被介绍给一类模型
多次减少。 这些模型是多状态模型的子集，并且是特殊的
从精算的角度来看，数学观点非常“频繁”。 一个这样的例子是
定义的养老金计划模型。 这是雇主提供的“法律承诺”
退休期间具有确定养老金收入的员工。 了解更多信息
请联系导师。



**第9页**

**标题：** 随机试验中停止规则的研究
**主管：** Gareth Ambler博士
**适应性：** 理学硕士统计学（取决于所选模块）和理学硕士医学统计学
**描述：**
中期分析通常在随机试验中进行，以研究安全性和有效性
虽然审判仍在进行中。 这些分析与统计停止一起进行
确保I类错误得到控制的规则。 在这个项目中，模拟（和分析
方法，如果可行的话）将用于调查几个方面的特征和性能
众所周知的停止规则，包括O'Brien和Fleming以及Pocock提出的规则。 在
此外，Lan–de Mets的'alpha支出'方法和停止无效的规则将是
调查。 此外，在（多臂）多阶段II期研究中使用停止规则可能
也有待探索。

**题目**： 探索在临床中引入HCM Risk–SCD风险模型的成本效益
实践
**主管**： Gianluca Baio和Rumana Omar
**适应性：** 硕士统计学和理学硕士医学统计学
**说明**：
HCM是一种常见的遗传性心肌病，可能影响多达200人中的1人。 它是一个
年轻人心源性猝死（SCD）的主要原因。 需要有高度SCD风险的患者
确定可以通过植入式心律转复除颤器为他们提供救生治疗
（ICD）。 HCM Risk–SCD风险模型用于估计患者的SCD和患者的5年风险
接受ICD植入的预测风险≥6%。 据计算，每13人计算一次
使用此阈值接受ICD的高风险患者，可能有1名患者可能被挽救
SCD。 HCM Risk–SCD计算器可用于避免低风险的不必要的ICD植入物
耐心。 该项目的目标是制定一个成本效益模型来估算价值 –
为风险模型的引入提供了资金。 该项目将涉及与临床医生和
专家创建一个模型来描述有无干预的患者通路; 一个
有针对性的文献综述，以获得该模型的主要参数信息; 和成本 –
效果分析。
**要求**： R（必要）知识; 贝叶斯建模（必不可少）; 健康经济

评估（可取）。



**第10页**

---

**标题**： 在成本效益分析中解决丢失的随机数据

**主管**： Gianluca Baio和Manuel Gomes

**适应性:** 硕士统计学和理学硕士医学统计学

**说明**：

缺失数据是成本效益分析（CEA）中的常见问题，并且经常被解决
假设数据“随机丢失”（MAR）。 但是，这种假设常常值得怀疑
需要进行敏感性分析以评估离开MAR的影响。 参考–
基于估算为进行这种敏感性分析提供了一种有吸引力的方法，因为
缺失的数据假设通过参考特定的子方式以可访问的方式构建
样本中的组。 例如，安慰剂控制的合理的非随机机制
试验将假设辍学的实验组参与者停止服用他们的
治疗，并与安慰剂组的结果相似。 最近的工作考虑了
通过假设成本效益终点是CEA中基于参考的插补方法
（联合）正常分配。 然而，这不太可能像素数的结果那样合理
对CEA的兴趣，例如成本和患者报告的结果，往往表现出很大的偏差
正常性（例如高度倾斜和多模态）。 这将导致误导性的推论，因为
基于参考的方法通常依赖于关于数据的合理分布假设。
基于现有的贝叶斯方法处理MNAR数据，该项目将考虑如何
扩展基于参考的方法以解决CEA中的非正常MNAR结果。 方法
将应用于IMPROVE研究，该研究评估血管内治疗策略与开放治疗相比
动脉瘤破裂患者的手术治疗。

**要求**： R（必要）知识; 贝叶斯建模（必不可少）; 健康经济
评估（可取）。

**标题**： 开发用于NMA试用级数据的交互式R / Shiny工具

**主管**： Gianluca Baio和Andrea Berardi

**适应性:** 硕士统计学和理学硕士医学统计学

**说明**：

该项目将涉及开发R / Shiny应用程序，用于传导荟萃分析
试用级数据。 疾病领域待确认，但一种选择是多发性硬化症
确诊残疾进展和年复发率的结果。 理想情况下分析
考虑Frequentist和贝叶斯NMA方法，但是Frequentist方法会
作为起点。 贝叶斯组件需要与外部程序链接



**第11页**

作为JAGS或WinBUGS用于Gibbs采样，因此将涉及更大的复杂性。 目的是
开发灵活的平台，其中用户可以选择包含在网络中的试验
只要网络保持连接。 该工具还将允许用户测试早期功效
对新产品的假设。
该工具的开发将在R / RStudio中使用Shiny包来开发用户
接口。
候选人的适合性：我们正在寻找一个勤奋，自我激励的人，热衷于此
对健康经济学和数据分析的兴趣加入团队的夏季安置。
理想的候选人将具有R编程的经验和对该编程的深刻理解
NMA的原则。 JAGS / WinBUGS的经验是可取的。 将提供技术支持。
**要求**：R / RStudio（基本），Microsoft Office（基本），JAGS和/或WinBUGS（优先）

这是一个位于伦敦办公室的全职工作，为期3个月。
关于安置地点 – PAREXEL International，Euston
卫生经济学模型组（HEMU）是PAREXEL Access Consulting的一个业务部门。
HEMU是一个专门的部门，在英国，美国和瑞典拥有35名员工，其目的是提供
卫生经济学咨询服务国际制药和医疗器械
公司。 主要活动是成本效益，预算影响建模和数据分析。

**标题：** 使用隐马尔可夫模型了解地震数据
**主管：** Alexandros Beskos博士
**适应性：** 理学硕士统计学或硕士数据科学
**描述：**
环境科学的一些工作使用隐马尔可夫模型来拟合相应的数据
地震事件。 潜在的马尔可夫状态可以被认为是“压力场”领先
到地震事件; 观测结果是推断的地震的大小。 该
项目将使HMM适应地震数据; 我们将看看最大似然和贝叶斯
方法，并尝试量化估计模型参数的不确定性并执行模型
选择。
先决条件：熟悉R（例如，如教授的STAT0030），应用贝叶斯方法（例如，如教导
在STAT0031）和马尔可夫链。 此外，愿意学习和应用计算技术，
使用例如STAN或OpenBUGS。



**第12页**

**标题：** 隐马尔可夫模型及其在金融和计算机中的应用
校准方法
**主管：** Alexandros Beskos博士
**适应性：** 理学硕士统计学或硕士数据科学
**描述：**
隐马尔可夫模型（HMM）构成了广泛使用的一类重要的统计模型
各种应用。 该项目将探索最近的蒙特卡罗技术（例如
马尔可夫链蒙特卡罗和粒子滤波器，称为粒子-MCMC），以使这些模型适合数据。

我们将研究HMM在金融模型中的应用，例如多变量随机波动率
模型或GARCH类型。

先决条件：熟悉R（例如，如STAT0030中所教导的），应用贝叶斯方法（例如，如
在STAT0031中讲授）和马尔可夫链。 此外，愿意学习和应用计算
与HMM相关的技术; 例如STAN或OpenBUGS。

**标题：** 汇率的多元随机波动率模型及其校正
使用STAN

**主管：** Alexandros Beskos博士

**适应性：** 理学硕士统计学或硕士数据科学

**描述：**

最近在计量经济学方面的研究工作研究了多元汇率的联合模型（或
随机波动率背景下的其他时间序列。 这些模型带来了挑战
由于未观测到的随机波动率和高维数，它们对观测值进行校准。
但是，最近的计算软件包如STAN允许非高级MCMC的专家
方法，应用强大的计算算法来校准这些模型并执行完整的
贝叶斯推理和预测。

先决条件：熟悉R（例如，如STAT0030中所教导的），应用贝叶斯方法（例如，如
在STAT0031中讲授，马尔可夫链。 此外，愿意学习和应用MCMC技术，使用
像OpenBUGS或STAN这样的软件包。



**第13页**

**标题：** 变点模型的贝叶斯推断

**主管：** Alexandros Beskos博士

**适应性：** 理学硕士统计学或硕士数据科学

**描述：**

统计学中的许多应用涉及从可用数据学习变化点模型。 这样
由于潜在组件的存在，模型对其校准提出了挑战
和高维度。 然而，最近的计算进步允许快速计算
统计方法 – 在贝叶斯框架中 – 发现变化点。 潜在的应用
包括生物学或金融学。

先决条件：熟悉R（例如，如STAT0030中所教导的），应用贝叶斯方法（例如，如
在STAT0031中讲授，马尔可夫链。 此外，愿意学习和应用蒙特卡罗技术，
例如MCMC或粒子过滤，使用OpenBUGS或STAN等软件包。

**标题：** 台风模拟西北太平洋

**主管：** 理查德钱德勒教授

**适合：** 理学硕士统计学，MSc数据科学

**描述：**
该项目将使用中国气象局“热带气旋最佳轨道”的数据
数据“开发可用于热带气旋（台风）形成和运动的模型
评估气候变化时东亚风暴破坏风险的变化。 这项工作将建立在一个
其他学生项目，将需要良好的计算技能（至少在“A”级）
STATG003）以及对广义线性模型（STATG001或STATG006）和a的熟悉程度
愿意学习新的随机建模技术。



**第14页**

---

**标题：** 对采样不均匀的空间点过程的推断
**主管：** 理查德钱德勒教授
**适合：** 理学硕士
**描述：**
该项目涉及（二维）空间中点的模式分析
没有观察到所有点的情况。 这是由于出现的问题所致
解释历史地震报告，其中没有具体的报告
位置可能是因为那里没有发生任何事件，或者因为那里没有人
体验它。 学生将被要求学习一些关于空间推理的新材料
点过程，也可能使用BUGS将任何新方法应用于数据集
过去英国和/或中国的地震。

