**底稿：**

**0、假设：数据假设&模型假设**

**符号：另外处理**

**一、数据处理&模型简介&状态的描述**

数据分为两类：一类描述具体比赛得分状态，一类描述本球技术类型

分类展示（）

**模型解释：（假设里会有公式，待补充）**

我们认为，选手的技术（含体力等因素）与心态会影响到对局得分

为了利于分析，我们在技术层面分析所有的原始技术变量，即心态层面只涉及到对于比分的影响

这两个因素都会随着时间而变化，我们假设其满足马尔可夫性假设。

而这两者是“得分”这一观察变量的隐变量

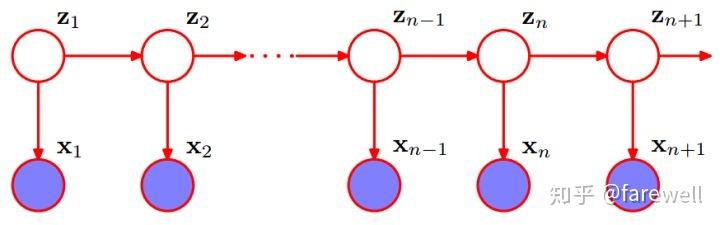
假设2观测独立性假设即任意时刻的观测状态只仅仅依赖于当前时刻的隐藏状态。

即得分仅依靠于此时的技术与心态

则将随机变量作为结点，以两个随机变量相关或者不独立连边，构成一个网络。则（网络有特殊性，待补充）。

注意我们认为此网络为有向网络，即每一球的心态不受下一球的印象，且本球的得分不决定本球的技术与心态

（网络示意图）



(有水印，4重画)

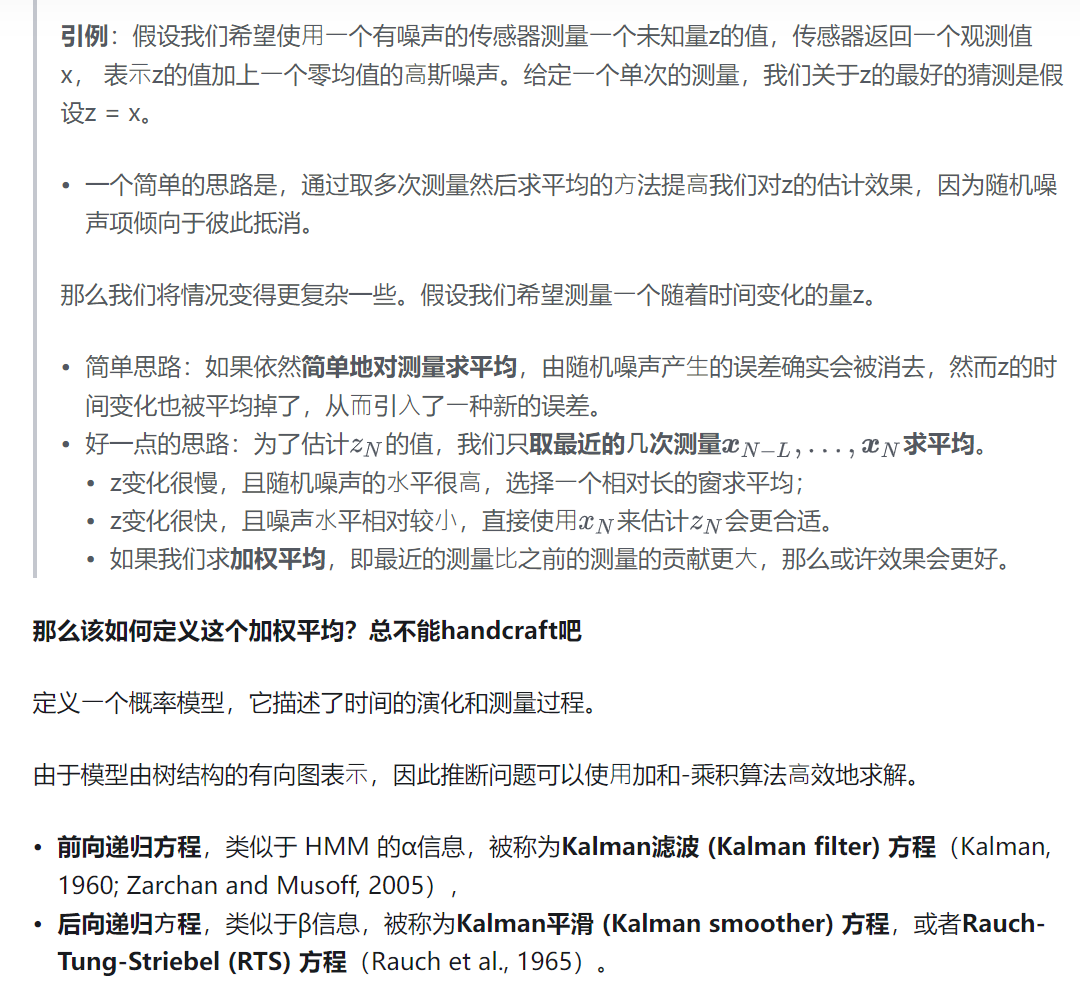
对于顺序数据来说，之前的概率图描述了两个重要的模型。如果潜在变量是离散的，那么我们得到了隐马尔科夫模型。但此时潜在变量和观测变量都是离散变量，那么我们就得到了动态系统。特别的，如果认为潜在变量和观测变量满足线性高斯条件（结点的条件概率分布对于⽗结点的依赖是线性⾼斯的形式），则此时系统为线性动态系统（linear dynamical system）。

（下面是更多说明，可能会调整，因为介绍的是线性动态，我们现在不能做出线性假设）

动态系统是一种时间序列模型，我们可以在这个模型中对某个变量进行后验估计，这个变量叫**状态变量**。但是，我们无法直接获取到状态变量的信息，只能间接地获取到与状态变量相关的另一变量的信息，并且这个信息往往是带有噪声的。这里的“另一变量”叫**观测变量**，而这里的“相关”是指，假如我知道状态变量的取值了，那么观测变量的分布就随之确定了，这个条件概率分布叫作**发射概率模型**。其中，“分布”体现出观测过程是有噪声的，因为如果没有噪声的话，观测变量就是一个确切的值而不是分布了——“分布”，其实就是观测噪声的不确定性模型。另外我们知道，状态变量是可以在不同的取值之间变化的，这一变化叫作**状态转移**。如果我们知道上一个观测时刻状态变量的取值，那么在新的观测发生的时刻，状态变量的分布也是知道的，这个条件概率分布叫作**转移概率模型**。其中，“分布”体现出转移过程是有噪声的，其实就是转移噪声的不确定性模型。我们的目标就是根据一系列带有噪声的观测量，去估计一系列的状态变量是多少，每一个观测对应输出一个状态。顺便提一下，在非贝叶斯概念下，发射概率模型也叫**观测方程**，转移概率模型也叫**状态（转移）方程**。

（例子或许添加，便于解释）

为了说明线性动态系统的概念，让我们考虑下⾯这个简单的例⼦，它经常在实际问题中出现。假设我们希望使⽤⼀个有噪声的传感器测量⼀个未知量z的值，传感器返回⼀个观测值x，表⽰z的值加上⼀个零均值的⾼斯噪声。给定⼀个单次的测量，我们关于z的最好的猜测是假设z = x。然⽽，我们可以通过取多次测量然后求平均的⽅法提⾼我们对z的估计效果，因为随机噪声项倾向于彼此抵消。现在，让我们将情况变得更复杂。假设我们希望测量⼀个随着时间变化的量z。我们可以对进⾏常规的测量x，从⽽我们得到了x 1 , … , x N ，我们希望找到对应的z 1 , … , z N。如果我们简单地对测量求平均，那么由于随机噪声产⽣的误差会被消去，但是不幸的是我们会仅仅得到⼀个单⼀的平均估计，对z的变化进⾏了平均，从⽽引⼊了⼀种新的误差。



