基于MCM数据洞察类问题的潜变量建模与多维推断模型：从历年优秀论文到2026年C题的系统化迁移策略

在数学建模竞赛（MCM）的演进历程中，“C题：数据洞察（Data Insights）”自2016年引入以来，已成为衡量参赛团队处理复杂、异构、大规模数据集能力的核心基石[1]。这类问题的本质在于，不仅要求对已知数据进行统计描述，更要求通过构建数学模型来推断那些隐藏在数据表象之下的“潜变量（Latent Variables）”或“隐状态（Hidden States）”[1]。2026年的MCM Problem C《与星共舞（Dancing with the Stars）》正是这一建模范式的典型代表，它要求通过公开的评委评分和淘汰结果，去逆向推导被视为“最高商业机密”的粉丝投票数[3]。为了攻克这一难题，深入分析2020年至2025年间Problem C的优秀论文（Outstanding Papers）所采用的模型体系，并将其逻辑严密地迁移至2026年的场景中，是建立高水平解决方案的必由之路。

历年MCM Problem C 题目类型与优秀论文模型谱系

MCM Problem C的题目设计具有高度的连贯性，即通过提供多维度的历史观测数据，要求团队挖掘数据间的因果关系、演化趋势及潜在干扰因子[5]。通过对近年来优秀论文的深度识别，可以发现其模型应用主要集中在贝叶斯推断、状态空间模型、集成机器学习以及因果推断四个维度。

2020年：海量市场数据的特征挖掘与评论情绪量化

2020年的C题《数据的财富（A Wealth of Data）》提供了亚马逊在线市场的海量评论、星级评分及帮助度评估数据，要求识别产品声誉的时间规律及特征关联[7]。优秀论文（如Team 2010638）构建了一套多层级的信息挖掘体系，直接应对了非结构化文本与结构化数值的混合分析压力[3]。

该论文采用的核心模型包括自然语言处理（NLP）中的情感分析技术，通过NLTK工具包进行分词并结合余弦相似度（Cosine Similarity）剔除噪声数据[3]。在特征关联阶段，该团队引入了多元逻辑回归模型（Multinomial Logistic Regression），通过构建似然函数探讨了星级评分、评论长度与用户反馈帮助度之间的非线性关系[3]。更为深入的是，他们使用了隐含狄利克雷分布（LDA）模型，这是一种三层贝叶斯概率模型（文档-主题-词），用于从评论文本中提取潜在的主题特征，并计算单个评论与这些特征的相似度得分[3]。此外，为了预测产品声誉的未来演变，采用了时间序列分析（Time Series Analysis）和分布滞后模型（Distributed Lag Model）配合Almon方法，这为处理具有长期记忆效应的数据序列提供了经典范式[3]。

2023年：贝叶斯框架下的概率分布预测

2023年的C题《预测Wordle结果（Predicting Wordle Results）》要求利用推特上用户报告的Wordle游戏尝试分布（从1次到6次及失败），预测特定日期的结果分布及单词难度[10]。这一问题的核心在于，数据呈现为高度约束的百分比分布，且受到单词本身的词汇特性（如词频、信息熵）的影响[12]

。

在优秀论文（如Team 2301192）中，研究者构建了基于马尔可夫链蒙特卡罗（MCMC）方法的贝叶斯层级模型 [14]。他们通过定义单词的信息熵（WIE）和使用频率（FREQ）作为先验输入，利用XGBoost回归模型预测结果分布的直方图特征 [12]。该团队还巧妙地结合了ARIMA模型来捕捉游戏参与人数的周期性波动，确保了模型既能处理横截面上的单词差异，也能处理时间序列上的热度更迭 [12]。在分类任务中，K-means聚类与决策树（Decision Tree）的结合使用，使得原本抽象的“难度”变量被量化为简单、中等和困难三个逻辑层级，这种“聚类+分类”的组合策略在处理缺乏标签的监督学习任务时表现优异 [12]。

