arXiv:2302.02304v1 [stat.ME] 5 Feb 2023

利用潜在因子建模子群结构的众包因子建模

Qi Xu、Yubai Yuan、Junhui Wang 和 Annie Qu*

摘要

众包已成为收集大规模标签的替代方案。然而，招募的工作者大多不是领域专家，因此他们提供的标签可能存在噪声。本文提出一种两阶段模型，用于预测众包多类别分类任务的真实标签。在第一阶段，我们使用潜因子模型拟合观测标签，并通过多质心分组惩罚项对任务和工作者的子群结构进行建模。我们引入组特定旋转，使处理不同任务类别的工作者相互对齐，以解决多类别众包任务。在第二阶段，我们提出一种基于一致性的方法，用于识别高质量工人子群组，并依靠这些子群组为任务分配标签。从理论上，我们证明了潜在因子的估计一致性以及所提方法的预测一致性。模拟研究表明，在假设任务和工作者内部存在子群结构的情况下，所提方法优于现有竞争方法。我们还展示了所提方法在现实问题中的应用，并证明了其优越性。

关键词：双聚类；基于一致性的方法；标签聚合；低秩近似。

*Qi Xu 是加利福尼亚大学欧文分校统计系的博士生，地址为加利福尼亚州欧文市，邮编 92697（电子邮件：qxu6@uci.edu）。Yubai Yuan 是宾夕法尼亚州立大学统计系的助理教授，地址为宾夕法尼亚州大学公园市，邮编 16802（电子邮件：yvy5509@psu.edu）。Junhui Wang 是香港中文大学统计系的教授，地址为香港沙田（电子邮件：junhuiwang@cuhk.edu.hk）。Annie Qu 是加利福尼亚大学欧文分校统计系的校长讲席教授，地址为加利福尼亚州欧文市，邮编 92697（电子邮件：aq2@uci.edu）。

1 引言

近年来，人们对带有高质量标签的大规模数据的需求日益增长，而这对于训练高性能机器学习模型至关重要。然而，完全依靠专家为大规模数据标注，成本可能非常高，耗时也很长。作为一种广受欢迎的替代方案，众包旨在利用众包工作者完成任务时提供的大量众包标签，预测真实标签；例如，在图像分类任务中，众包工作者需要将任务归入不同类别，并指定详细标签。在过去二十年间，随着 MTurk 等在线平台的发展，众包已应用于许多需要大量标注的领域，例如计算机视觉 [8, 19, 32], 自然语言处理 [4, 24, 29],和医学诊断 [13, 20, 25]。众包面临的主要挑战是，由于任务难度级别各异，且工作者在特定任务上的经验不同，众包标签可能存在噪声。例如，高质量工作者很可能更准确地标注任务，而低质量工作者在标注任务时的表现则更像随机猜测。因此，要实现有效的众包，关键在于考虑任务和工作者之间的异质性，以提高标签预测准确率。

一种简单的众包方法是多数投票 [1, 2, 26], ，即将多数工作者投出的标签指定为预测标签。然而，当低质量标注者所占比例相对较大时，多数投票并非最优。为纳入工作者之间的异质性，近年来研究者开发了基于混淆矩阵的模型，例如通过 EM算法 [7, 43], 进行最大似然估计、贝叶斯分类器组合 [17,37], 以及成对共现分解 [16]。然而，这些方法忽略了任务之间的异质性，并假设同一类别中的任务具有相同的正确标注概率。为适应任务难度和特征的异质性，研究者提出了潜在因子模型，例如标签、能力和难度生成模型（ GLAD） [41] 以及多维智慧

[1]https://www.mturk.com

模型 [39]。也就是说，每个任务或工作者都由一个潜在因子表示，该因子编码其特征，例如任务难度和工作者能力 [41]。尽管现有的潜在因子模型能够同时适应任务和工作者之间的异质性，但它们仅适用于二分类任务，不适用于多类别任务。

众包中的另一项关键挑战是标签指定。即使某种方法能够正确地将任务聚类到各类别中，也仍可能无法为每个类别正确指定标签。例如，对于基于混淆矩阵的模型 [7, 43], ，即使随机打乱类别标签，似然函数也保持不变。因此，通常需要对类别标签设定作出假设。基于混淆矩阵的模型 [7, 43] 采用的一种常见假设是，多数投票的标注结果是正确的。另一种做法是，Ghosh 等人 [14] 假设已知有一位高质量工作者，其对类别的标注能力优于随机猜测。然而，这些假设相对严格；当低质量标注者普遍存在时，它们可能不成立，或者可能无法确定哪位工作者是高质量工作者。这促使我们开发一种能够恢复类别标签设定的新框架，使所提方法在不牺牲标签预测准确率的情况下，能够更稳健地应对低质量工作者。

本文提出了一种新颖的众包框架，统一了二元和多类别标签预测。具体而言，我们将任务和工作者嵌入同一潜在空间，从而能够将众包标签建模为由相应任务与工作者的潜在表示之间的一致性参数化的分类分布。所提方法的一项主要创新是利用潜在因子建模刻画任务和工作者之间的异质性，同时促使相似的任务和工作者聚集在一起。基于一致性的规则这一类似概念已被应用于基于混淆矩阵的方法 [37],，但由于基于混淆矩阵建模的特性，任务的子群结构被忽略了。为了同时捕捉任务和工作者的子群结构，我们引入了多质心分组惩罚项，该惩罚项能够自适应地捕捉任务和工作者之间的子群结构，并推动来自同一子群体的任务或工作者的潜在因子彼此靠近。为适应多类别标签，我们引入了一组组特定的

旋转矩阵，以将来自不同类别的工作者的潜在因子与任务的潜在因子对齐。为恢复类别标签设定，我们通过基于一致性的规则识别高质量工作者；该规则能够准确且稳健地分配标签。

与现有众包方法相比，所提方法有两大优势。首先，所提方法利用潜在空间中的子群结构，提高了标签预测准确率。子群划分惩罚项旨在捕捉内在的子群结构，同时区分潜在空间中的不同任务类别和工人组。除此之外，子群划分惩罚项还会对潜在因子施加约束，从而控制模型复杂度，减轻模型可能对含噪众包标签过拟合的问题。此外，组特定旋转矩阵能够捕捉同一组中工作者对不同任务类别的共同意见，并利用工作者的子群结构形成简约的模型结构。从统计效率的角度来看，子群划分惩罚项和组特定旋转矩阵通过利用同一子群的信息，提高了估计效率。

其次，即使低质量工作者占多数，或高质量工作者身份未知，所提出的标签指定仍然有效。这是因为，基于高质量工作者比其他工作者更有把握正确标注这一假设，所提方法能够同时识别高质量工人子群并恢复正确的类别标签设定。无论工人子群的规模大小，也无论个体工作者的子群归属如何，都可以通过潜在空间中任务与工人子群之间的一致性来刻画标注不确定性。因此，对于少数专家与大量低质量标注者混杂的情况，所提方法尤其有效。

为实现所提方法，我们提出一种可扩展算法，用于求解非凸优化问题。为调节模型中的超参数，我们提出一种BIC型准则，该准则基于所选高质量工人子群组提供的标签，

并假设这些标签是真实标签的替代物。模拟研究表明，所提方法优于现有竞争方法，尤其是在低质量标注者占所有工作者多数的情况下。从理论上看，我们证明了模型参数在海灵格距离下具有一致性，并进一步证明了标签预测的一致性。

本文其余部分安排如下。第2节介绍基于潜在因子的众包模型的背景，并说明全文使用的符号。第3节提出用于多类别众包的框架，并给出相应的标签指定程序。第4节阐述所提方法的可扩展优化算法及实现细节。第5节建立所提方法的理论性质。第6节在模拟设置下，对所提方法与现有方法进行数值比较。第7节在多个众包基准数据集上考察所提方法的表现。第8节对所提方法进行讨论，并给出总结性结论。

2 背景与符号

本节介绍基于潜在因子建模的众包方法的背景，并说明全文使用的符号。我们将$\mathbf{R} = \{r_{ij}\}_{m\times n}$ 定义为众包标签矩阵，其中 r_{ij} 表示第 i个任务由第 j位工作者给出的标签（$i = 1, ..., m$； $j = 1, ..., n$）。对于二元众包，我们假设 $r_{ij} \in \{0, 1\}$；对于包含 C个类别的多类别众包，我们假设$r_{ij} \in \{0, 1, ..., C-1\}$。实际中，工作者通常不会标注所有任务，因此$\mathbf{R}$ 中只有部分数据可观测，我们将该子集记为$\Omega = \{(i, j)$： r_{ij}可观测$\}$。我们将$\mathbf{Z} = (Z_1, Z_2, ..., Z_m)'$记为任务的真实标签；若 $Z_i = c$，则称第 i个任务为标签-c 任务。众包的目标是通过汇总观测到的 r_{ij}来预测 Z_i。

（GLAD） [41], 是一种基于任务和标注者特征的逻辑模型：

$$\log\left\{\frac{P(r_{ij}=Z_i)}{1-P(r_{ij}=Z_i)}\right\}=\mu_i\nu_j, \tag{1}$$

其中，$\mu_i \in (0, \infty)$ 是衡量第 i 个任务难度级别的参数，μ_i 越大表示第 i 个任务越简单；$\nu_j \in (-\infty, \infty)$ 是衡量第 j 位工作者能力的参数，其值越高，标注任务的能力越强。如果 $\mu_i\nu_j \to \infty$，概率 $P(r_{ij}=Z_i)$ 将趋近于 1；也就是说，对于简单任务，高质量工作者正确标注的概率很高。GLAD 模型通过引入两个潜变量，同时纳入任务与标注者异质性，但它仅适用于二元众包，而且一维近似在实际应用中受到限制。

3 方法

本节介绍我们为多类别众包提出的两阶段模型。第一阶段使用潜因子模型拟合观测标签，第二阶段识别高质量工作者并进行预测。在我们的框架下，二元众包问题是一个特例。关于二元众包的详细讨论见补充材料。

3.1 模型框架

对于一个 C–类别众包问题，我们假设 r_{ij} 服从类别分布：

$$r_{ij} \sim categorical(p_{ij}^{(0)}, p_{ij}^{(1)}, ..., p_{ij}^{(C-1)}), \quad \sum_{c=0}^{C-1} p_{ij}^{(c)} = 1, \quad (i,j) \in \Omega, \tag{2}$$

