一种基于评价能力的概率模型同行评估技术

（1、应做到简洁明了，突出文章创新之处，一般不超过20个汉字，中英文题目应一致。题目应是所研究的主题以及具体内容的准确描述，尽量避免繁琐、含糊、抽象。2、题目采用修饰词+定语的结构）

作者

许 嘉，杨攀原，吕 品

（作者单位）

（单位从大到小，省市 邮编；两个单位之间用“；”隔开，例如：

广西科学院,国家非粮生物质能源工程技术研究中心，非粮生物质酶解国家重点实验室，广西生物炼制重点实验室，广西南宁　530007；广西科学生命科学与技术学院，广西南宁　530004）

摘要：（摘要既要有高度的信息浓缩性，又具有可读性，还要结构完整、篇幅简短及独立成篇，研究类论文摘要从目的、方法、结果、结论4个方面进行撰写，摘要中不应包括论文中没有出现的内容。）

关键词：(选择能准确反映文章主要内容，能引起读者兴趣，并为大家熟知的专业词汇来作为关键词，以使文章能被准确检索。一般为5-8个，中英文一一对应)

中图分类号：

【计量单位】

1、以《中华人民共和国法定计量单位》为准，溶液浓度不用M，N，ppm；bar(巴)改用Pa(帕)；能量、功、热均用J(焦)表示。单位符号采用标准化符号，如s(秒)、min(分)、h(小时)、d(天)、mol(摩)、m(米)、m2(平方米)、L(升)、kg(千克)等。

2、文中首次出现的生物名称要注明拉丁名（斜体）；统计学符号（如*P*，*t*，*n*等）和变量符号须用斜体。

3、除了%和℃外，其他计量单位与数量间空半字符，如：10 kg、15 m2、5 min、5%、25℃。

4、数值范围请统一用一字线，即“6.31~7.30”应为“6.31-7.30”，小数点前后没3位数空半字符，即“1000”应为“1 000”，“0.0001”应为“0.000 1”

**0引言**

（研究类论文引言从研究意义、前人研究进展、本研究切入点、拟解决的关键问题4个方面进行撰写，应以简短的篇幅介绍论文的写作背景和目的，以及相关领域内前人所做工作和研究的概况，说明本研究与前人工作的关系，目前的研究热点、存在的问题及作者工作的意义，引出本文的主题给读者以引导。引言也可以点明本文的理论依据实验基础和研究方法，简单阐述其研究内容；三言两语预示本研究结果、意义和前景，但不必展开讨论。）

随着互联网技术的不断发展，大规模的开放在线课程（MOOC）等在线教育平台逐渐兴起，为学生提供高质量的免费教育资源。一门优秀的在线课程拥有数万参与者，这给平台的教师们带来了巨大的挑战：他们需要对大量的作业进行批改。尽管对于那些客观作业，例如选择题、判断题来说，自动化批改很容易实现，但是对于那些没有标准答案的开放式作业，例如简答题、作文题来说，目前还没有统一、有效的自动评分技术。【】考虑到开放式作业对提高学生创新思维能力和语言组织能力至关重要【】，目前主流的MOOC平台，包括Coursera1、edX2和ICourse3，都采用同行互评的方法来帮助教师评估学生提交的大量开放式作业。

同行互评方法与众包类似，它要求所有提交作业的学生扮演同行评价者，根据教师提供的评分细则，以匿名的方式对所有提交的作业的一小部分进行评价。最后对每份作业的所有评价结果进行汇总，得到每份作业真实分数。其中，我们仅考虑基数估计的同行互评技术，即同行评价者必须给每份评价作业一个数值型的分数。

