arXiv:2301.00909v1 [stat.ML] 3 Jan 2023

协同过滤风格的多维项目反应理论

Yoav Bergner[+], Peter F Halpin[*] & Jill-Jênn Vie[×]

2023年1月4日

[+]纽约大学； [*]北卡罗来纳大学教堂山分校； [×]Inria

协同过滤风格的多维项目反应理论

摘要

本文介绍了一种用于多维项目反应理论（MIRT）的机器学习方法。多维项目反应理论是一类潜因子模型，可利用观测到的评估数据对学生表现进行建模和预测。受协同过滤启发，我们定义了一类通用模型，其中包含许多多维项目反应理论模型。我们讨论如何使用惩罚联合最大似然（JML）估计各个模型，并通过交叉验证选出性能最佳的模型。利用分批处理技术可以优化这一模型评估流程，从而高效分析稀疏的大规模数据。我们通过模拟数据和真实数据（包括一个大规模开放在线课程（MOOC）的实例）展示了我们的方法。针对这一大型稀疏数据集拟合的高维模型，不适合采用传统的因子解释方法。类比推荐系统应用，我们提出对因子模型进行另一种“验证”，即利用开卷考试期间学生查阅项目的受欢迎程度等辅助信息。

关键词：项目反应理论、多维性、机器学习、协同过滤、联合最大似然

引言

信息技术在教育领域的广泛应用催生了学习分析和教育数据挖掘等新兴研究领域。在学习分析领域，有些问题明确涉及学生能力及其发展的建模——这是心理测量学研究长期关注的主题——但其估计方法主要借鉴了机器学习和人工智能框架（Reye，2004；Martin、Mitrovic、Mathan 和 Koedinger，2010；Chrysafiadi 和 Virvou，2013）。事实上，一些作者已直接探讨机器学习方法与项目反应理论（IRT）等心理测量学框架之间的联系（Desmarais 和 Pu，2005；Cen、Koedinger 和 Junker，2006；Pel´anek，2016）；测量领域成熟理论的价值也日益受到学习分析领域的广泛认可（Bergner，2017）。本文通过阐释一种植根于机器学习的多维项目反应理论（MIRT），具体而言是基于模型的协同过滤（Bergner、Droschler 和 Kortemeyer，2012；Hu、Zhou、Wang、Li 和 Shen，2009；Su 和 Khoshgoftaar，2009；Zhou、Wilkinson、Schreiber 和 Pan，2008），介绍这两个研究领域之间的一种具体桥梁。本文的总体目标是展示协同过滤如何为预测性多维项目反应理论提供通用框架，讨论这种方法相对于现有心理测量学文献的优势与不足，并通过模拟数据和真实数据示例说明其应用。

我们提出一类模型，其中既包含许多著名的项目反应理论模型（例如单参数逻辑模型、双参数逻辑模型、多维双参数逻辑模型），也包含若干此前尚未命名的模型。本文重点讨论补偿性多维模型，不过这并非协同过滤本身固有的限制。通过按参数复杂度排列候选模型，可以系统地搜索模型空间。就实际操作而言，这里的情形与探索性因子分析（EFA）相似：r维模型嵌套在所有维度高于 r的模型中。然而，不同于

探索性因子分析的传统方法，维度是根据样本外预测准确率而非样本内拟合优度来确定的。因此得到的模型具有最大的预测能力，但不一定具有解释性。

预测准确度是机器学习中广泛使用的模型拟合优度评判标准，也是协同过滤中一种直观的度量指标。在当前情境下，目标是评估各模型预测缺失项目反应的效果。这通常称为矩阵补全；在协同过滤中，它常被理解为向用户推荐内容。我们认为，尽管预测准确度尚未普遍用于评估项目反应理论中的模型拟合优度，但项目反应理论的传统观点也明确关注预测准确度。Lord（1980）对项目反应理论目的的阐述很好地概括了这一观点："通过项目参数描述项目，通过受测者参数描述受测者，从而能够以概率方式预测任何受测者对任何项目的反应，即使相似的受测者此前从未做过相似的项目"（第11页）。为推动预测准确度在项目反应理论情境中的应用，我们介绍了有关项目反应理论模型平均预测准确度上界的一些新结果，这些结果有助于解读模型预测的质量。

在模型估计中，我们采用一种联合最大似然（JML）方法（Lord、Novick 和 Birnbaum，1968）的变体，并使用 L_2 范数对项目参数和个体参数进行惩罚。我们采用惩罚化联合最大似然，缘于正则化相关文献；正则化的早期根源可追溯至岭回归（Alldredge 和 Gilb，1976）。近来，它已成为机器学习中的常用方法，主要用于避免对含有大量潜在预测变量的回归模型进行过拟合（Tibshirani，1996；Hastie、Tibshirani 和 Friedman，2009）。在此，我们将基本相似的方法用于预测性多维项目反应理论中的惩罚化联合最大似然。

关于将惩罚化及相关方法应用于多维项目反应理论及相关模型的研究日益增多。许多方法采用惩罚化（通常为 LASSO 或其变体）来使模型参数具有稀疏性。例如

对因子载荷进行惩罚化，可以解决探索性模型中的旋转不确定性（例如，Hirose & Yamamoto, 2015年；Jin、Moustaki 和 Yang-Wallentin，2018年；Sun、Chen、Liu、Ying 和 Xin，2016年；Trendafilov & Adachi，2015年；Trendafilov、Fontanella & Adachi，2017年）；对残差协方差进行惩罚化，则可为验证性情境中的模型修正提供一种高效方法（例如，Pan、Ip 和 Dub´e，2017年、2019年）。本研究不关注因子旋转问题，而是关注对项目反应矩阵中缺失元素的概率性恢复。

众所周知，JML会受到Neyman-Scott问题的影响（Haberman，1977），自从边际最大似然（MML）变得易于计算以来（Bock和Aitken，1981），心理测量学文献便普遍不再采用JML。对于维度数量适中的情况，采用MML的多维项目反应理论在计算上是可行的（例如，Cai，2010）；然而，它通常不适用于协同过滤中遇到的题目规模，因为数百万受试者要回答数千个项目（Zhou等，2008）。因此，协同过滤方法往往以牺牲统计严谨性为代价，换取可扩展算法和实用的解决方案；协同过滤采用惩罚化联合最大似然，却以其他名称称呼的情况也并不少见（例如，Zhu、Shen和Ye，2016）。近期，Chen、Li和Zhang（2019a，2019b）重新考察了探索性和验证性情境下的JML，此时受试者数量和项目数量均趋于无穷。与本研究最相关的是，作者证明，在维度数量固定的模型中，JML的一个受约束版本可以确保因子载荷的（某种意义上的）渐近一致性。

本文提出，惩罚化联合最大似然在机器学习典型的预测应用中能够发挥重要作用；随着大规模开放在线课程等开放教育资源成为教育领域的常态，这一作用可能会日益凸显。联合最大似然速度快，无需在后处理步骤中提取潜变量的取值；与恰当的惩罚函数结合时，还能兼顾预测准确性与

稀疏因子解的优势。

在下文中，我们将介绍协同过滤，并进一步探讨它与项目反应理论的关系。随后，我们将讨论参数估计和正则化的作用，并介绍如何通过样本外预测准确率进行模型选择。数据分析包括一项数据模拟，用于说明模型选择和参数恢复。我们还提供了两个真实数据应用案例。第一个案例是一个传统的量表开发问题，我们针对力概念测验（Hestenes、Wells 和 Swackhamer，1992）估计多维项目反应理论模型。第二个案例考察从大规模开放在线课程中收集的数据，并在几分钟内校准最多包含 20 个潜在维度的多维项目反应理论模型。我们在模拟数据集和真实数据集上，将我们的方法与现有的联合最大似然计算程序进行比较。我们还通过一种利用该大规模开放在线课程辅助数据的自动化方法，获得了支持所学习的 20 维模型效用的收敛性证据。