**Data Cleaning and Visualization**

考虑对所有比赛的技术数据做分析，但2023-wimbledon-1310除外（该场的rally和speed\_mph

全部是无效值），采用刚刚的第二种方法，即选择两位选手体力总/最近20球体力消耗的对比而非总里程

**（在这里的数据处理包含了发球，特殊写入）**

**数据挖掘&特征工程：**

记状态向量为Z

（默认所有变量为关于N的函数，即代表第N球的状态，如Z应为Z(N)，后全部省略“(N)” ，但N的含义保留）

Z的分量为：

近期得分情况导致的心态波动

技术性因素

分别记为V，Q，即Z=[ V, Q ]

其中V是取值为 [0,1] 的离散变量，0代表劣势，1代表优势

，Q是取值为 [0,1] 的离散变量，0代表优势，1代表劣势

* **V**

V=WV-AVG( V0 , V1 , V2) ——加权平均

V0 是关于近 mp (memorized points) 球的变量，取加权平均Qv0

前mp球的每球表现 对应下一球的预测

V1 是关于近 1 game 得分 的变量，设本game得分为n，本人得分m

前game的每球表现 对应下一game的预测

V2 是关于本set的拿下game数的变量，由马尔科夫链得到的获胜概率

前set的每game表现 对应本set的预测

其中V0 关注得分的位置/次序/时间，V1 、 V2 不关注，只关注最终比例

V0的确定采用**滑动窗口**的思路处理数据

即：

整match的前2球，取0

剩余所有时间，取【本game球数，3】的最小值，作为记忆 mp

其中指向此盘前i-1盘是否胜利，权重Qv0遵循

也即：按mp球来算，第一球占比mp，第二球占比mp-1，第三球占比mp-2， 。。。 ，第mp球占比1，后除以综合求出权重

**需要满足**

**即，若某game中，player1’s point : player2’s point = 6:4，**

**则在下一game中，对于player1，n=10，m=6，**

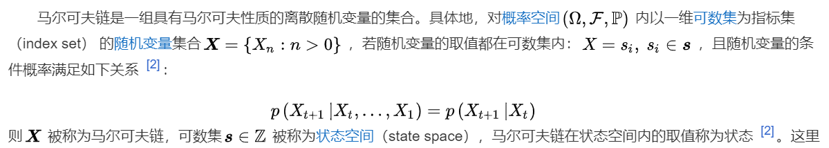
**考虑杨辉三角第9层：**

**1 9 36 84 126 126 84 36 9 1**

**player1获得前6个数的权重，player2获得前/后4个数，**

**即player1的V1==**

马尔可夫链模型（Markov Chain Model）: 将比赛状态建模为马尔可夫链，推断获胜的概率作为M的值——





以本次比赛的数据分析每种状态的状态转移函数，做出了状态转移图**。。。（有实际关系的压缩的图）**

此马尔科夫链是有限的，为一种吸收链。所以我们可以通过此模型，逆序递推出每种状态的最终胜率

递推公式：

结果（表）。。。（列表0,1,2,3,4,5,6）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 0 | 0.50 | 0.57 | 0.57 | 0.92 | 0.99 | 0.99 |  |
| 1 | 0.43 | 0.50 | 0.66 | 0.79 | 0.99 | 0.99 |  |
| 2 | 0.22 | 0.34 | 0.50 | 0.69 | 0.90 | 0.95 |  |
| 3 | 0.08 | 0.21 | 0.31 | 0.50 | 0.72 | 0.98 |  |
| 4 | 0.01 | 0.15 | 0.10 | 0.28 | 0.50 | 0.76 |  |
| 5 | 0.01 | 0.01 | 0.05 | 0.02 | 0.24 | 0.50 | 0.65 |
| 6 |  |  |  |  |  | 0.35 | 0.50 |

综合：

以一定权重——WV 确定为7:2:1综合V0 , V1 , V2，得到V综合，数据如下：

。。。（折线可视化结果）

* **Q：**

主成分分析PCA

简介：为了对“技术指标”降维，我们使用主成分分析，得到权重最重要的10个指标，对每个向量按照实际意义处理与解释，经过平滑化、标准化（正太化）、集中化的数据处理，得到每球的Q值

解释PCA模型：

共使用数据清洗后的6966-314个样本

29个指标分别为（改进列举）

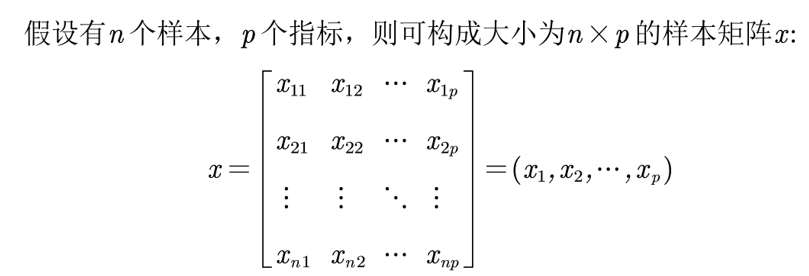
who\_to\_serve consumption\_of\_strength(最近20球消耗体力比值) consumption(总体力消耗比值)

p1\_ace p2\_ace p1\_winner p2\_winner p1\_double\_fault p2\_double\_fault p1\_unf\_err p2\_unf\_err

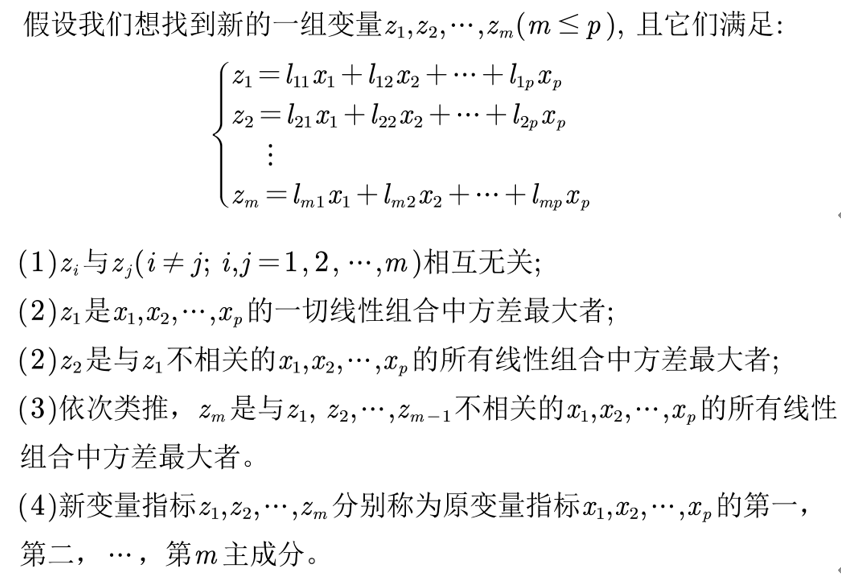
p1\_net\_pt p2\_net\_pt p1\_net\_pt\_won p2\_net\_pt\_won p1\_break\_pt p2\_break\_pt p1\_break\_pt\_won

p2\_break\_pt\_won p1\_break\_pt\_missed p2\_break\_pt\_missed p1\_serve\_speed\_ave

p2\_serve\_speed\_ave width\_1 width\_2 depth\_1 depth\_2 depth\_return\_1 depth\_return\_2的数据



