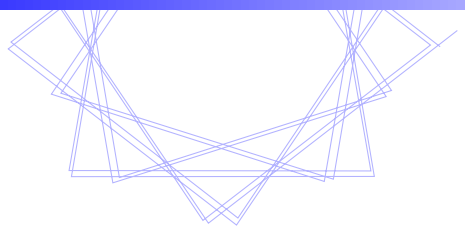
编委

目录

第 I 部分



数统札记

珞珈数学
Mathematics of Luojia March 2018

写在本书的最前面

齐民友

数学是为了什么？很难有一个答案。因为各种不同的答案其实都是相互交叉的。但是，大体上有两种：一是为了某个实际的目的。这不一定不好。例如学习某个其他学科需要一点数学知识。我们有针对性地学习有关数学知识，这是很正常的。也就是说，把数学当作一块敲门砖，这不是不可以，问题在于敲什么样的门。大家都在批评应试教育，好像责任全在学校和教师，这就很不公平。我们都看到有的人就是为了混一个什么才来学数学，甚至“动机”很不怎么样，难道也怪应试教育吗？也还有另外一种学数学的人。目的应该说是很高尚的：为了祖国和人民，为了进一步探索科学真理，等等。还有很多很复杂的情况，不可一概而言。更不能简单地用好坏二字划分。

但还有一个比较简单的划分方法。有一种情况是：把现有的教材学到能 OK 就行了。另一种情况就是，还需要或自己感到有一种要求，要多懂一些，多学一些，也希望自己能参与创造进一步的知识。现在提出建设创新型国家，这里有许多政策性质，经济性质的问题，对于一个学校里的学生和教师，似乎有一点距离。但是，把学校生活，把学数学的过程，变得更充满活泼的创造气氛，是我们共同的愿望。只有这样，才能做到多懂一些，多学一些，自己也能参与数学的创造。而且在大学里生活的更充实一些，更愉快一些，更有意思一些。创造有三种境界：创造的欲望，创造的磨练（说句大家听起来可能有点烦的话，做习题是最简单也最不可少的磨练，数学不是看懂的而是算懂的），创造的愉悦。希望《珞珈数学》这个刊物能在这方面满足读者。

现在很流行一种“傻瓜书”，例如 WINDOWS XP 傻瓜书，英文是 WINDOWS XP for dummies。特点就是，规定要记忆的几条，背得下来，就叫做学会了。对于傻瓜书也不能一概否定。例如，如果有人要学一点线性代数，目的就是例如一门什么课程，出现了矩阵记号。这样，找来一本线性代数 for dummies 就能满足要求，问题在于有一些书，有一些教材，本来不是傻瓜书，而有丰富的内涵，可是把它当作傻瓜书来读，也是常见的事情。总的特点就是，完全不需思考。所以现在也有反对傻瓜书的网站。希望《珞珈数学》这个刊物能帮助读者们，学会不把很好的课程或者好书当作傻瓜书来读，学会把数学作为一门充满创造精神的科学来体验，享受。

珞珈数学 Mathematics of Luojia March 2018

浅薄的代数学启示录

珞珈数学研习会　基础数学专栏

第一作者：2018 级　数学基地班　胡威

第 1 节　前言

分析、代数和几何是基础数学研究相对较为重要的三大分支，而代数以其抽象性和结构美吸引了无数学者投身其中。大多数中国大学的本科课程往往较少开设如分析般丰富的代数课程，偏微分方程、泛函分析等分析学中较困难的课程，即是非基础数学专业的学生也大多必须要学习，但绝大多数非基础数学专业的学生只学过线性代数，甚至没有学过抽象代数，而基础数学专业的学生也往往对交换代数、同调代数和李代数等学科了解相对不足。本文旨在科普这些代数学中相对而言较为主流，而且十分有用的课程的一些常识，分享一些学习经验，并推荐一些参考书目/文章，以使有兴趣学习更深入代数课程的同学，或是和我一样在代数学学习过程中碰到过困难和阻碍的同学，能够得到些许帮助。

本科阶段一般会接触到的代数课程有线性代数、抽象代数（其实我更愿意称呼为基础代数）、伽罗瓦理论、交换代数、同调代数、李代数和一些简单的群表示论。不同的代数学科之前有着千丝万缕的联系，而且代数本身和分析、几何就有着密不可分的关系，在应用数学中也愈发体现其价值。交换代数本身就脱胎于代数几何和代数数论，伽罗瓦理论源于人们对于解方程千百年来的探索。代数本身有着自己独特的魅力，也在其它学科中展现着其威力。

第 2 节　基础代数

基础代数又称为抽象代数、近世代数，往往会开设在大二，主要讲授群论、环论和域论的基础知识，有时也会介绍模论和伽罗瓦理论。作为第一门系统的介绍和引入代数学的思维和方法的学科，基础代数之于代数学的学习有着地基般的重要意义。实际上，读者对于基础代数的具体内容或已有着充分了解，至少也应有所耳闻，因此笔者只略笔介绍一下三门较为主要的分支的来源、些许脉络、应用和一些学习经验。

2.1　群论

群论顾名思义为研究名为群的代数结构的学科，其在代数中具有基本的重要地位，因为许多代数结构，包括环、域和向量空间等可以看作是在群的基础上添加新的运算和公理而形成的，进而许多群论的结果和方法在这些代数结构中也会有相对应的结论。举例而言，群的同态基本定理、第一同构定理和第二同构定理，几乎可以照搬到环、模等代数结构中。而在群论，通过群的自同构研究其自身的结构、通过研究群作用在不同的代数结构上（这一操作有着极其广泛的英语）来研究群或其它代数结构的应用，所衍生的代数学最为基本的思维方法（研究一个对象不仅需要研究其本身，还应该研究其与其它事物的“关系”），几乎贯彻了整个代数学体系。

群论是如何产生的呢？实际上，群论在历史上主要有三个来源：数论，方程理论和几何学。数论中出现的对群的研究始于欧拉，之后由高斯在对模算术和与二次域相关的乘法和加法的研究中进行了发展。倘若读者了解一些基本的初等数论，在学习完群论后，再回顾初等数论，定会会心一笑。而且，诸如费马小定理等定理，在群论中的证明显然较初等数论的证明方法更为容易。方程中出现的群论主要关联到一种特殊的群——置换群，最早出现在拉格朗日、鲁非尼和阿贝尔等人关于高次方程一般解的工作中。1830 年，伽罗瓦第一个用群的观点来确定多项式方程的可解性。伽罗瓦首次使用了术语“群”，并在新生的群的理论与域论之间建立起了联系。这套理论即为著名的伽罗瓦理论。几何中的群论最早出现在射影几何，并在之后的非欧几何（例如双曲几何）中起到了极其重要的作用。克莱因用群论的观点，在不同的几何学（如欧氏几何、双曲几何、射影几何）之间建立了联系，这就是爱尔兰根纲领。

群论的重要性不止体现为作为代数学学习的地基上，还体现在物理学和化学的研究中，因为许多不同的物理结构，如晶体结构和氢原子结构可以用群论方法来进行建模。于是群论和相关的群表示论在物理学和化学中有大量的应用。

值得一提的是，群论中的有限单群分类是 20 世纪数学最重要的结果之一。该定理的证明是一大批数学家集体努力的结果，它的证明出现在 1960 年和 1980 年之间出版的超过 10,000 页的期刊上。

初学者学习群论时，一定要耐心理解不同概念及其联系，并且一定要掌握用重要例子去理解一些结论。实际上，对于抽象的代数学而言，例子可以说是学习最重要的“救命稻草”，如果不了解一些例子，大部分情况下学完后很容易快速遗忘。同时，群论本身有一些技巧性很强的证明和习题，如果初学时卡住了，可以不用过于较真，着重理解主干的知识和体系，等“数学成熟度”（这一指标或许是真实存在的）到达一定程度时，再回头看一定会豁然开朗。

推荐参考书/资料：

- 丘维声《近世代数》北京大学出版社
  尽管是介绍基础代数的书，但群论占据了其中相当大的篇幅，本书较为全面，而且对初学者极为友好，常常从高等代数、初等数论的一些常识引入新概念，并且整个体系很清晰。

- 徐明曜《有限群导引》科学出版社
  被可靠的学长推荐过的较为进阶的群论的书，适合已经学过基础的群论，想要有更深入的了解的同学。

2.2 环论与模论

环论的研究对象主要是环这一比群更特殊的代数结构，粗略可分为交换环和非交换环，而这两者更多出现在其它代数学科中，事实上，交换环是交换代数主要研究的对象。因此，在基础代数的书中，环论的篇幅往往较小。而模论则是在环论的基础上自然而然需要学习的对象。在交换代数中，往往与环的理想相关的结论，都会有在模上的推广。

在基础代数的环论中，主要是介绍理想的概念及性质，一些特殊的理想（素理想、极大理想），然后是有关于整环的整除性，最后会有一些和域论相关的内容（有限域的构造等），最核心的主线还是理想。其余内容可以看成交换代数、代数数论和域论等学科的铺垫，因而相较于群论，环论“独立”的内容可能较少，而和其它学科相关的知识可能初学时有些难以理解，可以结合相关学科的内容一起学习，并多参考课后习题，但总体而言难度不大。

模实际上是线性空间的推广，因而线性代数的许多结论在模论中都有着相应的对应。可能相较于群论和环论，模论才是现代代数学的基础。在基础代数中的模论往往较为基础，基本上是定义的介绍和印证。初次接触模的概念时，初学者可能觉得有些抽象，不妨类比线性代数的结论进行理解，并多掌握一些例子，这会有助于对概念的理解。

其中主理想的有限生成模理论就是著名的 Jordan 标准型的推广，非常值得了解。

在基础代数的诸多板块中，环与模可能是最为简单的，其中定义部分相对而言最为重要，初学时最好多举一些例子来印证定义，并多做一些习题来巩固定义的理解（这部分的习题可能相较于群论，技巧性没有那么浓），往往经历群论的考验之后，不会觉得这一板块十分困难。

推荐参考书/资料：

- 聂灵沼、丁石孙《代数学引论》
  难度相对而言较大的基础代数书籍，但胜在内容丰富，尤其是基础的环论和模论知识，习题难度较大，但做一做对夯实基本功很有帮助。

- 其它较为全面的基础代数书籍，参考交换代数有关交换环的部分、同调代数有关模的部分。

2.3 域论和伽罗瓦理论

由于域本身更为特殊的代数结构，域没有类似环论中理想的结构，因而我们不再着眼于一个域的特殊子集，而是关注包含某一域的更大的域，这就是域扩张，域论最核心的研究内容。基础代数的域论内容，也大致包括超越扩张、代数扩张、分裂域、正规扩张、可分扩张，进而是伽罗瓦扩张，实际上，有限伽罗瓦扩张就是有限正规可分扩张。不同扩张之间的性质、关联、应用，较为繁琐，笔者就不赘述了。有趣的是，域论能很好地解决一些初等数学难以解决的问题，例如某两个无理数相加是否为有理数、古希腊数学三大难题（三等分角、立方倍积、化圆为方），运用域论解决这些问题是不难的。尽管域论的概念并非十分繁复，域论的习题难度是较高的，有些题目可能需要通过模仿来掌握解决问题的方法，甚至可以说，不进行一定量习题的训练，是很难掌握域论的。

伽罗瓦理论则是本科数学中相当精巧的一个小高峰，由天才数学家伽罗瓦创立。伽罗瓦本人传奇而短暂的一生，也是为人们所广为谈论的。伽罗瓦理论提供了域论和群论的联系，应用伽罗瓦理论，域论中的一些问题可以化简为更简单易懂的群论问题。而伽罗瓦理论最著名的结果即为：

在特征为 0 的域 $\mathbb{F}$ 上，多项式 $f(x)$ 的根可用根式解的充分必要条件是 $f(x)$ 的分裂由 E/F 的伽罗瓦群是可解群。

这一结果广为人知的推论即为：

当 $n>4$时，一般的 n 次多项式无根式解（“一般”意谓将多项式系数视为独立变元），原因是对称群 S_n 在 $n>44$ 时在不可解。

笔者对伽罗瓦理论的掌握也较为浅薄，无法提供较为深刻的洞见，但可以确定的是，在学完群论和域论之后，用伽罗瓦理论作为基础代数学习的检验，本身是很有价值的，而且对于数学能力的锻炼也是相当好的。课堂教学由于课时原因，几乎不可能较为系统地讲解伽罗瓦理论，所以读者不妨在假期挑选一段空闲的时间，好好欣赏几百年前天才创立的杰作。

推荐参考书/资料：

- 章璞《伽罗瓦理论：天才的激情》高等教育出版社
  相当全面和精巧的小书（仅有 135 页），解释性注解充分，证明也很详实，甚至具有一定美学气息。

- GTM167 《Field and Galois Theory》
  据称是讲解伽罗瓦理论最简单易懂的一本书，没有中文译本，并不是很薄（300 页左右），精力充沛且阅读英文教材能力较强的同学可优先采用本书。

第 3 节 交换代数

交换代数最早研究的动机就是为了解决费马大定理，最早的交换代数以交换环的理想（一种特殊的子集，类似群的正规子群）为研究对象，因而最早被称为“理想论”，由戴德金最先开始研究。后来随着代数数论和代数几何的发展，交换代数也逐渐发展成一门成熟的学科。代数几何、代数数论及交换代数在本质上连结的非常紧密，因此有时很难去区分某特定数学原理属于哪个领域。例如希尔伯特零点定理是代数几何的基本定理，但是陈述及证明时都是以交换代数的方式进行。而费马大定理问题的形式是以基本的算术方式（属于交换代数的一部分）呈现，但其证明用到很深的代数几何及代数数论。

传统的交换代数主要包括交换环的一些结构和性质，模论初步，分式环（模）和局部化方法，诺特环与阿廷环—两类重要的交换环，赋值环与戴德金整环，完备化和维数理论等等，具体可以参考交换代数最知名的教材—Atiyah 和 MacDonald 所著的交换代数导引。交换环的性质和模论较为贴近基础代数的内容，可以看做基础代数所学知识的深入和推广，甚至可以作为学习完基础代数的环论和模论后，以自学的方式来检验自己学习成果的一次测试；其中有关于理想的根、素理想与极大理想、准素理想和支集，以及准素分解等的内容，是相当有趣的，而且理解起来不太需要其它更深入的学科，尤其推荐各位学习完基础代数，对更深入的交换环理论感兴趣的同学尝试着了解一下。

而其他内容，尤其是完备化和维数理论等本身有着很强的使用背景，在初学时可能会觉得晦涩抽象，而且很容易遗忘，如果能结合具体的应用背景进行学习，效果会好很多。笔者自己在这一块的理解也不够，不敢妄言，所以也不敢写下自己的理解，有志于学好交换代数的同学不妨找相应的教授寻求指点。

另一处交换代数学习可能遇到的困难是里面所广泛使用的同调代数方法，越现代的书里使用的同调方法越多，初学时无法深刻理解这些方法（可能只知道怎么操作，但不知道为什么要这样操作）可以不用担心，在后续学习同调代数时，便会豁然开朗，甚至可以考虑同时学习同调代数和交换代数，尽管这样精力会不够。

交换代数是的的确确需要学两遍以上的学科，一方面，初学交换代数时可能因为对一些理论的动机和应用不了解而学得粗糙，另一方面，在学习对应的学科时，又会经意不经意间再学一遍交换代数。交换代数可以说是本科阶段难度最大的课程之一了（指学校开设的课程，自学的话多难的课程你都可以自学到），也是最具备代数美学的课程之一，所以在初学时不用过于锱铢必较，有些证明自己无法复现是极其正常的，只需要做到理解证明的核心操作，理解定理和定义并且会使用，必要时可以找一些习题，用模仿的方式来演练学过的内容（实际上，很多习题就是让你模仿一下定理的证明，以达到增进理解的目的），感觉学得不如其它分析学科通畅也没关系，这是极其自然的。

推荐参考书/资料：

- Atiyah 《Introduction to Commutative Algebra》
  最经典的交换代数教材，相对而言比较简练，正文难度不是很高，习题难度较大，但有些处理方法不够现代，如果使用这本教材感觉不是很适合可以考虑换别的教材。
- Allen Altman and Steven Kleiman 《A Term of Commutative Algebra》
  用较为现代的语言和处理改写 Atiyah 的书的一本讲义，在作者主页有 pdf 下载和习题解答，排版十分适合阅读，行文也很清晰，是笔者最为推荐的一本教材。
- Serre 《Local Algebra》
  难度相当相当大的一本书，自学起来难度尤其高，适合已经学过一遍 Atiyah 书的同学提升自我使用。
- 其他中文交换代数书，如冯克勤《交换代数基础》，其实大部分是 Atiyah 的翻译版，有些书会增添一些内容，如果英文阅读某些部分有困难时可以找这些书参考一下相应的部分，但不太建议作为主教材。

第 4 节 李代数

李代数实际上只需要高等代数的基础就可以学习，但想真正理解和运用李代数，当然需要李群和流形的知识。

尽管和李群结合的李代数才能发挥其功力，李代数本身的课题已经足够吸引人。幂零李代数和可解李代数（是不是想到了幂零群和可解群），Cartan 子代数及其判定，半单李代数的 Cartan 分解（弄清一类对象结构的分解往往是吸引人的），以及根系和邓肯图、单李代数的分类、表示理论等，其中，复单李代数的分类和九种邓肯图的一一对应是相当漂亮的结果。这些内容和之前我们所学过的代数内容其实也有这样千丝万缕的联系。可以说，通过李代数这一基础不需要太高的学科来磨炼代数功底是十分恰当的。

本人在李代数上造诣十分浅，所以无法提供有效的建议。但是李代数这一门学科难度还是不小的，可能需要很强的抽象思维和理解能力，尤其是在不具备李群和几何基础时学习。如果在没有学过这些课程，只学过高等代数就学习李代数的话，可以更多关注李代数和高等代数结合较为紧密的部分（幂零和可解李代数，Cardan 分解等等），作为高等代数的练习，遇到难以理解的内容可以放一放，等到学到相应的内容再学习便会豁然开朗。

相较于交换代数和同调代数，李代数的应用价值（对其它学科而言）是更高的，尤其是物理和计算机视觉等领域，此处不多赘述，可以在知乎等平台了解相关应用。所以从实际角度出发，这一门年轻、深浅均可、应用广泛的学科，是值得各位尝试的。

推荐参考书/资料：

- 万哲先《李代数》（第 2 版）高等教育出版社
  中规中矩的教科书，内容较为全面，但可能对初学者而言有一点点难。
- 苏育才等《有限维半单李代数简明教程》科学出版社
  一本很友善的李代数入门书。
- 其它英文版教材，如 GTM9
  有人称赞，但可能过于简明，对初学者不太友善，如果本身数学成熟度较高，可自行探索。
- 朱富海老师的公众号“数林广记”。强烈推荐，尤其是对于低年级的同学而言，其中除了李代数之外的文章也值得一看。

第 5 节 同调代数

同调代数可能是本科接触的代数课程中，抽象程度最高的一门课了，也是笔者而言学习难度最大的一门课，相较而言，同调代数可能更偏向于“语言”。同调代数是一门相对年轻的学科，其源头可追溯到代数拓扑（单纯形同调）与抽象代数（合冲模）在十九世纪末的发展，这两门理论各自由庞加莱与希尔伯特开创。同调代数的发展与范畴论的出现密不可分。大致说来，同调代数是（上）同调函子及其代数结构的研究。“同调”与”上同调”是一对对偶的概念，它们满足的范畴论性质相反（即：箭头反向）。数学很大一部分的内在构造可通过链复形理解，其性质则以同调与上同调的面貌展现，同调代数能萃取这些链复形蕴含的资讯，并表之为拓扑空间、层、群、环、李代数与 C*-代数等等“具体”对象的（上）同调不变量。谱序列是计算这些量的有力工具。作为工具，同调代数是抽象的，也是有力的。

同调代数在代数拓扑中扮演非常重要的角色。代数拓扑的一个重要任务便是运用代数工具给不同（胚）的拓扑空间分类，同调和同论便是其两大支柱。同调代数的影响日渐扩大，目前已遍及交换代数（这在交换代数的教材中有很明显的体现）、代数几何、代数数论、表示理论等前沿学科。

而本科阶段的同调代数书籍和课程，主要包括模论（更深入的）、范畴论、复形和导出函子、群的同调、谱序列等等，不同的教科书可能内容安排差异较大，而且入门难度可能要比李代数（以高等代数为基础）、交换代数（以交换环论为基础）难许多，这时结合具体的例子，也就是代数拓扑进行学习可能会好很多，所以建议在学习代数拓扑等学科，遇到需要学习同调代数的时候，再找相应的材料进行学习，不然陷入“抽象废话”的瓶颈，可能会既消耗精力，又消耗耐心。

不同于交换代数主要为代数几何服务，同调代数对基础数学研究的意义可能更重要，应用也更为广泛，甚至在学会这门语言之后，你可以重新用这门语言解释你之前学过的许多知识，以期更深刻的理解。在学习时遇到挫折是很正常的，越抽象的学科越需要耐心，理解概念，多用例子去验证，多使用定理去证明,逐渐下来抽象的东西会变得具体起来。(坦白讲,笔者在同调代数里还没有做到这一步)

推荐参考书/资料：

- 陈志杰《代数基础：模、范畴、同调代数与层》华东师范大学出版社
  作为高等代数教材的编写者，陈志杰教授同样在编写这本入门书上做到了简明易懂，非常适合初学者入门。
- 佟文廷《同调代数引论》
  被可靠的学长推荐过的同调代数教材，只能说四平八稳，没有太难。
- Weible《An Introduction to Homological Algebra》
  非常经典的教材，或许有些难度，但如果想掌握扎实的同调代数基础，还是需要阅读的，可以考虑先从简单的中文教材读起。

珞珈数学 Mathematics of Luojia March 2018

数学科普：传递美学与趣味

珞珈数学研习会　程子英

第 1 节　数学之美

提及数学，大多数人可能不会把它和“美”这个字眼联系起来，毕竟晦涩高深的理论，庞杂繁复的分支，完备精细的公式，都不会让初见者一开始就感受到它独特的魅力。

然而，亚里士多德曾经说过“硬说数学科学无美可言的人是错误的。美的主要形式是秩序、匀称与明确”，庞加莱也对数学的美多加赞赏“感觉到数学的美，感觉到数与形的协调，感觉到几何的优雅，这是所有真正的数学家都清楚的真实的美的感觉”。

数学科普的最终目的，就是要以轻松简单的方式，向大家介绍数学的美，消除所谓的刻板印象，让大家能够产生对数学的兴趣，燃起自我探索的欲望。

第 2 节　科普之法

数学科普要达到较好的效果，可以从以下几个方面切入。

第一，从历史的角度。对任何一个学科的介绍，都必定离不开对它发展历史的回溯，漫步在历史长河河畔，我们可以从零开始，跟随着先哲们一同见证，一砖一瓦是如何建成如今数学学科的高楼大厦，这里面的参与感，让我们减少了对数学的恐惧与疏离，在这样的心理下，大家自然更能专注于其精妙绝伦之处。此外，数学的发展历程中使用过的方法、前人们走过的弯路，也都为初学者今后的自我探索提供了宝贵的经验。

第二，从学术的角度。数学作为理工科的基础，其学术性不容忽视。在进行科普工作时，我们不能一味地只介绍轻松简单的一面，刻意忽略了其学术性，乃至与之相伴的一定程度的晦涩性。要让初学者明白，数学之所以在科学研究中享有不可小觑的地位，正是由于其严谨性、基础性之强，为其他学科夯实了坚实的理论基础，而这种可靠性必然源自于其超凡的严谨性。这样提前了解到数学学术性的一面，做足了心理建设，才不会在日后学业研究里轻易打退堂鼓。

第三，从未来的角度。几年前的金融，如今的计算机专业炙手可热，大家追捧的未必仅仅是专业本身，它们光明的发展前景也必然在众多簇拥者的考量范围内。因此，数学科普不能仅仅停留在现有的专业知识的介绍上，而应将数学学科的发展前景全面而客观地展现在大家面前，让大家不再因为旁人“学数学不赚钱”、“学数学没出路”的消极评判声中心生疑虑。

珞珈数学 Mathematics of Luojia March 2018

生于悖论，终于革新：三次数学危机

珞珈数学时空堂　数学科普专栏

第一作者：2019 级　数学基地班　刘焱澍
第二作者：2019 级　统计学　加欣怡
第三作者：2019 级　数学与应用数学　崔晗宇

第 1 节　引言

数学，作为一切科学之母，是人类理性思维的最大结晶，是人类认识世界的逻辑基础。很自然的，人们坚信数学是严密、准确、完美的象征，充满了确定性。但回顾数学史，事实却不令人称心如意。数学中的悖论，这直觉与逻辑相冲突的产物，总是似毒龙尼德霍格一般隐藏于数学发展的深根中，令数学之树的理论基础变得危机四伏，数学之树的理论根基不再牢靠之时，就是数学危机的爆发之时。人类重建数学理论基础的过程，也是人类数学智慧的集萃。因此，我们希望通过了解数学危机，帮助我们了解数学的发展与数学智慧。

第 2 节　第一次数学危机

时维今日，实数，尤其是无理数的概念，早已深入人心，显得理所当然。数学分析里，“证明 $\sqrt{n}$（n 非完全平方数）不是有理数”等有关无理数的命题，也已成为“这也要证”的“典范”之一。

但，无理数的出现，真的这么理所当然，自然合理吗？

回想一下数的出现历史：我想知道今天获得了多少个橘子，于是我用自然数去计算；我还想知道要把它分给几个人，要怎么分公平，于是我拿分数去计算。这些都是显然而合理的，于是，毕达哥拉斯说：“万物皆数”(指整数或整数之比)。就是说有了有理数，就够！既能修桥，也能做几何题，又严谨又好。

不得不说，毕达哥拉斯这套理论，挺成功的。结合勾股定理，不仅算数用得上，建筑也用得上，连音乐里的五度相生律，都被他搞了出来。看起来，所有的数学活动，真的只需要有理数！

然而，似乎在嘲笑他的成果，他最出名的成果，那“毕达哥拉斯定理”（中国称勾股定理，也是初等几何里的一个最美的定理，不知多少人听到“勾三股四”，会自动说出“弦五”），其实和他的信条背道而驰。

大约在公元前 5 世纪，毕达哥拉斯学派的希帕索斯从勾股定理出发，提出了：等腰直角三角形的直角边与其斜边不可通约。

直角三角形的直角边与其斜边不可通约，这个简单的数学事实的发现使毕达哥拉斯学派的人感到迷惑不解。

违背了毕达哥拉斯派的信条：“一切量都可以用有理数表示”。所以，通常人们就把希帕索斯发现的这个矛盾，叫做希帕索斯悖论。而新发现的数，由于和之前的所谓“合理存在的数”——有理数形成了对立，所以被称作“无理数”。在这一发现后不久，希帕索斯就死于海难，后世有人怀疑，这其实是毕达哥拉斯学派的一次“光荣谋杀”。

这次危机的出现，标志着世界关于无理数的研究、乃至演绎法在数学中广泛使用的开始。

回顾在此以前的各种数学，无非都是“算”，以应用。即使在古希腊，数学也是从实际出发，应用到实际问题中去的。至于埃及、巴比伦、中国、印度等国的数学，并没有经历过这样的危机和革命，也就继续走着以算为主，以用为主的道路。而由于第一次数学危机的发生和解决，希腊数学则走上完全不同的发展道路，从此希腊人开始从“自明的”公理出发，经过演绎推理，并由此建立几何学体系。

在第一次数学危机的空前烈火后，公理化的纯粹数学开始茁壮成长。

第 3 节 第二次数学危机

当我们翻开富含知识的微积分学课本时，极限往往是最先被要求理解的一个概念，想必不少同学也曾深受极限习题的折磨，心想如果没有极限该多好啊！如果你确实曾这样想过，那么穿越到几百年前的大学数学课堂绝对是一个不错的选择。

古希腊的阿基米德曾提出了“穷竭法”，这大概是最早的微积分学方法了，中国南北朝时期的祖暅与其父为解决球的问题，提出了“幂势既同则积不容异”的祖暅原理，这说明微积分学思想其实有着悠久的历史，但很显然当时的人们并没有提出极限这一概念。

在十七世纪前后，即微积分学初创之时，极限这一概念仍未被提出，当时人们大规模的使用微积分，取得了很多成果，但很少探求微积分基础是否合理，其中最关键的问题是：“无穷小量是不是零？无穷小及其分析是否合理?”两位创始人曾做如此解释：牛顿于 1669 年说它是一种常量；1671 年又说它是一个趋于零的变量；1676 年它被“两个正在消逝的量的最终比”所代替。但是，他始终无法解决上述矛盾。莱布尼兹曾试图用和无穷小量成比例的有限量的差分来代替无穷小量，但是他也没有找到从有限量过渡到无穷小量的桥梁。

后来，越来越多的人开始批判微积分，其中最著名的是：英国大主教贝克莱于 1734 年写文章，攻击流数（导数）“是消失了的量的鬼魂……能消化得了二阶、三阶流数的人，是不会因吞食了神学论点就呕吐的。”他说，用忽略高阶无穷小而消除了原有的错误，“是依靠双重的错误得到了虽然不科学却是正确的结果”。当然，他是出自对科学的厌恶和对宗教的维护，不过也抓住了当时微积分、无穷小方法中一些不清楚不合逻辑的问题。事实上，18 世纪的数学思想的确是不严密的、直观的，强调形式的计算而不管基础的可靠。其中特别是：没有清楚的无穷小概念，从而导数、微分、积分等概念不清楚；无穷大概念不清楚；发散级数求和的任意性等等；符号的不严格使用；不考虑连续性就进行微分，不考虑导数及积分的存在性以及函数可否展成幂级数等等。

事实上，第二次数学危机的缘起——无穷，一直是人类思想的矛盾，大约公元前 450 年，古希腊芝诺注意到由于对无限性的理解问题而产生的矛盾，提出的关于时空的有限与无限的四个悖论，这间接引发了第二次数学危机：

“两分法”：向着一个目的地运动的物体，首先必须经过路程的中点，然而要经过这点，又必须先经过路程的 1/4 点，如此类推以至无穷。——结论是：无穷是不可穷尽的过程，运动是不可能的。

“阿基里斯追不上乌龟”：阿基里斯总是首先必须到达乌龟的出发点，因而乌龟必定总是跑在前头。这个论点同两分法悖论一样，所不同的是不必把所需通过的路程一再平分。

“飞矢不动”：意思是箭在运动过程中的任一瞬时间必在一确定位置上，因而是静止的，所以箭就不能处于运动状态。

“操场或游行队伍”：A、B 两件物体以等速向相反方向运动。从静止的 c 来看，比如说 A、B 都在 1 小时内移动了 2 公里，可是从 A 看来，则 B 在 1 小时内就移动了 4 公里。运动是矛盾的，所以运动是不可能的。

芝诺揭示的矛盾是深刻而复杂的。前两个悖论诘难了关于时间和空间无限可分，因而运动是连续的观点，后两个悖论诘难了时间和空间不能无限可分，因而运动是间断的观点。芝诺悖论的提出可能有更深刻的背景，不一定是专门针对数学的，但是它们在数学王国中却掀起了一场轩然大被。它们说明了希腊人已经看到“无穷小”与“很小很小”的矛盾，但他们无法解决这些矛盾。其后果是，希腊几何证明中从此就排除了无穷小。

直到 19 世纪 20 年代，一些数学家才比较关注于微积分的严格基础。从波尔查诺、阿贝尔、柯西、

狄里赫利等人的工作开始，到威尔斯特拉斯、狄德金和康托的工作结束，中间经历了半个多世纪，基本上解决了矛盾，为数学分析奠定了一个严格的基础。

波尔查诺给出了连续性的正确定义；阿贝尔指出要严格限制滥用级数展开及求和；柯西在 1821 年的《代数分析教程》中从定义变量出发，认识到函数不一定要有解析表达式；他抓住极限的概念，指出无穷小量和无穷大量都不是固定的量而是变量，无穷小量是以零为极限的变量；并且定义了导数和积分；狄里赫利给出了函数的现代定义。在这些工作的基础上，威尔斯特拉斯消除了其中不确切的地方，给出现在通用的极限的定义，连续的定义，并把导数、积分严格地建立在极限的基础上。

19 世纪 70 年代初，威尔斯特拉斯、狄德金、康托等人独立地建立了实数理论，而且在实数理论的基础上，建立起极限论的基本定理，从而使数学分析建立在实数理论的严格基础之上。

第 4 节 第三次数学危机

第三次数学危机，一定程度上不像数学，而像逻辑学危机，它的缘起是罗素悖论，其通俗版本是“理发师悖论”，即：在某个城市中有一位理发师，他的广告词是这样写的：“本人的理发技艺十分高超，誉满全城。我将为本城所有不给自己刮脸的人刮脸，我也只给这些人刮脸。我对各位表示热诚欢迎！”来找他刮脸的人络绎不绝，自然都是那些不给自己刮脸的人。可是，有一天，这位理发师从镜子里看见自己的胡子长了，他本能地抓起了剃刀，你们看他能不能给他自己刮脸呢？如果他不给自己刮脸，他就属于“不给自己刮脸的人”，他就要给自己刮脸，而如果他给自己刮脸呢？他又属于“给自己刮脸的人”，他就不该给自己刮脸。

罗素悖论的基本思想是：对于任意一个集合 $\mathbf{A}$，$\mathbf{A}$ 要么是自身的元素，即 $\mathbf{A} \in \mathbf{A}$；$\mathbf{A}$ 要么不是自身的元素，即 $\mathbf{A} \notin \mathbf{A}$。根据康托尔集合论的概括原则，可将所有不是自身元素的集合构成一个集合 S_1，即 $S_1 = x : x \notin S_1$。理发师悖论与罗素悖论是等价的：如果把每个人看成一个集合，这个集合的元素被定义成这个人刮脸的对象。那么，理发师宣称，他的元素，都是城里不属于自身的那些集合，并且城里所有不属于自身的集合都属于他。那么他是否属于他自己？这样就由理发师悖论得到了罗素悖论。反过来的变换也是成立的。