一类用于评估数据的协同过滤模型

协同过滤

协同过滤通常用于所谓的推荐系统。其目标是根据其他受试者对项目的评分，以及相关受试者此前的评分信息，向其推荐不熟悉的项目（Su & Khoshgoftaar, 2009）。例如，Netflix 奖使电影推荐问题受到广泛关注（Bennett & Lanning, 2007）。协同过滤背后的基本思想是，当许多受试者与彼此重叠的项目子集发生交互时，这些信息便可用于推断潜在的新交互。在实践中，我们观察到一个$m \times n$反应矩阵 $\boldsymbol{U} = \{U_{ij}\}$，其各行表示受试者 i 对每个项目 j的反应向量。在通常的情形下，评分是连续值，例如 0 到 5 之间的小数。对于二元评分，假设交互$U_{ij} \in \{0, 1\}$ 对所有 i 和 j均成立。然而，所有可能交互中只有一小部分

已被观察到，因此 $\boldsymbol{U}$ 是稀疏的。协同过滤要解决的问题是预测缺失条目。因此，在通常的情形下，可以对反应矩阵 $\boldsymbol{U} \simeq \boldsymbol{\Theta}\boldsymbol{X}^T$ 进行因子分解，其中 $\boldsymbol{\Theta}$ 是一个 $m \times r$ 矩阵，表示所有受试者的潜在因子； $\boldsymbol{X}$ 是一个 $n \times r$ 矩阵，表示所有项目的潜在因子。$\boldsymbol{\Theta}$ 和 $\boldsymbol{X}$ 可利用数据通过交替最小二乘法或随机梯度下降进行学习（Koren & Bell, 2015），随后二者的乘积可用于填补 $\boldsymbol{U}$中的缺失数据。

对于二元评分，可以通过选择一个链接函数来解决这个问题。一个方便的选择是逻辑函数 σ，

$$\sigma(z_{ij}) = \frac{1}{1 + e^{-z_{ij}}}, \tag{1}$$

其中，参数 z_{ij} 旨在捕捉人物 i 和项目 j 的特征或协变量，这些特征或协变量与预测二者交互的结果有关。 z_{ij} 的函数形式以及可用的协变量取决于具体应用。

本文的基本观点是，基于模型的协同过滤与项目反应理论之间存在天然的共鸣，即使协同过滤和项目反应理论的目标看起来有根本区别。从当代使用项目反应理论进行量表开发的心理测量学家的角度来看，潜在因子结构可能比对未观测反应的预测更为重要，而后者的价值仅居于次要地位。反过来，协同过滤的应用，例如预测电影评论，主要关注预测，而对因子的解释则只居于次要地位。尽管这些方法显然各自更偏重其中一个目标，但我们注意到，解释模型理应具有预测能力，而预测模型也可能为解释性洞见打开大门。如引言中所述，Lord 在其关于项目反应理论的开创性著作（1980年）中，将潜在结构的作用与预测价值视为相互交织。当代教育工作者往往认为，掌握复杂学科需要具备多种技能，而这些技能的发展速度可能各不相同。预测模型或许能为这种结构提供线索，进而可以.....

使用标准的探索性和验证性方法进行细化。下面我们将通过一个示例来说明：利用从作业数据中学习到的高维模型，预测学生在开卷考试期间会参考哪些项目。

用于评估数据的潜因子模型

在协同过滤中，潜因子模型通常称为奇异值分解（SVD）模型（Billsus 和 Pazzani，1998年）。这类模型通常采用回归形式，而非分类形式。

为获得反应矩阵 $\boldsymbol{U}$ 的概率模型，令 $\boldsymbol{\theta}_i \in \mathbb{R}^r$ 为表示个体 i 特征的向量，$\boldsymbol{x}_j \in \mathbb{R}^s$ 为表示项目 j 特征的向量。对于 $\boldsymbol{U}$ 的每个元素，定义

$$p_{ij} = \Pr\{U_{ij} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{x}_j\} = \sigma(z_{ij}) \tag{2}$$

其中 σ 是方程（1）中的逻辑函数（而 z_{ij} 是 $\boldsymbol{\theta}_i$ 和 $\boldsymbol{x}_j$ 的函数）。此外，我们还要求在给定 $\boldsymbol{\theta}_i$ 和 $\boldsymbol{X}_j$ 的条件下，U_{ij} 独立同分布，从而得到如下因子分解

$$\Pr\{\boldsymbol{U} = \mathbf{1} \mid \boldsymbol{\Theta}, \boldsymbol{X}\} = \prod_{i,j} p_{ij}. \tag{3}$$

其中 $\boldsymbol{\Theta} = (\boldsymbol{\theta}_i)_{1 \le i \le m}$，$\boldsymbol{X} = (\boldsymbol{x}_j)_{1 \le j \le n}$，且 $\mathbf{1}$ 是一个与 $\boldsymbol{U}$ 阶数相同的全 1 矩阵。请注意，“个体参数”和“项目参数”的参数复杂度都会随 $\boldsymbol{U}$ 的相应维度增长。

方程（3）中的因子分解，可以参照 Lord 和 Novick（1968）关于潜变量局部独立性的讨论来理解。他们的基本论点是，$\boldsymbol{\theta}_i$ 的维度总可以选得足够大，从而满足局部独立性。这一依据适用于诸如本文所提出的预测应用，而探索性因子分析中的维度选择也基于同样的基本依据，

（Bartholomew、Knott 和 Moutsaki，2011）。

为简化这种情况，我们定义列向量

$$\boldsymbol{\theta}_i = (1 \;\; \theta_{i0} \;\; \theta_{i1} \;\; \cdots \;\; \theta_{ir}\,)^T \qquad \boldsymbol{x}_j = (x_{j0} \;\; 1 \;\; x_{j1} \;\; \cdots \;\; x_{js})^T$$

并要求 求 $r = s$ ，使得 $\boldsymbol{\theta}_i$ 和 $\boldsymbol{X}_j$ 的维度适用于内积 ct

$$z_{ij} = \langle \boldsymbol{\theta}_i, \boldsymbol{x}_j \rangle = x_{j0} + \theta_{i0} + \sum_{k=1}^{r} \theta_{i\,k}\; x_{j\,k}. \tag{4}$$

将模型写成一般形式，可得:

$$\Pr\{U_{ij} = 1 \mid \boldsymbol{\theta}_i, \boldsymbol{x}_j\} = \sigma\left(x_{j0} + \theta_{i0} + \sum_{k=1}^{r} \theta_{ik} x_{jk}\right) \tag{5}$$

这种对数几率 z_{ij}的选择，结合方程（1）–（3），定义了本文所研究的模型类别。对于 r的每个取值，我们允许题目截距（x_{j0}）、个体截距（θ_{i0}）或二者均为零。通过选择 r 以及决定是否纳入截距项，可以得到具体模型。总和 z_{ij} 的可识别性将在下一节讨论。