则



这组z能较好的反映x变量的特征，在尽可能减少取消信息熵的前提下，完成降维，便于处理。

通过SPSS软件分析，得到的主成分与贡献值累加和为（可视化结果，显示10项）

为了保证准确性，拟打算取10项，得到80%的贡献

对每个向量，忽略一些影响较小的因素，尽可能的每个变量做到线性相关，如图，并给出指标的理解

注意到可以正相关可以负相关，但是在一个向量内必须同向

（可视化结果）

数据处理的方法：

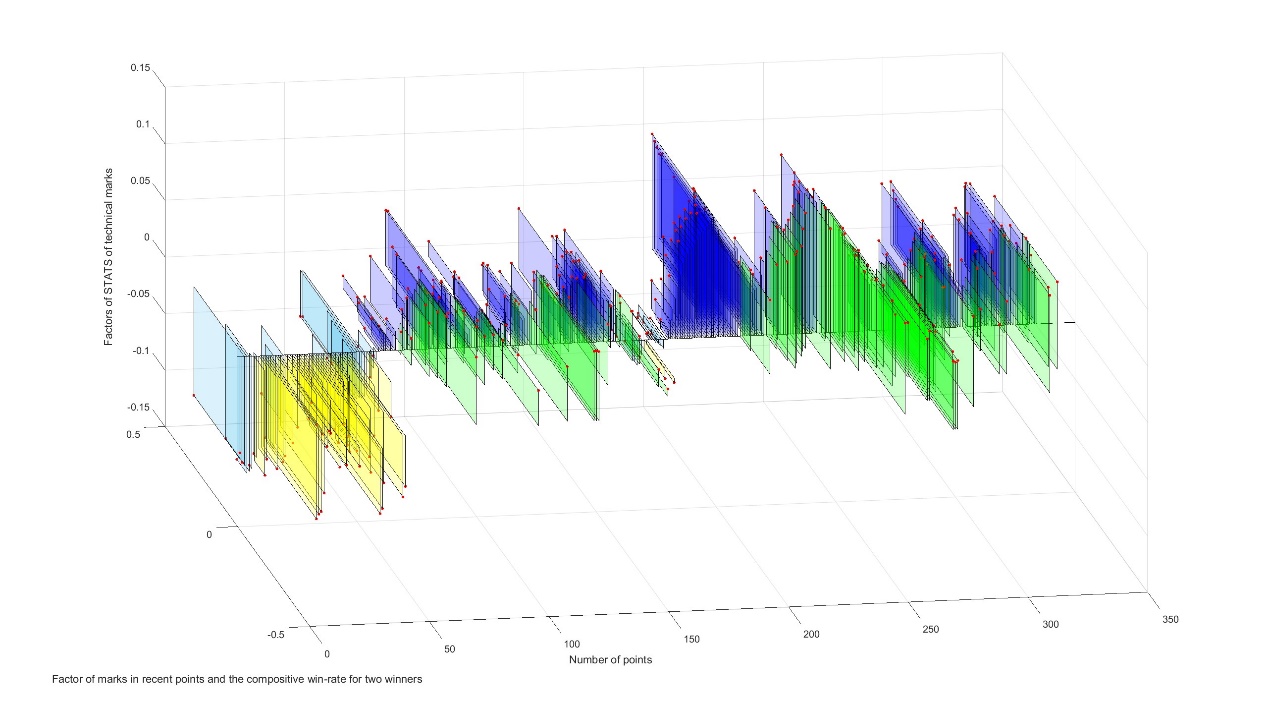
经过平滑化、标准化（正太化）、集中化

（数据可视）

* **综合可视**

折线图可视

最后可视化：



并依照图例分析经典对局：

**轴解释**

**颜色解释**

**面积解释**

第一局，黄块多，德约尽显优势

第二局，绿块多，且接近轴，说明双方难舍难分，但有些许蓝块，阿尔卡艰难赢下此局

第三局，蓝块明显，阿尔卡尽显优势

第四局，绿块多，偏向德约

第五句，最后还是蓝块多，阿尔卡拿下终局

**二、V数据处理的评估&教练的反驳**

**swings in play分析**

STEP 1 swings in play的量化

swings in play理解为优势转换，

定义为当点数差大于5时，局势判定为有利，则swings in play发生在局势转换的时候

STEP 2 蒙特卡洛模拟200次

考虑到构建教练的模型，涉及到对选手实力的评估，每球的随机因素，以及发球方

选手实力的评估，直接考虑一段够长的时间内选手得分的比率，我们在数据处理中直接用终局得分的比例，最能接近选手的实力比

考虑到发球方，我们对是否自身发球，乘以“优势因素”（不给具体数值）

之后模拟200次

最后得到的分布为：**（player1、player2柱状可视化结果展示）**

STEP 3 分析比较

对于player1：

M预测筛选后的均值0.654，方差0.02

R预测筛选后的均值0.568，方差0.13

**（散点可视化结果比较展示）**

方差更小更稳定的预测，趋向也更符合（趋近1）

按照逻辑斯蒂得分预估：

M的得分为7，R的得分为1.4

对于player2：

M预测筛选后的均值0.317，方差0.012

R预测筛选后的均值0.508，方差0.100

**（散点可视化结果比较展示）**

方差更小更稳定的预测，趋向也更符合（趋近0）

按照逻辑斯蒂得分预估：

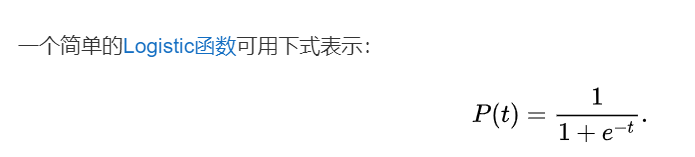
M的得分为8.9，R的得分为0.6

考虑到希望对这两种预测模型的更综合的评分，我们构建了以**逻辑斯蒂函数为基础的预测评分模型。**

**逻辑斯蒂函数的介绍与特性：**

**逻辑斯蒂函数**是一种常见的S函数，连续函数、单调递增、当自变量为0时，因变量为0.5，

当自变量接近正无穷，因变量接近1，当自变量接近负无穷，因变量接近0，



**评价体系：**

**我们考虑逻辑斯蒂的反函数，他有如下特性：**

**预测为0.5，值为0，函数越接近线性**

**预测越接近1，值越接近**正无穷，增势平缓

**预测越接近-1，值越接近**负无穷，减势平缓

考虑到单次预测不能对整体的得分造成太多的影响，设置一个单次得分的限度，本模型设置为[-2,2]，当得分偏离限度时，取值**超过限度时，控制为最接近得分的限度值。**

**逻辑斯蒂函数、y’函数与y关系、复合后函数，画图展示**

对每个预测值评分后平均，得到对预测模型的评估

**runs of success分析**

1 runs of success的量化

runs of success理解为连续得分，则

ROS\_player1 (N) = PTS\_player1(N) & PTS\_player1(N-1) & PTS\_player1(N-2)

