硕士学位
论　文

基于认知诊断的同行互评关键技术研究

李秋云

广西大学

二〇二一年六月

硕士　李秋云　基于认知诊断的同行互评关键技术研究　2021

分类号_____ 密级_____

UDC _____ 编号_____

硕士学位论文

基于认知诊断的同行互评关键技术研究

李 秋 云

学科专业 计算机科学与技术

指导教师 许嘉 副教授

论文答辩日期 _____ **学位授予日期** _____

答辩委员会主席 _____

广西大学学位论文原创性和使用授权声明

本人声明所呈交的论文，是本人在导师的指导下独立进行研究所取得的研究成果。除已特别加以标注和致谢的地方外，论文不包含任何其他个人或集体已经发表或撰写的研究成果，也不包含本人或他人为获得广西大学或其它单位的学位而使用过的材料。与我一同工作的同事对本论文的研究工作所做的贡献均已在论文中作了明确说明。

本人在导师指导下所完成的学位论文及相关的职务作品，知识产权归属广西大学。本人授权广西大学拥有学位论文的部分使用权，即：学校有权保存并向国家有关部门或机构送交学位论文的复印件和电子版，允许论文被查阅和借阅，可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或其它复制手段保存、汇编学位论文。

本学位论文属于：

☑保密，在 2023 年解密后适用授权。

☐不保密。

(请在以上相应方框内打“ √ ”)

论文作者签名：　　日期：

指导教师签名：　　日期：

作者联系电话：　　电子邮箱：

基于认知诊断的同行互评关键技术研究

摘 要

大规模开放式在线课程（MOOCs）给教师带来了严峻的教学挑战，因为一名教师可能需要批改上千名学生提交的主观题作业。同行互评是解决大规模主观题作业批改问题的主流技术，具体分为基数估计和序数估计。这两种同行互评估计技术有不同的互评优势：一方面基数估计要求评价者对主观题作业给出绝对分数，比序数估计更能准确量化作业之间的差距；另一方面序数估计要求评价者对主观题作业给出相对排序，比基数估计来说，非专家的同行更容易做出相对判断。因此，基数估计技术和序数估计技术均是目前智能教育研究的关注热点。然而现有同行互评技术未充分考虑基于认知诊断得到的学生对于主观题作业知识点的掌握程度给真实分数估计带来的影响，因此很难准确的估计主观题作业的真实分数。针对现有研究工作的不足，本文研究了基于认知诊断的同行互评评判的关键技术，旨在基于基数和序数两种同行互评评判技术以提高作业真实分数估计的准确性。本文的研究内容包括：

（1） 研究了基于认知诊断的同行互评基数估计技术。首先以评价者的历史答题结果为输入，基于流行的认知诊断 DINA 模型量化评价者对主观题作业知识点的掌握程度；其后同时以评价者的知识点掌握程度以及在主观题作业中取得的真实分数对评价者的评分可靠性进行建模；最后结合对评价者评分偏见的建模提出了两个估计主观题

作业真实分数的概率图模型 PG_8 和 PG_9。基于多次真实课堂实验收集到的主观题互评数据进行实验评估，结果表明本文提出的基于认知诊断的同行互评基数估计技术对主观题作业真实分数的估计比相关基数估计技术更为准确，真实分数的估计误差平均降低了 42%。

（2） 研究了基于认知诊断的同行互评序数估计技术（命名为 BT+CD 技术）。首先基于诊断得到的评价者对主观题作业知识点的掌握程度设定评价者可靠性的先验分布；其后以序数估计活动中评价者对作业的配对排名顺序为输入，建立基于认知诊断的序数估计模型；最后提出有效的算法估计主观题作业的排名和评价者的可靠性。基于多次真实课堂实验收集到的主观题序数同行互评数据对 BT+CD 进行实验评估。实验结果表明，本文提出的 BT+CD 技术对主观题作业的估计比相关技术更具优势，比相关技术在配对顺序正确占比上平均提高了 18.38%。

基于认知诊断的同行互评技术是值得深入研究的问题，具有重要的理论意义和应用价值。本文研究了基于认知诊断的同行互评基数估计技术和序数估计技术。基于真实课堂实验收集的主观题基数和序数同行互评数据集进行实验评估，实验结果验证了本文提出的基于认知诊断的同行互评各个关键技术的有效性。

关键词：同行互评　认知诊断　DINA 模型　主观题　真实分数估计

RESEARCH ON KEY TECHNOLOGIES OF PEER GRADING BASED ON COGNITIVE DIAGNOSIS

ABSTRACT

The popularity of massive open online courses (MOOCs) has brought a severe teaching challenge to teachers, because a teacher may need to evaluate thousands of subjective questions submitted by students. Peer grading is the mainstream technology to solve the grading problem of large-scale subjective questions, which is divided into cardinal estimation and ordinal estimation. These two peer grading techniques have different advantages: on the one hand, cardinal estimation requires graders to give absolute scores for subjective questions, which is more accurate than ordinal estimation to quantify the different between subjective questions; on the other hand, ordinal estimation requires graders to give relative ranking for subjective questions, which is easier for peer graders of non expert to make relative judgments than cardinal estimation. Therefore, both cardinal estimation and ordinal estimation are the focus of current intelligent education research. The existing peer grading technology does not fully consider the impact of students' competency to subjective questions obtained by cognitive diagnosis on the true score estimation, so it is difficult to accurately estimate the true score of subjective questions. In view of the shortcomings of existing

research work, this paper studies the key technology of peer grading based on cognitive diagnosis, aiming to improve the accuracy of true score estimation based on cardinal estimation and ordinal estimation. This thesis mainly covers the following two research aspects.

(1) This thesis proposes a cardinal peer grading technology based on cognitive diagnosis. First, a grader's competence to a subjective question is computed, based on a popular cognitive diagnosis DINA model. Second, a grader's grading reliability to a subjective question is modeled by using both of the grader’s competence information to the question and the true score information the grader may obtained in the question. Finally, by combining the modeling of a grader's bias, two probability models (named as PG_8 and PG_9) are proposed to estimate the true scores of subjective questions in peer grading activities. A real-world peer grading dataset, which is collected based on peer grading activities of subjective questions, is employed to conduct the experimental evaluation. Experimental results show that the proposed cardinal peer grading technology based on cognitive diagnosis gets more accurate estimates for the true scores of subjective questions compared with the state-of-the-art technologies, and our proposal reduces the error of true score estimation by on average of 42% compared with state-of-the-art technologies.

(2) This thesis proposes a ordinal peer grading technology based on

cognitive diagnosis (named as BT+CD). First, a prior distribution of grader reliability is set based on grader's competence to subjective question. Secondly, an ordinal peer grading model is established with the input of the grader's multiple pairwise preferences in ordinal estimation activity. Finally, an effective algorithm is proposed to estimate the true score of the subjective questions and the grader's reliability. The proposed technology is evaluated through a real classroom practice participated by several teaching classes. Experimental results demonstrate that the proposed BT+CD technology successfully improves the accuracy of true ordinal estimation for pairwise preferences by on average of 18.38%, compared with related technologies.

Peer grading technology based on cognitive diagnosis is a problem worth further study, having important theoretical and practical significance. This thesis studies the cardinal peer grading technology based on cognitive diagnosis and ordinal peer grading technology based on cognitive diagnosis. Based on the data set of subjective question peer evaluation cardinal and ordinal collected from real classroom experiments, the experimental results verify the effectiveness of the proposed key technologies in peer grading based on cognitive diagnosis.

KEY WORDS: peer grading; cognitive diagnosis; DINA model; subjective question; true score estimation

目　录

第一章 绪论

1.1 研究背景和意义

随着大数据、云计算和互联网技术的不断发展，以 Coursera、edX、中国大学 MOOC 和学堂在线为代表的在线教育平台的兴起给平台上的任课教师带来了严峻的教学挑战。一个最突出的教学挑战在于教师如何高效批改大规模选课学生在平台上提交的作业。鉴于做作业能够帮助学生巩固和内化知识，是至关重要的教学活动，各大在线教育平台都提供了客观题（例如选择题和判断题）的自动批改功能，减轻了任课教师的教学负担。相对于客观题，主观题（例如简答题和应用题）更能考察学生的语言表达能力、知识运用能力与创新思维能力，所以主观题的考察对于很多在线课程而言是必不可少的[1]。然而，由于没有唯一标准答案，主观题的批改很难由计算机自动完成[2]，需要任课教师花费大量精力逐份手工批改，导致他们无法将精力用于课程内容及活动的改进提高。可见，如何减轻任课教师的主观题批改负担是当前教育研究领域亟待解决的重要问题。

为了有效降低任课教师的主观题作业批改负担，国内外各大在线平台与科研机构提出了不少主观题评判的技术，这些技术可分为两类：基于自然语言处理的评判技术[3][4][5]和基于同行互评的评判技术[6][7][8][9][10]。其中，基于自然语言处理的评判技术通过分析学生答案与教师给的参考答案之间的匹配程度来实现主观题的自动判分。然而，基于自然语言处理的评判技术通常依赖于特定领域的知识，只适用于解决面向特定领域的主观题评分问题，因此鲜有在线教育平台提供基于自然语言处理的主观题评判功能。基于同行互评的评判技术是当下不少主流在线教育平台（例如 Coursera 和中国大学 MOOC）提供的主观题评判功能。该类技术将主观题批改任务的子集分派给每个学生，然后基于多名学生对某主观题的评分来估计该题的真实分数。基于同行互评的主观题评判技术对于教师与学生而言都有积极益处：一方面减轻了任课教师的主观题作业批改负担；另一方面要求学生评判他人的主观题作业，不但能够让他们学习到不同的解题思路，还能提高他们的课程参与度[11][12]。因此，基于同行互评的主观题评判技术成为当下解决主观题评判问题的主流技术和目前智能教育研究的关注热点。

同行互评估计场景分为两种：基数估计和序数估计。基于基数估计或绝对评判的同行互评场景，即每名同行评价者针对每道主观题给出一个数值型的评价分数，这种场景是目前大多数 MOOC 平台采取的同行互评估计方式。基于同行互评的主观题基数估计方法的目标是最小化和真实分数之间的绝对估计偏差，研究难点在于如何利用多个同行给出的评价分数估计被评价者的真实分数。与基数估计不同，序数估计要求每名同行评价者针对评价的主观题给出一个相对的评价排序。基于同行互评的主观题序数估计方法的研究难点在于如何利用多个同行给出的评价排序得到每个被评价者的相对排名。这两种同行互评估计场景有不同的互评优势：一方面基数估计要求评价者对主观题作业给出绝对分数，比序数估计更能准确量化作业之间的差距[13][14][15]；另一方面序数估计要求评价者对主观题作业给出相对排序，比基数估计来说，非专家的同行评价者更容易做出相对判断[15][16]。因此，基于同行互评的基数估计技术和序数估计技术均是目前智能教育的研究热点。

在同行互评的基数估计中，大多数在线教育平台只是简单基于各个评价分数的均值或中位数来估计被评价者的真实分数。然而，由于同行评价者的打分质量受其可靠性、偏见等因素的影响[17]，简单用各个评价分数的均值或中位数估计被评价者的真实分数往往不够准确[18]。近年来，研究人员将同行评价者的评分可靠性及评分偏见作为模型的随机变量，构建了估计被评价者的主观题作业真实分数的概率图模型，能够利用变量间的依赖关系提高估计的准确性[6][7][8][9]。在同行互评的序数估计中，有学者提出根据评价者的可靠性建立同行互评序数估计模型[19]，以评价排序和可靠性、真实分数的概率关系估计全局排序。然而，现有同行互评基数和序数估计的研究方法假设同行评价者的可靠性只与其当前作业的答题情况相关，均未考虑同行评价者对主观题考察的知识点的掌握程度（由其历史答题结果数据诊断得到）对其评分可靠性造成的影响，因而存在局限性。

图 1-1 展示了 284 名同行评价者针对三道主观题作业给出的 2109 条互评打分记录的统计分析结果。以这些同行评价者的历史答题结果数据为输入并利用流行的认知诊断 DINA 模型[20]诊断得到他们对主观题考察的知识点的掌握程度，并进而估计每个同行评价者对每道主观题作业的潜在正确作答概率（等于评价者对主观题考察的各个知识点的掌握程度的乘积值）。以得到的每个同行评价者对每道主观题作业的潜在正确作答概率为依据将 2109 条互评打分记录分为五类，并统计每类中各条互评打分记录给出的评价

分数与真实分数之间的均方根误差（RMSE）。如图 1-1 所示，同行评价者的评分质量受该评价者对主观题考察的知识点的掌握程度的影响：掌握程度越低（即潜在正确作答概率越小），则评价者的平均评分误差越大，可靠性越低；掌握程度越高（即潜在正确作答概率越大），则评价者的平均评分误差越小，可靠性越大。因此，在基数和序数同行互评场景下，如何基于同行评价者对主观题考察的知识点的掌握程度信息对同行评价者的评分可靠性进行建模，从而基于同行互评的主观题评判方法有效提高评分准确性是需要研究解决的重要问题。

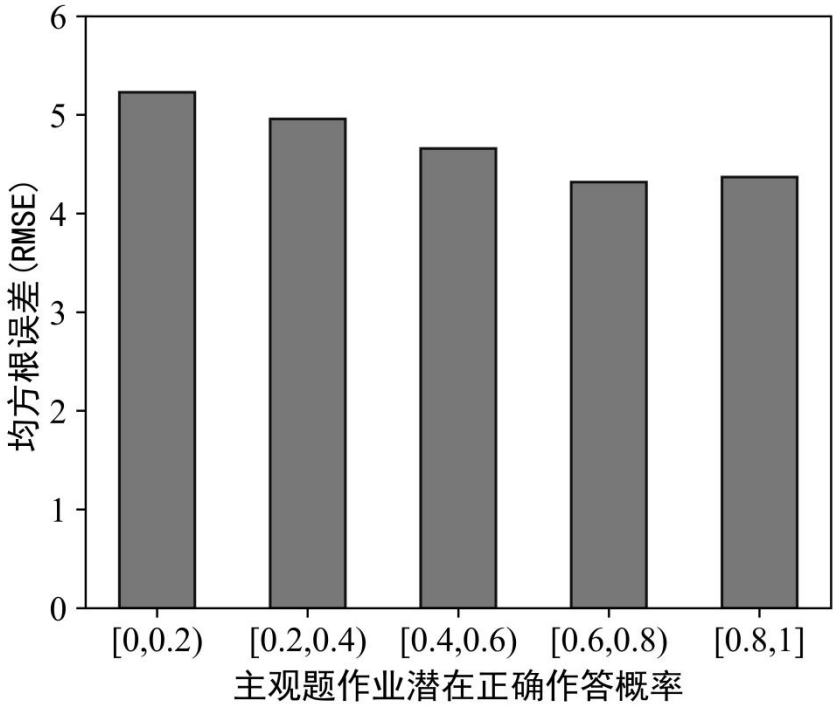


图 1-1 评价者对知识点的掌握程度与评分误差之间的关联性

Fig. 1-1 Correlation between graders' competence to knowledge points and RMSE

鉴于此，本文提出了基于认知诊断的同行互评关键技术，包括同行互评基数估计技术和序数估计技术。在同行互评基数估计技术中，提出了两个基数估计概率图模型 PG_8 和 PG_9。该技术在现有概率模型[9]的基础上，同时基于同行评价者在本次作业中的答题表现（对应于本次作业取得的真实分数）以及评价者的历史答题表现（对应于基于历史答题记录诊断得到的该评价者对本次作业题的掌握程度）对评价者的可靠性进行建模，以期在基数估计场景中最终提高概率模型估计主观题作业真实分数的准确性。PG_8 和 PG_9 的主要区别在于：PG_8 假设评价者的评分可靠性服从伽马分布；PG_9 则假设评价者的

评分可靠性服从高斯分布。在同行互评序数估计 BT+CD 技术中，提出了序数估计概率模型。BT+CD 技术基于认知诊断得到的评价者对作业知识点的掌握程度对其可靠性建模，并且基于互评评价排序、真实分数、可靠性的概率关系得到主观题作业的全局排序，从而提高序数估计场景中主观题作业排名顺序估计的有效性。

综上所述，本文围绕主观题互评的评判问题提出了基于认知诊断的同行互评基数估计技术和序数估计技术，能够有效解决大规模在线教育平台上针对主观题作业互评的真实分数或排名估计问题，具有重要的理论意义和应用价值。

1.2 国内外研究现状

1.2.1 信息聚合技术概述

在基数或序数同行互评中的真实分数估计问题可以看作一种特殊的信息聚合问题，同行互评中的聚合是将互评评价信息聚合以得到真实分数或全局排名。目前信息聚合方法已经广泛应用于不同的领域，下面分别从众包、论文和基金评审、教育领域的同行互评三个方面展开对信息聚合技术的描述。

（1） 众包

在众包中打标签者可能具有广泛的专业知识水平，众包的研究目标在于组合同一个项目的多个标签以估计众包项目的实际真实标签，该领域的研究目标和本文研究问题的联系最为密切。众包领域中的聚合方法主要分为：概率模型[21][22]、神经网络方法[23][24][25]、加权求和方法[26][27]、大多数投票[28][29]。

概率模型通过建立隐含变量之间的条件依赖关系，以可见变量为输入，从先验分布中取样以估计各个隐含变量的值。例如文献[22]基于建立每个打标签者的专业知识能力、每张图像的难度和图像标签是否正确的概率关系，以最大似然算法推断得到每张图像的真实标签、每个打标签者的专业知识能力和每张图像的难度值。神经网络方法是以每个众包项目的特征向量(如图像像素值)为输入，预测其真实标签。如在文献[25]中，提出了一种由神经网络层、真实分数层和考虑工人能力与偏见的众包层的三层网络构成的深度神经网络框架，其中神经网络层包括卷积层和完全连通的稠密层，对众包项目的输入特征向量进行一系列的非线性变换，并将变换结果输入到真实分数层，以预测每个项目

的真实标签的概率值。神经网络方法分类的准确度高，并行分布处理能力强，对噪声数据有较强的鲁棒性和容错能力，能够充分逼近复杂的非线性关系。但是神经网络方法还存在如下缺陷：需要大量的参数；不能观察中间学习过程；输出结果较难解释，还会影响到结果的可信度；需要较长的学习时间，当数据量较大时，学习速度会约制其应用。对于加权求和的方法而言，有研究工作提出通过一个问题的每个评价的质量估计对不同的评价分配权重，再使用加权求和的聚合方法推断真实值[26]。但是该方法存在如下缺陷：无法求解多个变量；权重确定不够精确；没有考虑影响权重的其他因素如偏见、评价者的专业知识能力、众包项目的难度。大多数投票方法是一种简单的启发式方法，表现为以大多数人的意见为最终的聚合结果，但是该方法不能够对评价结果建模，而是通过一些聚合规则直接识别真实标签[28][29]。

在以上聚合方法中，概率模型和另外的几种聚合方法不同，该方法具有如下优点：模型清晰，通过其表示随机变量之间的条件依赖关系，能被直观地理解，并且有较强的解释性。因此，本文研究基于概率模型的聚合方法对教育同行互评中的主观题作业的分数和排名进行估计，以此提高同行互评估计的有效性。

（2） 论文和基金评审

同行评审通常应用于期刊会议论文的录用和基金项目的选择。其中，期刊会议论文的同行评审的研究更关注于评审双方的匿名性[30][31]和论文审稿人的分配问题[32][33][34][35]，在聚合同行评审的意见时，大多直接简单地综合审稿人的意见作为最终的结果。而在基金项目选择中研究同行评审观点聚合的众多方法中，决策理论中的证据推理方法是目前应用最广泛的方法。证据推理方法可以系统的对不确定（模糊或不完整）的观点信息进行建模[36]，在不改变证据性质的情况下处理相互矛盾的评估信息[37]。证据推理规则是一种普遍的概率推理过程，可以同时考虑证据的权重和可靠性以组合多个独立的证据。许多学者利用 ER 方法对基金项目选择中的多个专家评价信息聚合进行了扩展研究并取得不错的效果[38][39][40][41]。

论文和基金项目评审中评审者的专业知识背景依赖较强，而主观题作业的同行互评中评价者的知识水平在同一知识框架下可能存在差异。因此，论文和基金项目评审的观点聚合方法不适用于同行互评中真实分数或排名顺序估计的研究。

（3） 教育领域的主观题评判技术

教育领域的主观题评判技术是一种聚合问题，该技术是当下的研究热点并已取得了不少研究成果，可分为两类：基于自然语言处理的评判技术和基于同行互评的评判技术。下面将分别对这两类技术的研究现状进行总结。

① 基于自然语言处理的评判技术

基于自然语言处理的主观题评判技术从题目本身的特性出发，利用自然语言处理、机器学习等技术实现主观题的自动评判。例如，文献[5]基于自然语言处理技术对开放式数学问题的每一个解答转变为数字特征，再通过聚类分析发现解答中正确、部分正确以及不正确的解答结构，从而实现了对该类问题的自动判分。文献[3]针对英文论文写作题给出了自动判分的解决方案，该方案利用潜在语义分析和学习向量量化算法来提升自动判分的准确率。文献[42]针对英语简答题设计了自动判分方法，该方法利用同义词词典和衡量语义距离的两种自然语言处理方法来解决标准文本相似度衡量方法对于同义词的匹配不够准确的问题。文献[4]则基于潜在语义分析的奇异值分解策略设计了日语短文的自动评分系统。

基于自然语言处理的主观题评判技术为主观题的自动评分提供解决思路，也取得了不错的评分效果。然而，该类技术通常依赖特定领域的知识来优化自然语言的处理过程，从而保证自动判分的准确性，因而只适用于解决特定领域的主观题自动判分问题，很难在其它领域推广使用。

② 基于同行互评的评判技术

基于同行互评的主观题评判技术[6][7][8][9][10]是当下解决主观题评判问题的主流技术和目前智能教育研究的关注热点。该技术是利用众包的思想，让每名同行评价者对分配给其的一部分主观题作业进行评判，最终基于各个评价者反馈的评判信息估计每份主观题作业的质量。按照同行评价者给出的评判内容形式的不同，基于同行互评的主观题评价方法可分为序数估计技术（Ordinal Peer Grading）和基数估计技术（Cardinal Peer Grading）两类。

教育领域中同行互评作业的分数或排名估计可以看作一种特殊的信息聚合。与众包、论文和基金评审两个领域应用场景下的信息聚合问题不同，教育领域的同行互评估计具有独特性。1）评审人不同。在众包领域中工人和打标签者是区分开来的，而在论