表1对模型空间进行了编码，并介绍了一些具体模型。对于每个模型，我们将个体变量标记为一个有序对，其中第一项表示 $\boldsymbol{\theta}$ 的维度，第二项表示是否使用截距项。题目变量也采用类似的记法。为清晰起见，对数几率列描述了 z_{ij}的形式，其中省略了个体和题目下标。这些模型以斜截式呈现，因此题目截距映射为维度 k 的常用难度参数，即 $\beta_j = -x_{j0}/x_{j1}$ (Reckase，2009)。

同时也说明了对数几率与一些项目反应理论模型之间的对应关系。Holland（1990）描述了个体独立模型，该模型确保每个项目的正确作答比例都能被完美拟合，但不依赖于 $\boldsymbol{\theta}$。它的

TABLE 1.模型示例及其参数数量

# $\boldsymbol{\theta}$ 参数	# $\boldsymbol{x}$ 参数	对数几率	模型名称
0	1	x_0	个体独立模型
1	0	θ_0	项目独立模型
1	1	$x_0 + \theta_0$	Rasch模型（单参数逻辑模型）
1	2	$x_0 + \theta_1 x_1$	2PL
2	1	$\theta_0 + \theta_1 x_1$	非标准
r	$r+1$	$x_0 + \sum_k \theta_k x_k$	M2PL
$r+1$	r	$\theta_0 + \sum_k \theta_k x_k$	非标准

注释 缩写 1PL 和 2PL 分别指单参数和双参数逻辑模型，M2PL 则指多维双参数逻辑模型。

对立模型是一个此前未命名的模型，我们称之为项目独立模型。该模型确保每个人的正确作答比例都能被完美拟合，但不依赖于 $\boldsymbol{x}$。从某种意义上说，这两个模型都乏味，因为它们并未对一阶边际分布之外的依赖关系进行建模。Rasch模型由这两个独立模型结合而成。文中还提到了双参数逻辑模型（双参数逻辑模型）和 Reckase多维双参数逻辑模型（多维双参数逻辑模型），以及它们各自未命名的对立模型。

表1所述的模型空间具有与探索性因子分析类似的嵌套结构。例如，两个独立模型都嵌套于Rasch模型中。Rasch模型嵌套于双参数逻辑模型中，也嵌套于双参数逻辑模型的非标准对应模型中。对于多维模型，将 r 视为模型的阶数，可以看出，所有阶数为 r 的模型都嵌套于所有阶数为 $r+1$或更高的模型中。由于模型空间具有嵌套结构，模型的阶数也可用作衡量其参数复杂度的指标。

我们尚未为通过公式（4）定义的任何模型提供实质性的依据。从机器学习的角度来看，目标只是选择对项目反应具有最佳样本外预测能力的模型。因此，模型选择问题分为两个步骤：训练和验证。训练是指估计模型参数，将在下一节讨论。验证用于

为每个模型选择正则化调优参数，同时也用于选择 r 的最佳取值；验证将在后续章节讨论。

参数估计与正则化

设 $\boldsymbol{\zeta}$ 为包含 $\boldsymbol{\theta}_i$ 和 $\boldsymbol{X}_j$ 的随机向量。$\boldsymbol{\zeta}$ 的长度受 $(m+n) \times (r+1)$ 的限制，具体取决于项目、受试者或二者是否存在截距项。我们还令 $f(\boldsymbol{\zeta})$ 表示 $\boldsymbol{\zeta}$ 的概率密度函数，并令 $p_{ij}(\boldsymbol{\zeta}) = \Pr\{U_{ij} = 1 \mid \boldsymbol{\zeta}\}$。于是，观测反应矩阵 $\boldsymbol{U}$ 的模型隐含分布可写为

$$L(\boldsymbol{U}) = \int \prod_{i,j} p_{ij}(\boldsymbol{\zeta})^{U_{ij}} (1 - p_{ij}(\boldsymbol{\zeta}))^{(1-U_{ij})} f(\boldsymbol{\zeta})\, d\boldsymbol{\zeta}. \tag{6}$$

公式（6）中的积分无法解析求解，因此任何需要对模型参数（或其子集）进行边缘化的估计策略都涉及数值积分。Cai（2010）回顾了将项目参数视为固定参数（即非随机参数）时可用的方法。当项目和受试者的参数均视为随机参数时，Cho 和 Rabe-Hesketh（2011）从贝叶斯视角对这一情形进行了讨论。在搜索大型模型空间时，这些方法在计算上并不具有吸引力，因此我们寻求一种折中方案。

传统上，人们使用条件最大似然来避免公式（6）中所涉及的计算难题。一般而言，条件最大似然指任何使用数据的条件密度代替其边缘密度的最大似然估计量（Palmer, 2004）。在项目反应理论文献中，“条件最大似然”这一名称专指 Andersen（1970）提出的方法，而基于公式（2）的估计则称为联合最大似然。尽管联合最大似然并不受欢迎，正如引言中所述，我们仍主张在当前应用中使用它，因为对于眼前的问题而言，这是一种计算效率更高的方法。

JML函数的对数为

$$\ell(\boldsymbol{\zeta};\boldsymbol{U}) = \sum_{i,j} \left(U_{ij} \log p_{ij}(\boldsymbol{\zeta}) + (1 - U_{ij}) \log(1 - p_{ij}(\boldsymbol{\zeta}))\right). \tag{7}$$

与边际方法不同，$\boldsymbol{\zeta}$ 被视为待估计的参数向量，而非随机变量。这种方法在应用中仍存在一些困难。首先，由于 $\boldsymbol{\zeta}$ 是模型参数的双线性函数，ℓ 并非凹函数，这使得方程（7）的数值优化更加复杂。边际方法避免了这一问题，因为它们只需对项目参数进行数值优化。针对双线性逻辑回归，研究者已提出类似于 Birnbaum（1968）最初提出的分块坐标下降法，但这些方法只能保证收敛到局部极大值（Shi、Xu 和 Baraniuk，2014）。Cai（2010）建议根据项目四分相关矩阵的初始探索性因子分析结果，为项目参数选择初始值，以解决多极大值问题。这种初始化方法同样适用于当前方法，但随机初始化似乎效果相当。就收敛时间而言，优化参数所需时间略有增加，但省去了探索性因子分析阶段的时间。

第二个困难与旋转不确定性有关。众所周知，z_{ij} 可通过 $k\times k$ 可逆矩阵 $\boldsymbol{T}$ 进行旋转（例如，Browne，2001；Reckase，2009）。因此，在 k维情况下，总共有 k^2 个约束。通常通过要求 COV$\{\boldsymbol{\theta}\} = \boldsymbol{I}$ 来施加$k\left(k+1\right)/2$ 个约束，这意味着正交性$\boldsymbol{T}^{-1} = \boldsymbol{T}'$。其余的 $k\left(k-1\right)/2$个约束则施加于 $\boldsymbol{X}_j$。通常，在模型估计完成后，会进一步寻求一种具有可解释性的选择。

在机器学习文献中，处理模型复杂度问题的一种方法是采用稀疏度（针对模型，而非数据）的思路（Lan, Waters, Studer, & Baraniuk, 2013；Shi 等，2014；Sahebi, Lin, & Brusilovsky, 2016；Doan & Sahebi, 2019）。其思路是使用标准的

回归方法，对非零模型参数进行惩罚（Tibshirani, 1996；Fan & Li, 2001）。在当前应用中，我们从公式（7）中减去一个惩罚项。于是，要最大化的目标函数变为：

$$\ell(\boldsymbol{\zeta};\boldsymbol{U}) = \sum_{i,j}\left(U_{ij}\log p_{ij}(\boldsymbol{\zeta}) + (1-U_{ij})\log(1-p_{ij}(\boldsymbol{\zeta}))\right) - \lambda\left(\sum_i ||\boldsymbol{\theta}_i||_2^2 + \sum_j ||\boldsymbol{x}_j||_2^2\right) \quad (8)$$