2 蒙特卡洛模拟200次**（player1、player2柱状可视化结果展示）**

3 分析比较

对于player1：

M预测筛选后的均值0.624，方差0.02

R预测筛选后的均值**0.612**，方差0.13

**（散点可视化结果比较展示）**

方差更小更稳定的预测，趋向也更符合（趋近1）

按照逻辑斯蒂得分预估：

M的得分为5.6，R的得分为3.1

对于player2：

M预测筛选后的均值0.387，方差0.02

R预测筛选后的均值0.468，方差0.08

**（散点可视化结果比较展示）**

方差更小更稳定的预测，趋向也更符合（趋近0）

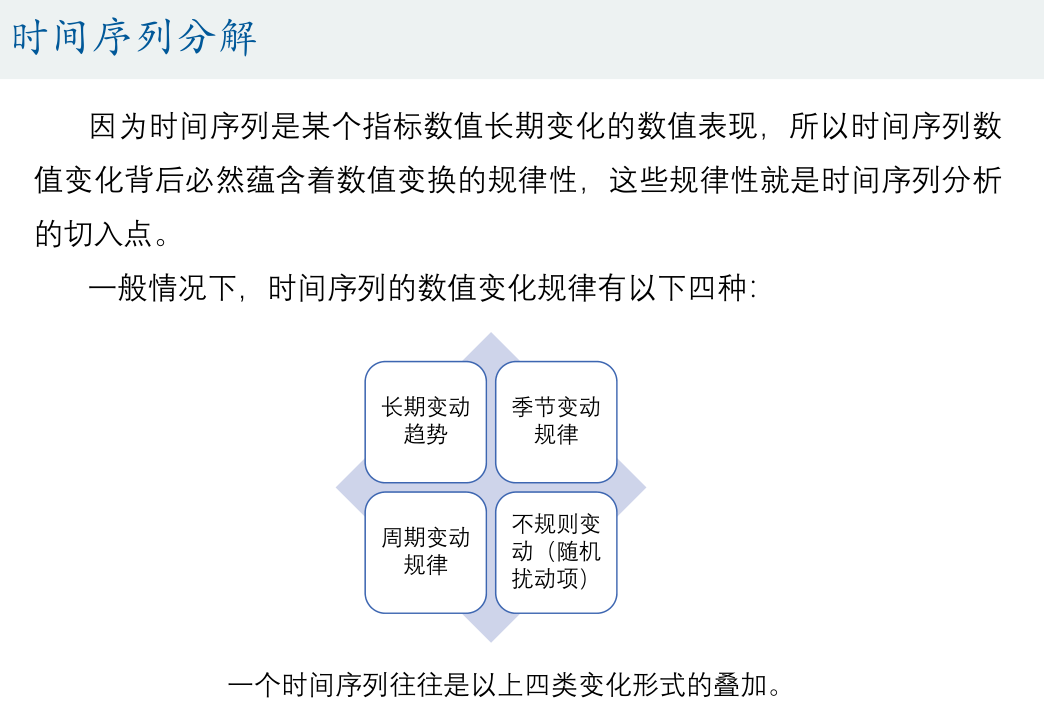
按照逻辑斯蒂得分预估：

M的得分为5.7，R的得分为1.9

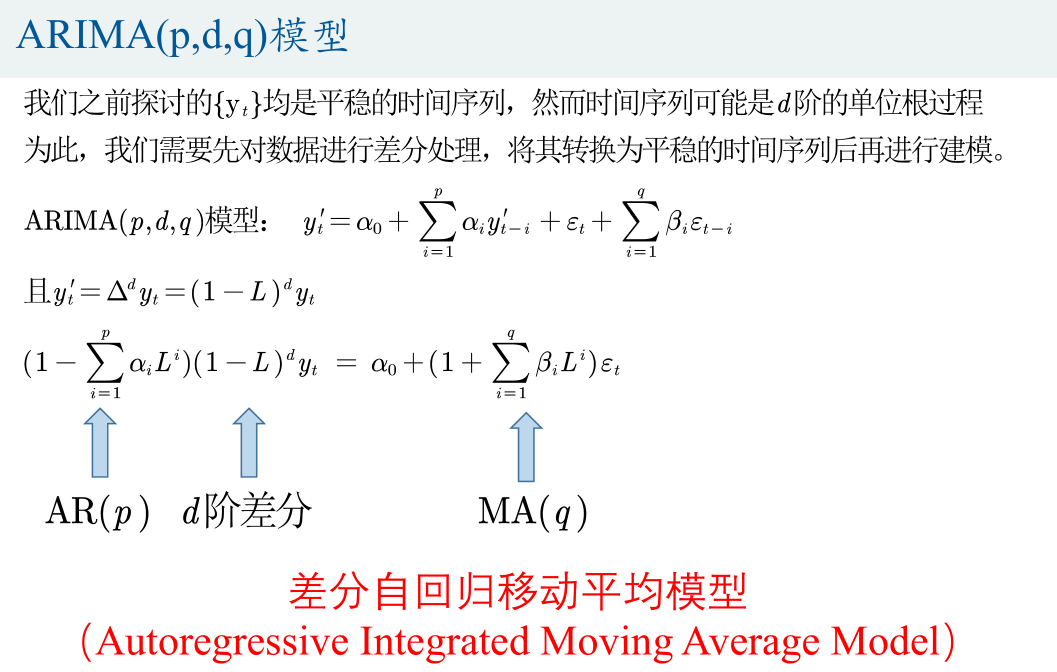
**momentum也不是随机的，而是与时间/赛程相关（ARIMA）**

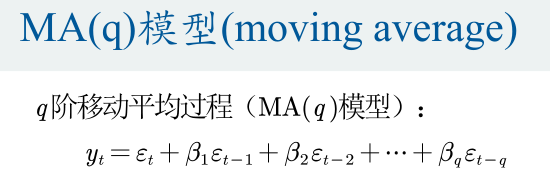
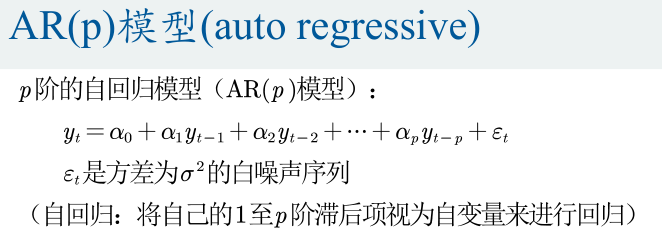
介绍

