备赛篇：

虽然不知道佬们打算什么时候开始准备！但个人感觉比较粗略地学习完模型并且写出论文（不论好坏，只论完整性）需要 21 天左右。主要分为以下三个阶段：

1、审美建立（4 天）

队员研读学长姐的获奖论文并进行讨论，建立对论文的审美，从自己的位置出发，了解好论文应当是什么样的，为后面针对性的努力定下方向。

*注：此处学长姐论文以本校为佳。本校甚至本专业的学长姐更知道大家的专业水平和缺漏部分。

**注注：美赛论文对于论文手对英文学术性语言的表达有很高的要求，同时对于画图的同学也是很大的挑战~建议论文手和编程手在这块可以拉长战线。

2、各自准备（15~21 天）

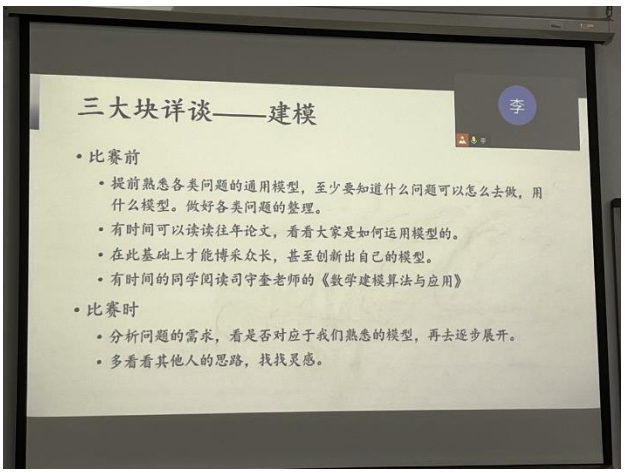
参考 1. 各位置确定自己的技能提升点，在暑假进行针对性的技能提升。以下是各个位置的可能学习重点。