数学家们通过将集合的构造公理化来排除了这样的集合的存在性。

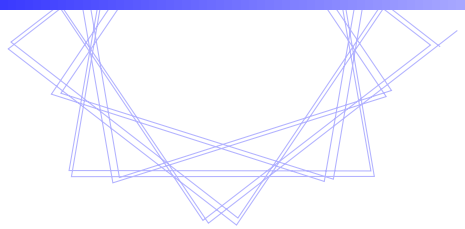
例如，在策梅洛（Zermelo）和弗伦克尔（Fraenkel）等提出的 ZF 公理系统（也称 ZFC 公理系统）中，严格规定了一个集合存在的条件（简单地说，存在一个空集【空集公理】；每个集合存在幂集【幂集公理】；每个集合里所有的集合取并也形成集合【并集公理】；每个集合的满足某条件的元素构成子集【子集公理】；一个”定义域“为 A 的”函数“存在“值域”【替换公理】等），这样无法定义出悖论中的集合。

第三次数学危机就此在一定程度上解决。但数学逻辑的发展仍未告结束，诸如哥德尔不完备定理也是其的成果之一。

第 5 节 总结

三次数学危机，如果只从“算”的角度看，其实都不是危机，因为没有无理数的发现，也不影响我们绘建筑图，不影响我们记账；没有极限的提出，牛顿等物理学家也用了微积分近百年；没有 ZFC 等公理系统的提出，我们仍可以放心大胆的使用集合论，毕竟现实中不会有那种奇怪的理发师。但从数学学习的角度看，我们最需要学习的应当是数学的严谨性，而数学危机的历史就是这种严谨性的绝佳体现。

第 II 部分



数海拾珠

珞珈数学 Mathematics of Luojia March 2018

微分方程：横亘理论与应用的数学

珞珈数学编辑部　　马金韬　邓琪雯

第 1 节　简介

微分方程，是指含有未知函数及其导数的关系式，解微分方程就是找出未知函数。微分方程是数学研究中的一个重要的分支。与第 25 期期刊中介绍的代数学与几何学相比，微分方程在工程问题中的应用更为广泛。在物理运动学、动力学（如最速落径问题、悬链线问题）、化学、工程学、经济学等领域都有广泛而深刻的应用（如人口发展模型、交通流模型）。因而微分方程的研究是与人类社会密切相关的。

第 2 节　发展历程

微分方程历史久远。牛顿和莱布尼茨创造微积分时就处理过与微分方程有关的问题。人们用微积分学研究几何学、力学、物理学所提出的问题时，微分方程问题应运而生。牛顿解决了二体问题：在太阳引力作用下，一个单一的行星的运动。

数学家们最初把精力集中于微分方程的通解，后来证明这一般不可能。

第一，能求得通解的方程显然是很少的。在常微分方程方面，一阶方程中可求得通解的，仅为线性方程、可分离变量方程和用特殊方法变成这两种方程的方程。

第二，当人们要明确通解的意义的时候碰到严重的含糊不清之处，达布在他的教学中经常提醒大家注意这些困难。这主要发生在偏微分方程的研究中。

第三，微分方程在物理学、力学中的重要应用，不在于求方程的任一解，而是求得满足某些补充条件的解。这些补充条件即定解条件。求方程满足定解条件的解，称之为求解定解问题。

于是科学家们转向定解问题：初值问题、边值问题、混合问题等。但是，即便是一阶常微分方程，初等解 (化为积分形式) 也被证明不可能，于是转向定量方法（数值计算）、定性方法，而这首先要解决解的存在性、唯一性等理论上的问题。

第 3 节　联系与展望

常微分方程的形成与发展是和力学、天文学、物理学，以及其他科学技术的发展密切相关的。数学的其他分支的新发展，如复变函数、李群、组合拓扑学等，都对常微分方程的发展产生了深刻的影响，当前计算机的发展更是为常微分方程的应用及理论研究提供了非常有力的工具。

在本章节中，我们将系统地介绍常微分方程与偏微分方程的重要数学理论，同时介绍针对微分方程数值解法的发展，利用常见的 MATLAB 软件给出数值解法的演示程序。希望读者结束本章节学习后，可以对微分方程理论有所熟悉，同时可以根据自己的兴趣上机进行程序编写。

珞珈数学 Mathematics of Luojia March 2018

浅谈常微分方程初值问题的数值解法

2018 级　数学基地班　宣源昊

数学上，凡是表示未知函数的导数以及自变量之间的关系的方程，就叫做微分方程。微分方程是研究函数变化规律的主要工具，应用十分广泛。遗憾的是，只有少数简单的微分方程可以求得解析解。在多数情况下，微分方程难以求出其解析解，故而寻求解析解的近似数值解在研究和应用中就十分重要。

当未知函数是一元函数时，该微分方程即称作常微分方程。其中初值问题是常微分方程求解问题中很重要的一类。本文首先介绍了常微分方程及初值问题的概念，进而推导了其他情况向标准形式的转化。然后给出了常微分方程初值问题数值解的定义，并介绍了两类常见的数值解法。最后，本文通过一个实例，详细介绍了利用 MATLAB 求解常微分方程初值问题的过程及相关语法格式。

关键词: 常微分方程数值解、MATLAB 实现、Euler 法、Runge-Kutta 方法

第 1 节　常微分方程初值问题数值解基础理论

1.1　常微分方程简介

凡含有参数，未知函数和未知函数导数 (或微分) 的方程，就称为微分方程。微分方程是研究函数变化规律的主要工具，应用十分广泛。当未知函数是一元函数时，该微分方程即称作常微分方程。一般的 n 阶常微分方程具有如下形式：

$$F(x, y, y', \cdots, y^{(n)}) = 0 \tag{1}$$

遗憾的是，只有少数简单的微分方程可以求得解析解。在多数情况下，微分方程难以求出其解析解，故而寻求解析解的近似数值解在研究和应用中就十分重要。

1.2　常微分方程初值问题与数值解的涵义

我们称形如下式问题为常微分方程的初值问题：

$$(E): \begin{cases} \frac{dy}{dx} = f(x, y) \\ y(x_0) = y_0 \end{cases} \tag{2}$$

对于初值问题，我们并不求出解析解 $y = y(x)$，而是在一系列离散点 $x_0 < x_1 < \cdots < x_n < \cdots$ 上求 $y(x_i)$ 的近似解 y_i。我们称 $\{(x_i.y_i)\}$ 为常微分方程初值问题 (E) 的数值解。通常我们固定步长为 h，即令 $x_n = x_0 + nh$。

1.3 常微分方程组的标准形式及其他情况的转化

常微分方程组的标准形式与化简

一个 n 阶的标准常微分方程组具有如下形式：

$$\begin{cases} \frac{dy_1}{dx} = f_1(x, y_1, y_2, \cdots, y_n) \\ \frac{dy_2}{dx} = f_2(x, y_1, y_2, \cdots, y_n) \\ \cdots \\ \frac{dy_n}{dx} = f_n(x, y_1, y_2, \cdots, y_n) \end{cases} \tag{3}$$

其中 f_i 是变元 $(x, y_1, y_2, \cdots, y_n)$ 在某个区域 D 内的连续函数。
接下来我们通过向量记号，将上述方程组进行化简，令 $y = (y_1, y_2, \cdots, y_n)$：

$$\begin{cases} f_i(x, \mathbf{y}) = f_i(x, y_1, y_2, \cdots, y_n), \quad (i = 1, 2, \cdots, n) \\ \mathbf{f}(x, y) = (f_1(x, \mathbf{y}), f_2(x, \mathbf{y}), \cdots, f_n(x, \mathbf{y})) \end{cases} \tag{4}$$

则上述微分方程组可简记为：

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x, \mathbf{y}) \tag{5}$$

在这里需要指出的是，在常微分方程的理论与实际求解过程中，因变量是函数还是向量函数在大多数情况下是没有区别的。

一般高阶常微分方程（组）与标准形式的转化

对 n 阶常微分方程：$y^{(n)} = F(x, y, y', \cdots, y^{(n-1)})$，我们令：$y_i = y^{(i-1)} (i = 1, 2, \cdots, n)$，即可化为标准的常微分方程组：

$$\begin{cases} \frac{dy_1}{dx} = y_2 \\ \frac{dy_2}{dx} = y_3 \\ \cdots \\ \frac{dy_n}{dx} = F(x, y_1, y_2, \cdots, y_n) \end{cases} \tag{6}$$

对于高阶常微分方程组，亦可通过这种方式转化为标准的常微分方程组。

第 2 节 求解常微分方程初值问题数值方法

2.1 常微分方程初值问题数值解的涵义

对于初值问题 (E)，我们并不求出解析解 $y = y(x)$，而是在一系列离散点 $x_0 < x_1 < \cdots < x_n < \cdots$ 上求 $y(x_i)$ 的近似解 y_i。我们称 $\{(x_i, y_i)\}$ 为常微分方程初值问题 (E) 的数值解。通常我们固定步长为 h，即令 $x_n = x_0 + nh$。下面我们给出初值问题 (E) 的一些常见数值解法。

2.2 常用数值解法

Euler（欧拉）法

Euler 法的思想是利用差商近似替代微商，即如下式近似：

$$\frac{y(x_{n+1})-y(x_n)}{x_{n+1}-x_n} \approx \frac{dy}{dx}|_{x=\xi} = f(\xi, y(\xi)), \quad \xi \in [x_n, x_{n+1}] \tag{7}$$

我们取不同的 ξ 值，就可以导出不同的 Euler 法。

向前（显式）Euler 法:

取 $\xi = x_n$(左端点)，得：

$$y(x_{n+1}) \approx y(x_n) + (x_{n+1}-x_n) f(x_n, y(x_n)) \tag{8}$$

从而我们有向前 Euler 公式：

$$y_{n+1} = y_n + hf(x_n, y_n) \tag{9}$$

可以证明，显 Euler 法是一阶精确的。

向后（隐式）Euler 法：

取 $\xi = x_n$(左端点)，得：

$$y(x_{n+1}) \approx y(x_n) + (x_{n+1}-x_n) f(x_{n+1}, y(x_{n+1})) \tag{10}$$

从而我们有向后 Euler 公式：

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \tag{11}$$

为求得 y_{n+1},，需解上述非线性方程，通常我们使用下述迭代解法：

$$\begin{cases} y_{n+1}^{(0)} = y_n + hf(x_n, y_n) \\ y_{n+1}^{(k+1)} = y_n + hf(x_n, y_{n+1}^{(k)}) \\ y_{n+1} = \lim\limits_{k\to\infty} y_{n+1}^{(k+1)} \end{cases} \tag{12}$$

同样的，我们可以证明，隐 Euler 法也是一阶精确的。

Runge-Kutta(龙格-库塔) 方法

从 Euler 法的格式中，我们可以得到一些启发。如果我们考虑在区间 $[x_n, x_{n+1}]$ 取多个点，并将各点处的微商加权平均，就有可能会构造更高精度的算法。于是 Runge-Kutta 方法诞生了。我们常用的有 2 阶 Runge-Kutta 方法和 4 阶 Runge-Kutta 方法。其数值格式如下：

2 阶 Runge-Kutta 方法：

$$\begin{cases} y_{n+1} = y_n + \frac{h}{2}(K_1 + K_2) \\ K_1 = f(x_n, y_n) \\ K_2 = f(x_n + h, y_n + hK_1) \end{cases} \tag{13}$$

2 阶 R-K 方法具有 2 阶精度。

4 阶 Runge-Kutta 方法：

$$\begin{cases} y_{n+1} = y_n + \frac{h}{6}(K_1 + 2K_2 + 2K_3 + K_4) \\ K_1 = f(x_n, y_n) \\ K_2 = f(x_n + \frac{h}{2}, y_n + h\frac{K_1}{2}) \\ K_3 = f(x_n + \frac{h}{2}, y_n + h\frac{K_2}{2}) \\ K_2 = f(x_n + h, y_n + hK_3) \end{cases} \tag{14}$$

4 阶 R-K 方法具有 2 阶精度。

第 3 节 利用 MATLAB 求解常微分方程初值问题

3.1 利用 MATLAB 求解常微分方程初值问题的一个求解实例

首先我们通过一个实例，来了解 MATLAB 求解常微分方程的一般过程。
这是空气动力学中的 Lorenz 模型（其中 β, ρ, σ 为参数）

$$\begin{cases} \dot{x}(t) = -\beta x(t) + y(t)z(t) \\ \dot{y}(t) = -\rho y(t) + \rho z(t) \\ \dot{z}(t) = -xy(t) + \sigma y(t) - z(t) \end{cases} \tag{15}$$

我们记 $r = (x, y, z)^T$ 并选取两组参数 $(\beta, \rho, \sigma) = (2, 5, 20)$ 和 $(\beta, \rho, \sigma) = (5, 10, 100)$，编写程序如下：

```
function r_dot = fun_Lorenz(t,r,beta,rho,sigma)
r_dot = [-beta * r(1) + r(2)*r(3);
-rho * r(2) + rho * r(3);
-r(1)*r(2) + sigma * r(2) - r(3)];
end
```

```
t_final = 50;
r0 = [0,1e-5,0];
beta = 2;   rho = 5;    sigma = 20;

options = odeset('RelTol',1e-7);
[t,r] = ode45(@fun_Lorenz,[0,t_final],r0,options,beta,rho,sigma);
```

运行上述程序，并使用 plot 分别绘制方程的解（如下图 1所示）：
我们也可以利用 comet3 绘制方程在相空间中的运动轨迹（如下图 2所示）：

```
figure()
comet3(r(:,1),r(:,2),r(:,3));
title('beta=2,rho=5,sigma=20');
hold on;
grid on;
```

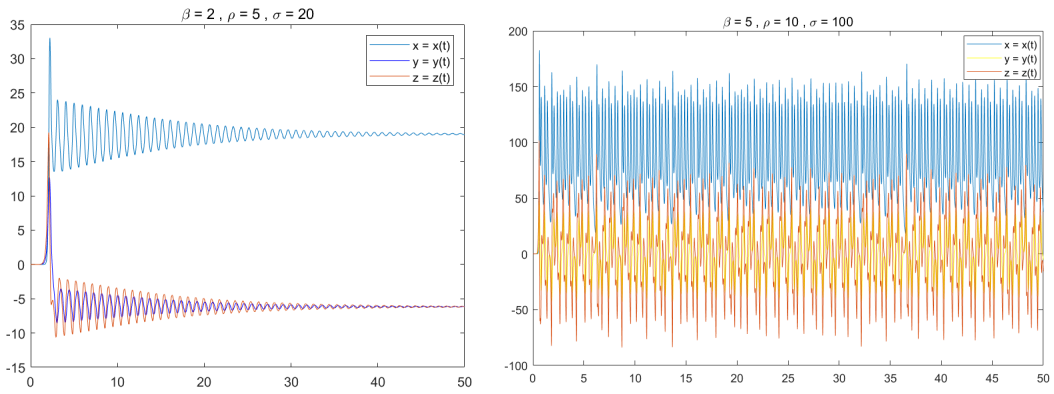


图 1: 不同参数下的解示意图

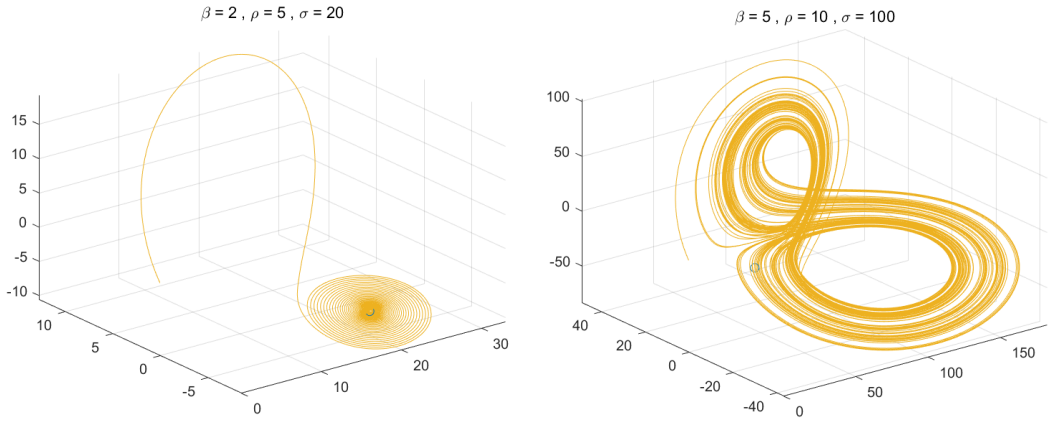


图 2: 不同参数下的解示意图

接下来，本文将逐一拆解使用 MATLAB 求解常微分方程初值问题的各个环节。

3.2 MATLAB 求解常微分方程初值问题的基本语法格式

MATLAB 求解常微分方程初值问题有三种基本的语法格式，具体如下：

格式 1：直接求解

```
[x,y] = ode45(@fun_name,[x_0,x_f],y0);
```

格式 2：带有控制参数

```
[x,y] = ode45(@fun_name,[x_0,x_f],y0,options);
```

格式 3：带有附加参数

```
[x,y] = ode45(@fun_name,[x_0,x_f],y0,options,p1,p2,⋯);
```

其中：fun_name 为微分方程（组）的描述函数，一般通过匿名函数或新建函数文件定义（具体请看本节的第三小节）；$[x_0, x_f]$ 为求解区间；y_0 即为该初值问题在初始时刻 x_0 处给定的值；options 为一结构体变量，用来控制算法的各项性质（具体请看本节的第五小节）；$p_1, p_2, \cdots$ 为附加参数。

3.3 微分方程（组）的描述函数 fun_name 格式介绍

格式 1：匿名函数

```
fun_name = @(x,y)(x*y);
```

格式 2：新建函数文件

```
function dy = fun_name(x,y,flag,p1,p2,⋯)
```

其中：x 是自变量（有时用 t 表示，亦称为时间变量）；y 是状态向量；dy 是状态向量的导数；$flag$ 用于指定初值，控制求解过程（部分版本不可缺省，即哪怕初值不用指定，也必须有此变量占位）；$p_1, p_2, \cdots$ 为主函数中设置的附加参数。

3.4 其他 MATLAB 求解常微分方程边值问题的函数简介

除了最常用的 ode45 函数，MATLAB 针对问题的不同特性以及求解的预设精度，还提供了一系列函数。具体的函数名及其性质，如下表 1所示。

表 1: 常见微分方程求解函数及说明

函数名	使用问题类型	求解精度	备注
ode45	非刚性	中等	首选的求解器
ode23	非刚性	低	用于求解具有粗略误差容限的问题或中等刚性问题
ode113	非刚性	低到高	用于求解具有严格误差容限的问题或计算密集型问题
ode15s	刚性	低到中	用于求解刚性问题。当 ode45 求解缓慢时，可尝试使用
ode23s	刚性	低	用于求解质量矩阵恒定的具有粗略误差容限的问题
ode23t	中等刚性	低	用于求解没有数值阻尼的中等刚性问题
ode23tb	刚性	低	用于求解具有粗略误差容限的问题或中等刚性问题

3.5 控制选项 options 简介

表 3: 属性类别与属性名称

属性类别	属性名称
Error control	RelTol,AbsTol,NormControl
Solver output	OutputFcn,OutputSel
Jacobian matrix	Jacobian,JPattern
Step-size	InitialStep,MaxStep
Mass matrix and DAEs	Mass,MStateDepedence
Event location	Event
sode15s-specific	MaxOrder

最后介绍参数设置的方法：

直接设置法：

格式：options = odeset(' 属性名', 值); 例如：

```
options = odeset('RelTol',1e-7);
```

面向对象的参数设置方法：

格式：options = odeset;options. 属性名 = 值; 例如：

```
options = odeset;
options.RelTol = 1e-7;
```

第 4 节 小结

1. 本文主要针对常微分方程初值问题，介绍了一些常见的的数值解法以及如何利用 MATLAB 软件实现对常微分方程初值问题的求解。在实际的应用过程中，如何选择恰当的函数、设定合适的参数，还需要大家在实践中不断摸索。

2. 此外，MATLAB 还可以使用 bvp4c 函数求解常微分方程的边值问题，有兴趣的同学可以自行查阅资料学习。

珞珈数学 Mathematics of Luojia March 2018

积分变换与偏微分方程理论

珞珈数学研习会　基础数学专栏

第一作者：2018 级　数学基地班　杨宇鹏

偏微分方程的研究是数学研究的一个重要方向，本文针对 (半) 无界区域上的偏微分方程理论及其研究成果进行介绍。在引入相关恒等式与 Fourier 变换的知识后，针对三类主要方程的 Cauchy 问题进行求解，最后对解的验证过程给出推导证明。

关键词：偏微分方程理论、三类主要的偏微分方程、Fourier 变换、解的验证

第 1 节　前言

在前文介绍的常微分方程基础上，偏微分方程的研究同样也是数学研究的一个重要方向，其中对于有界区域上的方程定解问题可以使用变量分离方法进行求解。而若定解问题推广到 (半) 无界区域上，可以利用 Fourier 变换将偏微分方程化为常微分方程求解，之后使用逆变换即可还原微分方程的解。针对三类主要方程的 Cauchy 问题，由于 Fourier 变换对函数光滑性提出了假设，在求解完毕后需要对解进行验证。本文将主要根据如上内容，对偏微分方程的数学理论进行简要介绍。

第 2 节　Fourier 变换相关知识回顾

由于 Fourier 变换为无界区域上的偏微分方程化简的主要工具，因此首先回顾 Fourier 变换与逆变换的定义。为方便叙述，本小节中仅给出常用性质，如读者有兴趣进一步深入了解其他性质或证明过程，可以阅读文献 [1]：

定义 1

给定函数 $f \in L^1(\mathbb{R}^n)$，记 n 维向量 x,ξ 的欧氏空间内积运算为 $<x,\xi>$ 定义其 Fourier 变换 $\hat{f}$ 如下：

$$\begin{cases} \hat{f}(\xi) = \int_{\mathbb{R}^n} f(x) e^{-2\pi i <x,\xi>} dx \\ <x,\xi> = \sum_{k=1}^{n} x_k \xi_k \end{cases} \tag{1}$$

在特定的条件下，可以定义 Fourier 变换的逆变换：

$$f(x) = \int_{\mathbb{R}^n} \hat{f}(\xi) e^{2\pi i <x,\xi>} d\xi \tag{2}$$

接下来回顾 Fourier 变换的三条基本性质：

- **Prop1**：$(\alpha f \hat{+} \beta g) = \alpha \hat{f} + \beta \hat{g}$
- **Prop2**：$\|\hat{f}\|_{\inf} \leq \|f\|_1$　$\hat{f}$ **连续**
- **Prop3**：$\lim\limits_{|\xi| \to \inf} \hat{f}(\xi) = 0$　**(Riemann-Lebesgue)**

其次，回顾函数间卷积的定义如下：

定义 2

对任意给定 $f_1(x), f_2(x)$ 定义两函数的卷积如下：

$$f(x) = f_1 * f_2 = \int_{\mathbb{R}^n} f_1(y) f_2(x-y) dy \tag{3}$$

在引入卷积后，可以推导出 Fourier 变换的如下三条性质：

- **Prop4**：$(f_1 \hat{*} f_2) = \hat{f}_1 \hat{f}_2$
- **Prop5**：$(\widehat{\frac{\partial f}{\partial x_j}})(\xi) = 2\pi i \xi_j \hat{f}(\xi)$
- **Prop6**：$(\widehat{-2\pi i x_j f})(\xi) = \frac{\partial \hat{f}}{\partial \xi_j}(\xi)$

第 3 节　相关恒等式回顾

在上述 Fourier 变换基本性质的证明与下文主要偏微分方程的求解计算过程中，很多积分恒等式发挥着化简、推导的重要作用。因此在本部分，展开对常用恒等式的探讨与证明：

性质 1

若 $f(x) = e^{-\pi|x|^2}$，则 $\hat{f}(\xi) = e^{-\pi|\xi|^2}$

上述恒等式给出了一种函数形式，在使用 Fourier 变换并转换变量的前后，能保持相同的函数形式，本性质的证明可以如下给出：

证明:

维数 n=1 时，函数 $f(x) = e^{-\pi x^2}$ 满足如下的微分方程：

$$\begin{cases} u' + 2\pi x u = 0 \\ u(0) = 1 \end{cases}$$

由上述性质 **Prop5、Prop6** 可知 $\hat{f}$ 具有相同的微分方程且有常数初值如下计算：

$$\hat{u}(0) = \int_{\mathbb{R}^n} u(x) dx = \int_{\mathbb{R}^n} e^{-\pi x^2} dx = 1$$

因此由常微分方程解的唯一性可以得到 $\hat{f} = f$。

在 $n > 1$ 时，$\hat{f}$ 计算过程中由 n 个相同积分得到结果，因此使用 $n = 1$ 时的结论可以得到 $\hat{f} = f$，证明完毕。　□

性质 2

$\forall \alpha > 0$，有如下积分等式成立：

$$\begin{cases} \frac{2}{\pi}\int_0^{+\infty}\frac{\cos\beta x}{a+x^2}dx = e^{-\beta} \\ \frac{1}{\sqrt{\pi}}\int_0^{+\infty}\frac{e^{-u}}{\sqrt{u}}e^{-\frac{\beta^2}{4u}}dx \end{cases} \tag{4}$$

证明：

取 $f(z)=\frac{e^{i\beta z}}{1+z^2}$ 可以利用复变函数留数定理得到如下式：

$$\int_{-R}^{R}\frac{e^{i\beta x}}{1+x^2}dx+\int_{\gamma_R}\frac{e^{i\beta z}}{1+z^2}dz = 2\pi i Res(f,i)$$

当 $R\to+\infty$ 时有：

$$\begin{aligned} |\int_{\gamma_R}\frac{e^{i\beta z}}{1+z^2}dz| &\le \frac{1}{R^2-1}\int_0^{\pi}e^{-\beta R\sin\theta}Rd\theta \\ &\le \frac{2}{R^2-1}\int_0^{\frac{\pi}{2}}e^{-\frac{2\pi R\theta}{\pi}}Rd\theta \\ &= \frac{2\pi}{\beta(R^2-1)}(1-e^{-\beta R})\to 0 \end{aligned}$$

结合留数公式 $Res(f,i)=\frac{e^{-\beta}}{2i}$，第一个恒等式成立。接下来直接使用该恒等式得到如下计算过程：

$$\begin{aligned} e^{-\beta} =& \frac{2}{\pi}\int_0^{+\infty}\frac{\cos\beta x}{a+x^2}dx \\ & \frac{2}{\pi}\int_0^{+\infty}\cos\beta x(\int_0^{+\infty}e^{-u}e^{-ux^2}du)dx \\ & \frac{2}{\pi}\int_0^{+\infty}e^{-u}(\frac{1}{2}\int_{-\infty}^{+\infty}e^{-ux^2}\cos\beta x dx)du \\ & \frac{2}{\pi}\int_0^{+\infty}e^{-u}(\frac{1}{2}\sqrt{\frac{\pi}{u}}e^{-\frac{\beta^2}{4u}})du \end{aligned}$$

上述最后一步由复变函数积分知识保证，由上述推导可知本性质成立。 □

性质 3

$\forall \beta > 0$，有如下积分等式成立：

$$\int_{\mathbb{R}^n}e^{-2\pi|y|\alpha}e^{-2\pi i<t,y>}dy = \frac{\Gamma(\frac{n+1}{2})}{\pi^{\frac{n+1}{2}}}\frac{\alpha}{(\alpha^2+|t|^2)^{\frac{n+1}{2}}} \tag{5}$$

证明：

首先考虑 $\alpha=1$ 的情形，由上述性质，可以进行如下推导：

$$\begin{aligned}\int_{\mathbb{R}^n} e^{-2\pi|y|}e^{-2\pi i<t,y>}dy &= \int_{\mathbb{R}^n}(\frac{1}{\sqrt{\pi}}\int_0^{+\infty}\frac{e^{-u}}{\sqrt{u}}e^{-\frac{4\pi^2|y|^2}{4u}}du)e^{-2\pi i<t,y>}dy\\ &= \frac{1}{\sqrt{\pi}}\int_0^{+\infty}\frac{e^{-u}}{\sqrt{u}}((\sqrt{\frac{u}{\pi}})^n e^{-u|t|^2})du\\ &= \frac{1}{\pi^{\frac{n+1}{2}}}\frac{1}{(1+|t|^2)^{\frac{n+1}{2}}}\int_0^{+\infty}e^{-s}s^{\frac{n-1}{2}}ds\\ &= \frac{\Gamma(\frac{n+1}{2})}{\pi^{\frac{n+1}{2}}}\frac{1}{(1+|t|^2)^{\frac{n+1}{2}}}\end{aligned}$$

对 $\alpha>0$，可以利用上述情形的结论，给出如下推导过程，令 y 变换为 $\frac{1}{\alpha}y$，直接推出：

$$\begin{aligned}\int_{\mathbb{R}^n} e^{-2\pi|y|\alpha}e^{-2\pi i<t,y>}dy &= \frac{1}{\alpha^n}\int_{\mathbb{R}^n}e^{-2\pi|y|}e^{-\frac{2}{\alpha}\pi i<t,y>}dy\\ &= \frac{1}{\alpha^n}\frac{\Gamma(\frac{n+1}{2})}{\pi^{\frac{n+1}{2}}}\frac{1}{(1+|\frac{t}{\alpha}|^2)^{\frac{n+1}{2}}}\\ &= \frac{\Gamma(\frac{n+1}{2})}{\pi^{\frac{n+1}{2}}}\frac{\alpha}{(\alpha^2+|t|^2)^{\frac{n+1}{2}}}\end{aligned}$$

至此本性质证完。 □

性质 4

$\forall\beta>0$，给定 $f(x)=e^{-\pi\alpha|x|^2}$，有 $\hat{f}(\xi)=\frac{1}{\alpha^{\frac{n}{2}}}e^{-\frac{\pi|\xi|^2}{\alpha}}$

证明：

$$\begin{aligned}\hat{f}(\xi) &= \int_{\mathbb{R}^n}e^{-\pi\alpha|x|^2}e^{-2\pi i<\xi,x>}dx\\ &= \prod_{i=1}^n\int_{\mathbb{R}}e^{-(\pi\alpha x_i^2+2\pi i\xi_i x_i)}dx\\ &= \prod_{i=1}^n\int_{\mathbb{R}}e^{-\pi\alpha(x_i+\frac{i\xi_i}{\alpha})^2-\frac{\pi\xi_i^2}{\alpha}}dx_i\\ &= \prod_{i=1}^n\frac{\sqrt{\pi}}{\sqrt{\pi\alpha}}e^{-\frac{\pi\xi_i^2}{\alpha}} = \frac{1}{\alpha^{\frac{n}{2}}}e^{-\frac{\pi|\xi|^2}{\alpha}}\end{aligned}$$

□

第 4 节 三类方程的 Cauchy 问题

在回顾上述知识后，我们将目光聚集于偏微分方程理论中的一个重要部分，即三类常见的微分方程的初值问题。

4.1 椭圆型方程

各种物理性质的定常（即不随时间变化）过程，都可用椭圆型方程来描述。其典型、简单的形式是泊松 (Poisson) 方程，如下：

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x,y) \tag{6}$$

特殊地，在 $f(x,y)$ 恒为 0 时，原方程即为拉普拉斯 (Laplace) 方程，又称为调和方程，带有稳定热源或内部无热源的稳定温度场的温度分布，不可压缩流体的稳定无旋流动及静电场的电势等均满足这类方程。拉普拉斯方程的 Cauchy 问题可以如下描述：

$$\begin{cases} \Delta w = 0 \quad on \quad \mathbb{R}_+^{n+1} \\ w(x,0) = f(x) \\ \lim\limits_{|x|^2+t^2 \to +\infty} w(x,t) = 0 \end{cases} \tag{7}$$

接下来对上式进行简单的求解，首先对前两个等式进行 Fourier 变换可以得到如下式：

$$\begin{cases} \frac{d^2\hat{w}}{dt^2} - (2\pi|\xi|)^2 \hat{w} = 0 \\ \hat{w}(x,0) = \hat{f}(x) \end{cases}$$

上述式将原方程化简为了关于 t 的常微分方程，求解上述常微分方程得到：

$$\begin{cases} \hat{w} = C_1 e^{2\pi|\xi|t} + C_2 e^{-2\pi|\xi|t} \\ C_1 + C_2 = \hat{f}(\xi) \end{cases}$$

又由上文 Fourier 变换部分的 **Prop3** 可知：

$$\lim_{|\xi| \to +\infty} \hat{f}(\xi) = \lim_{|\xi| \to +\infty} \hat{w}(\xi) = \lim_{t \to +\infty} \hat{w}(x,t) = 0$$

结合上述三个等式可以得到方程解为：

$$\hat{w} = \hat{f} e^{-2\pi|\xi|t}$$

由 Fouirer 变换 **Prop4** 与逆变换可以得到如下结果即为本方程定解问题最终的解：

$$w = f(x) * p_t(x)$$

其中 $p_t(x) = \frac{\Gamma(\frac{n+1}{2})}{\pi^{\frac{n+1}{2}}} \frac{t}{(t^2+|x|^2)^{\frac{n+1}{2}}}$，称为 Poisson 核。

4.2 抛物型方程

在研究热传导过程，气体扩散现象及电磁场的传播等随时间变化的非定常物理问题时，常常会遇到抛物型方程。如下给出高维热传导方程的基本形式与解的求法：

$$\begin{cases} \frac{\partial w}{\partial t} = \Delta w \quad on \quad \mathbb{R}_+^{n+1} \\ w(x,0) = f(x) \quad x \in R^n \end{cases} \tag{8}$$

对上述等式两边同时作 Fourier 变换可以得到如下结果：

$$\begin{cases} \frac{d\hat{w}}{dt} = |2\pi i\xi|^2\hat{w} \\ \hat{w}(x,0) = \hat{f} \end{cases}$$

同样转换为常微分方程的求解问题，类似 Poisson 方程的求解方法对上述常微分方程进行求解可以给出如下求解结果：

$$\hat{w}(\xi,t) = \hat{f}(\xi)e^{-4\pi^2|\xi|^2t}$$

进一步，对上述表达式进行 Fourier 逆变换，结合 **Prop4** 可以得到如下结果，即为光滑有界条件下抛物线型方程的解：

$$w(x,t) = (\frac{1}{2\sqrt{\pi t}})^n e^{-\frac{|x|^2}{4t}} * f(x)$$

4.3 双曲型方程

最后要介绍的一类方程是双曲型方程，在物理中的应用常为波动方程，如下给出高维双曲型方程的理论：

$$\begin{cases} \frac{\partial^2 w}{\partial t^2} = \Delta w \quad on \quad \mathbb{R}_+^{n+1} \\ w(x,0) = f(x) \quad x \in \mathbb{R}^n \\ \frac{\partial w}{\partial t}(x,0) = 0 \quad x \in \mathbb{R}^n \end{cases} \tag{9}$$

类似其他方程，首先对波动方程的两边同时作 Fourier 变换，可以得到如下结果，并将初值条件化简为如下形式：

$$\begin{cases} \frac{d^2\hat{w}}{dt^2} = |2\pi i\xi|^2\hat{w} \\ \frac{d\hat{w}}{dt} = 0 \\ \hat{w}(x,0) = \hat{f} \end{cases}$$

进一步使用常微分方程理论进行上述表达式的求解可以得到如下结果：

$$\hat{w} = \hat{f}\cos|2\pi\xi|t$$

进行 Fourier 逆变换即可还原并解出 w，其中特殊地在一维下称为 D'Alembert 公式，可以通过如下推导得到：

$$\begin{aligned} w(x,t) &= \int_{\mathbb{R}} \hat{f}\cos(2\pi\xi t)e^{2\pi ix\xi}d\xi \\ &= \int_{\mathbb{R}} \hat{f}\frac{e^{2\pi it\xi} + e^{-2\pi it\xi}}{2}e^{2\pi ix\xi}d\xi \\ &= \int_{\mathbb{R}} \hat{f}e^{2\pi i\xi(x+t)} + e^{-2\pi i\xi(x-t)}d\xi \\ &= \frac{1}{2}[f(x+t) + f(x-t)] \end{aligned}$$

第 5 节 解的验证

由于利用 Fourier 变换假设 f 具有良好的光滑性，因此上文给出的解是形式解，需要对所得到的解进行验证，在此以热传导方程为例，考虑维数为一的情形 (高维度情形可以类似证明)。如下为一维热传

导方程 (f 光滑有界) 及其形式解：

$$\begin{cases} \frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial x^2} \\ w|_{t=0} = f(x) \end{cases} \tag{10}$$

$$\tilde{w}(x,t) = \frac{1}{2\sqrt{\pi t}} \int_{-\infty}^{+\infty} f(y) e^{-\frac{(x-y)^2}{4t}} dy \tag{11}$$

证明：

不妨设 $|f(x,y)| \leq M$，可以立即给出如下式证实 $\tilde{w}$ 有界：

$$|\tilde{w}(x,t)| \leq M \int_{-\infty}^{+\infty} \frac{1}{2\sqrt{\pi t}} f(y) e^{-\frac{(x-y)^2}{4t}} dy = M$$

由于积分因子 $e^{-\frac{(x-y)^2}{4t}}$ 保证了一致收敛性能，因此可以对形式解求偏导数，可以立即验证出原方程成立。因此只需要验证初值条件的正确性：

$\forall x_0 \in (-\infty, +\infty), \epsilon > 0, \exists \delta > 0$，使得 $|x - x_0| \leq \delta$ 时有如下式成立：

$$|f(x) - f(x_0)| < \frac{\epsilon}{3}$$

在形式解中，设 $\lambda = \frac{y-x}{2\sqrt{t}}$，则可以改写形式解为如下形式：

$$\tilde{w}(x,t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} f(x + 2\sqrt{t}\lambda) e^{-\lambda^2} d\lambda$$

则可以推出如下不等式关系：

$$\begin{aligned} |\tilde{w}(x,t) - f(x_0)| &= |\frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} f(x + 2\sqrt{t}\lambda) e^{-\lambda^2} d\lambda - \frac{f(x_0)}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-\lambda^2} d\lambda| \\ &\leq \frac{1}{\sqrt{\pi}} (\int_{N}^{+\infty} + \int_{-N}^{N} + \int_{-\infty}^{-N}) |f(x + 2\sqrt{t}\lambda) - f(x_0)| e^{-\lambda^2} d\lambda \end{aligned}$$

其中 N 充分大且满足 $\int_{-\infty}^{-N} e^{-\lambda^2} d\lambda \leq \frac{\epsilon}{6M}, \int_{N}^{+\infty} e^{-\lambda^2} d\lambda \leq \frac{\epsilon}{6M}$，而在 $-N \leq \lambda \leq N$ 时，取 $t \to 0$ 且 $|x + 2\sqrt{t}\lambda - x_0| < \frac{\delta}{2}$，有如下不等式成立：

$$\frac{1}{\sqrt{\pi}} \int_{-N}^{+N} |f(x + 2\sqrt{t}\lambda) - f(x_0)| e^{-\lambda^2} d\lambda < \frac{\epsilon}{2\sqrt{\pi}} \int_{-N}^{+N} e^{-\lambda^2} d\lambda < \frac{\epsilon}{3}$$

结合以上式可以直接得到 $|\tilde{w}(x,t) - f(x_0)| < \epsilon$，因此命题证明成立。 □

第 6 节 参考文献

[1] Stein E M, Weiss Guido. Introduction to fourier analysis on Euclidean spaces[J]. 1971.