除了能够减少教师的工作量之外，同行互评还可以带来其他的教育价值，可以激发学生的学习兴趣【】，加强学生的课程参与【】，提高学生的责任感【】。

同行互评的最主要的研究问题之一是同行互评的分数汇总方式，即根据同行评价者的对作业评分集合得到尽可能接近该作业的真实分数的结果。目前，大多数的MOOC平台（例如Coursera和中国大学M00C）的汇总方式只是简单地计算作业所有评价的均值和中位数，作为该作业的最终分数。然而，同行评价者的评分质量乘次不齐，受到其可靠性、偏见等因素的影响【】，其中，偏见意味着同行评价者的评分有放大或缩小的习惯性倾向，而可靠性表示了同行评价者的评分与真实分数之间的随机偏差。因此用均值和中位数来作为最终评分并不准确【】。目前，一些研究者提出一系列的概率模型，将同行评价者的评分可靠性和偏见设置成了随机变量，通过变量之间的依赖关系来更加准确地估计每份作业的真实分数。然而现有的概率模型中，对同行评价者的可靠性建模只与该评价者作为答题者，在本次作业上的真实分数有关。虽然学生的作业分数在理想情况下可以反应一个学生的评价能力，但是他们并没有太强的依赖关系，例如，一个成绩很好的学生不喜欢批改其他人的作业，因此在互评中批改的并不仔细，导致其同行评分与真实分数仍存在较大的差距。因此，将作业得分作为可靠性的建模标准并不准确。而真正可以直观的衡量评价者可靠性的，是学生的评价能力。评价能力可以由互评分数与真实分数之间的绝对差体现，绝对差越小，评价能力越高，可靠性也就越高。

如图我们发现，评价者在同一次的评分表现结果是很接近的，如果他在其中一份作业中的评价表现很好，那么他大概率在其他几份作业中的评价表现也很好。根据实验结果，我们发现一个评价者的历史表现良好，那么他在历史上的表现也很好，如果表现不好，这个学生的表现也一直不好。对此，评价的可靠性能力是相对固定的。

鉴于此，本文利用学生的评价能力，提出了两个新的概率模型RPG6和RPG7。这两个模型在现有模型的基础上，将可靠性的建模中的真实分数替换成了学生的评价能力，以提高模型在推导真实分数的准确性。RPG6和RPG7的区别在于，RPG6假设评价者可靠性服从伽马分布，而RPG7假设评价者可靠性服从正态分布。我们还提出了一种方式来量化学生的评价能力，并且通过教师抽查的方式，对其迭代更新，提高准确性，最后通过真实数据集与传统的概率模型进行了对比，验证了该模型的有效性。

本文的主要贡献有：

**1 相关工作**

目前主流的主观题作业的自动判别技术主要有两种，基于自然语言处理的技术和基于同行互评的技术。

根据同行评价者反馈的内容进行分类，可以将主观题的同行互评的汇总估计技术分为序数估计技术和基数估计技术。

**1.1 序数估计技术**

序数估计技术要求每个同行评价者给出其评价作业的优劣排序，最后将所有同行评价者的部分作业排序进行汇总，得到所有作业的最终排名。序数估计的主要研究点在于将多个部分的排序序列整合，推断出整体序列。Shah等人在Bradley-Terry(BT)模型【】的基础上，提出了RBTL模型【】，将每个从学生那里收集的有序比较对，看做是两两比较的成对偏好分布，通过两两间的关系推导出最终的排名。而Raman【】等人将一些传统的排序聚合模型（包括BT【】、MAL【】、Thurstone【】、PL【】）应用到了序数估计模型中，并且引入了同行评价者的评分可靠性变量，有效提升了序数估计模型对作业真实分数估计的准确性。Nicola等人则基于模糊数学理论定义了一个有序同行评价模型FOPA【】，与只有“大于”、“小于”传统的序数估计不同，FOPA模型映入更多的关系符号，例如“远远大于”、“大于等于”、“约等于”等，以此提高序数估计的准确性。

序数估计技术将评审活动转化为了一项简单的比较任务，降低了互评的难度。但是，同行互评还存在一些的问题：对于一些水平相近的作业，序数估计在对其量化时会扩大它们的实际差距。此外，学生提交的比较对也可以通过直接打分的方式排序得到，这说明和直接打分相比，序数估计包含了更少的信息。