其中，$p_{ij}^{(c)}$ 是 r_{ij} 等于 c 的概率。这里，我们对观测标签的概率进行建模，而不是像式 (1) 中那样对正确标注 $\mathbb{1}(r_{ij} = Z_i)$ 的概率进行建模。我们的方法在多类别标签建模方面更加灵活。

为刻画任务和工作者内部的异质性，我们引入一个全参数化模型：

$$\log\left\{\frac{p_{ij}^{(c)}}{p_{ij}^{(0)}}\right\} = \mathbf{a}_i'\mathbf{b}_{j,c} - \mathbf{a}_i'\mathbf{b}_{j,0}, \quad c = 1, ..., C-1, \quad (i,j) \in \Omega, \tag{3}$$

其中，$\mathbf{a}_i \in \mathbb{R}^k$ 是第 i个任务的潜在因子，表示该任务的潜在特征。在图像分类问题的示例中，$\mathbf{a}_i$可能编码对象和纹理信息。潜在因子$\mathbf{b}_{j,c} \in \mathbb{R}^k$ 表示第 j个工作者对标签-c 任务的看法。内积$\mathbf{a}_i'\mathbf{b}_{j,c}$ 根据第 j个工作者的看法，衡量第 i个任务与标签-c 任务之间的一致性。与 GLAD (1) 不同，真实标签 Z_i 隐式地嵌入在 $\mathbf{a}_i$ 中，而且同一类别中的任务可能具有相似的潜在特征。因此，任务的潜在因子预期会形成聚类，我们用$\mathbf{U} = \{U_i\}_{i=1}^m$(
$U_i \in \{0, 1, ..., C-1\}$)表示任务成员关系。请注意，任务成员归属$\mathbf{U}$ 是对真实标签$\mathbf{Z}$的一个排列，而$\mathbf{U}$ 与$\mathbf{Z}$ 之间的对应关系将在第二阶段恢复。

全参数化模型（3）类似于多项逻辑回归 [10]，其中每个类别 c 对应一组类别特定参数 $\mathbf{b}_{j,\ c}$。然而，将全参数化模型（3）应用于众包问题时，它存在若干缺点。首先，该模型无法捕捉不同工作者之间的相关性，而这种相关性在现实世界的应用中十分常见 [5, 21]；例如，具有相同教育背景的工作者可能会对特定任务持有相似的意见。其次，模型（3）存在参数过多的问题，其中每个工作者都与 C 潜在因子相关联，导致工作者潜在因子的估计效率低下，因为该模型只能利用相应任务类别的信息。

为解决上述问题，我们提出了一种新方法，利用子群

结构来刻画任务和工作者。假设我们有 D 个工人组，工作者的群组成员关系表示为 $\mathbf{V} = \{V_j\}_{j=1}^{n}$，其中 $V_j \in \{1, 2, ..., D\}$。来自同一子群体的工作者对任务往往持有相似的意见，因此应考虑工作者意见之间的组内相关性。对于工作者子群组 $d \in \{1, 2, ..., D\}$，我们引入一组旋转矩阵

$$\mathbf{O}^{(d)} = \{\mathbf{O}_c^{(d)} \mid \mathbf{O}_c^{(d)'}\mathbf{O}_c^{(d)} = \mathbf{I}_{k\times k}, \mathbf{O}_c^{(d)} \in \mathbb{R}^{k\times k}, c = 0, 1, ..., C-1\}, \tag{4}$$

以使第 d个工人子组的潜在因子与各任务类别的潜在因子对齐。具体而言，对于属于第 d 个子群体的第 j位工作者，其对标签-c 任务的意见通过旋转矩阵$\mathbf{O}_c^{(d)}$对其关于参考类别的意见进行旋转来建模（本文默认将标签0任务视为参考任务）：

$$\mathbf{b}_{j,c} = \mathbf{O}_c^{(d)}\mathbf{b}_j, \quad c = 0, 1, ..., C-1, \quad d = V_j, \tag{5}$$

其中，为了实现可识别性，我们指定$\mathbf{O}_0^{(d)} = \mathbf{I}_{k\times k}$，使得$\mathbf{b}^j = \mathbf{b}_{j,0}$ 表示第 j位工作者对参考组的意见。我们将所有旋转矩阵记为$\mathbf{O} = \{\mathbf{O}^{(1)}, \mathbf{O}^{(2)}, ..., \mathbf{O}^{(D)}\}$。基于式（3）中对$\mathbf{b}_{j,c}$所采用的建模方式，多类别群体标签服从逻辑模型:

$$\theta_{ijc} = \log\left\{\frac{P(r_{ij} = c)}{P(r_{ij} = 0)}\right\} = \mathbf{a}_i'\mathbf{O}_c^{(d)}\mathbf{b}_j - \mathbf{a}_i'\mathbf{b}_j, \quad c = 1, ..., C-1, \quad d = V_j, \quad (i, j) \in \Omega. \tag{6}$$

为便于表示，我们将$\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m)'$记为任务潜在因子的集合。类似地，我们将$\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_n)'$记为工作者潜在因子的集合。图1展示了所提模型（6）：旋转矩阵$\mathbf{O}_1^{(1)}$ 和 $\mathbf{O}_2^{(1)}$ 分别将第一子群体工作者对参考任务的意见与标签1和标签2任务的潜在因子对齐。

与全参数化模型（3）相比，所提方法具有以下优势。首先，所提模型能够通过工人子群组共享的$\mathbf{O}_d^{(c)}$ 捕捉同一子群体中工作者之间潜在的相关性。纳入这些依赖关系后，我们便能利用同一子

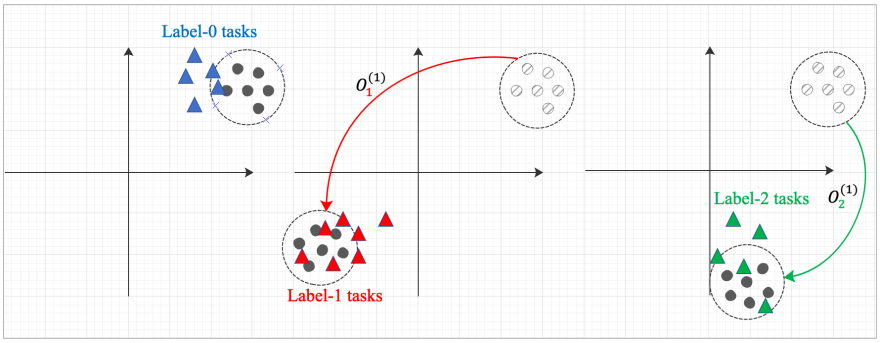


图1：所提出的多类别模型（6）的示意图。三个面板中的三角形分别表示标签0任务、标签1任务和标签2任务的潜在因子。左侧面板中的圆点表示一组工作者的潜在因子，代表他们对标签0任务的意见。中间和右侧面板中的旋转圆点分别表示工作者对标签1任务和标签2任务的意见。

群体中工作者观测到的众包标签信息，从而提高估计准确率和效率。其次，所提方法能够发现并增强工作者之间的子群结构，这对众包而言是一项重要优势。具体来说，嵌入式子群体划分程序能够帮助我们识别每个任务类别中的高质量工人子群，并进一步仅基于高质量工人子群提高标签预测准确率。第三，所提方法比全参数化模型更加简约：所提方法需要 $(m+n) \times k + (C-1) \times k^2$ 个参数，而全参数化模型需要 $(m+nC) \times k$。当 $n \gg k$ 时，这一优势尤为显著，而这在实践中很常见。

然而，拟合模型 (6) 并不能直接预测真实标签 $\mathbf{Z}$，而是恢复任务的子群成员关系 $\mathbf{U}$。因此，在第二阶段，所提方法旨在恢复任务成员关系 $\mathbf{U}$ 与真实标签 $\mathbf{Z}$之间的正确对应关系。对于一个 C 类众包问题，现有研究 [7, 43] 假设大多数工作者都能以较高概率正确标注所有类别的任务，这一假设在实践中相当严格。相反，我们放宽了这一假设，只要求一部分工作者能够以较高准确率区分某些类别。因此，我们提出为每个任务类别识别高质量工人子群组，并根据高质量工作者提供的标签恢复对应关系。

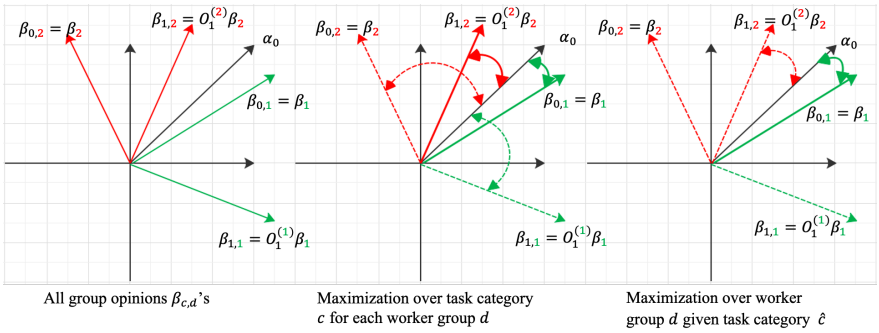


图2：识别高质量工作者并为 α_0 分配标签的示例。给定两个工人组，潜在因子质心 β_1，$\mathbf{O}_1^{(1)}\beta_1$、 β_2，$\mathbf{O}_1^{(2)}\beta_2$ 分别表示两个工人组对两个任务类别的意见。根据假设1和高质量工作者识别准则，质心 β_1 是 α_0 的高质量工人组，而 α_0 类别中的任务被指定为标签0任务。

对于许多特定领域应用 [25, 13], ，与其他工作者相比，高质量工作者在提供正确标签时往往更有把握。在所提出的潜因子模型 (6) 中，工作者的标注确定性可量化为该工作者的潜在因子与任务潜在因子之间的一致性。例如，在潜在空间中靠近某项特定任务质心的工作者子群组，其标注确定性高于其他工作者子群组。下文中，我们将 $\boldsymbol{\alpha}_I$，$I = 0, 1, \ldots\ldots, C-1$ 表示为任务的潜在因子质心，将 $\boldsymbol{\beta}_{I,J}$，$I = 0, 1, \ldots\ldots, C-1$，$J = 1, 2, \ldots\ldots, D$ 表示为工作者 J 在各任务类别 I 下所属工作者子群组的潜在因子。图 2 展示了识别高质量工作者子群组，并为质心为 $\boldsymbol{\alpha}_0$ 的任务类别分配标签这一核心概念。对于第一个工作者子群组，质心 $\boldsymbol{\beta}_{0,1}$ 与 $\boldsymbol{\alpha}_0$ 的一致性最大，这表明第一个工作者子群组会将 $\boldsymbol{\alpha}_0$ 类任务标注为标签 0 任务。对于第二个工作者子群组，对应的质心 $\boldsymbol{\beta}_{1,2}$ 与 $\boldsymbol{\alpha}_0$ 的一致性最大，但该一致性小于 $\boldsymbol{\beta}_{0,1}$ 与 $\boldsymbol{\alpha}_0$ 之间的一致性。基于上述思路，我们依据以下关于高质量工人组的假设，为每个任务子群体指定唯一标签：