其中，$||\cdot||_2^2$ 表示平方 L^2 范数，λ 是一个通过交叉验证确定的超参数。采用这种方法，除个体和题目截距外，每个非零模型参数都会对惩罚似然函数产生贡献。然而，这并不能保证所得参数化不存在旋转不确定性。具体而言，显然，符号变换不会影响惩罚项。涉及交叉乘积的惩罚项（例如用于探索性因子分析中的因子旋转的惩罚项（Browne，2001））并非凹函数；根据我们的经验，这会使数值优化变得复杂。我们的目标与 Chen 等人（2019a）中的约束联合最大似然估计类似。该文作者选择在一个紧集$\forall i,\ j,\ ||\boldsymbol{\theta}_i||^2 \leq C^2$, $||\boldsymbol{X}_j||^2 \leq C^2$ 中施加优化约束，其中 $C = 5\sqrt{r}$ 是预先指定的，但可以调整。由于我们使用的是惩罚似然而非约束参数空间，因此无需在每一步都将参数投影到可行集上，并且 λ 通过交叉验证选定。

–

度量指标与交叉验证

从预测准确度的角度来看，选择最佳模型时会出现两个问题：似然面的非凸性，以及过于复杂的模型对数据的过拟合。通常使用多次重新启动来解决凸性问题，而调整惩罚超参数则可以解决过拟合问题。下面我们介绍执行这一过程的两种算法。我们首先概述这两种方法，然后在下文中详细说明。

第一种方法是将数据集按行传统地划分为训练集、交叉验证集和测试集。我们将其称为条纹方法。请注意，我们的预测问题不同于简单的分类或回归问题，后者使用特征向量来生成预测。在我们的案例中，这是一个矩阵补全问题，其中一些条目 U_{ij} 已知，而其他条目 U_{ij} 需要推断。由于我们是按行划分数据集的，因此这不同于协同过滤中的通常设置。一旦模型得到训练，即超参数和项目参数得到估计，仍必须执行留出法（例如 n折交叉验证）作为性能测试。留出一些项目，基于剩余项目推导个体参数估计值，对留出的项目进行预测，最后检查这些项目上的性能。对于不同的留出项目集合，重复这一过程；如果使用 n 折，则重复 n 次。

在分层算法中，这项 n 折性能测试会进行两次。首先，在验证集上应用交叉验证，以寻找最佳超参数。然后，在测试集上应用交叉验证。使用训练数据、验证数据和测试折重新拟合模型，直到惩罚联合最大似然收敛。最后，如下文算法细节所述，使用 ROC 曲线下面积（曲线下面积）（Bradley，1997）评估性能。

鉴于预测问题具有个体-项目交互的性质，学习超参数和项目参数的另一种方法是按元素而非按行抽取数据留出。训练和交叉验证同时进行，这可能被视为一种缺点，因为局部极大值的影响与超参数的选择混杂在一起。但从曲线下面积的角度来看，这未必是个问题。在逐元素方法中，每次重新启动时，都会随机抽取一部分矩阵元素留出（视为缺失）。由于个体参数和项目参数都会被估计，因此可以在每次重新启动后评估留出项目上的性能。完成预设次数的重新启动后，对预测性能取平均值

（如同 n 折交叉验证的情况）。最后，与之前一样，将这些值用于一组新的受试者测试集。

对于这两种方法，在第一阶段都有一个验证集可用。因此，只要验证集上的性能评分不再提升，就可以停止惩罚联合最大似然的优化；这种技术被称为提前停止（Prechelt，1998年；Yao、Rosasco 和 Caponnetto，2007年）。

算法形式化

设模型空间记为 $\Omega = (\omega_j)_j$，并将每个模型所考虑的超参数空间记为 ω_j as $\Lambda_j = (\lambda_{ij})_{1 \leq i \leq n_j}$。设有一个估计程序 ESTIM（例如 JMLE），该程序给定反应矩阵 $\boldsymbol{U}$、模型 ω 和超参数向量 $\boldsymbol{\lambda}$，返回项目参数 $\boldsymbol{X}$ 和个体参数 $\boldsymbol{\Theta}$的估计向量。函数形式为：

$$(\boldsymbol{X}, \boldsymbol{\Theta}) \leftarrow \mathrm{ESTIM}(\boldsymbol{U}, \omega, \boldsymbol{\lambda}).$$

参数估计完成后，我们可以进行性能测试 PT，计算 S （通常为 ROC 曲线下面积）在 $\boldsymbol{U}$ 的非缺失元素的预测概率与实际元素之间的得分。

$$S \leftarrow \mathrm{PT}(\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{\Theta}, \omega, \boldsymbol{\lambda}).$$

设有一个分层性能测试 SPT，它接收反应矩阵 $\boldsymbol{U}$、模型 ω 和超参数向量 $\boldsymbol{\lambda}$，通过对 $\boldsymbol{U}$ 中的项目进行 k 折交叉验证，返回预测性能度量 S。也就是说，将 $\boldsymbol{U}$ 的项目（列）划分为 k 个大小相等的集合 $\boldsymbol{U}_{(1)\,val}, \ldots, \boldsymbol{U}_{(k)\,val}$（“折”）；通过最大化似然来估计受试者的能力（ $\boldsymbol{U}$ 的行），每次排除一折中的项目。

在留出折中的项目上估计正确作答的概率，然后计算性能误差。最后，对所有折的性能误差取平均值，参见算法1。目前，交叉验证折数 k 被设为全局参数，而不是 SPT 中的可变参数。因此，

$$S \leftarrow \mathrm{SPT}(\boldsymbol{U}_{tr}, \boldsymbol{U}_{cv}, \omega, \boldsymbol{\lambda}).$$

条纹方法

算法1以伪代码形式描述了一种在模型 ω_j 中调整超参数并在空间Ω内选择最佳模型的算法。对于该算法，假设已将 $\boldsymbol{U}$ 矩阵按行划分为训练集$\boldsymbol{U}_{tr}$、交叉验证集、 $\boldsymbol{U}_{cv}$和测试集 $\boldsymbol{U}_{ts}$ （比例可以例如为50%/25%/25%）。

逐元素方法

在算法1中，估计项目参数、在交叉验证集上调整超参数，最后选择最佳模型之间呈线性递进关系。这些步骤分别需要数据 $\boldsymbol{U}$的一个独立部分，即训练集 $\boldsymbol{U}_{tr}$、交叉验证集 $\boldsymbol{U}_{cv}$和测试集 $\boldsymbol{U}_{ts}$。

在此算法的一种变体中，前两个步骤经过重新安排，只需将 $\boldsymbol{U}$ 一次划分为训练集 $\boldsymbol{U}_{tr}$ 和测试集 $\boldsymbol{U}_{ts}$ 。对训练集 $\boldsymbol{U}_{tr}$ 的拟合实际上基于从训练集中随机抽取的逐元素样本。未抽样的数据构成交叉验证集，并且每次随机重启时都会变化。对于给定模型，首先根据多次重新启动的平均性能确定最佳超参数 λ_j^* 。最后一步与之前相同：使用全部 $\boldsymbol{U}_{tr}$ 重新估计参数，并固定 λ_j^* ，同时在测试集上进行分层性能测试。