**1.2 基数估计技术**

基数估计技术要求每个同行评价者根据评分准则，给评价的作业一个数字评分。之后通过一些方式汇总每份作业获得的所有评分，得到最终分数。目前基数估计是主流MOOC平台采用的同行互评的评判方法。对比序数估计，基数估计更能体现作业之间的绝对分数差，可以更加准确的的量化作业之间的分数差距。基数估计的汇总方法主要有两种：加权聚合模型和生成概率模型。加权聚合模型主要通过计算每个同行评价者的准确性作为权重，然后将同行评价者的评分加权聚合，得到作业的真实分数。Fang等人提出了RankwithTA模型【】，通过教师半监督抽查的方式确定每个同行评价者的评分准确性。在该模型中，教师将抽查部分学生的作业打分，以教师的评分作为学生的权重，并通过这部分学生与其他学生间的评价交集得到所有学生的权重。而Wang认为将学生的作业成绩等价为评审能力是不全面的，于是在此基础上提出了SSPA模型【】。SSPA在构建同行评价者的准确性时，将同行评价者与教师评分的差距也作为其影响因素，结合学生的作业成绩来共同决定学生的准确性，进而加权聚合得到学生的真实分数。虽然加权聚合模型比起均值、中位数更加准确，但是它还是没有很好的量化同行评价者之间的个体差异。而生成概率模型则是将学生作业的真实分数、同行评价者的其他属性建模为含隐变量概率分布模型，然后通过可观测的评分信息推演隐变量的值。生成概率模型最早由Piech等人提出，他们一共提出了三种概率模型PG1，PG2，PG3。【】其中效果最好的PG3认为同行评价者的可靠性与其作业的真实分数之间存在联系，于是在两者之间建立了线性关系进行转化，在实验的数据中，PG3的预测分数的准确性在三个模型中是最好的。而Mi等人认为，线性关系不能准确的概括真实分数与可靠性的联系，所以他们在PG3的基础上，将同行评价者的可靠性建模为均值为真实分数的概率分布模型，提出了PG4和PG5【】，以此提高了预测最终分数的准确性。其中PG4假设可靠性的概率服从伽马分布，而PG5假设可靠性的概率服从高斯分布。然而，上诉模型都只依靠同行互评者打分这一个可观测的变量，使得预测模型存在数据稀疏的缺陷。因此，Wang在PG4和PG5上加入了相对分数等级，即同一评分者不同提交成绩的绝对同伴等级的差异，构成了PG6和PG7【】。两者在原有的模型基础上，加入了一个可观测变量，有效地解决了数据稀疏性带来的问题。然而，现有的概率生成模型在对学生的可靠性建模时，将学生作业的真实分数作为其主要的影响因素。然而学生的评价能力和作业成绩没有直接联系。例如一个具有高水平学术能力的学生在没有认真评价的情况下，依然会给出一个低质量的评价结果。所以，直接使用真实分数作为可靠性的建模标准会产生一定的误差。对此我们提出了一种学生的评价准确性的定义方法，将其作为可靠性的建模方法。评价的准确性可以直接体现在学生评分与真实分数的差距上，差距越小，说明学生评价的可靠性越高。并且，在一个周期之内，学生的可靠性不会发生太大的变化。

我们将学生的在互评结束之后的汇总分数看做学生的真实分数，通过差值来计算学生的可靠性。而且我们还使用了一种半监督的方式，让教师在互评时对部分学生打出真实分数，以此能够更好的对模型呈现更新性。最后，我们在真实的实验数据集中验证了我们的模型的有效性。

\*基金项目

**【作者简介】**

第一作者 (出生年—)，性别，职称，主要从事研究方向。（我刊不接受共同第一作者。）

【\*\*通信作者】

姓名 (出生年—)，性别，职称（需与第一作者同级或更高），主要从事研究方向，E-mail：

**2 同行互评概率模型**

在本文中，我们会首先介绍使用的符号，然后介绍我们自定义的互评者准确性的计算方法，最后介绍我们的同行互评概率模型。

在本文中，我们将使用以下符号。对于所有在线上平台上提交了开放式作业的学生的集合，我们使用U表示，ui表示U中的任意学生。我们用H表示所有提交的开放式作业的集合，hi表示ui提交的作业。我们通常用v表示评价者，通常情况下,所有提交作业的人学生都将作为同行评价者为作业打分，v\in U。接下来是对概率模型的重要概念的定义，这些概念是概率模型中可观测和未观测的潜在变量。表一总结了以下变量。