文和基金领域的同行评审中更多关注于评审人，更依赖于评审人的知识背景、历史表现等对评审人进行分配或评审意见的聚合。而在教育领域中的同行互评中，每位学生即是同行评价者也是被评价者。2）评审目的不同。众包应用的主要目的是获得正确的标签，不需要分析打标签者本身。在论文和基金领域的同行评审应用的目的在于 n 个项目中里面选前 k 个进行录用或资助。而教育领域的同行互评具有特殊的教育价值：一方面，学生可以从其他同伴的评价反馈中受益，互评过程中可以学习解决同一问题的不同思路；另一方面，教师可以从偏见和可靠性反馈中得知学生的学习情况，进一步有针对性的调整教学方案。

1.2.2 同行互评中的基数估计技术

同行互评中的基数估计技术是当下的研究热点并已取得了不少研究成果。基数估计技术要求每名评价者对分配给其的每份主观题作业都给出一个量化分数，系统继而基于不同评价者针对同一份作业给出的多个评价分数估计作业的真实分数。主流的基数估计方式有两种：加权求和估计法和基于概率模型的估计法。下面将对这两种基数估计方法进行介绍。

（1） 加权求和估计法

加权求和估计法依据同行评价者的评分准确性和信任度给他们赋以不同的权重，然后以同行评价者针对主观题作业给出的评价分数为输入，通过加权求和的方法来估计该作业的真实分数。Alfaro 等人提出了 *Vancouver* 聚合算法[43]，该算法中的学生最终得分是互评得分、被评价的学生对同行评价者给出的反馈帮助分以及同行互评的评判准确性分数三个方面的加权得分。在每次同行互评评判后，系统会根据同行评价者在新的互评活动中的评分表现来迭代更新其权重信息。Walsh 提出了另一种迭代加权算法 *PeerRank*[44]，该算法的提出是受到了对网页进行排序的 *PageRank* 算法[45]启发，假设一个评分者的作业分数反映了评价能力，基于同行评价者的分数对每一份提交作业的多个评价者的评价分数进行加权，即学生提交作业的真实分数依赖于其他学生的作业分数。Gutierrez 等人则提出了一种基于信任的估计算法，通过学生之间的互动度量学生的被信任程度以对互评分数进行加权计算[46]。

基于加权求和的估计方法只是对学生的最终作业分数采取不同权重的计算，忽略了学生在评价过程中的可靠性和带有偏见的评估状态，并且在分数加权的权重设定上存在

一定的主观因素。

（2） 概率模型估计法

概率模型估计法是通过构建概率模型来估计主观题作业的真实分数。本文提出的基于认知诊断的主观题互评基数估计技术就属于这类方法。这类方法的主要实现思路是将待估计的主观题作业的真实分数、同行评价者的可靠性及偏见都建模为满足一定概率分布的隐含变量，然后基于能观察到的同行评价者的评分信息来推演以上各个隐含变量的值。具体而言，Piech 等人[6]首先提出了估计主观题作业真实分数的三个概率模型，即 PG_1（考虑了评价者当前的可靠性和偏见），PG_2（在 PG_1 的基础上考虑了评价者的历史偏见），PG_3（在 PG_1 的基础上将评价者当前可靠性设定为评价者当前作业真实分数的线性函数的随机变量）。考虑到 PG_3 模型所设置的评价者的可靠性是关于评价者真实分数的线性函数这一假设过于严格，Mi 等人将评价者的可靠性建模为满足形状参数为其真实分数的伽马分布或均值为其真实分数的高斯分布，分别得到了 PG_4 模型和 PG_5 模型[7]。研究表明一名同行评价者的评分偏见会受到其朋友的评分偏见的影响[47][48]，为了提高对评价者偏见建模的准确性，有学者利用学堂在线平台上收集到的学生间的社交关系信息优化对评价者偏见的建模，扩展了 PG_1、PG_4、PG_5 这三个概率模型[8]。然而上述概率模型均认为评价者针对不同主观题作业给出的评价分数之间是相互独立的，存在局限性。因此，Wang 等人在概率建模时引入了评价者的相对分数信息（即同一个评价者对不同作业评分之间的差值），提出了 PG_6 模型（构建在 PG_4 之上），PG_7 模型（构建在 PG_5 之上）[9]。这两个概率模型由于引入了同行评价者的相对分数信息，降低了数据稀疏性给参数估计带来的负面影响，从而有效提高了对主观题真实分数估计的准确性。然而，PG_6 模型与 PG_7 模型仅基于同行评价者针对当前主观题作业取得的真实分数对其可靠性进行建模。PG_6 模型与 PG_7 模型是当前估计效果最好的同行互评基数估计概率模型，实验部分将针对这两种相关模型进行比较分析。

综上，基于概率模型的方法是目前实现主观题作业的同行互评基数估计的主流方法，近年来研究人员们提出了不少相关工作。然而，现有研究工作在概率建模时均未同时考虑影响同行评价者评分可靠性的两大因素，即其在本次作业中的答题表现（对应于本次作业取得的真实分数）以及其历史答题表现（对应于基于历史答题记录诊断得到的该评价者对本次作业题的掌握程度），因而限制了对主观题真实分数的估计准确性。

1.2.3 同行互评中的序数估计技术

同行互评中的序数估计技术是当下的研究热点并已取得了不少研究成果。序数估计是一种特殊的排名聚合技术，在排名聚合领域已有大量的研究工作，文献[19]对排名聚合方法进行了的总结。这些排名聚合方法包括 Mallows（MAL）方法[49]，该方法提出定义一组作业排名的概率分布为排序距离的函数。此外该方法还使用肯德尔等级相关系数度量指标快速地进行计算，从而可以从一组部分观察到的排名中找到全局排名的最大似然估计。MALS 是一种扩展 Mallows 的方法，该方法估计两两配对项目之间的排名距离，以提高推理的鲁棒性。另一个重要的方法是 Plackett-Luce（PL）[50]，该方法基于经典的配对排序聚合模型 Bradley-Terry（BT）[51]扩展，逻辑损失函数具有凸函数特性的优点，可以利用优化技术快速求解。

与基数估计技术不同，序数估计技术是一种只需做出相对评判的技术，即要求每名同行评价者对分配给其的主观题作业给出表征作业质量高低的排名反馈，系统则基于所有同行评价者给出的作业间的偏序排名信息估计每份作业的质量[52]。序数估计的方法不要求同行评价者给出主观题作业的具体分数，降低了评价者的评判难度。现有的序数估计技术通常利用贝叶斯生成法和矩阵分解方法、基于配对比较等方法来估计主观题作业的质量。为了更好地将本文工作同相关研究工作进行对比，下面将对现有方法分析同行互评中的序数估计技术的研究现状。

Waters 等人[53]基于贝叶斯方法解决同行互评中的序数估计问题，利用排序的同行互评评价数据提出一种新的序数估计模型，基于 MCMC 方法推断出所有模型参数和参数的可靠性信息。和传统的最大似然估计方法不同，基于贝叶斯方法的序数估计技术能够利用参数之间的概率关系对模型参数的可靠性进行很好的估计，然而该方法未完全利用排序配对的互评数据信息，这大大影响了序数估计结果的准确性。而 Luacesa 等人[54]提出了基于因式分解法实现同行互评的技术，是一种在序数估计和基数估计之间寻求折衷的方法。该方法学习评价者的偏好评判，避免了绝对的数值型分数带有的主观性。除了评价者本身的偏好之外，还包括由平时成绩显著性不同的作业引起的偏好。该方法具有较快的速度处理大量的作业互评，然而未对作业真实分数和偏好之间的关系准确建模，并且没有对学生的可靠性进行分析，存在局限性。近年来还有学者提出利用模糊数学理论解决序数同行互评中的群决策问题，基于模糊群决策的序数估计不仅对作业的好坏进

行排序，还利用变量表示评价者对要评估的主观题作业之间的偏好程度[55]。

配对比较法指的是在同行互评的过程当中，评价者对要评价的作业进行两两比较，基于配对比较分析两份作业的评价排序和真实分数之间的概率关系。Shah 等人[16]提出改进经典的配对比较模型 BT，以结合评价者的评估能力得到扩展的序数估计模型 RBTL，从有序的配对比较中学习学生的潜在能力并且执行交叉验证预测同行评估偏好的实验。Raman 等人[19]提出扩展一些统计的排名聚合模型（包括 BT[51]、MAL[49]、Thurstone[56]、PL[50]）用以解决同行互评中的序数估计问题，并且在扩展的模型中引入了同行评价者的可靠性变量，使用迭代交叉最大似然估计技术得到真实分数和评价者的可靠性，最终实验表明引入了可靠性变量的序数估计模型提升了真实分数估计的有效性和精确性。

综上所述，基于配对比较的方法是目前实现主观题作业的同行互评序数估计的主流方法，近年来研究人员们提出了不少相关工作。然而，现有研究工作在对可靠性建模时未考虑到同行评价者对主观题作业知识点的掌握程度，这大大影响了主观题作业的同行互评序数估计技术的有效性和精确性。

1.2.4 国内外研究现状总结

根据上述相关研究工作的成果分析，目前国内外有较多的工作在不同的领域对同行互评技术展开研究，并且有较多学者在同行互评中的基数和序数估计两个方面展开研究，并取得了一定的成果。但是这些研究工作仍存在以下几点不足和可以改进的地方：

（1）在同行互评中的基数估计方面，基于概率模型的估计技术是目前主观题作业的同行互评基数估计主流技术，现有的基于概率模型的研究工作要么是未考虑同行评价者的可靠性，要么是仅仅基于当前作业的真实分数对可靠性建模，未对同行评价者的可靠性准确建模，导致相关工作存在局限性，影响了同行互评基数估计的准确性。

（2）在同行互评中的序数估计方面，现有研究工作中多数是基于配对比较的方法实现主观题作业的序数估计，虽然在基于配对比较的研究工作中有学者提出在扩展的模型中引入了同行评价者的可靠性变量，但是没有考虑到在主观题作业中评价者对主观题作业知识点的掌握程度对于可靠性的影响，导致对评价者可靠性的估计不够准确，进而使得在序数同行互评中的对主观题作业的估计无法达到较高的精度。

1.3 本文研究工作

鉴于 1.2.4 小节所提出的国内外研究工作的不足之处，本文研究与实现了基于认知诊断得到的学生对于知识点的掌握程度，并基于认知诊断技术提出新颖的同行互评基数估计技术和序数估计技术，旨在有效提升基数同行互评和序数同行互评两方面的估计准确性。本节将在 1.3.1 节详细介绍本文的研究内容，在 1.3.2 节阐述研究工作的创新点。

1.3.1 研究内容

（1）基于认知诊断的基数估计。鉴于同行互评的基数估计技术相关工作中的不足，本文提出了一种基于认知诊断的通同行互评基数估计技术。该技术首先基于认知诊断模型得到同行评价者对知识点的掌握程度，以评价者的历史答题表现（对应于基于历史答题记录诊断得到的该评价者对本次作业题的掌握程度）以及同行评价者在本次作业中的答题表现（对应于本次作业取得的真实分数）对可靠性建模，再结合评价者的评分偏见提出了两个基数估计概率图模型，命名为 PG_8 和 PG_9。这两个模型的主要不同在于：PG_8 模型的可靠性先验分布为伽马分布，PG_9 模型的可靠性先验分布为高斯分布。其次，对概率图模型中的隐含变量进行估计，包括被评价者的真实分数、同行评价者的可靠性和偏见。推断得到两个 PG_8 和 PG_9 模型中所有隐含变量的近似后验分布。最后，以可见变量互评分数、相对分数、评价者对知识点掌握程度信息为输入，基于流行的 Gibb 采样技术对两个概率模型的隐含变量进行采样估计，得到每个被评价者的主观题作业的真实分数，将估计的真实分数和教师给出的作业分数对比以评估本文提出技术的有效性。

（2）基于认知诊断的序数估计。针对同行互评的序数估计技术相关工作中的不足，本文提出了一种基于认知诊断的同行互评序数估计技术，命名为 BT+CD。BT+CD 技术首先基于流行的 DINA 模型诊断得到每个同行评价者对主观题作业知识点的掌握程度，再以掌握程度信息对评价者的评分可靠性建模。其次，以同行互评活动中的配对排名顺序对为输入，建立配对排名顺序和评分可靠性、被评价者提交的主观题作业真实分数之间的概率关系。最后，采用最大后验估计法估计序数同行互评中的评分可靠性和作业真实分数，并以随机梯度下降法优化两个隐含变量的估计。

1.3.2 创新之处

（1） 发现基于认知诊断模型诊断得到同行评价者对主观题考察的知识点的掌握程

度信息，可以有助于提高基于同行互评的主观题评判方法的评分准确性。

（2） 提出一种基于认知诊断的同行互评基数估计技术。该技术提出了改进现有基数同行互评概率模型的思路，即应同时以认知诊断得到的同行评价者对主观题的掌握程度信息和评价者在该主观题中取得的真实分数信息作为评价者评分可靠性的建模依据，并结合评价者的评分偏见，设计了两个概率图模型 PG_8 和 PG_9，还基于 Gibbs 采样技术推断概率图模型中的隐含变量，设计并实现了基于认知诊断的基数估计推断算法，以期进一步提高在基数同行互评活动中主观题作业真实分数的估计准确性。

（3） 提出了一种基于认知诊断的同行互评序数估计技术 BT+CD。该技术基于认知诊断 DINA 模型得到的评价者对主观题的掌握程度信息提出了序数估计模型，设计并实现了基于认知诊断的序数估计算法，进而提升了在同行互评活动中主观题作业序数估计的准确性。

（4） 通过组织计算机科学与技术专业的 284 名本科生于自主研发的在线教学服务系统中完成 3 次主观题作业收集真实的互评数据集，此外还收集了 40 道历史测试客观题数据集。基于学生提交的客观题答案使用流行的认知诊断 DINA 模型计算学生对知识点的掌握程度。继而基于主观题基数同行互评数据对本文所提出基于认知诊断的同行互评基数估计技术进行实验，分析该技术相对于其它同行互评基数估计技术的准确性，主观题作业的真实分数估计误差平均降低了 42%。同时使用主观题作业中互评收集到的偏序对信息对本文所提出基于认知诊断的序数估计 BT+CD 技术进行实验，通过与相关工作的对比证明了本文所提出技术的有效性和准确性，BT+CD 技术比相关的序数估计技术在配对顺序正确占比上平均提高了 18.38%。

1.4 论文组织结构

论文一共分为五章，论文结构如图 1-2 所示。

第一章是绪论，详细介绍了本文的选题研究背景和意义，分析了当前国内外研究现状，并对本文的研究工作进行总结。

第二章是背景知识和相关技术，详细描述了认知诊断 DNIA 模型对学生知识状态的诊断和同行互评技术的重要概念，并对同行互评基数估计和序数估计的相关技术进行了介绍。

第三章是基于认知诊断的同行互评基数估计技术，主要阐述了本文提出的基于认知诊断的基数估计概率图模型（包括 PG_8 模型和 PG_9 模型）的具体设计与模型推断，详细描述了概率模型的生成过程以及基于 Gibbs 采样技术设计的隐含变量推断算法等，并通过真实课堂实验证明了该技术的有效性。

第四章是基于认知诊断的序数估计技术 BT+CD，详细介绍了本文提出的基于认知诊断的序数估计模型的具体设计与实现算法，并最终通过真实课堂中的同行互评活动数据集验证了该技术的有效性和可行性。

第五章是总结与展望，主要对全文进行归纳和总结，并对未来研究工作进行了展望。

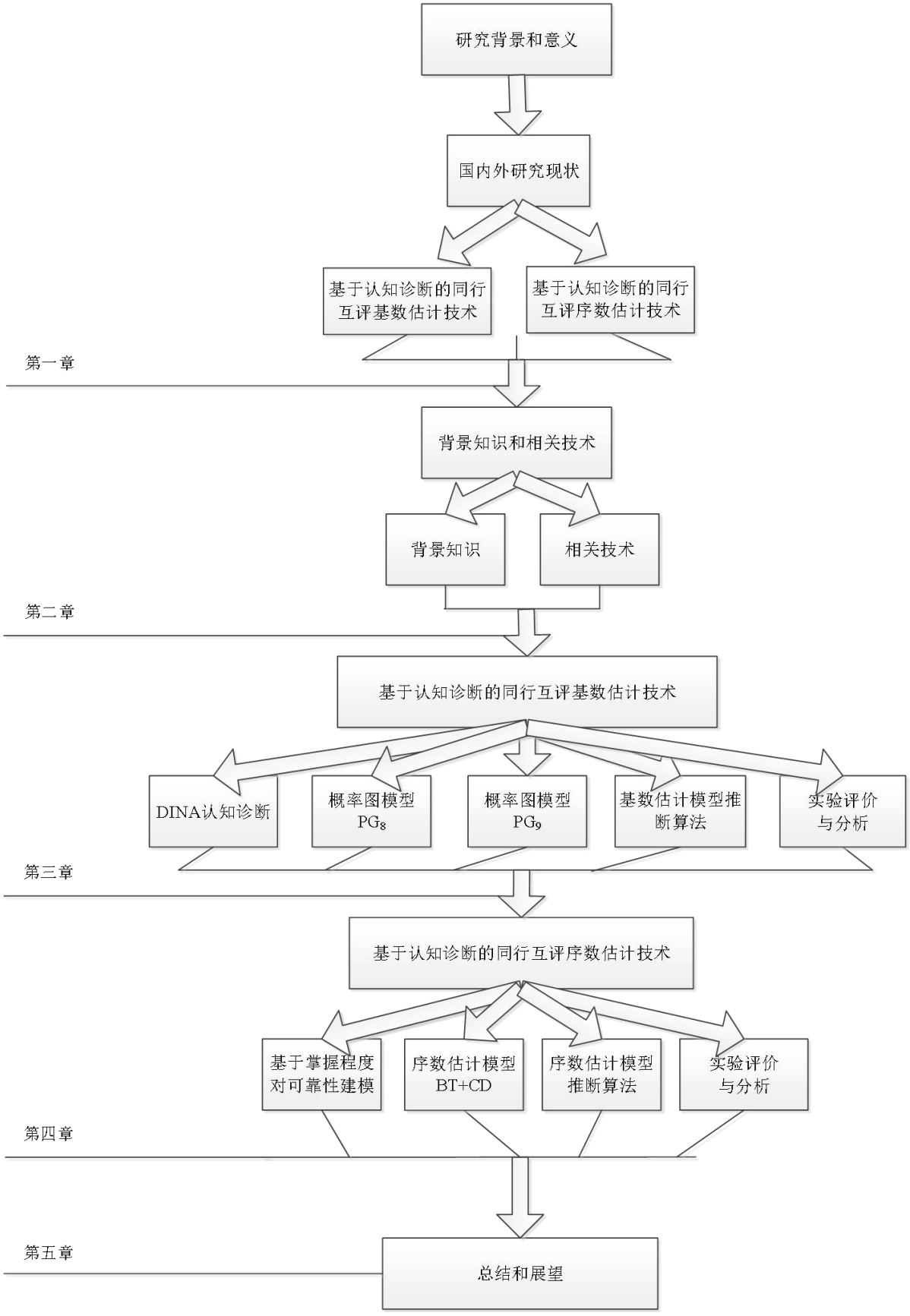


图 1-2 论文框架图

Fig. 1-2 The framework of the thesis

第二章 背景知识和相关技术

本章首先介绍本文涉及的相关背景知识，包括认知诊断技术和同行互评技术的重要概念，其次对本文所进行实验对照的相关技术进行原理分析，包括同行互评基数估计技术中位数、均值、PG_6 和 PG_7、序数估计技术 BT、RBTL、BT+G。

2.1 背景知识

2.1.1 认知诊断技术

认知诊断以认知心理学和心理计量学为理论基础，通过构建具有认知诊断功能的心理计量模型，能够基于被试的历史答题结果数据诊断其对不同技能（知识点）的掌握程度，从而为教学提供重要依据，是当下教育评估领域的研究热点[57][58][59]。作为最流行的认知诊断模型之一，DINA 模型[15]在实现对被试知识点掌握程度的精准建模的同时具有较好的解释性，近年来受到广泛的关注和研究[60][61]。以同行评价者的历史答题结果数据为诊断基础，本文正是基于 DINA 认知诊断模型来量化评价者对主观题作业的掌握程度。

给定被试集合 $C=\{c_1, ..., c_M\}$，习题集合 $E=\{e_1, ..., e_N\}$，则记录被试和其答题结果之间关联关系的响应矩阵 $\boldsymbol{R}$ 可表示为 $\boldsymbol{R}=[r_{mn}]_{M\times N}$，其中 $r_{mn}=1$ 表示被试 c_m 答对了习题 e_n（$r_{mn}=0$ 则表示答错了该题）。设习题集合 E 考察的知识点集合为 $KP=\{kp_1, \ldots, kp_K\}$，则记录习题与其考察的知识点之间关联关系的 $\boldsymbol{Q}$ 矩阵可表示为 $\boldsymbol{Q}=[q_{nk}]_{N\times K}$，其中 $q_{nk}=1$ 表示习题 e_n 考察了知识点 KP_k（$q_{nk}=0$ 则表示未考察该知识点）。DINA 模型将被试 c_m 的知识状态描述为一个向量 $\boldsymbol{\alpha}_m=\{\alpha_{m1}, ..., \alpha_{mK}\}$，称为被试 c_m 的知识点掌握程度向量。其中，α_{mk} 表示被试 c_m 对知识点 kp_k 的掌握程度，且 $\alpha_{mk}\in[0,1]$。$\alpha_{mk}=1$ 说明被试 c_m 完全掌握了第 k 个知识点；$\alpha_{mk}=0$ 则说明被试 c_m 完全没有掌握第 k 个知识点。DINA 认知诊断模型的项目反应函数为：

$$P(r_{mn}=1\,|\,\alpha_m)=guess_n^{(1-\delta_{mn})}(1-slip_n)^{\delta_{mn}} \tag{2-1}$$