珞珈数学 Mathematics of Luojia March 2018

偏微分方程的数值解法

珞珈数学研习会　基础数学专栏

第一作者：2018 级　信息与计算科学　周睿涵
第二作者：2019 级　信息与计算科学　时宇辰
第三作者：2019 级　数学基地班　刘宇佳

随着社会进步和科技发展，越来越多复杂的计算问题有待人类解决，因此，数值方法便以解决复杂问题为目的应运而生，而其中至关重要的一部分便是偏微分方程的数值解。而偏微分方程数值解中最重要的方法便是以下 3 种：有限差分法、有限元方法、有限体积法。本文将着重介绍有限差分法，并对这种方法的应用范围、基本思路以及解题步骤做简要的介绍。同时简单介绍其他方法的主要思路，并将其横向进行对比，以便更好地让读者理解这 3 种数值解法。

关键词：偏微分方程的数值解法、有限差分法

第 1 节　三类偏微分方程初边值问题回顾

在介绍偏微分方程的数值解法之前，首先对前篇提出的三类偏微分方程主要初边值问题进行回顾，为后文数值解给出理论基础。

1.1　椭圆型方程

低维形式椭圆型方程如下：

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=f(x,y)\qquad (x,y)\in\Omega \tag{1}$$

其中第一边值问题的边界条件可以如下描述：

$$u(x,y)|_{(x,y)\in\Gamma}=\varphi(x,y)\qquad \Gamma\in\partial\Omega \tag{2}$$

第二、三类边界条件可以如下统一描述：

$$\left(\frac{\partial u}{\partial n}+au\right)|_{(x,y)\in\Gamma}=\varphi(x,y) \tag{3}$$

1.2 抛物型方程

简单的抛物型方程形式如下：

$$\frac{\partial u}{\partial t}-a\frac{\partial^2 u}{\partial x^2}=0 \qquad (a>0) \tag{4}$$

首先初值条件是确定的如下式：

$$u(x,0)=\varphi(x) \qquad -\infty<x<+\infty \tag{5}$$

第一类边界条件如下式给出：

$$u(0,t)=g_1(t) \qquad u(l.t)=g_2(t) \qquad 0\le t\le T \tag{6}$$

第二类边界条件和第三类边界条件如下式：

$$[\frac{\partial u}{\partial x}-\lambda_1(t)u]|_{x=0}=g_1(t) \qquad 0\le t\le T,\lambda_1(t)\ge 0 \tag{7}$$

$$[\frac{\partial u}{\partial x}-\lambda_2(t)u]|_{x=l}=g_2(t) \qquad 0\le t\le T,\lambda_2(t)\ge 0 \tag{8}$$

其中 $\lambda_1(t)=\lambda_2(t)=0$ 时称为第二类边界条件，否则为第三类边界条件。

1.3 双曲型方程

双曲型方程的初值问题如下：

$$\begin{cases}\frac{\partial^2 u}{\partial t^2}=a^2\frac{\partial^2 u}{\partial x^2} & t>0,-\infty<x<+\infty\\ u(x,0)=\varphi(x) & -\infty<x<+\infty\\ \frac{\partial u}{\partial t}|_{t=0}=\phi(x) & -\infty<x<+\infty\end{cases} \tag{9}$$

双曲型方程的边界条件也一般有三类，其中最简单的一类边界条件如下：

$$u(0,t)=g_1(t) \quad u(l,t)=g_2(t) \quad 0\le t\le T \tag{10}$$

第 2 节 数值解法应用范围及基本思路

一个问题多种解决方法的本质区别在于求解思想的差别，由于每种解决方法的求解思想不同，一些方法的基本思路由于更利于大众接受而被广泛利用，当然不能排除一些因素，例如解决方法的适用范围。另外，解题人的偏爱和解题方法操作的难易程度也会对偏微分方程数值解法的选择产生影响。因此，对于任何数值解问题，我们都应该仔细思考它属于那种解题方法的应用范围，并掌握它的各种解决思路，以便对症下药，为其寻找最佳的解决方法。

第 3 节 有限差分法基本思路及应用范围

有限差分方法 (FDM) 是计算机数值模拟最早采用的方法，至今仍被广泛运用。该方法将求解域划分为差分网格，用有限个网格节点代替连续的求解域。有限差分法以 Taylor 级数展开等方法，把控制方程中的导数用网格节点上的函数值的差商代替进行离散，从而建立以网格节点上的值为未知数的代数方程组。该方法是一种直接将微分问题变为代数问题的近似数值解法，数学概念直观，表达简单，是发展较早且比较成熟的数值方法。

这种方法的主要步骤如下：

- **Step1: 首先将需要求解的领域进行分割，划分为不同的网格利用有限的网格节点来代替需要持续计算求解的领域。**
- **Step2: 通过开展不同的方法，将网格节点上的不同数值间的差商来替代方程中的数值，进行缩小，达到需求数值组建代数方程组的目的，并运用包含可以计数的差分方程中的未知量，逐步接近并且渐渐产生可以代替的数值的微分方程和定解条件。**
- **Step3: 我们在差分方程求得的结果，可以作为所需求的近似解，进一步把以前方程中出现的微分和在边界条件中出现的微分，使用差分来寻求近似。**

同时，近似值也可以运到机械求积公式中，进而将其逐步运用不同的条件转化成为差分方程组。在有限差分法中，最简便的方法就是把微分问题变成代数问题，进一步求得近似值。可以说，有限差分法是一种发展较早同时比较成熟的数值方法。

3.1 椭圆型方程第一边值问题的差分解法

Step1: 求解区域的划分

对于椭圆型方程的第一初值问题，首先进行二维平面的划分，使用两组平行线 $x = x_k = kh, y = y_j = j\tau (k, j = 0, 1, -1, 2, -2, \cdots)$ 将定界区域分成矩形网络，节点记为如下：

$$R = \{(x_k, y_j) | x_k = kh, y_j = j\tau, \quad i, j \in N^+\} \tag{11}$$

定解区域内部的节点称为内点，记内点集 $R \cap \Omega$ 为 $\Omega_{h\tau}$。边界 Γ 与网格线的交点称为边界点，边界点全体记为 $\Gamma_{h\tau}$。与节点仅差一个步长的点称为相邻节点，若一个内点的四个相邻接点均属于 $\Omega \cup \Gamma$，称为正则内点，正则内点全体记为 $\Omega^{(1)}$，非正则内点全体记为 $\Omega^{(2)}$。

接下来步骤的关键就是求解全体内点上方程的数值解。简记 $(x_k, y_j) = (k, j), f_{k,j} = f(x_k, y_j)$。

Step2: 构造差商公式

对正则内点，考虑如下二阶中心差商公式：

$$\begin{cases} \frac{\partial^2 u}{\partial x^2}|_{(k,j)} = \frac{u(k+1,j)-2u(k,j)+u(k-1,j)}{h^2} + O(h^2) \\ \frac{\partial^2 u}{\partial y^2}|_{(k,j)} = \frac{u(k,j+1)-2u(k,j)+u(k,j-1)}{\tau^2} + O(\tau^2) \end{cases} \tag{12}$$

因此对每一个正则内点，可以直接推导出与 Poisson 方程相近似的差分方程如下：

$$\frac{u_{k+1,j} - 2u_{k,j} + u_{k-1,j}}{h^2} + \frac{u_{k,j+1} - 2u_{k,j} + u_{k,j-1}}{\tau^2} = f_{k,j} \tag{13}$$

上述过程完毕后，对边界条件使用直接转移或线性插值即可建立边界条件的差分方程，进一步进行方程求解，并在 **Step3** 进行递推求解即可得到椭圆型方程的解。

3.2 抛物型方程初边值问题的差分解法

抛物型方程的求解区域网格化可以完全类似椭圆型方程进行，因此在此不进行赘述，主要考虑方程的差分处理与初边值条件的差分处理。

微分方程的差分近似

在网格内点 (k,j) 处，对 $\frac{\partial u}{\partial t}$ 分别采用向前、向后及中心差商公式，对 $\frac{\partial^u}{\partial x^2}$ 采用二阶中心差商公式，因此可以将一维热传导方程化为如下三种主要的差分近似形式：

$$\begin{cases} \frac{u_{k,j+1}-u_{k,j}}{\tau} - a\frac{u_{k+1,j}-2u_{k,j}+u_{k-1,j}}{h^2} = 0 \\ \frac{u_{k,j}-u_{k,j-1}}{\tau} - a\frac{u_{k+1,j}-2u_{k,j}+u_{k-1,j}}{h^2} = 0 \\ \frac{u_{k,j+1}-u_{k,j-1}}{2\tau} - a\frac{u_{k+1,j}-2u_{k,j}+u_{k-1,j}}{h^2} = 0 \end{cases} \tag{14}$$

初边值条件的处理

第一类边界条件的处理：

第一类边界条件较简单，可以直接函数离散化得到如下式：

$$\begin{cases} u_{k,0} = \varphi_k \\ u_{0,j} = g_{1j} \\ u_{n,j} = g_{2j} \end{cases} \tag{15}$$

第二、三类边界条件的处理：

第二、三类边界条件处理的关键在于如何利用差商代替偏导数 $\frac{\partial u}{\partial x}$，可以分别使用右边界的向后差商或中心差商代替，两种形式如下所示：

$$\begin{cases} \frac{u_{1,j}-u_{0,j}}{h} - \lambda_{1j}u_{0,j} = g_{1j} \\ \frac{u_{n,j}-u_{n-1,j}}{h} + \lambda_{2j}u_{n,j} = g_{2j} \end{cases} \tag{16}$$

如果需要使用如下的中心差商法，需要将函数的定义域拓展到边界的相邻点，此时假定热传导方程在边界也成立即可。

$$\begin{cases} \frac{u_{1,j}-u_{-1,j}}{2h} - \lambda_{1j}u_{0,j} = g_{1j} \\ \frac{u_{n+1,j}-u_{n-1,j}}{2h} + \lambda_{2j}u_{n,j} = g_{2j} \end{cases} \tag{17}$$

3.3 双曲型方程初边值问题的差分解法

以二阶波动方程 $\frac{\partial^2 u}{\partial t^2} = a^2\frac{\partial^2 u}{\partial x^2}$ 为例，引入代换 $v_1 = \frac{\partial u}{\partial t}, v_2 = \frac{\partial u}{\partial x}$，则原方程可以化为如下一阶线性双曲型方程组：

$$\begin{cases} \frac{\partial v_1}{\partial t} = a^2\frac{\partial v_2}{\partial x} \\ \frac{\partial v_2}{\partial t} = \frac{\partial v_1}{\partial x} \end{cases} \tag{18}$$

记 $v=(v_1,v_2)^T$，则方程组可以表示为矩阵形式：

$$\frac{\partial v}{\partial t}=\begin{pmatrix}0 & a^2\\ 1 & 0\end{pmatrix}\frac{\partial v}{\partial x}=A\frac{\partial v}{\partial x} \tag{19}$$

因此做变换 $w=Pv=(w_1,w_2)^T$，并记 $Lambda=diag\{a,-a\}$，可以得到如下变形式，即由两个独立的一阶双曲型方程联立而成：

$$\frac{\partial w}{\partial t}=\Lambda\frac{\partial w}{\partial x} \tag{20}$$

因此主要考虑一阶双曲型方程的初值问题，如下：

$$\begin{cases}\frac{\partial u}{\partial t}+a\frac{\partial u}{\partial x}=0 & t>0 \quad -\infty<x<+\infty\\ u(x,0)=\varphi(x) & -\infty<x<+\infty\end{cases} \tag{21}$$

此时可以使用类似前两种方程的求解方法首先划分网格，进一步通过不同的差商代替偏导数项，进一步对所有网格点建立离散化方程，如下给出三种常用的差分近似式：

$$\begin{cases}\frac{u_{k,j+1}-u_{k,j}}{\tau}+a\frac{u_{k+1,j}-u_{k,j}}{h}=0\\ \frac{u_{k,j+1}-u_{k,j}}{\tau}-a\frac{u_{k,j}-u_{k-1,j}}{h}=0\\ \frac{u_{k,j+1}-u_{k,j}}{\tau}-a\frac{u_{k+1,j}-u_{k-1,j}}{2h}=0\end{cases} \tag{22}$$

第 4 节 其他数值解法的应用范围及基本步骤简介

4.1 有限元法基本思路及应用范围

有限元法也叫有限单元法（finite element method, FEM），是随着电子计算机的发展而迅速发展起来的一种弹性力学问题的数值求解方法。五十年代初，它首先应用于连续体力学领域—飞机结构静、动态特性分析中，用以求得结构的变形、应力、固有频率以及振型。由于这种方法的有效性，有限单元法的应用已从线性问题扩展到非线性问题，分析的对象从弹性材料扩展到塑性、粘弹性、粘塑性和复合材料，从连续体扩展到非连续体。

有限元方法的基本理论主要是变分原理和加权余量法。它主要是将所需计算的领域通过划分，变成可以计数的并不重复的单元。在不同的单元内，需求适合的节点作为插值点，最终得到一系列插值函数组成的线性表达方式。以主要理论为基础，将微分方程分散求解，在选取了不同的数值之后，会形成不同的有限元方法。通过利用得到的线性组合不断接近方程的精确值，所有计算域内的解就能够看成是由所有单元上的近似解组成的。使用有限元法解题过程中，可以把求解域人为地分成许多有限元的小的、相互接近的子域组成。接着，假设一个比较简单的近似值，针对所划分的小单元，逐步地演算出这个领域需要的条件，进而得到我们需要的答案。但是，求得的结果并不是精确值，而是近似的。总的来说，有限元法在计算精度上算是很高的，并且可以应对各种不同复杂的形状，是使用最多也是最有效的方法。

有限元法在工程中最主要的应用形式是结构的优化，如结构形状的最优化，结构强度的分析，振动的分析等等。有限元法的出现，使得传统的基于经验的结构设计趋于理性，设计出的产品越来越精细，尤为突出的一点是，产品设计过程的样机试制次数大为减少，产品的可靠性大为提高。压力容器的结构应力分析和形状优化，汽车试制过程中的碰撞模拟，发动机设计过程中的减振降噪分析，武器设计过程中爆轰过程的模拟、弹头形状的优化等等，都是目前有限元法在工程中典型的应用。

4.2 有限元法解题步骤

偏微分方程中的有限元法在求解过程中，可以比较随意的配置离散点，选取合适的数值和单元剖分密度，从而达到要求中的计算精度。具体运用步骤如下：

a. 剖分：首先把需要的区域每进行分裂，分割为可以计数的要素集合。每一个小的单元，原则上形状是可以随意的。这样可以使得计算更加简便，结果更加准确。一般情况下，二维问题通常使用的形状为三角形或者是矩形；三维空间使用的则是多面体等;

b. 单元分析：在分割的不同区域中，插入我们研究所得的数值，也就是说把任意单元中的任意点进行展开计算，从而建立一个线性的插值函数;

c. 求解近似变分方程：在可以计数的单元将连续体的数值进行相应缩小，提高计算的时空效率，同时分区域插值解决各种需要解决的问题。杆系结构的形状是一个杆件，而连续体的形状可以是三角形、四边形或者是六面体等。不同的单元中，包含着一些可以计数的简单函数。这些简单函数集合是整个连续体函数的元素合。接着，通过精确的计算，可以得到所需求的数值。

现在，有限元法已经应用于各种大型或是专用程序。随着时间的推移，有限元法也不断衍生出更多解法，以便于解决更多问题。

4.3 有限体积法基本思路及应用范围

有限体积法（FVM）又称为控制体积法。其基本思路是将所需要计算的区域分割成为一系列不重叠的可控制的体积。同时，不同网格点的四周都得到一个控制体积，接着运用一定的方法将需要解决的方程进行计算，得到一组离散方程。假如需要求出控制体积的积分，则设定假设值，并将其插入网格点间的分布剖面上。因此，可以得到有限体积法的基本方法就是子区域法。

有限体积法的基本思路易于理解，并能得出直接的物理解释。离散方程的物理意义，就是因变量在有限大小的控制体积中的守恒原理，如同微分方程表示因变量在无限小的控制体积中的守恒原理一样。有限体积法得出的离散方程，要求因变量的积分守恒对任意一组控制体积都得到满足，对整个计算区域，自然也得到满足。这是有限体积法吸引人的优点。有一些离散方法，例如有限差分法，仅当网格极其细密时，离散方程才满足积分守恒；而有限体积法即使在粗网格情况下，也显示出准确的积分守恒。

就离散方法而言，有限体积法可视作有限单元法和有限差分法的中间物。有限单元法必须假定值在网格点之间的变化规律（既插值函数），并将其作为近似解。有限差分法只考虑网格点上的数值而不考虑值在网格点之间如何变化。有限体积法只寻求的结点值，这与有限差分法相类似；但有限体积法在寻求控制体积的积分时，必须假定值在网格点之间的分布，这又与有限单元法相类似。在有限体积法中，插值函数只用于计算控制体积的积分，得出离散方程之后，便可忘掉插值函数；如果需要的话，可以对微分方程中不同的项采取不同的插值函数。

4.4 有限体积法解题步骤

有限体积法易于人们理解和使用，并且可以得到合理的解释。它的最大意义在于，使用有限体积法得到的离散方程，完美地体现了守恒性，就同在微分方程中因为不断变化的量而产生不断变小的体积的原理守恒是同样的道理。同时，假设可以更具灵活性，解决了泰勒由于离散产生的一些缺点。其具体步骤如下：

a. 体积分割：在计算过程当中，将需要计算的区域分割成为一连串的具有不重复的控制体积，使各个得以控制的体积都有一个作为代表的节点，把需要求出的方程在随意的控制体积内或是具体的时间间隔内作积分;

b. 提出不同的假设：面对需要求解的函数或导数，通过对它们的时间或空间的变化线做出可能的需要假设，进一步提高计算的精准度，达到所需要的数值，使得工程能够得到更好的解决;

c. 整理：对以上步骤中出现的一系列线型进行类别划分，作出不同的整理。总结出节点上不可知量的离散方程的形式。

这样得到的数值，最大限度满足了守恒，数值更加精确，整个计算推导过程更加清晰。

第 5 节 小结

偏微分方程的数值解法在数值分析中占有很重要的地位，很多科学技术问题的数值计算包括了偏微分方程的数值解问题。近三十多年来，它的理论和方法都有了很大的发展，而且在各个科学技术领域中的应用也愈来愈广泛，求解偏微分方程已经成为科学与工程计算的核心内容，包括一些大型的计算和很多已经成为常规的计算。顺带要提的是，原则上，可以用 FORTRAN 或 C 语言来完成这些计算，但由于成本太高、编程太复杂，这些方法是不太适用的。而 MATLAB 是一种用于算法开发、数据可视化、数据分析以及数值计算的高级技术计算语言和交互式环境，所以使用 MATLAB 可以较使用传统的编程语言更快地解决技术计算问题。

第 III 部分

统计传习

珞珈数学 Mathematics of Luojia March 2018

统计模型：数学与数据间的信使

珞珈数学编辑部　刘禹希　杨夜姿

第 1 节　发展历程

统计学是一门很古老的科学，一般认为其学理研究始于古希腊的亚里士多德时代，迄今已有两千三百多年的历史。它起源于研究社会经济问题，在两千多年的发展过程中，统计学至少经历了“城邦政情”、“政治算数”和“统计分析科学”三个发展阶段。

“统计分析科学”课程的出现是现代统计发展阶段的开端。1908 年 t 分布论文的诞生创立了小样本代替大样本的方法，开创了统计学的新纪元。

在十九世纪初，数学家们逐渐建立了观察误差理论、正态分布理论和最小平方法则。至此，现代统计方法便有了比较坚实的理论基础。

第 2 节　学科展望

统计学的研究通常基于描述性统计方法展开，通过参数估计方法的建立、假设检验的完备流程，最终得到令人信服的结论。在如今大数据量、多变量的应用背景下，如何挖掘样本或变量间的关系、如何描述量化这种关系、如何准确的提取数据中蕴含的信息成为了亟待解决的问题。

在传统理论统计学的基础上，科学家进一步建立统计学习理论，与机器学习知识结合，希望在数据挖掘领域有更长足的进步。在本章节中，我们首先介绍传统统计学的基本方法，希望读者在学习后可以对统计学的原理有一定的熟悉，同时可以使用 R 软件进行简单的模型实现。在后文的交叉视野中，我们使用分类模型领域内的前沿改进举例，有兴趣的读者可以进行深入的学习。

21 世纪的统计学呈现出大繁荣态势。当数学在传统工业中以微分方程作为精细刻画与量化计算的根本工具时，统计学在计算机科学与大数据的助力下，成为新时代里大体量数据处理方式优化、人工智能算法革新的重要基石。

第 3 节　专业软件简介

R 是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能。

珞珈数学 Mathematics of Luojia March 2018

变量关系量化分析理论

珞珈数学研习会　理论统计专栏

第一作者：2018 级　统计学　　陈泽昱
第二作者：2019 级　数学基地班　　刘禹希
第三作者：2019 级　统计学　　彭睿骁

相关性分析是分析变量间关系的第一步，是挖掘随机变量间潜在联系的重要手段，在相关性系数计算的基础上，变量典型相关性可以更深入地挖掘变量组之间的相关性关系。而在较强相关性的基础上，回归方程的建立可以量化变量间的联系，同时可以量化某些随机变量对其他随机变量变化的影响，基于最基础的线性回归，根据整数、0-1 变量、不稳定变量等不同的特点，可以给出多种回归模型解决不同的问题。在本文中，还将给出 R 软件的实现程序。

关键词：典型相关分析、广义线性模型、R 软件

第 1 节　变量组间关系的挖掘：典型相关分析

典型相关分析将相关系数的计算推广到多维随机变量，思想在于寻找一组随机变量的线性组合使之最大程度上代表两组随机变量的相关程度。首先，以第一组典型变量为例说明如何求解典型相关变量：

假设随机向量 $X=(X_1,X_2,\cdots,X_p)^T$ 与 $Y=(Y_1,Y_2,\cdots,Y_q)^T$ 满足：(X,Y) 均值为 0，协方差矩阵为正定阵 Σ，则求解典型相关分析问题转换为如下优化问题：

$$\max \rho(V,W) \tag{1}$$

$$s.t.\begin{cases} V=\alpha^T X \quad W=\beta^T Y \\ Var(\alpha^T X)=1 \quad Var(\beta^T Y)=1 \end{cases} \tag{2}$$

1.1　典型相关分析基本概念

一般地，在如下定理中给出典型相关分析的一般性结果：

设 $Z=(X^T,Y^T)^T$，其中 $X=(X_1,\cdots,X_p)^T$ 为 p 维随机向量，$Y=(Y_1,\cdots,Y_q)^T$ 为 q 维随机向量，不妨设 $p\le q$。已知：

$$E(Z)=0 \quad D(Z)=\Sigma=\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}>0 \tag{3}$$

记 $T=\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ 并设 p 阶方阵 TT^T 的特征值依次为 $\lambda_1^2\geq\lambda_2^2\geq\cdots\geq\lambda_p^2(\lambda_i>0)$ 而 $l_1,l_2,\cdots,l_p$ 为相应的单位正交特征向量，则如下表示 X,Y 的第 k 对典型相关变量系数：

$$a_k=\sigma_{11}^{-\frac{1}{2}}l_k\quad b_k=\lambda_k^{-1}\Sigma_{22}^{-1}\Sigma_{21}a_k \tag{4}$$

因此 $V_k=a_k^TX,W_k=b_k^TY$ 即为 X，Y 的第 k 对典型相关变量，λ_k 为第 k 个典型相关系数。

由以上定理内容立即可以得到如下两结论：

1. 同组典型相关变量间的相关性关系：

性质 1

$\rho(V_k,V_i)=0=\rho(W_k,W_i)\quad(i<k)$
$\rho(V_k,V_k)=1=\rho(W_k,W_k)$

2. 不同组典型相关变量间的相关性：

性质 2

$\rho(V_i,W_j)=0\quad(i\neq k)$
$\rho(V_k,W_k)=\lambda_j$

基于以上性质，给出典型结构 D(Z) 的概念如下：

定义 1

设 V_k,W_k 为 X,Y 的第 k 对典型相关变量，设 $V=(V_1,\cdots,V_p)^T,W=(W_1,\cdots,W_p)^T,Z=(V^T,W^T)^T$，则由以上定理表明如下结论：

$$D(Z)=\Sigma=\begin{pmatrix}I_p & \Lambda\\ \Lambda & Ip\end{pmatrix}\quad\Lambda=diag(\lambda_1,\lambda_2,\cdots,\lambda_p) \tag{5}$$

基于典型结构的定义，可以考虑对随机向量作线性变换，可以进一步得到第三条性质：

性质 3

设 X 为 p 维随机向量,Y 为 q 维随机向量，令 $X^*=C^TX+d,Y^*=G^TY+h$, 则有如下结论:

(1)X^* 和 Y^* 的典型相关变量为 $(a_i^*)^TX^*$ 和 $(b_i^*)^TY^*$，且有关系:$a_i^*=C^{-1}a_i,b_i^*=G^{-1}b_i$，其中 a_i,b_i 为 X 与 Y 的第 i 对典型相关变量的系数。

(2) 线性变化不改变相关性，即：$\rho[(a_i^*)^TX^*,(b_i^*)^TY^*]=\rho(a_i^TX,b_i^TY)$。

1.2 典型相关系数的显著性检验

首先，如果 X 与 Y 变量间无相关关系，则无法进行典型相关分析，因此要检验的第一项即为 X 与 Y 的协方差矩阵是否为 0 矩阵。零假设为协方差矩阵为 0 矩阵，设总体 $Z\sim N_{p+q}(\mu,\Sigma)$，用似然比方法可以得到如下统计量:

$$\Lambda=\frac{|S|}{|S_{11}||S_{22}|}=\prod_{i=1}p(1-\hat{\lambda_i}^2) \tag{6}$$

其中 p+q 阶矩阵 S 是 Σ 的最大似然估计，S_{11},S_{22} 分别是 Σ_{11},Σ_{22} 的最大似然估计。在常规情形

下上述统计量的精确分布难以写出，但是可以进行近似分布的计算，下面给出 Box 的检验方法：

$$-m\ln(\Lambda) \sim \chi^2(f) \tag{7}$$

其中 $m = n - 1 - \frac{1}{2}(p+q+1), f = pq$，拒绝域取为右侧拒绝域，显著性水平为 α。

如果第一次检验通过，可以得出至少第一个典型相关系数非 0。但是如果第一组典型相关变量已经提取了大多数信息，则余下的部分近似不相关，因此需要检验 λ_k 的非 0 性。这里采用 Bartlett 提出的大样本卡方检验如下选取统计量：

$$Q_k = -[n-k-\frac{1}{2}(p+q+1)]\sum_{i=k}^{p}\ln(1-\hat{\lambda_i}^2) \tag{8}$$

则上述统计量在零假设即 $\lambda_k = 0$ 下，上述统计量有如下分布：

$$Q_k \sim \chi^2(f_k) \tag{9}$$

其中 $f_k = (p-k+1)(q-k+1)$，拒绝域取右侧拒绝域，显著性水平为 α。

在上述过程中，所有拒绝原假设的 k 都可以被认为可以进行对应阶数的典型相关变量的提取。

1.3 典型冗余分析---信息提取量的确定

可以类似主成分分析中信息量的概念，定义总变差百分比如下：

定义 2

1. 如下给出解释 X 随机向量本组的典型变量信息量比例：

$$R_d(X;V_1,\cdots,V_m) = \frac{1}{p}\sum_{k=1}^{m}\sum_{j=1}^{p}r^2(X_j,V_k) \tag{10}$$

2. 再给出解释 Y 随机向量本组的典型变量信息量比例：

$$R_d(Y;W_1,\cdots,W_m) = \frac{1}{q}\sum_{k=1}^{m}\sum_{j=1}^{p}r^2(Y_j,W_k) \tag{11}$$

类似地可以定义解释另一组随机变量的典型变量的信息量比例：

定义 3

1. 给出利用典型变量 W 解释 X 随机向量的信息量比例：

$$R_d(X;W_k) = \frac{1}{p}\sum_{j=1}^{p}r^2(X_j,W_k) = \lambda_k^2 R_d(X,V_k) \tag{12}$$

2. 再给出利用典型变量 V 解释 Y 随机向量的信息量比例：

$$R_d(Y;V_k) = \frac{1}{q}\sum_{j=1}^{p}r^2(Y_j,V_k) = \lambda_k^2 R_d(Y,W_k) \tag{13}$$

如上定义的 $R_d(X;W_k)$ 解释原变量组的变差被第二组中典型变量重复解释的百分比，称为第一组典

型变量的冗余测度。如上定义的 $R_d(Y;V_k)$ 解释原变量组的变差被第一组中典型变量重复解释的百分比，称为第二组典型变量的冗余测度。

冗余测度可以体现两组变量间的相关程度，冗余测度越大变量的相关程度越高。

1.4 典型相关分析的 R 软件代码实现

典型相关分析是经典的统计关联分析方法，因此在 R 软件中有封装的程序组，只需要掌握调用方法即可轻松实现典型相关分析有关的计算。

```
#Input:
#X,Y即为两组随机变量数据框，格式为：行样本，列随机变量指标
CCA.result = cancor(X, Y)
#Output:
#cor即为典型变量的相关系数
#xcoef第i列为第i个典型变量的系数向量
X_CCA = X %*% CCAresult$xcoef[,1:dim]
Y_CCA = Y %*% CCAresult$ycoef[,1:dim]
#第k列即为第k个典型变量
lambda = CCAresult$cor
```