建模手

① 确定三位队员均需要掌握应用场景、使用方法的模型，掌握常见问题的思路、模型

② 了解在比赛过程中有多少新知识需自学，熟悉自学新知识/收集解题思路的主要途径

③ 从学长姐的论文出发，了解如何在做题的过程中将小问们联系到一起，如何在问与问之间实现方法优化



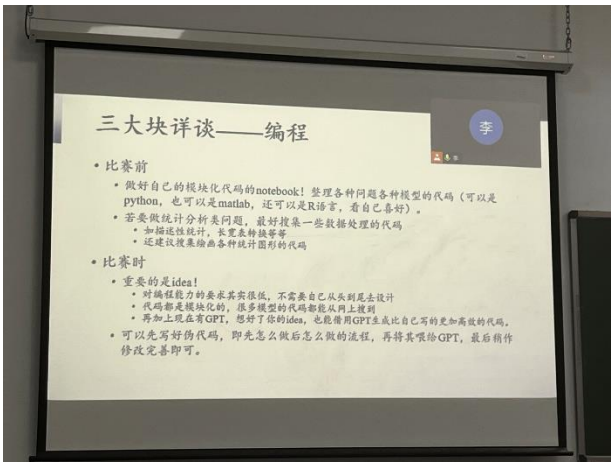
编程手

① 了解、封装常用的数据处理方法，使用往年数据熟悉数据处理操作，确定在比赛时数据处理是否需要多人共同参与

② 对绘图软件 Origin，熟练绘图操作，确定在比赛时绘图是否需要多人共同参与

③ 同建模手沟通，对需掌握应用场景、使用方法的模型整理好模块化代码

④ 熟练借助 GPT 和伪代码完成新思路代码生成、调试的操作

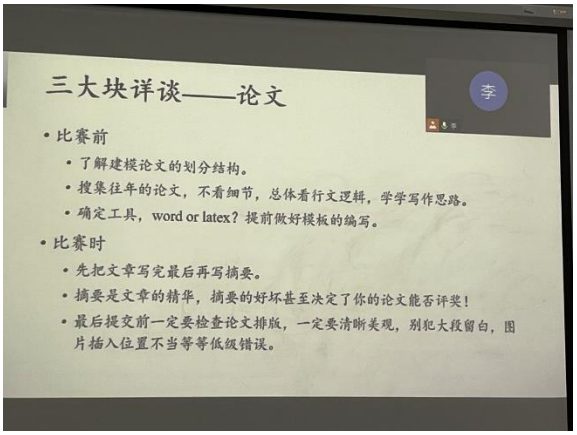


论文手

① 研读学长姐的论文，学习论文结构、论文内容、写作思路

② 熟悉需掌握模型的应用场景、使用方法，练习单个模型的局部写作

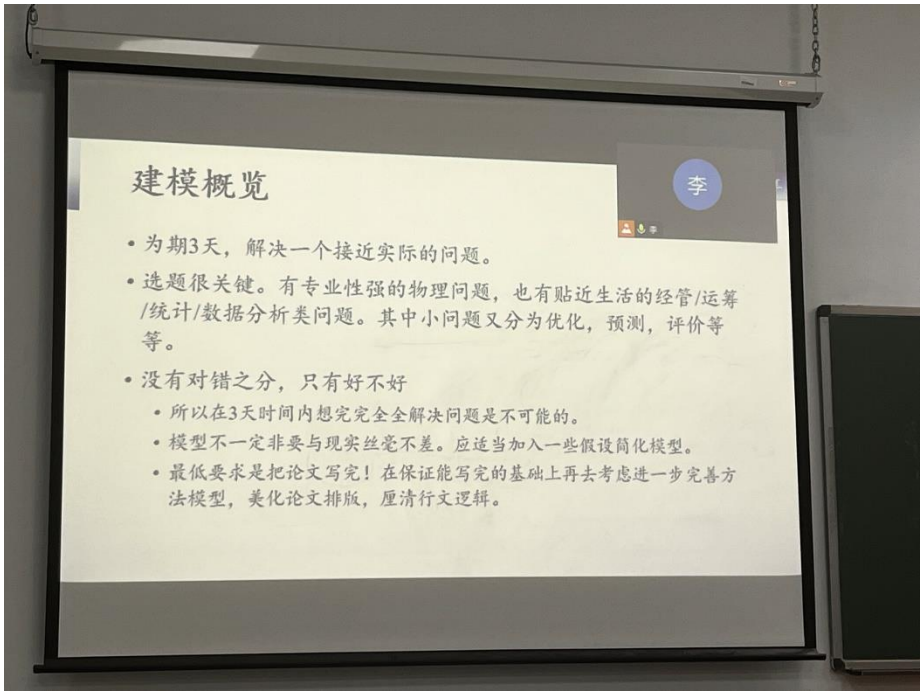
③ 熟悉 word 排版操作，准备论文模板

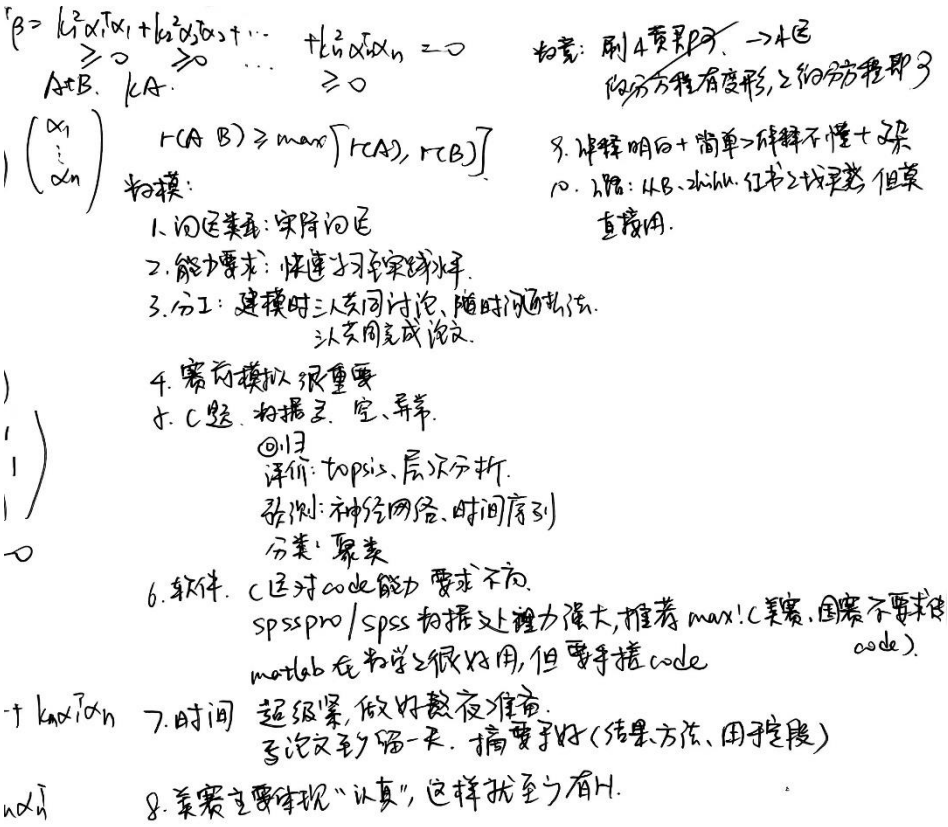


3、模拟训练

找往年题，三人线下，掐时间完成。了解怎么安排时间，怎么控制进度，怎么抓重点，怎么写论文。

将自己的论文同学长姐同年论文进行对比，明确如何将学长姐论文的亮点运用到自己的论文上。





另：针对美赛，图形的美观性大于科研风的要求。所以大家什么 flourish 之类的花花绿绿画图网页＆软件也可以用上（）重要的是！色调统一！整体美观！
然后个人认为复刻往年的国奖论文也非常重要。建模手如何想到这样的思路？编程手有了思路后是否能给出可实现的代码？论文手解释模型的能力如何？这都是需要在模拟、复刻的过程中不断反思的 ww

赛中篇

放在最前面的注意事项：规律作息、不要熬夜！规律作息、不要熬夜！规律作息、不要熬夜！俺们国赛每天 8 点开工，到晚上 10 点结束，最后提早提交。美赛时间更加充裕，所以一定补药熬夜哇！
在比赛的过程中注意相互补位。建模手和编程手的界限其实本来就不是很明显（？），在建模手思路可能卡顿时，编程手也需要及时出来参与讨论；在数据处理出现问题时，建模手和论文手也需要帮忙（毕竟数据是一切的基础..）总之最重要的就是一体同心，不因为队友的错误/疏漏而埋怨，要相信一次比赛下来最重要的是收获了两位共生死的伙伴(
然后在比赛前可以拟出一个（理想化的）timeline，尽量遵守，但实在超时也没有关系！要相信人在 ddl 之前的潜力是无限的！

赛后篇

复盘工作每次模拟比赛后都要进行。（最好在赛后三天内）
然后！
吃饭睡觉放松！

附录

1、数学模型整理（贫弱经管类学生版）

这个是咱们的模型索引！有想法的时候用 WPS 查找就可以！

1. （定性）变量：定类、定序：有分类变量，前者无序（性别国籍），后者有序（优良中差），无法计算
   定距、定比：有分类，偏定量，前者可加减、后者加减乘除都可

注意：定距变量没有绝对零点。
定量变量在数据清洗阶段分为离散、连续型。使用定距（定比）直方图/箱线图
需要着重了解对“对象”的分类，从而剔除某些“不合群”的情况（单独讨论 or 删除），也算是假设的一部分。