**标题：** 环境统计中的其他主题
**主管：** 理查德钱德勒教授
**适合：** 理学硕士统计学，MSc数据科学

**描述：**欢迎学生在任何与环境相关的项目主题中提出建议
应用。 任何希望选择此选项的学生都需要考虑其可用性
合适的数据集。

**第15页**

**标题：** 条件和部分制度转换Copula：基于的近似
混合物分布。
**主管：** Alex Donov博士
**适应性：** 理学硕士
**描述：**
部分相关系数提供关于关联程度的信息
两个随机变量以一组控制随机变量为条件。 介绍一个
部分copula概括了这种方法，并允许条件的多维建模
依赖。 由于这种依赖性可能会随着时间而改变，因此该项目的目标是结合起来
部分copula方法与政权转换，并提出了一个Copula切换模型
允许随时间推移的不同条件依赖结构

**标题：** 用极值理论和Copulas评估市场风险
**主管：** Alex Donov博士
**适应性：** 理学硕士
**描述：**
风险价值（VaR）被金融机构广泛用于管理其市场风险。 要正确
评估一定范围内投资组合的可能损失，准确衡量
投资组合中资产收益之间的依赖关系至关重要。 这项研究将纳入
极值理论（EVT）模拟回归分布的尾部和各种copulas构建
联合分配回报。

**标题：** 使用制度转换藤来模拟金融资产的依赖结构
系词。
**主管：** Alex Donov博士
**适应性：** 理学硕士
**描述：**
现在人们普遍认为，金融时间序列在时期表现出不同的行为
市场压力。 特别是，在市场低迷期间，资产回报表现出更大的依赖性
市场上涨。 这种现象通常被称为不对称依赖。 的目标
本研究旨在将copula理论与马尔可夫切换相结合，以便更准确地描述
危机时期金融资产收益的时变依赖结构。

**第16页**

**标题：** 因子copula在多元系统风险建模中的应用。

**主管：** Alex Donov博士

**适应性：** 理学硕士

**描述：**

为了量化系统性风险，Adrian和Brunnermeier（2008）提出了CoVaR测量。
但是，通过考虑孤立的来衡量特定机构的系统性风险影响
这个机构的痛苦并不能解释通过他人传播的痛苦
机构。 这项研究的目的是扩展CoVaR措施，以解决困境
多个金融机构在评估其系统性风险贡献时的溢出效应。 该
多CoVaR系统风险度量通过因子copula指定。 因子copulas可以
理解为因子模型隐含的copulas，它提供了潜在的降维。

**标题：**用于光谱校准的R程序

**主管** ： Tom Fearn教授

**适用性** ： 统计或数据科学（统计流）

**说明** ：

近红外（NIR）光谱学提供了一种测量完整组成的快速方法
食品和农产品，以及许多其他材料。 测量是间接的和
需要校准，其中导出预测规则，或者预测定量
样本的组成或基于高的分类将它们分类为两个或多个组中的一个
尺寸吸收光谱。 主成分回归（PCR）和PCR等方法
偏最小二乘回归（PLSR）是标准的，但非线性方法如人工
还使用了神经网络或高斯过程回归。
我主要使用Matlab工作，但（因为它是免费的！）R是一个很有吸引力的选择。 有一些R.
周围的程序实现像PLS这样的方法，但是很难找到R程序
例如，实现常见的光谱预处理，如衍生物或散射校正。 另一个
问题是交叉验证的实现，通常用于调整方法。 当数据有
在嵌套结构中，能够以灵活的方式指定要省略的块是有用的
包通常不允许这样。 该项目的想法是调查可用的内容，
比较存在的选项，提出一些建议，并填补任何明显的差距
有一些很好的文档代码。 你显然需要真正喜欢R来接受这个。



**第17页**

**标题** ： 从样本估算模式

**主管** ： Tom Fearn教授

**适用性** ： 统计或数据科学（统计流）

**说明** ：

我的化学家合作者有兴趣估计一组的分布模式
数据来自。 有一个名为modeest的R包似乎有几种方法

这样做。 这个项目将涉及尝试一些或所有这些方法在真实和
模拟数据以了解它们的工作情况。

**标题：** 一个或两个不同的高维光谱数据分析项目，
标题将取决于项目涵盖的确切领域

**主管**： Tom Fearn教授

**适用性**： 统计或数据科学（统计流）

**说明**：

近红外（NIR）光谱学提供了一种测量完整组成的快速方法
食品和农产品，以及许多其他材料。 测量是间接的和
需要校准，其中导出预测规则，或者预测定量
样本的组成或基于高的分类将它们分类为两个或多个组中的一个
尺寸吸收光谱。 主成分回归（PCR）和PCR等方法
偏最小二乘回归（PLSR）是标准的，但非线性方法如人工
还使用了神经网络，支持向量机或高斯过程回归。
我有各种NIR数据集，包括定量（例如猪胴体中脂肪酸的测量）和
定性（动物饲料成分的分类，伊比利亚火腿的分类）和一些想法
尝试它们的方法，当然学生可能有他们自己喜欢的方法
我想试试。 我很乐意与有兴趣的学生讨论选择。



**第18页**

**标题：** 分析开放式问题的文件

**主管**： Arthur Guillaumin和Mirko Bernardoni

**适用性**： 统计学或数据科学

**说明**：

Clifford Chance（www.cliffordchance.com）的全球知名律师事务所提供的主要项目
在五大洲拥有显着的深度和资源范围，并成为“魔术师”的成员
圆圈“领先的英国律师事务所。
变压器的双向编码器表示是一种预训练语言的新方法
在各种自然语言中获得最新结果的表述
处理（NLP）任务。 我们有兴趣拥有一个能够合法和合法的模型
安全文件和执行阅读理解。
项目目标可以从能够对文档和段落分类，具体提取
通过开放式问题训练数据点到更复杂的完整阅读理解。
该项目正在考虑使用该领域的前沿研究，如考虑斯坦福
问答数据集（SQuAD）排行榜。

**标题：** 使用GPU的高斯过程仿真器

**主管：** Serge Guillas教授

**适应性：** 硕士统计学和理学硕士数据科学

**描述** ： 仿真器取代了计算上昂贵的仿真器（例如气候，地球物理学）
或工程模型）。 使用高斯过程（GP）回归给出典型的仿真器
使用实验设计选择的点。 不幸的是，输入中的点数和
输出空间可能很大，从而减慢了GP的安装速度。 这个项目是关于使用
允许使用图形处理单元（GPU）加速拟合的最新代码。
将访问英国最快的集群，以说明这些体系结构的优势
一些合成问题。 使用的语言是python（一些知识会好但不是
必要）。 参考：https：//github.com/UCL/gp_emulator



**第19页**

**标题：** 使用分层结构的高斯过程模拟器

**主管** ： Serge Guillas教授

**适应性：** 硕士统计学和理学硕士数据科学

**描述：**

仿真器取代计算昂贵的仿真器（例如气候，地球物理或工程）
楷模）。使用选择的点通过高斯过程（GP）回归给出典型的仿真器
使用实验设计。不幸的是，输入和输出空间中的点数
可能很大，从而减慢了GP的安装速度。这个项目非常关注利用
最近的代码允许使用协方差的分层结构来加速拟合
功能。 该项目涉及在一些例子上评估这些方法的技能。
参考：https：//github.com/jiechenjiechen/RLCM

**标题：** 用于稳健推理的GAS模型

**主管**：Jim Griffin教授

**适应性：** 硕士统计学和理学硕士数据科学

**描述：** 广义自回归（GAS）模型是构建时间序列的方法
具有自回归结构的模型用于广泛概率模型的参数（例如
作为$ t $分布的方差或泊松分布的均值）。这是通过
在评分函数上建立一个自回归结构并且已被证明具有良好的性能
统计特性。在这个项目中，您将为重尾的方差实施GAS模型
分发并将它们应用于合适的数据。该项目将有机会进行学习
基础理论，计算方法和数据应用的实现。

**标题：** 具有相互作用的线性模型中的贝叶斯变量选择

**主管**：Jim Griffin教授

**适应性：** 硕士统计学和理学硕士数据科学

**描述：** 变量选择方法旨在找到一些可用变量的子集
对预测响应很有用。贝叶斯方法很有吸引力，因为它们附着在后面
所有可能子集的概率，而不是选择单个子集。在这个项目中，你会看

现代算法中的线性模型中的贝叶斯变量选择，并考虑它们是如何形成的
扩展到具有相互作用的线性模型 这些模型中的推论通常很强
遗传（如果包括两个相关的主效应，则只能包括交互）。 这个
依赖性使得发现后验分布的挑战更具挑战性。这个项目
提供了在这些中考虑贝叶斯推理中的计算挑战的机会
楷模。



**第20页**

**标题：** 脉冲响应函数的局部预测
**主管**：Jim Griffin教授
**适应性：** 硕士统计学和理学硕士数据科学
**描述：** 宏观计量经济学家经常面临着预测其影响的挑战
在其他变量上更改一个变量。例如，中央银行预测a的影响
通货膨胀，增长和失业率的变化。随着时间的推移测量该效果
通常称为脉冲响应函数。在这个项目中，您将看一个基于回归的
用于脉冲响应函数的方法称为局部投影。这个项目提供了机会
看一下现代回归技术在估计脉冲响应中的作用
功能。

**标题：** 研究点模式数据的抑制
**主管：** Thomas Honnor博士
**适应性：** 适用于理学硕士统计和理学硕士数据科学
**描述：** 基点过程模型是泊松点过程，为此
地点是独立分布的。通过增加了复杂程度
成对交互点过程的规范，通过它来模拟抑制
减少对附近点数可能性的贡献。最简单的模型是硬核
点不能以小于指定最小值的距离分开的过程。数据如此
因为公共可用的树木和手机桅杆的位置可能是由这样建模的
过程，有生物和经济论据，表明观察地点应该是
抑制。
学生应该期待阅读点过程理论，获得对它的理解
泊松点过程并将其扩展到成对交互点过程。真实数据
预期显示抑制行为然后将使用这些过程来建模
推断抑制的形式和程度以及数据集之间的差异。
数据分析需要熟练使用R，并且对随机过程的理解也是如此
有用。