其中，

$$\delta_{mn}=\prod_{k=1}^{K}\alpha_{mk}^{q_{nk}} \tag{2-2}$$

在公式（2-2）中，δ_{mn}表示知识状态为$\boldsymbol{\alpha}_m$的被试c_m对习题e_n的潜在正确作答概率，即可被定义为被试c_m对习题e_n的掌握程度值；$slip_n=P(r_{mn}=0 \mid \delta_{mn}=1)$表示被试掌握习题$e_n$考察的所有知识点但是答错该题的概率，被称为失误参数；$guess_n=P(r_{mn}=1 \mid \delta_{mn}=0)$指被试没有掌握习题$e_n$考察的任何一个知识点时但答对该题的概率，被称为猜测参数。DINA 模型利用 EM 算法最大化公式（2-1）的边缘似然值，从而得到被试c_m的知识点掌握程度向量$\boldsymbol{\alpha}_m$。

本文假设参与主观题互评活动的同行评价者在进行主观题作业评判之前完成了该主观题考察的知识点所对应的客观题的习题练习，因而作业互评测试系统能够收集到他们对于这些知识点对应的客观习题的答题结果数据。以某同行评价者的历史答题结果数据和表征习题和主观题作业知识点间考察关系的$\boldsymbol{Q}$矩阵为输入，利用 DINA 认知诊断模型即可求得该同行评价者的知识点掌握程度向量$\boldsymbol{\alpha}$。然后基于$\boldsymbol{\alpha}$和主观题作业所考察的知识点信息即可以利用公式（2-2）求得该同行评价者对于该主观题的掌握程度值。

2.1.2 同行互评问题定义

在一个典型的同行互评场景中，给定提交主观题作业的被评价者（学生作为被评价者）集合$U=\{u_1, \ldots, U_{|U|}\}$，被评价者提交的主观题作业集合$A=\{a_1, \ldots, A_{|A|}\}$（如简答题、期末作业报告），参与互评的同行评价者集合$V=\{v_1, \ldots, V_{|V|}\}$，考虑到实际教学实践中一般要求提交主观题作业的被评价者都参与该作业的互评活动，因而有$|U|=|V|$。主观题作业评价由同行评价者v完成，每个评价者评估主观题作业的子集A_v（$A_v \subset A$）。每个评价者对要评价的主观题作业是均匀随机分配的，或者遵循确定性或按一定的顺序原则进行分配。在这两种互评作业分配的情况下，任何同行评价者评估的主观题作业数$|A_v|$远小于作业总数$|A|$。每位同行评价者对要评价的主观题作业A_v提供反馈。根据同行评价者给出的反馈类型上的不同分为基数同行互评和序数同行互评。

（1）基数同行互评（Cardinal Peer Grading）：在基数同行互评中，每个同行评价者v对于要评价的每一份作业$d \in A_v$给出基数分值反馈，基数评估是一种绝对的质量评估方式。下面对基数同行互评技术涉及的重要概念定义和研究目标进行阐述。

真实分数（True grade）：假设每份被评价者提交的主观题作业对应一个该作业的真实分数，且用s_i表示被评价者$u_i \in U$所提交作业的真实分数。

可靠性（Reliability）：可靠性（记为τ_v），表示同行评价者$v \in V$对主观题作业的

评分精度。评价者 v 的可靠性实际反映了 v 给出的主观题作业的评价分数基于其偏见 b_v 修正后的分数与主观题作业真实分数之间的接近程度。给定某主观题作业，基于认知诊断的基数估计技术分别假设评价者 v 对于该作业的评分可靠性 τ_v 满足形状参数为 $\theta_1\delta_v+\theta_2 s_v$ 的伽马分布，得到 PG8 模型。同时假设 τ_v 满足均值为 $\theta_1\delta_v+\theta_2 s_v$ 的高斯分布，得到 PG9 模型。其中，δ_v 表示基于认知诊断得到的 v 对该作业的潜在正确作答概率。即同行评价者 v 对该作业的潜在正确作答概率越高其对该作业的评分可靠性越高。

偏见（Bias）：偏见（记为 b_v）是量化同行评价者 $v\in V$ 评分时表现出其评分高于真实分数或其评分低于真实分数的常量。例如，假设被评价者 u_i 的主观题作业的真实分数 s_i=10，评价者 v 的偏见 b_v=-2，则评价者 v 针对该主观题作业给出的互评分数 z_i^v 的均值为 $z_i^v = s_i + b_v = 10 + (-2) = 8$。

互评分数（Peer grade）：互评分数（记为 z_i^v）表示同行评价者 $v\in V$ 针对被评价者 u_i 提交的主观题作业给出的评价分数。设定所有评价者的互评分数集合为 $Z = \{z_i^v \mid u_i \in U, v \in V\}$。

相对分数（Relative peer grade）：相对分数（记为 d_{ij}^v）表示同行评价者 $v\in V$ 对被评价者 $u_i\in U$ 和 $u_j\in U$ 的主观题作业给出的互评分数间的差值。记面向所有评价者的相对分数集合为 $D = \{d_{ij}^v \mid u_i, u_j \in U, v \in V\}$。

基于认知诊断的同行互评基数估计技术与仅基于主观题作业互评分数信息构建的估计主观题作业真实分数的概率模型不同[6][7][8]，该技术考虑了基于同行评价者历史答题结果数据诊断得到的知识点掌握程度信息对主观题作业真实分数估计的影响。综上，该技术的研究目标是通过构建能够表征互评分数、相对分数、评价者偏见、评价者可靠性与被评价者提交的主观题作业的真实分数之间的关系的概率图模型，以有效估计主观题作业的真实分数。具体而言，该技术的研究问题的形式化定义为：已知所有同行评价者的互评分数集合 Z，面向所有评价者的相对分数集合 D，所有评价者的知识点掌握程度向量 $\boldsymbol{\alpha}$ 构成的矩阵 $\boldsymbol{M}_{|V|\times|KP|}$，通过训练基数估计技术的概率图模型，求解出每个同行评价者（即 $\forall v\in V$）的可靠性 τ_v、偏见 b_v 以及每个被评价者（即 $\forall u_i\in U$）提交的主观题作业的真实分数 s_i。

（2）序数同行互评（Ordinal Peer Grading）：在序数同行互评中，每个同行评价者 v 对于要评价的作业 A_v 给出评价的序数排名反馈（可能存在相同排名的反馈），和基数评估不同，序数评估是一种相对的质量评估方式。下面对基数同行互评技术涉及的研究目标进行阐述。

本文研究发现基于同行评价者历史答题结果数据诊断得到的知识点掌握程度信息，有助于提高同行互评主观题的评分准确性。鉴于此，利用认知诊断得到评价者对知识点的掌握程度信息不仅能够优化前文阐述的基数估计技术，还能应用于同行互评序数估计技术。因而，本文提出了基于认知诊断的同行互评序数估计技术，与序数估计相关工作未考虑可靠性因素或者仅仅简单的对可靠性建模不同，该技术基于掌握程度信息对同行评价者的评分可靠性建模，进而提升被评价者的作业排名估计的准确性。综上，该技术的研究目标是已知所有同行评价者的序数反馈信息和所有评价者的知识点掌握程度向量$\boldsymbol{\alpha}$构成的矩阵 $\boldsymbol{M}_{|V|\times|KP|}$，通过训练序数估计模型，求解出每个同行评价者（即$\forall v \in V$）的可靠性$\tau_v$、以及被评价者提交的所有主观题作业 D 的排名顺序。

2.2 相关技术

2.2.1 相关基数估计技术

（1）中位数

基数估计技术中位数采用一份主观题作业所获得的所有评价分数的中位数估计该作业的真实分数，这也是当今大多数提供主观题互评功能的 MOOC 平台（例如中国大学 MOOC 和 Coursera）采用的估计主观题作业真实分数的方法。

（2）均值

均值估计技术用一份主观题作业所获得的所有评价分数的均值估计该作业的真实分数，该技术是一种较简单的同行互评基数估计方法。

（3）PG_6 和 PG_7

PG_6 和 PG_7[9]均是解决主观题同行互评问题的现有最先进基数估计技术。这两种技术均引入了概率模+型对同行评价者的互评可靠性和互评偏见进行建模，且利用同行评价者对不同主观题作业打分的相对分数信息来提高概率模型对真实分数估计的准确性。

PG_6 技术与 PG_7 的区别在于：PG_6 假设同行评价者互评可靠性取值的先验分布为伽马分布，PG_7 则假设同行评价者互评可靠性取值的先验分布为高斯分布。本文提出的 PG_8 与 PG_9 模型分别是在 PG_6 和 PG_7 模型的基础上对评价者可靠性进行了建模优化。具体而言，PG_6 和 PG_7 模型在评价者可靠性时仅考虑了其在当前主观题作业中的答题表现，而 PG_8 与 PG_9 模型在对评价者的可靠性进行建模时不但考虑了其在当前作业中的答题表现，还考虑了基于其历史答题表现诊断得到的评价者对待评价作业的掌握程度信息，从而提高概率模型对主观题作业真实分数估计的准确性。需要说明的是：1）PG_8 与 PG_6 相对应，均假设同行评价者互评可靠性服从的先验分布为伽马分布；2）PG_9 与 PG_7 相对应，均假设同行评价者互评可靠性取值服从的先验分布为高斯分布。

2.2.2 相关序数估计技术

（1）BTL

BTL[51]是 1952 年 Bradley 提出的一个经典的配对比较模型。BTL 模型的一般表达式如下：

$$P_{BTL}(j \succ l) = \frac{1}{1+\exp(-(w_j - w_l))} \tag{2-3}$$

其中，同行评价者（学生作为同行评价者）i 评估另外两个学生 j 和 l 的主观题作业。w_j 表示被评价者 j 本身具有的内在能力（如对作业的答题能力）。BTL 以同行互评的排序对为输入，基于公式（2-3）的概率关系估计每一个学生的真实答题能力。

（2）RBTL

RBTL[16]是一种扩展 BTL 模型的序数估计方法。该方法类似于同行互评基数估计模型 PG_3 的方式结合了同行评价者的评价能力 g_i，一般表达式如下：

$$P_{RBTL}(j \succ l) = \frac{1}{1+\exp(-g_i(w_j - w_l))} \quad where\ g_i = aw_i + b \tag{2-4}$$

其中，a 和 b 为两个超参数，评价能力 g_i 由同行评价者 i 本身具有的内在能力 w_i 的线性表达式构成（$g_i=aw_i+b$），以更有效的衡量被评价者 j 和 l 的两份主观题作业之间的概率关系，可见 g_i 的值随着同行评价者 i 本身具有的内在能力 w_i 线性增加。g_i 的值有不同的物理意义：当 g_i 远远大于 0 时表示一个有能力评判的同行评价者，会预测 $j > l$ 当

且仅当 $w_j > w_l$；当 $g_i<0$ 时表示一个恶意评判的同行评价者，对主观题作业排名顺序的预测趋势和正确排名顺序是相反的；当 $g_i=0$ 时表示一个随意评判的同行评价者，对被评价者 j 和 l 的主观题作业的排序评判有同等的选择概率。从公式（2-4）中可看出，当赋值 $a=0$ 且 $b=1$ 时 RBTL 模型等同于 BTL 模型。

（3）BT+G

BT+G[19]是一种类似于同行互评基数估计模型 PG_1 的方式结合同行评价者的评分可靠性的序数估计方法。基于在真实课堂收集得到 170 个学生互评的主观题作业基数分数数据，将主观题基数分数转化为主观题序数排名顺序（可能存在相同的排名顺序）。BT+G 对同行评价者的可靠性建模，该方法以作业的排名顺序为输入实现了对评价者的可靠性和作业真实分数的估计。该方法的具体实现过程：定义一个要估计的隐含变量可靠性 $\hat{\eta}_g$，并且使用伽马分布作为可靠性变量的先验分布，并设定伽马分布的形状参数（Shape parameter）为 10，尺度参数（Rate parameter）为 0.1。可靠性变量的一般表达式如下：

$$\hat{\eta}_g \sim Gamma(10,0.1) \tag{2-5}$$

基于同行评价者的评分可靠性对序数估计模型优化，BT+G 模型的一般表达式如下：

$$P_{BT+G}(i \succ j) = \frac{1}{1+\exp(-\hat{\eta}_g(s_{d_i} - s_{d_j}))} \tag{2-6}$$

2.3 本章小结

本章围绕与本文研究内容相关的背景知识和相关技术进行了分析。2.1 小节介绍了认知诊断概念和典型的代表模型、同行互评技术的重要概念。2.2 小节阐述了对本文实验对比的两类同行互评估计相关技术，其中 2.2.1 节详细分析了四种基数估计技术：传统的基数估计技术中位数和均值、最新且具有较好实验效果的 PG_6 和 PG_7 技术；2.2.2 节详细阐述了三种与本文研究最相关的序数估计技术：BT、RBTL、BT+G。通过对相关技术的原理和机制的分析，充分体现了本文提出方法的创新之处。

第三章 基于认知诊断的同行互评基数估计技术

本章将详细介绍基于认知诊断的同行互评基数估计技术，分别阐述提出的两个基数估计概率图模型 PG_8 和 PG_9 的设计思想、模型的推断思路，并且对提出的基数估计技术的推断算法流程进行了阐述。最后通过真实的同行互评数据集，验证了提出的技术对主观题作业中真实分数估计的有效性和准确性。

3.1 问题的分析与提出

同行互评是目前大多数国内外在线教育平台（如中国大学 MOOC、学堂在线、Coursera）用以解决大规模作业批改问题的重要手段。受评价者的评分可靠性和评分偏见的影响，基于多个评价者给出的评分值估计主观题作业的真实分数充满挑战。近年来，研究人员基于概率模型对评价者的评分可靠性和评分偏见进行建模，提高了主观题作业真实分数估计的准确性。易知，评价者针对某主观题作业的评分可靠性受两方面因素的影响：一是其在本次作业中的答题表现（对应于本次作业取得的真实分数）；另一方面是其的历史答题表现（对应于基于历史答题记录诊断得到的该评价者对本次作业题的掌握程度）。然而，现有同行互评基数估计技术在对评价者的评分可靠性进行建模时均未同时考虑这两方面的影响因素。因此，本文提出了基于认知诊断的同行互评基数估计技术，同时以评价者对主观题作业的掌握程度信息以及评价者在该主观题作业中取得的真实分数信息实现对评价者的评分可靠性进行建模，以期提高主观题作业中真实分数估计的有效性和准确性。

3.2 基于认知诊断的同行互评基数估计技术实现流程

本文提出的基于认知诊断的同行互评基数估计技术，其实现框架如图 3-1 所示，整个实现流程包含三个重要步骤：

（1）认知诊断。以所有同行评价者的历史客观题答题矩阵 $\boldsymbol{R}$ 和历史客观题-知识点关联矩阵 $\boldsymbol{Q}$ 为输入，利用认知诊断 DINA 模型得到记录了他们对所有知识点的掌握程度

信息的矩阵 $\boldsymbol{M}$。

（2）基于概率图模型推理隐含变量的后验分布。本文基于认知诊断提出了两个基数估计概率图模型 PG_8 和 PG_9。概率图模型中的各个变量是相互联系的，因而基于模型中观测变量的观测值（包括同行评价者 v 对知识点的掌握程度、互评分数和相对分数）推断模型中隐含变量（包括同行评价者的偏见 b_v、可靠性 τ_v 和被评价者 u_i 的主观题作业的真实分数 s_i）的后验概率分布是一个循环推理的过程，最终推理得到 PG_8 和 PG_9 模型中各个隐含变量的近似后验分布。

（3）真实分数基数估计。以互评分数集合、相对分数集合和步骤一得到的知识点的掌握程度矩阵 $\boldsymbol{M}$ 为输入，以 Gibbs 采样技术为采样框架并利用步骤二得到的各个隐含变量的近似后验分布得到概率模型中每个隐含变量的多个样本值。最后整合概率模型中的每个隐含变量的多个样本值，进而得到主观题作业的真实分数的基数估计值。

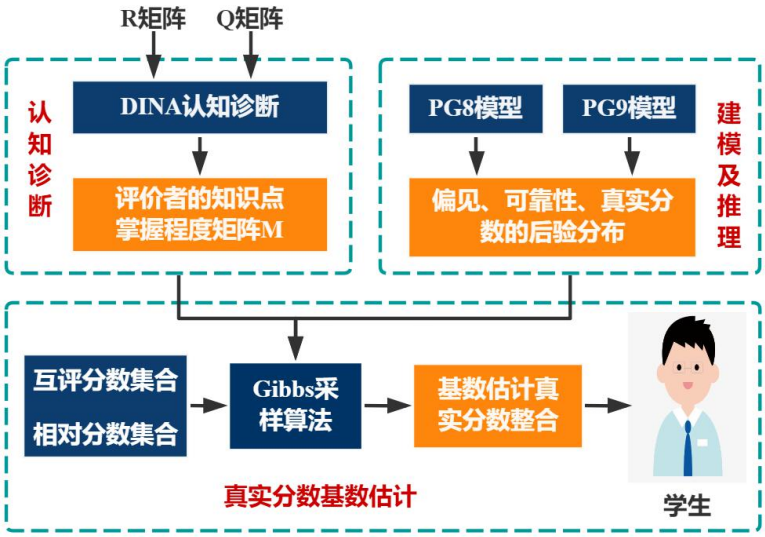


图 3-1 基数估计技术实现框架

Fig. 3-1 The implementation framework of the cardinal estimation technology

3.3 基于认知诊断的同行互评基数估计模型

针对现有的同行互评基数估计技术的不足，本文提出的基于认知诊断的主观题基数估计技术设计了两个估计主观题作业真实分数的概率图模型，即 PG_8 和 PG_9。PG_8 和 PG_9 均假设评价者的评分可靠性与利用认知诊断方法得到的评价者对主观题考察的知识点的掌握程度和本次作业取得的真实分数有关。下面分别介绍这两个概率模型。

3.3.1 PG_8 模型

PG_8 概率模型中各个变量之间的条件依赖关系如图 3-2 展示的概率图模型所示。可见，同行评价者 v 针对被评价者 u_i 的主观题作业给出的互评分数 z_i^v、v 针对被评价者 u_i 和被评价者 u_j 给出的评价分数之间的相对分数 d_{ij}^v、v 的潜在正确作答概率 δ_v 是概率图模型中的观测变量。而 u_i 的主观题作业的真实分数 s_i、v 的偏见 b_v、v 的可靠性 τ_v 则是概率模型估计的隐含变量，且这些隐含变量的先验分布由超参数 μ_0、γ_0、θ_1、θ_2、η_0 和 β_0 所确定。PG_8 模型的生成过程表述如下，模型中给出了不同变量的先验分布信息。

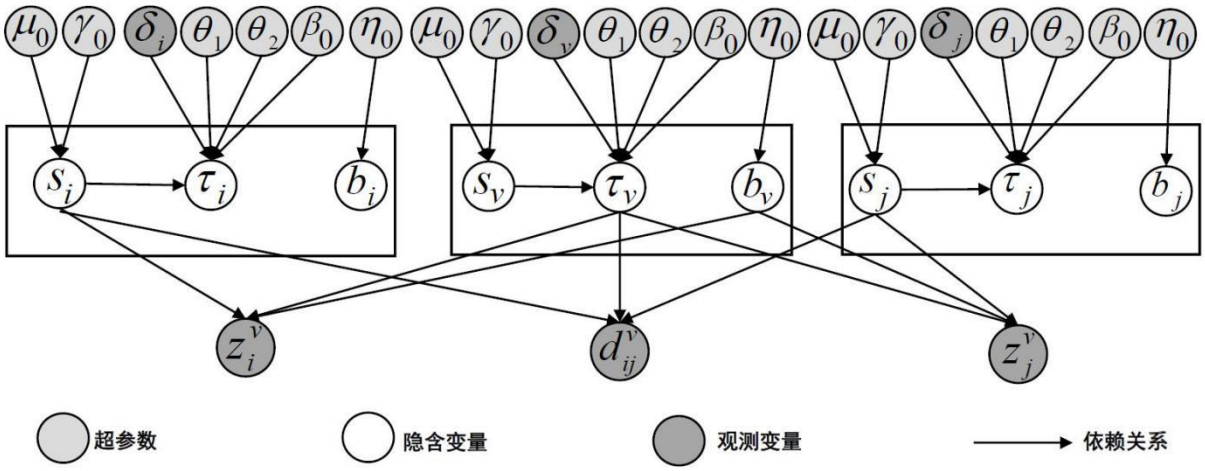


图 3-2 PG_8 和 PG_9 的概率图模型

Fig.3-2 Probabilistic graphical model for PG_8 and PG_9

- 对于第 i 个被评价者 u_i 提交的每份主观题作业
  - → 定义隐含变量 s_i（即 u_i 的真实分数）：

$$s_i \sim N(\mu_0, 1/\gamma_0) \tag{3-1}$$

- 对于每个同行评价者 v
  - → 定义隐含变量 τ_v（即 v 的可靠性）：

$$\tau_v \sim \Gamma(\theta_1\delta_v + \theta_2 s_v, 1/\beta_0) \tag{3-2}$$

  - → 定义隐含变量 b_v（即 v 的偏见）：

$$b_v \sim N(0, 1/\eta_0) \tag{3-3}$$

- 对于每个互评分数 z_i^v

  → 定义可观测变量 z_i^v：

$$z_i^v \sim N(s_i + b_v, 1/\tau_v) \quad (3\text{-}4)$$

- 对于每个相对分数 d_{ij}^v

  → 定义可观测变量 d_{ij}^v：

$$d_{ij}^v \sim N(s_i - s_j, 2/\tau_v) \quad (3\text{-}5)$$

可见，PG_8 模型假设被评价者 u_i 提交的主观题作业的真实分数 s_i 服从均值为μ_0、方差为 $1/\gamma_0$ 的高斯分布。考虑到互评活动中不同的同行评价者的偏见不同（有些给分偏高，有些则给分偏低），因此认为所有评价者的偏见值的均值为 0，即假设同行评价者 v 的偏见 b_v 服从均值为 0 且方差为 $1/\eta_0$ 的高斯分布。由于 PG_8 在对评价者可靠性建模时同时考虑了评价者的对当前作业答题表现（对应$\theta_2 s_v$ 部分）和评价者的历史答题表现（对应$\theta_1 \delta_v$ 部分），PG_8 模型中假设同行评价者 v 的评分可靠性τ_v 服从于形状参数为 v 对该主观题考察的所有知识点的掌握程度乘积（即 v 对该主观题的潜在正确作答概率δ_v）与评价者 v 的真实分数 s_v 的结合即$\theta_1 \delta_v + \theta_2 s_v$、尺度参数为超参数$\beta_0$ 的伽马分布。由伽马分布的特性可知，同行评价者 v 的评分可靠性τ_v 的均值为$(\theta_1 \delta_v + \theta_2 s_v)/\beta_0$。同行评价者 v 针对被评价者 u_i 的主观题作业给出的互评分数 z_i^v 则服从于高斯分布，且该分布的均值等于作业的真实分数 s_i 与 v 的评分偏见 b_v 之和、方差反比于 v 的评分可靠性τ_v。同行互评中可观测的变量除了同行评价者针对主观题作业给出的互评分数之外，还能观测到评价者针对不同被评价者的主观题作业给出的互评分数之间的差值，即相对分数 d_{ij}^v。相对分数的引入有利于提高对被评价者主观题作业真实分数估计的精度。PG_8 模型中，相对分数 d_{ij}^v 被设定为满足均值为两份被 v 评价的主观题作业的真实分数之差（即 s_i-s_j）且方差为 $2/\tau_v$ 的高斯分布。