2. 标准差：反映数据离散程度，值越高越离散
3. 标准误：标准差是单次抽样得到的，用单次抽样得到的标准差可以估计多次抽样才能得到的标准误差
4. 统计性 t 值：反映样本是否具有代表性，样本量是否足够大。正常区间：-3 到+3

p 值直接与显著性水平相关，而 t 值!! 一定需要换算成 p 值才能判断

5. 显著性水平是在统计假设检验中使用的一个指标，与数据无关，一般用 α 来表示，一般是 0.05。如果与数据直接相关的 p 值小于 α 值，即原假设成立的条件下获得当前结果的概率低于依靠现有数据推翻原假设的出错概率，则推翻原假设。
6. p 值本质上是连续的概率值，准确来说是「零假设成立的条件下获得当前结果的概率」
7. 优化类问题
8. 基础知识：递归——将一个母问题拆分成若干个同类的更简单的子问题，不断拆分直到遇到无需拆分的基本情况，完成求解
9. 动态规划

   底层逻辑：从基本问题解起，一步一步建构层级更高的问题，直到求得设问层级问题的（最优）解（循环迭代而非递归，减少了拆解问题的时间复杂度；无需检查这个子问题是否计算过，进一步降低时间复杂度）

   递归转化为递推，算法部分参见：教你彻底学会动态规划——入门篇-CSDN 博客

   应用条件：①最优子结构问题——一个问题如果想最优，那这个最优和你如何求解它的子问题相关②重叠子问题——子问题们相互勾连，解决一个字问题涉及到更多子问题的求解　e.g.线性问题——切割钢条令收益 max

   应用效果：①可找到全局最优解

   无后效性。
10. 贪心算法，由于其约束条件较多，注意其合理性的论证

    底层逻辑：整体最优=在每一步都做到当下的最优

    应用条件：①最优子结构问题②贪心选择性质：局部最优求和==全局最优③无后效性：后面的不影响前面

    应用效果：①在贪心选择性质成立时，可找到全局最优解
11. 线性规划:应用条件：①约束条件和目标函数必须是线性函数
12. 多目标规划问题
13. 问题特征：①有多个目标需要实现，目标之间往往相互矛盾　e.g.工资和闲暇②存在约束条件——解必须在某个范围中③注意可比性——不同量纲的目标如何统一
14. 常用思想：①帕累托最优②权衡与协调（罚款 or 直接约束）③遗传算法（全局寻优）
15. 遗传算法
16. 底层逻辑：进化论——适者生存和繁殖的概率大于不适者，由交换和突变保证基因多样性
17. 问题特征：①全局复杂问题——这块都用暴力随机生成和比较产生最优解了②解空间范围大
18. 投影寻踪：一种用于高维复杂数据的算法，通过将高维数据投影到低维，从而发现最有意思的结构，再进行研究（与 PCA 等方法相比，更能保留非线性结构。当然这需要我们对数据的结构有一定了解）
19. 优化/规划模型：用于解决决策问题，由决策变量（派多少人）、目标函数（要求什么最大化）、约束条件（正整数个人，最多几个人）组成。主要包括：多目标规划、整数规划（0-1），归一化与正则化，复杂网络优化，排队论

优化和规划模型具体思想和写作格式可参考 21C283 论文

20. 多目标规划

    前置知识：帕累托最优：无法在不损害他人的情况下提升自己的状态

    　　帕累托改进：在不损害他人的情况下提升自己的过程

    应用场景：有多个目标函数，希望同时取最值。然而，取各个函数的最值是互相妨碍的。需要找到一个最优解 x，让这个解相较其他解，在目标函数 A、B 上都更优；或找到一群

满意解，他们彼此在 A、B 上各有优势，无法在其中找到最优解 x。e.g.效率与平等

应用原理：①线性加权，把 f1、f2……合成 F 函数，转化成单目标规划

②优先级法，确定各个函数优先级，先确定优先级最高的函数的最优解，再在这个最优解的基础上，寻找次优先函数的最优解，以此类推

③理想点法，让最终解到每个函数最优解的距离之和最小

④人工智能求解

局限性：停留在满意解上，难以找到最优解

21. 0-1 整数规划

前置知识：整数规划：决策变量被限制为整数时的规划问题

整数规划的解决算法：分支定界法、割平面法、迭代改进算法

应用场景：当决策变量仅有 0 或 1 两种形式时，需要应用 0-1 整数规划 e.g.背包问题（物品唯一，只可能放 or 不放）、指派问题（员工唯一，只可能派 or 不派）

应用原理：分支定界法、割平面法、迭代改进算法

局限性：没有一种方法能有效地求解一切整数规划

22. 归一化

定义：用特定线性变化（lg 函数、最大最小值等），将数据范围缩小到为 0 到 1 之间，

作用：消除量纲对数据的影响，同时不改变数据的分布。像是“拍扁”

标准化

定义：用某种概率统计上的方法，令数据形成均值为 0，方差为 1 的正态分布

作用：无法固定数据范围，但改变了数据分布

正则化：在算法中加入先验条件，从而防止过拟合发生

另：觉得这一块是数据处理、机器学习方面的内容。有人把归一化、正则化归到优化模型里，可能是想用计算机的方法解释数学问题

23. 排队论

应用场景：通过概率理论解决排队过程中的调度和规划问题

应用原理：确定(1)顾客的抵达分布情况 (2) 服务台的服务情况 (3) 排队原则 (4) 系统容纳量 (5) 服务台数量 (6) 服务流程数量后，通过算法进行暴算