第 2 节 广义回归模型与修正方法

传统的线性回归已经建立了完备的统计分析方法，从模型到参数估计、假设检验、拟合优度判别等理论具有较高的理论价值。但是在现实的应用中，线性回归由于要求变量为连续且服从正态分布的变量、需要大量的标签数据、无法处理数据与时间的联系、模型对数据的波动敏感等特点而在使用中受到了很多局限。因此近年来提出的广义线性模型理论是对传统线性模型的重要补充与拓展。

2.1 泊松回归简介

如果待拟合变量为计数型变量或类别型变量，直接使用线性回归的方式会损失待拟合变量本身的分布信息，这样的回归模型大多无法通过正态性检验，因此有必要引入一种针对计数型变量的回归模型。考虑到泊松分布是对计数型变量的分布描述，因此引进泊松回归解决这类问题。

假设 X 为一组相互独立的随机变量组成的随机向量，Y 为响应变量，因此在假设 Y 服从泊松分布时，可以如下通过期望值建立回归方程：

$$\log(E(Y|X)) = \theta_T X \tag{14}$$

参数极大似然估计

由于 Y 的泊松分布假设，且在回归表达式中给出了均值的估计，因此可以完全确定 Y 的样本联合密度函数，以此作为似然函数如下：

$$P(y_1,\cdots,y_m|x_1,\cdots,x_m,\theta) = \prod_{i=1}^{m} \frac{e^{y_i\theta^T x_i - e^{\theta^T x_i}}}{y_i} \tag{15}$$

进一步可以构造似然函数的负对数作为损失函数：

$$L(\theta|X,Y) = -\sum_{i=1}^{m}(y_i\theta^T x_i - e^{\theta^T x_i}) \tag{16}$$

通常利用 Newton-Raphson 算法进行损失函数的最优化求解得到最小值点，该算法将在逻辑回归的小节中详细介绍。

过度离势检验

类似线性回归中对残差正态性进行检验，在泊松回归中，假设原变量 Y 服从泊松分布，因此至少应有期望与方差相等的性质，而如果不满足这个必要条件，称 Y 的观测数据发生了过度离势现象。在原变量服从泊松分布时，观测数据依然可能产生过度离势现象，产生过度离势的原因主要有如下几种：

1. 在泊松回归中遗漏了很重要的解释变量。
2. 在对原始随机变量进行观测时，没有做到样本点间观测的独立性。
3. 在纵向数据分析中，重复测量的数据可能产生过度离势，因为其具有内在群聚特性。

通过如下定理给出对过度离势数据的粗略判别：

定理 1

设 poisson 回归模型拟合后的残差平方和为:dev，设残差自由度为:df。则首先通过如下等式计算判断指标：

$$\Phi = \frac{dev}{df} \tag{17}$$

如果计算值在较大程度上远离 1，则原序列存在严重的过度离势。

过度离势检验在 R 软件中可以直接使用 qcc 包中的函数实现，如果出现过度离势线性，此时在拟合模型时需要考虑假设响应变量为类泊松分布。

泊松回归的实现

泊松回归在 R 软件的封装分类中属于广义线性模型类，因此实现时通过调用线性模型软件组 *glm* 函数可以实现相关结果的计算，代码示例如下。

```
#package:
library(qcc)
#Input:
#data 每一列均为自变量的数据框
#data$y.logistic 泊松回归整数因变量
#namek 第k列的data数据框名
#未修正过度离势:
poisson.model =
glm(data$y.poisson ~ name1 + name2, data = data, family = poisson())
#已修正过度离势:
poisson.od.model =
glm(data$y.poisson ~ name1 + name2,
data = data, family = quasipoisson())
#Output:
#poisson.coef回归系数poisson.coef.confint置信区间
#poisson.deviance过度离势检验量poisson.df.residual
#poisson.value 回归的P值
poisson.coef = as.numeric(poisson.model$coefficients)
```

```
poisson.deviance = poisson.model$deviance
poisson.df.residual = poisson.model$df.residual
poisson.value = poisson.model$fitted.values
#Test:
fai = poisson.deviance/poisson.df.residual
qcc.overdispersion.test(data$y.poisson, type = "poisson")
```

2.2 正则化回归模型

在传统的回归模型中，可以通过最小二乘法建立系数的估计，同时可以证明这个估计具有很好的性质，如无偏性以及进一步的线性最优估计性质等性质。但是在实际的应用中，如果得到的系数值过大，回归值会被自变量的微小扰动影响从而导致模型对数据过于敏感，为了克服这种敏感性，在损失函数中引入正则项，同时也牺牲了无偏性。

$$E(w) = E_{ols}(w) + \lambda J(w) \tag{18}$$

其中 λ 为超参数，在不同的回归模型中可以取不同的值以获得不同的意义。

而根据正则项的形式可以分为：岭回归、Lasso 回归、ElasticNet 回归，如下分别对其损失函数与回归估计结果进行介绍：

$$J(w) = \frac{||w||^2}{2} \tag{19}$$

上述模型为岭回归模型的损失函数，如此构造损失函数可以修正回归方程中的不稳定系数，进而可以使用类似最小二乘的梯度求法得到最终的系数估计值如下：

$$w_{rig} = (Z^T Z + \lambda I)^{-1} Z^T y \tag{20}$$

对回归模型中出现大量接近 0 系数的求解结果，可以采用 Lasso 回归正则项进行修正，如下构造损失函数：

$$J(w) = ||w||_1 \tag{21}$$

由于不可求导，Lasso 回归只能使用最优解搜索算法进行最优化处理，因此计算复杂度较高，但是可以有效解决系数求解的稀疏问题。而 ElasticNet 回归则是这两种回归方式的综合结果，可以如下构造损失函数：

$$J(w) = \rho||w||_1 + (1-\rho)||w||^2 \tag{22}$$

当超参数值接近 1 时，对稀疏系数的修正效果更明显，接近 0 时，对不稳定的回归系数修正更为明显，对于这种损失函数的构造方式，也只能使用最优化的搜索方法进行求解。

如下以参数 0.5 时的 ElasticNet 回归为例，给出 R 软件实现含正则项的回归模型代码：

```
#package:
library(glmnet)
library(MASS)
#Input:
#x 自变量形成的矩阵，列为指标行为向量
#y 待回归变量
#family 回归类型，默认为guassian线性最小二乘
#alpha 正则项参数
elasticnet.model = glmnet(x, y, family = "binomial", alpha = 0.5)
#Output:
```

```
#lambda 选取零变量的比例
#dev.ratio 拟合优度
#coef 回归方程中的系数
elasticnet.coef = coef(elasticnet.model,
 s=elasticnet.model$lambda[16])
```

2.3 偏最小二乘回归模型

传统线性回归方法需要大量的标签数据且只能处理单一变量与一组随机变量的回归关系，在现实中如果需要建立两组随机变量间的回归关系或在样本量较小情形下需要建立回归过程，近年来提出的偏最小二乘回归 (PLS) 可以作为一个重要的方法。

设两组变量值分别为 $\{Y_1,\cdots,Y_p\}$ 为因变量而 $\{X_1,\cdots,X_m\}$ 为自变量，PLS 思想为首先分别提取两组变量间相关性最大的一组变量再分别对两组变量建立回归关系，在残差矩阵满足一定条件时结束建模，具体过程如下：

线性变量的提取

设 $X=\{X_1,\cdots,X_m\}, Y=\{Y_1,\cdots,Y_p\}$, 分别提取第一对成分 T_1, U_1，应该获得变量组中尽可能多的变异信息：

$$T_1 = w_1^T X \qquad U_1 = v_1^T Y \tag{23}$$

$$\begin{cases} w_1^T w_1 = 1 \qquad v_1^T v_1 = 1 \\ \max t_1^T u_1 \end{cases} \tag{24}$$

对于上述求解使用拉格朗日数乘法，这里直接给出最终的求解结论作为定理：

定理 2

假设矩阵 $M = X^T Y Y^T X$ 最大的特征值为 θ_1^2，则对应的特征向量即为 w_1，进一步 v_1 可以按照如下等式进行计算：

$$v_1 = \frac{1}{\theta_1} Y^T X w_1 \tag{25}$$

建立多维回归模型

分别建立 Y 变量组与 X 变量组对 T_1 的回归，如下：

$$\begin{cases} X = t_1 \alpha_1^T + E_1 \\ Y = t_1 \beta_1^T + F_1 \end{cases} \tag{26}$$

其中由于 X 与 Y 变量组均有样本观测值，因此可以使用最小二乘估计得到系数的估计值，如下可以得到回归模型的系数估计结果：

$$\begin{cases} \alpha_1 = (t_1^T t_1)^{-1} X^T t_1 \\ \beta_1 = (t_1^T t_1)^{-1} Y^T t_1 \end{cases} \tag{27}$$

因此可以计算残差值矩阵 E_1, F_1，如果 F_1 中元素的绝对值近似为 0，认为第一成分的回归已经可以满足需要，可以停止抽取成分，如果无法满足精度要求，可以用上述残差值矩阵代替对应的变量组样

本观测矩阵重复以上步骤提取其他阶成分，最终可以得到如下的回归方程组：

$$\begin{cases} X = t_1\alpha_1^T + \cdots + t_r\alpha_r^T + E_r \\ Y = t_1\beta_1^T + \cdots + t_r\beta_r^T + F_r \\ T_k = w_{k1}X_1 + \cdots + w_{km}X_m \quad (k = 1, \cdots, r) \end{cases} \tag{28}$$

由于 T、Y、X 三者的关系，同时对标准化的变量进行还原，可以最终建立 X 与 Y 的回归方程：

$$Y_j = a_{j0} + a_{j1}X_1 + \cdots + a_{jm}X_m \quad (j = 1, 2, \cdots, p) \tag{29}$$

确定抽取的成分个数 l

如果使用全部的 r 个成分，通常带来计算复杂度很高的问题，因此在实际应用中，仅对前 l 个成分进行选取即可得到预测能力较强的回归模型。如下介绍一种确定抽取成分 l 的方法：舍一交叉验证法。

每次舍去第 i 个观测值，用余下的 n-1 个样本观测值建立 PLS 模型，最后将舍去的第 i 个观测点代入使用 k 个成分拟合的回归方程得到 Y 中所有变量在第 i 个观测点上的预测值 $y_{j(i)}(k)$，重复 i 为 1 至 n，可以如下计算 Y 的预测残差平方和 (PRESS)：

$$PRESS(k) = \sum_{j=1}^{p} PRESS_j(k) = \sum_{j=1}^{p}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{j(i)}(k))^2 \quad (k = 1, \cdots, r) \tag{30}$$

令 k 从 1 取到 r，最终选取其中 PRESS 最小的 k，令 l 取为 k 即可得到抽取的成分数 l。

第 3 节 参考文献

[1] 许欢, 苏树智, 颜文婧, 邓瀛灏, 谢军. 面向图像识别的测地局部典型相关分析方法 [J/OL]. 电子与信息学报:1-6[2020-09-20].http://kns.cnki.net/kcms/detail/11.4494.TN.20200729.1017.006.html.

[2] Muhammad Qasim, B. M. G. Kibria, Kristofer Månsson, et al. A new Poisson Liu Regression Estimator: method and application. 2020, 47(12):2258-2271.

[3] McEligot Archana J, Poynor Valerie, Sharma Rishabh, et al. Logistic LASSO Regression for Dietary Intakes and Breast Cancer.. 2020, 12(9)

[4] 蒋程, 寿旦, 俞忠明, 许平翠, 王绪平, 陈礼平, 张晓芹, 王娜妮. 基于紫外光谱和偏最小二乘回归算法的畲药地稔中浸出物和 6 种活性成分快速预测方法 [J]. 中国现代应用药学,2020,37(13):1574-1579.

珞珈数学 Mathematics of Luojia March 2018

数据降维与信息提取理论

珞珈数学研习会　理论统计专栏

第一作者：2018 级　统计学　　吴晨
第二作者：2019 级　数学基地班　　马金韬
第三作者：2019 级　金融数学　　邓琪雯

在大数据应用背景下，高维度随机向量分量间隐含的关系难以直接通过可视化的方法寻找，因此首先需要通过降维方法对数据进行处理。主成分分析通过对变量进行线性组合，去除变量间的强线性关系，以达到分解其中蕴含信息的目的；因子分析通过剖析变量组内与共同的线性成分，给出变量间共线性的表达式，达到信息浓缩的目的；最后通过对应分析可以对因子的信息浓缩结果进行统计描述。本节介绍的三个方法均为数据降维的常用方法。

关键词：主成分分析、因子分析、对应分析

第 1 节　主成分分析

主成分分析（Principal Component Analysis，PCA），是一种统计方法。首先是由 K. 皮尔森（Karl Pearson）对非随机变量引入的，H. 霍特林将此方法推广到随机向量的情形。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分，信息的大小通常用离差平方和或方差来衡量。

首先可以给出主成分的基本定义。

定义 1

设 $X=(X_1,X_2,\cdots,X_p)^T$ 为 p 维随机向量，Σ为X的协方差矩阵，称 $Z_i=a_i^TX$ 为 X 的第 i 主成分 $(i=1,2,\cdots,p)$，如果：

(1)　$a_i^Ta_i=1$;

(2)　当 $i>1$ 时，$a_i^T\Sigma a_j=0\quad(j=1,\cdots,i-1)$

(3)　$Var(Z_i)=\max Var(a^TX)$

在如上的主成分定义下，主成分的求解转换为条件极值问题的求解，因此在均值为 $E(X)$，方差为 Σ 时，可以通过如下方程组求解之：

$$\begin{cases}\phi(a_i)=Var(a_i^TX)-\lambda(a_i^Ta_i-1)\\ \frac{\partial\phi}{\partial a_i}=2(\Sigma-\lambda I)a_i=0\\ \frac{\partial\phi}{\partial\lambda}=a_i^Ta_i-1=0\end{cases}\tag{1}$$

求解上述等式并如下描述求解结果：

1. 设 $X=(X_1,\cdots,X_p)^T$ 是 p 维随机向量，且协方差矩阵为 Σ，特征值为 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p\geq0$,

对应的单位正交向量为 $a_1, a_2, \cdots, a_p$，则 X 的第 i 主成分为：

$$Z_i = a_i^T X \quad (i = 1, 2, \cdots, p) \tag{2}$$

2. 设 $Z = (Z_1, \cdots, Z_p)^T$ 是 p 维随机向量，则其分量 $Z_i (i = 1, 2, \cdots, p$ 依次是 X 的第 i 主成分的充分必要条件为：

(1) $Z = A^T X$，A 为正交矩阵;

(2) $D(Z) = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$，即随机向量 Z 的协方差矩阵为对角矩阵;

(3) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

1.1 主成分的基本性质

基于前一小节对主成分基本概念的理解，本文将在本节提出主成分的基本性质，为后文阐述主成分分析应用做充足的准备。

记 $\Sigma = (\sigma_{ij}), \Lambda = diag(\lambda_1, \lambda_2, \cdots, \lambda_p)$，其中 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ 为 Σ 的特征值，$a_1, a_2, \cdots, a_p$ 是相应的单位正交特征向量，记正交矩阵 $A = (a_1, a_2, \cdots, a_p)$。主成分 $Z = (Z_1, \cdots, Z_p)^T$，其中 $Z_i = a_i^T X (i = 1, 2, \cdots, p)$。

性质 1

$D(Z) = \Lambda$，即 p 个主成分的方差为：$Var(Z_i) = \lambda_i$，且所有主成分不相关。

性质 2

$\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i$，这个求和结果通常称为原总体 X 的总方差（或总惯量）。

性质 3

主成分 Z_k 与原始变量 X_i 的相关系数 $\rho(Z_k, X_i)$（又称因子负荷量或因子载荷量）为：

$$\rho(Z_k, X_i) = \frac{\sqrt{\lambda_k} a_{ik}}{\sqrt{\sigma_{ii}}} \quad (k, i = 1, 2, \cdots, p) \tag{3}$$

上述定义的因子载荷满足如下性质：

$$\begin{cases} \sum_{k=1}^{p} \rho^2(Z_k, X_i) = 1 \quad (i = 1, 2, \cdots, p) \\ \sum_{i=1}^{p} \sigma_{ii} \rho^2(Z_k, X_i) = \lambda_k \quad (k = 1, 2, \cdots, p) \end{cases} \tag{4}$$

1.2 样本主成分及其应用

主成分分析的一个重要的应用是对样本进行排序或对系统进行综合评估，样本主成分分析模型包括

数据的标准化、样本相关系数矩阵的确定、标化数据主成分分解、信息贡献率的计算、综合得分的计算五个主要步骤，下文将逐一进行介绍：

数据的标准化与相关系数求解：

由于各个指标间的量纲存在较大的差异，因此对于行指标为样本编号、列指标为指标编号的数据矩阵需按照下面公式中的方法进行列标准化，以此获得相同量纲与 0 均值的标化样本。

$$\tilde{x_{ij}} = \frac{x_{ij} - \bar{x_j}}{s_j} \quad (i = 1, 2, \cdots, n; \quad j = 1, 2, \cdots, m) \tag{5}$$

$$\begin{cases} x_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \\ s_j = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2 \end{cases} \tag{6}$$

对于此时的样本矩阵 $\tilde{X} = (\tilde{x_{ij}})_{n\times m}$, 每一列的均值为 0，方差为 1，因此可以直接依下式计算标化样本的相关系数矩阵，其中 r_{ij} 为第 i 个与第 j 个指标的相关系数：

$$R = \frac{1}{n-1}\tilde{X}^T\tilde{X} \tag{7}$$

样本主成分分析过程：

根据第一小节对主成分解的介绍，假设 R 矩阵的特征值为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$，对应特征向量为：$u_1, u_2, \cdots, u_m$，其中 $u_j = (u_{1j}, u_{2j}, \cdots, u_{nj})^T$，则可以建立如下等式计算主成分解：

$$\begin{cases} y_1 = u_{11}\tilde{x_1} + u_{21}\tilde{x_2} + \cdots + u_{n1}\tilde{x_n} \\ y_2 = u_{12}\tilde{x_1} + u_{22}\tilde{x_2} + \cdots + u_{n2}\tilde{x_n} \\ \cdots \\ y_m = u_{1m}\tilde{x_1} + u_{2m}\tilde{x_2} + \cdots + u_{nm}\tilde{x_n} \end{cases} \tag{8}$$

信息贡献率的确定：

结合第 2 小节对特征值的介绍，可以如下定义信息贡献率：

定义 2

假设提取 p 维随机向量的 m 个主成分，对应的特征值分别为 λ_i，则可以将 $b_k = \frac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ 定义为主成分 Z_k 的信息贡献率，进一步，可以将 $\frac{\sum_{k=1}^{m}\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ 定义为主成分 $Z_1, Z_2, \cdots, Z_m$ 的累计贡献率。

因此当累计贡献率在选取了 m 个主成分之后接近 1，则可以只选取 m 个主成分包含几乎所有的样本信息。

对每个样本计算主成分得分：

在如上定义信息贡献率之后，可以依据步骤 2 中的主成分值与步骤 3 中的信息贡献率结合计算每个样本的主成分得分，如下式：

$$Z = \sum_{j=1}^{m} b_j y_j \tag{9}$$

对每一个样本计算主成分得分后可以对 Z 进行排序以确定样本的顺序，由于主成分排序靠前的变量

所涵盖的变异信息较多，因此主成分的加权求和可以体现从主成分特征的角度对样本进行的打分，假设样本中每一个指标较高的数值代表较好的状态，这个排序的意义是：状态越好得分越高。

1.3 主成分分析的 R 软件实现

主成分分析是经典的数据降维算法，因此常被用作处理大数据量的正交化等问题，在 R 软件中封装的 prcomp 函数可以帮助使用者实现之。

```
#Input:
#data    变量：行为样本、列为因子
#retx    是否返回旋转变量
#center  均值归0化，scale方差归1化
#rank    主成分最大数量，在主成分数量远小于矩阵维数时有用
PCA.result = prcomp(data,retx = TRUE,
 center = FALSE, scale. = FALSE, rank. = NULL)
#Output:
#scale   每一个变量的方差 center 均值
#rotation 主成分系数
#x        主成分输出(在不标准化下右乘rot等于原数)
#sdev     主成分标准差（奇异值开方） sqrt(lambda_i)
#prop of var 方差占比
 data.scale = PCA.result$scale
 data.center = PCA.result$center #样本均值与方差
PCA.data = PCA.result$x      #主成分样本数据
PCA.rotation = PCA.result$rotation #主成分系数矩阵
PCA.var = PCA.result$sdev #主成分的方差值
```

第 2 节 因子分析

因子分析由英国心理学家 C.E. 斯皮尔曼提出，是从变量群中提取共性因子的统计模型。因子分析可在许多变量中找出隐藏的具有代表性的因子，将相同本质的变量归入一个因子，可减少变量的数目，还可检验变量间关系的假设。

2.1 因子降维基本概念

首先介绍正交因子模型相关理论，设 $X=(X_1,X_2,\cdots,X_p)^T$ 是可观测的随机向量，且 $E(X)=\mu, D(X)=\Sigma$, 设 $F=(F_1,F_2,\cdots,F_m)^T$ 为不可观测的随机向量，且 $E(F)=0, D(F)=I_m$, 又设 $\epsilon=(\epsilon_1,\epsilon_2,\cdots,\epsilon_p)$ 与 F 互不相关且 $E(\epsilon)=0, D(\epsilon)=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)$，首先提出以下两条基本假设：

1. 特殊因子互不相关，且 $D(\epsilon)=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)$。
2. 特殊因子同公共因子不相关，即 $Cov(\epsilon,F)=O_{p\times m}$

在上述假设下可以依照下式建立正交因子模型：

$$\begin{cases} X_1-\mu_1=a_{11}F_1+a_{12}F_2+\cdots+a_{1m}F_m+\epsilon_1 \\ X_2-\mu_2=a_{21}F_1+a_{22}F_2+\cdots+a_{2m}F_m+\epsilon_2 \\ \cdots \\ X_p-\mu_p=a_{p1}F_1+a_{p2}F_2+\cdots+a_{pm}F_m+\epsilon_p \end{cases} \tag{10}$$

可以有以下简单的矩阵表示方法：

$$X = \mu + AF + \epsilon \tag{11}$$

下面基于以上正交因子模型给出部分概念的定义：

定义 3

1. 公共因子：$F_1, F_2, \cdots, F_m$
2. 特殊因子：$\epsilon_1, \epsilon_2, \cdots, \epsilon_p$
3. 因子载荷：a_{ij} 成为第 i 个变量在第 j 个因子上的载荷，A 称为因子载荷矩阵。

对因子分析模型两边随机变量求协方差可以得到如下等式 (称为协方差结构)：

$$\Sigma = E[(X-\mu)(X-\mu)^T] = D + AA^T \tag{12}$$

$$\begin{cases} \sigma_{jj} = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + \sigma_j^2 \\ \sigma_{jk} = a_{j1}a_{k1} + a_{j2}a_{k2} + \cdots + a_{jm}a_{km} \quad (j \neq k) \end{cases} \tag{13}$$

$$Cov(X, F) = E[(X - E(X))(F - E(F))^T] = A \tag{14}$$

$$\rho_{ij} = Cov(X_i, F_j) = a_{ij} \tag{15}$$

基于上述理论，可以如下建立变量共同度与因子方差贡献的概念：

定义 4

1. 共同度：表现全部的公共因子对变量 X_i 的总方差的贡献, 可以用于刻画模型的解释能力。

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2 (i = 1, 2, \cdots, p) \tag{16}$$

2. 方差贡献：表现第 j 个公因子 F_j 的重要性，利用列平方和求得。

$$q_j^2 = \sum_{i=1}^{p} a_{ij}^2 (j = 1, 2, \cdots, m) \tag{17}$$

值得注明的是：量纲变化过程即为用对角阵乘 D；由于对 F 施行正交变换不改变 X 的形态，因此可以认为载荷矩阵 A 不唯一，利用不唯一性可以进行因子旋转，在一小节后着重介绍。

2.2 主成分法估计参数

假设样本协方差矩阵为 S，具有特征值 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$，对应的单位正交特征向量分别为 $l_1, l_2, \cdots, l_p$，则 S 在最后 p-m 个特征值较小时有如下近似谱分解式：

$$S \approx \lambda_1 l_1 l_1^T + \cdots + \lambda_m l_m l_m^T + D \tag{18}$$

因此可以近似得到如下的主成分解：

$$\begin{cases} A = (\sqrt{\lambda_1}l_1, \cdots, \sqrt{\lambda_m}l_m) \\ \sigma_i^2 = s_{ii} - \sum_{t=1}^{m} a_{it}^2 \quad (i = 1, 2, \cdots, p) \end{cases} \tag{19}$$

下文补充说明如何确定 m，通常预先确定一个阈值 P，如果再选取满足如下等式的最小整数 m 即可：

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_m + \cdots + \lambda_p} \geq P \tag{20}$$

2.3 方差最大化因子旋转

由于因子载荷接近时的因子分析模型不具有现实的解释意义，因此考虑对因子进行正交旋转。因子旋转的合理性基于以下定理：

定理 1

设因子模型为 $X = AF + \epsilon$，令 $Z = \Gamma^T F$(Γ为任意m阶正交矩阵)，则 $X = A\Gamma Z + \epsilon$，此时 $A\Gamma$ 为 Z 的因子载荷矩阵。

可以利用共同度结合载荷矩阵得到 A 的每列方差之和作为 A 的总方差，计算式如下：

$$V(A) = \sum_{j=1}^{m} V_j = \frac{1}{p^2}\{\sum_{j=1}^{m}[p\sum_{i=1}^{p}\frac{a_{ij}^4}{h_i^4} - (\sum_{t=1}^{p}\frac{a_{tj}^2}{h_t^2})^2]\} \tag{21}$$

方差越大则表明因子载荷值越趋于 0 或 1，也就更具有简化结构，因此我们希望通过旋转使上述方差最大化, 即求解如下条件极值问题：

$$\begin{cases} B = A\Gamma \quad A\text{为已知量} \\ \max\limits_{\Gamma} V(B) \end{cases} \tag{22}$$

2.4 巴特莱特因子得分简介

如果考虑将公因子表示为变量的线性组合或对每一个已知的样本计算公共因子的估计值，即为计算因子得分的过程，本文主要介绍巴特莱特因子得分的主要内容：

对于常规的因子分解模型：$X = AF + \epsilon$，将特殊因子看做误差，由于其方差往往不同，因此采用加权最小二乘方法估计因子 F 的值，即使用误差方差的倒数作为权重的误差平方和:

$$\phi(F) = \sum_{i=1}^{p}\frac{\epsilon_i^2}{\sigma_i^2} = \epsilon^T D^{-1}\epsilon = (X - AF)^T D^{-1}(X - AF) \tag{23}$$

只需要最小化上式即对 F 求偏导数置 0，可以解出如下 F 的估计值：

$$\hat{F} = (A^T D^{-1} A)^{-1} A^T D^{-1} X \tag{24}$$

在实际应用中，A、D 的值往往是未知量，因此必须考虑利用估计值代替真值，而利用主成分法估

计时，令 D 为单位阵、令 A 为主成分解即可化为普通最小二乘解，同时对比样本因子得分与样本主成分得分可以观察到如下式：

$$f_{ij} = \frac{z_{ij}}{\sqrt{\lambda_j}} \quad (i = 1, \cdots, n; j = 1, \cdots, m) \tag{25}$$

2.5 因子分析的 R 软件实现

因子分析通常需要经过：因子模型建立、参数估计、因子旋转、因子得分计算等步骤，在实际应用中较为繁琐，R 软件提供了 factanal 函数对上述过程进行封装，如下进行代码演示：

```
#Input:
#data    变量：行为样本、列为因子
#factor.num    指定的因子个数
#scores        因子得分计算方法
#rotation      因子旋转的方法
EFA.result = factanal(data, factors = factor.num,
 scores = "Bartlett",rotation = "varimax")
#Output:
#loading   因子载荷
#rotation.matrix  因子旋转矩阵
#stange.Var       特殊方差值
loadings = EFA.result$loadings #载荷+方差解释量
loading = loadings[,1:factor.num] #载荷
rotation.matrix = EFA.result$rotmat
strange.Var = EFA.result$uniquenesses #特殊方差
```

第 3 节 对应分析

在上文介绍的因子分析中，往往只涉及到针对样本的因子提取或针对指标的因子提取，且因子分析方法只对连续性变量值最有效。在本节中，将尝试将因子分析的方法推广到多分类变量，同时试图同时对样本与指标进行因子提取并将该因子提取结果绘制在同一张图中以研究变量间的相关关系。在本节中，我们引入对应分析方法，又称 R-Q 因子分析。

3.1 数据标准化处理

在分析进行之前，需要对连续性或计数型数据矩阵 $X = (x_{ij})_{n\times p}$ 进行概率化处理，如下进行：

$$\begin{cases} X_{i\cdot} = \sum_{k=1}^{p} x_{ik} \\ X_{\cdot j} = \sum_{k=1}^{n} x_{kj} \\ T = \sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij} P = \frac{X}{T} \end{cases} \tag{26}$$

在如上的处理过程中，提取了数据的边缘分布与概率格式，得到的矩阵 P 为所有元素求和为 1 的概率形式矩阵。之后同时考虑数据的行列边缘分布，可以进一步进行标准化处理得到标准化数据矩阵 Z 如

下式：

$$z_{ij}=\frac{x_{ij}-\frac{X_{i\cdot}X_{\cdot j}}{T}}{\sqrt{X_{i\cdot}X_{\cdot j}}}\quad(i=1,\cdots,n;j=1,\cdots,p) \tag{27}$$

3.2 计算行列轮廓分布与重心

分别消除行列出现概率不同的影响，即可以得到行列轮廓分布：

$$r_{ij}=\frac{x_{ij}}{X_{i\cdot}} \tag{28}$$

通过上式消除行间概率不同的影响，得到行轮廓分布为 np 维矩阵 R，表示为：$R=(R_1,\cdots,R_n)^T$。

$$c_{ij}=\frac{x_{ij}}{X_{\cdot j}} \tag{29}$$

通过上式消除列间概率不同的影响，得到列轮廓分布为 np 维矩阵 R，表示为：$R=(C_1,\cdots,C_p)^T$。

$$N(R)=\sum_{i=1}^{n}P_{i\cdot}R_i=(\sum_{i=1}^{n}p_{i1},\sum_{i=1}^{n}p_{i2},\cdots,\sum_{i=1}^{n}p_{ip})^T=(P_{\cdot 1},P_{\cdot 2},\cdots,P_{\cdot p})^T \tag{30}$$

N(R) 即为 n 个行轮廓构成点集的重心。

3.3 总惯量与相关性的判定

首先定义第 k 个与第 l 个样本的加权距离，如下式：

$$D^2(k,l)=\sum_{j=1}^{p}\frac{(\frac{p_{kj}}{P_{k\cdot}}-\frac{p_{lj}}{P_{l\cdot}})^2}{P_{\cdot j}} \tag{31}$$

下面给出总惯量的定义式：

$$Q=\sum_{i=1}^{n}P_{i\cdot}D^2(i,N(R)) \tag{32}$$

进一步通过简单计算可以基于列联表卡方统计量得出如下等式:

$$Q=\sum_{i=1}^{n}\sum_{j=1}^{p}z_{ij}^2=\frac{\chi^2}{T} \tag{33}$$

其中的卡方统计量可以用于检验行点与列点间是否互不相关。

3.4 求解因子载荷矩阵

首先求解 R 型因子分析与 Q 型因子分析的因子载荷矩阵。

设 Z 为标准化数据矩阵，可以如下求解矩阵的奇异值与标准化特征向量。记 $S_R=Z^TZ$，特征值为：$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m>0$, 对应的特征向量为：$v_1,v_2,\cdots,v_m$，令 $d_j=\sqrt{\lambda_j}$; 记 $S_R=ZZ^T$，特征值为：$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m>0$, 对应的特征向量为：$u_1,u_2,\cdots,u_m$，令 $d_j=\sqrt{\lambda_j}$, 再设 $D_r=diag(P_{1\cdot},P_{2\cdot},\cdots,P_{n\cdot})$，$D_c=diag(P_{\cdot 1},P_{\cdot 2},\cdots,P_{\cdot p})$。

R 型因子载荷矩阵有如下两种表述方法：

$$F=(d_1a_1,d_2a_2,\cdots,d_ma_m) \tag{34}$$

$$F_{ij} = \frac{d_j v_{ij}}{\sqrt{P_{\bullet i}}} \tag{35}$$

Q 型因子载荷矩阵有如下两种表述方法：

$$G = (d_1 b_1, d_2 b_2, \cdots, d_m b_m) \tag{36}$$

$$G_{ij} = \frac{d_j u_{ij}}{\sqrt{P_{i\bullet}}} \tag{37}$$

3.5 行列轮廓坐标的确定

基于上一小节因子载荷矩阵的求解，可以得到 F 与 G 的前两列含有充足的信息，可以提取 R 型因子矩阵的前两列作为变量点（列点）的坐标，提取 Q 型因子矩阵的前两列作为样品点（行点）的坐标，以此建立平面直角坐标系上的 n+p 个点坐标。

值得注明的是：二维图中各行点之间的欧氏距离与行轮廓之间的加权距离对应，但是行轮廓与列轮廓对应的点没有如坐标图中的直接的距离关系。

3.6 对应分析的 R 软件实现

类似主成分分析与因子分析，R 软件同样提供了对应分析可直接调用的函数 corresp，如下为实现代码，供有兴趣的读者复现：

```
Corr.result = corresp(data,nf=2)
row.point = Corr.result$rscore
col.point = Corr.result$cscore
lambda = Corr.result$cor
```

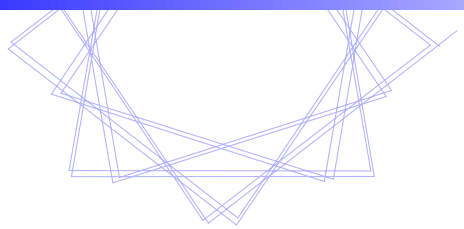
第 4 节 参考文献

[1] 朱聃, 王一凡. 基于主成分分析的机组威胁与差错管理评价模型 [J]. 民航学报,2020,4(05):74-79.

[2] 朱光婷, 潘晓琳. 基于因子分析和 SVM 的网络舆情危机预警研究 [J]. 重庆工商大学学报 (自然科学版),2020,37(05):94-100.

[3] 李阳超, 雍歧东, 顾宏波, 刘洲. 基于对应分析的军队油库安全隐患聚类挖掘研究 [J]. 军事交通学院学报,2020,22(08):88-94.