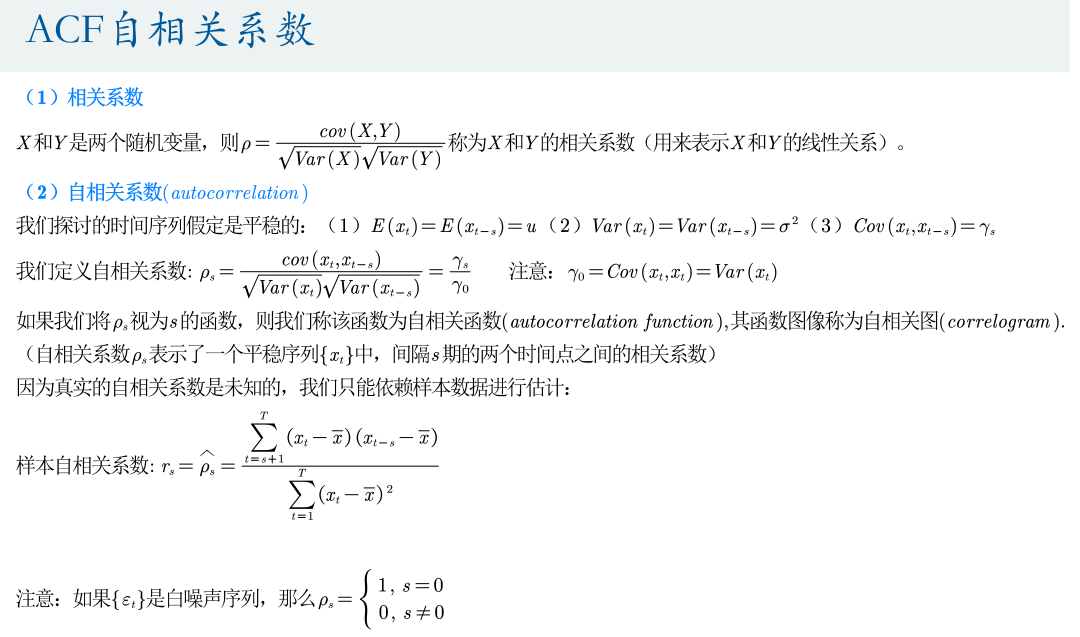
1 momentum是时序的



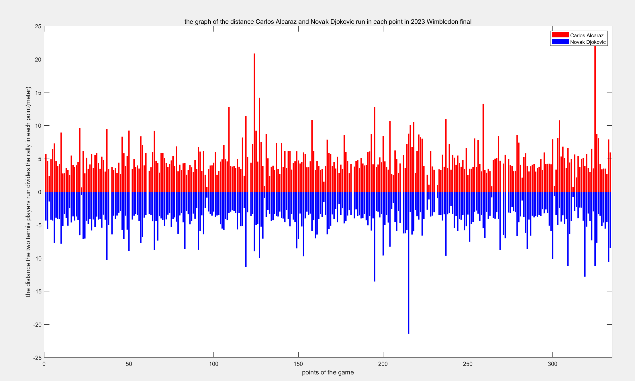
2 ARIMA







**模型结果**



**三、Q处理的评估&教练的期望**

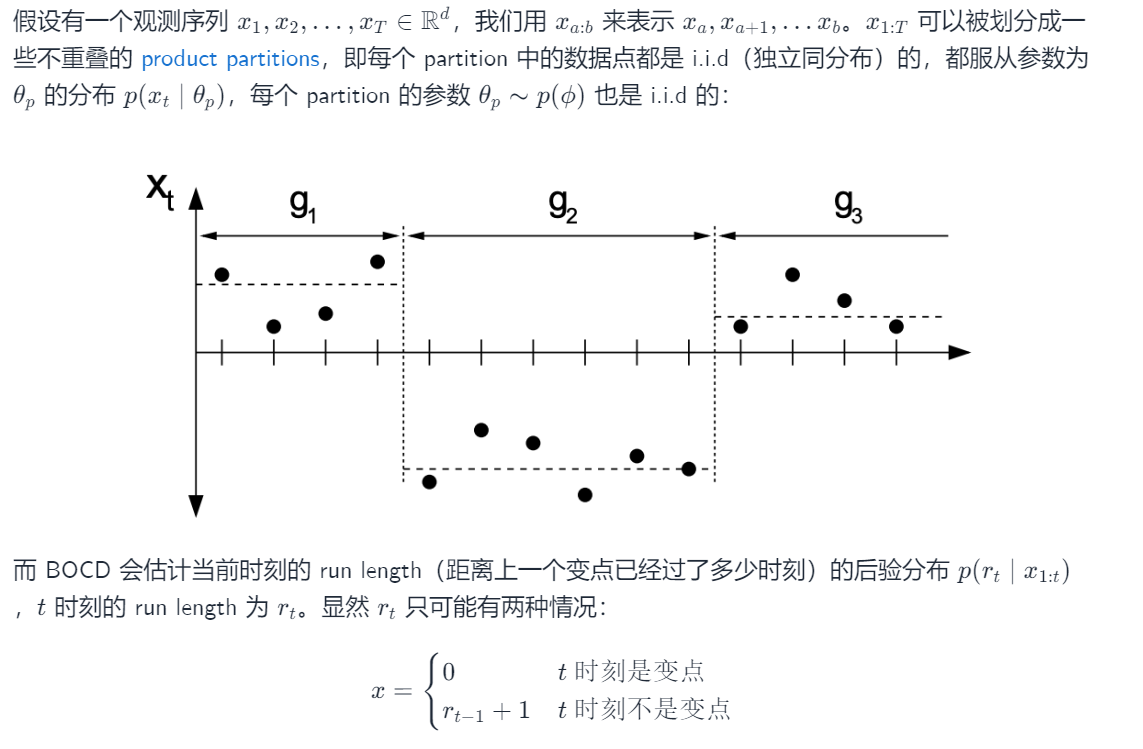
简介：为了分析the flow of play转变的指标，我们使用BOCD：Bayesian Online Changepoint Detection 处理每场合适比赛的Q，得到所有Q的显著断点具体信息，并对这些断点的产生情况进行分析，最后分析出5项最容易产生the flow of play转变的指标。

注意到这里选择Q，是因为为了取消具体得分失分对选手的影响。选手得分失分会在一定程度上一般化的影响选手，而单纯的单次得分失分，在高等级的比赛中，不会对选手的心理造成显著影响。如何得分失分才是我们所关注的。

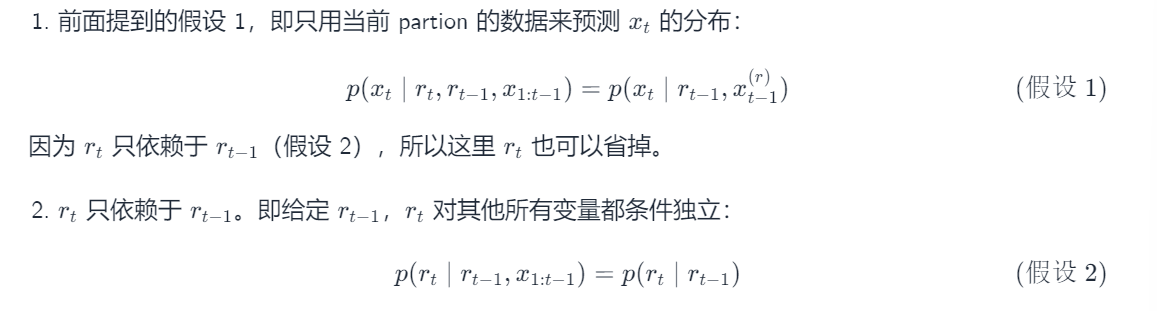
* 贝叶斯变点检测

解释：~~https://zxh.io/posts/zh/2020-08-22-bocd/~~

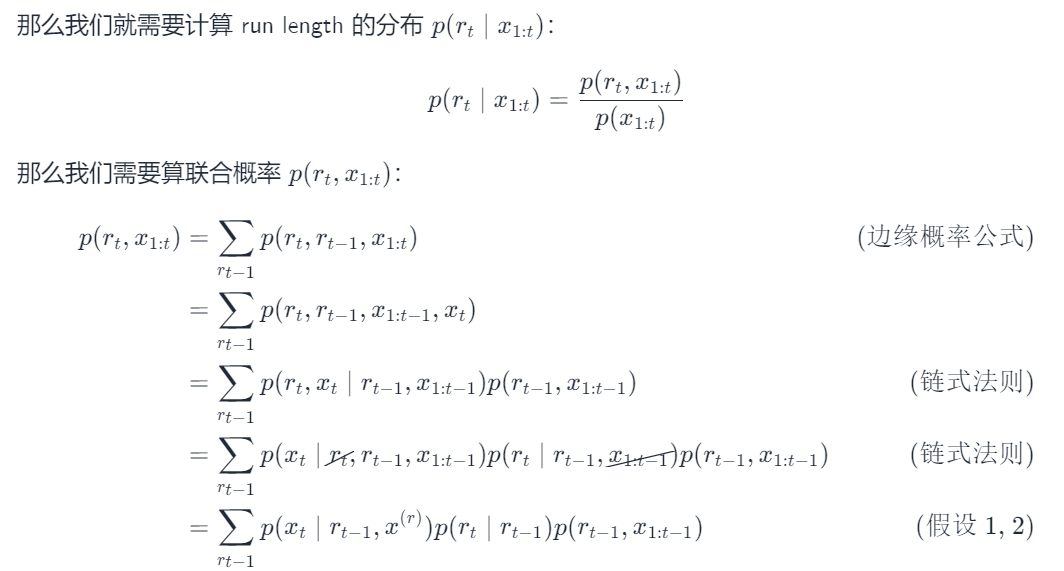
给定一个数据序列，在某个时间点，数据的某个（或某些）参数可能由于系统性因素（而非偶然性因素）而突然发生变化，那么这个时间点被称为**变点**（changepoint）。**变点检测**（changepoint detection）就是要估计出变点的位置。