假设1. 对于每个任务类别，如果某个工作者子群体的成员在该类别任务上的一致性高于其他组，则该子群体就是该任务类别的高质量工人组。该任务类别的高质量工人组所指定的标签

就是真实标签。

根据假设1，$\boldsymbol{\beta}_{0,1}$ 是 $\boldsymbol{\alpha}_0$的高质量工人子群的质心，因此，与$\boldsymbol{\beta}_{0,1}$相关联的工人组将 $\boldsymbol{\alpha}_0$ 指定为标签0任务。一般而言，我们可以根据假设1所依据的以下准则，识别高质量工人子群，并为第 i个任务指定标签：

$$(\hat{c}_i, \hat{d}_i) = \underset{\substack{c\in\{0,1,\ldots,C-1\},\\ d\in\{1,2,\ldots,D\}}}{\arg\max} \boldsymbol{\alpha}'_{U_i}\boldsymbol{\beta}_{c,d}, \tag{7}$$

ˆ其中，d_i 是第 i个任务的高质量工人组，ˆc_i 是ˆ高质量工人组 d_i指定的标签。式（7）中的最大化可以通过剖面法完成。对于每个固定的工人组 $d = 1$，2，...，D，对 c 的最大化对应于将第d组工作者的意见与 α_{U_i}对齐。对 d 的最大化对应于识别第 i个任务的高质量工人组。

所提方法依赖于识别高质量工人及其类别标签设定。与基于混淆矩阵的方法 [7, 43] 采用基于多数投票的类别标签设定相比，所提方法对低质量工作者所占比例的变化更具稳健性。此外，与 Ghosh 等人 [14]要求已知高质量工作者不同，所提方法是一种数据驱动方法，因为它直接从数据中识别高质量工作者。

3.2 参数估计

在本小节中，我们介绍如何通过标注矩阵 $\mathbf{R}$ 获取模型参数的估计值。此外，我们引入一种新颖的分组惩罚项，以探寻潜在空间中的子群结构。

基于对观测标签的分类分布所作的假设，负对数

观测标签的似然为:

$$\begin{aligned}\mathcal{L}(\mathbf{A},\mathbf{B},\mathbf{O}) &= -\sum_{(i,j)\in\Omega}\sum_{c=0}^{C-1}\mathbb{1}(r_{ij}=c)\log\tilde{r}_{ij,c}\\ &= -\sum_{(i,j)\in\Omega}\sum_{c=0}^{C-1}\mathbb{1}(r_{ij}=c)\log\frac{\exp(\mathbf{a}_i'\mathbf{O}_c^d\mathbf{b}_j)}{\exp(\mathbf{a}_i'\mathbf{b}_j)+\sum_{c'=1}^{C-1}\exp(\mathbf{a}_i'\mathbf{O}_{c'}^d\mathbf{b}_j)}.\end{aligned}$$

此外，我们采用多质心分组惩罚项，以探索任务和工作者的子群结构。具体而言，我们将多质心分组惩罚项定义如下:

$$\mathcal{G}_\lambda(\mathbf{A},\mathbf{B},\mathbf{U},\mathbf{V}) = \lambda\Big\{\|\mathbf{A}-\mathcal{U}\mathbf{A}\|_F^2+\|\mathbf{B}-\mathcal{V}\mathbf{B}\|_F^2\Big\}, \tag{8}$$

其中， λ 是用于惩罚的调节参数，矩阵 $\mathcal{U}$ 和 $\mathcal{V}$ 是与群组成员关系$\mathbf{U}$ 和$\mathbf{V}$ 相关联的投影矩阵，通过一一映射 [15] 分别计算任务和工作者的子群质心。具体而言，投影矩阵 $\mathcal{U}$和 $\mathcal{V}$ 将每个个体的潜在因子$\mathbf{a}_i$ 和$\mathbf{b}^j$ 投影到其对应的质心上，即$\mathcal{U}\mathbf{a}_i = \boldsymbol{\alpha}_{U_i}$和 $\mathcal{V}\mathbf{b}^j = \boldsymbol{\beta}_{0,\ V_j}$。因此，多质心分组惩罚等价于

$$\mathcal{G}_\lambda(\mathbf{A},\mathbf{B},\mathbf{U},\mathbf{V}) = \lambda\Big\{\sum_{I=0}^{C-1}\sum_{\{i:U_i=I\}}\|\mathbf{a}_i-\boldsymbol{\alpha}_I\|^2+\sum_{J=1}^{D}\sum_{\{j:V_j=J\}}\|\mathbf{b}_j-\boldsymbol{\beta}_{0,J}\|^2\Big\}.$$

此类惩罚项最初由 [34] 在纵向研究场景中提出，并已应用于有向网络社区检测 [42]。此外，其他惩罚函数，如融合套索 [35] 和OSCAR [3], ，也可用于潜在因子的子群分组；然而，这些方法可能面临计算负担和估计偏差。因此，我们在所提方法中选择使用多质心分组惩罚。

多质心分组惩罚具有以下优点。首先，它在潜在空间中寻找子群结构，并降低估计偏差和计算成本。传统的子群结构寻求惩罚项，例如融合套索 [35]和 OSCAR[3], ，由于需要进行成对比较而面临估计偏差和计算成本问题。在

相比之下，多质心分组惩罚将计算成本从 $O(m^2 + n^2)$ 降至 $O(m + n)$。同时，多质心分组惩罚不会将不同子群体中的潜在因子也拉到一起，而是将彼此不同的潜在因子分开，以实现分组。此外，它通过引入多个收缩方向来降低估计偏差。与一般的 L_2 惩罚项相比，多质心分组惩罚以数据自适应的方式确定质心，使所提方法比将质心固定在某个点的标准方法更灵活地利用观测数据。

通过引入多质心分组惩罚，我们可以通过最小化带惩罚项的负对数似然，联合估计潜在因子$\mathbf{A}$、$\mathbf{B}$ 以及群组成员关系$\mathbf{U}$、$\mathbf{V}$：

$$(\hat{\mathbf{A}}, \hat{\mathbf{B}}, \hat{\mathbf{O}}, \hat{\mathbf{U}}, \hat{\mathbf{V}}) = \underset{\mathbf{A},\mathbf{B},\mathbf{O},\mathbf{U},\mathbf{V}}{\arg\min}\ \mathcal{L}(\mathbf{A}, \mathbf{B}, \mathbf{O}) + \mathcal{G}_\lambda(\mathbf{A}, \mathbf{B}, \mathbf{U}, \mathbf{V}). \tag{9}$$

在估计模型参数后，根据假设1，我们可以在第二阶段针对每个任务类别识别高质量工人子群，并通过式（7）预测真实标签。最小化（9）的详细算法及其实现见第4节。

4 算法与实现

本节针对多类别情况，介绍通过交替最小化损失函数（9）来估计潜在因子、旋转矩阵和群组成员关系的可扩展算法。此外，当真实标签信息不可用时，我们还提供了一种用于调节超参数的新准则。

我们首先介绍求解（9）的算法。由于旋转矩阵具有非凸性和正交约束，（9）的优化颇具挑战。接下来，我们提出一种交替最小化算法，以最小化（9）中的联合损失函数。具体而言，在第（$t + 1$）次迭代中，我们依次更新（$\hat{}\mathbf{A}$，$\hat{}\mathbf{B}$，$\hat{}\mathbf{O}$）

给定估计值（$\hat{\mathcal{U}}^{(t)}$，$\hat{\mathcal{V}}^{(t)}$），如下所示：

$$\begin{aligned}
\hat{\mathbf{A}}^{(t+1)} &= \arg\min_{\mathbf{A}} \mathcal{L}(\mathbf{A}, \hat{\mathbf{B}}^{(t)}, \hat{\mathbf{O}}^{(t)}) + \lambda \|\mathbf{A} - \hat{\mathcal{U}}^{(t)}\mathbf{A}\|_F^2, \\
\hat{\mathbf{B}}^{(t+1)} &= \arg\min_{\mathbf{B}} \mathcal{L}(\hat{\mathbf{A}}^{(t+1)}, \mathbf{B}, \hat{\mathbf{O}}^{(t)}) + \lambda \|\mathbf{B} - \hat{\mathcal{V}}^{(t)}\mathbf{B}\|_F^2, \\
\hat{\mathbf{O}}^{(t+1)} &= \arg\min_{\mathbf{O}} \mathcal{L}(\hat{\mathbf{A}}^{(t+1)}, \hat{\mathbf{B}}^{(t+1)}, \mathbf{O}),
\end{aligned} \tag{10}$$

其中，关于 O 的最小化通过执行 Cayley变换 [40]来实现，从而保证迭代过程中满足正交约束。具体而言，我们引入 G $= \{\mathrm{G}_0^{(1)}, \mathrm{G}_1^{(1)}, ..., \mathrm{G}_{C-1}^{(1)}, ..., \mathrm{G}_{C-1}^{(D)}\}$ 作为 L 关于 O 的梯度，其中 $\mathrm{G}_c^{(d)} = {}^{\partial}\mathrm{L}\,{}_{\partial}\mathrm{O}_c^{(d)}$。接着定义一组矩阵 S $= \{\mathrm{S}_0^{(1)}, \mathrm{S}_1^{(1)}, ..., \mathrm{S}_{C-1}^{(1)}, ...,\mathrm{S}_{C-1}^{(D)}\}$，其中 $\mathrm{S}_c^{(d)} = \mathrm{G}_c^{(d)}\mathrm{O}_c^{(d)\prime} - \mathrm{O}_c^{(d)}\mathrm{G}_c^{(d)\prime}$。随后，按照 Cayley变换[40]更新 $\mathrm{O}_c^{(d)}$:

——

$$\mathbf{O}_c^{(d)} \leftarrow (\mathbf{I} + \frac{\eta}{2}\mathbf{S}_c^{(d)})^{-1}(\mathbf{I} - \frac{\eta}{2}\mathbf{S}_c^{(d)})\mathbf{O}_c^{(d)}, \quad d = 1, ..., D, \quad c = 1, ..., C-1, \tag{11}$$

其中，η 为正学习率。我们根据式（11）迭代更新$\mathbf{O}_c^{(d)}$，直至算法收敛。值得注意的是，对于 η 的任意正值，$\mathbf{O}_c^{(d)}$在每次迭代中都保持正交；我们根据实证将其设为 $\eta = 0.1$，这在保持较低训练误差的同时实现了快速收敛。此外，潜在因子$\mathbf{a}_i$和$\mathbf{b}_j$的更新可以并行进行，以加快计算速度。进一步地，投影矩阵 $\mathcal{U}$ 和 $\mathcal{V}$ 的更新分别等价于将任务和工作者潜在因子聚类为 C 个和 D 个子组。因此，我们使用 K 均值算法更新 $\mathcal{U}$ 和 $\mathcal{V}$。所提出的多类别众包算法概述于算法1中。

算法1 交替最小化算法

初始化：初始化 $\mathbf{A}^{(0)}$、$\mathbf{B}^{(0)}$、$\mathbf{O}^{(0)}$，$\mathcal{U}^{(0)}$，$\mathcal{V}^{(0)}$；设置 λ、k、D、$t=0$、学习率 η和停止容差 ϵ。

步骤 1. 在第（$t+1$）次迭代中，给定第 t 次迭代得到的 $\hat{}\mathbf{O}^{(t)}$、$\hat{}\mathcal{U}^{(t)}$、$\hat{}\mathcal{V}^{(t)}$，通过式（10）更新潜在因子（$\hat{}\mathbf{A}^{(t+1)}$、$\hat{}\mathbf{B}^{t+1}$）。

步骤 2. 在第（$t+1$）次迭代中，给定步骤 1 得到的 $\hat{}\mathbf{A}^{(t+1)}$、$\hat{}\mathbf{B}^{(t+1)}$，通过式（11）更新正交矩阵 $\hat{}\mathbf{O}^{(t+1)}$。

步骤 3. 在第（$t+1$）次迭代中，给定步骤 1 和步骤 2 得到的 $\hat{}\mathbf{A}^{(t+1)}$、$\hat{}\mathbf{B}^{(t+1)}$、$\hat{}\mathbf{O}^{(t+1)}$，通过 K 均值算法更新投影矩阵（$\hat{}\mathcal{U}^{(t+1)}$、$\hat{}\mathcal{V}^{(t+1)}$）。

步骤4. 重复步骤1-3，直到 $|\{\mathcal{L}(\hat{}\mathbf{A}^{(t+1)}, \hat{}\mathbf{B}^{(t+1)}, \hat{}\mathbf{O}^{(t+1)}) + \mathcal{G}_\lambda(\hat{}\mathbf{A}^{(t+1)}, \hat{}\mathbf{B}^{(t+1)}, \hat{}\mathcal{U}^{(t+1)}, \hat{}\mathcal{V}^{(t+1)})\} - \{\mathcal{L}(\hat{}\mathbf{A}^{(t)}, \hat{}\mathbf{B}^{(t)}, \hat{}\mathbf{O}^{(t)}) + \mathcal{G}_\lambda(\hat{}\mathbf{A}^{(t)}, \hat{}\mathbf{B}^{(t)}, \hat{}\mathcal{U}^{(t)}, \hat{}\mathcal{V}^{(t)})\}| \leq \epsilon$.

由于式(9)中的损失函数是非凸的，算法1在不同初始化下会收敛到不同结果。为了得到较好的局部极小值并加快收敛速度，我们研究了若干初始化方案。具体而言，我们提出一种信息丰富的方案，使用 Dawid Skene（DS）估计量初始化任务成员归属$\mathbf{U}^{(0)}$，并根据基于 DS 估计的准确率评估结果初始化工作者群组归属$\mathbf{V}^{(0)}$。此外，根据$\mathbf{U}^{(0)}$ 和$\mathbf{V}^{(0)}$给出的群组成员关系，使用多个具有不同均值向量的高斯分量初始化潜在因子。最后，将初始旋转矩阵$\mathbf{O}$ 设为单位矩阵。我们还在模拟研究中将这一信息丰富的方案与随机初始化进行了比较，结果表明，该信息丰富的初始化所需迭代次数更少，且预测准确率更高。关于初始化方案及相应实证研究的更多详细讨论见第6.3节。

我们算法中的一个技术问题是，工人子群组的数量 D 未知，需要预先指定；这对所提出的框架至关重要，因为它与高质量工作者子群组识别密切相关。当指定的 D较小时，高质量工人子群组的表现往往会受到影响，因为一些低

质量工作者可能会被混入高质量工人子群组。另一方面，当指定的 D较大时，来自同一子群体的工作者可能会被划分到多个子群组中，从而削弱高质量工作者的效能。

在实践中，工人子群组的数量可以根据先验信息确定，例如工人的专业水平 [9]。如果没有先验信息，我们可以采用数据驱动方法选择子群体数量，例如跳跃统计量 [33],、间隙统计量 [36], 和基于信息准则的方法 [22]。本文采用修正贝叶斯信息准则 [38]来调整工人子群组的数量 D,

$$BIC(D) = \log(\frac{\hat{\mathcal{L}}}{|\Omega|}) + (D + k - 1)\frac{\log|\Omega|}{|\Omega|}, \tag{12}$$

^ ^ ^ ^ 其中， |Ω| 表示 Ω 的基数，而 $\mathcal{L} = \mathcal{L}$(**A**、**B**、**O**) 表示给定观测标注矩阵 **R** 时的对数似然。此外，我们还在第 6.2 节讨论了所提方法在工人子群组数量不同时的稳健性，结果表明，工人组数量相对较少更为可取。

众包面临的另一项挑战是真实标签信息缺失，因此传统的交叉验证 [30] 并不适用。下文提出一种新颖的超参数调整准则，无需任何真实标签信息。该准则利用假设1，即某些工作者提供的标签比其他工作者更准确。因此，我们可以根据式 (7) 识别高质量工人子群。由此，我们提出基于这些已识别的高质量工作者进行标签拟合，以选择式 (9) 和式 (6) 中的超参数 λ 和 k。具体而言，假设高质量工作者给出的标签索引集合为 $\Omega_h = \{(i, j) \in \Omega | d_i = V_j\}$。则该准则定义为

$$Q_h(\lambda, k) = \frac{\sum_{i,j\in\Omega_h} \mathbb{1}(r_{ij} = \hat{r}_{ij}(\lambda, k))}{|\Omega_h|}, \tag{13}$$

其中，$\hat{r}_{ij}(\lambda, k)$ 表示 在给定惩罚系数 λ的情况下，通过最小化 (9) 得到的拟合标签

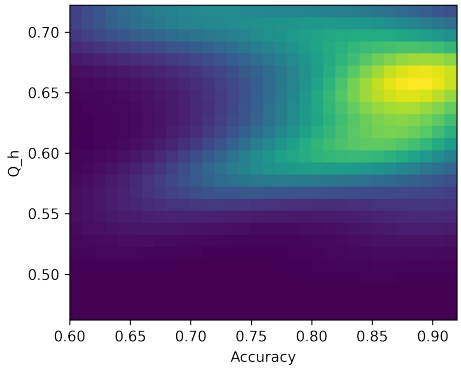


图 3：基于真实标签的 Q_h 与准确率的核密度图。该实验基于 Bluebird 数据[39]，其中 λ 和 k 有 500 种不同的组合，且 D 固定。

以及潜在因子的维度 k。准则 Q_h 用于衡量估计标签与高质量工作者给出的标签之间的一致性，并可用于近似基于真实标签的准确率，因为高质量工作者提供的众包标签更准确，可以很好地替代真实标签。图 3 显示， Q_h 与基于真实标签的准确率高度相关，这表明在缺少真实标签信息时，它可以作为有效的超参数调优准则。

为应用式 (13) 中提出的准则，在调整超参数 λ 和 k之前，需要识别出高质量工作者。我们首先通过修正贝叶斯信息准则 (12) 选择合适数量的工人子组。实证研究表明，在选定合适数量的工人组后，工人组成员关系对 λ 和 k并不敏感，也就是说，高质量参考标签 Ω_h 相当稳定，不过我们仍可以通过网格搜索选出 λ 和 k的最优超参数组合。

5 理论

本节中，我们为所提方法估计潜在因子和群组成员关系提供理论保证。具体而言，我们证明，与施加多质心分组

惩罚项相比，所提方法能够实现更快的估计收敛速率。我们还基于对高质量工作者的假设，证明了所提方法在恢复真实标签方面的一致性。此处我们仅提供多类别众包的理论结果，二元众包的结果可作为一种特殊情况推导得出。

考虑具有以下真实参数 $\boldsymbol{\theta}^*$ 的潜因子模型 (6)

$$\theta^*_{ijc} = \log\left\{\frac{p^{(c)}_{ij}}{p^{(0)}_{ij}}\right\} = \mathbf{a}_i^{*\prime}\mathbf{O}_c^{(d)^*}\mathbf{b}_j^*, \quad d = V_j^*, \tag{14}$$

其中 $\mathbf{a}_i^* \in \mathbb{R}^k$，$\mathbf{b}_j^* \in \mathbb{R}^k$, $\mathbf{O}_c^{(d)^*} \in \mathbb{R}^{k\times k}$，且每个 $\mathbf{O}_c^{(d)^*}$ 都是正交矩阵。任务和工作者的真实群组成员关系分别表示为 $\mathbf{U}^*$ 和 $\mathbf{V}^*$，其对应的投影矩阵分别表示为 $\mathcal{U}^*$ 和 $\mathcal{V}^*$。

令 $\mathcal{PL}(\mathbf{A}, \mathbf{B}, \mathbf{O})$ 为惩罚似然函数：

$$\mathcal{PL}(\mathbf{A}, \mathbf{B}, \mathbf{O}) = \mathcal{L}(\mathbf{A}, \mathbf{B}, \mathbf{O}) + \lambda_{|\Omega|}\mathcal{G}(\mathbf{A}, \mathbf{B}),$$