3.3.2 PG_9 模型

由于 PG_9 模型在对评价者可靠性建模时也同时考虑了评价者对当前作业的答题表现和历史答题表现，因此 PG_9 模型中各个变量间的条件依赖结构与 PG_8 模型一致，详见图 3-2 所示。PG_9 模型的生成过程表述如下，模型中给出了不同变量的先验分布信息。

- 对于第 i 个被评价者 u_i 提交的每份主观题作业

  → 定义隐含变量 s_i（即 u_i 的真实分数）：

$$s_i \sim N(\mu_0, 1/\gamma_0) \tag{3-6}$$

- 对于每个同行评价者 v

  → 定义隐含变量 τ_v（即 v 的可靠性）：

$$\tau_v \sim N(\theta_1\delta_v + \theta_2 s_v, 1/\beta_0) \tag{3-7}$$

  → 定义隐含变量 b_v（即 v 的偏见）：

$$b_v \sim N(0, 1/\eta_0) \tag{3-8}$$

- 对于每个互评分数 z_i^v

  → 定义可观测变量 z_i^v：

$$z_i^v \sim N(s_i + b_v, \lambda/\tau_v) \tag{3-9}$$

- 对于每个相对分数 d_{ij}^v

  → 定义可观测变量 d_{ij}^v：

$$d_{ij}^v \sim N(s_i - s_j, 2\lambda/\tau_v) \tag{3-10}$$

PG_9 模型与 PG_8 模型相比具有两点不同之处。首先，PG_9 模型假设同行评价者 v 的评分可靠性 τ_v 服从高斯分布而 PG_8 模型则假设 τ_v 服从于伽马分布。其次，由于 PG_9 模型将变量 τ_v 的概率分布设定为高斯分布，因此 τ_v 的取值大小由正确作答概率 δ_v 确定。由于正确作答概率变量不可被调节，使得同时依赖于 τ_v 的互评分数变量 z_i^v 的方差以及相对分数变量 d_{ij}^v 的方差均不可被调节。因此，PG_9 模型引入了超参数 λ，并设定 z_i^v 服从于方差为 λ/τ_v 的高斯分布且设定相对分数 d_{ij}^v 服从于方差为 $2\lambda/\tau_v$ 的高斯分布，使得这两个高斯分布的方差均可被调节。

3.4 模型推断

本小节首先阐述了概率图模型的推断思路、概率图模型 PG_8 和 PG_9 的近似后验分布，

其次描述了基数估计模型的推断算法。

3.4.1 近似后验分布

上节给出了两种基于认知诊断的同行互评基数估计的概率模型 PG_8 和 PG_9，模型中给出了各个变量的先验分布信息。下一步则是利用模型中的观测变量（即互评分数 z_i^v、相对分数 d_{ij}^v、同行评价者对主观题作业的知识点掌握程度向量 $\boldsymbol{\alpha}_v$）的观测值来推断模型中各个隐含变量（即同行评价者的评分偏见 b_v、评分可靠性 τ_v、被评价者提交的主观题作业的真实分数 s_i）的后验分布，从而得到每个被评价者提交的主观题作业的真实分数的估计值，该模型推断问题可以形式化表示为 $P(\{b_v|v\in V\}, \{\tau_v|v\in V\}, \{s_i|u_i\in U\}|Z, D, \boldsymbol{M})$。

由于概率模型中各个变量之间是相互联系的，因而基于模型中观测变量的观测值推断模型中隐含变量的后验概率分布是一个循环推理的过程。例如，一方面由概率模型间变量的依赖结构可知，只有准确估计出每名给 u_i 提交的主观题作业进行评分的同行评价者 v 的可靠性 τ_v，才能准确估计出 u_i 提交的主观题作业的真实分数 s_i。另一方面概率模型间变量的依赖结构又表明只有准确估计出某同行评价者 v 所评价的每份作业的真实分数 s_i，才能更准确估计出该同行评价者的可靠性 τ_v。本文采用 Gibbs 采样技术[62]来解决该循环推理问题。具体而言，Gibbs 采样技术：首先基于每个隐含变量的近似后验分布信息运行若干次 Gibbs 采样以生成该变量的若干个样本，得到该变量的样本集；其后，当隐含变量样本的分布逐渐趋于收敛和稳定时，基于隐含变量的样本集推断变量的真实值。例如，假定基于 Gibbs 采样技术所得到的被评价者 u_i 的主观题作业真实分数 s_i 的样本集为 $\left\{s_i^1, s_i^2, ..., s_i^{I_G}\right\}$ 且 I_G 为采样的次数，最终在运行 t 次运行后得到的估计的真实分数 $\hat{s}_i$ 基于样本集中样本的平均值计算，如公式（3-11）所示。考虑到 Gibbs 采样过程存在老化阶段（Burn-in 阶段），这时得到的隐含变量的样本不准确，因而基于 Gibbs 采样技术生成隐含变量的样本集时需要丢弃在老化阶段生成的样本（一般为样本集中的前 n 个样本）。

$$\hat{s}_i = \frac{1}{I_G}\sum_{t=1}^{I_G} s_i^t \tag{3-11}$$

由于概率模型 PG_8 中的隐含变量 s_i 没有闭式解（close-form solution），因而采用近

似离散推断的策略得到该隐含变量的近似后验分布。概率模型 PG_8 中所有隐含变量的近似后验分布的推断结果如下：

$$s \propto \frac{\beta_0^{\theta_2 s_i}\tau_i^{(\theta_2 s_i-1)}}{\Gamma(\theta_1\delta_i+\theta_2 s_i)}\times\exp(R(s_i-\frac{Y}{R})^2)$$

$$\text{其中 } R=\gamma_0+\sum_{v\in V_{u_i}}\tau_v+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v}{2}, \tag{3-12}$$

$$Y=\mu_0\gamma_0+\tau_v(\sum_{v\in V_{u_i}}(z_i^v-b_v)+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{(d_{ij}^v+s_i)}{2})$$

$$\tau\sim\Gamma(\theta_1\delta_v+\theta_2 s_v+\frac{|U_v|^2}{2},\beta_0+\frac{\sum_{u_i\in U_v}(z_i^v-s_i-b_v)^2+\sum_{u_i,u_j\in U_v}\frac{1}{2}(d_{ij}^v-s_i+s_j)^2}{2}) \tag{3-13}$$

$$b\sim N(\frac{\sum_{u_i\in U_v}\tau_v(z_i^v-s_i)}{\eta_0+|U_v|\tau_v},\frac{1}{\eta_0+|U_v|\tau_v}) \tag{3-14}$$

由于概率模型 PG_9 中的隐含变量 s_i 和τ_v 没有闭式解（Close-form solution），因而采用近似离散推断的策略得到该隐含变量的近似后验分布。概率模型 PG_9 中所有隐含变量的近似后验分布的推断结果如下：

$$s \propto \frac{\beta_0^{\theta_2 s_i}\tau_i^{(\theta_2 s_i-1)}}{\Gamma(\theta_1\delta_i+\theta_2 s_i)}\times\exp(R(s_i-\frac{Y}{R})^2)$$

$$\text{其中 } R=\gamma_0+\sum_{v\in V_{ui}}\frac{\tau_v}{\lambda}+\sum_{v\in V_{ui}}\sum_{u_j\in U_v}\frac{\tau_v*(|U_v|-1)}{2\lambda}, \tag{3-15}$$

$$Y=\mu_0\gamma_0+\frac{\tau_v}{\lambda}(\sum_{v\in V_{ui}}(z_i^v-b_v)+\sum_{v\in V_{ui}}\sum_{u_j\in U_v}\frac{(d_{ij}^v+s_i)}{2})$$

$$\tau\propto\tau_v^{\frac{|U_v|^2}{2}}\times\exp\left(-\frac{\beta_0}{2}(\tau_v-Y)^2\right),$$

$$\text{其中 } Y=\theta_1\delta_v+\theta_2 s_v+\sum_{u_i\in U_v}\frac{(z_i^v-s_i-b_v)^2}{\lambda\beta_0}+\sum_{u_i,u_j\in U_v}\frac{(d_{ij}^v-s_i+s_j)^2}{2\lambda\beta_0}. \tag{3-16}$$

$$b\sim N(\frac{\sum_{u_i\in U_v}\frac{\tau_v}{\lambda}(z_i^v-s_i)}{\eta_0+|U_v|\frac{\tau_v}{\lambda}},\frac{1}{\eta_0+|U_v|\frac{\tau_v}{\lambda}}) \tag{3-17}$$

对于上述 PG_8 模型和 PG_9 模型中各个隐含变量后验分布的推理过程详见附录。

3.4.2 基数估计技术推断算法

以 3.3 节提出的同行互评基数估计技术为基础，本节首先给出了基于认知诊断的同

行互评基数估计技术的算法描述；然后详细分析了算法的时空复杂度。

（1）算法描述

算法 3-1 给出了基于认知诊断的主观题互评基数估计技术的算法伪代码。如算法 3-1 所示，以习题-知识点关联矩阵 $\boldsymbol{Q}$ 和同行评价者-历史客观题得分矩阵 $\boldsymbol{R}$ 为输入，首先利用流行的认知诊断 DINA 模型计算得到记录了所有同行评价者对所有知识点的掌握程度信息的矩阵 $\boldsymbol{M}$（行 1）。其次，基于本文提出的同行互评基数估计技术的概率模型（即 PG_8 或 PG_9）分别设定隐含变量 s_i、τ_v 和 b_v 的先验概率分布（行 2）。之后，运行 I_G 轮 Gibbs 采样得到各个隐含变量的样本集（行 3-13）。在每一轮 Gibbs 采样过程中：若同行互评概率模型为 PG_8，则分别基于公式（3-12）、（3-13）、（3-14）采样得到隐含变量 s_i、τ_v 和 b_v 的样本集；若同行互评概率模型为 PG_9，则分别基于公式（3-15）、（3-16）、（3-17）采样得到隐含变量 s_i、τ_v 和 b_v 的样本集（行 4-13）。记每轮采样所得到的包含了各个隐含变量样本集的集合为$\zeta^{(t)}$，则最后删除掉产生于老化阶段（Burn-in 阶段）的$\zeta^{(t)}$（即删除所有$\zeta^{(t)}$且 $t \le \theta$），然后再以各个隐含变量样本集中剩余样本的均值作为该隐含变量真实值的估计值并返回这些估计值（即 $\hat{s}_i, \hat{\tau}_v, \hat{b}_v$）（行 14-15）。

算法 3-1 基于认知诊断的同行互评基数估计算法

输入： 被评价者集合 U；同行评价者集合 V；知识点集合 KP；习题-知识点关联矩阵 $\boldsymbol{Q}$；同行评价者-历史客观题得分矩阵 $\boldsymbol{R}$；互评分数集合 Z；相对分数集合 D；Gibbs 采样次数 I_G；老化阶段阈值θ；同行互评概率模型 PG_x。

输出： 所有被评价者所提交的主观题的真实分数 s_i 和所有同行评价者的评分可靠性τ_v及偏见 b_v。

```
/* 诊断得到所有评价者对各个知识点的掌握程度*/
1.  M = DINA(slip0, guess0, Q, R);
/*基于同行互评概率模型设置隐含变量的先验分布*/
2.  s_i, τ_v, b_v = setDistribution(PG_x);
3.  for t = 1 → I_G do
4.      for each s_i 且 u_i ∈ U do
        /*得到真实分数 s 的一个样本*/
5.          s' = gradeSampling(Z, D, M);
6.          s_{u_i} ← s';
```

7. **for each** τ_v 且 $v \in V$ **do**

/*得到评价者可靠性τ的一个样本*/

8. $\tau' = \text{reliaSampling}(Z, D, \boldsymbol{M})$;

9. $\tau_{v_i} \leftarrow \tau'$;

10. **for each** b_v 且 $v \in V$ **do**

/*得到评价者偏见 ***b*** 的一个样本*/

11. $b' = \text{biasSampling}(Z)$;

12. $b_{v_i} \leftarrow b'$;

13. $\xi^{(t)} \leftarrow (\{s_i \mid u_i \in U\}, \{\tau_v \mid v \in V\}, \{b_v \mid v \in V\})$

/*去除老化阶段的样本后得到各隐含变量的估计值*/

14. $(\{\hat{s}_i \mid u_i \in U\}, \{\hat{\tau}_v \mid v \in V\}, \{\hat{b}_v \mid v \in V\}) \leftarrow \frac{1}{I_G - \theta} \sum_{t=\theta+1}^{I_G} \xi^{(t)}$;

15. **return** $(\{\hat{s}_i \mid u_i \in U\}, \{\hat{\tau}_v \mid v \in V\}, \{\hat{b}_v \mid v \in V\})$;

（2）算法复杂度分析

基数估计技术推断算法主要包含两大功能模块：（1）基于认知诊断 DINA 模型计算所有同行评价者对于所有知识点的掌握程度（算法 3-1：行 1）；（2）运行 I_G 轮 Gibbs 采样（算法 3-1：行 3-13）。由于 DINA 模型是利用 EM 算法进行求解，因此算法的第一个功能模块的时间复杂度是 $O(|V|*2^{|KP|}*I_C)$，其中$|V|$是同行评价者的数量、$|KP|$是知识点的数量、I_C 是 EM 算法的迭代次数。算法的第二个功能模块的时间复杂度为 $O(|U|*|V_{ui}|*I_G+|V|*|U_v|*I_G)$，其中$|U|$是被评价者的数量，$|V_{ui}|$是评判被评价者 u_i 提交的主观题作业的同行评价者数量，$|U_v|$是同行评价者 v 所评价的被评价者的数量，I_G 是 Gibbs 采样次数。考虑到实际教学实践中一般要求提交主观题作业的被评价者都参与该主观题的互评，因而有$|U|=|V|$，且$|V_{ui}|$一般近似等于$|U_v|$，故算法的第二个功能模块的时间复杂度最终为 $O(|U|*|V_{ui}|*I_G)$，进而得到算法的时间复杂度为 $O((|V|*2^{|KP|}*I_C + |U|*|V_{ui}|*I_G)$。

假设被评价者的数量$|U|$远大于知识点的数量$|KP|$和同行评价者所做的历史客观题的数量$|E|$，算法的空间消耗主要来自于存储 I_G 轮 Gibbs 采样得到的 I_G 个隐含变量的样

本集集合（即$\{\zeta^{(t)} \mid 1 \leq t \leq I_G\}$），且该样本集集合的大小为$I_G*(|U|+2|V|)$。考虑到$|U|=|V|$，则推断算法的空间复杂度为$O(I_G*|U|)$。

3.5 实验评价

3.5.1 实验设置

为了验证本文提出的基于认知诊断的同行互评基数估计技术对于主观题评判的有效性，基于自主研发的“会了吗”在线教学服务系统收集计算机专业核心主干课“数据库原理”中“关系数据库规范化理论”这一节的真实教学数据，得到涉及关系数据库规范化理论相关知识点的主观题基数同行互评数据集以及客观题测试结果数据集。基于真实采集的主观题基数同行互评数据集，本节对本文提出的基于认知诊断的同行互评基数估计技术PG_8、PG_9和相关的主观题同行互评基数估计技术进行了实验比较。下面将分别讨论实验设置的三个部分：数据集、参数设置、软硬件设置。

（1）数据集

主观题基数同行互评数据集。在会了吗在线教学服务系统中实现了主观题作业的布置功能和基数同行互评功能。通过给“数据库原理”五个本科平行教学班的 284 名学生布置考察了关系数据库规范化理论的三次主观题作业并组织他们进行同行互评从而得到主观题基数同行互评数据集。每次主观题作业仅包含一道主观题，且布置的三次主观题作业涉及考察关系数据库规范化理论的 11 个知识点，这些知识点和它们的编号分别为：（1）一范式；（2）二范式；（3）三范式；（4）BC 范式；（5）主属性；（6）传递函数依赖；（7）决定因素；（8）函数依赖；（9）码；（10）部分函数依赖；（11）非主属性。这些知识点是数据库原理这门课的教学难点，而主观题形式的作业比客观题形式的作业能更好地帮助学生巩固对这些知识点的学习。图 3-3 给出了记录了三次主观题作业所考察知识点信息的$\boldsymbol{Q}$矩阵。

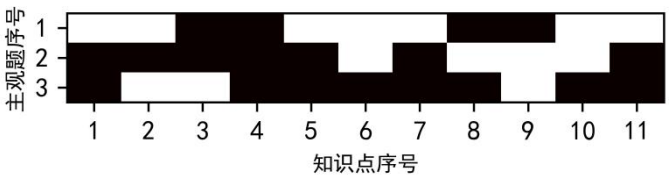


图 3-3 主观题作业的 Q 矩阵

Fig.3-3 The Q matrix of subjective assignments

在主观题作业的互评教学实践中，每名学生既是提交主观题作业的提交者（即被评价者）又是评判同行提交的主观题作业的评价者。具体而言，每次主观题作业的互评教学实践包含以下三个教学活动：

1）评价训练：即每个同行评价者在批改别人的主观题作业之前需要先完成一个评价训练。互评训练以一份含有部分错误的主观题作业为例，要求评价者基于给定的评价规则对该作业进行评分，从而帮助该评价者熟悉作业互评的规则和流程。

2）作业互评：评价者完成评价训练后会收到系统随机给其派发的 3 份主观题作业，并要求其遵循教师制定的评分指导规则完成对这 3 份主观题作业的判分。需要说明的是，为了保证互评的质量，互评活动采用双盲的方式进行，且教师为每道主观题制定了多条细化的评分指导规则并给出每条评分指导规则的分值。

3）成绩：当所有评价者均完成主观题作业的互评打分后，对于每份学生提交的主观题作业，系统将基于多个同行评价者给出的评价分数估计该作业的真实分数（例如以多个评分的均值估计真实分数），并以该估计值作为该主观题作业的成绩。

为了评估不同主观题互评基数估计技术对于主观题作业真实估计的准确性，邀请拥有 6 年以上“数据库原理”课程教学经验的教师对每份学生提交的主观题作业进行评价打分，并以教师的评分作为该主观题作业的基准分数。表 3-1 给出了基于三次主观题作业的基数互评教学实践收集到的同行互评数据集的相关统计信息。

表 3-1 主观题基数同行互评数据集的统计信息

Table 3-1 Summary statistics of subjective assignments for cardinal peer grading datasets

	作业 1	作业 2	作业 3
作业提交数	268	269	267
教师评价数	268	269	267
同行评价者数	254	260	254
同行互评数	694	725	690
作业分值	20	20	20

历史客观题测试结果数据集。本文基于流行的认知诊断 DINA 模型提出了主观题同行互评基数估计技术。为了能够基于 DINA 模型诊断学生对主观题考察的知识点的掌握程度，要求学生们在“会了吗”在线教学平台上完成客观题形式的在线测试。该在线测试包含 40 道客观题，覆盖了三次主观题作业所考察的关系数据库规范化理论的 11 个知识点。基于在线测试活动得到的每名学生的客观题测试结果数据和记录了每道客观题考察的知识点信息的 $\boldsymbol{Q}$ 矩阵（详见图 3-4 所示），即可基于 DINA 模型诊断得到该学生对 11 个知识点的掌握程度值。

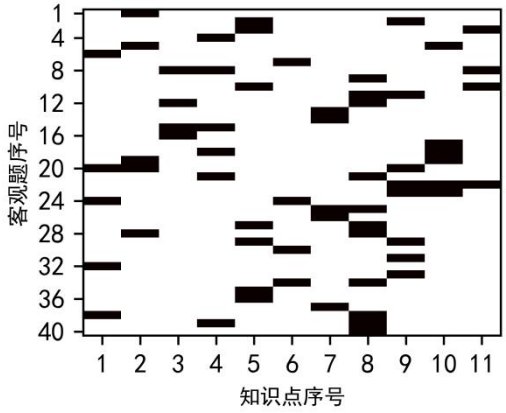


图 3-4 每道客观题考察的知识点信息的 Q 矩阵

Fig.3-4 The Q matrix of objective questions

（2）参数设置

本文提出的主观题同行互评基数估计技术和相关主观题同行互评基数估计技术 PG_6 和 PG_7 均是利用概率模型对同行评价者的互评可靠性和互评偏见进行建模，因而都使用

了一些超参数。为这些超参数设置合理的值对准确估计主观题作业的真实分数非常重要。对于概率模型中的真实分数变量 s_i，由于 PG_8、PG_9、PG_6 和 PG_7 均假设 s_i 满足高斯分布，实验中统一将其所满足的高斯分布的超参数，即均值μ_0和方差 $1/\gamma_0$，分别设置为所有主观题作业互评分数的均值和方差。鉴于 PG_8 和 PG_6 中同行评价者可靠性τ_v所满足的伽马分布中的尺度参数β_0、PG_9 和 PG_7 互评分数 z_i^v 所满足的高斯分布中的参数λ分别是这些技术中最关键的超参数[9]，直接影响它们估计主观题作业真实分数的准确性，因此本文需要调整 PG_8 和 PG_6 中的参数β_0，PG_9 和 PG_7 中的参数λ。根据文献[7][9]的参数设置，本文的具体调整策略为：对于 PG_8 和 PG_6，在其它参数取值固定的前提下，以 50 为步长尝试超参数β_0在[150，400]范围中的不同取值，然后以该技术所得到的对真实分数最准确的估计值为该技术的最终估计值；对于 PG_9 和 PG_7，在其它参数取值固定的前提下，以 0.2 为步长尝试超参数λ在[0.6，1.6]范围中不同取值，然后以该技术所得到的对真实分数最准确的估计值为该技术的最终估计值。对于以上四种基于概率模型的同行互评基数估计技术均涉及的超参数η_0，遵循文献[7]提出的微调策略本文设置在范围[0.04, 0.2]中调整η_0的取值。对于 PG_7 和 PG_9 涉及的超参数β_0，依据文献[9]将其设置为 0.1。由于基于概率模型的同行互评基数估计技术在估计主观题作业真实分数时具有一定的随机性，因此对于超参数集合的每种设定，每种技术都执行 10 次真实分数的推断算法。对于基于概率模型的同行互评基数估计技术中每个需要估计的隐含变量，推断算法均运行 600 次 Gibbs 采样获取隐含变量的样本值，并设定前 60 次采样得到的样本为老化阶段的样本，这些老化阶段的样本将不参与对真实分数的估计运算。