**第21页**

**标题：** 多点模式数据依赖性研究

**主管：** Thomas Honnor博士

**适应性：** 适用于理学硕士统计和理学硕士数据科学

**描述：**

多点点模式由点位置组成，每个点位置被分配给一个离散点
类别集合。传统的现实世界的例子是两个光检测的分布
眼内的细胞 – 每个细胞都有一个位置，可以检测光的存在或不存在
光。 然后，一个感兴趣的问题是，是否可以将点类别视为实现
独立点过程和任何依赖的形式。数据如位置
按公司分类的超市可能会被调查为多重点过程
经济论据表明偏好类似的高人口地区或不同地区
竞争程度降低。
学生应该期待阅读点过程理论，获得对它的理解
非参数方法，用于总结基本点模式并将其扩展到multitype
点过程。然后将调查可以对点进行分类的真实数据
存在和依赖的形式决定。数据分析需要熟练使用R和a
对随机过程的理解也很有用。

**标题：** 点模式数据的研究

**主管：** Thomas Honnor博士

**适应性：** 适用于理学硕士统计和理学硕士数据科学

**描述：**

点过程理论可以应用于广泛的现实世界场景中
更好地了解事件在空间上的分布情况。对多种类型观察的扩展
事件的位置，事件的位置所依赖的过程以及事件的比较
分布都是可能的，并且在潜在的MSc项目范围内。提出这个项目
没有参考特定的数据集 – 学生可能有特定的应用程序/数据集
他们自己的想法，应该随时与我联系，更详细地讨论我们的方向
可能需要。
学生应该期待阅读点过程理论，获得对它的理解
非参数方法，用于总结基本点模式并将其扩展到multitype
点过程。然后将调查可以对点进行分类的真实数据
存在和依赖的形式决定。数据分析需要熟练使用R和a
对随机过程的理解也很有用。



**第22页**

**标题：** 统计工具箱的工作流架构和统一界面设计
存在结构时的建模和机器学习（包括时间和层次）

**主管：** 弗兰兹基拉利博士

**适应性：** 硕士统计学和理学硕士数据科学

**描述：**

尽管结构化建模/学习任务非常重要，例如时间序列预测或
分层建模，工具箱支持目前很差 – 可能是由于所需的交叉

软件工程，统计建模和机器学习方面的专业知识。
支持良好的建模任务类型或多或少都假定为“表格”数据，即
适合excel表或数据框的数据，最值得注意的是监督预测任务
其中包括突出的工作流程/工具箱包，如python / sklearn和R / mlr。

有多个项目可用于扩展工具箱对新任务或新建模方面的支持：

- 时间结构的监督和无监督学习任务，例如：时间序列
  预测，面板数据预测，异常检测和变点检测
- 概率建模包括贝叶斯和复合频率预测，密度
  估计，极值分析和预测尾部分布建模
- 风险建模，风险/严重性综合建模，时间 - 事件建模
- 模型调整，模型选择，模型比较和模型性能量化
- 工作流程编排，异构数据源集成，流水线操作，基准测试
- 并行化和与高性能计算内部结构的集成

根据兴趣和所解决的任务类型，可以使用用例数据集
作为框架测试案例，来自能源，金融和健康领域。
该项目与艾伦图灵研究所合作，实现了实习机会
夏天也可以（单独）。过去的项目包括：
用于概率监督建模的skpro工具箱：
https://github.com/alan-turing-institute/skpro

用于分层数据类型和特征提取的xpandas数据容器接口
https://github.com/alan-turing-institute/xpandas



**第23页**

学生应具备统计/ ML建模和软件经验的扎实背景
在R，python或Julia中开发。他们将最佳地（但不一定）参加2018年
STATG019模块的迭代。
**标题：** 广义误差估计的理论和实证研究，定量
模型验证和模型比较元方法
**主管：** 弗兰兹基拉利博士
**适应性：** 硕士统计学和理学硕士数据科学
**描述：**
用于外部，域无关的模型验证的元方法对于成功控制至关重要，
特别是在有数以千计的建模策略可供选择的时代，但很多
它们是没有内在保证或模型选择量词的黑盒子。同样，理论
比较基于平等基础上广泛不同的理论假设的建模策略
在任何“最佳选择”不明确的实际情况中都至关重要。
在这种情况下，有许多主题可供使用，可以在任何地方进行研究
理论/数学重和经验/实验重的方法之间的频谱，例如：

- 模型 - 内在与模型 - 外在绩效保证的比较研究
- 模型不可知估计预测策略的泛化误差

- 用于比较预测策略和预测功能的假设检验
- 无监督建模策略的定量，模型不可知验证
- 中心极限定理和在顺序设置中进行模型比较的保证
- 自动化，模型无关的超参数调整的原则方法

候选人应该在统计学的理论基础上有扎实的背景
在R或python中进行模拟研究的概率和/或经验。他们将是最佳的
（但不一定）参加了STAT3019模块的2018年迭代。

**标题：**用于贝叶斯计算的高级马尔可夫链蒙特卡罗方法
**主管：** 塞缪尔利文斯通博士
**适应性：** 硕士统计学和理学硕士数据科学
**描述：** 马尔可夫链蒙特卡洛（MCMC）是一种适用于贝叶斯拟合的有效方法
模型，以及统计/机器学习研究的蓬勃发展领域。有几个方向
这可以在这个项目中取决于学生，也可能不止一个项目
根据兴趣可用。该项目将包括学习，实施和
改进在不同设置中使用的各种MCMC方法。该项目将需要一个



**第24页**

对马尔可夫链的基本理解以及对数学，编程的热情
或两者兼而有之（取决于学生）。

---

**标题：** 贝叶斯模型和遗传祖先的推断
**主管：** Samuel Livingstone博士和Ioanna Manolopoulou博士
**适应性：** 硕士统计学和理学硕士数据科学
**描述：** 推断出两种生物体具有多长时间的回溯时间
共同的祖先是进化生物学（特别是系统发育学）中的一个关键问题
是试图回答这些问题的各种新颖的统计和机器学习方法
使用遗传数据和适当的概率模型。学生将学习一些
使用的基本模型，将这些算法拟合到数据的算法，并比较一些不同的方法
常用于实践中。还有一些空间来测试一些新的想法。该项目将有一个
有趣的建模和计算/编程任务的良好组合，所以适合学生谁
对这些事情感兴趣，以及处理真实的科学数据和回答重要的事情
与之相关的问题。
**先决条件：** R中的编程
**参考文献：** Hein，J.，Schierup，MH和Wiuf，C。Gene Genealogies，Variation and Evolution – A
聚结理论入门。牛津大学出版社，2005年。
**标题：** 采样算法的收敛性
**主管：** 塞缪尔利文斯通博士
**适应性：** 硕士统计学和理学硕士数据科学
**描述：**
该项目适合那些有兴趣了解用于量化和数学的数学的人
比较拟合模型中使用的算法的性能与当前非常的数据
在概率机器学习和统计中很受欢迎（特别是在更复杂的模型中）
需要一些正规化）。我们将重点关注抽样方法（最有可能是马尔可夫
连锁蒙特卡洛），并使用各种不同的指标来看看最近的融合发展

关于概率分布的空间，包括瓦瑟斯坦距离，总变差，
最大平均差异，可能还有其他一些差异。这适合对学生充满热情的学生
主题的数学方面。

---



**第25页**

**标题：** 预测天气变化对儿童救护车需求的影响
大伦敦的服务

**主管：** 塞缪尔利文斯通博士

**适应性：** 所有课程

**描述：**

临床医生与大奥蒙德街的儿童急性转运服务（CATS）有关
医院认为，对服务的需求存在强烈的短期波动，尤其是
在冬季，根据天气情况而定。他们要求我们对此进行调查
提供了12年的日常需求数据。学生会尝试通过设计来捕捉这种效果
并使用灵活的广义加法模型框架拟合几种不同的模型
（GAMS）。如果一切按计划进行，工作有可能在期刊上发表。没有真正的预
需要必要条件，但要很好地理解基本统计模型和获得意愿
他们的手数据与一些真实的数据和实验将是非常有利的。克里斯蒂娜博士
来自伦敦大学学院临床运筹学部门的Pagel也将参与该项目。

**第26页**

**标题：** 脓毒症中器官功能障碍的模式
**主管：** Giampiero Marra博士和Edward Palmer博士
**适应性：** 所有课程
**描述：**
抽象
脓毒症是一种危及生命的疾病，可以杀死全球数百万人。这是一个非常不同的
条件; 患者存在于不同的时间尺度，具有不同程度的严重程度，并有反应
与治疗不同。目前正在进行重大研究，以确定其模式
可以解释这种异质性的生理学，从而为临床试验设计提供信息。我们有一个
大型数据资源，包含来自英国的40,000多个重症监护事件。我们感兴趣
使用该数据资源对脓毒症患者的生理模式进行建模。信息性审查
是数据的主要组成部分，学生需要解决这个问题。该项目将是
由Ed Palmer博士（重症监护医师）临床监督，并由Dr.博士统计监督
Giampiero Marra。作为该项目的输出，应鼓励学术出版物。

介绍
脓毒症是一种由感染引起的危及生命的器官功能障碍的异质综合征。 尽管
无数潜在的治疗载体已经在基础科学研究中显示出前景，但没有一个
被发现在翻译成人类时是有效的。因此，现有疗法的主要支柱
保持广谱抗生素和支持治疗; 血管活性药物，流体和技术
旨在增加或暂时替换失败的器官。
描述和理解器官功能以及功能障碍是复杂的。目前流行
描述器官功能障碍的方法是基于过时的专家共识意见。 数据
驱动描述确实存在，但它们尚未被广泛采用，目前还不清楚它们是否存在
比专家主导的描述更好。在子日分辨率，所有当前的方法
这个问题开始崩溃了。
我们正在寻找一个有动力的个人，将统计学习技术应用于这个问题。
目标是开发一个简约的评分系统，准确描述患者的
器官功能。