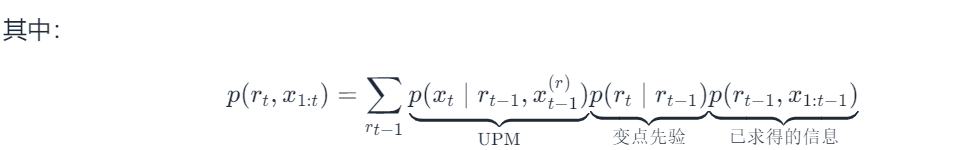
最后一步能成立依赖于以下两个假设：

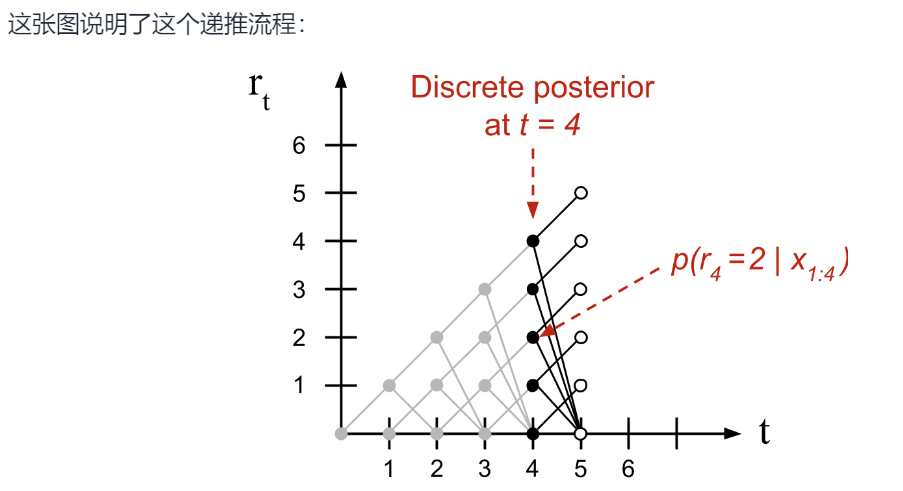


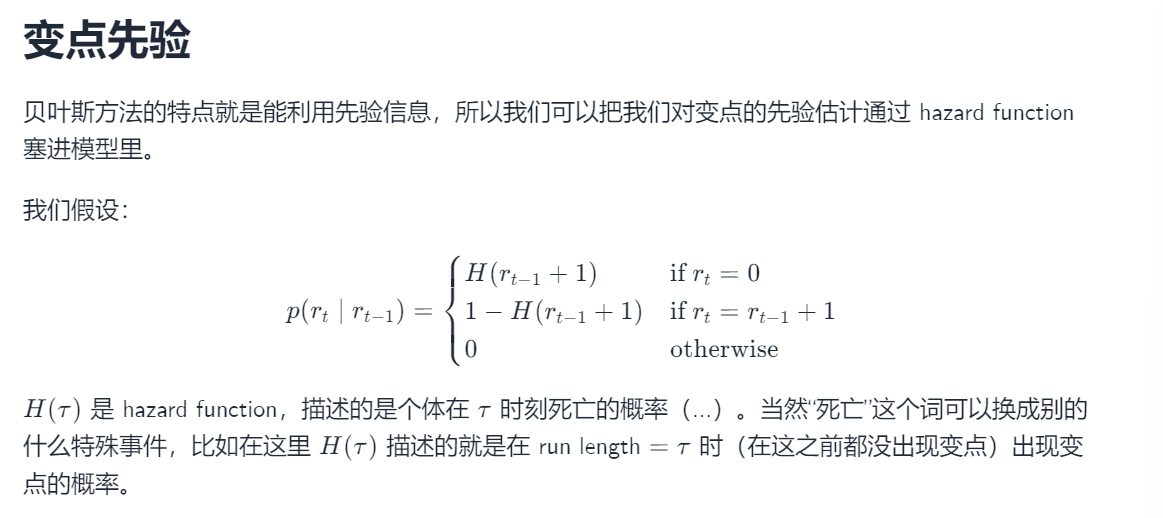
（下图第一行是贝叶斯公式

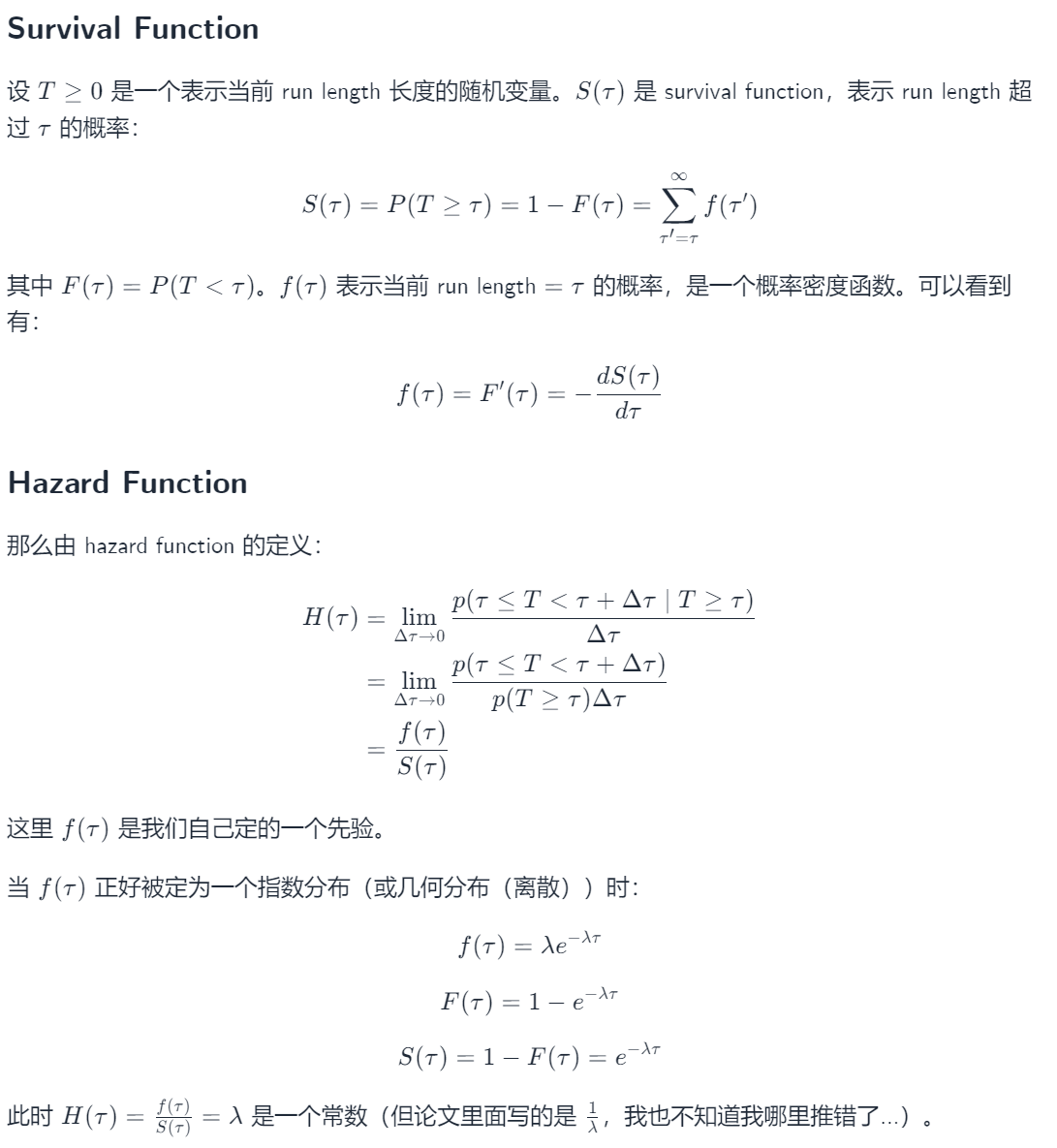
（放头尾即可）

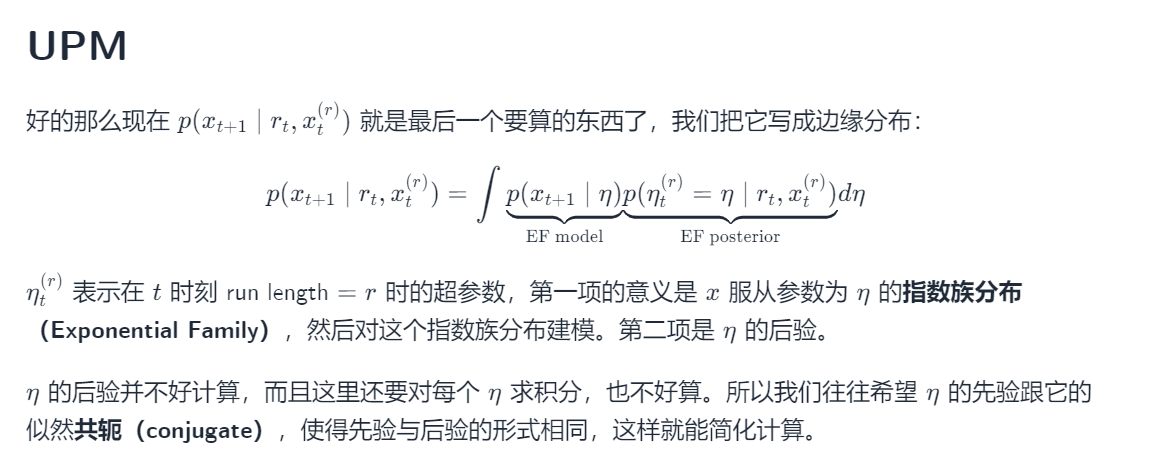
由此方程得知应使用递推求解，在这里我们对**比赛数据建模**，游戏开始的时刻一定是个变点。那么显然有递推初始值