（3）软硬件设置

所有参与比较的主观题同行互评基数估计技术均基于 Python（v3.7）语言实现，并在配备了 i5-8500 3GHZ CPU、8GB 内存、1TB 硬盘，运行了 64 位 Windows 10 操作系统的服务器上进行统一实验测试。

3.5.2 评价指标

采用不同技术给出的对主观题真实分数的估计值和主观题作业基准分数之间的均方根误差 RMSE（Root Mean Square Error）作为不同主观题同行互评基数估计技术有效

性的评估指标。RMSE 被广泛应用于评估同行互评基数估计技术有效性[6][8]，其计算公式如下：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i-\hat{s}_i)^2} \tag{3-18}$$

其中，s_i 表示教师针对学生 $u_i \in U$ 提交的主观题作业给出的分数（作业的基准分数），$\hat{s}_i$ 则表示同行互评基数估计技术对该主观题作业真实分数的估计值，n 为提交主观题作业的学生数量且 $n=|U|$。

3.5.3 实验结果与分析

（1）估计真实分数的准确性

表 3-2 展示了不同主观题同行互评基数估计技术估计主观题作业真实分数的准确性。其中，表格中的 RMSE 和 STD 分别是指每种主观题同行互评基数估计技术执行 10 次推断算法所得到的真实分数估计值相对于基准分数得到的 10 个 RMSE 值的均值以及这 10 个 RMSE 值的标准差。由表 3-2 可知，中位数技术和均值技术对主观题作业真实分数的估计误差最大，这是因为它们在估计真实分数时均未考虑同行评价者的互评可靠性和互评偏见这两个重要因素。由于同时考虑了同行评价者在本次作业中的答题表现以及评价者的历史答题表现对其评分可靠性的影响，本文提出的基于认知诊断的主观题同行互评基数估计技术 PG_8 和 PG_9 对三次主观题作业真实分数的估计误差 RMSE 均明显低于其它技术。特别地，PG_9 技术对三次作业真实分数的平均估计误差比中位数技术（即当今慕课平台上采用的主流互评技术）的平均估计误差降低了 69.6%。同时还可观察到，在大多数情况下（即作业 1 和作业 2），PG_9 比 PG_8 的 RMSE 值小，这说明假设同行评价者评分可靠性的先验分布为高斯分布能够更好拟合本实验中大多数主观题作业的同行互评数据集。

由于 PG_8 与 PG_6 相对应，均假设同行评价者互评可靠性取值的先验分布为伽马分布，且 PG_9 与 PG_7 相对应，均假设同行评价者互评可靠性取值的先验分布为高斯分布，下面对这两组技术分别进行比较分析。

PG_8 vs. PG_6：由表 3-2 可知，对于三次主观题作业，这两种基数估计技术估计作业真实分数的误差 RMSE 的标准差 STD 取值相近，且 STD 的取值都不大。这说明它们对

不同主观题作业真实分数估计误差较为稳定。同时还可观察到，PG_8 对三份主观题作业真实分数的估计误差 RMSE 均明显低于 PG_6，且 PG_8 对三次作业真实分数的平均估计误差比 PG_6 降低了 42%。

PG_9 vs. PG_7：由表 3-2 可知，PG_7 和 PG_9 这两种基数估计技术对于所有主观题作业的 RMSE 的标准差 STD 取值相同，且 STD 取值最大仅为 0.01。且 PG_9 对三份主观题作业真实分数的估计误差 RMSE 均明显低于 PG_7。具体而言，PG9 对三次作业真实分数的平均估计误差比 PG_7 降低了 43%。

通过以上对比分析可知，PG_8 技术和 PG_9 技术比其对应的 PG_6 技术和 PG_7 技术对主观题作业真实分数的估计更为准确，实验结果证实了结合本次作业中的答题表现以及评价者的历史答题表现建模可靠性对于同行互评基数估计的有效性。

表 3-2 估计真实分数的误差

Table 3-2 The Error of true score estimation

	作业 1		**作业 2**		**作业 3**	
	RMSE	STD	RMSE	STD	RMSE	STD
均值	4.61	0.00	4.16	0.00	4.53	0.00
中位数	5.09	0.00	4.59	0.00	5.04	0.00
PG_6[9]	3.32	0.01	2.67	0.01	3.32	0.02
PG_8	2.31	0.01	1.69	0.01	**1.30**	**0.01**
PG_7[9]	2.46	0.01	2.56	0.00	2.82	0.01
PG_9	**1.57**	**0.01**	**1.28**	**0.00**	1.63	0.01

（2）估计真实分数的最大误差

表 3-3 展示了不同主观题同行互评基数估计技术给出的主观题作业真实分数的估计值与基准分数之间最大评分偏差。由表 3-3 可知，均值技术与中位数技术的最大评分偏差均比基于概率模型的同行互评技术更大。这是因为均值技术与中位数技术仅依赖同行评价者给出的互评分数信息来直接估计作业的真实分数，这无疑使得它们对真实分数的估计准确性依赖于同行评价者的评分质量：若一组同行评价者的评分质量均不高，这两种技术将给出与基准分数偏差很大的估计值。相比而言，基于概率模型的同行互评基数

估计技术在估计主观题作业的真实分数时不但考虑了同行评价者给出的互评分数信息，还考虑了同行评价者的评分可靠性及评分偏见信息，因而能更准确地估计作业的真实分数。同时还可观察到，基于认知诊断的同行互评基数估计技术（PG_8 和 PG_9）比另外两个基于概率模型的同行互评基数估计技术（PG_6 和 PG_7）的最大评分偏差要小，表明了同时考虑影响可靠性的两方面因素（即同行评价者在本次作业中的答题表现以及评价者的历史答题表现）能更有效地保障对每个学生的主观题作业真实分数的估计准确性。

表 3-3 真实分数估计值与基准分数间的最大评分偏差

Table 3-3 Maximum deviation between an estimated grade and ground truth for all students

	作业 1	**作业 2**	**作业 3**
均值	18	8	16
中位数	18	8	16
PG_6[9]	10.87	6.31	10.46
PG_7[9]	10.74	6.63	10.81
PG_8	6.03	5.38	**4.26**
PG_9	**5.54**	**4.12**	4.94

（3）超参数的敏感性

该部分分别分析了超参数β_0的取值对 PG_8 技术的影响以及超参数λ的取值对 PG_9 技术的影响。基于超参数β_0（或λ）的不同值训练 PG_8（或 PG_9）中的概率模型并观察不同参数值下该技术给出的主观题作业真实分数的估计值相对于基准分数的估计误差 RMSE。需要注意的是，超参数β_0（或λ）取值变化时，其它参数的取值不能改变。将 PG_8 中的超参数β_0 设置在[150, 400]范围内以 50 为步长变化，得到图 3-5；将 PG_9 中的超参数λ设置在[0.6, 1.6]范围内以 0.2 为步长变化，得到图 3-6。由图 3-5 和图 3-6 可知：在合理的取值范围内，这两种同行互评基数估计技术对超参数值具有鲁棒性，它们对主观题作业真实分数的估计误差都控制在可接受的范围。

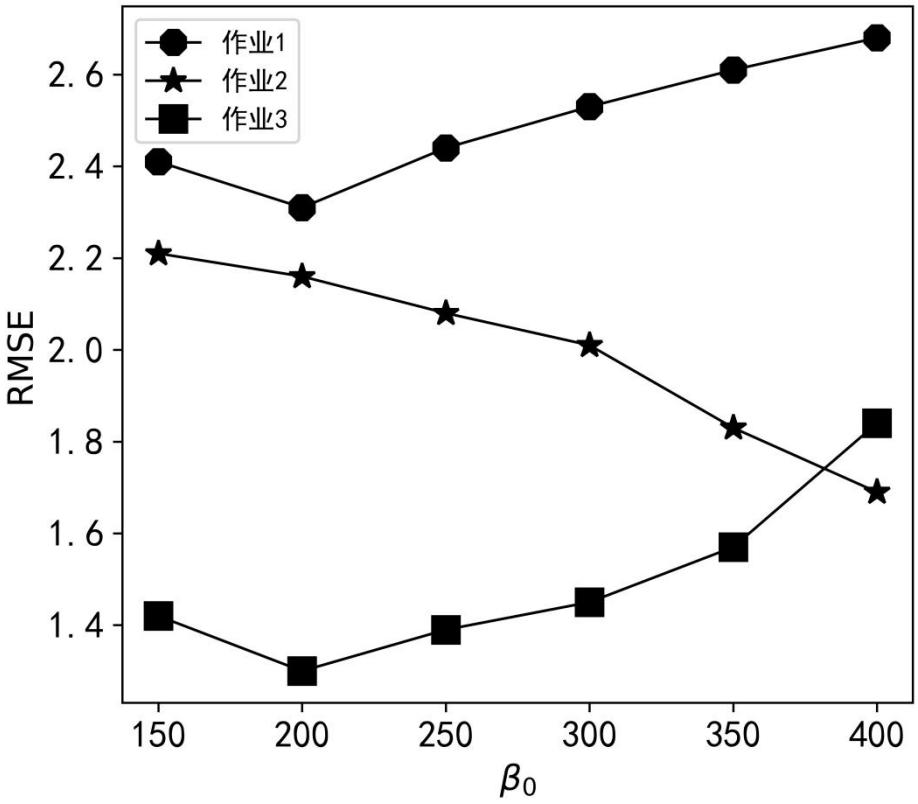


图 3-5 PG_8 技术的超参数敏感性分析

Fig.3-5 Sensitivity analysis of hyper-parameter for PG_8

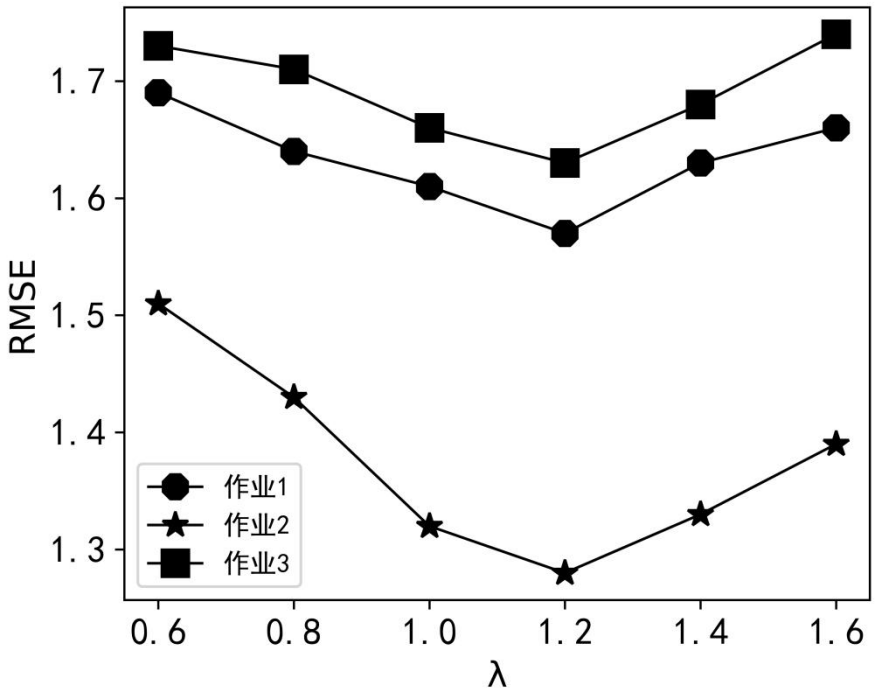


图 3-6 PG_9 技术的超参数敏感性分析

Fig.3-6 Sensitivity analysis of hyper-parameter for PG_9

（4）执行时间

均值技术和中位数技术只是简单用互评分数的均值或中位数来估计主观题作业的真实分数，因此它们的执行时间非常短，不在此处进行横向比较。图 3-7 展示了相同参数设置下，PG_6、PG_7、PG_8 和 PG_9 基数估计技术的真实分数推断算法运行 10 次的平均运行时间。如图 3-7 所示，对于不同主观题作业，各个基于概率模型的主观题同行互评基数估计技术的推断算法运行时间均大于 1 分钟，这是因为它们均需要运行 600 次 Gibbs 采样来估计主观题作业真实分数的取值。其中，各个基数估计技术在推断作业 2 的真实分数时均耗时最长，这是因为作业 2 收集到了最多的同行互评反馈（详见表 3-1），而同行互评数越大则推断算法的运行时间会越长。同时还可观察到对于不同作业，PG_9 技术的推断算法耗费均最长，这是因为 PG_9 的概率模型基于隐含变量的先验分布信息只能得到隐含变量偏见的近似后验分布的闭式解（Closed-form），因而推断时间较长。而其它三种技术在推断过程中均只有一个隐含变量的近似后验分布不存在闭式解的情况，因而缩短了推断时间。

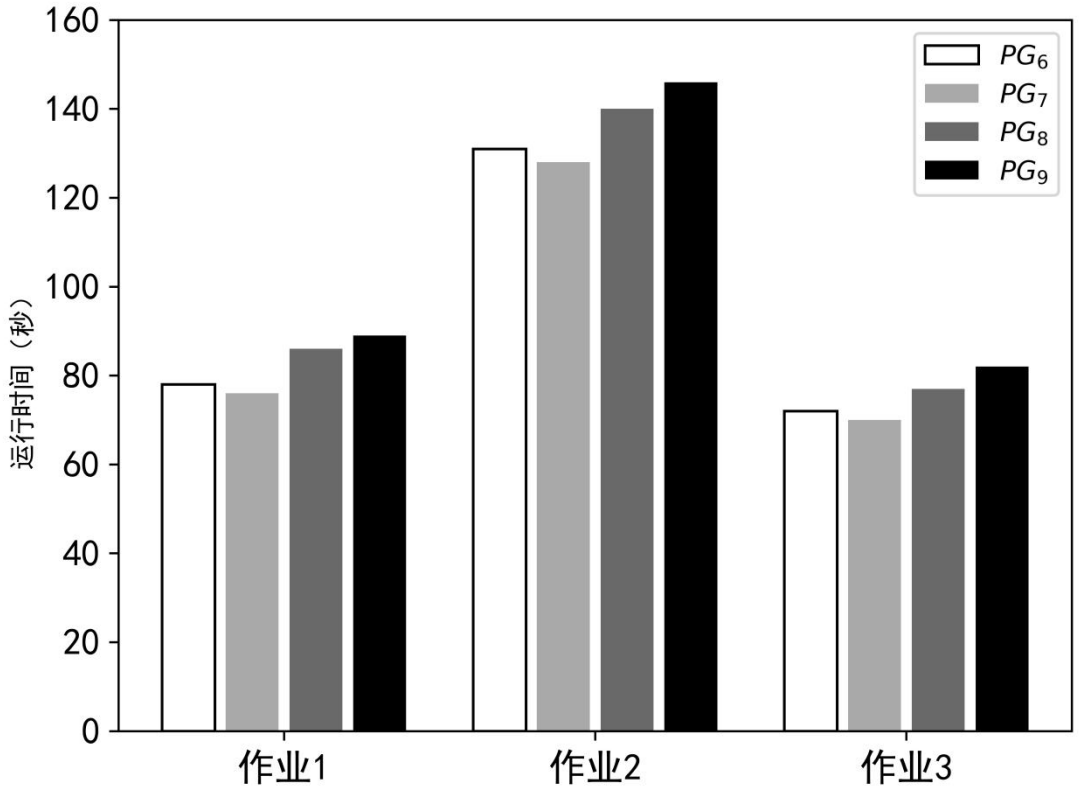


图 3-7 运行时间分析

Fig.3-7 Analysis of running time

3.6 本章小结

针对主观题作业的同行互评技术能够大大减轻教师批改大规模主观题作业的负担，具有重要的研究意义。然而，在基数的同行互评场景中，现有的主观题同行互评基数估计技术均没有考虑评价者对主观题知识点的掌握程度对于作业评分误差的影响，从而不能够准确的建模同行评价者的可靠性，进而影响了评价者最终的真实分数的准确估计。针对现有研究工作的不足，提出了一种基于认知诊断的同行互评基数估计技术，其中基于认知诊断得到的评价者在本次作业中的答题表现（对应于本次作业取得的真实分数）以及评价者的历史答题表现（对应于基于历史答题记录诊断得到的该评价者对本次作业题的掌握程度）对评价者的可靠性进行建模，再结合评价者的偏见，提出了两个概率图模型 PG_8 和 PG_9。通过组织学生参与主观题作业和同行互评活动收集学生对主观题作业的同行互评数据，并基于收集的数据集对基于认知诊断的同行互评基数估计技术进行实验分析，实验结果表明提出的 PG_8 和 PG_9 技术能够有效提升主观题基数同行互评中真实分数估计的准确性。

第四章 基于认知诊断的同行互评序数估计技术 BT+CD

本文第三章提出了一种基于认知诊断的同行互评基数估计技术，包括两个概率图模型 PG_8 和 PG_9，PG_8 和 PG_9 基于认知诊断得到同行评价者对主观题作业知识点的掌握程度，并基于评价者的偏见构建了有效的同行互评的真实分数估计概率模型，提高了基数估计对真实分数估计的准确率。然而，评价者在同行互评序数估计中比基数估计更容易做出判断，序数估计是另外一种重要的同行互评评估方式。因此，基于认知诊断得到的评价者对知识点的掌握程度能够减少估计误差的研究对序数估计技术进行优化，以提升序数估计的有效性和准确性，故本章将集中探讨基于认知诊断的同行互评序数估计技术。

4.1 问题的分析与提出

比较于基数估计方法要求同行评价者对主观题作业给出绝对的分数反馈来说，序数估计方法只要求同行评价者给出相对的排序反馈，对非专家的同行评价者而言更容易做出相对判断，该方法减少了同行评价者的偏见，广泛的证据表明同行互评中序数反馈比基数反馈更容易且更可靠[19,35]。序数估计技术的研究难点在于如何利用多个同行评价者给出的绝对排序得到每个被评价者的作业相对排名。随着大规模在线教育的发展，基于同行互评的序数估计技术成为了研究热点并取得了不少研究成果。现有的相关工作针对同行互评中的序数估计问题采用了一些经典的排名聚合方法，然而这些工作仅仅是对排序做简单的聚合，没有考虑同行评价者在主观题作业互评中的可靠性问题。虽然少数研究工作在主观题同行互评的序数估计中引入了同行评价者的可靠性以能够更准确的估计相对排名，然而这些研究工作对可靠性的建模不够准确，没有考虑到评价者对主观题作业知识点的掌握程度对可靠性的影响，因而对评价者的可靠性估计不够准确，进而影响了作业排名的序数估计。因此，本章基于认知诊断 DINA 构建同行互评序数估计概率模型（命名为 BT+CD 技术）。BT+CD 技术根据 DINA 模型得到同行评价者对知识点的掌握程度，并利用掌握程度信息对评价者的可靠性建模，从而实现在同行互评序数场景

中的主观题作业排名估计和可靠性估计。使用真实的同行互评活动中收集到的评价者对主观题作业互评的相对评价数据和用于计算掌握程度的客观题数据集对本文所提出的BT+CD 技术进行评估，通过对比 BT+CD 技术对被评价者作业的估计排名和教师给出的作业排名，结果表明该技术能够有效提高主观题同行互评序数估计的准确性。

4.2 BT+CD 技术的设计思想

基于认知诊断的序数估计 BT+CD 技术的流程图如 4-1 所示，实现过程共包括三个步骤：

（1）计算评价者的知识点掌握程度。以所有评价者的客观题答题矩阵 $\boldsymbol{R}$ 和客观题-知识点关联矩阵 $\boldsymbol{Q}$ 为输入，利用 DINA 模型诊断得到他们对所有知识点的掌握程度信息的矩阵 $\boldsymbol{M}$。

（2）主观题作业同行互评降序排名。首先根据序数同行互评活动中每个评价者对主观题作业做出的相对评价，统计每个评价者的评判作业的降序排名（可能包括相同排名的两份作业）。其次筛选有效的降序排名作业，即相同排名的作业数据不参与算法的下一个步骤。

（3）基于认知诊断的序数估计。以步骤一得到的所有评价者的知识点掌握程度矩阵 $\boldsymbol{M}$ 对评价者的可靠性进行建模，再以步骤二得到的有效作业互评降序排名为输入，计算序数估计 BT+CD 模型的损失函数。最后基于随机梯度下降法优化估计评价者的可靠性和作业的真实分数，将作业分数转化为最终排名。