2024年：状态空间模型中的潜变量提取

2024年的C题《网球运动中的动量（Momentum in Tennis）》是与2026年“粉丝投票推导”关联度最高的问题之一 [2]。该题目要求根据温布尔登网球赛的得分记录，量化球员在比赛过程中的“动量（Momentum）”——这是一个典型的不可直接观测的隐藏变量 [2]。

优秀论文（如Team 2410482）采用了状态空间模型（State Space Model）配合卡尔曼滤波（Kalman Filter）算法 [3]。在这一框架下，动量被建模为随时间演化的隐状态，而球员的得分、非受迫性失误及跑动距离被视为观测变量。通过卡尔曼滤波的预测和更新步骤，模型能够实时推断出当前时刻动量的概率分布。同时，团队使用马尔可夫模型（Markov Model）来计算在特定动量状态下的获胜概率，通过逻辑回归（Logistic Regression）验证了动量对比赛流向的显著性影响 [3]。这种将“不可见变量”置于动态方程核心的处理方式，为2026年推导随周次变化的粉丝投票提供了直接的技术路径。

2025年：复杂集成模型与因果推断

2025年的C题《奥运奖牌榜模型（Models for Olympic Medal Tables）》侧重于多源异构数据的集成与宏观预测 [16]。优秀论文（如Team 2500759和Team 2510862）展示了如何处理具有强烈社会学背景的复杂数据，如“主场效应”和“顶级教练效应” [3]。

这些论文采用了堆叠集成模型（Stacking Ensemble Model），融合了XGBoost、LightGBM、随机森林（Random Forest）和支持向量机（SVM），实现了高达0.88的 R^2 预测精度 [3]。在特征处理上，主成分分析（PCA）被用于降维，处理具有高度相关性的宏观经济和社会指标 [3]。特别值得借鉴的是，为了量化“教练效应”，团队采用了双重差分法（Difference-in-Differences, DID）和倾向得分匹配（PSM-DID），这在因果推断领域是识别政策或特定干扰项（如DWTS中的专业舞者更换）对结果影响的权威工具 [3]。此外，通过Bootstrap方法产生的预测区间，有效地量化了模型在面对未来（2028年洛杉矶奥运会）预测时的不确定性 [3]。

面向2026年C题的模型迁移与构建策略

2026年的MCM Problem C要求对《与星共舞》的粉丝投票进行推算、对两种投票计算方法进行公正性评估，并分析明星特征对结果的影响 [3]。基于上述历年优秀论文的分析，可以构建如下模型体

系来应对该问题的核心挑战。

核心任务一：粉丝投票的隐状态推断模型

任务要求估计每位参赛者在每周比赛中获得的未知粉丝票数，并评估模型的一致性和确定性 [3]。这与2024年网球动量模型及2023年Wordle分布模型高度契合。

1. 状态空间模型与卡尔曼滤波的迁移应用 借鉴2024年Team 2410482的思路，可以将粉丝投票数看作一个受多种因素影响的隐状态向量 X_t [3]。

- 状态转移方程：假设粉丝投票具有惯性，本周的投票数受上周表现、明星知名度及积累的观众好感度影响。
- 观测方程：淘汰结果和最终排名是由于“评委评分百分比+粉丝投票百分比”之和的排序决定的 [3]。通过构建一个排序似然函数，利用卡尔曼滤波或粒子滤波（Particle Filter）在受限的解空间内搜索最符合淘汰观测序列的粉丝投票数值序列。
- 确定性度量：滤波算法中自带的状态协方差矩阵可直接用于量化投票估计的不确定性 [3]。

2. 贝叶斯层级模型与**MCMC**采样 参考2023年Team 2301192的Wordle预测模型，可以将粉丝投票建模为基于Dirichlet分布的比例推断问题 [14]。

$$P(\text{Votes}|\text{Scores}, \text{Eliminations}) \propto \ P(\text{Eliminations}|\text{Scores}, \text{Votes})P(\text{Votes})$$