其中 $\lambda_{|\Omega|}$ 是惩罚参数，$|\Omega|$ 是 Ω 的基数，$\mathcal{G}$ 是潜在因子 A 和 B 的惩罚函数。在后续分析中，我们将任务潜在因子记为 $\boldsymbol{\gamma}_t = (\mathrm{a}_1, \mathrm{a}2, ..., \mathrm{a}_m)$，将工作者潜在因子记为 $\boldsymbol{\gamma}_w = (\mathrm{O}^{(V_1)}\ 0\ \mathrm{b}_1, \mathrm{O}^{(V_1)}\ 1\ \mathrm{b}1, ..., \mathrm{O}^{(V_1)}\ C-1\mathrm{b}1, ..., \mathrm{O}^{(V_n)}\ C-1\mathrm{b}_n)$。我们考虑两种惩罚函数及其对应的渐近性质：$\mathcal{G}_1(\mathrm{A}, \mathrm{B}) = \|\mathrm{A}\|_F^2 + \|\mathrm{B}\|_F^2$ 和 $\mathcal{G}_2(\mathrm{A}, \mathrm{B}) = \|\mathrm{A} - \mathcal{U}\mathrm{A}\|_F^2 + \|\mathrm{B} - \mathcal{V}\mathrm{B}\|_F^2$。

我们将潜在因子 $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_t, \boldsymbol{\gamma}_w)$ 的参数空间 $\mathcal{S}$ 定义如下。令 $\tau_1 > 0$ 为控制潜在因子惩罚项的常数。我们假设 $\mathcal{G}(\boldsymbol{\gamma}) < \tau_1$，其中 $\mathcal{G}$ 可以是 $\mathcal{G}_1$ 或 $\mathcal{G}_2$。此外，由于每个标签的概率恰好为 0、1 或 $C-1$ 的可能性很低，因此假设 $\|\boldsymbol{\gamma}\|_\infty \le \tau_2$ 是合理的。接下来，我们将参数空间定义为：

$$\mathcal{S}(\tau_1, \tau_2) = \{\boldsymbol{\gamma} : \|\mathcal{G}(\boldsymbol{\gamma})\| \le \tau_1, \|\boldsymbol{\gamma}\|_\infty \le \tau_2\},$$

其中，上界 τ_1 可能取决于样本量 m、n 以及潜在因子的维度 k。

为了衡量 $\boldsymbol{\gamma}$ 与 $\tilde{\boldsymbol{\gamma}}$ 在 $\mathcal{S}$（τ_1， τ_2）之间的差异，我们引入 [28]中的海灵格距离：

$$\mathcal{H}_{\mathcal{S}}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}) = \Big\{ \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [\int (f^{1/2}(r_{ij}; \boldsymbol{\gamma}) - f^{1/2}(r_{ij}; \tilde{\boldsymbol{\gamma}}))^2 d\nu(r_{ij})] \Big\}^{1/2},$$

其中， $f(r_{ij}; \boldsymbol{\gamma})$ 是给定参数 $\boldsymbol{\gamma}$ 时 r_{ij} 的似然函数， $\nu(\cdot)$ 是概率测度。海灵格距离 $\mathcal{H}_{\mathcal{S}}(\cdot, \cdot)$ 可以证明是一个有效的距离函数，且有 $\|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}\|_F^2$ 的上界。

我们考虑满足 $\mathcal{PL}(\hat{\boldsymbol{\gamma}} \mid \mathbf{R}) \leq \arg\min_{\boldsymbol{\gamma} \in \mathcal{S}(\tau_1, \tau_2)} \mathcal{PL}(\boldsymbol{\gamma} \mid \mathbf{R}) + \tau_{|\Omega|}$ 的估计量 $\hat{\boldsymbol{\gamma}}$，其中 $\lim_{|\Omega| \to \infty} \tau_{|\Omega|} = 0$。在所提方法中，损失函数 $\mathcal{PL}(\hat{\boldsymbol{\gamma}}|\mathbf{R})$ 有非负下界，因此其下确界始终存在且有限。定理1给出了理论保证：在一定条件下，估计量 $\hat{\boldsymbol{\gamma}}$ 依概率收敛到真实参数 $\boldsymbol{\gamma}^0$ 。

定理1. 假设 $\lambda_{|\Omega|} < {}_{\tau_1}{}^{1}{}_{(m+n)} \epsilon_{|\Omega|}^2$, 则存在一个正常数 c_1 ，使得

$$P(\mathcal{H}_{\mathcal{S}}(\hat{\boldsymbol{\gamma}}, \boldsymbol{\gamma}^0) \geq \epsilon_{|\Omega|}) \leq 7 \exp(-c_1 |\Omega| \epsilon_{|\Omega|}^2), \tag{15}$$

具有尽可能最佳收敛速率的 $\hat{\boldsymbol{\gamma}}$ 是

$$\epsilon_{|\Omega|} \sim \frac{\sqrt{(m+n)k}}{|\Omega|^{1/2}} \left[\log \left\{ \frac{\tau_1 |\Omega|}{\sqrt{mnk}(m+n)} \right\} \right]^{1/2}. \tag{16}$$

此处，(16) 中的 τ_1 由惩罚函数的选择决定，该函数用于量化参数空间 $\mathcal{S}(\tau_1, \tau_2)$ 的体积。考虑惩罚函数 $\mathcal{G}_1(\cdot)$、 $\tau_1 \sim O((m+Cn)k)$ 以及惩罚函数 $\mathcal{G}_2(\cdot)$、 $\tau_1 \sim O((m_0+n_0)k)$，其中 m_0 是任一类别中的最大任务数， n_0 是任一子群体中的最大工作者数。推论1给出了在 $\mathcal{G}_1$ 和 $\mathcal{G}_2$ 下的收敛速率，并表明采用多质心分组惩罚 $\mathcal{G}_2$ 的所提方法，相较于不进行子群体分组时的 L_2 惩罚 $\mathcal{G}_1$ ，具有更快的收敛速率。

推论1。 在定理*1*的假设下，*(15)* 关于 $\mathcal{G}_1$ 的最佳收敛速率为

$$\epsilon_{|\Omega|} \sim \frac{\sqrt{(m+n)k}}{|\Omega|^{1/2}} \left[\log \left\{ \frac{\sqrt{k}|\Omega|}{\sqrt{mn}} \right\} \right]^{1/2}, \tag{17}$$

对于惩罚函数 $\mathcal{G}_2$，最佳收敛速率为

$$\epsilon_{|\Omega|} \sim \frac{\sqrt{(m+n)k}}{|\Omega|^{1/2}} \left[\log \left\{ \frac{\sqrt{k}|\Omega|(m_0+n_0)}{\sqrt{mn}(m+n)} \right\} \right]^{1/2}. \tag{18}$$

注释1. 上述定理和推论基于满足条件 $\mathcal{PL}(\hat{}\boldsymbol{\gamma} \mid \mathbf{R}) \leq \operatorname{argmin}_{\boldsymbol{\gamma} \in \mathcal{S}(\tau_1,\tau_2)} \mathcal{PL}(\boldsymbol{\gamma} \mid \mathbf{R}) + \tau_{|\Omega|}$ 的估计量 $\hat{}\boldsymbol{\gamma}$，其中不要求 $\mathcal{PL}(\cdot \mid \mathbf{R})$ 的全局极小值存在；对于非凸优化而言，这是一个相对宽松的假设。

注释2. 式（*15*）中的收敛是基于参数空间 $\mathcal{S}$ 上的海灵格距离 $\mathcal{H}_{\mathcal{S}}(\cdot,\cdot)$。在 L_2 准则下的收敛性质比海灵格距离下的收敛更严格。具体而言，在海灵格距离下，估计量的收敛仅取决于参数空间 $\mathcal{S}$ 的大小和惩罚系数 $\lambda_{|\Omega|}$。相比之下，在 L_2 距离下的收敛要求更强的局部性质，即 $Var(L(\boldsymbol{\gamma}) - L(\boldsymbol{\gamma}^0))$ 在参数空间上有界。

注释3。与 $\mathcal{G}_1(\cdot)$ 相比， $\mathcal{G}_2(\cdot)$ 在收敛速率方面的改进体现在式（*18*）对数项中的 $\frac{m_0+n_0}{m+n}$。在潜在因子的大小具有一致上界 τ_2 的条件下，当底层存在子群结构时，由 $\mathcal{G}_2(\cdot)$ 约束的参数空间与包含最大子群体的参数空间处于相同量级。相比之下，惩罚项 $\mathcal{G}_1(\cdot)$ 是所有潜在因子仅属于一个组时的特殊情形。因此，由 $\mathcal{G}_2(\cdot)$ 约束的参数空间小于由 $\mathcal{G}_1(\cdot)$约束的参数空间。海灵格距离的优势在于，$\hat{}\boldsymbol{\gamma}$ 的收敛速率仅取决于参数空间的大小和惩罚系数 $\lambda_{|\Omega|}$ *[28]*，因此，在 $\mathcal{G}_2(\cdot)$ 下的收敛速率快于在 $\mathcal{G}_1(\cdot)$下的收敛速率。

注释4。在　　为了获得模型参数的相合估计，并使　　e

任务标签，我们要求随着任务数量 m 和工作者数量 n 趋于无穷大，$\epsilon_{|\Omega|}$ 趋于零。因此，我们可以根据推论*1*给出的最佳收敛速率推导 $|\Omega|$ 的阶。具体而言，如果使用 L_2 惩罚项来估计潜在因子，那么 $|\Omega| \geq \mathcal{O}((m+n)k\log\{(m+n)k\})$可保证$\hat{\gamma}$收敛，并保证任务和工作者群组成员关系的相合估计；如果使用所提出的多质心分组惩罚，那么 $|\Omega| \geq \mathcal{O}((m+n)k\log\{(m_0+n_0)k\})$（其中 m_0和 n_0分别是任务组和工作者组的最大基数）可保证$\hat{\gamma}$及群组成员关系具有相同的理论性质。值得注意的是，与 L_2 惩罚项相比，采用*MCGP*所需的观测标签更少，因为所提出的*MCGP*能够利用来自同一组个体的信息。