算法1 条纹方法

输入： $\boldsymbol{U}_{train}$，$\boldsymbol{U}_{val}$，模型 ω, $\boldsymbol{\lambda}$ 正则化超参数
输出： 验证分数 S

过程 SPT（$\boldsymbol{U}_{tr}$，$\boldsymbol{U}_{cv}$，ω，$\boldsymbol{\lambda}$）
  对于 $n = 1$ 到 N_f **执行** ▷ 对每一折
    $\boldsymbol{X}^{(n)}, \boldsymbol{\Theta}^{(n)} \leftarrow$ ESTIM($\boldsymbol{U}_{tr} \cup \boldsymbol{U}_{cv} \setminus U^{(n)}_{fold}, \omega, \boldsymbol{\lambda}$)
    $S^{(n)} \leftarrow$ PT($U^{(n)}_{fold}$，$\boldsymbol{X}^{(n)}$，$\boldsymbol{\Theta}^{(n)}$，$\omega$，$\boldsymbol{\lambda}$)
  结束循环
  $S \leftarrow 1/N_f \sum_n S^{(n)}$ ▷ 各折的平均分数
  返回 S
结束过程

输入： 候选模型 Ω，超参数 Λ_j 对每个模型 ω_j
输出： 最佳模型 ω^*
过程 BESTMODELSTRIATED(Ω，$(\Lambda_j)j$)
  对 ω_j Ω 中 **执行**
    对 λ_{ij} 中 Λ_j **执行**
      $S^{(i)}_{cv} \leftarrow$ SPT($\boldsymbol{U}_{tr}$，$\boldsymbol{U}_{cv}$，ω_j，λ_{ij})
    结束循环
    $\lambda^*_j \leftarrow \lambda_{ij}, \quad i = \text{argmax}\, S^{(i)}_{cv}$ ▷ 给定 ω_j时的最佳超参数 λ
    $S^{(j)}_{test} \leftarrow$ SPT($\boldsymbol{U}_{tr} \cup \boldsymbol{U}_{cv}, \boldsymbol{U}_{ts}, \omega_j, \lambda^*_j$)
  结束循环
  $\omega^* \leftarrow \omega_j, \quad j = \text{argmax}\, S^{(j)}_{test}$ ▷ 最佳模型，ω^*
  返回 ω^*
结束过程

我们将更详细地说明逐元素过程，从定义开始，最后给出算法 2。

设 $\boldsymbol{U}$ 的维度为 $n \times m$。定义一个与其维度相同的随机矩阵 B，对于给定的概率 p，使得

$$B_{ij}(p) = \begin{cases} 1 & \text{with probability} \quad p \\ 0 & \text{with probability} \quad 1-p. \end{cases} \tag{9}$$

B 可用作从 $\boldsymbol{U}$ 中进行采样的指示矩阵，每次重新启动 ESTIM 时

。我们定义样本矩阵 $\boldsymbol{U}^{(s)}$ 及其补集 $\boldsymbol{U}^{(c)}$

$$\boldsymbol{U}_{ij}^{(s)} = \begin{cases} U_{ij} & \text{if } B_{ij} = 1 \\ \text{NA} & \text{otherwise} \end{cases} \qquad \boldsymbol{U}_{ij}^{(c)} = \begin{cases} U_{ij} & \text{if } B_{ij} = 0 \\ \text{NA} & \text{otherwise.} \end{cases} \tag{10}$$

$\boldsymbol{U}^{(s)}$ 大约有 p 的矩阵元素非缺失（例如 $p = 0.7$）。ESTIM 函数将 $\boldsymbol{U}^{(s)}$ 中缺失的矩阵元素视为随机缺失。对数似然和参数估计值均按通常方式获得。

$$(\boldsymbol{X}, \boldsymbol{\Theta}) \leftarrow \text{ESTIM}(\boldsymbol{U}, \omega, \boldsymbol{\lambda}).$$

我们在补集矩阵元素 $\boldsymbol{U}^{(c)}$ 上检验通过 $\boldsymbol{U}^{(s)}$ 估计的参数的性能。

预测准确度

在介绍应用之前，我们想先谈几句如何将预测准确度用作性能度量。预测准确度，即正确分类为 0 或 1 的人-项目元素所占的比例，是一种简单直观的度量，但也有其局限性。具体而言，给定模型和估计参数后，每个矩阵元素都被视为具有某一期望值的随机抽样。如果矩阵元素的概率估计值达到或超过 0.5，则预测为 1。预测准确度不像交叉熵损失（等价于负对数似然）或均方根误差那样能够反映置信程度。例如，如果实际观察到的响应为错误响应，那么即使预测正确响应的概率为 0.51 或 0.95，受到的惩罚也相同。

重要的是，即使已知“真实”的参数值，预测准确度也

算法 2逐元素方法

输入： 候选模型 Ω，超参数 Λ_j 对于每个模型 ω_j

输出： 最佳模型 ω^*

过程 BESTMODELELEMENTWISE(Ω,$(\Lambda_j)j$)
  对于 ω_j 在 Ω **执行**
    对于 λ_{ij} 中 Λ_j **执行**
      对于 $n = 1$ 至 N_r **do** ▷ 每次重新启动
        $B^{(n)} \leftarrow$ 随机矩阵 $B(p)$
        $\boldsymbol{X}^{(n)}, \boldsymbol{\Theta}^{(n)} \leftarrow$ ESTIM($\boldsymbol{U}_{tr}^{(s,n)}, \omega_j, \lambda_{ij}$)
        $S_{ij}^{(n)} \leftarrow$ PT($\boldsymbol{U}_{tr}^{(c,n)}$, $\boldsymbol{X}^{(n)}$, $\boldsymbol{\Theta}^{(n)}$, ω_j, λ_{ij})
      结束循环
    结束循环
    $S_{ij} \leftarrow 1/N_r \sum_n S_{ij}^{(n)}$ ▷ 逐元素样本的平均分数
    $\lambda_j^* \leftarrow \lambda_{ij}$, $i =$ 最大值参数 S_{ij} ▷ 最佳超参数 λ，给定 ω_j
    $S_{test}^{(j)} \leftarrow$ SPT($\boldsymbol{U}_{tr}, U_{test}, \omega_j, \lambda_j^*$)
  结束循环
  $\omega^* \leftarrow \omega_j, j =$ 使.....最大化 $S_{test}^{(j)}$ ▷ 最佳模型, ω^*
  返回 ω^*
结束过程

当然也不会接近 1，除非每个项目对每个人来说都无限简单或无限困难。预测准确度存在一个可预期的上限，这有助于判断使用高维模型所带来的性能渐进提升是否应被视为显著改进。这个上限无法预先得知（模拟研究除外），因为它取决于模型以及项目参数和个体参数的分布。据我们所知，目前还没有关于项目反应模型预期可达到准确率的解析结果，因此我们在下文给出了推导和示例。

如果期望矩阵中随机变量 P 的分布由分布函数 $f_P(p)$ 给出，那么期望准确率得分可表示为以下“平均值”

“平均值”

$$S = \int_0^{0.5} (1-p) f_P(p) dp + \int_{0.5}^1 p f_P(p) dp \tag{11}$$

其中，第一项表示预测错误的矩阵元素，第二项表示预测正确的矩阵元素。如前所述，$f_P(p)$的形状取决于模型中人员和项目参数的分布以及链接函数。以Rasch模型或单参数逻辑斯蒂模型为例，正确作答的概率服从方程（13）所示的分布。