24. 拟合度检验

25. 是对已制作好的预测模型进行检验，比较它们的预测结果与实际发生情况的吻合程度。通常是对数个预测模型同时进行检验，选其拟合度较好的进行试用。

26. 拟合度 r2 表示模型多大程度上拟合了数据。取值范围 0 到 1，1 的拟合度最高

27. F 值：看 a、b 相关性，a、b 方差相除，越大越相关（多运用于线性约束条件）

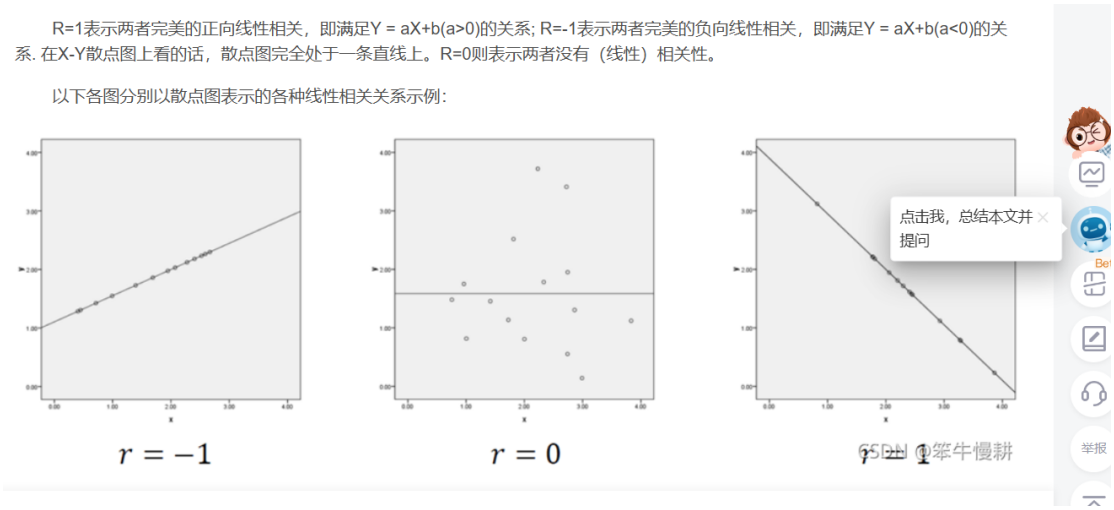
28. AIC 值和 BIC 值用于对比两个模型的优劣时使用，此两个值均为越小越好

29. 相关性分析

看变量 X、Y 的相关性如何。在此，沿用显著性水平、P 值的判断方式

基础方法：Pearson 系数法、Spearman 系数法、Kendall 系数法

Pearson 系数：用于进行线性相关分析，最常用，需要数据满足正态分布，定量数据。两个连续变量间呈线性相关时，使用 Pearson 积差相关系数。由于其是在原始数据的方差和协方差基础上计算得到，所以对**离群值**比较敏感。即使 pearson 相关系数为 0，也只能说明变量之间不存在线性相关，但仍有可能存在曲线相关。



另外两种统计量：不满足积差相关分析的适用条件时，使用 Spearman 秩或者 kendall 相关系数来描述，他们都是更为一般性的非参数方法，对离群值的敏感度较低，因而也更具有耐受性，度量的主要是等级变量之间的联系。

Speraman：考察两个定类变量单调性之间的联系

在不一定正态分布、非线性关系、一个变量为序数型（第一第二第三第四）、变量内有离群值时可用 spearman，检验效能较 Pearson 系数低

Kendall:定类变量、序数型、样本量小、必须有序或连续至少满足一个、单调关系必须有且只有一种

30. 卡方检验

用于无序分类变量独立性/相关性检验 e.g.22C 的纹饰、颜色，p《0.05 就拒绝原假设，推出相关（原假设是不相关）

Pearson 卡方检验：总样本数》40，自由度 df>1 时，期望次数小于 5 的字段《=总字段的 20%

Yates 校正卡方检验：总样本数》40 不满足 pearson 卡方检验期望次数时就用 yates 校正卡方检验

Fisher 精确卡方检验：样本总量小于 40，或任何格子出现期望频数 T<1，或检验所得的 P 值接近于检验水准α

31. 评价模型
32. 按照目标，确定准则及其权重，进而给各个方案打分的模型

主要包括：层次分析、topsis、模糊综合评价、秩和比

33. 确定权重的方法：上网查找别的研究报告，发问卷做调查，找专家赋权，层次分析法
34. 层次分析：

应用场景：不知道各准则的权重，没有各个方案的数据；

应用方法：①给各准则两两比较，出判断矩阵，通过一致性检验，出一致性矩阵，根据一致性矩阵算各准则权重②方案两两比较，看各方案在每个准则下的权重③综合①②，给各个方案打分