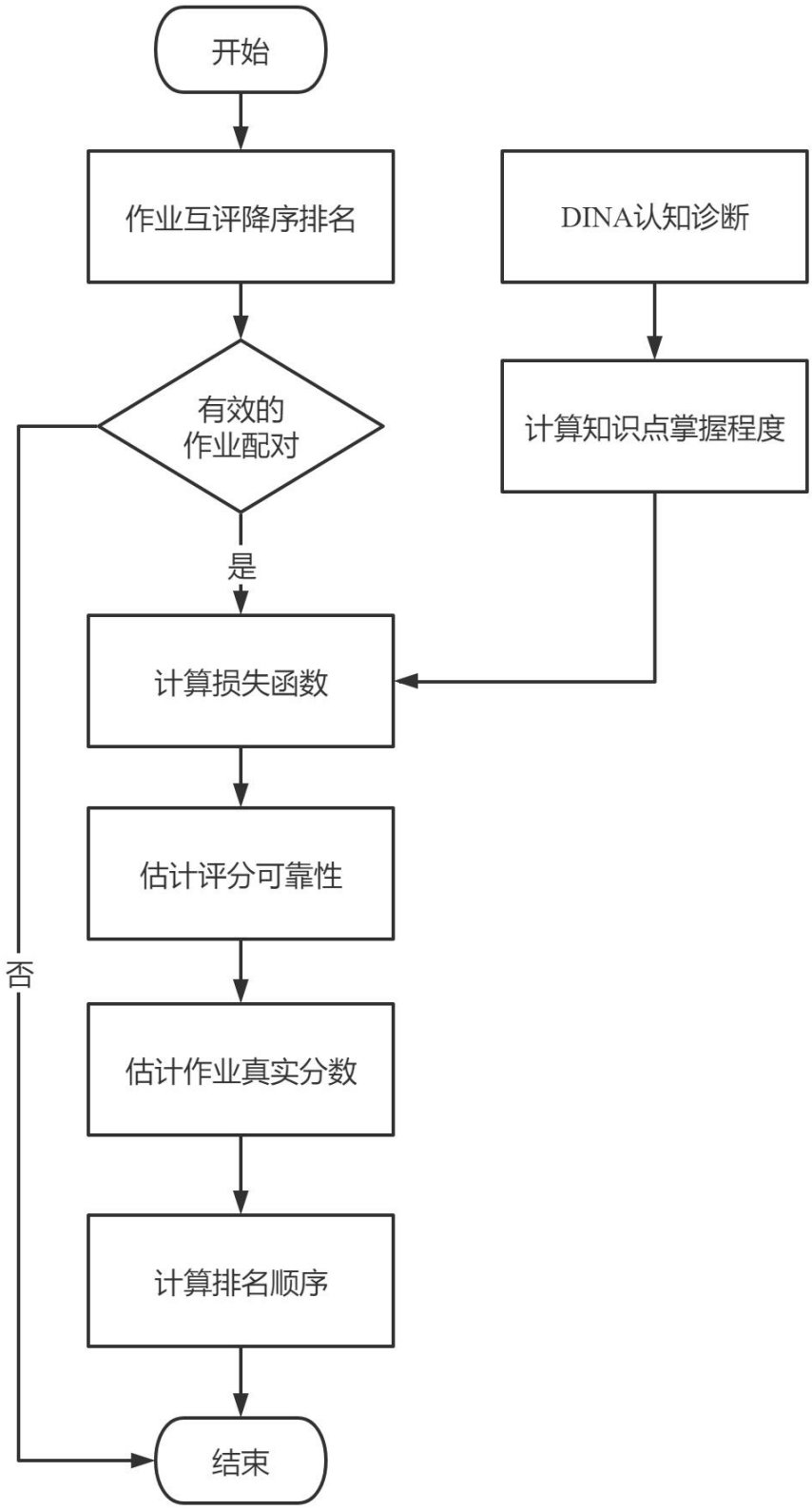


图 4-1 BT+CD 技术的流程图

Fig. 4-1 The flow chart of the proposed BT+CD technology

4.3 BT+CD 技术的具体实现

由于前文已经就认知诊断DINA模型得到同行评价者对知识点的掌握程度进行过介绍，故本节不再赘述。因此，本节将依次对基于认知诊断的同行互评序数估计模型、BT+CD 技术的推断算法进行详细阐述。

4.3.1 基于认知诊断的同行互评序数估计模型

n 份主观题作业的排序也可以看作配对排序比较（任意取 *n* 份作业中的两份作业）的集合，本文提出的BT+CD是一种基于配对比较概率分布的作业排名聚合模型。BT+CD 和 2.2.2 节描述的相关序数估计 BT 技术、RBTL 技术、BT+G 技术不同，该技术结合了评分可靠性τ_v提出基于认知诊断的同行互评序数估计模型，给定作业降序排名的概率分布关系表示如下：

$$hypothesis = P(u_i \succ_{\rho(v)} u_j) = \frac{1}{1+\exp(-\tau_v(s_i - s_j))} \tag{4-1}$$

其中，s_i和 s_j分别是被评价者 u_i和 u_j提交主观题作业的真实分数，$u_i \succ_{p(v)} u_j$表示评价者 v 对于被评价者 u_i的主观题作业的排名在 u_j的作业排名之前，$p(v)$为同行评价者 v 的配对评价排序集合，$u_i \succ_{p(v)} u_j$的概率关系定义为两个被评价者 u_i和 u_j的真实分数差值 s_i-s_j和同行评价者的可靠性τ_v的逻辑函数。

真实分数估计的最大后验估计如下：

$$\hat{s} = \arg\max_{s,\tau}\{prior(s,\tau_v)\prod_{v\in V}\prod_{u_i \succ_{\rho(v)} u_j}\frac{1}{1+e^{-\hat{\tau}_v(s_i - s_j)}}\} \tag{4-2}$$

其中，$prior(s, \tau_v)$为真实分数 s 和评价可靠性τ_v的先验。本文假设同行评价者 v 的可靠性τ_v的先验分布为满足于均值为 v 对主观题作业所有知识点的掌握程度$\prod_{k=1}^{K}\alpha_{vk}^{q_k}$、方差为 $1/\varphi_0$的高斯分布，其中φ_0为超参数，形式化表达如下：

$$\tau_v \sim N(\prod_{k=1}^{K}\alpha_{vk}^{q_k}, 1/\varphi_0) \tag{4-3}$$

假设所有被评价者提交主观题作业真实分数 s 的先验分布为服从于均值为μ和方差为 σ^2的高斯分布，BT+CD 模型的损失函数（cost function）表示如下：

$$L=\frac{\lambda}{2\sigma^2}\sum_{u_i\in U}(s_i-\mu)^2-\sum_{v\in V}\sum_{u_i\succ_{p(v)}u_j}\log(hypothesis) \quad (4\text{-}4)$$

其中，公式中的第一项式是避免过拟合（overfitting）的正则化项，第二项式是对数似然函数（data likelihood）项。该损失函数对于所有要估计的隐含变量而言都是联合凸函数，因而 BT+CD 技术采用随机梯度下降（stochastic gradient descent）解决估计隐含变量（被评价者 u_i 的真实分数 s_i 和同行评价者的可靠性τ_v）的优化问题。由于数据的稀疏性问题，数据的最大似然估计可能会产生非连续的估计，而基于真实分数先验分布定义的正则化项对于解决该问题起到了关键的作用。

4.3.2 BT+CD 技术的实现算法

以 4.3.1 节提出的同行互评序数估计 BT+CD 技术为基础，本节首先对基于认知诊断的同行互评序数估计模型的算法进行描述；然后对算法的时空复杂度进行了分析。

（1） 算法描述

算法 4-1 给出了序数估计 BT+CD 推断算法伪代码。如算法 4-1 所示，以习题-知识点关联矩阵 $\boldsymbol{Q}$ 和同行评价者-历史客观题得分矩阵 $\boldsymbol{R}$ 为输入，首先利用 DINA 模型诊断得到记录了所有同行评价者对作业知识点的掌握程度信息矩阵 $\boldsymbol{M}$（行 1）。其次，分别设定估计的两个隐含变量真实分数和可靠性的先验分布（行 2）。根据被评价者集合 U、同行评价者集合 V、互评活动的排名评价集合σ' 为输入信息，计算得到所有评价者的互评评价降序排名比较数 np（行 3）。之后，训练 $maxepoch$ 次被评价者的真实分数和评价者的评分可靠性的序数估计（行 4-10），在每一次的训练中，随机打乱训练样本数据（行 5），并且将评价降序排名比较对顺序随机打乱且重新设置学习率 eta，eta 随着迭代训练次数的增加而下降（行 6）。每一次的训练执行 np 次的参数估计，基于公式（4-2）的排名概率关系 $hypothesis$ 得到的损失函数 L，基于随机梯度下降法迭代估计评价者的可靠性τ_v和真实分数 s（行 7-10）。最后，将估计的主观题作业真实分数转化为排名顺序$\hat{\sigma}$（行 11）并返回这些估计值（$\hat{s}$，$\hat{\sigma}$，$\hat{\tau}_v$）（行 12）。

算法 4-1 BT+CD 推断算法

输入： 被评价者集合 U；同行评价者集合 V；知识点集合 KP；习题-知识点关联矩阵 $\boldsymbol{Q}$；同行评价者-历史客观题得分矩阵 $\boldsymbol{R}$；所有评价者的排名评价集合σ'；真实分数 s 和评价可靠性τ_v的先验 $prior$；排名顺序的概率分布 $hypothesis$；最大训练次数 $maxepoch$。

输出： 所有被评价者所提交的主观题作业的真实分数 $\hat{s}$ 和排名 $\hat{\sigma}$；所有评价者的评分可靠性 $\hat{\tau}_v$。

1. $\boldsymbol{M}$ = DINA($slip_0$, $guess_0$, $\boldsymbol{Q}$, $\boldsymbol{R}$); //诊断得到所有同行评价者对各个知识点的掌握程度
2. s, τ_v=$prior$(s, τ_v, M); //真实分数 s 和评价可靠性τ_v 的先验
3. np=$numOfPair$(U, V, σ'); //所有评价者的互评评价降序排名比较数
4. **for** $epoch = 1 \rightarrow maxepoch$ **do**
5. P=$randperm$(σ'); //随机打乱的训练样本数据
6. eta= 1/sqrt($epoch$); //学习率设置
7. **for** $i = 1 \rightarrow np$ **do**
8. $L \leftarrow$ $getCostfunction$($hypothesis$, s, τ_v, P); //损失函数
9. $\hat{\tau}_v \leftarrow SGD_G(L, s, \tau_v, eta, P)$; //基于随机梯度下降迭代更新可靠性的值
10. $\hat{s} \leftarrow SGD_S(L, s, \tau_v, eta, P)$; //基于随机梯度下降迭代更新真实分数的值
11. $\hat{\sigma} \leftarrow getOrder(\{\hat{s}_i \mid u_i \in U\})$; //将分数转化为排名顺序
12. **return** $\hat{s}$, $\hat{\sigma}$, $\hat{\tau}_v$;

（2） 复杂度分析

BT+CD 技术推断算法主要包含两大功能模块：（1）基于 DINA 模型诊断所有同行评价者对于所有作业知识点的掌握程度（算法 4-1：行 1）；（2）训练 *maxepoch* 次被评价者的真实分数和评价者的评分可靠性的序数估计（算法 4-1：行 4-9）。由于 DINA 模型是利用 EM 算法求解掌握程度，所以算法的第一个功能模块的时间复杂度是 $O(|V|*2^{|KP|}*I_C)$，其中$|V|$是同行评价者的数量、$|KP|$是知识点的数量、I_C 是 EM 算法的迭代次数。算法的第二个功能模块的时间复杂度为 $O(maxepoch*np)$，其中 *maxepoch* 是算法最大训练次数，*np* 是所有评价者 *v* 的互评评价降序排名比较数。综合两个功能模块得到算法的时间复杂度为 $O(|V|*2^{|KP|}*I_C + maxepoch*np)$。

假设序数同行互评中被评价者的数量$|U|$远大于知识点的数量$|KP|$和评价者所做的历史客观题的数量$|E|$，那么 BT+CD 推断算法的空间消耗主要来自于存储每次训练中序数估计得到的被评价者的真实分数和排名顺序、评价者的可靠性估计值集合，且该估计值集合的大小为$(2|U|+|V|)$。考虑到实际同行互评活动中一般要求提交主观题作业的被评价者都参与该主观题的互评，因而有$|U|=|V|$，进而得到算法的空间复杂度为 $O(|U|)$。

4.4 实验评价与分析

4.4.1 实验设置

为了验证本文提出的基于认知诊断的同行互评序数估计技术对于主观题评判的有效性，基于自主研发的“会了吗”在线教学服务系统收集计算机专业核心主干课“数据库原理”中“关系数据库规范化理论”这一节的真实教学数据，得到涉及关系数据库规范化理论相关知识点的主观题序数同行互评数据集以及客观题测试结果数据集。基于真实采集的主观题序数同行互评数据集，本节将本文提出 BT+CD 技术与当前 MOOCs 平台常用的均值和中位数两种估计技术进行实验对比，还与时下较为相关的 BT、RBTL 和 BT+G 三种序数估计技术进行实验对比。其中 BT 技术是利用两个配对比较的被评价者的排名顺序和他们作业真实分数之间的概率分布关系来进行序数估计；RBTL 技术是在 BT 技术的基础上引入了评价者的评价能力，且评价能力是由同行评价者本身具有的内在能力（表现评价者作业的真实分数）的线性表达式构成；而 BT+G 技术则是引入了评价者的评分可靠性并设为可变参数，对评价者的可靠性和被评价者的真实分数采用最大似然估计法进行计算。下面将分别讨论同行互评序数估计实验设置的两个部分：

（1）数据集

主观题序数同行互评数据集。在会了吗在线教学服务系统中实现了主观题作业的布置功能和序数同行互评功能。通过给“数据库原理”五个本科平行教学班的 284 名学生布置考察了关系数据库规范化理论的三次主观题作业并组织他们进行序数同行互评。其中，三次主观题作业所考察知识点和 3.5.1 节描述的相同，主观题数据的 $\boldsymbol{Q}$ 矩阵详见图 3-3。

主观题作业的序数互评教学实践和 3.5.1 节中描述的相似，每名学生既是提交主观题作业的提交者（即被评价者）又是评判同行提交的主观题作业的评价者，每次主观题作业的互评教学实践均包含评价训练、作业互评、成绩三个教学活动，但是与基数同行互评不同的是：在作业互评的设置中，序数同行互评要求同行评价者在评分规则指导下对评判的主观题作业给出相对的评判顺序。

为了评估不同主观题互评序数估计技术对于主观题作业真实排名估计的准确性，邀请拥有 6 年以上“数据库原理”课程教学经验的教师对所有提交的主观题作业进行评价打

分，并以教师的评分作为该主观题作业的基准分数，在序数估计实验中将教师所给的评价分数转化为相对的基准排名顺序。表 4-1 给出了基于三次主观题作业的序数互评教学实践收集到的同行互评数据集的相关统计信息。

表 4-1 主观题序数同行互评数据集的统计信息

Table 4-1 Summary statistics of subjective assignments for ordinal peer grading datasets

	作业 1	**作业 2**	**作业 3**
同行评价者数	194	248	183
同行互评数	421	649	389
互评配对比较数	15575	25277	10729

历史客观题测试结果数据集。本文基于认知诊断提出了主观题同行互评序数估计技术。为了测试本文提出技术的有效性，本文基于第三章真实课堂实验收集的覆盖相同知识点的历史客观题测试数据集来分析，且基于认知诊断 DINA 模型得到学生对主观题考察知识点的掌握程度。基于在线测试活动得到的每名学生的客观题测试结果数据集对应的 $\boldsymbol{Q}$ 矩阵详见图 3-4。

（2）软硬件设置

所有参与比较的主观题同行互评序数估计技术均基于 Matlab R2016a 实现，并在配备了 i5-8500 3GHZ CPU、8GB 内存、1TB 硬盘，运行了 64 位 Windows 10 操作系统的服务器上进行统一的实验测试。

4.4.2 评价指标

当前，主要用于评估同行互评序数估计技术精确性的指标是肯德尔等级相关系数和均方根误差 RMSE。为了更好的评估序数估计模型的性能，本文用这两个评价指标对于方法的精确性和合理性进行全面考察。RMSE 的具体描述见 3.5.2 节，下面对肯德尔等级相关系数评价指标进行阐述。

肯德尔等级相关系数 KTRCC（Kendall Tau Rank Correlation Coefficient）：可以用于评估不同技术给出估计顺序相对于基准顺序的好坏。具体而言，给定两个排序后的元素序列 X 和 Y 且它们的元素数量为$|X|=|Y|=n$，则 X 和 Y 两个序列之间的肯德尔等级相关

系数的计算公式如下：

$$KTRCC = 2 \times \frac{C_{consist}(X,Y) - C_{inconsist}(X,Y)}{n(n-1)} \quad (4\text{-}5)$$

其中，$C_{consist}(X, Y)$表示在 X 和 Y 两个序列中相对顺序保持一致的元素对的数量，$C_{inconsist}(X, Y)$表示在 X 和 Y 两个序列中相对顺序不一致的元素对的数量，分母则表示两个序列全部元素对的数量。由公式（4-5）可看出，肯德尔等级相关系数数值越大表示序列 X 和 Y 在元素排序上的一致性越高。

在序数同行互评设置下，将肯德尔等级相关系数用于评估不同序数估计技术给出的被评价者作业的排名顺序相对于基准作业顺序的好坏，其中将教师对主观题作业的评价顺序作为作业的基准顺序。将公式（4-5）应用在序数同行互评估计问题中，针对序数同行互评的肯德尔等级相关系数的计算公式如下：

$$KTRCC = 2 \times \frac{\left|\left\{(s_i,s_j) \mid s_i,s_j \in S; i \neq j; \hat{s}_i \succ \hat{s}_j; s_i \succ s_j \right\}\right|}{n(n-1)} \quad (4\text{-}6)$$

其中，s_i 和 s_j 分别表示被评价者 u_i 和 u_j 提交的主观题作业的基准分数，$\hat{s}_i$ 和 $\hat{s}_j$ 分别表示被评价者 u_i 和 u_j 提交的主观题作业的估计分数，n 为被评价者的数量。可见，分母为在序数同行互评中作业的基准顺序和估计顺序一致的全部配对的数量，KTRCC 数值越大表示主观题作业排序上的一致性越高。

4.4.3 实验结果与分析

（1）序数估计技术的配对顺序正确占比

表 4-2 展示了不同主观题同行互评序数评判技术对主观题作业估计的配对排名顺序正确占比，即肯德尔等级相关系数 KTRCC。实验中每种主观题同行互评序数估计技术执行 10 次算法，以估计的作业配对排名顺序相对于基准配对排名顺序（教师对作业的评判）计算得到 10 个 KTRCC 的均值。由表 4-2 可知，中位数技术和均值技术对主观题作业估计的配对排名顺序正确占比最小，这是因为这两种技术在估计真实分数时仅对互评分数进行简单的计算，均未考虑同行评价者和评价者之间的评价关系对主观题作业真实分数的影响。由于考虑了同行评价者对主观题作业知识点的掌握程度对其评分可靠性的影响，本文提出的基于认知诊断的同行互评序数估计技术 BT+CD 对三次主观题作业估计的排名顺序正确占比 KTRCC 均明显比其它技术更高。特别地，BT+CD 技术对三

次作业估计的配对顺序正确占比比中位数技术平均提高了 94.8%。同时还可观察到，在三份主观题作业中，相关序数估计技术 BT、RBTL、BT+G 的 KTRCC 值差距不大，其中它们的 KTRCC 差值均未超过 0.02，且 BT+CD 技术比相关技术估计的配对顺序正确占比平均提高了 18.38%，这表明基于认知诊断得到的同行评价者对主观题作业知识点的掌握程度能够有效的提升序数估计技术的准确性。

表 4-2 不同序数估计技术的 KTRCC 比较

Table 4-2 Comparison of KTRCC of different ordinal estimation technique

	作业 1	**作业 2**	**作业 3**
均值	0.4831	0.5516	0.4105
中位数	0.3333	0.4509	0.4051
BT[51]	0.6380	0.6346	0.6369
RBTL[16]	0.6436	0.6400	0.6486
BT+G[19]	0.6503	0.6484	0.6460
BT+CD	**0.7540**	**0.7642**	**0.7650**

（2）序数估计技术的评分误差

表 4-3 展示了不同主观题同行互评序数估计技术估计主观题作业真实分数相对于基准分数的均方根误差 RMSE。由表 4-3 可知，中位数技术和均值技术在估计真实分数时均未考虑同行评价者的评分可靠性，所以这两种技术的 RMSE 最大。由于考虑了基于认知诊断模型得到的评价者对知识点的掌握程度对其评分可靠性的影响，本文提出的序数估计技术 BT+CD 对三次主观题作业的估计误差 RMSE 均明显低于其它技术。其中，BT+CD 技术比中位数技术（目前 MOOC 平台采用的主流同行互评技术）对三次主观题作业的估计误差平均降低了 41.4%。同时还可观察到，在三次主观题作业中，BT+CD 技术比其他相关序数估计技术的作业估计更为准确，实验结果证实了结合评价者对知识点的掌握程度信息建模可靠性对于同行互评序数估计的有效性。

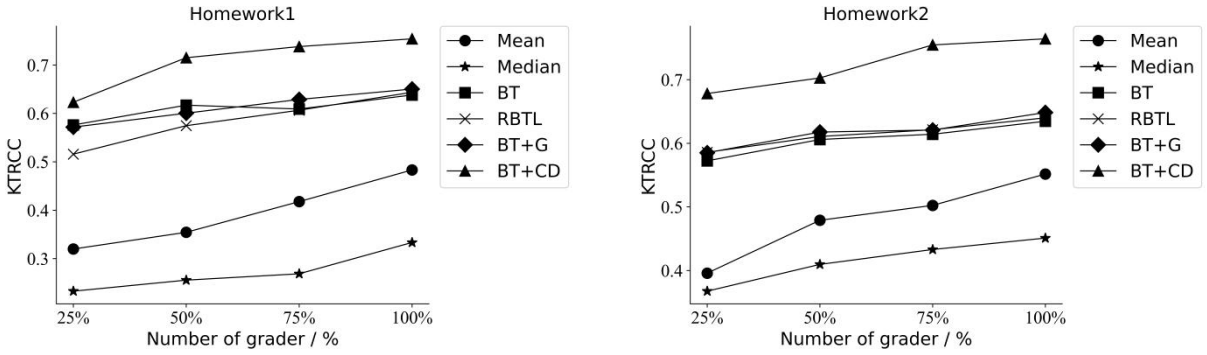
表 4-3 不同序数估计技术的 RMSE 比较

Table 4-3 Comparison of RMSE of different ordinal estimation technique

	作业 1	作业 2	作业 3
均值	4.61	4.16	4.53
中位数	5.09	4.59	5.04
BT[51]	3.58	3.53	3.62
RBTL[16]	3.45	3.47	3.41
BT+G[19]	3.34	3.42	3.40
BT+CD	**2.93**	**2.87**	**2.81**

（3）同行评价者数量和评分准确性的分析

为了分析同行评价者数量和评分准确性之间的变化，实验执行多个数据集下的同行互评估计。基于同行评价者的不同数量划分多个互评数据集，不同序数估计技术在多个互评数据集中的评分准确性如图 4-2 所示。其中，图 4-2(a)(b)(c)和图 4-2(d)(e)(f)分别表示各个序数估计技术在不同评价者数量下对三次作业估计的配对正确排序占比 KTRCC（值越高越准确）和评分误差 RMSE（值越低越准确）。如图 4-2 所示，当评价者逐渐减少时所有的估计技术评分准确性均有降低的趋势。



(a) (b)