接下来，我们证明子群成员关系的一致性。令 $M_k^* = \{i : \mathbf{U}_i^* = k\}$为第 k 个真实任务类别，$N_k^* = \{j : \mathbf{V}_j^* = k\}$ 为第 k 个真实工人子群组。$\hat{}$ 类似地，我们分别将 M_k 和 N_k 定义为第 k 个估计任务类别和工人子群组。注意到子群成员关系在标签置换下保持不变，我们将置换操作定义为 $\pi(\cdot)$，并将子群成员关系的估计误差定义为

$$err_{\text{task}} = \min_{\pi(\cdot)} \frac{1}{m} \sum_{k=0}^{C-1} |M_{\pi(k)}^* \backslash \hat{M}_k|, \quad err_{\text{worker}} = \min_{\pi(\cdot)} \frac{1}{n} \sum_{k=1}^{D} |N_{\pi(k)}^* \backslash \hat{N}_k|,$$

其中，$|\cdot|$ 是集合的基数，$A \backslash B$ 表示集合 $\{i：\ i \in A$ 和 $i \notin B\}$。

定理2. 假设定理*1*中的假设成立，且存在常数 $\eta_1 > 0$，使得 $\frac{1}{m}\|\mathbf{A}^* - \mathcal{U}^*\mathbf{A}^*\|_{2\to\infty} \leq \eta_1$且 $\frac{1}{n}\|\mathbf{B}^* - \underline{\mathcal{V}}^*\mathbf{B}^*\|_{2\to\infty} \leq \eta_1$，其中 $\|\mathbf{A}\|_{2\to\infty} = \sup_{\|\mathbf{x}\|=1}\|\mathbf{A}\mathbf{x}\|_\infty$ˆ是所谓的二到无穷范数 *[6]*。那么，估计的子群成员关系$\mathbf{U}$ 和ˆ$\mathbf{V}$ 是一致的：

$$err_{task} \to_P 0, \quad err_{worker} \to_p 0.$$

备注5。该 定理*2*中的常数 η_1 控制各个潜在因子与其对应 t

群组质心之间的距离，用以衡量任务组和工人组的集中程度。由于潜在因子建模 *(6)* 旨在捕捉任务和工作者内部的异质性，因此 $\mathbf{A}$ 和 $\mathbf{B}$ 需要满足正的 η_1 ，这与 *[42]* 中的网络嵌入截然不同。在理论分析中，他们采用随机块模型 *[27]* ，其中同一组中的节点共享相同的潜在因子；这与我们模型中的 $\eta_1 = 0$ 等价。

推论2. 假设定理*1*和定理*2*中的条件成立，且假设ˆ*1*（高质量工人子群识别假设）成立，则预测标签$\mathbf{Z}$ 依概率收敛到$\mathbf{Z}$ 。

定理2保证了所提出的子群成员关系估计量具有一致性，这表明来自不同类别的任务渐近地可以聚类到不同的子群体中。类似地，工作者的子群成员关系也可以渐近地恢复。总之，通过识别高质量工人子群，可以根据假设1渐近地恢复真实标签与任务成员关系之间的对应关系。因此，预测标签是真实标签的一致估计量。定理和推论的证明见补充材料。

6 模拟研究

本节通过多类别众包模拟来展示所提方法的性能。具体而言，我们将所提方法与三种现有方法进行比较，包括多数投票（MV）、以多数投票初始化的 Dawid-Skene估计器（DS-MV） [7], 以及以谱方法初始化的 Dawid-Skene估计器（DS-Spectral） [43]。在补充材料中，我们还考虑了更多模拟设置，以考察所提方法在二元众包问题中的性能，并与另外两种专为二元众包设计的方法进行比较。

表1：研究2中各工人组在任务类别上的专业水平分配和潜在因子质心。“高”表示正确标注特定任务类别的概率高，“低”表示错误标注特定任务类别的概率高，“平均”表示专业水平高于“低”但低于“高”。

	工作者组1	工作者组2	工作者组3
标签0任务类别 $(2, 0, 0)'$	high $(2, 0, 0)'$	low $(0, 1, 1)'$	平均 $(1, 1, 0)'$
标签-1任务类别 $(0, 2, 0)'$	平均 $(1, 1, 1)'$	high $(0, 2, 0)'$	平均 $(1, 1, 0)'$
标签-2 任务类别 $(0, 0, 2)'$	low $(1, 1, 0)'$	low $(1, 1, 0)'$	high $(0, 0, 2)'$

6.1 研究1：多类别众包

在本研究中，我们考虑包含三个任务类别（$C = 3$）和三个工人组（$D = 3$）的多类别众包。在多类别情形下，不同任务类别对应的高质量工人组可能并不相同。

令任务数量为 $m = 150$，每个类别包含 50 个任务；工作者数量为 $n = 150$，每个组包含 50 名工作者。任务的潜在因子 $\mathbf{a}_i$ 由多元正态分布 $\mathcal{N}(\boldsymbol{\alpha}_c, \sigma^2 \mathbf{I}_3)$ 生成，其中 $\boldsymbol{\alpha}_0 = (2, 0, 0)'$ 对应标签1任务，$\boldsymbol{\alpha}_1 = (0, 2, 0)'$ 对应标签2任务，$\boldsymbol{\alpha}_2 = (0, 0, 2)'$ 对应标签3任务。此外，方差 $\sigma^2 \in \{0.5, 2\}$ 控制任务异质性，其中 $\sigma^2 = 0.5$ 和 $\sigma^2 = 2$ 分别表示同质场景和异质场景。工作者的潜在因子 $\mathbf{b}_j$ 由多元正态分布 $\mathcal{N}(\boldsymbol{\beta}_{c,d}, \mathbf{I}_2)$ 生成，其中 $\boldsymbol{\beta}_{c,d}$ 是第 d 个工人组在标签-c 任务上的质心。质心 $\boldsymbol{\beta}_{c,d}$ 与各工人组在各任务类别上的专业水平相关；与任务质心具有高一致性，表明专业水平较高。表1列出了工作者在各任务类别上的专业水平以及工作者的潜在因子质心的各种设定。众包标签 r_{ij} 由参数为 $(\tilde{p}_{ij,1}, \tilde{p}_{ij,2}, \tilde{p}_{ij,3})$ 的多项分布生成，其中 $\tilde{p}_{ij,c} = \frac{exp(\mathbf{a}_i'\mathbf{b}_{j,c})}{\sum_{c=1}^{3} exp(\mathbf{a}_i'\mathbf{b}_{j,c})}$ 表示 r_{ij} 等于 c 的概率。每个众包标签

表2：所提方法与三种现有方法的标签预测准确率（标准误）比较。模拟结果基于200次独立实验。

σ^2	所提方法	MV	DS-MV	DS-Spectral
0.5	**0.928(0.050)**	0.647(0.089)	0.886(0.040)	0.889(0.039)
2	**0.745(0.030)**	0.599(0.129)	0.702(0.045)	0.701(0.051)

以 0.7 的概率随机缺失。

我们将所提方法与MV、DS-MV和DS-Spectral进行比较。对于所提方法，我们将潜在因子维度 k 从2调至5，将工人组数量 D 从2调至6，并将惩罚系数 λ 从0.001调至1。标签预测准确率和标准误见表2。显然，在同质和异质场景中，所提方法均优于其他方法。在同质场景中，所提方法的准确率比其他方法高4.38%（DS-Spectral）至43.4%（MV）；在异质场景中，所提方法的准确率比其他方法高6.13%（DS-MV）至24.4%（MV）。

所提方法的改进源于以下原因。首先，所提方法能够以较高的准确率识别每个任务类别中的高质量工人组，并在分配标签时高度依赖高质量工作者。具体而言，在同质和异质场景下，所提方法对工人组的分类准确率分别达到 75.5% 和 73.2%。相比之下，Dawid-Skene 估计量不考虑工作者的质量，而是依赖多数工人分配标签，因此预测准确率较低，尤其是在多数工人子群组中包含低质量标注者的任务上。其次，在异质场景下，所提方法的潜在因子建模纳入了同一类别内各任务之间的异质性，而多数投票和基于混淆矩阵的方法则未考虑这一点。

场景	#任务组	任务组质心	#工人组	工作者组质心
1	3	标签-1: (2,0,0) 标签-2：（0，2，0） 标签-3：（0，0，2）	2	高质量: (2，0，0)，(0，2，0)，(0，0，2) 中等质量: (2，0，0)，(0，2，1)，(0，0，2)
2	3	标签-1: (2，0，0) 标签-2: (0，2，0) 标签-3：（0，0，2）	2	高质量: (2，0，0)，(0，2，0)，(0，0，2) 低质量: (1，2，0)，(0，2，1)，(1，1，1)
3	3	标签-1：(2,0,0) 标签-2：(0,2,0) 标签-3：(0,0,2)	3	高质量: (2，0，0)，(0，2，0)，(0，0，2) 中等质量: (2，0，0)，(0，2，1)，(0，0，2) 低质量: (1，2，0)，(0，2，1)，(1，1，1)
4	3	标签-1：(2，0，0) 标签-2：(0，2，0) 标签-3：(0，0，2)	3	高质量: (2，0，0)，(0，2，0)，(0，0，2) 低质量: (2，0，0)，(1，0，1)，(0，2，0) 低质量: (1，2，0)，(0，2，1)，(1，1，1)

表3：多类别任务的模型设置

6.2 研究2：对工人子群组数量的稳健性

本节研究所提方法对工人组数量设定错误的稳健性。一般而言，选择适当数量的簇，对于得到可解释或合理的聚类结果至关重要。在所提出的框架中，任务组的数量预先设定为任务类别的数量，而工人组的数量则由我们的算法确定。接下来，我们将针对多类别任务，在若干场景下考察所提方法对工人组数量的稳健性。

我们考虑了四种不同于第6.1节设置的三类别任务情景。详细的模型设置见表3。与上述设置类似，我们假设存在高质量、中等质量和低质量的工作者子群。我们模拟了150个任务，每个类别包含50个任务，并将300名工作者平均分配到各组。每种情景均重复模拟200次。

模拟结果如表4所示，这表明所提方法相当

	指定的工人组数量			
场景	2	3	4	5
1	0.884(0.194)	0.888(0.187)	0.887(0.186)	**0.888(0.028)**
2	**0.893(0.116)**	0.892(0.181)	0.888(0.186)	0.886(0.017)
3	0.941(0.023)	0.946(0.028)	**0.956(0.021)**	0.953(0.022)
4	0.824(0.239)	0.945(0.027)	**0.945(0.026)**	0.944(0.026)

表4：模拟结果：对工人组数量的稳健性，多类别任务的标签预测准确率

能够稳健应对工人组数量设定错误的问题。对于全部四种情景，不同工人组数量下的预测准确率都相当稳定，而且工人组数量较多时，标准差甚至会显著降低。工人组数量较少时，一些低质量标注者可能会被划分到高质量工人组中，从而降低整个组的标注确定性。因此，我们建议在实际应用中使用相对较多的工人组。