数据
重症监护健康信息学协作（CC-HIC）是一个大型多中心研究项目，
汇总来自12个重症监护病房的重症监护病人的高保真纵向数据
英国的五个生物医学研究中心（伦敦大学学院，剑桥大学，牛津大学，盖伊大学和圣托马斯大学）



**第27页**

帝国）。有263个变量供研究，包括：人口统计学，急性病严重程度评分，
高分辨率床边监测，药物输注，微生物学，器官支持和结果。 在
目前，CC-HIC内部有近5亿个数据点。
输出
期望是学生将设计和验证新的器官功能障碍指标，使用
数据驱动技术。强烈鼓励出版和会议介绍

支持的。
任务

•关于该主题的小型文献综述

•设计合适的方法来模拟单个器官系统中的器官功能障碍

（心血管或呼吸系统）

先决条件

期望学生具备良好的R工作知识.Python不太可能

在研究期间。可以使用SPSS和SAS，但它们不是首选。

道德与治理

一旦项目名称正式化并与候选人达成一致，就需要这样做

注册由CC-HIC科学顾问委员会审查（Palmer博士将安排）。除此以外

道德审查已经到位。

支持

该项目将具有临床世界密切监督和协作的优势

败血症研究领域的领导者。拟议的直接监督结构将是：

•Edward Palmer博士：主要主管，临床/数据资源

•Giampiero Marra博士：主要监督员，统计数据



**第28页**

**标题：** 重新审视在成本效益分析中使用copula建模

**主管：** Giampiero Marra和Manuel Gomes

**适应性：** 所有课程

**描述：** 根据定义，成本效益研究通常对建立联合感兴趣

关于被比较的替代干预措施的有效性和成本的推论。 这个

通常涉及多变量建模或捕获成本联合性质的替代方法

和结果。这超出了成本与成本之间相关性的技术调整

效果。 例如，它解决了对某些模型进行联合假设检验的需要

系数（例如子组效应）。成本效益数据的另一个独特特征是它们

分配形式。例如，成本通常是高度倾斜或半连续的和结果

与健康相关的生活质量指标往往是左倾或多模式。

超出双变量正态情况的参数化联合模型实施起来相当复杂

并且通常需要在贝叶斯框架内使用MCMC方法。一个实用的选择

方法是使用copulas，通过使用可以用于构建多变量分布

单变量累积分布函数。这种方法的关键优势在于其灵活性

结合不同类型的边际分布，它可以模拟更复杂的依赖

成本效益终点之间。该项目的目的是重新审视copula的潜力
与传统的关节生产方法相比，成本效益分析建模
推论。copula方法将在REFLUX研究评估的CEA中说明
腹腔镜手术治疗反流病患者。对R的熟悉是
可取的。

**标题：** 在时间混杂的观察研究中解决缺失的数据
**主管：** Giampiero Marra和Manuel Gomes
**适应性：** 所有课程
**描述：**
现在人们越来越有兴趣使用大型观察性研究来估计
治疗效果，以帮助像NICE这样的机构提出有关哪些健康的建议
干预提供。依赖这种纵向，常规收集的主要问题
数据是（时变的）混淆的指示。这是一个经常出现的问题，因为患者
进展通常会影响未来的治疗和结果，但也受到以前的影响
治疗。 一个相关的问题是通常这些例行数据源是为响应而收集的
临床需要，关于感兴趣的结果和潜在的混杂因素的信息不完整，
这可以放大这些偏见，并为解决混乱问题提出额外的挑战。



**第29页**

缺少数据为使用标准方法解决时变问题带来了新的挑战
基于逆概率加权（IPW）的边际结构模型（MSM）等混淆
或G估计，尤其是因为缺失数据的模式往往是非单调的。 这个
项目将考虑更适合处理非单调缺失数据的替代方法，
例如多重插补（MI）。该项目将探索MI与MSM的结合并进行比较
使用传统的IPW审查权重。这些方法将在案例研究中说明
估计生物药物治疗类风湿性关节炎患者的有效性
来自美国国家风湿病数据库的数据。

**标题：** 预测足球博彩市场的运动
**主管：** Giampiero Marra
**适应性：** 硕士统计学和理学硕士数据科学
**描述：**
在体育博彩市场中，通常可以下注的金额随着您的增加而增加
接近开球。然而，与此同时，市场通常变得更有效率，领先
进行量 – 价格权衡。一个常见的问题是，是否放置低音但是慷慨
现在定价，或等待更高的价格，可能更糟糕的价格。
Smartodds是一家为专业赌徒提供服务的私人咨询公司，可以提供丰富的服务
主要投注线的赛前市场价格数据集，用于至高无上，总目标和比赛
足球比赛中的结果市场（例如，分别是曼城赢得超过2个
目标; 比赛进球超过3球; 从曼彻斯特城赢得的一系列顶级飞行
和二线欧洲联赛。对于每场比赛，我们可以提供有关这些价格的数据
投注线在开盘前的几分钟，几小时和几天内从开盘价开始（当时
市场开盘）到收盘价（比赛开始时）。
该项目将尝试回答以下问题：如果当前可用的赔率是X，那么如何
远离X可能是比赛实际开始时的几率？
可能有助于解决此问题的问题包括：例如：这取决于多长时间

比赛开始前还剩下什么？它取决于联盟还是市场类型？很早，很早
在特定方向上移动预测后期市场走势？
学生可以通过任何吸引人的方法来做到这一点，但一种选择可能是模拟赔率变动
作为一个随机游走，并试图评估，例如，比赛开始前的剩余时间，具体
联盟和市场类型会影响可能的移动规模。例如，更高档的联赛
通常会吸引更多的博彩行动，因此市场价格可能更不稳定。
学生应该能够用R或python编写代码。



**第30页**

**标题：** 评估足球博彩市场的效率
**主管：** Giampiero Marra
**适应性：** 硕士统计学和理学硕士数据科学
**描述：**
两个事实：（1）几乎所有足球联赛都可以赌博，（2）投注市场倾向于
随着时间的推移变得更有效 但是，在每个联盟中市场同样有效，它的速度有多快
改善并且在一个赛季结束时市场比一开始更好的预测？
Smartodds是一家为专业赌徒提供服务的私人咨询公司，可以提供胜利
对于足球结果的收盘市场赔率几率所暗示的得分概率
自2010年以来，各种全球足球联赛的比赛。例如，这些将告诉你什么
市场分配给团队赢得，抽签或输掉一场比赛（1X2）的概率是多少
他们被期望获得的平均目标（至高无上）以及你期望的目标数量
在比赛中得分（总进球数）。
该项目将使用回归技术（例如通用添加剂或线性混合模型）或
机器学习试图确定这些联赛的市场预测质量
随着时间的推移，它是否会随着联赛的不同而变化，是否随着时间的推移而变化
季节。 可以通过查看例如Brier得分来评估市场预测的质量
1X2获胜概率，以及RMSE的至高无上和总目标预测，但有很多方法
这可以根据学生的兴趣和专业知识来完成。
学生应该能够用Python或R编写代码。

**标题：** 学生提出的项目
**主管：** Giampiero Marra
**适应性：** 所有课程
**描述：**
学生在生存分析，copula回归建模领域提出的项目，
分布回归或惩罚样条回归。

**第31页**

**标题：** 流式细胞术中细胞亚型的混合物建模
**主管：** Ioanna Manolopoulou
**适应性：** 所有课程
**描述：**
在生物技术中，流式细胞仪可以有效地测量各种蛋白质水平
细胞表面; 典型的数据集包含数十到数十万个单元测量值。 在
传统的流式细胞仪，细胞通过人工检查分为亚型（也称为
（测量）测量分布。高斯混合建模允许我们自动进行
通过将每个细胞亚型与a相关联来识别潜在细胞亚型的结构和位置
高斯密度。该项目将研究将混合模型拟合为a的不同方法
流式细胞仪数据集并比较结果。

**先决条件：** R中的编程

**参考文献：** C。Chan，F。Feng，J。Ottinger，D。Foster，M。West，TB Kepler（2008）统计混合物
流式细胞术中细胞亚型鉴定的建模。Cytometry A.73（8）：693–701
G. McLachlan，D。Peel（2000）有限混合模型。威利。

---

**标题：** 使用高斯井模拟环形星系
**主管：** Ioanna Manolopoulou
**适应性：** 适用于理学硕士统计和理学硕士数据科学
**描述：**
该项目的目的是为环形星系构建一个模型，以便它们可以自动生成
确定和表征。该项目将使用二维'高斯井'的概念，
其中一个高斯分布由于较小的高斯分布而在中间有一个洞
减去，产生环状结构。将实施这种高斯井的概念
来自环形星系的成像数据，我们使用阈值将图像分成“开”和“关”
像素，'on'被视为观察。

**先决条件：** R中的编程

**参考文献：** Ioanna Manolopoulou，Thomas B. Kepler和Daniel M. Merl。“高斯混合物
井：理论，计算和应用。“计算统计和数据分析56.12（2012）：
3809–3820。



**第32页**

**标题：** 使用Hawkes流程建模计数流程
**主管：** Ioanna Manolopoulou
**适应性：** 适用于理学硕士统计和理学硕士数据科学
**描述：**
标准的霍克斯过程是由均匀的泊松过程构成的

事件根据一些令人兴奋的功能产生后代。霍克斯过程的概括
允许一个非均匀泊松过程的自激过程，具有不同的形式
令人兴奋的功能，以及该过程的多变量版本。这个项目将探索各种
霍克斯过程的各个方面，调查其不同口味的属性，并评估其适合
各种数据集。

**先决条件：** R编程。泊松过程的一般知识。

**参考文献：** Fierro，R.，Leiva，V。，＆Møller，J。（2015）。不同激动人心的霍克斯过程
函数及其渐近行为。Journal of Applied Probability，52（1），37–54。
Hawkes，Alan G.和David Oakes。“一个自我激励过程的集群过程表示。”
Journal of Applied Probability 11.03（1974）：493–503。
Møller，Jesper和Jakob G. Rasmussen。“霍克斯过程的完美模拟。”*进步
*应用概率* 37.03（2005）：629–646。