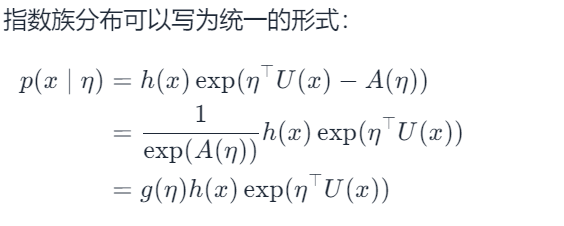
Underlying Probabilistic Model（UPM）

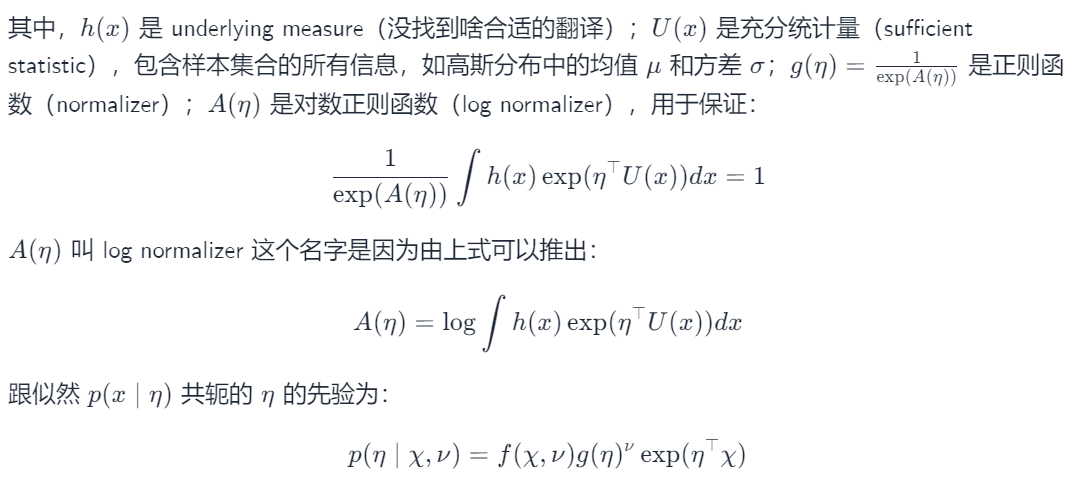
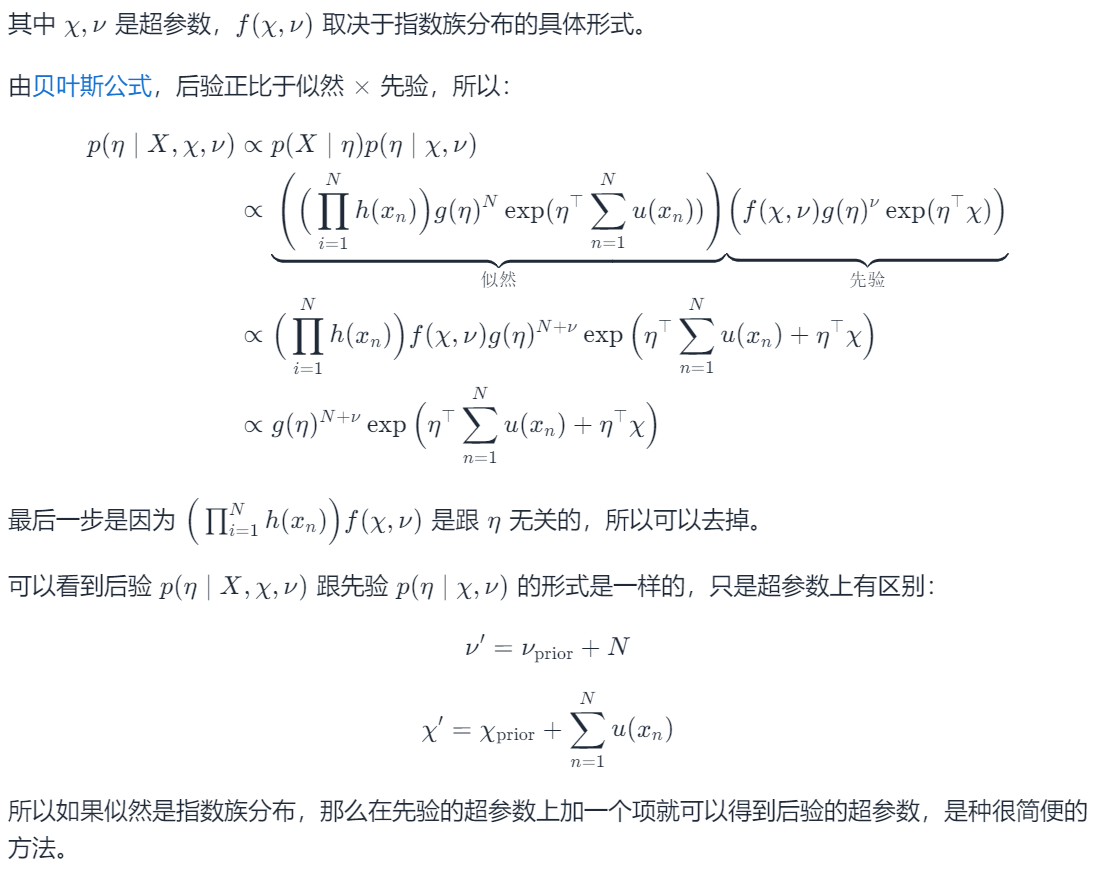










（整理压缩。。。）

执行程序进行BOCD，所得示意图如下

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

节选半决赛&四分之一决赛的，顺序为从左上到右下1501 1502 1503 1504 1601 1602

对所有Q的显著断点具体信息结果进行处理，共得到129个断点。统计每个断点的主要因素，去除发球方因素，其中选择5项频数最高的，也即最容易产生the flow of play转变的指标，并进行解释，如下

列表：

|  |  |  |
| --- | --- | --- |
|  | 频率 | 解释 |
| 对方breakpoint/发球局胜负 | 31/129 |  |
| 接发球 | 27/129 |  |
| 失误 | 23/129 |  |
| 体力比值 | 19/129 |  |
| 发球直接得分 | 18/129 |  |

**四、模型结果&预测**

* 卡尔曼滤波分析

假设我们的测量方式正确，测量结果存在噪声，即我们的测量结果可能会呈现高斯分布

只需要搞明白两件事情：

1、先验+实验观测->后验，或者说：后验结果

2、1中所述事实在一条[马尔可夫链](https://www.zhihu.com/search?q=%E9%A9%AC%E5%B0%94%E5%8F%AF%E5%A4%AB%E9%93%BE&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra=%7B%22sourceType%22%3A%22answer%22%2C%22sourceId%22%3A911535358%7D)上反复发生。