6.3 研究3：初始化的数值敏感性

由于优化问题（9）具有非凸性，通常无法通过数值方法获得全局极小值，而且算法所得解取决于初始化。因此，必须研究所提方法对不同初始化的敏感性，并提出一种高效的初始化方案，以获得良好的实证结果。

在本小节中，我们研究了几种潜在因子和子群归属初始化方案，并比较其模拟结果。具体而言，我们在所提模型中采用三种方案来初始化群组成员关系和潜在因子：随机、多数投票和 Dawid-Skene估计 [7], 其中后两种方案是信息性方案，可根据给定数据提供热启动初始估计。在每种方案下，我们首先初始化任务成员归属 $\mathbf{U}^{(0)}$，三种方案采用不同的任务成员关系初始化方式：随机生成成员关系、多数投票或 Dawid-Skene估计 [7]。接下来，我们根据三种已初始化任务成员关系的标注准确率来初始化工人成员关系。具体而言，第 j位工作者的群组

表5：采用不同初始化方案时，所提方法在多类别众包中的模拟结果。随机方案：随机初始化任务/工作者成员关系和潜在因子；多数投票方案：基于多数投票初始化任务成员关系及其对应的工作者成员关系和潜在因子；DS方案：基于DS估计初始化任务成员关系及其对应的工作者成员关系和潜在因子。这里，稳定性是每次实验中预测准确率均值和标准误的平均值。

初始化	正交矩阵初始化	收敛迭代次数	准确率	稳定性
随机方案	随机	14.82(6.98)	0.595(0.097)	0.354(0.115)
多数投票方案	单位矩阵	12.79(6.41)	0.697(0.059)	0.575(0.088)
DS方案	单位矩阵	**11.46(5.11)**	**0.745(0.030)**	**0.669(0.070)**

群组成员关系指定如下：

$$\mathbf{V}_j^{(0)} = \begin{cases} 1 & acc(\mathbf{R}_{\cdot j}, \mathbf{U}^{(0)}) \le q_{\frac{1}{D}} \\ 2 & q_{\frac{1}{D}} < acc(\mathbf{R}_{\cdot j}, \mathbf{U}^{(0)}) \le q_{\frac{2}{D}} \\ \cdots & \\ D & q_{\frac{D-1}{D}} < acc(\mathbf{R}_{\cdot j}, \mathbf{U}^{(0)}) \le 1 \end{cases},$$

其中，$acc(\mathbf{R}_{\cdot j}, \mathbf{U}^{(0)}) = \frac{1}{|\mathbf{R}_{\cdot j}|} \sum_i I(\mathbf{R}_{ij} = \mathbf{U}_i^{(0)})$ 是第 j 位工作者在初始化的任务组成员归属 $\mathbf{U}^{(0)}$ 下的准确率。阈值 $q_{\frac{1}{D}}$，$q_{\frac{2}{D}}$，……，$q_{\frac{D-1}{D}}$ 是所有工作者准确率经验分布的第 $\frac{1}{D}$，$\frac{2}{D}$，……，$\frac{D-1}{D}$ 分位数。

对于任务和工作者潜在因子的初始化，我们在随机方案和信息性（MV 或 DS [7]）方案中采用不同的方法。在随机方案中，无论群组成员关系如何，我们都按照 $\mathbb{R}^k$ 中的随机标准正态分布初始化潜在因子。在信息性方案中，我们使用 $\mathbb{R}^k$ 中的不同正态分布初始化潜在因子，其中由 $\mathbf{U}^{(0)}$ 或 $\mathbf{V}^{(0)}$ 指定的每个组都有各自的随机单位长度均值，协方差矩阵则指定为单位矩阵。此外，我们比较了多类别众包中初始化正交矩阵的两种方案：（1）所有 $\mathbf{O}_c^{(d)}$ 均使用单位矩阵；（2）除固定的参考组 $\mathbf{O}_0^{(d)}$ 外，均使用随机正交矩阵。

在接下来的　接下来，我们报告基于上述初始化

表 6：五个真实数据集的摘要统计量，包括类别数、任务数、工作者数、众包标签数以及众包标签矩阵的缺失率。

数据集	#类别	#任务	#工作者	#众包标签	缺失率
BlueBird	2	108	39	4,212	0.0%
青光眼	2	127	78	2,540	74.4%
乳腺癌	2	500	24	12,000	0.0%
Iris	3	100	24	2,000	16.7%
Dog	4	807	52	7,354	82.5%

方案的模拟结果。我们按照第 6.1 节中的异质性设定（$\sigma^2 = 2$）进行模拟。我们进行了 100 次独立实验，并在每次实验中生成 10 组不同的群组成员关系和潜在因子初始值。我们在表5中报告迭代次数、准确率和稳定性等性能指标。信息性方案比随机方案收敛更快，并取得了更高的预测准确率和更好的稳定性。

总之，我们的数值实验表明，信息性初始化方案对于确保更快收敛并提升实证表现至关重要。因此，我们建议在实践中采用信息性方案，以获得更快速、更准确的预测。

7 真实数据应用

本节中，我们将所提方法应用于五个真实数据集，并将其表现与几种竞争方法进行比较。在这五个数据集中，Bluebird [39], 青光眼[25]和乳腺癌 [31]是二分类任务，而鸢尾花 [12] 和狗 [8]是多类别任务。Bluebird、青光眼和狗数据集由在线平台上的人工标注者标注，而乳腺癌和鸢尾花数据集则由多个机器学习算法标注。五个数据集的摘要统计量见表 6。

蓝鸟数据集包含 108 张图像，工作者的任务是判断图像中是否能看到鸟。青光眼数据集包含正常和青光眼视盘图像，其疾病状态由专家预先确定。工作者经过简单教程培训后，负责判断视盘图像中的疾病状态。每张图像的真实标签

并不可用，不过，我们可以通过 [25] 中的摘要统计量来预测真实标签。乳腺癌数据集包含从乳腺肿块细针穿刺抽吸物（FNA）的数字化图像中提取的手工设计特征，以及相应的癌症状态，例如恶性或良性。该数据集包含 569 个观测值，其中前 69 个观测值用于训练附录中列出的 24 个不同的二元分类器。这 24 个分类器被视为众包工作者，而其余 500 个观测值的预测标签则被视为众包标签。

鸢尾花数据集包含三种鸢尾花及其特征测量值。我们使用前50个观测值训练24种不同的分类器，并用训练好的分类器预测其余100个观测值，作为众包标签。狗数据集由四种犬种组成，工作者需要将每张狗的图像标记为诺福克梗、诺维奇梗、爱尔兰猎狼犬或苏格兰猎鹿犬。

除了第6节中介绍的方法外，我们还将所提方法与另外两种竞争方法进行了比较：混合策略众包（MiSC [18]）和基于社区的贝叶斯分类器组合（CBCC， [37]）。MiSC方法对增广张量进行分解，而CBCC则利用子群结构来建模工作者混淆矩阵。这些方法的比较结果汇总于表7，结果表明，所提方法在标签预测准确率方面始终优于其他方法。具体而言，在蓝鸟数据集上，所提方法比其他方法提高了3.1%至29.5%；在青光眼数据集上提高了1.6%至76.9%；在乳腺癌数据集上提高了2.36%至4.95%。对于多类别数据集，所提方法在鸢尾花数据集上提升了3.2%至7.07%，在狗数据集上提升了0.0%至5.22%。

所提方法优于其他方法，原因如下。首先，同一类别中的任务可能存在异质性。例如，在图4中，左图展示了通过 t-分布随机邻域嵌入（t-SNE） [23]对乳腺癌数据进行降维后得到的前两个潜在因子的可视化结果，其中良性或恶性类别的潜在特征都分散在潜在空间中。所提方法能够利用潜在因子建模来纳入异质性。其次，工作者的专业水平

表7：所提方法与现有方法在真实数据上的真实标签预测准确率。“MV”是多数投票，“DS-MV”是以多数投票初始化的Dawid-Skene估计器，“DS-Spectral”是以谱方法初始化的Dawid-Skene估计器，“MiSC”是混合策略众包，“CBCC”是基于社区的贝叶斯分类器组合，“SVD”是奇异值分解，“GLAD”是生成式标签能力与难度模型。

数据集	类别	所提方法	MV	DS-MV	DS-Spectral	MiSC	CBCC	SVD	GLAD
BlueBird	2	**0.935**	0.759	0.889	0.899	0.907	0.889	0.722	0.789
青光眼	2	**0.913**	0.516	0.890	0.787	0.899	0.874	0.567	0.726
乳腺癌	2	**0.868**	0.827	0.848	0.848	0.844	0.838	0.840	0.836
Iris	3	**0.960**	0.890	0.920	0.920	0.930	0.920	-	-
Dog	4	**0.846**	0.804	0.833	0.831	**0.846**	0.814	-	-

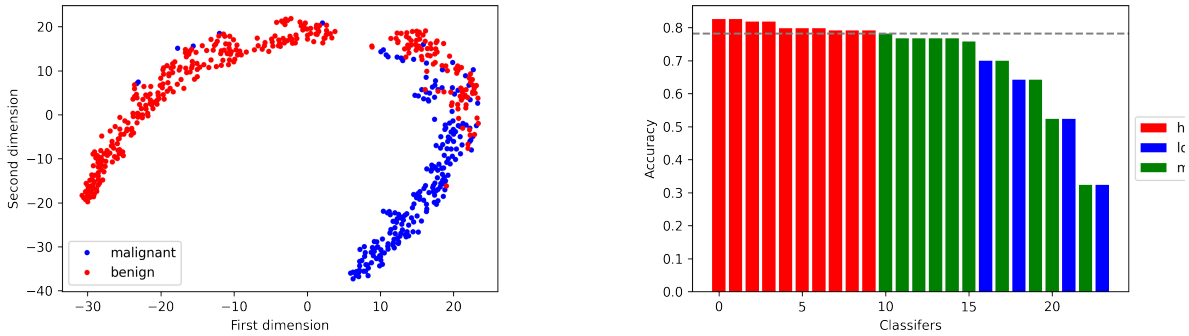


图4：乳腺癌数据集。左图：数据点经 t-SNE [23] 转换；右图：分类器准确率。

可能差异很大。图4右图展示了每个单独训练的分类器的准确率，表明某些分类器的表现优于其他分类器；所提方法能够识别高质量工作者，从而提高标签预测准确率。

8 讨论

本文提出了一种用于多类别分类的新型众包框架。所提方法能够通过潜在因子建模，将任务和工作者之间的异质性纳入考量。我们还利用多质心分组惩罚和组特定旋转矩阵，在潜在空间中利用子群结构。因此，任务