命题*1.* 设 Θ 和 Δ 是相互独立的随机变量，其概率密度分别为 $f_\Theta(\theta)$和$f_\Delta(\delta)$。给定该模型，

$$P = g(\Theta, \Delta) = \frac{1}{1 + e^{-(\Theta - \Delta)}}, \tag{12}$$

then

$$f_P(p) = \frac{1}{p(1-p)} \int_{-\infty}^{\infty} f_\Theta(\theta) f_\Delta \left(\theta + \ln \frac{1-p}{p} \right) d\theta. \tag{13}$$

证明. 根据定义，概率密度

$$f_P(p) = \frac{d}{dp} \Pr[P \le p] = \frac{d}{dp} \iint_{D_p} f_\Theta(\theta) f_\Delta(\delta) d\theta d\delta, \tag{14}$$

其中，$D_p = \{(\theta, \delta) \mid g(\theta, \delta) \le p\}$ 是(θ, δ)-平面上的积分区域。对方程（12）求逆，可将此约束写为

$$\delta \ge \theta + \ln \frac{1-p}{p}. \tag{15}$$

TABLE 2.预期准确率估计示例。

θ	β	预期准确率 S
$\mathcal{N}(0,\ 1)$	$\mathcal{N}(0,\ 1)$	0.7252
	$\mathcal{N}(0,\ 2)$	0.7946
	$\mathcal{N}\ (-5,\ 1)$	0.9833
	$\mathcal{N}\ (5,\ 1)$	0.9833

将该约束并入积分限,

$$f(p) = \frac{d}{dp}\int_{-\infty}^{\infty}\int_{\theta+\ln\frac{1-p}{p}}^{\infty} f_\Theta(\theta)f_\Delta(\delta)d\theta d\delta \tag{16}$$

$$= \frac{1}{p(1-p)}\int_{-\infty}^{\infty} f_\Theta(\theta)f_\Delta\left(\theta+\ln\frac{1-p}{p}\right)d\theta \tag{17}$$

w这里对 β 积分号下的表达式求导时，使用了莱布尼茨积分法则,

$$\frac{d}{dz}\left(\int_{a(z)}^{b(z)} f(x,z)\,dx\right) = \int_{a(z)}^{b(z)}\frac{\partial f}{\partial z}f(x,z)\,dx + f(b(z),z)\frac{db}{dz} - f(a(z),z)\frac{da}{dz}. \tag{18}$$

□

给定个体参数和项目参数的分布，便可据此推导出期望准确率得分。表2展示了一些简单示例，说明在一维模型中，个体参数和项目参数均服从正态分布时的预期准确率估计。我们发现，对于过于简单或过于困难的测验，准确率都非常高；但对于项目难度与受试者能力匹配良好的测验，预期准确率则明显较低。此处举例旨在说明，高预测准确度并非仅由估计算法决定。与测验对总体中的应试者来说过于简单或过于困难的情况相比，当项目难度与个体能力相匹配时，测验结果必然更难预测。此外，除非纯属偶然，预测准确度不可能超过这些理论上限。

数据应用

我们首先进行了一些小规模模拟研究，以展示如何使用所提出的方法进行模型选择和参数恢复，并与 R 中的 JML软件包（Chen 等人，2019a，2019b）进行初步比较。这些模拟研究旨在说明，在相对简化的情境下，所提出的方法能够按预期发挥作用。我们并未尝试对多维项目反应理论的各种竞争方法进行大规模比较。在模拟之后，我们给出两个真实数据示例。第一个是将多维项目反应理论应用于物理概念多维评估的标准案例。尽管该方法主要用于预测，我们仍表明，所学习模型的高维性与采用因子分析方法的相关研究结果一致。第二个案例则将该方法应用于从一门大规模开放在线课程中收集的大型稀疏数据集。在此，我们对预测模型应用了一种新颖的验证方法。具体而言，我们展示了如何以“推荐式”的方式使用该方法，识别哪些作业题目与考试题目具有内容相关性。最后这项分析使用了同一门在线课程中一次开卷考试的辅助数据。

模拟研究：多单维测验

我们首先介绍一项模拟研究，其中使用三维补偿型项目反应理论模型（3d2PL），生成仅在一个区分度维度上载荷的项目的作答：也就是说，这是一项多单维测验。从多元正态分布 $\mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ 中为 1000 名模拟受试者抽取能力，其均值为零，协方差为 $\boldsymbol{\mu} = \mathbf{0}$ 和 $\boldsymbol{\Sigma} = I_3$。模拟了 60 个项目，截距参数从区间 $[-2.5, 2.5]$ 中均匀抽取；区分度则针对一个随机索引服从 $\ln \mathcal{N}(0,\ 0。5)$ 分布（其他所有区分度均为零）。

使用模拟的反应矩阵，通过上述算法探索了维度为 $\{1, 2, 3, 4, 6, 9\}$ 的多维双参数逻辑模型 Ω。正则化参数集合

超参数为每个模型 Ω $\Lambda_j = \{0.02, 0.04, 0.06, 0.08\}$ $\omega_j \in$ 设定。参数优化采用 Adam（Kingma & Ba，2015年），这是一种带有自适应学习率的随机梯度下降变体，学习率初始化为 $\gamma = 0.1$。为节省内存并加快收敛速度，我们在训练集的 10 个批次上计算梯度。交叉验证采用 5 折，每次留出 12 个项目进行测试，并选择 ROC 曲线下面积（AUC）最高的模型。为衡量性能，我们报告准确率（ACC）、AUC、Goodman and Kruskal 的兰姆达（1954）（Λ_{GK}；误差比例缩减的度量）以及均方根误差（RMSE）。

结果见表3和表4。我们注意到，条纹方法和逐元素方法的预测性能相近。两种情况下，一维模型的表现都明显较差，二维模型略好，而维度为3及以上的模型表现相差不大。这表明，对数据欠拟合（即使用过于简单的模型）造成的性能下降，比对数据过拟合（即使用过于复杂的模型）更为严重。此外，使用已知的（模拟）参数时，准确率的解析上界为0.7664。因此，多维模型的表现几乎达到了最优水平。

TABLE 3.使用条纹方法对模拟数据集进行探索性分析的结果

dim	λ	ACC	AUC	Λ_{GK}	RMSE
3	0.04	**0.7452**	**0.8141**	**0.4013**	**0.4149**
4	0.04	0.7422	0.8111	0.3948	0.4161
6	0.06	0.7413	0.8107	0.3918	0.4204
9	0.06	0.7416	0.8107	0.3924	0.4206
2	0.02	0.7343	0.7993	0.3755	0.4205
1	0.06	0.7177	0.7761	0.3388	0.4346

在一台配备 2016 年款 2.6 GHz Intel Core i7 处理器的计算机上，条纹方法每个模型耗时 46 秒，逐元素方法耗时 47 秒。作为比较，对于一个规模较小、受试者数量减半（500）、项目数量为三分之一的数据集

TABLE 4.使用逐元素方法对模拟数据集进行的探索性结果

dim	λ	ACC	AUC	Λ_{GK}	RMSE
3	0.04	**0.7433**	**0.8116**	**0.4025**	**0.4170**
9	0.06	0.7400	0.8085	0.3951	0.4222
4	0.06	0.7394	0.8083	0.3936	0.4223
6	0.06	0.7390	0.8078	0.3930	0.4224
2	0.02	0.7323	0.7967	0.3793	0.4219
1	0.04	0.7183	0.7776	0.3472	0.4330