第 IV 部分



交叉视野

珞珈数学 Mathematics of Luojia March 2018

交叉视野：智能算法与机器学习

珞珈数学编辑部　时宇辰　刘宇佳

第 1 节　简介

随着数学与计算机、金融学、物理学、生物学等学科的交叉领域逐渐扩大，引入其他学科知识、利用计算机编程实现数值算法成为了研究的热点。

在 21 世纪的最优化领域与数据科学领域，传统的朴素方法已经对部分极端的问题无能为力。如较大的优化变量群与较强的约束条件组会大幅度减慢传统分支定界等最优化方法的运算效率与效果；过大或过小的数据体量使得传统的统计模型在数据分析中无法达到较好的拟合程度或出现模型不显著的现象。

第 2 节　智能算法与统计学习

传统的遗传算法是对达尔文进化理论的一次大胆迁移，一定程度上提升了多变量群优化的性能；但是其收敛慢、易早熟、单目标特点极为明显。因此科学家通过创建差分进化模板改进收敛的速度，使用不同的选择策略（精英保留、锦标赛、轮盘赌）解决早熟问题，构造不同的遗传、变异算子适应不同的问题背景。

蚁群算法也是常见的仿生算法，模拟蚂蚁爬行轨迹解决路径最优化问题。在路径规划问题中，蚁群算法要明显优于其他启发式算法。

与仿生学相对应地，科学家提出了模拟物理学过程的算法，其中模拟退火算法表现最为优良。此外，还有引力算法等其他物理过程的迁移，目的均为解决非线性函数最优化问题。

统计学习旨在解决传统统计模型对数据量的严苛要求与对变量关系的过度重视问题。当样本量低于 50 或多于 1000 时，设计对应的统计学习算法，可以有效的抽样（如马尔科夫蒙特卡洛方法）、分类（如决策树与支持向量机）、评价、预测。同时，对于传统统计学无法解决的问题，如：文本语义挖掘等领域，给出解决问题的基本方法。

近年来，深度学习知识的不断发展，给这些问题的解决提供了其他途径。如 BP 神经网络可以进行分类、卷积神经网络可以充分利用样本、循环神经网络可以进行预测，还有对抗神经网络等正在发展的方法。数据科学将不仅限与统计学的常见方法，未来可期。

第 3 节　学科展望

本文的篇幅有限，因此无法将所有上述涉及到的方法展现给读者。我们介绍方法的目的是，让读者掌握数学与计算机、经济等领域的交叉问题的常见解决思路，同时对遗传进化算法、分类问题的统计学习模型有自己的了解与认知。

交叉学科涉及到更多的编程技巧，读者可以在兴趣的基础上，掌握 Python、Matlab、R 等常用的软件，为算法的实现做好准备。

珞珈数学 Mathematics of Luojia March 2018

浅谈遗传进化算法在最优化问题中的应用

珞珈数学研习会　算法研究专栏

第一作者：2018 级　信息与计算科学　丁思哲
第二作者：2019 级　数学基地班　方礼喆
第三作者：2019 级　数学与应用数学　王　艺

遗传算法（Genetic Algorithm，GA）最早由 John holland 于 20 世纪 70 年代提出, 模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型，搜索最优解。该算法通过数学的方式, 利用计算机仿真运算, 将问题的求解过程转换成类似生物进化中的染色体基因的交叉、变异等过程。尤其在变量规模达到 100 个时求解性能优秀。

第 1 节　进化算法综述

遗传算法主要框架如下图 1，算法将可行解具象化为生物种群，通过模拟自然界中的进化实现最优解的寻找。

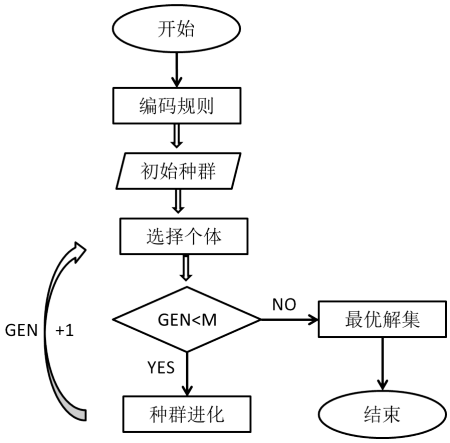


图　1: 进化算法基本框架

进化历程通常分为种群内部的基因变化与外部环境的自然选择，因此不同的进化算法设计的差异也在于基因变化规则的差异与外部自然选择法则的差异。下文将展开叙述不同的最优化求解算法，这些算法均基于基本的进化算法框架，算法首先将最优解的求解问题转化为种群进化问题；进一步，通过自然选择，求解最符合优化目标的种群；最后，给出最优解，将该种群还原回最优解集。主要的步骤如下：

1.1　Step1: 初始化

制定编码规则---将问题的解转换为生物种群

本步骤为进化算法的第一步，需要在最优化问题的解与生物种群间建立一一映射，将问题的解一一对应到生物种群内的个体基因片段。常见的方式为：将 N 维的解转换为 N 元数组，将每一个数位的实数直接对应到具体基因或将数位的整数通过二进制转换为基因片段。

编码规则主要会在两个过程应用。第一，问题的转换阶段：需要将求解最优解的问题转换为种群进化的问题，此时需要使用编码规则将初始解集转化为种群的个体；第二，给出最优解的阶段：得到进化完成的种群后，通过编码规则将种群一一映射到最优解集。

以上建立一一映射的过程称为：**编码**，建立的解集与种群间的一一映射称为：**编码规则**，最终由最优种群还原最优解集的过程为，**解码**。

参数初始化

最大进化次数 M 的指定：

现实中的种群进化是一个无穷尽的过程，但是，由于计算机运算能力的限制，需要指定一个很大的整数作为最大迭代进化次数，即为最大进化次数。

变异概率 p_v 与重组概率 p_c 的指定：

在现实的种群进化中，并不是所有的种群个体均会进行变异和基因重组，因此在模拟过程中需要指定变异和重组的概率，在每一次进化过程中仅有按变异概率 p_v 随机取出的部分个体需要进行变异操作，而每一对育种的个体仅有 p_c 概率会发生基因重组。

初始种群 G_0 的确定：

若问题为含约束条件的最优化问题，则在约束边界处随机取一组个体数为 n 的解，若为无约束问题则随机取定一组解；取定解后通过编码规则映射到种群个体，即可给出初始种群 G_0。

1.2 Step2: 种群进化

不同的进化算法选择进化育种父本的方式有所不同，在育种父代选择完毕后，通常通过基因重组与基因变异产生新的种群。

基因重组

基因重组操作通常将两个父本的部分基因片段进行交换，同时保证得到的新子代依然是可行解，通常根据组合方式与交叉位点数分为如下几种：recdis (离散重组)、recint (中间重组)、reclin (线性重组)、recndx (正态分布交叉)、recsbx (模拟二进制交叉)、xovbd (二项式分布交叉)、xovdp (两点交叉)、xovexp (指数交叉)、xovmp (多点交叉)、xovox (顺序交叉)、xovpmx (部分匹配交叉)、xovsec (洗牌指数交叉)、xovsh (洗牌交叉)、xovsp (单点交叉)、xovud (均匀分布交叉)。在后文中，将选取常见的基因重组方式进行详细解释。

基因变异

基因突变操作通常针对已经进行重组后的个体进行，在保证依然是可行解的条件下，改变个体中某个位点的基因或交换单一个体不同位点的基因，即为基因突变。通常使用的变异算子有如下几种：mutbga(遗传算法突变算子)、mutbin(二进制变异算子)、mutde(差分变异算子)、mutgau(高斯突变算子)、mutinv(染色体片段逆转变异算子)、mutmove(染色体片段移位变异算子)、mutpolyn(多项式变异)、mutpp(排列编码变异算子)、mutswap(染色体两点互换变异算子)、mutuni(均匀变异算子)。同样在后文中会对常见的基因突变方式进行解释。

1.3 Step3: 自然选择

自然选择通常也分为两个步骤，首先需要评估种群中的每一个个体对环境的适应能力，在单目标问题中通常给出适应度函数进行计算、在多目标问题中常转换为与目标的贴近程度计算；其次，需要根据个体对环境适应能力的不同淘汰、遴选个体进入下一代种群，即模拟自然界中“适者生存”法则。

适应度评估

适应度函数通常与个体对环境的适应能力 (解越接近最优目标则对应个体适应环境能力越强) 成正比。因此通常的构造方式即为与目标接近程度的倒数，如下给出数学的定义式：

单目标优化问题适应度计算：

设此时的种群个体 g 对应的解为 x，单值目标函数为 $A(x)$，最优目标为 $A(x_0)$，则可以如下定义个体适应度函数 $f(g)$：

$$f(g)=\frac{1}{|A(x)-A(x_0)|} \tag{1}$$

多目标优化问题适应度计算：

设此时种群个体 g 对应的解为 x，向量值目标函数为 $(A_1(x),A_2(x),\cdots,A_m(x)$，个体 g 对每一个目标 $A_i(x)$ 的适应度函数为 $f_i(g)$，则可以如下定义个体的适应度：

$$f(g)=\sum_{i=1}^{m}\lambda_{gi}f_i(g) \tag{2}$$

其中 λ_{gi} 为对个体 g 提出的多目标加权值，可以根据实际应用需要或者其他综合评价方法进行调整。

适者生存

根据计算得到的适应度函数值，适应度越高的个体越容易适应此时的环境，因此更容易被选择进入下一代种群；但是在现实中，适应能力差的个体也有一定几率存活，因此为增强种群多样性，需要引入选择过程的随机性。通常采取的方法有：精英选择法 (优先选择适应度较高的个体)、锦标赛选择法 (局部优先选择适应度高的个体)、轮盘赌方法 (适应度决定个体进入下一代种群的概率)。在后文中将对这三种方法进行详细的介绍。

1.4 Step4: 由最优解集给出最优解

在育种进行 M 代后，达到最大进化次数，此时得到的种群 G_M 即为最终的种群，通过编码规则即可一一转换为最优解集 X。

单目标最优解的确定：

对于单目标优化问题，直接根据适应度函数对最优解集中的个体进行排序，并选取适应度最高的个体即可作为最终的单目标最优解。

多目标最优解的确定：

对于多目标最优化问题，如果已知各个目标的重要程度可以给出加权，即可将多目标问题转换为单目标问题进行求解。但是在现实中，大多数情形只能给出最优解间的优劣关系判定，将解间的优劣关系定义为**支配关系**如下：

设 x_1 与 x_2 为以 $(A_1(x),A_2(x),\cdots,A_m(x)$ 为目标的最小化问题的两个最优解，若满足如下式，则称 x_1 支配 x_2：

$$A_i(x_1)\le A_i(x_2)\qquad \forall i=1,2,\cdots,m \tag{3}$$

若在解集 X 的一个子集 X_0 中任意两个解间不存在支配关系，则称子集 X_0 为一个非支配最优解集，同时也称其中元素构成了 Pareto 前沿平面。

根据定义可以明晰：Pareto 前沿平面中的解一定优于其他的解，因此最优解应在此子集中产生，而常见的一种方式即通过 Hurwicz 评判准则来根据选择者的意愿确定最优解，定义“乐观系数”α 并如下定义集中常见的策略：

在计算之前首先需要对每一个目标函数 $A_i(x)$ 在解集中的所有观测值 $A_i(x_j)$ 进行标准化，如下：

$$A_i'(x_j)=\frac{A_i(x_j)-\min\limits_{j=1,2,\cdots,n}A_i(x_j)}{\max\limits_{j=1,2,\cdots,n}A_i(x_j)-\min\limits_{j=1,2,\cdots,n}A_i(x_j)} \tag{4}$$

$\alpha=1$ **“乐观准则”：**

此时将每一个解所有目标函数值中最优的函数值提取出来，作为评估解优劣性的指标，确定解 x 的评价值为 $K(x)$ 如下，将 K 最小的个体选择为最优个体。此时的评价可能存在过于激进的问题：选择的最优解可能在某一个目标下极优，但是在其他目标下较差。

$$K(x)=\min_{i=1,2,\cdots,m}A_i'(x) \tag{5}$$

$\alpha=0$ **“悲观准则”：**

此时将每一个解所有目标函数值中最劣的函数值提取作为评估指标，将 K 最小的个体选择为最优个体。此时评价可能存在过于保守的问题：选择的最优解可能在所有目标下都不是最优，但是综合来看避免了在某一个目标下过差的情况。

$$K(x)=\max_{i=1,2,\cdots,m}A_i'(x) \tag{6}$$

$\alpha=0.5$ **“折衷准则”：**

此时将每一个解所有目标函数值中最劣的函数值与最优函数值加权求和提取作为评估指标，将 K 最小的个体选择为最优个体。此时可以通过调节 α 值向乐观准则或悲观准则进行偏移,α 越接近 1 则评价越乐观但是也越激进，α 越接近 0 则评价越保守，但是同时也越稳定，如果需要折衷评判可以设置为 0.5。

$$K(x)=\alpha\min_{i=1,2,\cdots,m}A_i'(x)+(1-\alpha)\max_{i=1,2,\cdots,m}A_i'(x) \tag{7}$$

第 2 节 按初始化分类：几种常见的编码方式

按照基因的不同类型通常需要给出不同的编码规则，对整数基因通常给出二进制编码方式与格雷编码方式，而对实数基因通常需要直接利用实数编码，同时还有其他的如排列编码的方式。本小节着重介绍四种常见的编码方式：二进制编码、实数编码、排列编码。

2.1 二进制编码

由于将整数展开为二进制表示后进行的重组变异更细节化，因此改变的意义更大，在基因为整数的条件下，可以将该整数转化为二进制，示例如下图：

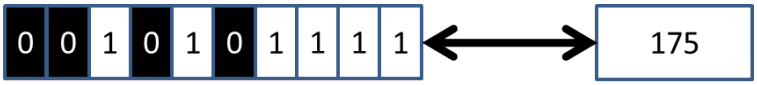


图 2: 二进制编码示意图

2.2 实数编码

对于实数基因，无法将该数表示为二进制数，因此直接使用各个数位作为编码片段或直接将此数作为编码，如下图：

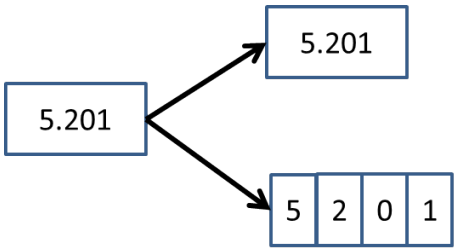


图 3: 实数编码示意图

2.3 排列编码

对于某些问题，涉及到的是元素间的排列，此时使用排列编码方式更为合适，如在 TSP 旅行商问题中，形成点 A,B,C,D,E 间的路线更适合用它此种方式编码，示意图 4如下，但是若每一个基因仅有 m 种取值最后的编码结果高达 $m!$，因此排列编码仅对部分小范围的基因有效。

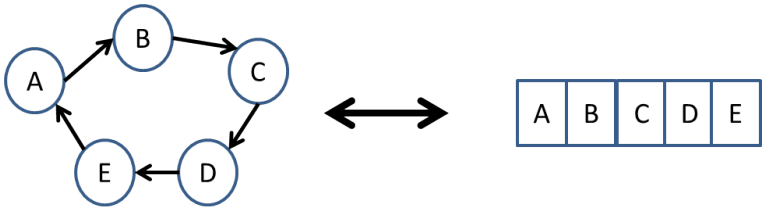


图 4: 排列编码示意图

第 3 节 按基因重组分类：几种常见的基因重组方法

基因重组是种群进化的第一个阶段，两个父代的个体通过交换基因片段得到两个子代个体，这里根据基因为整数与实数的不同编码方式给出最常见的六种重组方法：针对整数编码的：单点交叉、多点交叉、均匀交叉；针对实数编码的：离散重组、中间重组、线性重组。

3.1 针对整数二进制编码的基因重组

xovsp 单点交叉

单点交叉的基因重组仅随机指定一个交叉点，交换交叉点后所有的基因片段，如下图。易见，这种交叉重组方式是一种不稳定的育种方式，有利于提升种群的多样性但是相对不稳定。

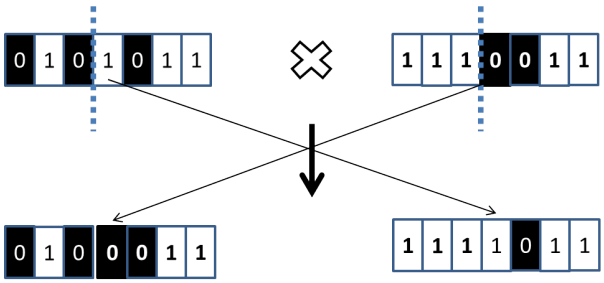


图 5: 单点交叉重组示意图

xovmp 多点交叉

多点交叉在单点交叉的基础上，增加交叉点（交叉点的选取依然是随机的），仅交换交叉点间的基因片段，有助于提高育种后代的性状稳定性，因此其特殊形式两点交叉为遗传算法的默认基因重组方式，多点交叉的示意如下：

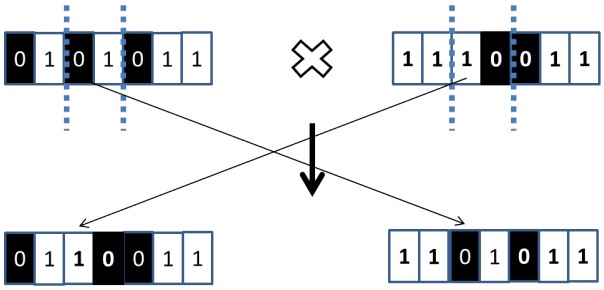


图 6: 多点交叉重组示意图

xovud 均匀分布交叉

若使用单点交叉依然无法收敛，则需要更灵活的交叉方式，此时可以使用均匀分布交叉重组，这种重组的子代每一个基因均随机地由两个父本对应位置基因二选一得到，随机性极强种群多样性最高，均匀分布交叉重组的示意如下：

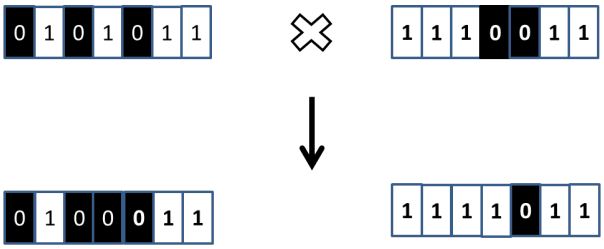


图 7: 均匀交叉重组示意图

3.2 针对实数编码的基因重组

recdis 离散重组

上文叙述的均匀分布交叉即为离散重组的一种方式，离散重组不仅限制各个位点基因全为整数，此时可以使用实数作为基因，按与 xovud 相同的方式进行基因重组即可。

reclin 线性重组

实数的线性重组充分利用实数的连续性特点，将两个父代个体代入线性公式计算得到子代个体，并且可以通过改变线性系数得到不同的子代，设父代个体为向量 p_1, p_2，则给定线性系数 a 即可依下式唯一确定子代个体 g。可以给出一个实例如下图：

$$g = p_1 + a(p_2 - p_1) \tag{8}$$

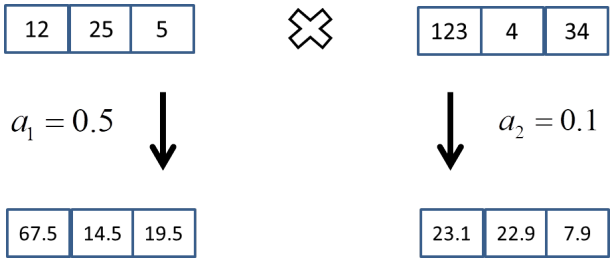


图 8: 线性重组示意图

recint 中间重组

线性重组对父本信息的利用率较高，但权重需要指定，因此随机性不足会导致种群多样性降低。若将线性系数 a 变为从区间 $[0,1]$ 内直接生成，即为中间重组的基因重组方式，此种方式对种群多样性有显著的提升。

第 4 节 按基因突变分类：几种常见的基因突变类型

由于基因重组只是在原有解元素中进行顺序调整，并不能产生新的解元素，因此需要进一步引入基因突变。但是由于基因突变的随机性过强，极易造成种群混乱或基因重组效果不明显等问题，因此这里分别针对实数编码与二进制编码，最常用的仅有如下两种基因突变方式。

4.1 实数编码 mutbga 突变算子

设待突变变量 X 的取值范围为 L，则如下可以给出突变后的变量 X' 的表达式。其中 p_0 以一半概率取 0.5、一半概率取-0.5，m 常取 20，$a(i)$ 以概率 $\frac{1}{m}$ 概率取 1，以 $1-\frac{1}{m}$ 概率取 0：

$$X' = X + p_0 L\Delta = X + p_0 L(\sum_{i=0}^{m} \frac{a(i)}{2^i}) \tag{9}$$

4.2 二进制编码 mutswap 两点互换算子

若变量以二进制编码确定，则通常只需要使用两点互换算子对整数位置进行操作。在基因片段中生成两个正整数 n_1, n_2，直接交换两个数位上的整数即可得到最终的突变子代，如下图：

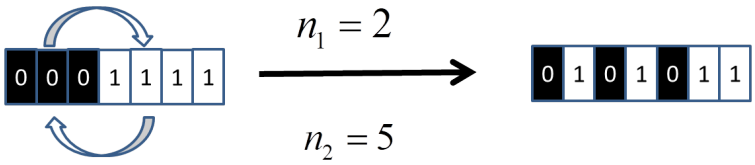


图 9: 互换突变算子示意图

第 5 节 按适者生存方式分类：几种常见的自然选择方式

在相同的基因重组与突变的方式下，自然选择方式为种群性质与算法结果最主要的决定步骤。若全选择适应度高的个体会导致种群缺乏多样性，从而导致早熟、过快收敛，更可能落入局部最优解；若选择过多适应度低的个体，算法不易收敛，最终可能无法搜索得到最优解。这里主要介绍三种选择方式：精英选择法、锦标赛选择法、轮盘赌选择法。

5.1 精英选择法

假设原种群内个体数为 n，其中有 $n-k$ 个体保持不变，此时原父代中有 k 个体参与了育种，而子代也对应有 k 个体。将参与育种的父代个体 g 通过适应度排序，将适应度最低的个体淘汰掉，并在子代 f 中遴选出适应度最高的个体选入新的种群。**精英选择方法**如下图 10：

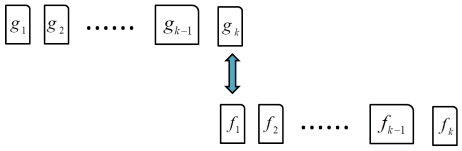


图 10: 精英选择法示意图

5.2 锦标赛选择法

精英选择法直接在所有育种对象中选择适应度较高的个体，这种方式收敛方式较快但是容易早熟落入局部最优解，因此需要缩小选择范围到参与育种的 k 个个体的一个仅含 j 个个体的子集。每一次选择仅淘汰这个子集中适应度最低的父代、选择适应度最高的子代产生下一代种群，称为**锦标赛选择法**，如下图 11：

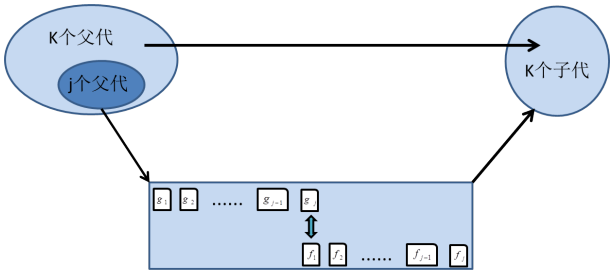


图 11: 锦标赛选择法示意图

5.3 轮盘赌选择法

易见，锦标赛选择法中适应度最低的 $j-1$ 个个体永远不会被选入新种群，因此在此种方法中依然可能出现局部最优解的情形，若想进一步减缓收敛速度，可以使用**轮盘赌选择法**。假设原种群参与育种的 k 个个体经过遗传变异后与父代合并得到的 $2k$ 个个体适应度为 $\{g_1, g_2, \cdots, g_k, f_1, \cdots, f_k\}$，则这 $2k$ 个个体的选择概率可以如下计算：

$$\begin{cases} p_i = \dfrac{g_i}{\sum\limits_{i=1}^{k}(f_i+g_i)} & \forall i = 1, 2, \cdots, k \\ p_{i+k} = \dfrac{f_i}{\sum\limits_{i=1}^{k}(f_i+g_i)} & \forall i = 1, 2, \cdots, k \end{cases} \tag{10}$$

进一步根据上述概率进行轮盘赌，共进行 k 次不重复选择，与未育种的个体集共同构成新的子代种群，示意图如下图：

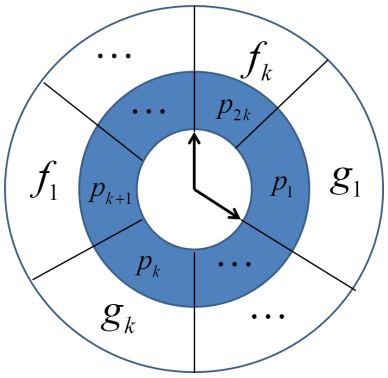


图 12: 轮盘赌选择法示意图

第 6 节 常见算法介绍

基于上述对编码、重组变异算子的定义与选择方式的定义，本部分将展开介绍单目标、多目标的进化算法。进化算法的差异主要体现为：参与育种的父代数量、重组变异前后是否有其他方式产生新个体、多目标非支配个体间的适应度排序方式。如下分别介绍可以最稳定且快速收敛到全局最优解的三个单目标进化算法与三个多目标进化算法。

6.1 单目标进化算法

所有的单目标算法共性在于个体排序依适应度进行，选择过程中优先选择适应度高的个体，其中最基本的算法是基础遗传算法 (GA)、差分进化算法在此基础上引入差分进化过程从而提高收敛速度、稳定遗传算法通过减少育种个体数更有可能收敛至全局最优。

基础遗传算法

基础遗传算法的进化过程可以由如下示意图给出，其中选择个体中使用适应度计算排序与轮盘赌选择方法，重组使用两点交叉重组、变异使用 Breeder 突变方法。

值得注意的是，遗传算法在使用的过程中并不是每一个父代均会发生重组或突变。在选择两个父代育种时仅有 p_c 的概率会发生基因重组，此时称 p_c 为交叉概率；同时仅有 p_v 概率发生基因突变，此时该概率成为突变概率。

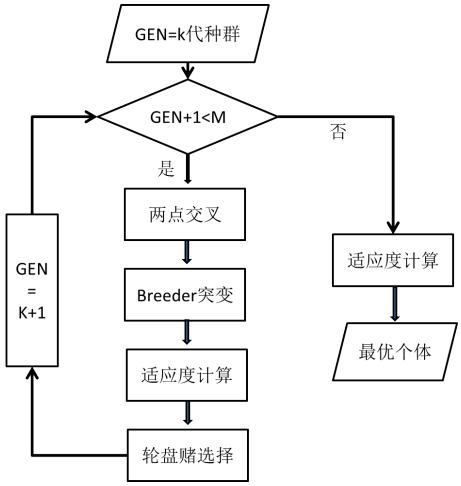


图 13: 基础遗传算法示意图

稳定遗传算法

若在某些问题的求解中基础遗传算法出现早熟现象或落入局部最优解，可以使用稳定遗传算法进行改进。

在 GA 算法中，所有的个体均参与了育种过程，但是在稳定遗传算法中仅使用其中的两个父代进行育种，其他个体均保持不变。易见，这种育种方式具有极强的稳定性，但是收敛速度也会对应地降低。

差分进化算法

由于算法需要提速，同时避免早熟现象，因此差分进化算法在设计中仅对进化过程中的变异与重组进行调整，设前代种群中有 m 个个体，均可以表示为 n 维向量，表示为：$X_i(k) = (X_{i,1}(k), X_{i,2}(k), \cdots, X_{i,n}(k))$。

变异操作：

首先从上一代种群中随机选择 3 个个体 $X_{p_1}(k), X_{p_2}(k), X_{p_3}(k)$，在区间 $[0, 2]$ 间随机取系数 F，按照如下公式生成新的个体：

$$H_i(k) = X_{p_1}(k) + F(X_{p_2}(k) - X_{p_3}(k)) \tag{11}$$

交叉重组操作：

首先指定重组概率为 p_v，则对个体 $X_p(k)$ 的每一位元素，有 p_v 概率替换为 $H_p(k)$ 对应数位上的基因。此过程完成基因重组。重组后根据如下公式对 $X_p(k)$ 进行每一个基因位点的更新，其中 f_i 为第 i 个目标函数，目标为最大化：

$$X_{p,i}(k+1) = arg\max(f(X_{p,i}(k)), f(H_{p,i}(k))) \tag{12}$$

差分进化算法对 GA 算法的改进主要体现在上述的基因重组与变异操作中，其余步骤均与 GA 算法相同。

6.2 多目标进化算法

多目标进化算法在选择过程中无法直接按照适应度进行，需要引入多目标意义下的非支配关系，对种群进行非支配层级划分，而在相同层级间需要进一步定义距离关系，定义方式的不同决定了不同的算法性能。

在 NSGAII 算法的自然选择过程中，首先根据非支配关系对全种群分层，算法的流程如下图，操作步骤如下：

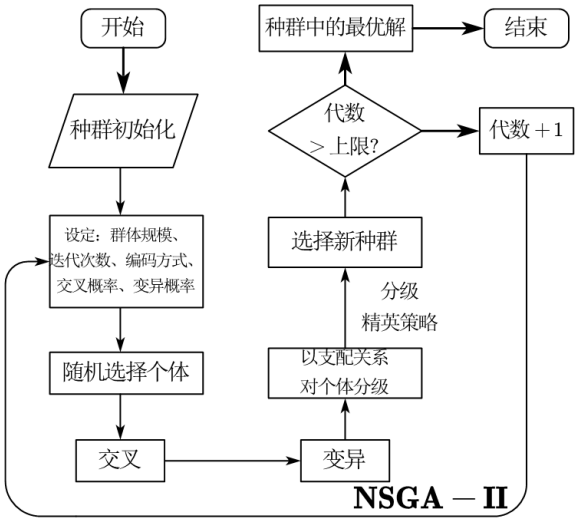


图 14: NSGAII 算法示意图

非支配分层

Step1: 记非支配层数 $l=1$，计算种群中所有个体 g 支配的个体数 n_g。

Step2: 将所有 $n_g=0$ 的个体加入第 l 层非支配集，令所有的 $n_g=n_g-1$，令 $l=l+1$，重复本步骤直至所有种群中的个体均分入某非支配层级。

增强精英选择方法

在**非支配分层**结束后，按照增强精英选择方法，由层数较低的非支配层起选入容量为 n 的新种群至第 K 层为止，若记第 l 层的个体数为 n_l，则 K 应满足如下式：

$$\sum_{l=1}^{K} n_l \le n < \sum_{l=1}^{K+1} n_l \tag{13}$$

此时若新种群中仍有空位，需在第 $K+1$ 层中定义拥挤距离，根据拥挤距离进行个体的选择。定义个体 i 在目标函数 $f_j(x)$ 下的拥挤距离：将所有个体在目标函数 f_j 下的计算值依升序排列，并设函数值最小的个体与在此目标函数下的拥挤距离为 $+\infty$、函数值最大的个体与在此目标函数下的拥挤距离为 $-\infty$，并设最大与最小的函数值分别为：f_{jmax} 与 f_{jmin}；设个体 i 出现在第 t 个排序位置，并记第 $t-1$ 与 $t+1$ 个排序位置的个体函数值为 $f_j^{(t-1)}$ 与 $f_j^{(t+1)}$，可以根据下式计算个体 i 在目标函数 $f_j(x)$ 下的拥挤距离 d_{ij}：

$$d_{ij}=\frac{f_j^{(t+1)}-f_j^{(t-1)}}{f_{jmax}-f_{jmin}} \tag{14}$$

个体 i 的拥挤距离 d_i 即为目标空间中与 i 相邻两个个体 $i+1$ 与 $i-1$ 的距离，以个体 i 在各个目标函数下的拥挤距离之和计算，如下式：

$$d_i=\sum_{j=1}^{k} d_{ij} \tag{15}$$

在完成拥挤距离定义后，从临界的第 $K+1$ 层中根据拥挤距离从高到低依次取出个体填入新种群，直至新种群生成完毕。该选择过程如下示意图 15：

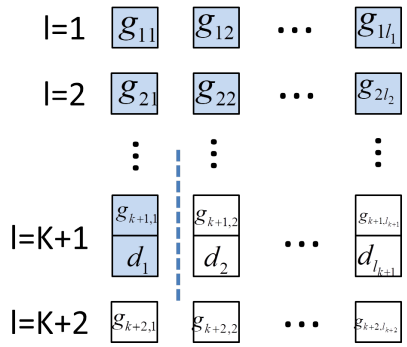


图 15: 拥挤距离选择方法示意图

若想进一步提升 NSGAII 算法的收敛速度，可以使用差分进化算法中的重组与变异算子代替原算法中的遗传算子，此时算法改进为 DE-NSGAII 算法。

NSGAIII 算法及差分形式

容易观察到，上述算法使用了增强精英选择策略配合拥挤度的计算，两种方式都是优中选优原则，因此种群易早熟从而落入局部最优解。为改善上述缺点，NSGAIII 算法取消了拥挤度的计算，转而使用**理想点法**排序选择下一代个体。理想点法主要包含四个步骤：参考点的产生、目标空间标准化、关联操作、环境选择。

Step1: 参考点的产生

在 m 维目标空间中，一个维度为 $m-1$ 的标准单纯形，它对所有的目标轴都有相同的倾斜度。如果考虑沿着每个目标方向分为 p 份，参考点 H 的总数为 C_{m+p-1}^{p}，且参考点在单纯形上均匀分布。下图为一个参考点示例：

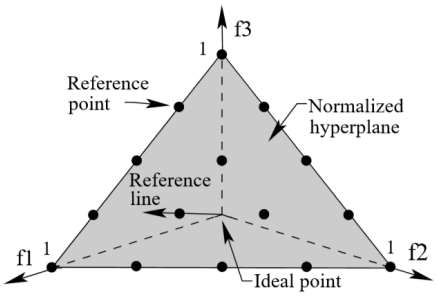


图 16: NSGAIII 算法参考点示例

Step2: 标准化目标空间

由于各目标空间存在不同的量纲，因此若做统一评估需要去单位化。设解 x 在第 i 个目标下的目标函数值为 $f_i(x)$，第 i 个目标的样本最大值为 z_i^{max}，最小值为 z_i^{min}，可以按如下式计算标准化后的目标函数值：

$$f_i'(x)=\frac{f_i(x)-z_i^{min}}{z_i^{max}-z_i^{min}} \tag{16}$$

上述操作将每一个目标的极小值点变换到了原点，此时 $m-1$ 维超平面对应发生变化，上述定义的参考点随超平面发生位置变化。

Step3: 关联操作与环境选择

首先作参考点与原点的连线，进一步，针对每一个点，计算点与所有参考线间的最短距离，并将该点关联到距离最小的参考点，此步骤完成关联。示意图如下图 17：

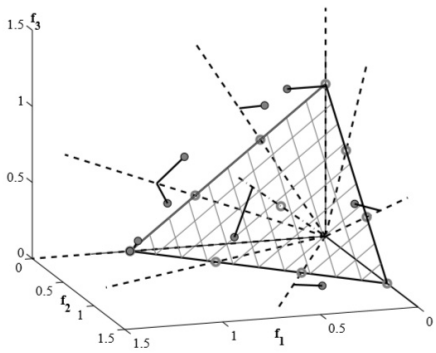


图 17: NSGAIII 算法关联操作示例

此时首先按照非支配关系得到的分层选择层数较低的个体进入下一代种群，在临界层中，按照关联规则选取：首先考虑有关联个体且关联数最少的参考点，按距离从近到远逐个选入下一代种群。若种群数量仍不足，移除该参考点后重复上述选择操作，直到种群选择完毕。示意图如下：

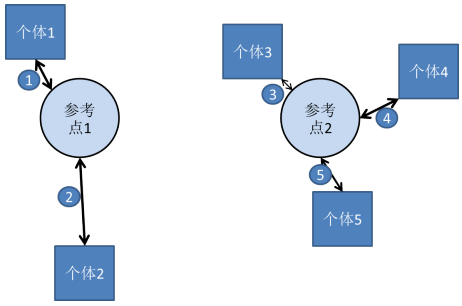


图 18: NSGAIII 算法理想点选择

易见，选择过程总体按精英选择策略进行选择，因此收敛速度有保证，局部引入了理想点代替拥挤度，允许种群的多样性，因此可以克服早熟与局部收敛。
同样地，若 NSGAIII 算法的收敛速度过慢，可以使用差分进化算法中的重组与变异算子代替原算法中的遗传算子，此时算法改进为 DE-NSGAIII 算法。

MOEA/D 算法

NSGA 族多目标算法经常需要结合 Hurwicz 准则共同使用，且运行速度略显不足，因此在定义域更广泛、局部最优点更多的多目标最优化问题中，研究出基于问题分解的 MOEA/D 算法。

假设原问题可以表述为定义域为 Ω 的 m 个目标函数 $(f_1(x), f_2(x), \cdots, f_m(x))$ 最大化问题，并设容量为 n 的种群集中可以取到的最大值分别为 $z_1, z_2, \cdots, z_m$，因此原问题可以转化为如下最小化问题：

$$g^{te}(x, \lambda) = \max_{1 \le i \le m} \{\lambda_i | f_i(x) - z_i|\} \tag{17}$$

Step1: 初始化

给出 n 个权重向量 $\lambda^1, \lambda^2, \cdots, \lambda^n$ 一一对应到所有的个体，计算与每一个权重向量欧式距离最近的 T 个权重，将序号索引即为 $B(i)$。初始化非支配解集 EP 为空集。

Step2: 种群进化

对每一个种群中的个体进行进化操作，在第 i 个个体的邻域 $B(i)$ 中随机取 j, k 两个个体，利用基础遗传算法中的交叉算子与变异算子进行操作，将得到的新解 y 作为进化结果。并对比每一个目标函数处的函数值，若第 j 个目标满足 $z_j < f(y)$ 则设置 $z_j = f(y)$。

Step3: 在邻域内更新种群

从原有的 EP 中移除被 y 支配的解，直至无被 y 支配的解后将 y 加入 EP。若 y 在 x_i 的邻域 $B(i)$ 内支配所有的解，则将 x_i 替换为 y。

以上为 MOEA/D 算法的主要步骤，算法流程中借助了基础遗传算法的交叉变异算子，最终得到的 EP 解集即为 Pareto 前沿平面。算法流程图如下图：

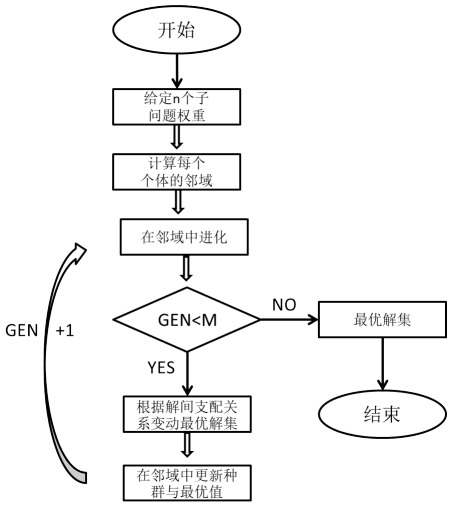


图 19: MOEA/D 算法流程示意图

第 7 节 参考文献

[1] Deb K , Jain H . An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints[J]. Evolutionary Computation, IEEE Transactions on, 2014, 18(4):577-601.