---

时间： 贝叶斯多状态转移率建模
**主管：** Owen Nicholas和Ardo Van Den Hout
**适应性：** 所有课程
**描述：**
通过一系列状态进行个体的进展，例如健康类别（没有
疾病，1期疾病，2期疾病，.....死亡），或就业（就业，失业，退休），
或者教育等，可以用连续时间马尔可夫多状态模型来描述。 这些是
对于理解状态之间的过渡速率以及模拟轨迹的真实有用
状态。

当涉及从数据转换率的参数贝叶斯分析时，有许多
数值和统计方面的挑战，包括矩阵的有效取幂和
从后验采样的有效方法，以及异质的潜在方法
个人之间的过渡率和先前的选择。该项目将开发贝叶斯马尔可夫



**第33页**

蒙特卡洛链接近转换率矩阵的采样方法，重点关注肺部数据
移植患者。建议使用R，Python或Matlab作为编码的合适包
掌握项目的技术方面。
时间： 创建R包
**主管：** Paul Northrop博士
**适应性：** 适用于理学硕士统计和理学硕士数据科学
**描述：**
R是一种免费提供的统计计算和图形语言和环境。包
是R代码的基本单元，包括可重用的R函数，描述的帮助文档
如何使用这些函数和样本数据。该项目的目的是创建一个R包
其他人，也许是该部门的其他学生，可以用来执行特定的统计
任务。 这将涉及创建一组函数来执行这些分析的各个方面，
注释此代码，提供如何使用每个函数的说明，给出一些说明
例子和写小插图。插图是指向用户提供的包的指南
该软件包解决的问题概述以及它如何解决这个问题。包的目的
需要仔细考虑，在选择之前需要与我讨论

项目。 该软件包不应该非常密切地复制现有R软件包的用途。 一
可能性是编写代码，以便同学们轻松地进行分析
您在统计模块中看到过。显然这个项目适合学生
编程的能力，谁喜欢学习如何在STAT0030中使用R并想要采取
他们的R编程进一步。



**第34页**

时间： 使用贝叶斯极值建模来预先形成开集分类
**主管：** Paul Northrop博士
**适应性：** 适用于理学硕士统计和理学硕士数据科学
**描述：** 在统计分类中，我们寻求识别新对象属于哪个类，
基于包含有关对象（特征）及其对象的定量信息的训练集
已知的类标签。在开集分类中，我们允许将新对象分类为
源自未知类，即未在数据中表示的类。在最近
论文（https://arxiv.org/abs/1808.09902）Vignotto和Engelke使用的模型执行此任务
源于极端价值理论。这个想法是量化一个新的功能有多远
观察来自训练集中的观察。Vignotto和Engelke使用频率论者
处理这个问题。该项目的目的是开发贝叶斯方法并进行比较
它以频率论的方式表现。参加STAT0017主题1的学生将是
熟悉该项目涉及的极值模型。但是，这个项目也是开放的
适合其他学生。

时间： 设置较低的极值阈值
**主管：** Paul Northrop博士
**适应性：** 理学硕士
**描述：** 在极端价值分析中，目标是对未来的极端做出推断
值。 例如，需要设计防洪装置以应对最大的影响
在未来很长一段时间内预计的极端降雨事件。通常，有必要
超出数据推断，即需要对更严重的条件进行保护
而不是记录。极值理论建议可以基于推断的模型。 一
model指定观察量的广义Pareto（GP）分布
超过适当的高门槛。该理论基于尾部的限制行为
当阈值增加时，基础分布高于阈值。在实践中，一个合适的

阈值是根据经验设定的。可以合理地设置阈值越低，数据越多
可用于推理，我们可以预期这些推论更精确。我们能有多低
设置阈值将取决于收敛到限制GP结果的速率。已经注意到了
（Wadsworth，et al 2010，https：//projecteuclid.org/euclid.aoas/1287409386），非线性
数据的转换将影响这种收敛速度。这个项目的目的是寻求一个
可以设置启用较低阈值（更多数据高于阈值）的转换。 学生们
参加STAT0017主题1的人将熟悉该项目涉及的极值模型。
但是，这个项目也对其他学生开放，并且适合其他学生。



**第35页**

**标题：** 具有非比例危险的事件时间建模
**主管：** Aidan O'Keeffe博士
**适应性：** 理学硕士统计学或理学硕士医学统计学
**描述：**
如果感兴趣的结果是直到预先指定的事件发生的时间，则经常进行推断
通过估计危险。危险被定义为事件的瞬时概率
在给定时间发生，并且当目的是比较两个或更多组的危险时，a
通常会假设比例风险。这个假设可能非常严格，可能
在实际数据中经常被违反。该项目将探索建模时间到事件数据的方法
危险可能不成比例的地方。

**标题：** 研究在随机实验中使用置换测试
**主管：** Aidan O'Keeffe博士
**适应性：** 理学硕士统计学或理学硕士医学统计学
**描述：**
当分析来自随机实验的数据时，例如并行组随机控制
试验，在零假设下具有已知分布的检验统计量（通常为零假设
制定并用于确定零假设是否为
被拒绝。 这种方法可以依赖于一些可能难以在某些方面验证的假设
场景。基于分布的测试的替代方案是置换测试，其中观察到
对实验数据进行重新采样，然后从所有可能的结构中构建采样分布
样本。 这些方法依赖于较少的假设，并且可以提供对I型和I型的更多控制
II型错误率。该项目将详细研究排列测试，并探讨其在中的应用
随机实验。将提供一些示例数据集。

**第36页**

**标题：** 比较临床表现和检测异常值
**主管：** Menelaos Pavlou博士和Gareth Ambler博士
**适应性：** 理学硕士统计学或理学硕士医学统计学
**描述：**
常规收集的数据通常用于比较医院绩效（例如，与 –
医院死亡率）并确定具有异常（差）结果的医院。几种统计方法
为此目的建议包括使用漏斗图和随机效应的方法
造型。一些方法也能够纳入风险调整（理想情况下使用经过验证的风险）
预测模型）考虑患者病例组合的差异。该项目将审查和
使用真实手术数据和模拟数据实施其中一些方法。特别是
兴趣是比较这些方法在不同情景中的表现并量化
检测到的真/假异常值的数量。

**标题：** 业余无线电竞赛数据分析
**适应性：** 理学硕士统计学，MSc数据科学
**主管：** Yvo Pokern博士
**描述：**
无线电爱好者经常进行“竞赛”，这是一种每个参与无线电的操作
业余爱好者试图与其他参与的无线电爱好者进行无数的无线电联系
在给定的时间跨度内（例如2小时，24小时或48小时）尽可能的世界。在每次接触中，a
交换了少量信息：标识所涉及的两名无线电业余爱好者的呼号
在接触中，接收信号的强度，操作的无线电频率和运行
标识符（即联系人数）。该信息由每个参与的无线电业余爱好者收集
在比赛日志中，通常通过使用标准软件来捕获日期和时间
联系发生了什么。然后将此日志以电子方式提交给比赛组织者
比赛结束。评分系统用于为每个联系人奖励积分并确定
比赛的获胜者。不幸的是，参与的无线电爱好者提交的数据没有
总是匹配：有时，无线电联系中的两方中只有一方承认
发生了联系，有时信息（频率，标识符，信号强度）没有
匹配，以便可能需要解决一些数据质量问题。

该项目旨在研究联系的频率 – 很可能是联系人
当电离层的反射和折射特性改善时，它变得更加频繁
所用频率的传播。数据可通过在线接口获得



**第37页**

使用脚本查询; 这方面的一个重要组成部分可能使该项目适合
数据科学理学硕士（具有统计学重点）。该项目有一个重要的探索性组成部分
以便学生在可获得的数据的支持下追求其他方向。
先决条件：
1。 愿意学习业余无线电竞赛操作和短波电台的基础知识

传播

2。 具有计算机的通用设施，用于提取数据，格式转换等。

3。 能够操作标准统计软件包来分析数据，例如R.

4。 具有脚本的工具（例如bash / python或类似工具），用于启用Web内容集合

在数据科学理学硕士的框架下进行

**标题：** 协方差矩阵锥上的几何及其在统计推断中的应用

**适应性：** 理学硕士

**主管：** Yvo Pokern博士

**描述：**

协方差矩阵，即正定对称矩阵，在许多领域发挥着重要作用
统计科学。从协方差矩阵的随机样本开始，可以简单地说
使用算术样本均值估计总体平均值。的确，这恰好与之相吻合
假设Wishart分布的最大似然估计。这个意思就是特殊情况
Frechet的更一般概念意味着应用于欧几里德度量。该指标捕获了
协方差矩阵的自然结构和其他距离测量（黎曼，Log-
欧几里德，平方根，乔尔斯基，.....）已被提出。对这些指标及其指标的第一次审核
应用视角的属性可用于

“协方差矩阵的非欧几里德统计量，应用于扩散张量成像”，Ian
L. Dryden，Alexey Koloydenko和Diwei Zhou，Ann。 申请 Stat。，Volume 3，Number 3（2009），1102-
1123。

先决条件：

1。 线性代数的坚实基础：你应该对特征值和特征向量感到放松，

二次形式，规范，内积等

2。 无论是之前暴露于黎曼几何的一些还是愿意接受裸露的

要点速度快



**第38页**

**标题：** 场效应晶体管的统计研究

**适应性：** 理学硕士

**主管：** Yvo Pokern博士

**描述：**

场效应晶体管（FET）是高频电子设备中常用的电子设备。
不幸的是，场效应和制造过程的敏感物理性质导致了这种情况
跨导的大变化，这是表征FET放大能力的参数
信号。除了随机变化之外，已知跨导取决于晶体管
温度很可能取决于制造商。记录的数据集约为50
晶体管可用。此外，该项目可能涉及统计调查的设计
温度效应，进行电子测量和进一步分析
但这是可选的。主要目的是调查不同的分布
不同制造商和FET类型的跨导。这个项目可能会少一些
技术上（数学上）比其他一些项目要求更高，但是也应该清楚地知道
交换更多的勤奋，计算技能和实践技能