（20），以及 $d = 6$，mirt 软件包（Chalmers，2012）耗时 100 秒，而我们的实现耗时 2 秒。mirt 软件包采用 Metropolis-Hastings-Robbins-Monro 算法（Cai, 2010），该算法不假设项目参数服从任何先验分布，并使用马尔可夫链蒙特卡洛对个体参数进行采样，同时优化项目参数。我们的实现可在 GitHub[1],上获取，使用 TensorFlow。

参数恢复

探索性搜索正确识别了生成数据的模型维度（$d = 3$）。现在我们来考察参数本身是否得到了恢复。可识别性仍然是一个问题，因为即使加入了 L_2 正则化项，模型仍存在旋转不变性。可以使用标准程序对估计能力进行正交化，该程序采用协方差矩阵奇异值分解得到的投影矩阵。完成这一操作后，正交能力向量似乎与模拟参数清晰对应，仅整体符号可能相反，见表5。估计值与真实值的各对相关系数绝对值介于0.83至0.84之间。方差比例及相应的决定系数见图1。

[1]https://anonymous.4open.science/r/913f850f-e12c-45b0-a897-41782504fb59/

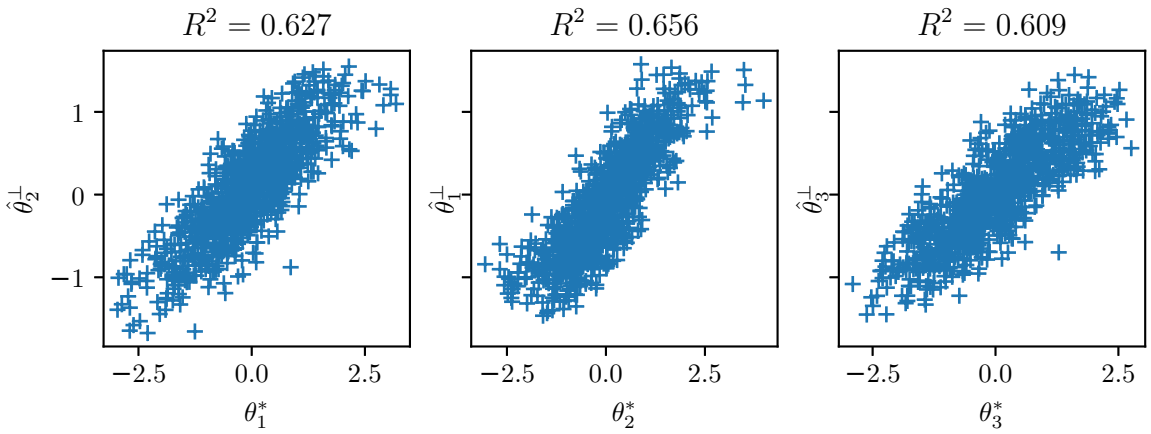


FIGURE 1.估计参数解释的方差比例，例如 ˆ $\theta_1^{\perp}$与 θ_2^*呈正相关。

TABLE 5.正交化后估计能力与模拟值的相关性

	$\hat{\theta}_1^{\perp}$	$\hat{\theta}_2^{\perp}$	$\hat{\theta}_3^{\perp}$	θ_1^*	θ_2^*	θ_3^*
$\hat{\theta}_1^{\perp}$	1.00	-0.02	0.00	0.26	**0.84**	-0.14
$\hat{\theta}_2^{\perp}$	-0.02	1.00	0.00	**-0.83**	0.24	0.26
$\hat{\theta}_3^{\perp}$	0.00	0.00	1.00	0.23	0.10	**0.83**
θ_1^*	0.26	-0.83	0.23	1.00	0.02	-0.06
θ_2^*	0.84	0.24	0.10	0.02	1.00	0.01
θ_3^*	-0.14	0.26	0.83	-0.06	0.01	1.00

总之，在使用不相关能力和多单维测验进行的模拟研究中，我们的探索性模型能够在总运行时间约为145秒的情况下，恢复正确的模型维度。经过正交化后，能力估计值与模拟值高度相关。图1显示，潜在特质估计值出现了明显的收缩；这在正则化情境中是意料之中的现象，使用潜在特质先验分布的项目反应理论评分方法也会出现这种情况（Reckase，2009）。

TABLE 6.与使用不同规模和稀疏度的模拟数据集的受约束的JMLE方法进行比较

Data	受试者数	项目数	稀疏度	受约束的JMLE（Chen 等人，2019a）			我们的非正则化方法 2 个批次		
				LL	Acc	AUC	LL	Acc	AUC
模拟的三维	1000	60	0%	-24085	0.79	0.87	-22530	0.81	0.89
模拟的五维	80	20	0%	-354	0.91	0.97	-244	0.95	0.98
			20%	-205	0.94	0.99	-76	0.98	1
			80%	-1.11	1	1	-0.02	1	1
模拟的五维	1000	50	0%	-18187	0.83	0.92	-17777	0.83	0.92
			20%	-13189	0.85	0.93	-12880	0.85	0.93
			80%	-1276	0.95	0.99	-491	0.98	1

模拟研究：不同稀疏度水平

我们使用不同规模和稀疏度水平的模拟数据集，将受约束的JMLE（CJMLE）方法（Chen 等人，2019a）作为基准进行比较。我们采用一个5维补偿型项目反应理论模型，从80名受试者和20个项目，或1000名受试者和50个项目中生成作答。个体能力、项目截距和项目区分度均从标准正态分布中抽样；数据按完全随机缺失的方式缺失，稀疏度为 $p \in \{0, 0.2, 0.8\}$。总共生成了六个此类模拟数据集。我们从似然、准确率、曲线下面积和运行时间等方面，对CJMLE与我们自己的惩罚化联合最大似然估计方法进行了比较。除上一节所述的多重单维数据集外，所有数据集的估计模型规模均为 $d = 5$；对于该多重单维数据集，则为 $d = 3$。模拟数据集的结果见表6。在所有模拟情形下，我们的估计代码都能达到更高的对数似然。稀疏度较高时，这一现象尤为显著，因为当观测到的样本较少时，CJMLE的超参数可能会过度限制参数空间。

相关因子和交叉载荷题项

也可以使用具有因子交叉载荷和相关能力的项目进行模拟。例如，我们使用了一组 30 个三维项目

这些项目在多维项目反应理论（Reckase，2009，表6.1）的经典著作中有所规定，其参数经过设计，既符合实际情况，又具有较高的估计难度。参照同一文献，我们模拟了2000 名受试者的作答，将能力的均值向量设为$[-0.4, -0.7, 0.1]$ ，并将方差-协方差矩阵设为

$$\begin{bmatrix} 1.21 & 0.297 & 1.232 \\ 0.297 & 0.81 & 0.252 \\ 1.232 & 0.252 & 1.96 \end{bmatrix}.$$

在此案例中，能力维度 1 和 3 高度相关（$\rho_{13} = 0.8$）。总体在维度 2 上的平均能力则显著较低。

TABLE 7.使用相关因子的探索性结果，条纹方法

dim	λ	ACC	AUC	Λ_{GK}	RMSE
3	0.08	0.7015	**0.7611**	0.2775	0.4441
6	0.08	**0.7038**	0.7610	**0.2831**	0.4439
9	0.08	0.7016	0.7608	0.2774	0.4440
2	0.06	0.7028	0.7595	0.2809	**0.4414**
4	0.08	0.7005	0.7595	0.2752	0.4445
1	0.06	0.6974	0.7531	0.2670	0.4440