可靠性：可靠性表示评价者给出互评分数的准确度，用\tau\_v表示。具体反映的是在经历了偏置的校准后，其平均倾向于接近提交的真实分数。在研究中，评价者的可靠性将取决于我们定义的评分准确性。评分准确性越高，表示可靠性越高。

偏置：偏置表示同行互评者的客观存在的偏见，用b\_v表示。具体反应了一个同行互评者存在对于所有互评分数夸大和减低一定分数的倾向。

真实分数：真实分数表示一个提交的作业按照评分准则的真实打分，用s\_i表示,它是未观察到的，需要估计。

互评分数：互评分数表示同行评价者给作业打出的分数。其中z\_i^v表示评价者v为u\_i打出的分数。互评分数是可以被观察到的，其集合被表示为Z={ziv|u\_i \in U, v \in U}

相对分数：相对同伴分数表示同一个同行评价者的互评分数之间的差距，用d表示。d^v\_{ij}表示评价者v对hi和hj的互评分数之间的差距，即z\_i^v和z\_j^v之间的绝对差。

现有的同行概率模型中，对同行评价者的可靠性建模由学生的真实分数决定。然而作业分数与学生评价能力之间并没有绝对的关系，根据XXX的描述，同行评价者的可靠性可以由其互评分数与教师之间的差距决定。因此我们定义了一种方式，通过同行评价者的互评分数与教师评价的真实分数，来作为可靠性的建模元素。我们将这个新定义的变量称为同行评价者的准确性。下面讨论计算准确性的方法。

我们通过同行评价者的互评分数与真实分数之间的差距来计算学生的评价能力。其公式为：

tj表示学生的真实得分，zij表示学生i对j的评分，Ti表示评价者i评价的所有集合中有真实分数的集合。

为了定义学生的评价能力，我们需要提前知道学生的真实分数。因此我们可以通过半监督的方式，让教师评价部分学生的作业资源，获取该作业的真实分数，这些具有真实分数的作业被称为审查资源，在互评流程中，我们可以尽量均匀分布审查资源，使得让更多的学生可以和真实分数对比获得评价能力。在这个而那些没有被评价审查资源的学生，将会沿用上一次计算留下的评价能力，而他们评价能力将会在互评结束之后，通过概率模型汇总的最终分数作为真实分数，进行更新，以用于下一次的同行互评。所以我们模型的流程图如下。

在拥有了评价能力之后，我们就可以对概率模型进行建模了。基于以上的符号表征。我们研究的问题是通过建模互评分数、相对分数、可靠性、偏置、评价能力以及作业的真实分数之间的关系，来有效的估计作业的真实分数。换句话说，我们将问题定义如下：给定同行分数Z、相对分数D和评价能力R，我们会学习所有评价者v\in U，的\tau和b，以及所有的si。

接下来我们介绍模型RPG6和RPG7.

RPG6拓展于现有的PG6模型，其中结合了评分者的评价能力。其中RPG6中的随机变量之间的条件依赖结构如图1中的图形模型所示。如图所示，互评分数z、相对分数d和同伴的评价能力r是模型中可观测的随机变量。学生ui提交的真实得分si、评分者的可靠性\tau和偏置b是模型中需要估计的潜在变量。这些潜在变量的先验分布与超参数、、、制定的分布如图所示。由图可知，这些隐变量是彼此联系的。

为了确定这些隐变量的值，需要基于每个隐含变量的后验分布信息进行取样，我们采用Gibbs采样技术，对每个隐变量的取值进行采样。我们采用可观测的随机变量，包括互评分数，相对分数和学生的评价能力来估计隐变量。具体而言，对于每个隐变量的值，我们通过可观测变量和其他隐变量的值的后验分布进行计算。在计算若干次之后，我们将生成的若干组潜在变量的均值作为其最终值。对于一组潜在变量。我们进行了600次吉布斯采样，在计算均值时，我们去除了前60次的老化采样结果，使得最终的结果更加准确。