珞珈数学 Mathematics of Luojia March 2018

分类问题：从传统模型到统计学习

珞珈数学研习会　算法研究专栏

第一作者：2018 级　统计学　　袁继超
第二作者：2019 级　数学基地班　　尹　鹏
第三作者：2019 级　数学与应用数学　杨夜姿

分类问题是生活中常见的一种数据统计问题，可能是样本的“物以类聚”，也可能是指标的“近似合并”，可能已经有了先验的知识、也可能一无所知地根据经验分类。通常按照样本量的大小将分类问题分为：大数据分类问题、小数据量分类问题；根据是否已有分类结果作监督分为：有监督分类与无监督分类；根据使用的方法分为：使用传统模型的聚类与利用机器学习方法的分类、根据目标类个数分为：二分类、多分类。

在本节中，我们将分别探讨：有监督分类与无监督聚类、小样本分类与大数据分类，对应不同的实际问题背景给出最合适的模型与算法。

第 1 节　传统二分类模型：Logistic 回归

logistic 回归又称 logistic 回归分析，是一种广义的线性回归分析模型，常用于数据挖掘，疾病自动诊断，经济预测等领域。由于估计过程利用最小二乘法，传统回归模型对稀疏变量的回归性能很差，因此在变量取值很稀疏（特殊地，只取为 0-1 变量）时，需要对回归模型进行改进以获得更好的拟合性能。改进后的模型将成为概率判别模型，可以利用为二分类或多分类的判别方法。

Logistic 回归是为解决二分类问题提出的统计模型，与其他统计模型一样，在处理中等数据量 (多于 50 且少于 1000) 的样本分类问题中表现良好。在本节中，介绍二分类判别 Logistic 回归模型。

1.1　Logistic 回归模型构建

假设原标签数据为 (x_n, t_n) 经过特征提取后的变量为 $z_n = \phi(x_n)$，现将变量分入 C_1, C_2 两个类，在介绍分类模型前首先引入一个重要的函数，sigmoid 函数：

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1}$$

此函数有如下两条重要的性质，由于证明很简单在这里直接给出结论：

$$\begin{cases} \sigma(-x) = 1 - \sigma(x) \\ \sigma'(x) = \sigma(x)(1-\sigma(x)) \end{cases} \tag{2}$$

因此在假设响应变量 y 服从二项分布的基础上，利用上述 sigmoid 函数可以给出回归判别模型：

$$\begin{cases} P(z \in C_1) = y = \sigma(w^T z) \\ P(z \in C_2) = 1 - P(z \in C_1) \end{cases} \tag{3}$$

1.2 参数估计：极大似然估计

由于上述模型的非线性，因此估计参数不能用最小二乘法进行估计，转而采取极大似然方法进行参数估计，如下进行似然函数与损失函数的构造：

似然函数：

$$p(t_1, \cdots, t_N|w) = \prod_{n=1}^{N} p(t_n|w) = \prod_{n=1}^{N} y_n^{t_n}(1 - y_n)^(1 - t_n) \tag{4}$$

损失函数（交叉熵):

$$E_D(w) = -\log(p(t|w)) = -\sum_{n=1}^{N}[t_n \log(\sigma(w^T z_n)) + (1 - t_n)\log(1 - \sigma(w^T z_n))] \tag{5}$$

在迭代中，使用 Newton-Raphson 迭代法，以 η 为步长，构造如下迭代式：

$$w^{(new)} = w^{(old)} - \eta \nabla E_D(w) \tag{6}$$

或使用二阶的 Newton-Raphson 迭代，H 为 $E_D(w)$ 的 Hessian 矩阵，如下：

$$w^{(new)} = w^{(old)} - H^{-1} \nabla E_D(w) \tag{7}$$

在本问题中，使用二阶方法进行推导，设 $y = (y_1, \cdots, y_N)^T, t = (t_1, \cdots, t_N)^T, Z = (z_1, \cdots, z_N)^T$，则可以得如下计算式：

$$\begin{cases} \nabla_w E_D(w) = Z^T(\sigma(Zw) - t) = Z^T(y - t) \\ H = Z^T R Z \end{cases} \tag{8}$$

由数值分析知识，这个迭代式一个收敛的迭代过程。

1.3 假设检验：回归系数检验与过度离势检验

首先进行回归系数显著性检验，在 Logistic 回归中成为 Wald 检验，零假设为：某个回归系数为 0，备择假设为该回归系数项不为 0，可以如下构造检验统计量：

$$\chi^2 = \frac{\beta}{\sqrt{\beta_1 + \beta_2 + \cdots + \beta_N}} \sim \chi^2(dim) \tag{9}$$

其中 dim 为卡方自由度, 数值上等于回归系数 w 的维数。进行单边卡方检验，在给定显著性的情形下判断是否拒绝回归系数为 0 的零假设。

由于本模型要求响应变量服从二项分布，因此需要对其进行检验，而二项分布数据方差为 $\sigma^2 = n\pi(1 - \pi)$，其中 π 为属于 1 值组的概率，如果响应变量的方差大于上述标准过多，则认为可以由数据显

著地拒绝零假设。在 Logistic 回归中通常以残差平方和与残差自由度 df 之比进行过度离势检验：零假设为：$\Phi = 1$ 即完全不存在过度离势现象。

$$\Phi = \frac{\epsilon_1 + \epsilon_2 + \cdots + \epsilon_N}{df} \tag{10}$$

如果计算值在较大程度上远离 1，则原序列存在严重的过度离势，此时在拟合模型时需要考虑假设响应变量为类二项分布。

上述检验只是进行粗略的检验，在假设为类二项分布后可以用卡方检验进行精确的过度离势检验，由于 R 软件中提供了对应的检验函数，因此这里不进行展开叙述。

模型诊断

在 Logistic 回归中，由于线性规律被打破，因此无法直接使用拟合优度 R-square 值，这里介绍三种可视化的检验方式：hatvalue、学生化残差值、Cook 距离，其中学生化残差值计算公式如下：

$$\hat{\epsilon_i} = \frac{\epsilon_i - E(\epsilon)}{\sqrt{Var(\epsilon)}} \tag{11}$$

设样本矩阵为 Z，则 Cook 距离的计算如下两式：

$$H = Z(Z^T Z)^{-1} Z^T \tag{12}$$

设 h_i 为 H 矩阵对角线上的第 i 个元素，则第 i 个样本的 Cook 距离为：

$$D_i = \frac{\epsilon_i^2 h_i}{MSE^2 p(1-h_i)^2} \tag{13}$$

第 2 节 拓展的多分类模型：Softmax 回归

在上述二分类问题基础上可以进行适当推广以得到多分类的回归方法，这里同样首先引入 K 维空间间的同构映射函数 Softmax 函数如下：

$$softmax(x_k) = \frac{exp(x_k)}{\sum_{j=1}^{K} exp(x_j)} \tag{14}$$

因此这里选取模型为：

$$P(z_j \in C_k) = Softmax(z_j)_k \tag{15}$$

可以进一步构造似然函数：

$$P(t_1, \cdots, t_N | w_1, \cdots, w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} Softmax(y_n)_k^{(t_n)_k} \tag{16}$$

得到多分类问题的交叉熵函数：

$$E_D(W) = -\sum_{n=1}^{N} \sum_{k=1}^{K} (t_n)_k \log(Softmax(W z_n))_k \tag{17}$$

进一步对 W 的第 j 列可以求梯度, 设所有 N 个样本标签的第 j 个变量组成的向量为 t_j，有：

$$\nabla_{w_j} E_D(w_1, \cdots, w_K) = \sum_{n=1}^{N} (y_{nj} - t_{nj}) z_n = Z(y_j - t_j) \tag{18}$$

可以进一步利用一阶梯度下降可以得到：

$$w_j^{(new)} = w_j^{(old)} - \eta \nabla_{w_j} E_D(w_1, \cdots, w_K) \tag{19}$$

如上即分别求得 K 个回归系数，可以得到梯度下降的系数结果，作为对 Logistic 回归的多分类推广，Softmax 回归同样可以作为广义线性模型的组成。

在模型的选取上，有一个很自然的问题，多分类问题模型过程中选择 Softmax 回归通常难以与 K 个 Logistic 回归区别开，如下提供一个简单的思路。

举一个计算视觉领域的例子，你的任务是将图像分到三个不同类别中。假设这三个类别分别是：室内场景、户外城区场景、户外荒野场景，三个类别是互斥的，更适于选择 softmax 回归分类器。当假设这三个类别分别：是室内场景、黑白图片、包含人物的图片，建立三个独立的 logistic 回归分类器更加合适。

作为 Logistic 回归的推广形式，Softmax 回归同样也在中等数据量的情形下表现良好。

第 3 节 有监督的大数据二分类模型：支持向量机

当样本量达到近万或十万量级时，传统的统计模型已经无法给出精确的分类判别，而在当今时代，这样的大样本数据量随处可见。因此，科学家通过研究，发展起了统计学习方法，用机器学习的思想处理面临的大数据量问题。

支持向量机是 Cortes 和 Vapnik 于 1995 年首先提出的一种基于统计学习的二类分类模型。它是一种监督学习方法，在学习过程中通过最大化分类间隔使得结构风险最小化。

在进行统计学习模型的使用前，首先注意应当合理构造特征。如果将原始数据直接使用，可能导致机器无法进行效果较好的分类操作。常见的特征工程方法有：比例变换、分箱变换，由于本节重点是介绍分类模型，因此将此类特征工程知识略过。

记已经构造的特征指标组为 X，以每一个任务点作为特征空间的样本点，此时的样本点在特征空间中线性不可分，需首先通过核函数映射到线性可分的空间中，进一步构造线性分类平面，完成分类。如下图为支持向量机模型示意图：

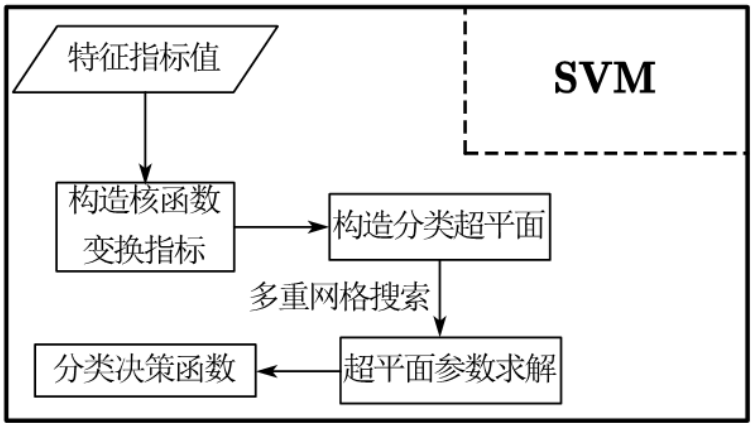


图 1: 支持向量机模型示意图

3.1 核函数的构造：将原样本空间映射到线性可分的样本空间

设原特征空间中向量 x_1, x_2 的内积为 $< x_1, x_2 >$，计算方式为 $x_1^T x_2$，构造如下高斯核函数对内积运算规则进行映射，可以使得线性不可分的样本点在新的特征空间中线性可分，其中 σ 为核函数中的参数：

$$K(x_1, x_2) = e^{-\frac{||x_1 - x_2||^2}{2\sigma^2}} \tag{20}$$

3.2 线性分类平面的构造：

平面一般方程的建立：

设任务点指标构成的数据集为：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$，其中 $y_i \in \{-1, 1\}$，取-1 时表示定价不合理，取 1 表示定价合理。在通过核函数构造的内积空间中，设样本点可以被如下超平面分为两类：

$$y =< \omega, x > +b \tag{21}$$

定义样本点集到超平面的距离：

对任意样本点 (x_i, y_i) 可以通过如下式计算该点到上述超平面的欧氏距离 γ_i：

$$\gamma_i = y_i(\frac{\omega}{||\omega||}x_i + \frac{b}{||\omega||}) \tag{22}$$

基于各样本点到超平面的欧氏距离，可以给出点集到超平面距离 γ 的定义如下式：

$$\gamma = \min_{i=1,2,\cdots,N} \gamma_i \tag{23}$$

超平面参数的求解：

根据模型的定义，可以通过最大化点集与平面距离确定超平面参数，因此参数求解问题转换为如下最优化问题：

$$\max_{\omega,b} \gamma \tag{24}$$

$$s.t. \quad y_i(\frac{\omega}{||\omega||}x_i + \frac{b}{||\omega||}) \geq \gamma \quad i = 1, 2, \cdots, N \tag{25}$$

将约束条件两边同时除以 γ，并如下式重新定义平面系数，上述约束条件转换为：

$$\begin{cases} \omega_0 = \frac{\omega}{||\omega||\gamma} \quad b_0 = \frac{b}{||\omega||} \\ y_i(< w_0, x_i > +b_0) \geq 1 \quad i = 1, 2, \cdots, N \end{cases} \tag{26}$$

此时，目标函数 γ 最大化等价为 $||\omega_0||$ 最小化，因此上述最优化模型可以化简为如下式：

$$\min_{\omega,b} \frac{1}{2} ||\omega||^2 \tag{27}$$

$$s.t. \quad y_i(< \omega, x_i > +b) \geq 1 \quad i = 1, 2, \cdots, N \tag{28}$$

此问题为含不等式约束的凸二次优化问题，因此可以使用拉格朗日数乘法转换为如下无约束最优化

问题，即求函数 $L(\omega,b,\alpha)$ 在任意 α 下的最小值：

$$\min_{\omega,b}\max_{\alpha}L(\omega,b,\alpha)=\min_{\omega,b}\max_{\alpha}\frac{1}{2}\left|\left|\omega\right|\right|^2-\sum_{i=1}^{N}\alpha_i(y_i(<\omega,x_i>+b)-1) \tag{29}$$

上述无约束最优化问题中，拉格朗日函数的最大化与最小化顺序可以变化，因此先求函数 $L(\omega,b,\alpha)$ 对 ω,b 的最小值，即对拉格朗日函数求偏导数：

$$\begin{cases}\frac{\partial L}{\partial\omega}=\omega-\sum\limits_{i=1}^{N}\alpha_i(y_i(<\omega,x_i>+b)-1)\\ \frac{\partial L}{\partial b}=-\sum\limits_{i=1}^{N}\alpha_i y_i\end{cases} \tag{30}$$

令上述偏导数等于 0，代入原拉格朗日函数进行化简整理可以得到如下求解结果：

$$L(\omega,b,\alpha)=\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j<x_i,x_j> \tag{31}$$

利用核函数构造的内积代替原式中的内积，同时确定约束条件，原最优化问题进一步转换为如下约束优化问题：

$$\begin{aligned}&\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i,x_j)-\sum_{i=1}^{N}\alpha_i\\ &s.t.\sum_{i=1}^{N}\alpha_i y_i=0\quad\alpha_i\geq 0,i=1,2,\cdots,N\end{aligned} \tag{32}$$

代入样本点数值，再次利用拉格朗日数乘法，即可求解出上述优化问题的解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)$，代回到式 (30) 即可求解得到分类超平面的系数 ω^* 与 b^*。

分类决策函数的确定：

基于上述对分类超平面的确定，记 $sign(t)$ 为符号函数：在 $t>0$ 时取 1，在 $t<0$ 时取-1，在 $t=0$ 时取 0。可以直接定义如下的分类决策函数：

$$f(x)=sign(x^T\omega^*+b^*) \tag{33}$$

最后通过计算原样本点的分类决策函数值，将取值为 1 的点划分为同一个类别，将其他取值的点划分为另一个类别，此时得到的类别可以作为评估任务定价合理性的依据，完成支持向量机模型的构建。

第 4 节 有监督的小样本二分类模型：随机森林

统计学习算法不仅可以在构造大数据分类模型，在面对数据量不足的情形时，也可以通过构造机器学习模型有效避免传统统计模型易不显著的劣势。

CART 决策树是使用二叉树结构、根据样本在各个影响因子处的观测值完成样本二分类的高性能分类器，通常可以对样本完成精确的二分类。

利用样本进行模型训练的主要步骤为：输入观测值、构造决策树结构与分类阈值、剪枝防止过拟合、输出分类结果。

决策树模型通常在小样本量下表现良好，因此本文将介绍文献 [1] 中决策树模型的改进：随机森林模型解决小样本量下的分类问题。本文将依照随机森林模型的建立顺序，首先介绍 CART 决策树的构造

过程，进而由决策树的重复构造形成随机森林模型。

4.1 CART 决策树原理介绍

本问题决策树模型的输入项为：容量为 n 的可重复样本集，样本 X_j 有 m 个维度作为特征，第 j 个样本的观测值分别是 $X_{j1}, X_{j2}, \cdots, X_{jm}$。

CART 决策树模型核心步骤为：

(1) 通过基尼不纯度确定特征的优劣与分类的阈值。

(2) 通过逐次选取最优的特征构建决策树。

(3) 对测试集样本使用决策树分类给出分类结果。

特征优劣与分类阈值的确定

对第 i 个有序多分类型特征求解分类阈值：

假设在第 i 个特征的样本观测值中有 k 个水平等级，则此时需要在其中选择一个等级作为二分类边界。依次计算在 k 个等级划分时对应的基尼不纯度，根据系数大小关系确定分类阈值。

设在第 t 个等级划分二分类，将样本集 D 划分为两个集合 D_1, D_2，分别在两个集合中如下式计算各自的基尼不纯度：

$$Gini(D_i) = 1 - \sum_{k=1}^{K} (\frac{|C_k|}{|D_i|})^2 \quad i = 1, 2 \tag{34}$$

其中 $|C_k|$ 表示 D_i 划分中第 k 个类别的样本数量。

进一步可以计算样本集 D 在按等级 t 划分时的基尼不纯度：

$$Gini(D, t) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \tag{35}$$

对比计算得到的 k 个基尼不纯度，其值越小不纯度越低，划分方法越好，因此使用基尼不纯度最低的划分方法对应等级作为分类阈值。

对第 i 个连续型特征求解分类阈值：

在 CART 决策树中，对连续型变量的基本处理方法为离散化，本问题共有 m 个特征，因此将连续型变量离散为 m 段区间。设选取的特征为 F，在样本中的最大值为 $\max\limits_{D} F = F_1$，最小值为 $\max\limits_{D} F = F_2$，其中第 i 个区间的上下界如下所示：

$$[F_2 + (F_1 - F - 2)\frac{i-1}{m}, F_2 + (F_1 - F - 2)\frac{i}{m}] \tag{36}$$

记落入第 k 个区间即为样本处于第 k 个等级水平，因此将连续型特征转换为了有序多分类型特征，类似上节处理方法即可求解出该特征的分类阈值。

特征优劣性的评估

设第 i 个特征的基尼不纯度为 $Gini^{(i)}(D)$，可以对每一个特征计算基尼不纯度，进行排序，其中基尼不纯度最小的特征为最优特征，基尼不纯度越高特征越差。

逐次选取最优特征构造决策树

首先给出决策树构造完毕的条件：在每一个叶子节点处，待分类的样本不超过一个，此时决策树构造完成。

Step1：构造根节点：

首先对全数据集构造根节点，然后计算每一个特征的基尼不纯度，选择最优的特征，将该特征及分类条件放置于根节点，将数据集根据根节点进行二分类。若根节点中仅有一个样本则构造完毕，若否前往下一步构造孩子节点。

Step2：递归地构造决策子树：

对二分类中每一个分类的数据集进行节点构造，将 D_1 产生的决策子树记为左子树，D_2 产生的决策子树记为右子树，分别对两个子树执行 **Step1** 构造子树的根节点。

最后，对测试集样本使用决策树分类给出分类结果。使用预先分出的 20% 测试集利用上述步骤构造的决策树模型进行二分类，输出分类结果即可。

如下图左图给出一颗 CART 决策树模型训练的主要流程，右图给出训练完成的一个 CART 决策树模型示例：

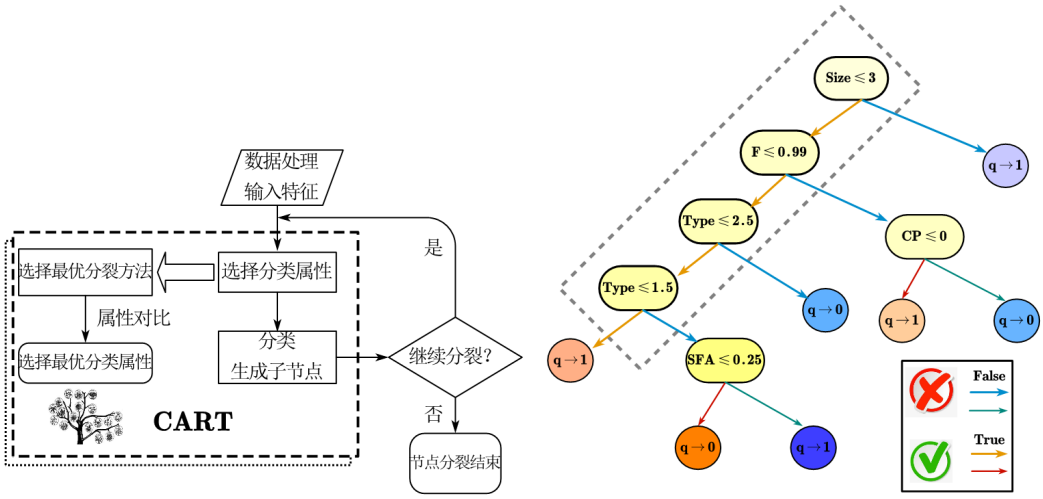


图 2: 决策树结构构造图

4.2 基于决策树构造的随机森林

随机森林是决策树的改进，通过引入 Bootstrap 非参数抽样方法解决小样本量下的二分类问题，最初由 Breiman 提出。

随机森林的主要思想是：首先通过随机选取影响因子、在样本集中随机抽样，构建大量决策树得到不同决策结果；然后综合决策树结果进行投票；最后将票数占比确定划分为对应分类结果的概率。

如下将叙述在本题中基于自变量观测样本集 X 构造随机森林进行样本二分类求解的主要步骤，下图为随机森林构建流程：

Step1：将测试集与训练集的随机划分为 8:2

Step2：基于 Bootstrap 算法对训练集样本进行重抽样

设共有 n 个样本需要进行二分类，根据 Bootstrap 原理需要有放回地对训练集 X 抽取 n 次构成一个新样本集 $X^{(1)}$。不断重复上述过程 n 次得到 n 个样本集 $X^{(t)}$。

Step3：针对每一个样本集构造 CART 决策树 $TR_t(X)$

首先，对每一个 Bootstrap 样本集 $X^{(t)}$，构造 CART 决策树模型，利用样本数据进行决策树模型训练，根据训练结果得到 n 个样本 (可能有重复) 的分类结果。然后，合并重复样本，记合并后样本在原样本集中的指标集合为 $\{index_t\}$。最后，可以给出决策树对样本 x_i 的分类结果为：

$$\begin{aligned} &TR_t(x_i) = \varphi_1, \varphi_2, \cdots, \varphi_{n_{ti}} \\ &\varphi_s \in \{0,1\}, s \in \{1,2,\cdots,n_{ti}\}, i \in \{index_t\} \end{aligned} \tag{37}$$

Step4：根据投票结果计算分类概率

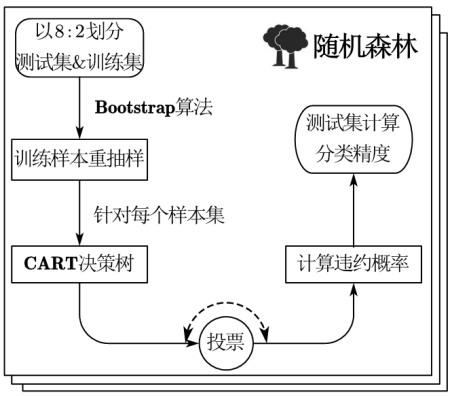


图 3: 随机森林流程图

首先合并上述 n 个决策树分类结果集合，得到每一个样本 x_i 所有的分类结果值如下：

$$\begin{aligned}&\{\varphi_1,\varphi_2,\cdots,\varphi_{n_i}\}\\&\varphi_s\in\{0,1\},s\in\{1,2,\cdots,n_i\},i\in\{1,2,\cdots,n\}\end{aligned}\tag{38}$$

分别求 $\varphi_s=1$ 的次数为 num_{i1}，$\varphi_s=0$ 的次数为 num_{i0}，可以根据如下式计算样本 x_i 分入对应类的概率：

$$P(x_i=1)=\frac{num_{i1}}{n_i}\quad i=1,2,\cdots,n\tag{39}$$

第 5 节 无监督的分类模型：聚类分析模型

聚类分析是一种无监督的分类方法，源于很多领域，包括数学，计算机科学，统计学，生物学和经济学。在不同的应用领域，很多聚类技术都得到了发展，这些技术方法被用作描述数据，衡量不同数据源间的相似性，以及把数据源分类到不同的簇中。

本节将由向量距离与类间距离的相关知识出发，介绍重心法、类平均与离差平方和两种系统聚类方法，进一步介绍有序样品聚类 Fisher 算法。

5.1 预备知识一：常用的距离计算方式

在实际应用中，数据的值受量纲、相关性的影响较大，因此直接在向量空间中计算欧氏距离常无法描述随机变量间的真实差异，这里引入闵式距离、兰氏距离、马氏距离三种各具特性的距离计算方法。

1.Minkowski 距离：

$$d_{ij}(q)=[\sum_{t=1}^{m}|x_{it}-x_{jt}|]^{\frac{1}{q}}\quad(i,j=1,2,\cdots,n)\tag{40}$$

2. 兰氏距离：

$$d_{ij}(L)=\frac{1}{m}\sum_{t=1}^{m}\frac{|x_{it}-x_{jt}|}{x_{it}+x_{jt}}\quad(i,j=1,2,\cdots,n)\tag{41}$$

3. 样本间的马氏距离：

$$d_{ij}(M)=(X_{(i)}-X_{(j)})^TS^{-1}(X_{(i)}-X_{(j)})\quad(i,j=1,2,\cdots,n)\tag{42}$$

其中 S 为两样本 $X_{(i)}, X_{(j)}$ 的协方差矩阵。

在上述距离中可以观察到以下特点：

1. 闵式距离中 q=1 时为绝对值距离，q=2 时为欧氏距离，q 趋于无穷时为切比雪夫距离。
2. 兰氏距离改进了闵式距离与各指标量纲有关的缺点，且兰氏距离更适合高度偏倚的数据。
3. 马氏距离可以进一步排除变量间的相关性的影响。

5.2 预备知识二：聚类间距离的计算方式

在聚类算法中，聚类间的距离计算直接影响最终的分类结果，因此合理选取聚类间距离的构造方法十分重要。最常见的构造方法有：重心法、类平均法、离差平方和方法。

重心法计算类间距离

最直观的定义方式即为：将两个聚类重心间的距离直接作为聚类间的距离。假设某一步后将 G_p, G_q 两类合并为 G_r，三个聚类包含的样本数分别为：n_p, n_q, n_r，各类中心分别为：$\bar{X}^{(p)}, \bar{X}^{(q)}, \bar{X}^{(r)}$，则首先由如下公式：

$$\bar{X}^{(r)} = \frac{1}{n_r}(n_p\bar{X}^{(p)} + n_q\bar{X}^{(q)}) \tag{43}$$

假设某一个聚类 G_k 重心为 $\bar{X}^{(k)}$，则它与新类 G_r 的类间距离可以依如下两种计算方法：

$$D^2_{rk} = (\bar{X}^{(k)} - \bar{X}^{(r)})^T(\bar{X}^{(k)} - \bar{X}^{(r)}) \quad (k \neq p, q) \tag{44}$$

$$D^2_{rk} = \frac{n_p}{n_r}D^2_{pk} + \frac{n_q}{n_r}D^2_{qk} - \frac{n_pn_q}{n_r^2}D^2_{pq} \quad (k \neq p, q) \tag{45}$$

类平均法计算类间距离

由于重心法并未充分利用每一个样品的信息，因此考虑使用样品两两间的平方距离的平均值作为类间距离，即提出如下的类平均方法：

$$D^2_{pq} = \frac{1}{n_pn_q}\sum_{i\in G_p, j\in G_q} d^2_{ij} \tag{46}$$

进一步，对将类 G_p, G_q 合并为 G_r 的过程有如下递推式：

$$D^2_{rk} = \frac{n_p}{n_r}D^2_{pk} + \frac{n_q}{n_r}D^2_{qk} \quad (k \neq p, q) \tag{47}$$

值得注明的是，类平均方法是使用很广泛、效果很好的一种类间距离计算方法。

离差平方和计算类间距离

这里再介绍 Wald 提出的离差平方和方法作为补充。

假定已有 k 分类的样品，分别记聚类为 $G_1, G_2, \cdots, G_k$，其中 $n_t, \bar{X}^{(t)}$ 表示聚类 G_t 的样品数与重心，且在这个聚类中第 i 个样品表示为 $X^{(t)}_{(i)}$，则可以依如下式定义聚类 G_t 的离差平方和：

$$W_t = \sum_{i=1}^{n_t}(X^{(t)}_{(i)} - \bar{X}^{(t)})^T(X^{(t)}_{(i)} - \bar{X}^{(t)}) \tag{48}$$

这里可以进一步计算 k 个类的总离差平方和：

$$W = \sum_{t=1}^{k} W_t \tag{49}$$

Ward 法把某两类 G_p, G_q 合并为 G_r 后的离差平方和增量作为类间的平方距离：

$$D_{pq}^2 = W_r - (W_p + W_q) \tag{50}$$

根据上述离差平方和的定义，可以将上式整理为：

$$D_{pq}^2 = \frac{n_p n_q}{n_r}(\bar{X}^{(p)} - \bar{X}^{(q)})^T(\bar{X}^{(p)} - \bar{X}^{(q)}) \tag{51}$$

因此在 G_p, G_q 合并为 G_r 后，G_r 与其他类 G_k 的距离可以按如下公式递推计算：

$$D_{rk}^2 = \frac{n_k + n_p}{n_r + n_k} D_{pk}^2 + \frac{n_k + n_q}{n_r + n_k} D_{qk}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2 \tag{52}$$

5.3 系统聚类模型

Step1: 聚类数量的确定

1. 在实际应用中往往会根据应用背景选择聚类的数量，或根据已有的先验知识确定最终的目标聚类数量。

2. 在对应用背景完全未知时，可以预先选取实数 d 作为类间距离的阈值，下文将讨论在这种情形下的系统聚类算法。

Step2: 求解样品距离矩阵

对未经任何处理的 n 个样品，如下式计算第 i 与 j 个样品间的距离可以得到样品间的距离矩阵 $D^{(0)}$，其中 d 可以选择上文三种样品距离计算方法的任意一种，常用欧式距离计算。

$$D^{(0)}(i,j) = d(X_{(i)}, X_{(j)}) \tag{53}$$

Step3: 类的合并

初始状态下，类的个数 k=n，类间距离即为样品间的距离，假设第 i 次聚类之后，可以计算的得到类间距离矩阵 $D^{(i-1)}$，则合并类间距离最小的两类为一个新的聚类，此时可以得到 k=n-i+1 个类。而如果 k 不大于预先给定的聚类数量或最大的类间距离小于给定的阈值 d，则可以认为聚类过程完成。

Step4: 聚类完成

最终绘制谱系聚类图，并统计各个类内的成员，计算类重心、类数量等参数，聚类完成。

5.4 有序样品的聚类算法

在有序聚类问题中，如果进行系统聚类会丢失样品间的排序信息，因此引入最优分割法解决这类聚类问题，又称 Fisher 算法。

定义聚类直径

假设有序样品为 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$，均为 m 维向量。设某一类 G 包含样品 $\{X_{(i)}, X_{(i+1)}, \cdots, X_{(j)}\}$，

记 G 为 {i,i+1,$\cdots$,j}, 可以依如下式定义这个类的均值向量与类直径：

$$\bar{X_G} = \frac{1}{j-i+1}\sum_{t=i}^{j} X_{(t)} \tag{54}$$

$$D(i,j) = \sum_{t=i}^{j}(X_{(t)} - \bar{X}_G)^T(X_{(t)} - \bar{X}_G) \tag{55}$$

定义损失函数

记 $b(n,k)$ 为将 n 个有序样品分为 k 类的某一种分法，表示为：$G_k = \{i_k, i_k+1, \cdots, i_{k+1}\}$, 进一步可以根据聚类直径定义损失函数如下：

$$L[b(n,k)] = \sum_{t=1}^{k} D(i_t, i_{t+1}-1) \tag{56}$$

递推计算最优解

Fisher 算法本质上是递推算法，通过已知的聚类直径信息可以由如下递推公式递归地产生聚类：

$$L[P(n,k)] = \min_{2\le j\le n}\{D(1,j-1) + D(j,n) \quad k \le j \le n\} \tag{57}$$

$$L[P(n,k)] = \min_{k\le j\le n}\{L[P(j-1,k-1)] + D(j,n) \quad k \le j \le n\} \tag{58}$$

上式计算得到所有的 D(i,j) 与 L[P(i,j)], 其中 P(n,k) 为使损失函数达到最小的分类方法，且 $1 \le i \le n, i \le j \le n$, 在计算过程中可以如下确定聚类：

1. 由 L[P(n,k)] 取最小的 j 确定第 k 个类 $G_k = \{j_k, j_k+1, \cdots, n\}$, 其中 $j_k = j$。
2. 由 L[P(j,k-1)] 取最小的 j_{k-1}，确定第 k-1 个类 $G_{k-1} = \{j_{k-1}, j_{k-1}+1, \cdots, j_k-1\}$。
3. 循环 2 步骤直到 G2 构造为 $G_2 = \{j_2, j_2+1, \cdots, j_3-1\}$。
4. 最终构造 $G_1 = \{1, 2, \cdots, j_2-1\}$。

以上步骤可以得到最终的聚类结果：$P(n,k) = \{G_1, G_2, \cdots, G_k\}$。

第 6 节 参考文献

[1] 张潇. 大数据背景下的车险索赔概率研究 [D]. 山东大学,2020.