局限性：①准则不能太多，否则判断矩阵难以通过一致性检验②你的判断不一定有代表性

35. TOPSIS 分析：

应用场景：有各个方案的数据

应用方法：①数据正向化②数据标准化③用出现的数据拼凑出最优解和最劣解④计算各个方案同最优解、最劣解的距离，将方案排序

局限性：我们默认所有指标对最终打分的重要程度是相同的，也就是他们的权重相同，但在现实中，这不一定能实现

36. 熵权法

应用场景：一种通过样本各指标下数据变异程度，确定评价指标权重，进而对方案进行评价的方法

应用方法：认为变异程度（波动程度）—信息量—权重呈正相关关系，故指标的变异程度越小（概率越高），所反映的现有信息量也越少，其对应的权值也越低

局限性：只从数据出发，不考虑问题的实际背景，确定权重时就可能出现与常识相悖的情况。以至于评分的时候，也会出现问题

37. 模糊综合评价

应用场景：需要将各个元素向模糊集合归类

应用方法：用某种方法整出来函数，从而算出各个元素对各个集合的隶属度，从而判定各元素属于哪个集合

38. **预测模型**

描述过去、分析规律、预测未来，此模型通过对历史数据的学习，估计新数据的数值

主要包括：时间序列分解模型、指数平滑模型、ARIMA 模型、计量

39. 时间序列分析相关基本概念

时间序列数据：同一观察对象在不同时间，连续观察所取得的数据

由时间要素和数值要素（测量对象的具体情况）组成；

分为时点时间序列（一点时间，加总无意义 e.g.温度）和时期时间序列（一段时间，加总有意义 e.g.GDP）

40. 时间序列分解模型

基本概念--时间序列分解：

① 季节趋势 S，季节（季、月、年）转变让数据指标发生周期性变化

② 长期趋势 T，统计指标在相当长一段时间内，受到长期趋势影响因素的影响，表现出持续上升 or 持续下降的趋势

③ 循环变动 C，经济周期

④ 不规则变动 I（白噪声），后面说

⑤ 指标数值的最终变动 Y：

可能是叠加模型 Y=S+T+C+I，可能是乘积模型 Y=STCI

模型选取：季节波动恒定：叠加模型；季节波动变动：乘积模型；无季节波动：都可

应用场景：数据有周期性（分月数据 or 分季度数据），包括长期趋势、季节变动、循环变动

应用方法：先画时间序列图，再判断时间序列包含的变动成分，再用时间序列分解成 S、T、C、I，再建时间序列分析模型，最后预测未来的指标数值

局限性：数据需要有周期性，且不能以年为周期

41. 指数平滑模型：具体选哪个交给 spss

① 简单指数平滑模型

应用场景：不含趋势和季节成分的数据

原理：通过对前 n-1 个数据进行加 α 权平均，从而整出来第 n 个数据的预测值

局限性：只能预测未来一期的数据

② 线性趋势模型

应用场景：数据具有线性趋势，且不含季节成分

原理：水平平滑方程+趋势平滑方程+预测方程

③ 阻尼趋势模型

应用场景：线性趋势逐渐减弱，且不含季节成分；可解决②对未来预测趋势过高的问题！在实际中更常用

原理：预测方程中加入了阻尼参数

④ 简单季节性模型

应用场景：含有稳定的季节成分，不含趋势

⑤ 温特加法模型

应用场景：含有稳定的季节成分和趋势

⑥ 温特乘法模型

应用场景：含有不稳定的季节成分和趋势（23C055，真的很想知道他们怎么用专家预测找到那些合适拟合度时序模型）

42. ARIMA 模型前置知识

【平稳时间序列：①均值为固定常数②方差存在且为常数③协方差只和间隔 s 有关，和 t 无关，方便建模。不平稳时间序列变形后可转化为平稳时间序列

白噪声序列：是平稳时间序列的特例，均值、协方差为 0，方差存在且为常数

差分方程：把某时间序列变量表示为该变量的滞后项、时间和其他变量的函数，这样的一个函数方程被称为差分方程（也就是将 Yt 用 Yt-1、Yt-2 等项表示）

差分方程的齐次部分：只包括 Yt 和 Y 的部分

滞后算子：一个符号 L^i，表示将 Yt 滞后，输出 Yt-i】

AR 模型，自回归模型，捕获具有较长历史趋势的数据的趋势，并据此进行预测。难处理有临时、突发的变化或者噪声较大的数据。忽略了现实情况的复杂性、忽略了真正影响标签的因子带来的不可预料的影响。

MA 模型，处理那些有临时、突发的变化或者噪声较大的时间序列数据。无法处理历史趋势对时间序列数据的影响

ARIMA 模型利用数据本身的历史信息（标签值+偶然事件）来预测未来。一个时间点上的标签值既受过去一段时间内的标签值影响，也受过去一段时间内的偶然事件的影响

ARIMA 模型假设：标签值是围绕着时间的大趋势而波动的，其中趋势是受历史标签影响构成的，波动是受一段时间内的偶然事件影响构成的，且大趋势本身不一定是稳定的

43. 灰色关联分析

应用场景：进行系统分析，判断影响系统发展的因素的重要性。第二个作用就是用于综合评价问题，给出研究对象或者方案的优劣排名

应用方法：通过计算，搞出来各个自变量的图像和因变量图像的相似度，挑选图像最像因变量的自变量，作为因变量的主要影响因素。生成有较强规律性的数据序列，然后建立相应的微分方程模型，从而预测事物未来发展趋势的状况。