可能需要对实际电子元件进行测量以获得类似的标记。

先决条件：

1。 愿意了解FET及其物理学的运作

2。 实验设计基础，线性模型/ ANOVA

**标题：** Smeary中心极限定理

**适应性：** 理学硕士

**主管：** Benjamin Eltzner博士（哥廷根大学外部）和Yvo Pokern博士（UCL）

**描述：**

从概率论，我们习惯于中心极限定理（CLT），它确定了a
零均值独立，相同分布的方可积随机变量序列，
它们的平均值乘以 包含的变量数的**平方根**收敛
分布到正态分布的随机变量，其均值为零且方差相同
序列中的变量。虽然这对于实值随机变量是正确的，但是随机的
在其他集上取值的变量，例如在单位球面或圆上可以表现 **得更慢**
收敛，即需要乘以比一半更小的功率（即平方根）和
随着样本量的增长，置信区间比平时缩小得更慢。得到的CLT是
由于Frechet均值的分布（样本的替换），称为“污点”



**第39页**

是指欧几里德以外的空间）。虽然一些人口分布在圈子和
已知球体导致“涂抹”CLT，一个悬而未决的问题是这些是多么普遍
分布是。这个问题对理论（分析/概率论）和理论都是开放的
经验（计算机模拟）研究。该项目旨在（i）审查CLT的标准版本
（ii）阅读两篇关于“污点”CLT的论文，以及（iii）对某些人进行模拟和/或计算
具体例子。

先决条件：概率论的坚实基础，包括严格的CLT证明。 一些
数学分析的背景，例如通过数学本科学位。

**标题：** 人体髋关节和膝关节运动的模型

**适应性：** 理学硕士统计学，医学统计学硕士

**主管：** Yvo Pokern博士

**描述：**在人类运动模式的研究中（对于医学问题很重要，
人体工程学和运动科学），兴趣在于推断潜在骨骼的运动
所谓的“皮肤标记”的运动 – 粘在皮肤上的反光贴片。有宝贵的
在基础骨骼运动和皮肤运动的情况下，可获得的数据很少
标记已同时记录。一些数据（也可用于此
项目）已经使用简单的普通最小二乘（OLS）估计进行了分析
开放到实质性的统计改进。数据和这些OLS估计的概述是
可用

“用于生成逼真的大腿软组织假象的髋关节运动学驱动模型”，V。
Camomilla等。，Journal of Biomechanics 46（2013）625-630。

该项目的目的是应用更复杂的模型和估算方法，例如
采用具有可变选择或可能是主成分的广义最小二乘法
分析寻找几个皮肤标记的共同运动。

先决条件：R的数据分析能力很强。愿意使用数据（已经发布
在科学文献中记录在人体尸体的实验中。



**第40页**

**标题：** 人口动态和SDE
**适应性：** 理学硕士
**主管：** 马蒂娜拉西亚斯
**描述：**
该项目的主要目的是研究经典的随机微分方程（SDE）模型
人口动态的背景。我们将关注两种不同类型的随机性：环境
和人口统计学。这是一个主要以理论为导向的项目，学生将接触到这个项目
随机微分方程和Itô演算。

**标题：** 随机微分方程及其应用
**适应性：** 理学硕士
**主管：** 马蒂娜拉西亚斯
**描述：**
请直接联系Rassias博士讨论可能性 （ m.rassias@ucl.ac.uk ）

**标题：** 组成属性的条件和测试
**适应性：** 硕士统计学和理学硕士数据科学
**主管：** Kayvan Sadeghi
**描述：**
众所周知，如果随机变量A和B以及随机变量A和D是独立的，则A是
不一定独立于（B，D）。在调节时的这种性质的概括
附加变量C称为\ emph {composition}属性。该项目的目标是
调查哪些发行版满足此属性。这可以通过理论方法完成
（特别是使用代数统计或指数族模型理论），也可以通过
模拟研究。另一种方法是设计测试组合是否的算法
满意的给定数据。

**第41页**

**标题：** 离散行列式点过程的限制性
**适应性：** 硕士数据科学
**主管：** Kayvan Sadeghi
**描述：**
离散决定点过程（DPPs）是用于排斥的随机过程，其模型在
数学上优雅和一般的消极方式
协会。 最初在量子物理学中发展，DPP在其他领域自然产生，例如
组合学，随机矩阵理论，概率和代数。该项目的主要目标是
通过研究它们相应的\ emph {kernel}矩阵，确定这些模型的限制性。
这可以通过理论方法完成，但最有可能通过模拟研究完成。 一个
在Matlab，Mathematica中需要良好的线性代数知识以及编程技巧，
或类似的编程语言。

**标题：** 用于指数随机图模型的模型多面体
**适应性：** 硕士统计学和理学硕士数据科学
**主管：** Kayvan Sadeghi
**描述：**
研究随机网络模型的几何形状提供了一个系统的
网络模型中的参数估计框架以及理解渐近性
这些模型的行为。特别是，找到模型多面体的角是一项重要的任务
用于最大似然估计。该项目的目标是理解和审查
关于网络模型多面体的文献。特别是，我们希望表明，对于非
指数族形式的可交换网络模型，模型多面体的角落是
通过相应的足够统计数据可唯一实现的图形。 一个例子
这个陈述是beta模型，其中模型多面体的角是图
可以通过度序列唯一地实现。
另外，基于对可交换网络的理解，我们可以说什么呢
楷模？需要一些凸几何和指数族模型的基本知识。



**第42页**

**标题：** 不完全竞争电力市场中不确定性下的产能扩张
**主管：** Afzal Siddiqui博士
**适应性：** 理学硕士
**描述：**

放松对电力行业的管制已经引发了竞争和波动的价格
部门。现在，利润最大化的电力公司对新一代产能进行投资
并且对采用时间有自由裁量权。因此，他们的决定必须考虑到
考虑竞争企业的行为和需求的不确定性。传统的真实
期权模型是短视的，因为它们没有装备来处理公司的投资这一事实
决定应取决于其现有的产能库存。通过扩展均衡建模
Gahungu和Smeers（2012）的框架，拟议的项目将解决方法论
关于现有技术的局限性和对寡头垄断行为的综合见解
放松管制的行业。

参考文献：普林斯顿大学AK Dixit和RS Pindyck的“不确定性投资”（1994年）
按; J. Gahungu和Y. Smeers撰写的“电力扩张的实物期权模型”（2012年）。

**标题：** 开关选项的评估
**主管：** Afzal Siddiqui博士
**适应性：** 理学硕士
**描述：**
许多资本密集型项目包括从一种操作模式切换到的灵活性
另一个。 例子包括在能源生产中替代输入燃料的可能性
选择制造业的组成部分。重视这些选择的灵活性需要
分析响应于不确定输入和/或的切换操作模式的最佳时序
产出价格。实物期权理论提供了一个合适的框架来执行这样的
评估并将部署在该项目中，以评估能源中的多个切换选项
制造业。

参考：普林斯顿大学的AK Dixit和RS Pindyck的“不确定性投资”（1994年）
按。



**第43页**

**标题：** 电力部门的风险管理
**主管：** Afzal Siddiqui博士
**适应性：** 理学硕士
**描述：**
向能源可持续性过渡需要电力公司采用更多可再生能源
能源技术，通常具有间歇性输出。同时，放松管制
电力行业增加了能源价格的不确定性。因此，电力公司需要
适当的风险管理策略，以投资风能等可再生能源。在这个项目中，一个随机的
将采用编程方法来表示能源价格和风速的不确定性
基于时间序列分析的场景。随后，将优化这些方案
用于找到最佳的金融对冲头寸。

参考：电力市场不确定性下的决策（2010年），作者：AJ Conejo，M。
Carrión和JM Morales，Springer。

**标题：** 美国通货膨胀的福利成本

**主管：** Cemil Selcuk博士

**适应性：** 理学硕士

**描述：**

即使在充分预期的情况下，通货膨胀也会带来社会福利成本。用罗伯特卢卡斯的话来说：“在一个
货币经济，试图让别人持有非货币符合每个人的私人利益
有息现金和储备金。但是有人必须坚持下去，所以所有这些努力都必须简单
抵消。我们每年都花费几个小时来完成这项工作，我们雇佣了数千名才华横溢的人才
和训练有素的人来帮助我们。这些人小时被简单地扔掉，浪费在任务上
根本不应该执行。“
该硕士论文的目的是估计上述美国的福利损失。 该
该研究所需的数据（GDP增长，M1，利率等）在圣路易斯FED公开发布
网站。

**参考文献：**

- “通货膨胀和福利”，作者：Robert E. Lucas，Jr.，*Econometrica,* Vol。68，第2号（2000年3月），第247-274页



**第44页**

**标题：** 英国通货膨胀的福利成本

**主管：** Cemil Selcuk博士

**适应性：** 理学硕士

**描述：**

即使在充分预期的情况下，通货膨胀也会带来社会福利成本。用罗伯特卢卡斯的话来说：“在一个
货币经济，试图让别人持有非货币符合每个人的私人利益
有息现金和储备金。但是有人必须坚持下去，所以所有这些努力都必须简单
抵消。我们每年都花费几个小时来完成这项工作，我们雇佣了数千名才华横溢的人才
和训练有素的人来帮助我们。这些人小时被简单地扔掉，浪费在任务上
根本不应该执行。“
该硕士论文的目的是估计上述英国的福利损失。数据
研究所需（GDP增长，M1，利率等）在ONS和BoE公开发布
网站。

参考文献：
“通货膨胀和福利”，作者：Robert E. Lucas，Jr.，Econometrica，Vol。68，第2号（2000年3月），第247页 – 274

**标题：** 能源市场中的博弈论投资分析

**主管：** Afzal Siddiqui博士

**适应性：** 理学硕士

**描述：**

为了获得能源，将需要投资新一代和输电能力
部门变得更加可持续。但是，电力公司和网络规划人员需要这样做
说明竞争和市场力量对设计其最佳能力的影响 –

扩张战略。由于互补性建模包括市场均衡条件
反映每个代理商的利润或福利最大化的一阶条件，它提供了一个解决方案
任何代理人都不会有单方面偏离的动机。在这个项目中，这样的游戏 -
理论方法将用于分析投资激励，例如可再生投资组合
标准化（RPS）或限额与交易（C＆T）系统，在放松管制的能源部门。