第 V 部分

优秀美赛论文欣赏

珞珈数学
Mathematics of Luojia March 2018

数学建模中的创造性

应用数学系　胡元明

摘要

通过多年实践，经过认真、反复地思考，我们从数学建模的学术角度对大学生培养中被公认是极端重要的创造性、创新能力的内涵进行了比较深入的探讨，逐渐从大量的感性认知中升华出一些理性的认识。

1　创造性的一种分类结果

数学建模的角度看，创造性根据创造积累的时间长度、所运用知识的深度大致可以分为两类。一类是原创性成果、重大发明中所包含的创造性，这些创造不是一朝一夕就可以实现的，都需要经过长时间的积累，甚至几代人的努力。例如，载人宇宙飞船的研制和发射、优质杂交水稻品种的培育和推广、概率论中的中心极限定理、哥德巴赫猜想的证明等。另一大类创造性种类繁多、情况各异，但也有共同的特点，就是“一听就能够明白，不听就是想不到，采用后作用重大”。这些创造性与前一类创造性的差别在于，它不需要特别高深的理论和复杂的知识背景，一般当事人已经具备或只需要稍加补充即可从事这些活动，甚至道理浅显近乎常识；它解决问题的过程也比较短暂，无需漫长的积累，甚至“立竿见影”；但采用这些创造性后，对困难的问题就能“势如破竹，迎刃而解”，“攻坚克难，如履平地”。例如，经济学中的投入产出理论，虽然对经济界产生重大的影响，但从代数理论上看并不高深，只是将众多原材料和产品之间的数量关系近似为线性关系并把这种关系用矩阵来表达，然后利用矩阵有关理论得出经济方面诸如各行业之间应协调发展等许多重要的经济规律。再例如数学建模的经典范例，牛顿推导著名的万有引力定律，就是先用简单的极坐标参数方程来表示在椭圆轨道上物体的运动，进而对这个方程进行简单求导，最高也就是二阶导数，最后将开普勒三大运动定律的结论代入上述求导的结果，就得出了万有引力定律，过程并不复杂。

全国大学生数学建模竞赛二十多年来，也不乏许多学科的前沿问题，然而大学生们在短短的 72 小时之内也做出了一些有价值的成果，有力地证明了第二类创造性在大学生数学建模活动中同样大量存在。

虽然上述两种创造性相互之间存在明显的差别，但它们之间的联系却是相当紧密。实际上，第一种创造性的基础就是第二种创造性，第二种创造性经过长期大量的积累可能升华为第一种创造性；反过来，第一种创造性中蕴含了大量的第二种创造性，第一种创造性的产生也会大大刺激第二种创造性的涌现。

第二种创造性因为不需要当事人有特别高深的理论和复杂的知识背景，限制比较少，存在的范围非常大，应该是大学生创造性培养的重点，同时也是数学建模活动力所能及、可以胜任的任务。又因为一旦培养出这些创造性，人们的能力就可以大幅提升，工作效率就会有惊人的提高。第二种创造性大量存在的事实，有力地说明创造性固然可以极大地提高效率，但创造性并不神秘，并非高不可攀。而通过数学建模活动有利于培养大学生们的第二种创造性，从而增强大学生从事科学研究的自信心、提高大学生解决实际问题的能力，所以值得重视。

2　创造性的另一种分类结果

根据二十多年来全国大学生数学建模竞赛和为我校大学生开设数学建模课程的长期实践，仅从学术角度分析、考察大学生数学建模活动和相关的学术研究活动，我们认为大学生数学建模活动中所需要、所培养的创造性大致可以归纳为以下十个方面。

2.1 敢于质疑、挑战权威，勇于猜测、标新立异，善于发现、提出有价值的问题

世界上的许多事物是错综复杂的，没有经验的人遇到这类问题经常会感到无从下手，甚至不知道应该解决什么问题，不知道应该向什么方向努力，更不知道会有什么结果，只能是“盲人骑瞎马”。所以提出有价值的问题或新的理论是创造的前提，也是重要的创造性。例如，数学上的费马大定理、概率论的中心极限定理，天文学的宇宙大爆炸的学说，化学上的元素周期律等都是因为猜测并提出有价值的问题而导致有关学科的迅速发展。许多定理、定律、学说都是现有命题、假设、猜想，经过理论或时间的证明才最终成为定理、定律、学说。创造性之所以被称为创造，就是因为从来没有人这么想过，没有人这么做过。因此它首先一定是大胆的猜测，固然要有一定的道理，但也不会有绝对的把握。猜是经验的升华，是跳跃式的思考，是前进的阶梯。猜来自敏锐的洞察力，猜的基础是对问题本质的研究。经常猜测有助于活跃思维，所以猜测是创造性的摇篮。

而要解决新问题特别是困难的问题，一定伴随着思想的突破与飞跃，经常会与主流观念发生激烈的冲突。如果不敢质疑权威，墨守成规，就不会有大胆的猜测，也就不会有质的变化。爱因斯坦如果不敢质疑几百年来一直占据统治地位的牛顿运动定律，就不会有相对论原理。因此猜测经常且必须和质疑紧密相连。

当然要猜测，首先要有猜测的对象、猜测的目标、猜测的可能结果，因此必须首先发现问题。人们处于大致相同的环境，接触基本相同的对象，接受大致相同的信息，甚至具备相同的学习经历，但多数人发现不了问题，更谈不上提出有价值的问题，少数善于观察、勤于动脑的人却可以发现并提出值得思考的问题，显示出显著地差别。正因为如此，大学生阶段创造性培养的重要内容之一就是让他们解放思想、突破束缚；敢于质疑、挑战权威；勇于猜测、标新立异；善于发现、提出新问题、新理念、新方法，这一点可能是我国大学生教育的薄弱环节。

2.2 善于经常从多个不同的视角观察、考虑问题，思维活跃，精于发现不同事物的相似之处，长于借鉴，移植，往往能够另辟蹊径地解决问题

为什么对问题会有不同的看法、不同的结论，一般是由于看问题的角度不同，关注的重点不同。有与众不同的视角，注意到被其他人忽略的关键，就有可能产生创新。为什么会有不同的做法、不同的途径，多数源于经历的不同、接受教育的不同、日常观察、考虑问题的方式不同。如果经常能够从表面上大不相同的事物发现它们的共性，甚至不同的本质，就容易借鉴、移植其他学科的方法和结论，另辟蹊径地解决问题，这种创造，相对而言比较容易实现。大学生应该训练多方位地观察事物，思考问题。青藏铁路中“以桥代路”的方法创造性地解决了高原活跃冻土带施工地世界性难题，就是由于另辟蹊径，穿过冻土层直接在岩石上建桩，在桩上架桥，在桥上铺铁路就能有效避免冻土层对铁路路基地破坏。

2.3 具体问题具体分析，正确选择问题“突破口”的能力是一种重要的创新能力

即使再困难的问题也肯定有相对薄弱的部分，选择从这些地方攻关，就可以取得突破，快速推进解决问题的进程。因为要解决的问题千姿百态、千变万化，要善于分析实际问题的特点，才能从中寻找出薄弱环节予以突破，所以如何选择“突破口”具有很强的创造性。此外也不是对每个问题“突破口”的选择都毫无规律可寻，只是在很大程度上依赖经验的积累，依赖当事人对类似、有部分相同或相似问题的处理经历，依赖当事人对成功解决问题全过程的了解，总之，“熟能生巧”。由于全国大学生数学建模竞赛的题目都是没有被解决过的、比较困难的实际问题，所以在选择“突破口”方面，为大学生提供了极好的锻炼机会。

2.4 善于把复杂问题恰当地分解为一系列简单问题的串并联，制定合适的技术路线是科技人员必备的创造性

解决复杂问题绝不能一蹴而就。解决复杂问题就好像攀登高山，要想成功登顶，一定要选择正确的登山路线，既要在前进中保持逐段向上，又要能不断地前进直至登顶。同样解决一个复杂问题，一定要制定一条合适的技术路线，要把技术上地整体跨度分解成若干个可达跨度来实现，把一个复杂问题恰当地分解为一系列简单问题的串并联；由于每一个子问题比较简单因而能够容易得到解决；当所有这些简单的子问题都解决了，则复杂问题就最终获得了解决。因此，这种创造性是高级科技人员必须具备的。

要制定正确的技术路线迫切需要创造性和敏锐的洞察力。应该不断用我们熟悉的事物去描述我们不熟悉的事物，不断用确定的内容去替换那些尚未确定的内容，不断以已经获得的结论为基础去扩大战果，要根据过去的经验去预测预期的成果和可能的结论，确定下一步的目标和步骤，直至问题的完全解决。而要能够实现这个过程只能依赖实践的熏陶。由于大学生数学建模竞赛的题目有相当的难度，要解决它们一定要制定恰当的技术路线，因此对培养大学生制定合适的技术路线的创造性很有帮助。

2.5 学科交叉是创造性的源泉之一，科研人员要能够将各学科知识融会贯通、灵活运用

实际问题和已经被抽出来的理论问题之间最大的区别就在于它不会仅仅属于某一个学科，它有许多具体的、各种各样的属性，它们的变化受到各种规律的支配。即使用某个学科最先进的成果来分析复杂的实际问题，也仅仅是从一些侧面、某些角度来进行考察，仍然可能无法对错综复杂的现象做出全面、合理、本质的解释。因此要解决这类问题，学科交叉、知识融合就是必不可少的。尤其在科学技术高度发达的今天，各门学科之间相互渗透、相互融合已经相当普遍；由于学科交叉，一门学科某个方面的突破带动其他学科进展的事例层出不穷；许多重大科技项目都因为多学科联合攻关而取得成功；不少重大科技成果都是多学科共同协作的结晶，有力地说明了学科交叉、知识融合是创造性的源泉之一。

然而大学生们尽管学习过多门学科的大量科学知识，但在他们的脑海里，各门学科的知识之间并没有做到融会贯通，与实际问题中各学科的规律紧密耦合成一体是迥然不同的，这大大制约了大学生创造性的发挥。由于在自然界一切小的规律都是受普遍规律支配的，而且不同的事物之间也不是截然不同的，经常发生的情况反而是不同的事物之间存在某种共性，不同的实际问题经常有相同的数学模型。因此牢记并熟练掌握重要的普遍规律，适当扩大知识面，在学习其他学科知识时经常联系本学科的有关问题，注意借鉴，可能会有意想不到的收获。

2.6 对知识深刻理解、灵活运用能够产生创造性

书本知识与实际问题之间总存在一定的距离。一般情况下，书籍特别是学术著作只介绍基本原理、基本方法，很少介绍将知识如何应用于解决具体的实际问题。即使介绍个别的具体应用事例，从使用角度看也很不全面。因此如果人们对知识理解不深、认识不透，对知识的运用更加生疏，在接触不熟悉的问题时，就会想不到或者想不出办法把已经学习过的数学知识运用到实际问题中去。

在各门学科的知识形成过程中都必然包含着巨大的创造，但是，其中创造性并不能简单地通过知识的传授就可以为受教育者所接受。显然学习并全部理解牛顿所创立的微积分和牛顿三大运动定律，甚至学习并理解牛顿的全部学术著作也绝不能够就成为牛顿那样伟大的科学家。现实中经常发生的却是在知识形成过程中的创造性被淹没在“以其昏昏”，无法“使人昭昭”的平庸教学和被动的单纯接受中，以至于历史上重大的科技进步能够在培养教育者的创造性中发挥积极作用的并不多见。现在不少大学生教材都只有结果、只有结论，不介绍探究的过程；不少书籍都是相互传抄，完全不见当年的创造性、突破性的思考，要初学者不折不扣地领会其中的创造显然不太现实。

创造性经常存在于从感性知识到理性知识地飞跃。从数学建模地角度看，推导和证明是其中的关键。可是往往多数大学生在学习过程中只关心结论，忽略了推导和证明，对证明中地创新理解不透，更谈不上掌握推导和证明的一般方法，使很多好的思路、结果无法上升为理论。例如，证明某个解是最优解，掌握证明规律后一般并不困难。如果未能给出证明，则这个结果只能作为经验处理，不能视为一般规律。因而显著地降低了论文的创造性和理论价值，即使把结果应用于所解决的实际问题，也有不良影响，非常可惜。

总之，知识中蕴含着大量的创造性，学懂知识，并不代表理解其中的创造性，必须经过认识上的升华。

2.7 洞察事物规律和抓准问题的主要矛盾也属于创造性的范畴

众所周知，错综复杂的事物内部有许多矛盾，但在一定时期内一定有一种矛盾是主要的，抓住这个主要矛盾，问题就迎刃而解了。要能够最终彻底解决困难的问题，必须对问题有本质的了解。但问题的本质又往往被许多表面现象所掩盖，甚至为一些假象所包裹，要抓住问题的本质必须撕开假象、透过表面现象去发现问题的本质。不同水平、不同层次的当事人也往往在这种情况下暴露出显著地差别。抓准主要矛盾、洞察其他人没有发现的规律就是创造性的体现。

在抓准问题的主要矛盾和发现事物规律方面，行之有效的办法，就是应该通过压缩问题的规模、降低问题的难度、固定一些原来可变的条件、暂不考虑一些影响结果的因素、构造出相对简单的情况，这样就容易发现问题的规律。通过简化、固定条件，增加复杂问题和简单问题之间的可比性，借用对简单问题已经知道的主要矛盾、客观规律、去猜测复杂问题的主要矛盾、客观规律。

2.8 问题的多种表达方式包含创造性

错综复杂的问题有许多侧面，包含众多现象，问题内部有许多元素，元素之间有复杂的关系，还不断发生变化。特别是对新问题而言，准确、简洁、全面、严格、通俗地把问题表达出来，本身就是创造。因为准确、全面、严格地表达问题是解决问题地前提，简洁、直观、本质地表达是创造性思想地“温床”。多种表达中蕴含了丰富的创造性，使人们从多个视角观察问题成为可能。

2.9 善于捕捉信息、有效地综合利用信息是信息社会时代地基本创造性

进入信息化社会，数据急剧膨胀。海量数据使人目不暇接，熟视无睹，人们对数据已经近乎麻木，人脑好像已经无法再存储，信息筛选成为无法回避地重大问题。虽然在统计数据以及熟悉模型的计算或仿真结果中蕴藏着大量有价值的信息，但拥有同样的数据、同样的结果，对不同的人却起着完全不同的作用。因此善于捕捉隐藏在海量数据中的重要信息，有效地利用数据就需要创造性。因为“巧妇难为无米之炊”，所以防止重要、宝贵的信息从手中不经意地滑走是科技工作者十分重要地品质。发现有有价值地信息有时是解决问题地关键创造性。当然仅发现信息也是不够地，要创造性地解决实际问题必须善于选择和综合利用已有的信息，否则一大堆杂乱无章地信息只会使人手足无措，这时综合信息地创造性就能够发挥重要地作用。现在互联网和计算机软件发展异常迅猛，给数据的收集、查找、整理、分析创造了极为有利的条件，充分、高效地综合利用这些条件是科技工作者应该具备的基本创造性。

2.10 对结果的分析、挖掘、推广同样需要创造性

实际问题的数学模型建立之后，将实际的数据带入或者进行仿真之后就会有一批数据输出，这些结果中也包含着大量有价值的信息，也蕴藏着事物的变化规律，刻画了问题的本质。这批数学建模的成果能否充分地被消化吸收对实现数学建模的最终目的有举足轻重的影响。所以必须加强对结果的分析、挖掘，以最大限度地发挥数学建模的功能。同时，在建立数学模型时都是有假设的，有的是为了简化问题，有的是开始时考虑不太周到。所以这些假设不一定总是成立的，当然会限制数学模型的使用。尤其是在得到非常理想的结果时，能够扩大原来数学模型的使用范围，降低使用这批数学模型的“门槛”具有重要的意义。这里同样需要创造性。

珞珈数学
Mathematics of Luojia March 2018

Race against time ---Emergency evacuation plan for the Louvre

方姝蘅　蓝梦婷　方启航

1 Summary

Emergenccy will result in a large scope of damage. If there is no effectice evacuation strategy, it is very likely to cause unnecessary or evensecondary damage during the evacuation process. To this end, it is necessary to generate evacuation plans, assign reasonable routes to people in dangerous locations as soon as possible, make maximum use of the capacity of the passage and reduce congestion so that visitors can reach safe places as soon as possible.

Our aim is to develop an**emergency evacuation plan**for the Lourve, for which we develop two models at the macro and micro levels. Then we develop an adaptive model related to potential threats. We obtain total evacuation time in different situations and optimal route by solving our model.

In Macro emergency evacuation model, we simplified the Louvre into a graph. To obtain optimal route, we use**time expanded graph**to transform the problem into minimum cost flow theory. To solve the problem, we use algorithm based on **Capacity Constrained Route Planning(CCRP).**Then we draw a conclusion that the time taken by the latest person arriving at the exit is 330s. The results of the optimal route are shown in Table 1. We also explore the utilization of app"Affluen-ces"to help us improve our plan.

In Crowd evacuation behavior model, we analyze people's **pre-movement time** and people's **movement velocity**, which will influence the whole evacuation process. What's more, we did some simulation of several evacuation scenarios in the Louvre.

In potential threats of circumstances model, we explore the reason that bottlenecks occur. We find that although people arrive at exit quickly, they have to wait for a long time to go out. It takes about 1860s. To improve the model,we introduce the node capacity and utilize additional exits. It takes about 450s to leave the Louvre by utilizing 2 exits at 150s and 4 exits at 180s. We also build an emergency personnel entering model to explore when they can enter the Louvre to prevent potential threats.

Besides, we do some sensitivity analysis about the speed of movement and potential threats. We also propose some policies and suggestions for the Louvre. Finally, we briefly state the application of our models in other large, crowded structures.

Keywords: Evacuation　CCRP　optimized route　potential threats　bottlenecks

2 Introduction

2.1 Problem Background

In recent years, an increasing number of terror attacks in Paris have triggered new thinking on urban security and contingency measures, especially how to safely and effectively evacuate people in public places. As one of the most visited landmarks in France, the Louvre is marked by rich and marvelous exhibitions which appeal to tens of thousands of people from across the globe.

In an effort to decrease the damages that result from emergencies, it is of utmost importance for the museum to design emergency evacuation plans. To come up with a comprehensive and optimal plan, there are lots of factors related to the actual situation of the Louvre to be considered.

2.2 Interpretation of the Problem

1)We should develop an emergency evacuation model that allows individuals egress to and through an optimal exit in order to empty the building as quickly as possible, while also allowing emergency personnel to enter the building as quickly as possible.

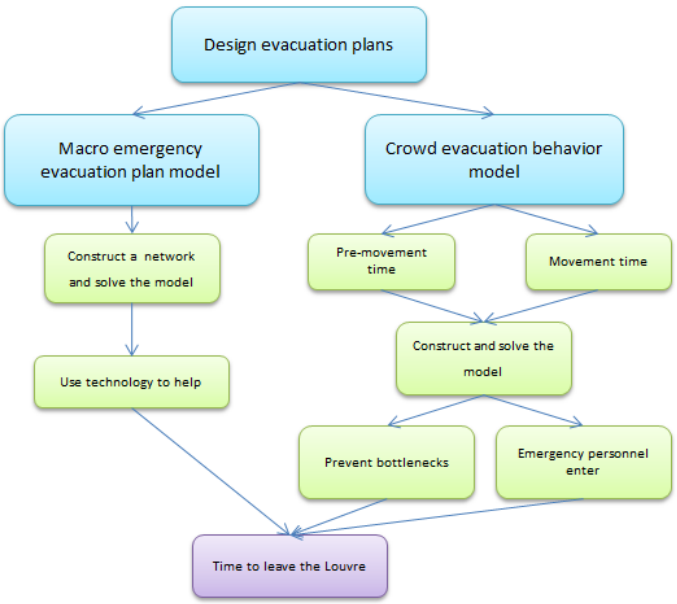
2)We should take the fact that the number of guests in the museum varies throughout the day and the diversity of visitors into consideration. Further more, we should figure out how technology such as Affluences could be used to facilitate our evacuation plan.

3)Considering the public awareness of total exit points(service doors, emplo-yee entrance, VIP entrances, and old secret entrances built by the monarchy, etc.) serves as a double-edged sword, we should analyze carefully when and how any additional exits might be utilized.

4)We should build an adaptable model that can address a broad set of considerations and various types of potential threats. Validate the model and discuss how the Louvre would implement it.

5)Propose policy and procedural recommendations including applicable cro-wd management and control procedures that are necessary for the safety of the visitors for emergency management of the Louvre.

2.3 Overview of our work



Figue 1: Overview of our work

First, we develop a Macro emergency evacuation model to explore the optimized evacuation routes. We also consider how technology helps to facilitate our plan.

Second, we analyze people's behavior during specific evacuation process, including pre-movement time and movement time. In this way, we can optimize our first model by considering something detailed. We also do some simulations of evacuation scenarios in the museum.

Then, we solve other problems raised by the topic.

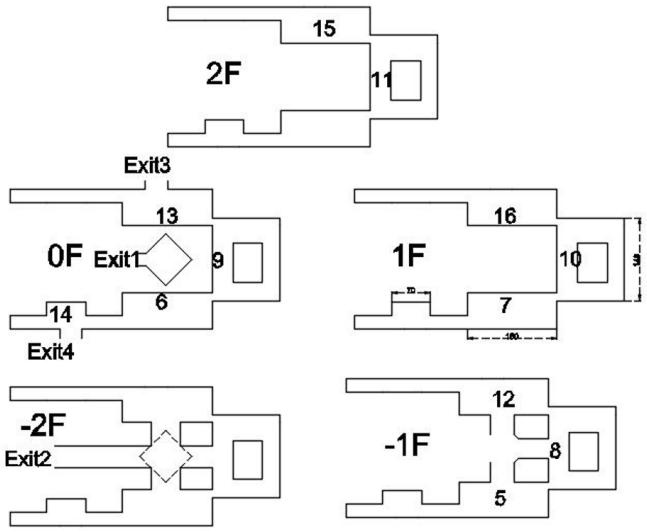
- How to prevent bottlenecks? To solve the problem, we should decide when and how to utilize additional exits.
- When should emergency personnel enter the museum.

Last, we make a sensitivity analysis about the speed of movement and potential threats.

3 Preliminaries

3.1 Constructing the Louvre evacuation network

According to the Louvre Museum Plane Map, the Louvre has five floors, and there are three exhibition halls on each of the upper four floors: RICHELIEU, DENON and SULLY. We assign a specific number to each exhibition hall and get a simplified floor plan of Louvre as shown in Figure 2.



Figue 2: The floor plan of Louvre

We begin to build the Louvre evacuation network by making some reasonable assumptions:

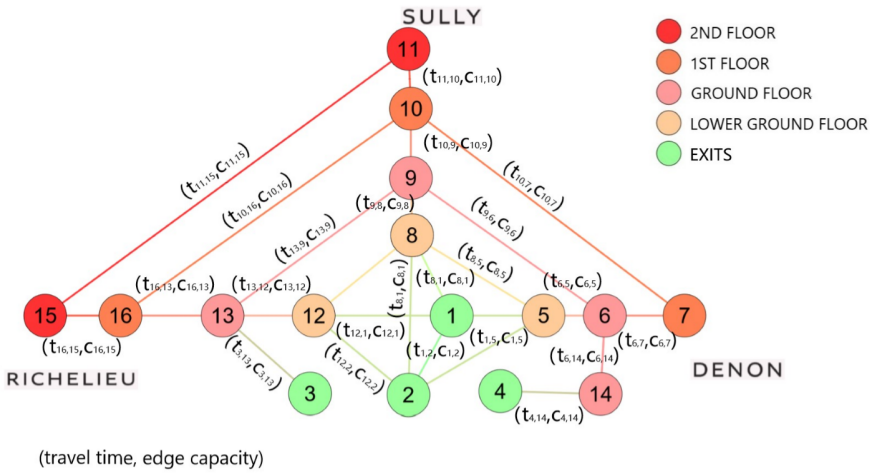
- Under normal circumstances, tourists gather in exhibition halls. For simplicity and without loss of closeness to reality, we do not consider the tourists scattered on the stairs and corridors before an emergency.
- The evacuation channel between the exhibition halls has a limited capacity and can only allow a certain number of people to pass at a time.

Define $V(G) = v_1, v_2, \cdots v_{16}$as the set of all exhibition halls and exits, where $v_1 \cdots v_4$ denotes the four main entrances/exits–the pyramid entrance, the Carrou-sel du Louvre entrance, the Passage Richelieu entrance and the Portes Des Lions entrance respectively, $v_5 \cdots v_{16}$denotes the three exhibition halls situated on the upper four floors of Louvre: RICHELIEU, DENON and SULLY. Define$E(G)$as the set of edges of the network. Each edge denotes all available evacuation routes between two nodes. $(v_i, v_j) \in E(G)$if one of the following holds:

- i and j are neighboring exhibition halls on the same floor.

- i and j are the same exhibition hall on the directly-related floors.
- i is an exhibition hall and j is a directly-connected exit.
- i and j are neighboring exits.

$G = V(G), E(G)$ is defined as the graph of Louvre evacuation network. Additionally, edge ij is marked by two attributes: travel time and edge capacity, which will be involved in our following models. We then visualize this network in Figure 3.



Figue 3: Louvre evacuation network

4 The Model

4.1 Macro emergency evacuation model

We introduce a Linear programming based method to model the emergency evacuation on the basis of the above Louvre evacuation network. Firstly, We summarize some basic rules:

- The emergency departments in authority broadcast emergency announcements which are always clearly heard by individuals. And people evacuate to safe places in the guidance of emergency personnels.
- The Louvre received 10.2 million visitors in 2018[1]. In this model, we assume that the number of visitors is uniform. (we will discuss the variety of number of guests in the following section.)
- We assign numbers of visitors to each exhibition hall based on their popularities among people. For example, the treasure "Mona Lisa Smile" on the first floor of DENON WING will attract more visitors than others.
- The distance between different exhibition halls is defined as the straight-line distance. Note that the distance between the same exhibition halls on directed-related floors is the path length of the stairs.
- Due to the lower security level at the additional exits compared to four main entrances, we prioritize the four main entrances in our original model.

Mathematical notations

In order to clearly illustrate our model, we now settle down some mathematical notations:

- n_i: the initial occupancy of visitors in node $i, i = 5...16$,which denotes an exhibition hall

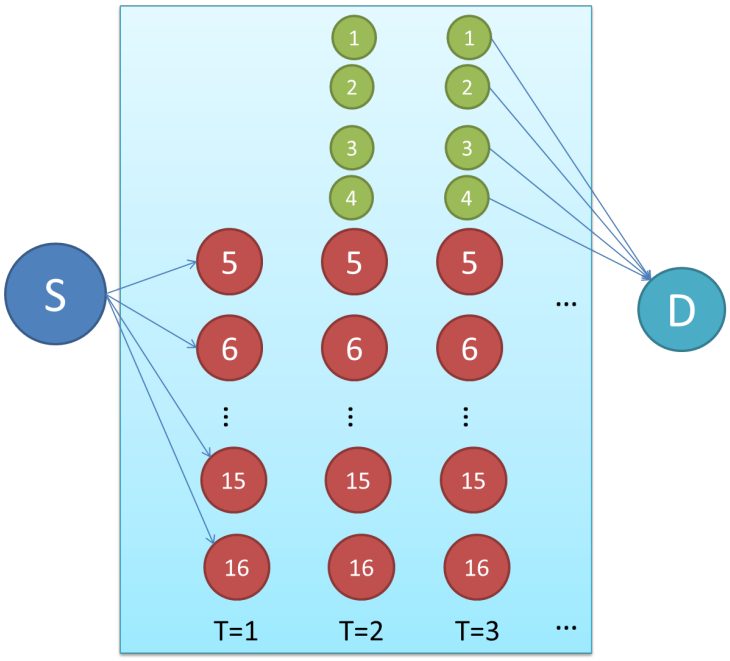
- D_{ij} : the length of path between node i and node j
- $v = \begin{cases} v_1, & \text{visitor's moving speed on the ground} \\ v_2, & \text{visitor's moving speed on the stairs} \end{cases}$
- t_{ij}: average time period required for a visitor to move from node i to node j.$t_{ij} = \frac{D_{ij}}{v}$
- U_{ij}: the maximum capacity of edge ij
- f_{ij}: the number of visitors evacuating through edge ij from node i to node j

The model construction

To solve the original problem, we introduce a **Linear programming based method.** More specifically, we transform the network into its time-expanded graph and solve the minimum cost flow problem on the graph.

- **Time expanded graph**

The time expanded graph needs to be expanded by a multiple of the time. We have to copy the nodes of the original image according to how many units of time that are required for the evacuation process. A time expanded graph based on the original network is shown in Figure 4.



Figue 4: Louvre evacuation network

As is shown in the graph, nodes 5-16 are source nodes, from which visitors evacuate.Nodes 1-4 are destination nodes, which represent the safe exits. S denotes the virtual Super source node. The population to be evacuated in S is the sum of the population to be evacuated in all source nodes. Similarly, D denotes the virtual Super destination node, it can accommodate all the population to be evacuated. Note that we have omitted the specific edges between the nodes for simplicity.

- **Minimum cost flow theory**

  We introduce a novel method to model the evacuation process, which is conceptually similar to Minimum cost flow theory.

  The objectve for the model is:

– the minimum time taken for all visitors to move from exhibition halls to exits.

The constraints for the model are:

– All the visitors must be evacuated to safe places. Thus total number of visitors evacuating to safe places(exits) must be equal to the total initial occupancies of all exhibition halls.

– During the evacuation process, the total number of people who flow out of the exhibition hall and that of people flowing into the exhibition hall are dynamically balanced, that is, the number of people who flow out minus that of people flowing into the exhibition hall is equal to total initial occupancies.