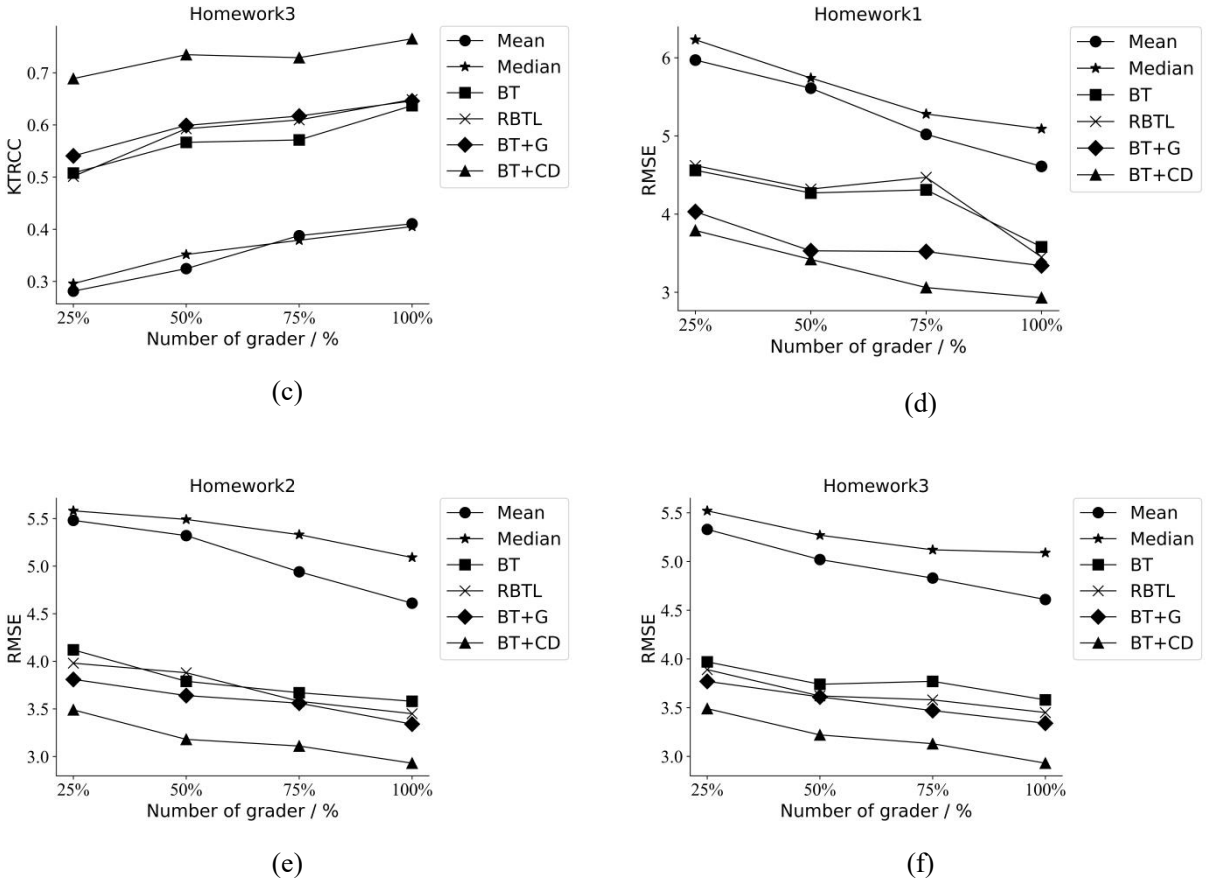


图 4-2 评价者数在不同技术下的 KTRCC 和 RMSE 分析

Fig.4-2 Analysis of KTRCC and RMSE of different techniques when varying the number of peer graders

（4）评价作业数量和评分准确性的分析

该部分分析同行评价者的评价作业数量对不同同行互评估计技术评分准确性的影响。基于同行评价者评价作业的不同数量划分多个同行互评数据集，不同序数估计技术在多个互评数据集中的评分准确性随评价作业数量（当数量减小时，评价者的批改负担会减小）的变化如图 4-3 所示。其中，图 4-3(a)(b)(c)和图 4-3(d)(e)(f)分别表示各个序数估计技术在不同评价作业数量下对三次主观题作业估计的配对正确排序占比 KTRCC（值越高越准确）和评分误差 RMSE（值越低越准确）。如图 4-3 所示，当评价者需要评价的主观题作业数量逐渐增加时所有同行互评估计技术的评分准确性均有增大的趋势。

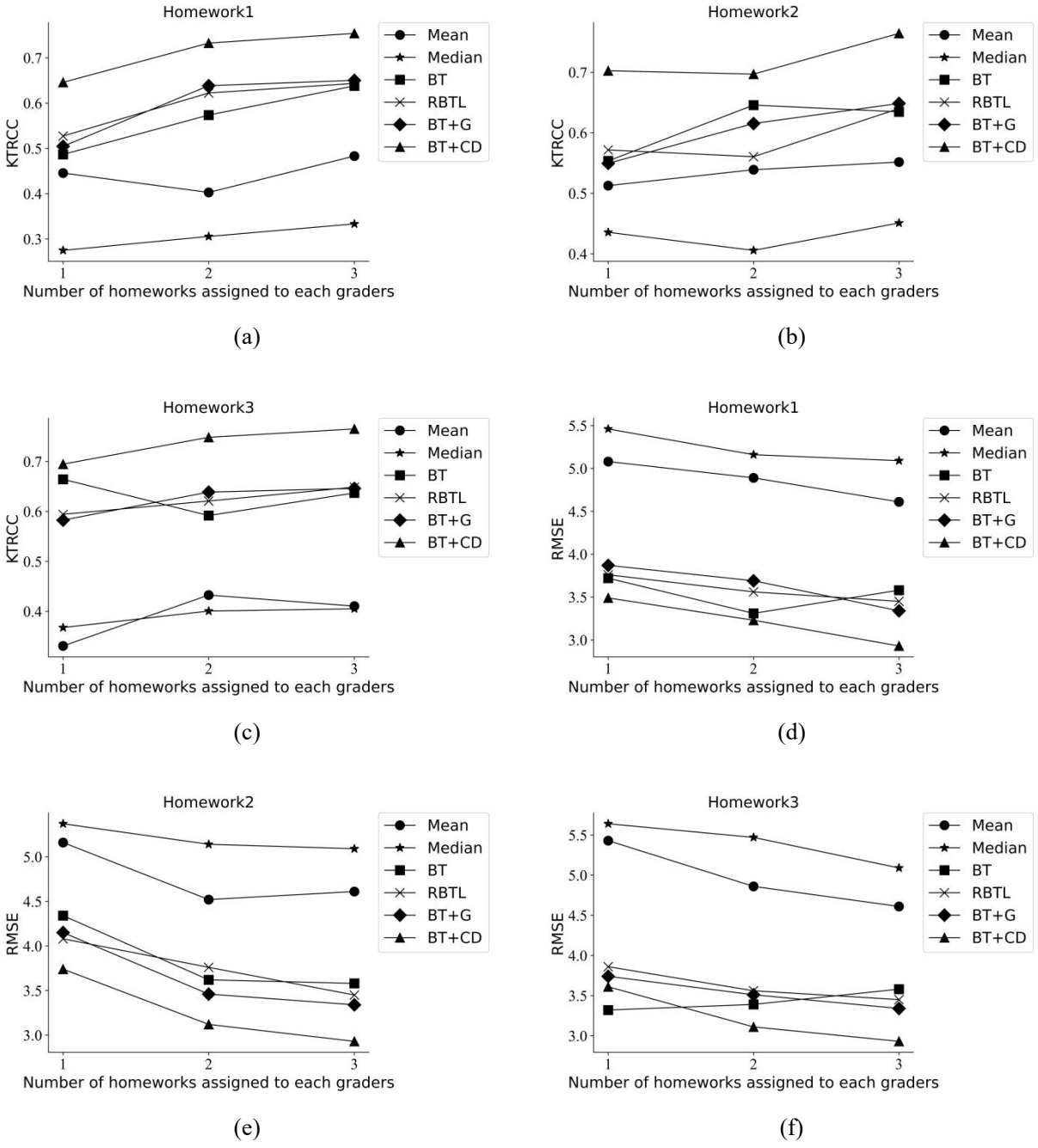


图 4-3 评价作业数量在不同技术下的 KTRCC 和 RMSE 分析

Fig.4-3 Analysis of KTRCC and RMSE of different techniques when varying the number of homeworks assigned to each graders

（5）序数估计技术的计算效率

表 4-4 展示了 BT、RBTL、BT+G 和 BT+CD 四种序数估计技术对三次主观题作业的推断算法分别运行 10 次的平均时间。由表 4-4 可知，所有的同行互评序数估计技术都是易于处理的，对三次主观题作业的推断算法运行时间均在秒数之内。其中，各个序

数估计技术对作业 2 进行推断时均耗时最长，这是因为作业 2 收集到了最多的同行互评配对比较数（详见表 4-1），而同行互评配对比较数越大时序数推断的运行时间会越长。同时还可以观察到在所有作业中，RBTL 技术耗时最长，这是因为 RBTL 技术需要基于每个同行评价者的评价能力和内在能力的线性关系对评价能力进行计算。

表 4-4 不同序数估计技术的运行时间(s)比较

Table 4-4 Comparison of runtime of different ordinal estimation technique in seconds

	作业 1	**作业 2**	**作业 3**
BT[51]	0.59	0.98	0.48
RBTL[16]	5.54	11.44	4.95
BT+G[19]	0.55	1	0.47
BT+CD	0.56	0.97	0.48

4.5 本章小结

针对主观题作业的序数同行互评技术能够降低学生互评评判的难度，具有重要的研究意义。然而，在序数的同行互评场景中，现有的主观题同行互评序数估计技术均没有考虑评价者对主观题知识点的掌握程度对于同行评价者的可靠性的影响，进而影响了作业估计的准确性。针对现有研究工作的不足，提出了一种基于认知诊断的同行互评序数估计技术 BT+CD，该技术首先基于 DINA 模型诊断得到的该评价者对本次作业题的掌握程度，以掌握程度信息对评价者的可靠性进行建模，最后基于给定作业的排名和评价者的可靠性的概率分布关系建立同行互评序数估计模型。通过对主观题作业的序数同行互评数据进行实验与分析，表明提出的 BT+CD 技术能够提升对主观题作业估计的配对顺序正确占比比例和降低评分误差，进而表明该技术能有效提升序数同行互评的评分准确性。

第五章 总结与展望

5.1 研究工作总结

随着大型开放式网络课程（MOOCs）的流行，全世界的学生都能访问到优质的教学资源，提升了教育的公平性。MOOCs 的流行也给授课教师带来了严峻的教学挑战，突出体现在一名授课教师可能需要批改上千名学生提交的主观题作业。同行互评是当前 MOOCs 平台解决大规模主观题作业批改问题的重要手段。同行互评估计方式分为基数估计和序数估计，在这两种估计方式下，同行评价者的评分偏见和评分可靠性是未知的，因此基于多个同行评价者给出的评价分数估计主观题作业的真实分数或真实排名是一个具有挑战的问题。现有同行互评技术利用概率模型对同行评价者的评分可靠性和评分偏见进行建模，有效提高了估计主观题作业的真实分数或真实排名的准确性。然而，这些基数和序数同行互评估计技术均未考虑同行评价者对主观题考察的知识点的掌握程度对其评分可靠性造成的影响。鉴于此，本文提出了基于认知诊断的同行互评关键技术，在同行互评中基数和序数的场景下分别实现了基于认知诊断的基数估计技术和基于认知诊断的序数估计技术。具体而言，本文主要完成了以下几个方面研究工作：

（1） 本文阐述了同行互评在众包、论文和基金评审、教育领域中的研究工作，并对同行互评中的基数估计和序数估计的相关技术进行了介绍和分析，对认知诊断理论、同行互评技术涉及的重要概念进行了详细的描述。

（2） 本文提出以认知诊断得到的同行评价者对主观题的掌握程度信息和评价者在该主观题中取得的真实分数信息作为评分可靠性的建模依据，并结合评价者的评分偏见，设计并实现了基于认知诊断的基数估计的概率模型 PG_8 和 PG_9，并且基于 Gibbs 采样技术得到学生的真实分数，使得在同行互评的基数估计场景下有效提高学生真实分数的准确率。

（3） 本文利用评价者的历史答题结果数据，基于认知诊断得到的学生对于主观题作业中知识点的掌握程度对评价者的可靠性建模，在同行互评的序数估计场景下设计并实现了序数估计模型 BT+CD，并且根据提出的同行互评序数估计的推断算法得到全局

排名，使得在同行互评的序数估计场景下有效提高真实排名的准确率。

（4） 通过真实在线同行互评系统先后组织了 284 名学生完成客观题和 3 份主观题作业和基数互评活动，基于收集的真实基数互评数据集进行实验，实验验证了基于认知诊断的基数估计技术 PG_8 和 PG_9 的评估准确性。真实课堂实验的实验结果表明，本文提出的 PG_8 和 PG_9 技术比相关技术平均提高了 42%的准确率。

（5） 将本文提出的基于认知诊断的序数估计技术 BT+CD 在真实序数互评数据集上进行实验，实验验证了 BT+CD 技术的有效性。通过实验分析证明，本文提出的基于认知诊断的序数估计技术 BT+CD 比相关技术具有更高的评分准确性，比相关的序数估计技术在配对顺序正确占比上平均提高了 18.38%。

综上所述，本文的创新点主要包括：

（1） 发现基于同行评价者的历史答题结果数据，并利用认知诊断模型诊断得到评价者对主观题考察的知识点的掌握程度信息，可以有助于提高基于同行互评的主观题评判方法的评分准确性。

（2） 提出一种基于认知诊断的同行互评基数估计技术。该技术提出了改进现有基数估计场景下的同行互评概率模型的思路，基于认知诊断 DINA 模型得到的同行评价者对于主观题作业中知识点的掌握程度，设计了两个概率图模型 PG_8 和 PG_9，还基于 Gibbs 采样技术推断概率图模型中的隐含变量，设计并实现了基于认知诊断的基数估计算法，以期进一步提高在基数同行互评活动中主观题作业真实分数估计的准确性。

（3） 提出了一种基于认知诊断的同行互评序数估计 BT+CD 技术。BT+CD 技术基于认知诊断技术得到的同行评价者对主观题的掌握程度信息对评价者的可靠性建模，提出了同行互评序数估计模型，设计并实现了基于认知诊断的序数估计推断算法，从而提升了在序数同行互评活动中主观题作业估计的有效性和精确性。

（4） 通过组织数据库原理教学班的 284 名本科生于自主研发的在线教育系统中完成 40 道客观题测试、3 次主观题作业和同行互评活动。基于学生提交的客观题答案使用流行的认知诊断 DINA 模型计算学生对知识点的掌握程度。继而基于收集的基数主观题互评数据集对本文所提出基于认知诊断的同行互评基数估计技术进行实验，实验验证并分析该技术的有效性，该技术在主观题作业的真实分数估计误差上相对于其它基数估计技术平均降低了 42%。同时使用主观题作业序数同行互评中收集的数据集对本文所提出

基于认知诊断的同行互评序数估计技术进行实验，通过与相关工作的对比证明了本文所提出技术的准确性，提出的序数估计技术比相关序数估计技术在配对顺序正确占比上平均提高了 18.38%。

5.2 展望

本文基于认知诊断的同行互评基数估计技术和同行互评序数估计技术虽然取得了较好的实验效果，但是仍然存在可以进一步完善和优化的地方，本文的未来工作包括：

（1） 本文发现了评价者对知识点的掌握程度对评分误差的影响，在今后的研究工作中还可以探索其他影响评价者的评分可靠性和评分偏见的因素，比如评价者的评语和评价题目难度，以进一步优化同行互评估计的概率模型，提高在基数或者序数同行互评活动中真实分数估计的准确率。

（2） 本文提出的同行互评关键技术中分析了评价者的真实分数会受到同行评价者的可靠性和评价偏见的影响，而评价者的真实分数是在学习活动中做作业得到的结果，因此未来工作还可以考虑分析评价者的可靠性和偏见和学习活动之间的关系，以预测学生辍学、学习成绩上升或下降的趋势。

（3） 本文在教育平台实现了对主观题作业的同行互评关键技术，除了将同行互评应用于教育领域之外，还可以应用于众包、论文和基金评审等领域中，希望能够进一步验证基数和序数估计技术应用于其他领域的有效性。

（4） 未来研究工作可以将本文提出的基于认知诊断的同行互评基数估计技术和序数估计技术在更大规模的课堂实践上推广应用，并利用更多的同行互评数据优化提出的技术。

参考文献

[1] Paré D E, Joordens S. Peering into large lectures: examining peer and expert mark agreement using peerScholar, an online peer assessment tool[J]. Journal of Computer Assisted Learning, 2008, 24(6): 526-540.

[2] Caragiannis I, Krimpas G A, Voudouris A A. Aggregating partial rankings with applications to peer grading in massive online open courses[C]. Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, 2015: 675-683.

[3] Ratna A A P, Raharjo B S, Purnamasari P D, et al. Automatic essay grading system with latent semantic analysis and learning vector quantization[C]. Proceedings of the 3rd International Conference on Communication and Information Processing (ICCIP), 2017: 158-163.

[4] Ratna A A P, Santiar L, Ibrahim I,et al. Latent semantic analysis and winnowing algorithm based automatic japanese short essay answer grading system comparative performance[C]. IEEE 10th International Conference on Awareness Science and Technology (iCAST), 2019: 1-7.

[5] Lan A S, Vats D, Waters A E, et al. Mathematical language processing automatic grading and feedback for open response mathematical questions[C]. Proceedings of the Second ACM Conference on Learning @ Scale (L@S), 2015: 167-176.

[6] Piech C, Huang J, Chen Z, et al. Tuned models of peer assessment in moocs[C]. Proceedings of the 6th International Conference on Educational Data (EDM), 2013: 153-160.

[7] Mi F, Yeung D Y. Probabilistic graphical models for boosting cardinal and ordinal peer grading in moocs[C]. Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI), 2015: 454-460.

[8] Chan H P, King I. Leveraging social connections to improve peer assessment in moocs[C]. Proceedings of the 26th International Conference on World Wide Web (WWW), 2017:

341-349.

[9] Wang T Q, Li Q, Gao J, et al. Improving peer assessment accuracy by incorporating relative peer grades[C].Proceedings of the 12th International Conference on Educational Data (EDM), 2019: 450-455.

[10] Song Y, Hu Z, Guo Y, et al. An experiment with separate formative and summative rubrics in educational peer assessment[C]. 2016 IEEE Frontiers in Education Conference (FIE), 2016: 1-7.

[11] Kulkarni C E, Wei W P, Le H, et al. Peer and self assessment in massive online classes[J]. ACM Transactions on Computer-Human Interaction, 2013, 20(6): 1-31.

[12] Gehringer E F. A survey of methods for improving review quality[C]. New Horizons in Web Based Learning-ICWL 2014 International Workshops, 2014: 92-97.

[13] Stewart N, Brown G D A, and Chater N. Absolute identification by relative judgment[J]. Psychological Review, 2005, 112(4):881-911.

[14] Carterette B, Bennett P N, Chickering D M, et al. Here or there[C]. In Advances in Information Retrieval, 2008, 16-27.

[15] Shah N. B, Balakrishnan S, Bradley J, et al. When is it better to compare than to score?[Z]. arXiv preprint arXiv:1406.6618, 2014.

[16] Shah N B, Bradley J K, Parekh A, et al. A case for ordinal peer-evaluation in moocs[C]. In NIPS Workshop on Data Driven Education, 2013: 1-8.

[17] Alfaro L D, Shavlovsky M. Dynamics of peer grading: an empirical study[C]. Proceedings of the 9th International Conference on Educational Data (EDM), 2016: 62-69.

[18] Capuano N, Caballé S. Towards adaptive peer assessment for MOOCs[C]. 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2015: 64-69.

[19] Raman K, Joachims T. Methods for ordinal peer grading[C]. The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2014: 1037-1046.

[20] Torre D L. DINA model and parameter estimation: a didactic[J]. Journal of Educational and Behavioral Statistics, 2008, 34(1): 115-130.

[21] Snow R, Connor B O, Jurafsky D, et al. Cheap and fast - but is it good? Evaluation non-expert annotiations for natural language tasks[C]. Conference on Empirical Methods in Natural Language Processing, 2008:254-263.

[22] Whitehill J, Ruvolo P, Wu T, et al. Whose vote should count more: optimal integration of labels from labelers of unknown expertise[C]. 23rd Annual Conference on Neural Information Processing Systems, 2009:2035-2043.

[23] Guan M Y , Gulshan V , Dai A M, et al. Who said what: modeling individual labelers improves classification[C]. Proceedings of the Thirty-Second Conference on Artificial Intelligence, 2018: 3109-3118.

[24] Isupova O, Li Y, Kuzin D, et al. BCCNet: Bayesian classifier combination neural network[Z]. CoRR abs/1811.12258, 2018.

[25] Rodrigues F, Pereira F C. Deep Learning from Crowds[C]. Proceedings of the Thirty-Second Conference on Artificial Intelligence, 2018: 1611-1618.

[26] Aydin B I, Yilmaz Y S, Li Y, et al. Crowdsourcing for multiple-choice question answering[C]. Proceedings of the Twenty-Eighth Conference on Artificial Intelligence, 2014: 2946-2953.

[27] Yue D , Yu G , Shen D , et al. A weighted aggregation rule in crowdsourcing systems for high result accuracy[C]. IEEE International Conference on Dependable, 2014.

[28] Li W , Huhns M N , Tsai W T , et al. Collaborative majority vote: improving result quality in crowdsourcing marketplaces[J]. 10.1007/978-3-662-47011-4(Chapter 8), 2015: 131-142.

[29] Li H, Yu B. Error rate bounds and iterative weighted majority voting for crowdsourcing[Z]. CoRR abs/1411.4086, 2014.

[30] Rooyen S V, Godlee F, Evans S, et al. Effect of blinding and unmasking on the quality of peer review[J]. Journal of General Internal Medicine, 1999, 14(10):622-624.

[31] Nobarany S, Booth K S, et al. Understanding and supporting anonymity policies in peer

review[J]. Journal of the Association for Information Science & Technology, 2017, 68(4):957-971.

[32] Li B, Hou Y T. The new automated IEEE INFOCOM review assignment system[J]. IEEE Network, 2016, 30(5):18-24.

[33] Price S, Flach P A. Computational support for academic peer review: a perspective from artificial intelligence[J]. Communications of the ACM, 2017, 60(3):70-79.