利用MCMC采样（如Metropolis-Hastings或Gibbs采样），在满足“最低综合分被淘汰”这一约束条件下，生成成千上万种可能的投票分布方案 [3]。通过分析这些样本的均值和标准差，不仅可以得到投票估计值，还能给出确切的置信区间，响应题目对“确切度（Certainty）”的要求 [3]。

核心任务二：投票合并方法的对比与社会选择分析

题目要求对比“排名法”与“百分比法”在不同赛季的效果，并探讨其对粉丝票数的偏好性 [3]。

1. 社会选择理论（**Social Choice Theory**）视角下的模型构建 参考2023年HiMCM关于投票机制的探讨，可以引入**Borda**计数法或**孔多塞准则（Condorcet Criterion）**作为评估基准 [19]。

- 灵敏度分析模型：构建一个仿真系统，输入相同的评委评分和模拟的粉丝投票分布，分别在排名法和百分比法下运行淘汰逻辑。通过计算不同方法下淘汰结果与“仅粉丝投票排名”或“仅评委评分排名”的相关系数（Spearman's Rho），量化哪种方法更倾向于保护人气明星（如Jerry Rice案例）或专业舞者 [3]。
- 公平性评价指标：引入基尼系数（Gini Coefficient）或赫芬达尔指数（HHI）来衡量两种投票系统对权力分配的集中度，分析是否存在“极端粉丝投票”绑架比赛结果的数学风险 [18]。

核心任务三：明星特征与专业舞者影响的因果推断

任务要求分析年龄、行业及合作伙伴对表现的影响，并判断对评委和粉丝的影响是否一致 [3]。

1. 因子分析与**SHAP**解释模型 借鉴2025年奥运奖牌榜模型（Team 2510862）[3]。

- 集成学习模型：使用XGBoost构建预测模型，以明星年龄、行业分类、家乡、合作伙伴为自变量，分别以评委平均分和估算的粉丝投票数为因变量。
- **SHAP**值解释：通过计算Shapley值，定量分析每个特征对预测结果的贡献度[3]。例如，可以直观地展示“运动员（Athlete）”这一行业背景对评委评分的正向贡献是否显著高于其对粉丝票数的贡献，从而回答题目中关于“影响方式是否一致”的问题[3]。

2. 双重差分法（**DID**）在专业舞者价值评估中的应用 参考2025年Team 2500759对“顶级教练”的分析逻辑[3]。

- 模型构建：选取在不同赛季更换了明星搭档的专业舞者数据，构建DID模型。通过比较同一位舞者在携带不同特征明星时的表现增量，可以剥离出舞者自身的“品牌溢价”对粉丝投票的拉动作用[3]。

综合数据分析中的模型选择建议

在处理2026年Problem C所提供的多维度CSV数据时，下表总结了从历年优秀论文中识别出的、最适合迁移的高级模型：

分析维度	推荐模型	来源参考	迁移至**2026**年**C**题的逻辑
粉丝投票推导	状态空间模型 (SSM)	2024 Team 2410482 [3]	将周次视为时间步，将淘汰结果视为观测约束，反演隐状态投票数。
投票分布模拟	贝叶斯层级模型 (BHDM)	2023 Team 2301192 [14]	在给定约束下使用MCMC采样，获取投票分布并量化不确定性。
特征贡献分析	XGBoost + SHAP	2025 Team 2510862 [3]	识别年龄、行业及舞者搭档对评分和粉丝票数的非线性影响。
政策/机制评估	双重差分法 (DID)	2025 Team 2500759 [3]	评估投票规则变更（如增加评委选择权）对最终结果的