可以根据其类别进行聚类，工作者也可以根据其对不同任务的意见加以区分。与现有众包方法相比，所提方法在假设限制较少的情况下识别出高质量工作者，并提供了一种对含噪众包标签更具鲁棒性的标签规范流程。

理论上，我们证明了模型参数在海灵格距离下的估计一致性。与 L_2 惩罚项相比，采用多质心组惩罚项的所提方法能够更快收敛。此外，我们还证明了标签预测渐近一致。对于有限样本问题，我们的模拟研究表明，所提方法通过考虑任务和工作者之间的异质性及子群结构，提高了标签预测准确率。当低质量标注者相对于高质量工人的比例较高时，这种提升更为显著，这得益于稳健的标签指定流程。

未来有几个值得进一步探索的拓展方向。本文仅考虑任务和工作者潜在因子之间的线性一致性。然而，在实际应用中，可以通过非线性矩阵补全技术 [11] 将其推广，以建模更复杂的任务与工人关系。此外，在 MTurk 等平台上，通常会通过激励鼓励高质量工作者标注更多任务，因此众包标签并非随机缺失。因此，新方法应能将信息性缺失机制纳入考量。此外，高质量工作者的假设1在统计上无法检验，多数投票也同样如此。因此，探索一种能够验证这一假设的有效流程也很有意义。

致谢

作者感谢编辑、副编辑和匿名审稿人提出的深刻建议与宝贵反馈，这些意见显著提升了本文质量。作者声明，本研究的直接应用未引发任何财务或非财务利益冲突。本研究得到美国国家科学基金会资助项目 DMS 2210640、DMS 1952406、香港研究资助局

资助项目 GRF-11304520、GRF-11301521、GRF-11311022 以及香港中文大学启动基金 4937091 的资助。

补充材料

补充材料提供了二元众包的方法、二元众包的模拟结果，以及定理和推论的证明。

参考文献

[1] Aydin, B. I., Yilmaz, Y. S., Li, Y., Li, Q., Gao, J., and Demirbas, M. (2014). 多项选择题问答中的众包。载于*AAAI*，第2946–2953页。Citeseer。[2] Berend, D. and Kontorovich, A. (2014). 加权多数投票的一致性。载于神经信息处理系统进展，第27卷，第3446–3454页。Curran Associates, Inc. [3] Bondell, H. D. and Reich, B. J. (2008). 预测变量的同步回归收缩、变量选择和监督聚类（OSCAR）。生物统计学，64(1):115–123。[4] Borromeo, R. M. and Toyama, M. (2015). 自动情感分析与众包情感分析。载于第*19*届国际数据库工程与应用研讨会论文集，第90–95页，纽约。[5] Cao, P., Xu, Y., Kong, Y., and Wang, Y. (2019). Max-mig：一种从众包数据中进行联合学习的信息论方法。*arXiv*预印本 *arXiv:1905.13436*。[6] Cape, J., Tang, M., and Priebe, C. E. (2019). 二到无穷范数与奇异子空间几何及其在高维统计学中的应用。统计学年鉴，47(5):2405–2439。[7] Dawid, A. P. and Skene, A. M. (1979). 使用EM算法对观察者错误率进行最大似然估计。英国皇家统计学会杂志：*C*辑（应用统计），28(1):20–28。[8] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet：大型层次图像数据库。载于*2009*年*IEEE*计算机视觉与模式识别会议，第248–255页。[9] Deng, K., Han, S., Li, K. J., and Liu, J. S. (2014). 基于顺序的排名数据的贝叶斯聚合。《美国统计学会杂志》，109(507):1023–1039。[10] Engel, J. (1988). 多项逻辑回归。荷兰统计学，42(4):233–252。[11] Fan, J. and Cheng, J. (2018). 通过深度矩阵分解进行矩阵补全。神经网络，98:34–41。[12] Fisher, R. A. (1936). 多重测量在分类学问题中的应用。优生学年鉴，7(2):179–188。[13] Foncubierta Rodríguez, A. and Müller, H. (2012). 医学影像中的真实标注生成：一种基于众包的迭代方法。载于*2012*年*ACM*多媒体众包研讨会论文集，第9–14页，纽约。[14] Ghosh, A., Kale, S., and McAfee, P. (2011). 谁来监督版主？用户生成内容中的众包滥用行为检测。载于第*12*届*ACM*电子商务会议论文集，第167–176页，纽约。[15] Hoaglin, D. C. and Welsch, R. E. (1978). 回归与方差分析中的帽子矩阵。美国

统计学家，32(1):17–22。[16] Ibrahim, S.、Fu, X.、Kargas, N. 和 Huang, K.（2019）。通过成对共现进行众包：可识别性与算法。载于《神经信息处理系统进展》，第32卷，第7847–7857页。[17] Kim, H.-C. 和 Ghahramani, Z.（2012）。贝叶斯分类器组合。第22卷，第619–627页，加那利群岛拉帕尔马。[18] Ko, C.-Y.、Lin, R.、Li, S. 和 Wong, N.（2019）。MiSC：混合策略众包。*arXiv*预印本 *arXiv:1905.07394*。[19] Kovashka, A.、Russakovsky, O.、Fei-Fei, L. 和 Grauman, K.（2016）。计算机视觉中的众包。*arXiv*预印本 *arXiv:1611.02145*。[20] Li, Y.、Du, N.、Liu, C.、Xie, Y.、Fan, W.、Li, Q.、Gao, J. 和 Sun, H.（2017）。从众包中获得可靠的医学诊断：从非专家中发现可信答案。载于《第十届ACM国际网络搜索与数据挖掘会议论文集》，第253–261页，纽约。[21] Li, Y.、Rubinstein, B. 和 Cohn, T.（2019）。利用工作者相关性进行众包中的标签聚合。载于《第36届国际机器学习会议论文集》，第97卷，第3886–3895页，长滩，加利福尼亚州。[22] Ma, S. 和 Huang, J.（2017）。一种用于亚组分析的凹成对融合方法。《美国统计学会杂志》，112(517):410–423。[23] Maaten, L. v. d. 和 Hinton, G.（2008）。使用t-SNE可视化数据。机器学习研究杂志，9(86):2579–2605。[24]MacLean, D. L. 和 Heer, J.（2013）。识别患者撰写文本中的医学术语：一种基于众包的方法。美国医学信息学会杂志，20(6):1120–1127。[25] Mitry, D.、Peto, T.、Hayat, S.、Blows, P.、Morgan, J.、Khaw, K.-T. 和 Foster, P. J.（2015）。利用众包作为筛查工具，通过数字摄影检测青光眼性视神经病变的临床特征。*PLOS ONE*，10(2):e0117401。[26] Nitzan, S. 和 Paroush, J.（1982）。不确定二分选择情境中的最优决策规则。国际经济评论，23(2):289–297。[27] Rohe, K.、Qin, T. 和 Yu, B.（2016）。对有向图进行协同聚类，以发现不对称性和方向性社群。美国国家科学院院刊，113(45):12679–12684。[28] Shen, X.（1998）。论惩罚方法。中华统计学志，第337–357页。[29] Snow, R.、O'Connor, B.、Jurafsky, D. 和 Ng, A. Y.（2008）。廉价且快速——但效果好吗？评估用于自然语言任务的非专家标注。载于《2008年实证

自然语言处理方法》，第254–263页，美国。[30] Stone, M.（1978）。交叉验证：综述。《统计学：理论与应用统计学杂志》，9(1):127–139。[31] Street, W. N.、Wolberg, W. H. 和 Mangasarian, O. L.（1993）。用于乳腺肿瘤诊断的核特征提取。载于《生物医学图像处理与生物医学可视化》，第1905卷，第861–870页。[32] Su, H.、Deng, J. 和 Fei-Fei, L.（2012）。用于视觉目标检测的众包标注。载于《人类计算——2012年AAAI研讨会论文，技术报告》，第40–46页。[33] Sugar, C. A. 和 James, G. M.（2003）。确定数据集中簇的数量：一种信息论方法。《美国统计学会杂志》，98(463):750–763。[34] Tang, X.、Xue, F. 和 Qu, A.（2020）。个体化多方向变量选择。《美国统计学会杂志》，0(0):1–17。[35] Tibshirani, R.、Saunders, M.、Rosset, S.、Zhu, J. 和 Knight, K.（2005）。通过融合套索实现稀疏性与平滑性。《英国皇家统计学会杂志：*B*辑（统计方法学）》，67(1):91–108。[36] Tibshirani, R.、Walther, G. 和 Hastie, T.（2001）。通过间隙统计量估计数据集中的簇数量。《英国皇家统计学会杂志：*B*辑（统计方法学）》，63(2):411–423。[37] Venanzi, M.、Guiver, J.、Kazai, G.、Kohli, P. 和 Shokouhi, M.（2014）。用于众包的基于社区的贝叶斯聚合模型。载于《第23届万维网国际会议论文集》，第155–164页。[38] Wang, H.、Li, R. 和 Tsai, C.-L.（2007）。平滑截断绝对偏差法的调参选择器。《生物统计学》，94(3):553–568。[39] Welinder, P.、Branson, S.、Perona, P. 和 Belongie, S. J.（2010）。多维群体智慧。载于《神经信息处理系统进展》，第23卷，第2424–2432页。[40] Wen, Z. 和 Yin, W.（2013）。一种带正交约束的优化可行方法。数学规划，142(1-2):397–434。[41] Whitehill, J.、Wu, T.-f.、Bergsma, J.、Movellan, J. R. 和 Ruvolo, P. L.（2009）。谁的投票应当拥有更大权重：对专业水平未知的标注者所提供标签的最优整合。载于《神经信息处理系统进展》，第22卷，第2035–2043页。[42] Zhang, J.、He, X. 和 Wang, J.（2021）。利用网络嵌入进行有向社区检测。《美国统计学会杂志》，第1–11页。[43] Zhang, Y.、Chen, X.、Zhou, D. 和 Jordan, M. I.（2016）。谱方法结合期望最大化算法：一种理论上最优的众包算法。《机器学习研究杂志》，17(1):3537–3580。