局限性：适用于少量数据、短期数据。（但也不适应过于少的数据环境）

Matlab 代码调用：【数学建模】灰色关联分析 +Matlab 代码实现_灰色关联分析 matlab 代码-CSDN 博客

计量经济分析检验模型

44. 似然比检验：反映检测结果的特异性和灵敏度。阳性似然比是特异性，越大越好；阴性似然比是灵敏度，越小越好

注意：如果设定约束条件一样，则两个模型给出的似然函数值应该近似相等。

同时注意 BIC 对模型参数惩罚较多，更倾向于选择参数少的模型。

45. Wald 检验是一种常用的假设检验方法，用于检验一个连续型变量的均值的显著性
    Wald 检验的结果通常会以统计量及其对应的 p 值的形式呈现
    先确定显著性水平：在判断 Wald 检验结果时，需要设定一个显著性水平，通常为 0.05。如果计算出的 p 值小于显著性水平，则可以拒绝原假设。
    再比较 p 值和显著性水平：如果 p 值小于显著性水平，则可以拒绝原假设；如果 p 值大于等于显著性水平，则不能拒绝原假设。
    最后注意假设检验的问题：在进行 Wald 检验时，需要注意假设检验的问题，即需要明确原假设和备择假设。通常，原假设是总体均值等于某个特定的值，备择假设是总体均值不等于该值

优点：只需估计无约束一个模型，适用于线性与非线性条件的检验

计量经济分析：计量经济学中的三大检验（LR， Wald， LM） _wald 检验-CSDN 博客

46. 一致性检验的含义用于确定构建的判断矩阵是否存在逻辑问题
47. 线性回归：用线性方程拟合自变量 abcde…和因变量 X 之间的关系，通过线性回归方程预测 X 的变化趋势
    多元线性回归：相较简单线性回归有更多自变量，且假设自变量之间不存在线性关系
    回归系数：表示 a 对 b 的影响程度，回归系数绝对值越大，影响程度越深
    拒绝回归系数为 0 的原假设：a、b 有相关性；接受~：a、b 无明显相关性
    共线性分析：看看模型的自变量们是否存在很强的线性关系，排除线性关系，提升模型的预测准确度。
    衡量多重共线性的指标是 VIF，VIF 越接近于 1，多重共线性越低，模型越好
48. 逻辑回归：探究一组自变量对分类因变量的影响程度
    过采样和欠采样：当数据出现类别不平衡时，模型可能会投机取巧以保证所谓的“准确度”（e.g.猫狗）。为了避免此问题，对占大头的数据欠采样降低占比，对占小头的数据过采样提升占比
49. 自由度是指当以样本的统计量来估计总体的参数时，样本中独立或能自由变化的数据的个数，称为该统计量的自由度
50. OR 值用于衡量两个分类变量之间的关联强度。具体而言，OR 值表示在某个分类变量上发生事件的风险，与另一个分类变量相比。其值为 1 表示两个分类变量之间没有关联，大于 1 表示一个分类变量对另一个分类变量的发生有影响，小于 1 表示一个分类变量对另一个分类变量的发生有保护作用
51. ● 准确率：预测正确样本占总样本的比例，准确率越大越好。
    ● 召回率：实际为正样本的结果中，预测为正样本的比例，召回率越大越好。--灵敏度
    ● 精确率：预测出来为正样本的结果中，实际为正样本的比例，精确率越大越好。--特异性
    ● F1：精确率和召回率的调和平均，精确率和召回率是互相影响的，虽然两者都高是一种期望的理想情况，然而实际中常常是精确率高、召回率就低，或者召回率低、但精确率高。若需要兼顾两者，那么就可以用 F1 指标。
    ● AUC：AUC 值越接近 1 说明分类效果越好。
52. 微分方程
    基本定义：含有导数、微分的方程
    阶数：微分方程中含有的导数 or 微分的最高阶
    特解：不含任意常数的解；通解：含任意常数的解

初值条件：帮助确认特解的条件
应用场景：人口预测、捕食者猎物、种群相互竞争、传染病/信息传播模型

> **SEIS 模型**：易感 - 暴露 - 感染 - 不免疫
>
> **SEIR 模型**：易感 - 暴露 - 感染 - 移出（免疫）
>
> **SIRS 模型**：易感 - 感染 - 短时免疫 - 易感

应用方法：通过专业知识 or 套用模型，列出微分方程，利用软件求函数解析式和对应函数值
局限性：离散型问题没法用

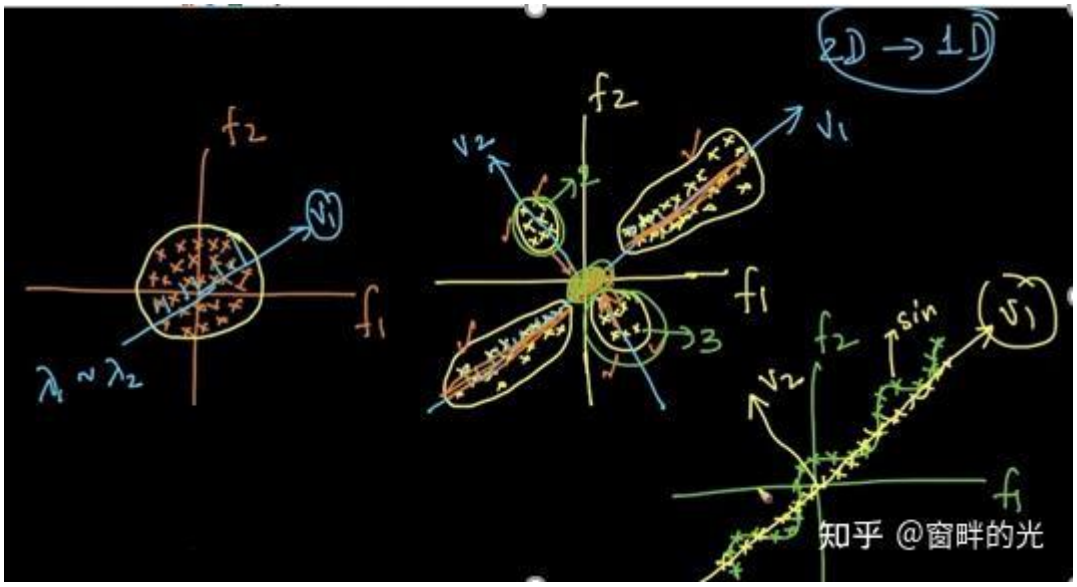
53. *元胞自动机（不常用、难度较大）
应用场景：生态系统动态变化的模拟、动物群体行为的模拟、生物群落的扩散模拟；经济危机的形成与爆发过程、非线性经济学研究、个人行为的社会性、服装流行色的形成、城市空间和交通流的复杂性、传染病的传播过程
应用原理：将二维平面分为若干个小个体，各个小个体遵循同一规则进行演化，各个小个体未来的状态同周围小个体的状态相关
局限性：