参考文献：能源市场中的互补模型（2012），作者：SA Gabriel，AJ Conejo，JD
富勒，BF霍布斯和C.鲁伊斯，斯普林格; 发电和发电投资
传输：不确定性下的决策（2016）由AJ Conejo，L。Baringo，SJ Kazempour，
和Siddiqui，Springer。



**第45页**

**标题：** 具有“弱联想”的因果网络
**主管：** 里卡多席尔瓦博士
**适应性：** 所有课程
**描述：**
因果网络是在某些条件下允许的因果关系的表示
使用观测数据估计因果效应。然而，这种结构可能很难
仅从背景知识中获取。存在允许的机器学习算法
估计部分结构，但如果数据中的关联较弱，则它们可能不可靠。 我们会
调查对弱关联有效的方法。 看到
http://auai.org/uai2017/proceedings/papers/229.pdf获取（专业）阅读的一个例子
一些情况，但我不希望学生在这个阶段完全理解这篇论文 - 这是更多
显示一些动机。

**标题：** 强化学习中的因果关系
**主管：** 里卡多席尔瓦博士
**适应性：** 所有课程
**描述：**
强化学习领域包括学习规划一系列动作的方法
根据在估算过程的任何阶段收集的数据，选择有希望的候选人。
这通常需要大样本量和与环境的持续交互。 在这
项目我们将探索利用观测数据来帮助强化学习的方法。 看到
https://arxiv.org/abs/1812.10576举个例子。

**标题：** 模拟社会和空间网络以及其他相关数据中的因果效应
**主管：** 里卡多席尔瓦博士
**适应性：** 所有课程

**描述：**
请参阅https://qaps.princeton.edu/sites/default/files/q-aps/files/spatial-kriging-2016-04-01.pdf
背景

**第46页**

**标题：** 使用自然实验模拟时间序列数据的因果效应
**主管：** 里卡多席尔瓦博士
**适应性：** 所有课程
**描述：**
有关问题和方法的示例，请参阅https://ai.google/research/pubs/pub41854。

**标题：** 以学生为主导的项目
**主管：** 里卡多席尔瓦博士
**适应性：** 所有课程
**描述：**
我对学生主导的项目持开放态度，例如深度学习图形，变分自动编码器等
以及与高效和有效的近似推理算法相关的其他问题
复杂的概率模型。

**标题：** Spotify数据对用户行为的大规模分析
**主管：** 里卡多席尔瓦博士
**适应性：** 硕士统计学和理学硕士数据科学
**描述：**
使用公共Spotify和相关数据开发用户行为模型的几个机会。
下面列出了感兴趣的数据集和感兴趣的问题的描述。学生是
鼓励他们提出他们自己的有趣问题的变种。请注意固体
编程技巧是必需的。
音乐流媒体会话数据集
https://arxiv.org/pdf/1901.09851.pdf
https://www.crowdai.org/challenges/spotify-sequential-skip-prediction-challenge
Spotify Million播放列表数据集
https://recsys-challenge.spotify.com/readme
Last.fm音乐事件数据集
http://www.cp.jku.at/datasets/LFM-1b/
WSDM 2018音乐推荐数据集
https://www.kaggle.com/c/kkbox-music-recommendation-challenge/data
WSDM 2018流失预测数据集
https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data



**第47页**

**标题：** 地震的Shazam – 区域地震目录的应用
**主管：** Katerina Stavrianaki博士
**适应性：** 硕士统计学和理学硕士数据科学

**描述：**
近年来，统计地震学正在经历数据量的快速增长。
地震检测 – 连续数据中地震事件的识别 – 是一个基础
观测地震学的操作。
缺少余震，特别是在大地震之后，可能会导致错误的估计
统计模型中的参数以及偏差预测和地震危险评估。
学生将使用指纹和相似度阈值（FAST），一种地震检测
基于歌曲匹配应用程序Shazam的算法，由斯坦福大学的一组地震学家开发
大学从他们选择的地震目录中识别失踪的地震。
该项目的目的是使用区域地震目录识别失踪的地震
通过处理地震中记录的波形来检测地震的算法
网络。
FAST算法将提供给学生，但算法仅在Linux系统上运行。
还需要了解Python。
参考文献：Yoon，CE，O'Reilly，O.，Bergen，KJ，＆Beroza，GC（2015）。地震检测
通过计算有效的相似性搜索。科学进步，1（11），e1501057。

**标题：** 解聚地震目录数据
**主管：** Katerina Stavrianaki博士
**适应性：** 硕士统计学和理学硕士数据科学
**描述：**
流行型余震序列（ETAS）模型是最广泛使用的统计模型
地震学。这是霍克斯型自激式随机模型及其相应的过程
描述地震发生相当于每个点都可以的分支过程
作为背景事件生成或由特定的先前地震触发。 这允许
我们确定地震群，这种识别对于地震灾害是有用的。目标是
这个项目是使用机器学习聚类方法并将结果与之比较
ETAS模型的结果。R编程语言技能是必要的。



**第48页**

**标题：** 参数时间依赖的多状态生存模型
**主管：** Ardo Van Den Hout博士和欧文尼古拉斯博士
**适应性：** 所有节目
**描述：**
多态模型是两态生存模型的扩展。而不是只有一个事件
时间（死亡的时间，比方说），有多个事件时间（状态之间的转换时间）。 一个
示例是心脏同种异体移植血管病变（CAV）等级的纵向数据模型。数据
CAV可从心脏移植患者的后续研究中获得。定义了四种CAV状态：
无CAV（状态1），轻度/中度CAV（状态2），严重CAV（状态3）和死亡（状态4）。
有趣的是灵活的参数模型，可以描述最初增加的过渡危险，
并在以后减少。
该项目首先探索基本多状态生存模型的R包，并对其进行编码
相应的似然函数。下一步是扩展到更灵活的模型。

**标题：** 双变量离散分布模型认知功能

**主管：** Ardo Van Den Hout博士

**适应性：** 所有节目

**描述：**

纵向数据可用于老年人群的认知功能。使用两个认知
测试，认知功能在双变量离散量表上重复测量。该项目是关于
使用双变量离散分布来描述随时间变化的认知功能
个人。
该项目从每个人一次测量定义的数据开始。二项分布
并将适用其延期。接下来，将为其指定随机效应模型
重复测量。
软件：R。



**第49页**

**标题：** 生存数据的广义时间依赖逻辑模型

**主管：** Ardo Van Den Hout博士

**适应性：** 所有节目

**描述：**

广义的时间依赖逻辑系列包括各种时间到事件的模型
数据。 该项目将研究这些模型，并将它们与标准模型进行比较，例如
Weibull和Gompertz。
该项目的数据来自肺部细支气管炎闭塞症综合征的纵向研究
移植接受者。可以将死亡的标准生存模型定义为事件。另外，一个
可以定义三态生存模型，包括两种生存状态（存在和不存在）
综合症）和第三个吸收死亡状态。对于这两种模型，广义时间
依赖后勤家庭将被调查。
该项目从标准生存模型和相应的似然函数开始。 该
可以使用通用优化器最大化似然函数。下一步是
扩展到三态模型。软件：R。

**标题：** 多次测试中的错误发现控制

**主管：** 王腾尧博士

**适应性：** 适用于理学硕士统计和理学硕士数据科学

**描述：**

近年来，统计假设检验的格局发生了重大变化。
数千种假设的同时测试现在在很多方面都是司空见惯的
应用领域。自Benjamini和Hochberg（1995）的开创性工作以来，许多程序都有
有人建议在多次测试的情况下控制错误发现率。在这个项目中，

候选人有望调查文献并比较现有方法的表现
（根据经验或理论上）。一个更有野心的候选人可以看看可能的修改
现有方法在某些特定设置中实现更好的性能。

**先决条件：**无。

**参考**

Benjamini，Y。和Hochberg，Y。（1995）控制错误发现率：实用且有力
多重测试的方法。 *J.罗伊。中央集权。Soc。，Ser。乙*，**57**，289--300。



**第50页**

**标题：** 矩阵完成

**主管：** 王腾尧博士

**适应性：** 适用于理学硕士统计和理学硕士数据科学

**描述：**

缺少数据是大数据时代的规则而不是例外。例如，在着名的
Netflix用户电影评级数据集，只观察到所有可能评级的很小一部分。 至
能够预测可观察的评级缺失评级是非常有价值的，Netflix曾经提供过
一百万美元的现金奖励，以找到最佳的缺失值预测算法。许多人的一个关键想法
表现最佳的方法是通过求解凸来执行低秩矩阵完成
优化问题。该项目的目标是概述低等级矩阵
从统计角度看完成方法。然后，候选人可以看看
矩阵完成方法的性能保证或将算法应用于现实世界
候选人选择的数据集。

**先决条件：**线性代数（特征分解，矩阵规范）和一个良好的命令
编程语言（例如Matlab，python）。

**标题：** 高维主成分分析

**主管：** 王腾尧博士

**适应性：** 适用于理学硕士统计和理学硕士数据科学

**描述：**

主成分分析（PCA），涉及将多变量数据样本投影到
由样本协方差矩阵的前导特征向量跨越的空间，是最古老的和
统计学中使用最广泛的降维技术。然而，约翰斯通的工作和
Lu（2009）和Paul（2007）表明，PCA在高维设置中崩溃了
在许多现代应用领域经常遇到。经典PCA的几种修改
此后提出在各种结构假设下解决这个问题。这个项目会
比较不同的高维PCA方法，并将一种特定方法应用于a
候选人选择的高维数据集。

**先决条件：**线性代数（特征分解，矩阵规范）和一个良好的命令
编程语言（例如R，Matlab，python）。

**参考文献：**

Johnstone，IM和Lu，AY（2009）关于校长的一致性和稀疏性
高维分量分析。 *J. Amer。中央集权。协会。*，**104**，682--693。
Paul，D。（2007）用于大尺寸尖峰协方差的样本特征结构的渐近性
模型。 *中央集权。报。*，**17**，1617--1642