			因果影响。
不确定性度量	Bootstrap 方法	2025 Team 2510862 [3]	通过重采样生成粉丝投票的置信区间，满足确定性度量要求。
文本与情绪挖掘	LDA + 情感分析	2020 Team 2010638 [3]	若引入社交媒体讨论数据，可量化明星的舆论好感度作为模型先验。

数据处理与稳健性考量

在应用这些模型之前，必须注意2026年数据中的特殊性，例如赛季15的“全明星”阵容以及部分周次没有淘汰的情况 [3]。优秀论文在处理类似复杂数据时，通常会进行以下预处理：

- 缺失值处理：针对第4位评委分数的N/A，采用基于K-近邻（KNN）或均值填补的策略，并在灵敏度分析阶段测试缺失值对模型结论的稳健性 [3]。
- 归一化与标准化：由于不同赛季评委的总给分范围可能因评委人数变化（3人或4人）而异，需将所有分数转换为百分比形式进行统一建模，这在2020年处理不同类别产品的评分时被证明是必要的 [3]。

推理模型的深化：面向“公平性”与“精彩度”的新系统建议

题目最后要求提出一种更“公平”或更“精彩”的投票系统 [3]。这一要求触及了计算社会科学的核心。

引入布拉德利-特里（Bradley-Terry）模型的改进方案

在2021年IM2C优秀论文中，研究者利用Bradley-Terry模型来选择“最伟大的运动员” [22]。该模型通过成对比较来推导个体的相对强度：

$$P(i \text{ beats } j) = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}}$$

对于DWTS，可以提议一种新的“天梯评分系统”，即粉丝投票不再是简单的票数累加，而是通过粉丝在明星之间的“选择倾向”来更新明星的动态ELO等级分 [24]。这种系统的优势在于：

1. 减少刷票影响：ELO系统对实力（评分）相差悬殊的对决有自动调节机制。
2. 动态性更强：能够捕捉明星在赛季中的“进步曲线”，增加节目后期的悬念和精彩程度 [23]。

通过将这种逻辑与2026年C题的百分比投票法进行模拟对比，可以为节目制作人提供具有深厚数

学背景的建议方案[3]。

结论

2026年MCM Problem C的核心挑战在于数据中的信息不对称。通过系统性地借鉴2020年至2025年优秀论文中的模型体系——特别是状态空间模型对隐变量的捕捉、贝叶斯层级模型对概率分布的刻画、以及集成学习配合SHAP值的特征解释力——参赛者可以构建出一个逻辑严密、具备统计稳健性的多维推断框架。这不仅能够精准地估算出粉丝的"隐秘投票"，更能深刻地揭示娱乐竞赛背后复杂的人际影响力和制度设计逻辑。

Works cited

1. MCM Problem C: Data Insights, accessed January 30, 2026, https://sci.cqu.edu.cn/__local/4/5F/82/1E6B5565E87046445639EF839B2_AD91DF65_1B33D.pdf?e=.pdf
2. 2024 MCM Problem B - COMAP - Contests, accessed January 30, 2026, https://www.contest.comap.com/undergraduate/contests/mcm/contests/2024/results/2024_MCM_Problem_B_Results.pdf
3. 2020-2025美赛优秀论文大合集
4. 2026 MCM Problem C: Data With The Stars - International Mathematical Modeling Challenge, accessed January 30, 2026, https://www.immchallenge.org/mcm/2026_MCM_Problem_C.pdf
5. A Comparative Analysis of the National College Student Mathematical Modeling Competition and the American College Student Mathematical Modeling Competition - Oreate AI Blog, accessed January 30, 2026, https://www.oreateai.com/blog/a-comparative-analysis-of-the-national-college-student-mathematical-modeling-competition-and-the-american-college-student-mathematical-modeling-competition/e0b5cfdd5d968e0e8221b37636017371
6. MCM Director's Overview of the Mathematical Contest in Modeling: What Advisors Need to Know. Abstract, accessed January 30, 2026, https://meetings.ams.org/math/jmm2021/mediafile/Handout/Paper2657/MCM%20Directors%20Overview%20HandOut.pdf
7. 2020 MCM Problem C - COMAP - Contests, accessed January 30, 2026, https://www.contest.comap.com/undergraduate/contests/mcm/contests/2020/results/2020_MCM_Problem_C_Results.pdf
8. 2020 MCM Problem C | PDF | Data | Computing - Scribd, accessed January 30, 2026, https://www.scribd.com/document/888031768/2020-MCM-Problem-C
9. 2020 MCM/ICM Problems - COMAP - Contests, accessed January 30, 2026, https://www.contest.comap.com/undergraduate/contests/mcm/contests/2020/problems/
10. 2023 MCM Problem A Results - COMAP - Contests, accessed January 30, 2026, https://www.contest.comap.com/undergraduate/contests/mcm/contests/2023/results/2023_MCM_Problem_A_Results.pdf
11. MCM/ICM Problems - COMAP - Contests, accessed January 30, 2026, https://www.contest.comap.com/undergraduate/contests/mcm/contests/2023/pr