1. 元胞形态：在标准元胞自动机中，元胞具备规则一致的形状，有规律地在元胞空间中排列。然而，在现实世界中很少有如此规则的状态。
2. 元胞空间的几何形状：在标准元胞自动机中，二维元胞空间可按照三角形、四边形、六边形等几种网格排列。然而，三角形网格在计算机显示与表达时困难，需要转变成四方网格；四方网格不能较好地模拟各向同性现象；六边形网格能较好地模拟各向同性现象，模型更加自然而真实，但是表达和显示上较困难、复杂。
3. 元胞邻域的定义：在一维元胞自动机中通常以半径r来确定邻域，在距离某个元胞r内的所有元胞被认为是该元胞的邻域。二维元胞自动机通常以规则的空间单元划分，按照 von Neuman型、Moore型和 Margolus型等方法进行处理。这直接导致元胞状态更新规则不能应用于更远的单元。
4. 应用领域的限制：元胞自动机虽然能够模拟一些复杂的系统，但它对于处理的问题仍有一些限制。例如，对于一些高度复杂的系统，元胞自动机可能无法提供足够的精度和复杂性。此外，对于一些动态变化非常快或者具有高度非线性的系统，元胞自动机可能无法捕捉到其动态行为。
5. 参数选择困难：元胞自动机的行为强烈依赖于参数的选择，如邻域类型、状态转换规则等。这些参数的选择可能会对模拟结果产生重大影响，因此需要仔细考虑和调整。

54. 统计模型
55. 主成分分析 PCA：
应用场景：给统计数据降维
应用原理：假设某对象 A 对应两数据（x1,x2）,则 A 为二维数据。通过改变坐标轴（利用线代，确定坐标轴到底改成什么样），让在新坐标轴下 ，$x2$接近 0，$x2$中蕴含的信息也大部分全集中到主元 $x1$上，最后，可在尽量保存 $x2$信息的前提下，将

$x2$删去，对象 A 仅对应一数据 $x1$，完成数据降维

局限性：



1. 第一幅圆形图中 $\lambda_1 \approx \lambda_2$ ，如果直接舍弃某一个特征向量，会直接丢失一半的原始信息。
2. 第二幅图中，明显有四个分类，如果直接映射到最大方差方向上，反而其中两个分类的信息更加模糊不清。
3. 第三幅图中，映射在（ V_1 ）后，呈现均匀分布，看不出规律信息。

56. 因子分析

应用场景：类似主成分分析，但好像比主成分分析更优越。在需要数据降维时采用即可——多特征多变量捆绑打包成单一特征

57. 聚类

应用场景：将数据的描述对象分类、分组，e.g.淘宝通过聚类算法，将用户分为“一老家庭”“新晋父母”