具体的算法流程如图。

算法步骤，算法流程图。

，RPG6模型的所有随机变量的分布如下所示：

在RPG6模型中，我们假设真实的分数si服从高斯分布，均值为\alpha，方差为1/\gamma。虽然评价者在同行互评中可能有不同的偏见，我们认为偏见的均值为0.并且假设b服从均值为0，方差为1/g的高斯分布。对于可靠性t，我们在RPG6中假设其服从伽马分布，其形状参数等于评价能力r，和率参数等于\beta。互评分数z，服从高斯分布，其均值等于si+bv，方差与评分者的可靠性成反比。相对分数的dvij，服从高斯分布，其均值等于si-sj，和方差等于2/tv

由于概率模型RPG6和RPG7中的隐含变量si没有闭式解（closeform 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。RPG6中的隐含变量的后验分布推断结果如下：

RPG7模型和RPG6的模型的区别在于RPG7模型假设同行评价者的可靠性满足伽马分布，而RPG7则假设同行评价者的可靠性满足高斯分布。RPG7模型拓展了现有的PG7模型，其所有随机变量的分布如下所示：

简介，不一定要

由于概率模型RPG6和RPG7中的隐含变量si没有闭式解（closeform solution），因而采用近似离散推断的策略，以间隔宽度为0.1来得到该隐含变量的近似后验分布。RPG7中的隐含变量的后验分布推断结果如下：

对此，我们需要得到学生的真实分数。由于我们需要在互评结果出现之前得到学生的

我们采用了半监督的方式，让教师对部分学生的作业进行打分，得到了部分分数。如果教师抽查的数量不够，就无法拥有更多的真实分数。所以我们进行了一个实验，确定了学生的评价能力不会有太大的变化。于是我们可以通过上一次计算的真实分数和互评分数对学生的评价能力进行定义。由于上一次的结尾，我们可以通过汇总分数得到，我们主要通过两种方式，第一是一种教师的打分，第二是通过模型推导得到的真实分数。首先对于一个学生，打分。因此，我们采用了半监督的方式，

**2.1**

2.1.1

2.1.1.1

……

**2.1**

2.2.1

2.2.1.1

……

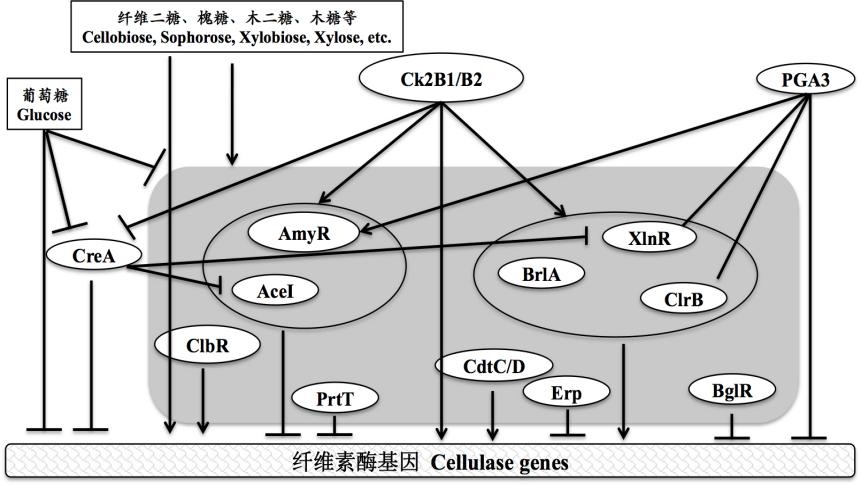
【图表】

1、只附最必要的，能用文字表述尽量不用图表，图表要求精心设计，力求美观，互不重复。给出图表，不能仅仅有图表而没有其他文字说明，且正文中必须遵循“先文后图表”的原则，即先有对图表的重要数据的描述文字，及由图表得出的结论，并在正文中明确出现“图1”或“表1”等文字，方可出现图表。

2、图题、标题、图注、标注，以及图表内所有的中文应补充对应的英文翻译。

3、图序、表序不分章节，完全按顺序统一编号，如图1、图2……图n，表1、表2……表n（不可用图1-1，表1-1等形式）

4、其中，图的编排顺序为图—图注—图题；表的编排顺序为标题—表—表注。图表中的单位用括号标注，如“（kg）”