[34] Charlin L, Zemel R S, Boutilier C. A framework for optimizing paper matching[C]. Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, 2011:86-95.

[35] Long C, Wong R C W, Peng Y, et al. On good and fair paper-reviewer assignment[C]. IEEE 13th International Conference on Data Mining, 2013:1145-1150.

[36] Xiong J, Zhang Q, Peng Z, et al. Double sample data fusion method based on combination rules[J]. IEEE Access, 2016, 4(99):7487-7499.

[37] Yang J B, Xu D L. Evidential reasoning rule for evidence combination[J]. Artificial Intelligence, 2013, 205(12):1-29.

[38] Zhu W D, Liu F, Chen Y W, et al. Research project evaluation and selection: an evidential reasoning rule-based method for aggregating peer review information with reliabilities[J]. Scientometrics, 2015, 105(3):1469-1490.

[39] Zhu W, Li S, Ku Q, et al. Evaluation information fusion of scientific research project based on evidential reasoning approach under two-dimensional frames of discernment[J]. IEEE Access, 2020, 8:1-1.

[40] Liu F, Zhu W, Chen Y W, et al. Evaluation, ranking and selection of R&D projects by multiple experts: an evidential reasoning rule based approach[J]. Scientometrics, 2017, 111(3): 1501-1519.

[41] Du Y W, Yang N, Ning J. IFS/ER-based large-scale multiattribute group decision-making method by considering expert knowledge structure[J]. Knowledge-Based Systems, 2018, 162(12):124-135.

[42] Omran A M B, Aziz M J A. Automatic essay grading system for short answers in english

language[J]. Journal of Computer Science, 2013, 9(10): 1369-1382.

[43] Alfaro L D, Shavlovsky M. Crowdgrader: a tool for crowdsourcing the evaluation of homework assignments[C]. The 45th ACM Technical Symposium on Computer Science Education (SIGCSE), 2014: 415–420.

[44] Walsh T. The peerrank method for peer assessment[C]. 2014-21st European Conference on Artificial Intelligence (ECAI), 2014: 909–914.

[45] Page L, Brin S, Motwani R. The pagerank citation ranking: bringing order to the web[R]. Stanford Digital Library Technologies Project, 1999.

[46] Gutierrez P, Osman N, Sierra C. Collaborative assessment[C]. In Proceedings of the 17th International Conference of the Catalan Association for Artificial Intelligence, 2014: 136-145.

[47] Singla P, Richardson M. Yes, there is a correlation: from social networks to personal behavior on the web[C]. Proceedings of the 17th International Conference on World Wide Web (WWW), 2008: 655-664.

[48] Yang S H, Long B, Smola A J, et al. Like like alike: joint friendship and interest propagation in social networks[C]. Proceedings of the 20th International Conference on World Wide Web (WWW), 2011: 537-546.

[49] Mallows C L. Non-Null ranking models. I[J]. Biometrika, 1957, 44(1/2):114-130.

[50] Plackett R L. The analysis of permutations[J]. Journal of the Royal Statistical Society Series C, 1975, 24(2):193-202.

[51] Bradley R A, Terry M E. Rank analysis of incomplete block designs: I. The Method of Paired Comparisons[J]. Biometrika, 1952, 39(3/4): 324-345.

[52] Wauthier F L, Jordan M I, Jojic N. Efficient ranking from pairwise comparisons[C]. Proceedings of the 30th International Conference on Machine Learning (ICML), 2013(3): 109–117.

[53] Waters A E, Tinapple D, Baraniuk R G. Bayesrank: A bayesian approach to ranked peer grading[C]. Proceedings of the Second ACM Conference on Learning @ Scale (L@S), 2015: 177-183.

[54] Luacesa O, Díeza J, Betanzosb A A, et al. A factorization approach to evaluate open-response assignments in moocs using preference learning on peer assessments[J]. Knowledge-Based Systems, 2015, 85(9): 322-328.

[55] Capuano N, Loia V, Orciuoli F. A fuzzy group decision making model for ordinal peer assessment[J]. IEEE Transactions on Learning Technologies, 2017, 10(2):247-259.

[56] Thurstone, L. L. The method of paired comparisons for social values[J]. Journal of Abnormal & Social Psychology, 1927, 21(4):384-400.

[57] Wu R Z, Liu Q, Liu Y P, et al. Cognitive modelling for predicting examinee performance[C]. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI), 2015: 1017-1024.

[58] Zhu T Y, Liu Q, Huang Z Y, et al. MT-MCD: A multi-task cognitive diagnosis framework for student assessment[C]. Database Systems for Advanced Applications-23rd International Conference, 2018(2): 318-335.

[59] Cheng S, Liu Q, Chen E H, et al. DIRT: Deep learning enhanced item response theory for cognitive diagnosis[C]. Proceedings of the 28th International Conference on Information (CIKM), 2019: 2397-2400.

[60] 朱天宇, 黄振亚, 陈恩红,等. 基于认知诊断的个性化试题推荐方法[J]. 计算机学报, 2017, 40(1):178-193.

[61] 王超, 刘淇, 陈恩红,等. 面向大规模认知诊断的 DINA 模型快速计算方法研究[J]. 电子学报, 2018, 46(5): 1047-1055.

[62] Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(6): 721-741.

符号说明

符号	意义
U	学生集合被评价者集合，u_i 表示第 i 个被评价者
V	同行评价者集合，$v \in V$ 表示某评价者
U_v	作业被评价者 v 评判的被评价者集合
V_{ui}	评判被评价者 u_i 作业的同行评价者集合
$\boldsymbol{R}$	评价者-历史客观题得分矩阵
$\boldsymbol{Q}$	历史客观题-知识点关联矩阵
$\boldsymbol{\alpha_v}$	评价者 v 的知识点掌握程度向量
$\boldsymbol{M}$	所有评价者的知识点掌握程度矩阵
δ_v	评价者 v 对主观题的潜在正确作答概率
s_i	被评价者 u_i 的主观题作业的真实分数
τ_v	评价者 v 的可靠性
b_v	评价者 v 的偏见
z_i^v	评价者 v 给被评价者 u_i 的主观题作业的评分
Z	所有评价者的互评分数集合
d_{ij}^v	即相对分数，表示评价者 v 给被评价者 u_i 的作业评分与给 u_j 的作业评分之差
D	所有评价者的相对分数集合

附录

基数估计概率图模型各个隐含变量的联合后验分布如下：

$$P(Z,D|\{s_i\}_{u_i\in U},\{b_v\}_{v\in V},\{\tau_v\}_{v\in V})=\prod_i P(s_i|\mu_0,\gamma_0)\cdot\prod_v P(b_v|\eta_0)\cdot P(\tau_v|\delta_v,s_v,\theta_1,\theta_2,\beta_0)\cdot\prod_{z_i^v} P(z_i^v|s_i,b_v,\tau_v)\cdot\prod_{d_{ij}^v} P(d_{ij}^v|s_i,s_j,\tau_v). \quad (1)$$

马尔科夫毯 MB（Markov Blanket）表示满足如下特性的一个最小特征子集：一个特征在其马尔科夫毯条件下，与特征域中所有其他特征条件独立。例如，MB(s_i)表示被评价者 u_i 的主观题作业真实分数 s_i 独立于其他特征变量（即评价者的评分偏见 b_v 和评分可靠性τ_v）。因此，在推断变量 s_i 时，需要求解变量 s_i 的值是固定不变的，而其他变量是可以随机初始化的。对 PG_8 模型中的各个隐含变量的推断采样过程将在（1）（2）（3）分别阐述，由于 PG_9 模型中真实分数 s 和评分偏见 b 的推理过程和 PG_8 相似，故不在此赘述。故将在（4）中对 PG_9 模型中的评分可靠性τ_v 的推理进行描述。

（1）PG_8 模型真实分数 s_i 的推断过程

考虑同行互评活动中的一个被评价者 u_i 且固定不变，那么对于 u_i 提交的主观题作业的真实分数 s_i 的采样步骤如下：

$$\begin{aligned}
s &\sim P(s_i\,|\,MB(s_i)),\\
&\propto P(s_i\,|\,\mu_0,\gamma_0)\cdot P(\tau_i\,|\,\delta_i,s_i,\theta_1,\theta_2,\beta_0)\cdot\prod_{v\in V_{u_i}} P(z_i^v|s_i,b_v,\tau_v)\cdot\prod_{v\in V_{u_i},u_j\in U_v} P(d_{ij}^v|s_i,s_j,\tau_v),\\
&\propto \frac{\beta_0^{(\theta_1\delta_i+\theta_2 s_i)}\cdot\tau_i^{(\theta_1\delta_i+\theta_2 s_i-1)}}{\Gamma(\theta_1\delta_i+\theta_2 s_i)}\times\exp\left(-\frac{1}{2}\gamma_0(s_i-\mu_0)^2-\beta_0\tau_i\right)\times\\
&\quad\exp\left(\sum_{v\in V_{u_i}}\left(-\frac{1}{2}\tau_v\left(z_i^v-(s_i+b_v)\right)^2\right)+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\left(-\frac{1}{4}\tau_v\left(d_{ij}^v-(s_i-s_j)\right)^2\right)\right), \quad (2)\\
&\propto \frac{\beta_0^{(\theta_2 s_i)}\cdot\tau_i^{(\theta_2 s_i-1)}}{\Gamma(\theta_1\delta_i+\theta_2 s_i)}\times\exp\left(-\frac{1}{2}\left(\gamma_0(s_i-\mu_0)^2\right)\right)\times\\
&\quad\exp\left(-\frac{1}{2}\left[\sum_{v\in V_{u_i}}\left(\tau_v\left(z_i^v-(s_i+b_v)\right)^2\right)+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\left(\frac{1}{2}\tau_v\left(d_{ij}^v-(s_i-s_j)\right)^2\right)\right]\right).
\end{aligned}$$

指数中的表达式是二次的，对此完全平方：

$$
\begin{aligned}
&\gamma_0(s_i-\mu_0)^2+\sum_{v\in V_{u_i}}\left(\tau_v\left(z_i^v-(s_i+b_v)\right)^2\right)+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\left(\frac{1}{2}\tau_v\left(d_{ij}^v-(s_i-s_j)\right)^2\right)\\
&=const.+\gamma_0\left(s_i{}^2-2\mu_0 s_i\right)+\\
&\qquad\sum_{v\in V_{u_i}}\tau_v\left((s_i+b_v)^2-2z_i^v(s_i+b_v)\right)+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\left(\frac{1}{2}\tau_v\left((s_i-s_j)^2-2d_{ij}^v(s_i-s_j)\right)\right),\\
&=const.+\left(\gamma_0+\sum_{v\in V_{u_i}}\tau_v+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v}{2}\right)s_i{}^2-2s_i\left(\mu_0\gamma_0+\sum_{v\in V_{u_i}}\tau_v\left(z_i^v-b_v\right)\right)-\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}s_i\tau_v\left(d_{ij}^v+s_j\right),\\
&=const.+R\left(s_i-\frac{Y}{R}\right)^2,\\
&where\ R=\gamma_0+\sum_{v\in V_{u_i}}\tau_v+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v}{2},\ and\ Y=\mu_0\gamma_0+\sum_{v\in V_{u_i}}\tau_v\left(z_i^v-b_v\right)+\sum_{v\in V_{u_i}}\sum_{u_j\in U_v}\frac{\tau_v\left(d_{ij}^v+s_j\right)}{2}.
\end{aligned}
\tag{3}
$$

因此，隐含变量 s 的样本分布如下：

$$
s\propto\frac{\beta_0^{(\theta_2 s_i)}\cdot\tau_i^{(\theta_2 s_i-1)}}{\Gamma(\theta_1\delta_i+\theta_2 s_i)}\times\exp\left(R\left(s_i-\frac{Y}{R}\right)^2\right)
\tag{4}
$$

（2）PG_8 模型评分可靠性τ_v的推断过程

考虑同行互评活动中的一个评价者 v 且固定不变，同行评价者的评分可靠性τ_v的采样步骤如下：

$$
\begin{aligned}
&\tau\sim P(\tau_v\mid MB(\tau_v)),\\
&\propto P(\tau_v\mid\delta_v,s_v,\theta_1,\theta_2,\beta_0)\cdot\prod_{u_i\in U_v}P(z_i^v\mid s_i,b_v,\tau_v)\cdot\prod_{u_i,u_j\in U_v}P(d_{ij}^v\mid s_i,s_j,\tau_v),\\
&\propto\tau_v^{(\theta_1\delta_v+\theta_2 s_v)}\times\exp\left(-\beta_0\tau_v+\sum_{u_i\in U_v}\sqrt{\frac{\tau_v}{2\pi}}\left(-\frac{\tau_v}{2}\left(z_i^v-(s_i+b_v)\right)^2\right)\right)\\
&\qquad\times\exp\sum_{u_i,u_j\in U_v}\sqrt{\frac{\tau_v}{4\pi}}\left(-\frac{\tau_v}{4}\left(d_{ij}^v-(s_i-s_j)\right)^2\right),\\
&\propto\tau_v^{(\theta_1\delta_v+\theta_2 s_v)+\frac{|U_v|^2}{2}}\times\exp\left[-\beta_0+\frac{1}{2}\left(\sum_{u_i\in U_v}\left(z_i^v-s_i-b_v\right)^2+\sum_{u_i,u_j\in U_v}\left(d_{ij}^v-s_i+s_j\right)^2\right)\right]\tau_v.
\end{aligned}
\tag{5}
$$

由此可知，隐含变量τ_v的样本分布是伽马分布：

$$
\tau\sim\Gamma(\theta_1\delta_v+\theta_2 s_v+\frac{|U_v|^2}{2},\ \beta_0+\frac{\sum_{u_i\in U_v}(z_i^v-s_i-b_v)^2+\sum_{u_i,u_j\in U_v}\frac{1}{2}(d_{ij}^v-s_i+s_j)^2}{2})
\tag{6}
$$

（3）PG$_8$模型评分偏见 b_v 的推断过程

评价者的评分偏见 b_v 的采样步骤如下：

$$\begin{aligned} b &\sim P(b_v \mid MB(b_v)), \\ &\propto P(b_v \mid \eta_0) \cdot \prod_{u_i \in U_v} P(z_i^v \mid s_i, b_v, \tau_v), \\ &\propto \exp\left(-\frac{1}{2}\eta_0 b_v^{\,2} - \frac{1}{2}\sum_{u_i \in U_v} \tau_v \left(z_i^v - (s_i + b_v) \right)^2 \right), \\ &\propto \exp\left(-\frac{1}{2}\left[\eta_0 b_v^{\,2} + \sum_{u_i \in U_v} \tau_v \left((s_i + b_v)^2 - 2 z_i^v (s_i + b_v) \right) \right] \right). \end{aligned} \tag{7}$$

指数中的表达式是二次的，对此完全平方：

$$\begin{aligned} &\eta_0 b_v^{\,2} + \sum_{u_i \in U_v} \tau_v \left((s_i + b_v)^2 - 2 z_i^v (s_i + b_v) \right) \\ &= const. + \left(\eta_0 + \sum_{u_i \in U_v} \tau_v \right) b_v^{\,2} - 2\left(\sum_{u_i \in U_v} \tau_v \left(z_i^v - s_i \right) \right) b_v, \\ &= const. + R\left(b_v - \frac{Y}{R} \right)^2, \\ &\quad where\ \ R = \eta_0 + \sum_{u_i \in U_v} \tau_v = \eta_0 + |U_v|\tau_v, and\ \ Y = \sum_{u_i \in U_v} \tau_v \left(z_i^v - s_i \right). \end{aligned} \tag{8}$$

所以隐含变量 b 的样本分布为高斯分布：

$$b \sim N\left(\frac{\sum_{u_i \in U_v} \tau_v (z_i^v - s_i)}{\eta_0 + |U_v|\tau_v}, \frac{1}{\eta_0 + |U_v|\tau_v} \right) \tag{9}$$

（4）PG$_9$模型评分可靠性 τ_v 的推断过程

考虑同行互评活动中的一个评价者 v 且固定不变，同行评价者的评分可靠性 τ_v 的采样步骤如下：

$$
\begin{aligned}
&\tau \sim P(\tau_v \mid MB(\tau_v)),\\
&\propto P(\tau_v \mid \delta_v, s_v, \theta_1, \theta_2, \beta_0) \cdot \prod_{u_i \in U_v} P(z_i^v \mid s_i, b_v, \tau_v) \cdot \prod_{u_i, u_j \in U_v} P(d_{ij}^v \mid s_i, s_j, \tau_v),\\
&\propto \exp\left(-\frac{\beta_0}{2}\left(\tau_v - \left(\theta_1 \delta_v + \theta_2 s_v\right)\right)^2 - \frac{1}{2} \sum_{u_i \in U_v} \sqrt{\frac{\tau_v}{2\pi}} \left(-\frac{\tau_v}{\lambda}\left(z_i^v - \left(s_i + b_v\right)\right)^2 \right) \right)\\
&\qquad \times \exp\left(-\frac{1}{2} \sum_{u_i, u_j \in U_v} \sqrt{\frac{\tau_v}{4\pi}} \left(-\frac{\tau_v}{2\lambda}\left(d_{ij}^v - \left(s_i - s_j\right)\right)^2 \right) \right),\\
&\propto \tau_v^{\frac{|U_v|^2}{2}} \times \exp\left(-\frac{\beta_0}{2}\left(\tau_v - \left(\theta_1 \delta_v + \theta_2 s_v\right)\right)^2 - \sum_{u_i \in U_v} \frac{\left(z_i^v - \left(s_i + b_v\right)\right)^2 \tau_v}{2\lambda} \right)\\
&\qquad \times \exp \sum_{u_i, u_j \in U_v} -\frac{\left(d_{ij}^v - \left(s_i - s_j\right)\right)^2 \tau_v}{4\lambda}, \qquad (10)\\
&\propto \tau_v^{\frac{|U_v|^2}{2}} \times\\
&\quad \exp\left(-\frac{\beta_0 \tau_v^{\,2}}{2} + \left(\theta_1 \delta_v + \theta_2 s_v\right)\tau_v + \sum_{u_i \in U_v} \frac{\left(z_i^v - s_i - b_v\right)^2 \tau_v}{\lambda} + \sum_{u_i, u_j \in U_v} \frac{\left(d_{ij}^v - s_i + s_j\right)^2 \tau_v}{2\lambda} \right),\\
&\propto \tau_v^{\frac{|U_v|^2}{2}} \times \exp\left(-\frac{\beta_0}{2}\left(\tau_v - Y\right)^2 \right),\\
&\quad where\ Y = \theta_1 \delta_v + \theta_2 s_v + \sum_{u_i \in U_v} \frac{\left(z_i^v - s_i - b_v\right)^2}{\lambda \beta_0} + \sum_{u_i, u_j \in U_v} \frac{\left(d_{ij}^v - s_i + s_j\right)^2}{2\lambda \beta_0}.
\end{aligned}
$$

致谢

光阴荏苒，时光飞逝，三年的研究生生活不知不觉就过去了，回想当年刚入校时的场景，仍然历历在目。在广西大学计算机与电子信息学院，经历了我人生中成长最快的时光，这段时光的点滴收获，都离不开老师的教诲，同学的帮助以及家人的支持。

首先，衷心地感谢我的恩师许嘉副教授，我想真诚的对老师说一声，您辛苦了！许老师一直以严谨的治学精神和出彩的人格魅力影响着我，许老师有着优秀的科研素质，积极严谨的科研态度，平易近人的人格魅力。在我入学以后，老师经常组织实验室组会进行学习，让我对计算机学科各个领域的知识都有所收获，组会作报告锻炼了我的表达和沟通能力；在探讨学术问题中，老师一直给予我鼓励和教诲，和许老师的每一次讨论，都觉得受益匪浅；在论文写作中，许老师工作繁忙但也从不忘对我的论文进行细致的指导，为我的研究指明方向，耐心的逐字逐句的帮我修改论文，为我提出宝贵的修改意见。同时，我也意识到自己很多方面，仍与老师严格的要求存在一些差距，今后在工作中，我会记得老师的教诲和教导，严谨，积极，认真，继续以许老师的精神和要求，时刻提醒自己。

其次，感谢实验室的吕品老师，吕老师在我心中是一位严谨，博学，正直，富有才华，事必躬亲的实干派老师，吕老师经常和许老师一起在百忙之中为实验室同学组织组会学习，解答大家的疑惑，推动大家的科研进度。在遇到的科研问题和人生的疑惑中，吕老师都给予我很多有益的建议。依旧记得在每年的中秋圣诞等节日中，两位老师为实验室同学送去祝福和月饼，让我们在科研的同时也感受到了家的温暖。师恩难忘，祝许老师，吕老师身体健康，万事顺意！

还要感谢实验室已经毕业的师兄师姐们，王俊斌，柳开弘，潘思羽，周悦等师兄师姐为我科研提供了很多指导和帮助，在与导师讨论之余，经常和师兄师姐们交流科研问

题和心得，从另一个角度让我的科研问题有所突破，受益匪浅。还有我的同届同学，李凯，莫晓坤，贺云艳，钟国樑以及研一研二和本科的师弟师妹等，你们都是很善良很优秀的人，我也祝福师弟师妹们能早日收获自己的科研成果，有一个美好的前程。

同时，感谢我的家人对我一路的支持，你们的经济支持让我有了继续深造的可能，让我安心的完成自己的学业，希望以后能够用自己的力量也回报你们的养育之恩。

最后，感谢各位专家和老师们对我论文的评审，感谢各位老师百忙之中参加我的论文答辩，您们的宝贵意见将使我的论文更加完善！

李秋云

2021 年 4 月

攻读学位期间发表论文情况

[1] Xu J, **Li Q Y**, Liu J, et al. Leveraging Cognitive Diagnosis to Improve Peer Assessment in MOOCs[J]. IEEE Access, 2021, 9:50466-50484, doi:10.1109/ACCESS.2021.3069055.

[2] 许嘉, **李秋云**, 刘静, 等. 一种基于认知诊断的主观题同行互评技术[J]. 小型微型计算机系统, 2021, 已录用.

[3] 许嘉, **李秋云**, 刘静, 吕品. 基于概率图模型的主观题同行互评系统的开发与实践[J]. 中国教育信息化, 2020, 已录用.

[4] 许嘉, **李秋云**, 刘静, 吕品. 软件著作权：教学互评系统，完成日期：2020 年 9 月 20 日, 软著登记号: 2020SR1578458.