**第51页**

**标题：** 优化的公民科学数据汇总，用于生物医学图像分析
**主管：** 薛景浩博士; 马丁琼斯博士（弗朗西斯克里克研究所）
**适应性：** 所有节目
**描述：**
众包的“公民科学”分析的使用已被证明是分析中的一个有价值的工具
大量数据，特别是在人类视觉处理仍然优于现有的任务中
计算方法。继其他研究领域的成功项目之后，如
Galaxy Zoo [1]，我们的项目Etch a Cell从数千名非专家那里获得图像分割
我们的体积电子显微镜数据的志愿者[2]。关键的一步是聚合这些
数据，其中来自多个用户的注释被组合以创建最终的高质量注释
对于每个图像。该项目旨在开发一种强大而优化的方法来实现这一目标
聚合，以帮助提供下游机器学习的培训数据。
[1] Lintott等人，银河动物园：从斯隆的星系视觉检查得到的形态
数字天空调查（2008年）MNRAS doi：10.1111 / j.1365-2966.2008.13689.x
[2] Peddie＆Collinson，探索第三个维度：体积电子显微镜来自
年龄（2014）Micron doi：10.1016 / j.micron.2014.01.009

---

**标题：** 半监督机器学习分类卵巢癌的拉曼图像
**主管：** 薛景浩博士; Geraint Thomas教授（伦敦大学学院细胞与发育生物学）
**适应性：** 所有节目
**描述：** 来自卵巢癌患者的高光谱拉曼数据集需要分类
成为两种癌症之一。然而，正如许多生物医学问题一样，它只有少数
标记的拉曼图像（每组9个）。该项目旨在开发一种半监督机器
用于对此数据集中的数据进行分类的学习方法。

---

**标题：** 对胃肠癌的聚类拉曼数据进行分类
**主管：** 薛景浩博士; Geraint Thomas教授（伦敦大学学院细胞与发育生物学）
**适应性：** 所有节目
**描述：**
拉曼光谱仪之间的一致性尚未建立，导致临床缓慢
采用该技术。在SMART数据集中，三种不同的拉曼光谱仪
中心用于将胃肠道（GI）癌症分为5组，产生三种相似的癌症
具有分层结构的数据集。该项目旨在开发一种分类方法
考虑此数据集的此结构。



**第52页**

**标题：** 从拉曼图像分类伪癌与息肉癌
**主管：** 薛景浩博士; Geraint Thomas教授（伦敦大学学院细胞与发育生物学）

**适应性：** 所有节目

**描述：**

描述：上皮错位是一种良性过程，可以在结肠样本时发生
处理。它们经常被误诊为腺癌，因为它们具有许多视觉特征
（渗入粘膜下层）。通过拉曼光谱成功分类可以减少
误报的数量。该项目旨在开发一种分类36拉曼数据集的方法
图像变成息肉癌或假性癌。

---

**标题：** 结肠癌评估的kappa统计的Meta分析

**主管：** 薛景浩博士; Geraint Thomas教授（伦敦大学学院细胞与发育生物学）

**适应性：** 所有节目

**描述：**

描述：评估癌症的病理学家之间的评估者间可靠性和评估者内部可靠性
早就注意到了。但是，尚未在荟萃分析中收集数据以证实这一点
统计效应。该项目专注于结肠癌，旨在开展系统的文献
考虑到统计异质性，对衍生的kappa统计数据进行回顾和荟萃分析
研究之间。

---



**第53页**

**标题：** 基于表示的分类

**主管：** 朱瑞和薛景浩

**适应性：** 理学硕士统计学，MSc数据科学

**描述：**

基于表示的分类方法已成功应用于各种领域，例如
作为人脸识别和高光谱图像分类。着名的代表性
分类方法包括基于稀疏表示的分类（SRC）[1]和协作
基于表示的分类（CRC）[2]。SRC和CRC首先使用all表示测试实例
训练实例，然后将其分类到具有最小重建误差的类。最近
研究建议选择k–最近的类（KNC）来表示测试实例，而不是使用
所有训练实例[3]。该项目旨在研究和改进基于表示的分类
方法，以及它们在真实世界数据上的比较。

**参考**

[1] Wright，John，Allen Y. Yang，Arvind Ganesh，S。Shankar Sastry和Yi Ma。“通过强大的人脸识别稀疏表示。“ *IEEE模式分析和机器智能交易* 31，no.2（2009）：210–227。

[2] 张磊，孟阳，冯祥初。“稀疏表示或协作表示：哪个帮助人脸识别？“ *2011年国际计算机视觉会议*，第471–478页.IEEE，2011。

[3] 郑成勇，王宁宁。“用k近似类的协作表示“ *模式识别字母* 117（2019）：30–36



**第54页**

以下内容摘自“研究生手册”第25–27页

**研究项目**

**准备和提交指南**

学生应该在开始之前计划在笔试后短暂休息
从事他们的项目。所有监管人员都可能在此期间不时离开
6月至9月期间，参加会议或度假。因此学生应该看到
他们的上司一旦考试结束，就能互相方便
开始他们的项目工作的安排。

在项目过程中，学生和主管应安排定期会面
（大约每周一次，尽可能）并且应该同意适当的时间表
完成工作并制作书面帐户。主管应该建议
学生开始写作，并询问主管对他们写作的反馈，
在这个时期的早期。

主管将至少对一份项目论文的整个草案提供反馈意见
在提交截止日期前至少三周内提供。
在此截止日期之后的任何反馈请求由主管决定。
主管应在两周内提供反馈。

最终（经过文字处理的）论文应该交给教学办公室
在广告日期的16:00（通常在9月初）。晚了
提交的内容将导致严重的“迟到”处罚（参见“迟交提交处罚”部分）
在第30页）。此外，论文的电子版应通过提交
Moodle在同一天（MSc Tutor将传递更接近的详细说明
日期）。

项目论文的篇幅取决于项目的主题，可能会有所不同
相当。长度在8,000到15,000字之间（不包括计算机程序，
表格，图表，公式和其他输出）通常是可以接受的。典型的项目是
长度在10,000到12,000字之间。

每篇论文应包括目录，介绍，结论或
讨论部分和参考列表。参考列表应包括所有参考
已被用于支持项目中报告的工作; 这些参考应该
在论文的文本中引用，以表明它们的使用位置，
遵循公认的引用惯例。页面应清楚编号和
应该有一个至少2厘米的左边距。审查员 *非常*重视
准确性，清晰度和整体演示质量。

除了项目论文，每个学生都需要进行演示
他们的研究。通常分配给每个演示文稿的时间是15分钟，不包括
的问题。 学生应参加并积极参加口头报告
其他学生。演讲通常在9月初举行; 因此学生
需要确保此时部门可以使用它们。

上述第三和第四段所述安排的具体日期将是
单独提供。 *请确保您了解它们。*



**第55页**

**评估指南**

项目论文由两名审查员独立阅读，其中一名通常是
候选人的项目主管。每位审查员都提供简短的书面评估。 一个
访问考官也会阅读论文的选择。最终标志得到了同意
整个考试委员会，其中包括访问考官。最后的标记应该是
按照第15页的指导说明进行解释。

审查员将确信论文是候选人的工作，并将
考虑到以下几点：

- 项目 的难度和新颖性;
- 学生所掌握的新方法/应用知识的数量
  需要学习;
- 项目主管要求的指导程度;
- 学生在整个项目中的进步。

根据这些总体标准，审查员将同时考虑论文的内容
及其呈现，更高优先级附加到内容。考虑的方面会
通常包括以下内容：

- *内容*： 完成的工作量; 理解程度
  证明; 推理的质量和准确性，解释的有效性，相关性
  结论; 批判性评估，限制的讨论和进一步的建议
  工作; 目标明确; 文献综述质量; 数据组织和质量
  收集（如适用）; 编程或使用软件的质量（如适用）。
- *演示*： 论文的布局和展示中的关注; 的结构
  论文; 在选择材料时使用适当的判断; 清晰的表达，
  可读性和连贯性; 语法和拼写的正确性; 充足性
  图表，图表和表格（如适用）; 数学的呈现质量
  材料（如适用）。

如果材料尽管是正确的，则将被判定为少于50的商标
以纯技术方式转载。

对于85岁以上的商标，预计该学生除了提交了一份以外的
提出论文，展示了对材料和a的良好理解
相对较高的工作量，也会表现出一些主动而非简单
以下说明。除此之外，90或更高的标记可能是合适的
技术或概念上的材料难度很高，或者在某些地方工作
可以被视为学生的原创研究。

项目论文的篇幅将取决于项目的主题，可能会有所不同
相当。长度在8,000到15,000字之间（不包括计算机程序，
表格，图表，公式和其他输出）通常是可以接受的。典型的项目是
长度在10,000到12,000字之间。超长的论文将受到惩罚（见
第30页）。通常要求完成和演示的工作量很高
足够的，并且材料以一种可以理解的方式呈现给同学们
相似的背景（所以8,000个单词可能只是一个非常适合的长度

理论或密集论文）。另一方面，论文不应该是
过于重复或包含不必要或不相关的细节，这可能导致标记。



**第56页**

虽然上面给出的字数不包括附录，表格和节目清单，
如果这些物品过多，也会受到处罚。

每个项目的介绍将由两名审查员进行评估。通常，都不是
审查员将成为候选人的主管。审查员对此进行独立记录
在讨论和商定商标之前的陈述。通常考虑的方面
包括以下这些：

- *内容*： 演示文稿有趣吗？它是否专注于重要方面
  逻辑上的工作和流程？是否有足够的细节可以在统计上理解
  有识字的听众谁没有对特定主题的深入了解？是
  有明确的目标和结论吗？
- *演讲技巧*： 口头表达自信且清晰可听
  不同的拐点？演示文稿是否与观众互动？是视觉辅助工具
  清晰，生产良好，使用良好？问题处理得当吗？是的
  适合允许时间的材料量？