(图片大小在80 mm左右，像素300-600)

图1 青霉属纤维素酶合成调控网络

Fig.1 Regulatory network of cellulase synthesis in *Penicillium*

表1 Ca-MMT含量对HDPE复合材料燃烧性能的影响

Table 1 Influence of Ca-MMT content on the combustion properties of composites HDPE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 试样  Sample | 极限氧指数LOI（％）  Limit oxygen index | 垂直燃烧等级  Vertical flame rating | 水平燃烧等级  Horizontal flame rating | 燃烧现象  Combustion phenomenon |
| A | 19.0 | 完全燃烧殆尽  Completely burned up | HB75 | 严重滴落  Severe dripping |
| B | 28.1 | V-2 | HB | 发烟量多，严重滴落  Large amount of smoke，severe dripping |
| C1 | 30.91 | V-0 | HB | 发烟量少，不滴落  Smoke less，no dripping |
| C2 | 32.51 | V-0 | HB | 发烟量少，不滴落  Smoke less，no dripping |
| C3 | 32.77 | V-0 | HB | 发烟量少，不滴落  Smoke less，no dripping |
| C4 | 32.28 | V-0 | HB | 发烟量少，不滴落  Smoke less，no dripping |

**3实验结果**

（主要是对文中研究结果进行解释和推断，评论所报道结果的重要性和创新性，并与其他相关研究进行比较，阐述作者的结果是否支持或反对某种观点，指出对原有结论有何新的启示，或从实验数据归纳出一般规律，要避免不作具体对比分析，就宣称本文结果“属于国内和先进水平，填补了国内的空白”等，必要时指出本试验的不足之处及未能解决的问题。要求围绕结果进行讨论，但不要重复结果，并与引言中所提出的问题前后呼应；叙述要简明扼要，推测要有依据，不能得出超出本文结果所能够支持的结论。不要在本节中介绍新的结果。）

**3.1**

**3.2**

**3.3**

……

**4结论**

（应该是最终的、总体的结论，而不应是正文中各段小结的简单重复。应准确、完整、明确、精炼。如果不可能导出应有的结论，也可以没有结论而进行必要的讨论。可以在结论或讨论中提出建议、研究设想、仪器设备改进意见、尚待解决的问题等。）

……

**参考文献：**

英文作者姓名，姓在前（全大写），名在后（只写第一个字母，且大写），例如“ZHANG Xiaohua”应写成“ZHANG X H”，“David William Russell” 应写成“RUSSELL D W”；英文题目只有第一个字母用大写，其余都用小写，专有名词除外，如China；英文出版社，出版地，期刊名，实词的第一个字母大写，虚词全部小写。

[1] 温远光. 桉树生态、社会问题与科学发展[M]. 北京:中国林业出版社，2008.

[2] 李国雷，刘勇，于海群，等. 油松人工林下植被发育对油松生长节律的响应[J].生态学报，2009，29(3)：1264-1275.

[3] WEN Y G，YE D，CHEN F，et al. The changes of understory plant diversity in continuous cropping system of Eucalyptus plantations, South China[J]. Journal of Forest Research, 2010,15(4):252-258.

[4] 上海材料研究所.GB/T 16421—1996 塑料 拉伸性能小试样试验方法[S]. 北京：中国标准出版社，1997.

[5] 曹青. 草酸青霉胞外β-葡萄糖苷酶多样性的研究[D]. 济南：山东大学，2014.

**Title**

（与中文一致，英文题目所有的实词第一个字母用大写，例如，“Study”；所有虚词用小写，例如“from”。）

Author

（姓名之间用逗号隔开，拼音格式姓全大写，名首字母大写，如“ZHANG Xiaohua，OUYANG Xiaohua”。）

Organization

（1、单位从小到大，市，省，邮编，国籍。 2、跟英文题目一样，英文格式所有的实词第一个字母用大写，所有虚词用小写。

例如：State Key Laboratory of Non-food Biomass and Enzyme Technology, National Engineering Research Center for Non-food Biorefinery, Guangxi Key Laboratory of Biorefinery, Guangxi Academy of Sciences, Nanning, Guangxi, 530007, China）

**Abstract:**

**Key words**：