应用原理：根据数据类型，选择计算数据距离的方法。先计算数据之间的距离，再按照网格/密度/层次/1k 均值等方式将数据归类，再将数据对应的描述对象进行分类

监督：给你一些有标签的数据，让你学习有标签数据，对无标签数据进行分类，或给你有标签数据，让你搞出来分类依据

SVM：

应用场景：二分类且高维度且数据量小

决策树：

应用场景：能够处理二分类和多分类问题，数据集有分类特征或连续特征，可以很好地处理数据中的非线性关系，能够处理缺失值并且不需要特征缩放，解释性强

58. R 聚类：软聚类，一个数据可以属于多个组

    层次聚类：自下而上、树形建构

    Kmeans 聚类：先随机搞几个质心，然后把剩余数据按照到质心的距离分类

    Dbscan 聚类：根据密度聚类

59. Q 聚类：硬聚类，一个数据只能属于一个组

60. 辨别分析

    应用场景：对目前已知分类的数据建立判别函数，将判别函数应用于新数据上，对新数据进行数据分类

    应用原理：主要有 fisher，协方差，贝叶斯判别法

61. 相关分析

    应用场景：判断随机变量 A、B 相关方向、相关强度的一种方法

    应用原理：根据数据类型，选择 Pearson、Spearman 等分析模式，判断数据相关方向、相关强度

62. 方差分析

    应用场景：研究观察对象 A、B 的某个对象的指标是否有显著差异 e.g.看大学老师和中

学老师的工资是否有显著差异

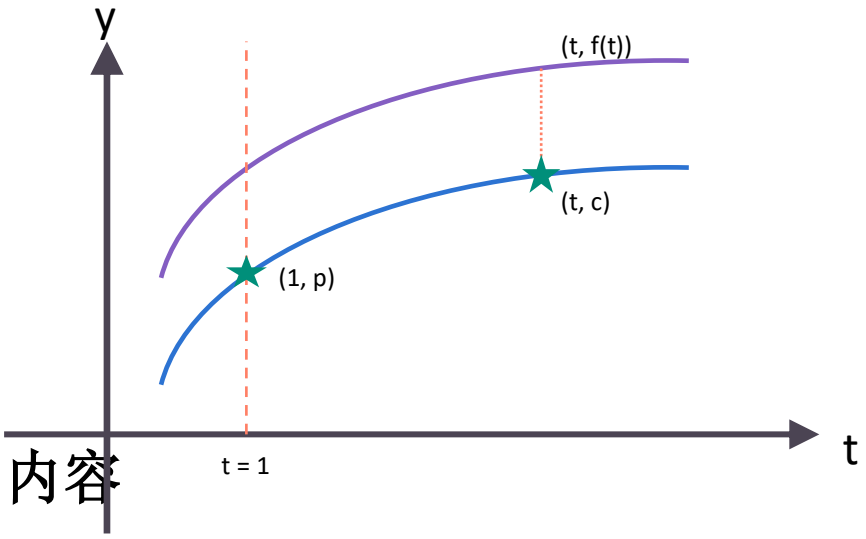
附录 2 论文手的要点（笔记）

排版知识点

CTRL 快捷键类：分页符 全选 选一句 恢复上一步操作 刷新（CTRL + A + Fn + f9）
ALT 快捷键类：先按 ALT，然后按照提示标签操作
放大镜类：显示编辑细节
表格记得关闭缩进
公式表格边框是白色，要用域编号，从头到尾都复制上一个表格
行内公式勿分行
行距设置为固定 18pounds

排版 tips:

① 准备好样式，中宋体，数字和英文新罗马
② 设定行距，标题和标题之间的行距>标题内部的行距
③ 表的标题和表之间一定要留够距离，表格前段段后空 0.5 行
④ 行内公式编辑器如何不扩大行距？
⑤ ppt 矢量图



题目：基于 A 方法/数据的 B 问题

摘要：针对问题一，本文首先对数据进行预处理，对某表中的数据进行某种处理以达到某个目的。对于第一小问题目要求做什么，本文对某些数据采用什么方法（并用什么反映结果）。对于某小文，本文通过什么方法令数据如何，并基于什么样的数据，通过什么方法建立什么模型以做了是什么。

针对问题二，第一小问要求什么，本文引入什么，求解结果发现某些指标有什么特征，说明模型性能优良。第二小问要求，本文采取什么方法。并通过在扰动范围内随意重新赋值，扰动范围处于某范围中，说明此模型敏感性良好。

针对问题三，题目要求什么，本文基于问题一/二的什么，对什么数据使用问题一/二中的什么，并对数据进行敏感度分析，结果表明在准确率和敏感性上，本文均有不错表现。

在解题以外可写模型的现实意义

关键词：数据处理方法、模型所属大类方法 e.g.写卡方检验，不写 Pearson 卡方检验或 Yates 校正卡方检验、4-6 个为宜

问题重述：有①问题背景②问题方法两个部分，在原题基础上做有限调整，避免查重。

问题分析：基本上是把摘要对应的部分写一遍。把结果描述删掉。

模型假设：①题目假设②对小概率事件的排除和对环境的假设③对无关变量的排除④模型要求的假设⑤对特定因果关系的排除⑥定义的明确⑦数据空或 0 的原因注意，只写假设但不进行描述

符号说明：只放重要变量，注意三线表行距统一，重要变量在文章中首次出现时也要做解释。

模型的建立与求解：

数据预处理：①剔无效数据（由题可知，数据应当有什么特征，哪些数据不符合这些特征，故予以剔除）②数据变换（由题可知，本题数据处于什么空间，由于此空间需满足什么，因此针对普通书觉得传统统计学分析方法对于本题数据不再适用。由文献，可知在什么空间上分析有什么问题：，基于上述问题，本文用什么变换处理数据，经过什么变幻的数据可以如何，数据计算公式如下所示）③为更好进行什么分析，本文分别对什么表单中的什么指标进行量化处理，量化处理结果如表 1 所示④空缺数据处理（原因+处理方法）⑤可以将具体数据在支撑材料中体现

问题与模型：基本内容为①题目要求+方法的介绍+方法与问题的结合②检查方法的前提条件③公式与必要解释④使用表格和图像展示结果⑤结果的分析和解释⑥可以单起一个小标题做结果的总结；求解可用 step123 三级标题用完后可用（1）（2）（3）；卡方检验：提出假设，构建卡方检验统计量X^2，计算出X^2的自由度。

灵敏度分析：从 22C155 看不出来咋写。读其他文章时注意点。

模型优缺点：

优点：①本文通过合理假设简化问题，其模型准确描述、巧妙解决了问题②本文内程序在已有问题规模下，时间复杂度、空间复杂度均较低，程序运行时间短，占用空间小③本文充分考虑到数据什么特点，对原始数据采用什么变换，变换后的数据有什么优点④本文采用随机扰动的方法对模型进行敏感性分析，可以使模型评价更客观⑤本文充分考虑到变量间的相关性，使用什么方法对原始数据进行将为，可以使什么更加客观⑥本文的什么模型对样本量的多少、样本量有无规律同样适用，计算量小，较为方便，且不会出现定量分析结果同定性分析结果相悖的情况

缺点：①本文使用的模型在数据量小时表现良好，但缺少大数据验证，延展性无法保证②本文的关系探究式未考虑定量变量与定性变量间的潜在联系

附录：分 ABCD，问题对应的运算代码和 excel spss 的交互式命令

支撑材料：在附录的基础上添加较长的（中间）结果图表，附录和支撑材料有什么关系

结果展示形式：

相关性分析/描述性统计/预测结果（点状）：三线表/折线图/热力图
占比：饼图
连续性数据散布范围与中心位置：箱线图
聚类：树状图/三线表
决策树：类似树状图
偏最小二乘判别分析：散点图