问题就可以简单明了的解释明白。

其中我们要预测的结果就是后验，先验是技术因素Q，而实验观测即M

先验就是之前的观测结果，即M，

它是一个在**隐马尔可夫链**上的**递归贝叶斯估计**，对于每一个时刻，上一时刻的状态转移作为本时刻的**先验**，建模为高斯分布，依据贝叶斯公式，通过本时刻建模为高斯分布的的**观测**对这个先验进行修正，最后得到本时刻的**后验**，这个后验是前面两个高斯分布的乘积，依然是高斯分布，新的高斯分布的参数：均值和方差可以通过比较系数推导出来。这个后验将作为下一个时刻的先验，如此反复进行。

模型方程：。。。

得到的胜率结果：

开始：N(0.5,0.009) ~ N(0.5,0.04)

胜率分布：

[[易懂]如何理解那个把嫦娥送上天的卡尔曼滤波算法Kalman filter? - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/77327349)

以及等等的原理

**五、模型评估&拓展**

**模型评估.**

1. 混淆矩阵分析：精确度召回率

- 模型在真正例和真负例上的表现较好，但可能存在一些假正例和假负例，需要进一步关注。

总结

混淆矩阵概述：

混淆矩阵是一个二维矩阵，用于总结分类模型在不同类别上的预测结果，包括 True Positive (TP)、False Negative (FN)、False Positive (FP)、True Negative (TN)。

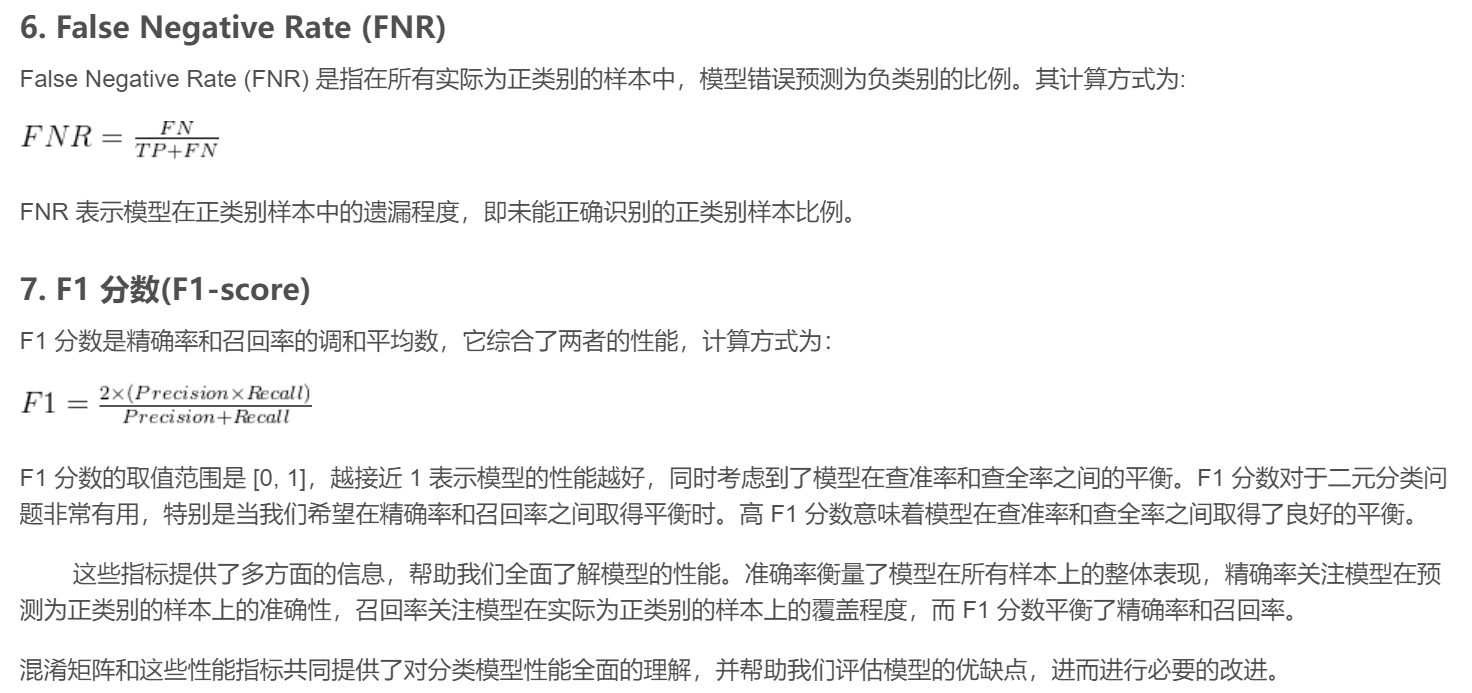
性能指标：

准确率（Accuracy）：模型正确分类的样本占总样本数的比例。

精确率（Precision）：模型预测为正类别的样本中有多少是真正的正类别。

召回率（Recall）：实际为正类别的样本中，有多少被模型正确预测为正类别。

F1 分数：精确率和召回率的调和平均数，综合考虑了查准率和查全率。



|  |  |  |  |
| --- | --- | --- | --- |
| 好测对 | 坏测对 | 好测错 | 坏测错 |
| 2565 | 1678 | 1750 | 956 |

**以下部分语文建模**

给出比赛建议

战术建议涉及到控制V（面对失分的积极心态）、Q（技术提高）、以及战术针对

修正建议

建议增加维度：天气、风速，等等

通用性分析

女子:q（局数无影响）  
球场表面:q变化

乒乓球:更强得分潮

**六、备忘录**

**报告结构建议**

1. **引言**：简要介绍研究的背景、目的和重要性。
2. **理论基础**：讨论势头在体育比赛中的角色，以及先前研究的相关理论。
3. **方法论**：
4. 数据收集：描述使用的数据集及其来源。
5. 特征工程：解释选择的特征及其对预测势头转换的重要性。
6. 模型选择与开发：概述所选模型的理由，包括模型的类型、训练过程和参数调整。
7. **结果分析**：
8. 势头分析：展示模型如何捕捉比赛中的势头变化。
9. 势头的影响：讨论势头变化对比赛结果的潜在影响。
10. 模型泛化能力：报告模型在不同比赛、运动和条件下的测试结果。
11. **讨论**：分析模型的局限性，提出未来研究方向。
12. **结论**：总结研究发现，并强调其对教练和选手准备的实际意义。
13. **附录**：包括数据集描述、代码实现和额外的图表或表格。

**模型总结与教练建议备忘录**

这一部分提供给教练的建议，基于上述分析的结果。

**建议摘要**

1. **势头的角色**：我们的分析显示，势头确实在比赛中扮演了一个重要角色。势头的增加与选手的表现提升相关联，而势头的减少可能预示着比赛流向的转变。
2. **势头变化的指标**：某些关键指标，如连胜点、重要得分时刻的表现，可以作为势头变化的前兆。
3. **准备选手**：
4. 加强心理训练，特别是在面对势头不利时，帮助选手保持冷静和专注。
5. 分析对手的比赛数据，识别可能引发势头转换的模式或弱点。
6. 在训练中模拟势头转换的情况，提高选手的适应能力。

基于模型结果，我们可以向教练提供以下建议：  
• 关注特定的比赛转折点指标，如得分差、发球权利用等，这些因素可能预示着比赛流向的即将变化。  
• 训练选手识别并利用比赛中的转折点，比如在连续得分或面临重要得分时的心理准备和战术调整。  
• 分析不同对手的比赛数据，以定制化的策略应对不同的比赛情况。

# Conclusion

# References（待整理）

[1]: Wikipedia contributors. "Wordle." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 22 Jan. 2024. Web. 23 Jan. 2024.

[2]: Bonthron, Michael. "Rank One Approximation as a Strategy for Wordle." *arXiv preprint arXiv:2204.06324*  (2022). Print.

**Bayesian Online Changepoint Detection.** Ryan Prescott Adams and David J.C. MacKay. arXiv 2007.