– The number of visitors evacuating through edge ij from node i to node j at a time must be no greater than the max capacity of edge ij.

These objectives and constraints are realized by:

$$\min \sum_{(i,j)\in G} t_{ij} f_{ij} \tag{1}$$

$$S.T. \begin{cases} \sum\limits_{k=1}^{4} \sum\limits_{j\in V(G)} f_{jk} = \sum\limits_{i=5}^{16} n_i \\ \sum\limits_{i=5}^{16} (\sum\limits_{j\in V(G)} f_{ij} - \sum\limits_{j\in V(G)} f_{ji}) = \sum\limits_{i=5}^{16} n_i \\ 0 \le f_{ij} \le U_{ij} \end{cases} \tag{2}$$

To solve the model

The method based on linear programming has certain drawbacks. For example, the increasing number of nodes and time steps in time expanded graph will introduce expensive computatinal costs. As a result, we introduce an algorithm based on **Capacity Constrained Route Planning(CCRP)[2].**

Step 1 Input a graph G with a set of nodes N and a set of edges E. Define S as the set of source nodes, D as the set of destination nodes, $S, D \subseteq N$. Each node n has a property: Initial node occupancy. Each edge e has two properties: Maximun edge capacity and Travel time.

Step 2 For a source node s $\in S$, if it still has evacuees, use Dijkstra Algorithm to find the route $R < n_0, n_1 \cdots n_k >$ with time schedule $< t_0, t_1 \cdots t_k >$ from source node s to all destinations, and then determine the Earliest Arrival Path.

Step 3 Determine the flow on the Earliest Arrival Path. It follows: $flow = Min\{$number of evacuees still at source node s, available capacity of edge $i\}$, where edge i lies on the Earliest Arrival Path.

Step 4 "Reserve" edges on the Earliest Arrival Path. Subtract the value of flow from the available capacity of each edge on the Earliest Arrival Path, and return to **Step 2.**

The basic results of evacuation routes are as Table 1.

Notes:the percentages in the table present the allocating proportion of evacuees on the corresponding route.

According to the analysis of Table 1, we can conclude:

Law one Two exhibition halls of SULLY and RICHELIEU on the 2nd floor need the longest evacuation time: 330 seconds, which is also the time required to evacuate all visitors in Louvre.

Law two More effective evacuation path is the path between the same exhibition halls on different floors, while the path between different wings on the same floor is almost not used. This is partly because of the long length of path between wings on the same floor. Further more, movement between different exhibition halls not significantly reduce evacuation time because the exits are on the lower floors.

Law three It can be seen that the No.2 exit (i.e. the Carrousel du Louvre entrance) is hardly used alone

because it is rather far. When the available capacity of the path to the No.1 exit is insufficient, the No.2 exit will be selected.

Law four No.1 exit (The pyramid entrance) is the most used exit for evacuation because it is located in the center of the Louvre and is near to all wings.

Table 1: Evacuation routes

Source node	Arrival time(s)	Routes
5	100	5→ 1
6	144	6→ 5→ 1(36%)
6	144	6→ 14 → 4(64%)
7	240	7→ 6→ 14→ 4(22%)
7	240	7→ 6→ 5→ 1(33%)
7	240	7→ 6→ 5→ 2(45%)
8	60	8→ 1
9	160	9→ 8→ 1(75%)
9	160	9→ 8→2(25%)
10	240	10→ 9→ 8→ 1(78%)
10	240	10→ 9→ 8→ 2(22%)
11	330	11→ 10 → 9→ 8→ 1(80%)
11	330	11→ 10→ 9→8 →2(20%)
12	50	12→ 1
13	96	13→ 3
14	64	14 → 4
15	330	15→ 16 → 13 →12→ 1(66%)
15	330	15→16 →13→3(34%)
16	180	16→ 13→12 →1(66%)
16	180	16→ 13→3(34%)

- **Using technology to facilitate evacuation plan**

The number of guests in the museum varies from day to day. Evacuation plan should consider the adjustments resulting from changes in the number of guests. We can use Affluences to help us estimate the number of guests in Louvre.

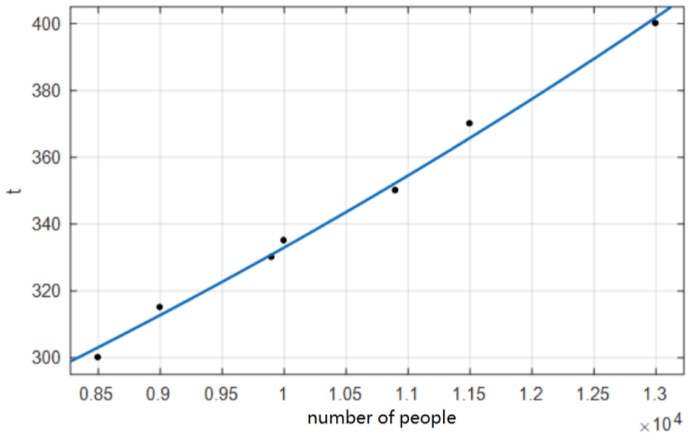
In macro emergency evacuation plan, we use the average number of guests in a certain period of time as the real-time visitors in the museum. In fact, the number of guests varies from weekday to weekend.

In weekday, the number of guests is less than the average. From Affluences, we know the average time required for queuing is 5 minutes and the longest is no more than 10 minutes. While in weekend, the number of guests is more than the average. And the average time required for queuing is 15 minutes and the longest is 20 to 30 minutes.

In the former case, we do not need to adjust the evacuation plan, and the final evacuation time will be less than the previous result. While in the latter one, adjustments should be made.

Queuing results from waiting for security check. We assume that the rate of guests passing the security check is 1 person per second. The peak period of the queuing time is 15 to 25 minutes more than the average, so the number of guests is 900 to 1,500 more than the average. Use Macro emergency evacuation model to calculate an increasing number of guests. Due to space limitations, the result is omitted. After calculating, it takes 350s-370s to get all the guests who arrive at the exit in the peak period. There is an increase in evacuation time compared to the initial plan.

Figure 5 shows relationship between museum guests and evacuation time, which can help facilitate evacuation plan.



Figue 5: the relation between population and evacuation time

In reality, people' s evacuation during emergency is a complex task. The implementation of evacuation plans is influenced by many factors. In the following sections, we will discuss how the behavior of evacuated people and the potential threats in the emergency environment affect the evacuation process.

4.2 Crowd evacuation behavior model

Our first model does not consider the complex terrain of the Louvre, so we develop the second model to analyze specific passing time during evacuation process. We first divide the evacuation process into two period: Pre-movement and Movement. The former describe people' s recognition and response to the emergency alert, while the latter involves the evacuating behavior after recognition.

Pre-movement

The response time to an emergency will largely determine the damage caused by the disasters. It is necessary to analyze pre-movement behavior in making evacuation plans.In this period, we focus on two factors: evacuation information and human relations.

- **Evacuation information**

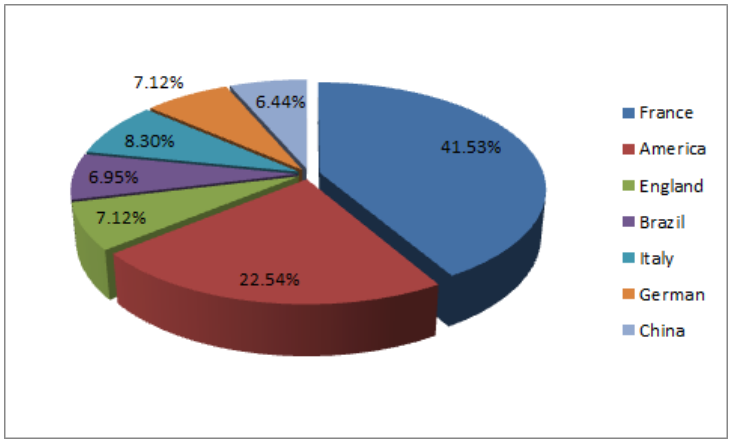
Evacuation information refers to signals of emergency events, emergency conditions, evacuation routes etc., which are usually released by announcements from those in authority and disseminate among evacuees. The more evacuees be infomed of the evacuation information in the pre-movement period, the more people safely evacuaed. To begin with, our added assumptions for the diversity of visitors of Louvre are listed below:

- Considering the international practices and cultural background, evacuation notices are only announced in English and French. Visitors from other countries may have difficulty in understanding the information rapidly.

- Because a large proportion of visitors come from France, The United Kingdom,The United States, Germany, Italy, Brazil and China, we neglect visitors from other countries to simplify our model.

As a result of the diversity of visitors—speaking a variety of languages, evacuation guidance information issued by relevant security departments cannot be quickly and effectively transmitted. We have the following proportion of tourists from different countries[3]:

According to Figure 6, although the visitors from France, the United Kingdom and the United States accounted for nearly 70%, a significant number of tourists come from countries that are not native



Figue 6: The proportion of tourists from different countries

Table 3: Groups classification

	proportion	numbers of people in a group(num)
$Biggroup(>=7)$	13.8%	$num(8, 4^2)$
Small group(2-6)	70%	$num(4, 2^2)$
Single	16.2%	$num = 1$

speakers of French or English. We analyze the effect of the diversity of tourists on the propogation of evacuation information from the following two aspects:

Understanding difficulty Due to the language diversity, different visitors have different levels of understanding on evacuation information. Some visitors may have difficulty in understanding the information quickly. From the individual level, we define U as the Understanding difficulty coefficient:

$$U = \begin{cases} 0, \text{visitors from France,The United States or The United Kingdom} \\ 1, \text{visitors from non-Franch or non-English speaking countries} \end{cases} \tag{3}$$

Communication among evacuees Visitors from the same country will continue to improve their evacuation information through communication during the evacuation process. People communicate and share information among themselves and adjust their actions and behaviors accordingly[4] . We calculate Communication coefficient C as:

$$C = \frac{1}{e^{kp}} \tag{4}$$

where p is the normalized proportion of visitors from different countries. k is a parameter reflecting the influence of communication. We set $k = 5$ in the follwing simulation in this part. The informtion tend to propogate at a fast speed. So the exponential term in this expression can depict the effect of communication.

We use weighted sum to incorporate the two factors, the Evavuation information indicator is defined by:

$$E = t_1 U + t_2 C \tag{5}$$

where $t_1 + t_2 = 1$.

- **Human relations**

When considering groups traveling together in Louvre, the factor of human relations influences evacuees' mental states in evacuation process. Furthermore, previous work shows that the pre-movement time could be significantly influenced by the phenomena of attachment to people (waiting to be reunited with family members and friends)[**?**].

For a visitor, the more people in his group, the more effects of attachment on his premovement time. The groups traveling together in Louvre are roughly classified as follows:

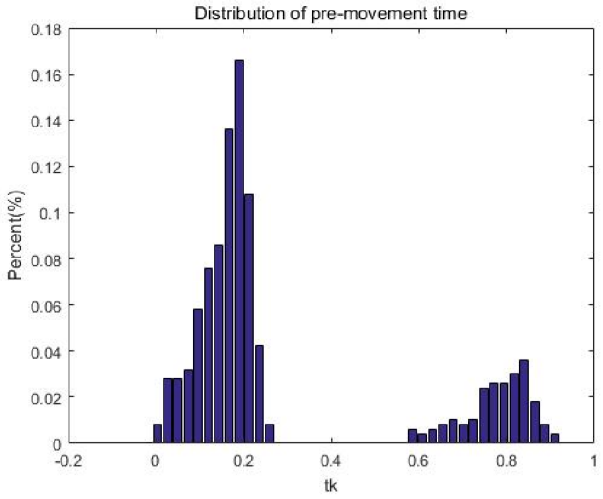
For a visitor, the Human relations factor affecting his pre-movement time in terms of attachment to people is defined as:

$$H = \frac{a \times num}{a \times num + 1} \tag{6}$$

where num presents the numbers of people in his group, a is a parameter.

- **To solve the model**

In this part, we present our simulation results of visitors' pre-movement time coefffi- cien which indicates the pre-movement time span by incorporating the two indicators of Evacuation information and Human relations. We have the coefficient of pre-movement time: $t_k = c_1E + c_2H$, and set $a = 3, t_1 = 0.3, t_2 = 0.7, c_1 = 0.7, c_2 = 0.3$. With 500 people involved in the simulation, Figure 7 shows the distribution of pre-movement time among visitors.



Figue 7: the distribution of pre-movement time

The pre-movement time of tourists from Franch and English speaking countries gather around 0.1, while that of tourists from other countries gather around 0.8. Both of two parts showed a normal distribution trend. It is because the tourists from Franch and English speaking countries have better and more timely understanding of the evacution information due to the language convenience and communication among evacuees. The normal distribution trend of two parts is attributed to the factor of Human relations.

Increased pre-movement time will lengthen evacuation time, so the time that the latest person arrives at the exit will be extended. Museum should try to reduce premovement time.

Movement

In this section, we introduce a velocity analysis model to ananlyze the movement of evacuees in the Louvre. Taking the instantaneous velocity, ecucuees density and collision in the process into consideration, we mainly focus on the movement direction of evacuees under the guidance of museum staffs. We first set some mathematical notations:

- r:Through physical abstraction, we see the individual as a sphere of radius r.
- Δt:a tiny time interval.
- $\rho^{(k)}(x, y)$: the crowd density distribution function at time $k\Delta t$.
- $\overrightarrow{r_i^{(k)}} = (x_i^{(k)}, y_i^{(k)})$:the displacement vector of ecacuee i at time $k\Delta t$,where $x_i^{(k)}, y_i^{(k)}$indicate position coordinates.

- $\overrightarrow{r_L^{(k)}}$:the displacement vector of museum staffs at time $k\Delta t$, who serve as evacuation guides.
- $\overrightarrow{v_i^{(k)}}$:the velocity vector of evacuee i at time $k\Delta t$.

- **The crowd density distribution function**

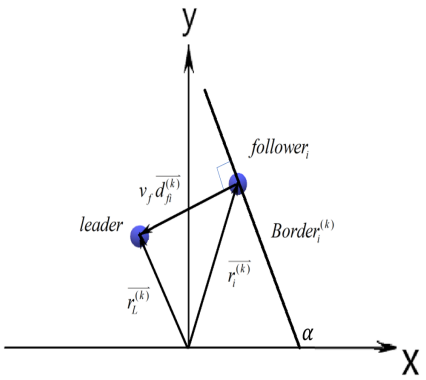
Population density distribution can be considered as continuous change. Based on Spatial Continuous Surface Model[**?**], we get the density distribution function of Louvre. The basic rules is summarized as follows:

- Divide the area into grids;
- Convert the population of the area into population density;
- Place a central point in each area and assign the population density to the center point;
- Use an interpolation method to interpolate the population density at the point into the grid surface.

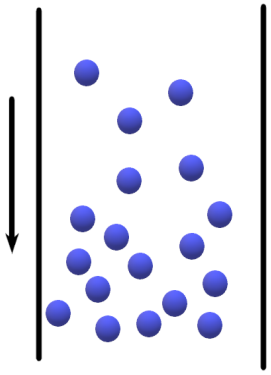
- **Determine the velocity at time $k\Delta t$**

The velocity of evacuee i at $k\Delta t$ is composed of three velocity components. We give the expression and then analyze the components respectively.

$$\overrightarrow{v_i^{(k)}} = v_f \overrightarrow{d_{fi}^{(k)}} + v_c \overrightarrow{d_{ci}^{(k)}} + v_r^{(k)} \overrightarrow{d_{ri}^{(k)}} \tag{7}$$



Figue 8: Spatial positional relations of vectors



Figue 9: positions

- **Following velocity:** $\overrightarrow{v_{fi}} = v_f \overrightarrow{d_{fi}^{(k)}}$,where $v_f$is a constant, and $\overrightarrow{d_{fi}^{(k)}}$is the unit vector in the direction of following velocity. During the evacuation process, the evacuated crowd will move under the guidance of museum officials and have a following velocity component. As Figure 8 shows,

$$\overrightarrow{d_{fi}^{(k)}} = \frac{\overrightarrow{r_L^{(k)}} - \overrightarrow{r_i^{(k)}}}{||\overrightarrow{r_L^{(k)}} - \overrightarrow{r_i^{(k)}}||} \tag{8}$$

- **Corrected velocity:**$\overrightarrow{v_{ci}} = v_c \overrightarrow{d_{ci}^{(k)}}$

Due to the large number of evacuees, it is obviously impossible to move straight toward the guidance. It can be reasonably assumed that the evacuated people will adjust the moving direction according to the surrounding environment while generally maintaining the following trend. They tend to move toward the smaller density area and to avoid walls and obstacles.

Next, we discuss the corrected direction $\overrightarrow{d_{ci}^{(k)}}$of evacuee i located in $\overrightarrow{r_i^{(k)}} = (x_i^{(k)}, y_i^{(k)})$at time $k\Delta t$. Considering the evacuee tend to move in the direction with the fastest decline in density, based on the knowledge of calculus, it is just the negative gradient direction of $\rho^{(k)}(x, y)$in$(x_i^{(k)}, y_i^{(k)})$i.e.$-\nabla\rho(x_i^{(k)}, y_i^{(k)}) = (-\frac{\partial\rho}{\partial x_i^{(k)}}, -\frac{\partial\rho}{\partial x_i^{(k)}})$.Thus we have

$$\overrightarrow{d_{ci}^{(k)}} = \frac{-\nabla\rho(x_i^{(k)}, y_i^{(k)})}{||\nabla\rho(x_i^{(k)}, y_i^{(k)})||} \tag{9}$$

In fact, as shown in Figure 9, in spite of the smaller density behind the crowd, it is obviously impossible for the evacuees to run backwards. Therefore, it is necessary to further optimize the corrected direction $\overrightarrow{d_{ci}^{(k)}}$.

As is shown in Figure 8, make a line $Border_i^{(k)}$ perpendicular to $\overrightarrow{d_{fi}}$through $(x_i^{(k)}, y_i^{(k)})$.We define the angel between $Border_i^{(k)}$ and the positive direction of x-axis as α,and that between $\overrightarrow{d_{ci}^{(k)}}$ and the positive direction of x-axis as θ.The direction with the fastest decline in density can be solved in a constrained optimization problem:

the objecive:the direction with the fastest-decreasing density, which correspond to the corrected velocity.

the constraint:the corrected direction must be roughly consistent with the direction of guidance.

Realized by:

$$\min f(\theta) \tag{10}$$

$$S.T. \begin{cases} f(\theta) = \frac{\partial\rho}{\partial x_i^{(k)}}\cos\theta + \frac{\partial\rho}{\partial y_i^{(k)}}\sin\theta \\ \theta \in [\alpha, \alpha+\pi] \end{cases} \tag{11}$$

- **Random velocity:** $\overrightarrow{v_{ri}} = v_{ri}\overrightarrow{d_{ri}^{(k)}}$. The actual evacuation direction is also affected by many factors such as the psychological state of ecacuees and the surrounding environment. For other complicated factors, we use random velocity to indicate. Herev_{ri}denote random velocity magnitude, v_{ri} $N(0, 0.1)$. The angle between$\overrightarrow{d_{ri}^{(k)}}$ and the positive direction of x-axis is defined asθ', and$\theta' \sim U(0, 2\pi]$.

- ## Movement process

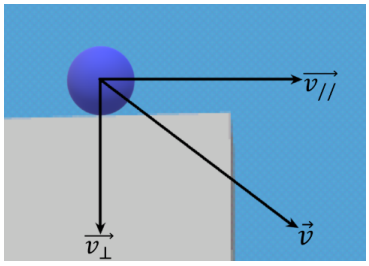
At each time interval of Δt,we can get a corresponding crowd density distribution function$\rho^{(k)}(x, y)$and the velocity $\overrightarrow{v_i^{(k)}}$. However, in reality there are some special situations that cannot be ignored. We need to do some necessary illustration.

- **Situation 1: Lean against the wall:** As is shown in Figure 10. When a person touches a wall (or obstacle) during evacuation, we decompose the velocity $v_i^{(k)}$ into a direction perpendicular to the wall $\overrightarrow{v_\perp}$and parallel to the wall $\overrightarrow{v_{//}}$.When the component in the direction perpendicular to the wall is not 0, he will change the original velocity into $\overrightarrow{v_i^{(k)}} = \overrightarrow{v_{//}}$.

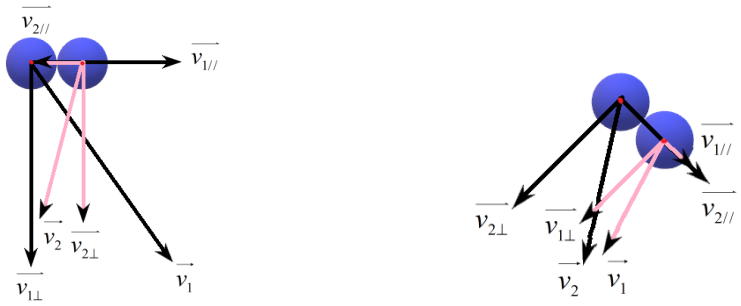
- **Situation 2 : Collision between evacuees:** When two people collide with each other, similarly, we decompose the velocity into two directions: parallel to the centroid connection$\overrightarrow{v_{1//}}, \overrightarrow{v_{2//}}$ and perpendicular to the centroid connection $\overrightarrow{v_{1\perp}}, \overrightarrow{v_{2\perp}}$. This situation is illustrated as follows:

  In collision 1,where the $\overrightarrow{v_{1//}}$and $\overrightarrow{v_{2//}}$hold th opposite direction,the original velocity change into $\overrightarrow{v_1} = \overrightarrow{v_{1\perp}}, \overrightarrow{v_2} = \overrightarrow{v_{2\perp}}$.In collision 2,where the $\overrightarrow{v_{1//}}$ and $\overrightarrow{v_{2//}}$ hold the same direction yet $\overrightarrow{v_{2//}} > \overrightarrow{v_{1//}}$,we set $\overrightarrow{v_{2//}} = \overrightarrow{v_{1//}}$ to correct the original velocity.

In fact, there are many more complicated situations such as collision within several people, we can deal with them in the same way as the above cases: correcting the velocity to meet the reality. Lack of space limits further discussion at this point.



Figue 10: Situation 1:Lean against the wall



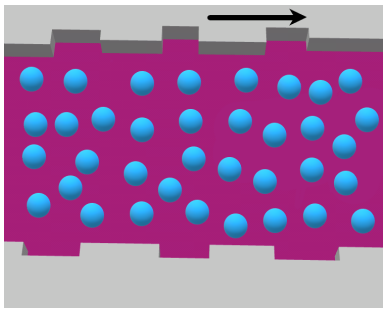
(a) collision 1 (b) collision 2

Figue 11: Situation 2:Collision between evacuees

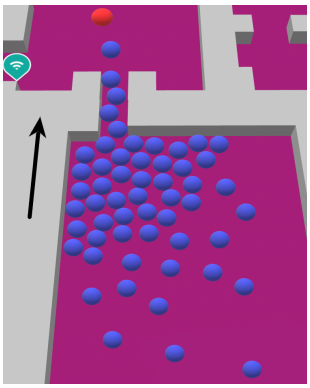
- **Simulation and analysis**

With the model above, we set proper initialization state and ran the simulation. Specifically we set $\Delta t = 1s, v_f = 1.8m \cdot s^{-1}, v_c = 0.5m \cdot s^{-1}$.

At the beginning(time after pre-movement process), the flow of people began to move orderly with the guidance of museum staffs. Figure 12 is a part of passage connecting RICHELIEU WING and SULLY WING on the 1st floor.



Figue 12: a passage connecting RICHELIEU WING and SULLY WING in the 1st floor



Figue 13: Clogging in the narrow escape of SULLY on the 1st floor. Red ball: Museum staffs; Blue ball: evacuees

Aswecansee, thereisasuitable distancebetweentheevacuees, andthecrowddensityisrelativelylowinanunobstructedevacuationprocess. Accordingtooursimulation results, it took about 25s for 500 people to cross the area. However, the orderly flow will transit to disorder in some specific situations.

Figure 13 is a narrow passage in the 1st floor of SULLY WING .The crowd formed a funnel-shape obstruction in this narrow passage. Based on our model, evacuees tend to move to a less dense

direction while generally following the guidance. What' s more, when a evacuee leans against the wall, his velocity remains only the component parallel to the wall. Our model helps to better explain this clogging phenomenon.

With a great amount of evacuees piling up, it' s rather hard for those who got stuck in the middle of the crowd to move(collision 1), so a lot of time would be wasted here. In our simulation results, it took more than 150s for 500 evacuees to cross this narrow escape.

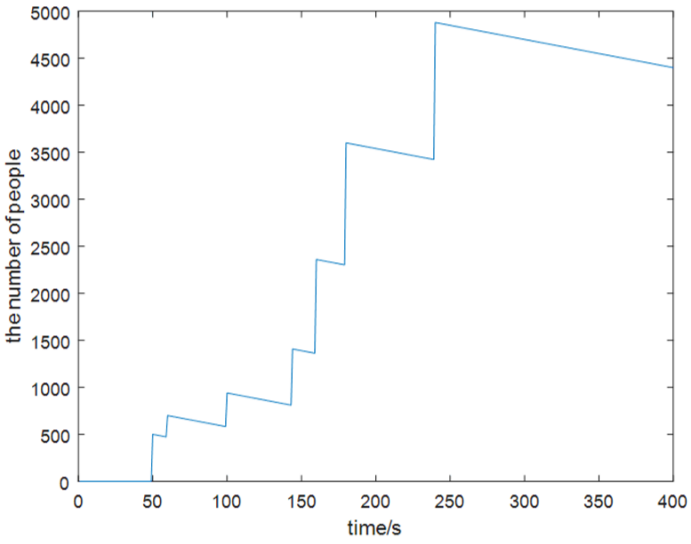
4.3 Potential threats of circumstances model

In addition to the behavior of the evacuated population, potential threats in the evacuation environment can also have an impact on the evacuation process. We improve our previous model by considering environmental factors.

- **The pyramid entrance—a bottleneck**

Bottlenecks refers to places where movement is dramatically slowed or even stopped. Due to space limitations, we take the Pyramid exit as an example.The pyramid exit is the most commonly used entrance and exit. It consists of two revolving doors and one automatic door, while the revolving door rotates at a slow speed, the number of passing evacuees is less at a time, the efficiency is lower during evacuation.

It can be reasonably assumed that when an emergency occur, no one will enter the Louvre through the pyramid entrance. all three doors will be used to evacuate tourists. Based on the revolving door speed and the passing flow, we estimate that the flow rate is 3 people/s, which is much lower than that in the exhibition hall. It may cause people to gather too much at the exit, generating a funnel-shaped blockage i.e. bottlenecks. The results are: it takes a long time of 1860s to evacuate all the visitors at the exit to safe places.



Figue 14: numbers of people at Pyramid entrance

In fact, the evacuation plan in the former Macro emergency evacuation model does not take into account the capacity of nodes, resulting in too many people gathering at the destination node. Moreover, the time spent on congestion at the exit should be involved in the total evacuation time. Taking the node capacity into consideration, we use the same method to find the optimal routes. Define V_i as the

capacity of node i, we get an added constraint:

$$f_{ji} - f_{ij} \leq V_i (i = 1.2.3.4) \tag{12}$$

- **To solve the bottleneck—additional exits utilization**

Based on the improved model, in order to evacuate people gathering at the exit quickly, additional exits near the main exits should be utilized. We assume that additional exits are evenly distributed in Louvre. Because these entrances and exits are not commonly used, security personnel are insufficient and the safety coefficient is low, we need to utilize them cautiously.

We calculate the average evacuees density within a certain area near the Pyramid entrance to determine how to utilize the additional exits to facilitate our evacuation plan.Specifically:

$$\bar{\rho} = \frac{1}{S} \iint_s \rho(x, y) d\sigma \tag{13}$$

we set the threshold: $\rho_t hreshold_1, \rho_t hreshold_2 \cdots \rho_t hreshold_k$.If$\bar{\rho} > \rho_t hreshold_k$,we then utilize k additional exits.Note that we will close the additional doors when the average evacuees density drops below the corresponding threshold.

Figure 15 shows: two additional exits are opened in around 150s, and the increasing trend of population at Pyramid entrance has slowed down compared with the original one, while the number is still rising. And then four exits are opened in about 180s, after which the number of people has dropped significantly. After closing two exits at about 240s, the number of people can still keep a relatively low level. It takes 450s for all people to safely evacuate. Overall, the strategy of utilizing additional exits based on actual conditions not only greatly reduces evacuation time, but also takes into account security issues.

- **Emergency personnel entering model**

Emergency personnel refers to people who help in an emergency, such as guards, fire fighters, medics, etc. Professional emergency personnel can help us effectively prevent potential threats of circumstances. To allow emergency personnel to enter the building as quickly as possible, we need to analyze the opotimal time when they enter the Louvre. Figure 16 illustrate the trend of numbers of people at the four main entrances. Emergency personnel should enter the building from each exit at the time when there are fewer evacuees.

Table 5: optimal entering time of emergency personnel

	optimal time
Exit 1	t<50s
Exit 2	t<160s
Exit 3	t<96s or 312s<t<330s
Exit 4	t<64s or t>320s

5 Policies and Suggestions

From our specific analysis of the evacuation plan, the following recommendations can be drawn:

- Broadcast in multiple languages when broadcasting emergency notifications, in order to make more people understand in time, reducing pre-movement time.
- Because the guests do not know much about the route of the museum, the evacuation plan must be guided by the museum staffs. They can lead guests to escape. What' s more, escape signs can be placed in a conspicuous place to attract peoples attention.

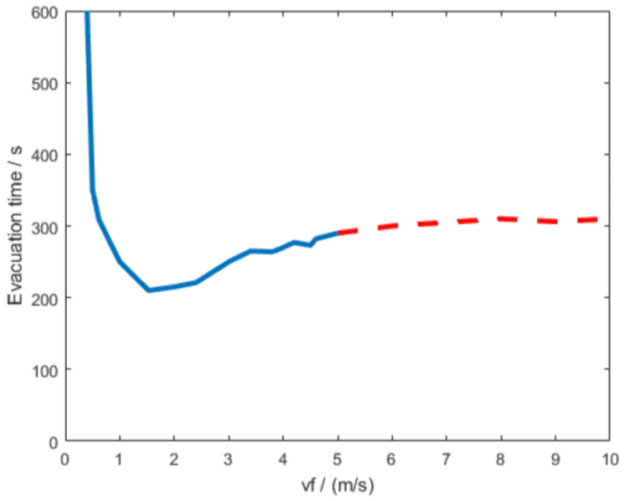
- During the process of evacuation, staffs should maintain an appropriate speed, which is not only beneficial to speed up evacuation but also prevent potential threats like stampede event.
- The Louvre can consider canceling the design of the revolving doors, for they can pass only a few people in unit time. It is not suitable for evacuation. For the same reason, the design of narrow passages should be minimized.
- During evacuation process, the museum should send certain staffs to help people who with disabilities due to their limited mobility and they should take full advantage of fire elevators.

6 Sensitivity Analysis

- **Crowd evacuation behavior model**

In the Crowd evacuation behavior model, we choose identical speed of all followers as$v_f = 1.8m \cdot s^{-1}, v_c = 0.5m \cdot s^{-1}$. Since the speed of all evacuees is a factor in determining the time through a narrow passage, we make a sensitivity analysis about it. Change the speed, the time changes a lot.

We choose different speed to run simulation and the result is shown in Figure 17.



Figue 15: the sensitivity analysis of v_f

Through the results, we can find as the speed increases, the evacuation time becomes significantly smaller firstly and then gradually increases. And the red dotted line indicates that it is virtually impossible to achieve such a large speed.

The reason for this situation is because of faster-is-slower effect[**?**] . Within a certain range, increasing the evacuation speed is conducive to speeding up evacuation, while too fast will play the opposite role. Too fast speed can overcrow the crowd and make it difficult to pass through narrow passages. So museum staffs should try to calm everyone down during the evacuation process.

- **Macro emergency evacuation model**

Potential threats may occur in the path which is of great importance to evacuation. Some threats have the potential to alter optimized route. We make an analysis about potential threats.

If threats occur in key path (e.g. path between node13 and node3), it will change the route to a large extent, because many guests in Richelieu rely on this path to escape. If it is damaged, they have to turn to other exits, which will cause guests in other wing to change routes.

If threats occurs on a channel that is not often used (e.g. path between node15 and node11), then the optimal path does not change greatly because of the threats.

7 Further study

The Macro emergency evacuation model we develop can be adopted and implemented for other large, crowded structures. Application and implementation steps are as follows:

- Determining the clustering of people in the structure and all the path to escape, simplified to a graph.
- Determine the length of time each escape route passes according to the specific details of the structure.
- Use the CCRP algorithm to calculate the optimal evacuation path and the shortest time.

So our adaptive model can apply to other large and crowded structures.

8 Strength and Weaknesses

8.1 Strengths

- Comprehensive:From both macro and micro perspective, we analyze the emergency evacuation plan. Macro evacuation plan help us make optimized route, while micro analysis during the process makes the evacuation plan more detailed and accurate.

8.2 Weaknesses

- Because few data are available, the data we estimated may not accurate enough.
- For the simplified graph of the museum is too simple, we can choose more nodes as a crowd gathering point.
- In crowd evacuation behavior model, we ignore the acceleration of peoples movement.

1 Appendices

Here are simulation programmes we used in our model as follow.

Input matlab source:

```
%Dijkstra Algorithm for calculating the shortest path
function [dis,pa] = Dijk( W,s,d )%W—adjacency matrix
%s—starting point d—destination
n=length(W);%number of nodes
D = W(s,:);
visitflag= ones(1,n); visitflag(s)=0;%determine if the node is accessing
parent = zeros(1,n);%record the previous node of each node
pa =[];
for i=1:n-1
temp = [];
%Starting from the starting point, find the next point of the shortest
%distance, do not repeat the original track every time
for j=1:n
if visitflag(j)
temp =[temp D(j)];
else
temp =[temp inf];
end
end
```

```
[value,index] = min(temp);
visitflag(index) = 0;
%if the index node is smaller from the starting point to the path of each node,
%it is updated, and the predecessor node is recorded, which is convenient for
%backtracking.
for k=1:n
if D(k)>D(index)+W(index,k)
D(k) = D(index)+W(index,k);
parent(k) = index;
end
end
end
dis = D(d);%the shortest distance
%Retrospective method
t = d;
while t~=s && t>0
pa =[t,pa];
p=parent(t);t=p;
end
pa =[s,pa];%the shortest path
end
```