oblems/
12. Puzzle Game: Prediction and Classification of Wordle Solution Words - arXiv, accessed January 30, 2026, https://arxiv.org/html/2403.19433v4
13. Puzzle game: Prediction and Classification of Wordle Solution Words indicates equal contribution. - arXiv, accessed January 30, 2026, https://arxiv.org/html/2403.19433v3
14. Fungi, Trash, Bikes, and Wordle: Remembrance of Models Past- American Mathematical Society, accessed January 30, 2026, https://meetings.ams.org/math/jmm2024/meetingapp.cgi/Paper/30716
15. Predicting the Guess Distributions and Number of Reported Plays for Wordle - DigitalCommons@Linfield, accessed January 30, 2026, https://digitalcommons.linfield.edu/symposium/2023/all/46/
16. 2025 MCM Problem C - COMAP - Contests, accessed January 30, 2026, https://www.contest.comap.com/undergraduate/contests/mcm/contests/2025/results/2025_MCM_Problem_C_Results.pdf
17. 2026 MCM Problem C: Data With The Stars - COMAP - Contests, accessed January 30, 2026, https://www.contest.comap.com/undergraduate/contests/mcm/contests/2026/problems/2026_MCM_Problem_C.pdf
18. How Eliminations are Calculated : r/dancingwiththestars - Reddit, accessed January 30, 2026, https://www.reddit.com/r/dancingwiththestars/comments/1gh0135/how_eliminations_are_calculated/
19. There is one best voting method, let's end the non-discussion : r/slatestarcodex - Reddit, accessed January 30, 2026, https://www.reddit.com/r/slatestarcodex/comments/1ahbspq/there_is_one_best_voting_method_lets_end_the/
20. Based on Math Alone, Where Does Andy Richter's 'DWTS' Run End? - Pajiba, accessed January 30, 2026, https://www.pajiba.com/tv_reviews/based-on-math-alone-where-does-andy-richters-dwts-run-end.php
21. The Voting premise is actually more critical than 50% : r/dancingwiththestars - Reddit, accessed January 30, 2026, https://www.reddit.com/r/dancingwiththestars/comments/1o1xg6j/the_voting_premise_is_actually_more_critical_than/
22. 2021 Solutions - The International Mathematical Modeling Challenge (IMMC), accessed January 30, 2026, https://immchallenge.org/Contests/2021/Solutions.html
23. Bradley–Terry model - Wikipedia, accessed January 30, 2026, https://en.wikipedia.org/wiki/Bradley%E2%80%93Terry_model
24. Statistical Extensions of the Bradley-Terry and Elo Models - LMArena, accessed January 30, 2026, https://lmarena.ai/blog/extended-arena/
25. Bradley-Terry and Elo scores are equivalent mathematical models! The fundamental... | Hacker News, accessed January 30, 2026, https://news.ycombinator.com/item?id=44595895