TABLE 8.使用相关因子的探索性结果，逐元素方法

dim	λ	ACC	AUC	Λ_{GK}	RMSE
3	0.08	**0.7039**	**0.7653**	**0.3045**	**0.4437**
4	0.08	0.7003	0.7650	0.2962	0.4443
9	0.08	0.7014	0.7646	0.2986	0.4442
2	0.08	0.7019	0.7645	0.2998	0.4441
6	0.08	0.7007	0.7640	0.2969	0.4444
1	0.04	0.6937	0.7532	0.2803	0.4438

维度为 $\{1, 2, 3, 4, 6, 9\}$ 的探索性模型结果见表7-8

其中，超参数是在交叉验证期间于 $\Lambda_j = \{0.02, 0.04, 0.06, 0.08\}$ 中选定的。当因子高度相关且存在交叉载荷题项时，真实的三维模型未必会被选为拟合最佳的模型，信号也不那么清晰。采用条纹抽样时，三维模型的 AUC分数最高，但二维嵌入的均方根误差较低，而六维模型的准确率和古德曼-克鲁斯卡尔 λ 较高。有趣的是，逐元素抽样算法在所有拟合优度指标上都将三维嵌入选为最佳模型。Reckase（2009）选择这些生成参数，是为了挑战当时可用的估计软件。事实上，研究发现，即使采用商业化的基于马尔可夫链蒙特卡罗的三维补偿模型估计，“对该数据集得到的基本上是一个二维解”（第175页）。接下来，我们将结合真实数据集的研究发现，再简要讨论这些结果。

物理概念测验

我们的第一个真实数据集采用了 $N \approx 13000$ 名学生在概念物理领域的力概念测验（Hestenes 等人，1992）中对 30 个项目的作答。缺失条目极少，不到 0.2%。我们纳入了维度为$\{1, 2, 3, 5, 7, 9, 20\}$的模型。每个模型的运行时间为 111 秒，其中在学得最佳超参数后只需 22 秒。结果见表9和表10。

TABLE 9.使用条纹方法对FCI数据集进行的探索性结果

维度 d 最佳正则化项 λ	ACC	AUC	Λ_{GK}	RMSE
20 **0.05**	**0.7409**	**0.8091**	**0.4022**	**0.4191**
9 **0.05**	**0.7399**	**0.8087**	**0.4000**	**0.4192**
5 0.05	0.7402	0.8079	0.4005	0.4199
7 0.05	0.7384	0.8069	0.3964	0.4203
3 0.05	0.7367	0.8053	0.3925	0.4215
2 0.02	0.7343	0.8026	0.3872	0.4216
1 0.02	0.7250	0.7934	0.3654	0.4266

TABLE 10.使用逐元素方法对FCI数据集进行的探索性结果

维度 d 最佳正则化项 λ	ACC	AUC	Λ_{GK}	RMSE
9	**0.05**	**0.7422**	**0.8132**	**0.4349**	**0.4177**
20	**0.05**	**0.7422**	**0.8127**	**0.4347**	**0.4180**
7	0.05	0.7395	0.8115	0.4292	0.4187
5	0.05	0.7400	0.8110	0.4300	0.4190
3	0.05	0.7384	0.8092	0.4263	0.4203
2	0.05	0.7373	0.8078	0.4240	0.4218
1	0.01	0.7268	0.7965	0.4015	0.4252

对于这两种方法，包含 $d = 9$ 和 $d = 20$ 的模型性能最高且相近，超参数为 $\lambda = 0.05$。用 20 个维度对 30 个项目建模，比例已危险地接近 1:1，我们几乎不会建议对每个项目单独建模。此处纳入$d = 20$ 模型有两个原因。首先，它表明使用这种方法可以处理高维性。其次，也是更重要的一点，它证明正则化过程按预期发挥作用。也就是说，样本外预测结果会随着模型复杂度的提高而改善，直到某个程度后趋于平稳。经过适当正则化的 $d = 20$ 模型不会明显过拟合数据。鉴于改进微乎其微，也有理由在九个、甚至五个潜在维度时停止。值得注意的是，其他作者近期采用传统探索性方法独立发现，FCI 可用九因子模型表示（Stewart、Zabriskie、Devore 和 Stewart，2018）。

大规模稀疏大规模开放在线课程数据集

该大规模开放在线课程数据集包含 $N \approx 30000$ 名学生回答197道作业题目的作答记录，总计200万条数据。然而，由于许多大规模开放在线课程的注册者只是偶尔参与，回答的作业题目很少，因此该数据集的稀疏度为64%（Seaton、Bergner、Chuang、Mitros和Pritchard，2014）。尽管学生

有多次机会回答家庭作业题目，但采用“最终答对”评分所得的结果方差很小，因此通常更有参考价值的是根据每个项目首次作答的正确性进行评分（Bergner、Colvin和Pritchard，2015）。我们在此采用了这种方法。我们估计了维度为 $\{1, 5, 10, 15, 20\}$的模型。每个模型的运行时间为7.5分钟，其中，在学习到最佳超参数后需要88秒。结果见表11和表12。

TABLE 11.使用条纹方法对大规模开放在线课程数据集进行的探索性结果

维度 d	最佳正则化项 λ	ACC	AUC	Λ_{GK}	RMSE
20	**0.02**	**0.7648**	**0.8323**	**0.3935**	**0.4005**
15	0.02	0.7620	0.8293	0.3864	0.4022
10	0.02	0.7572	0.8233	0.3744	0.4058
5	0.01	0.7481	0.8140	0.3514	0.4113
1	0.05	0.7290	0.7898	0.3020	0.4263

TABLE 12.使用逐元素方法对大规模开放在线课程数据集进行的探索性结果

维度 d	最佳正则化项 λ	ACC	AUC	Λ_{GK}	RMSE
20	**0.02**	**0.7641**	**0.8333**	**0.3974**	**0.4005**
15	0.02	0.7619	0.8305	0.3916	0.4022
10	0.02	0.7563	0.8239	0.3774	0.4062
5	0.02	0.7478	0.8148	0.3555	0.4114
1	0.01	0.7260	0.7862	0.3003	0.4248

对于此数据集，$d = 20$ 其中 $\lambda = 0.02$ 是表现最佳的模型。该大规模开放在线课程数据集包含种类更多的项目（$n = 197$），这些项目在14周内施测，参与者群体也十分多样，包括来自多个国家和机构的学生，以及教师、专业人士和爱好者。我们的应用结果表明，可能需要采用高维模型来解释这类大规模数据集中的方差。

TABLE 13.使用真实数据集与受约束的JMLE方法进行比较

Data	受约束的JMLE (Chen 等人，2019a)				我们的非正则化方法 2 个批次				我们的正则化方法 $\lambda = 0.02$，10 个批次			
	LL	Acc	AUC	Time	LL	Acc	AUC	Time	LL	Acc	AUC	Time
力学概念测验训练集	-116369	0.84	0.93	57″	**-113792**	0.84	0.93	42″	-140406	0.82	0.91	12″
力学概念测验测试集	-31181	0.74	0.76		-49723	0.73	0.75		**-21461**	0.75	0.77	
大规模开放在线课程训练集	-773746	0.79	0.88	33′18″	**-771703**	0.79	0.88	5′12″	-843627	0.79	0.86	2′58″
大规模开放在线课程测试集	-108280	0.74	0.79		-109805	0.74	0.8		**-104761**	0.74	0.80	

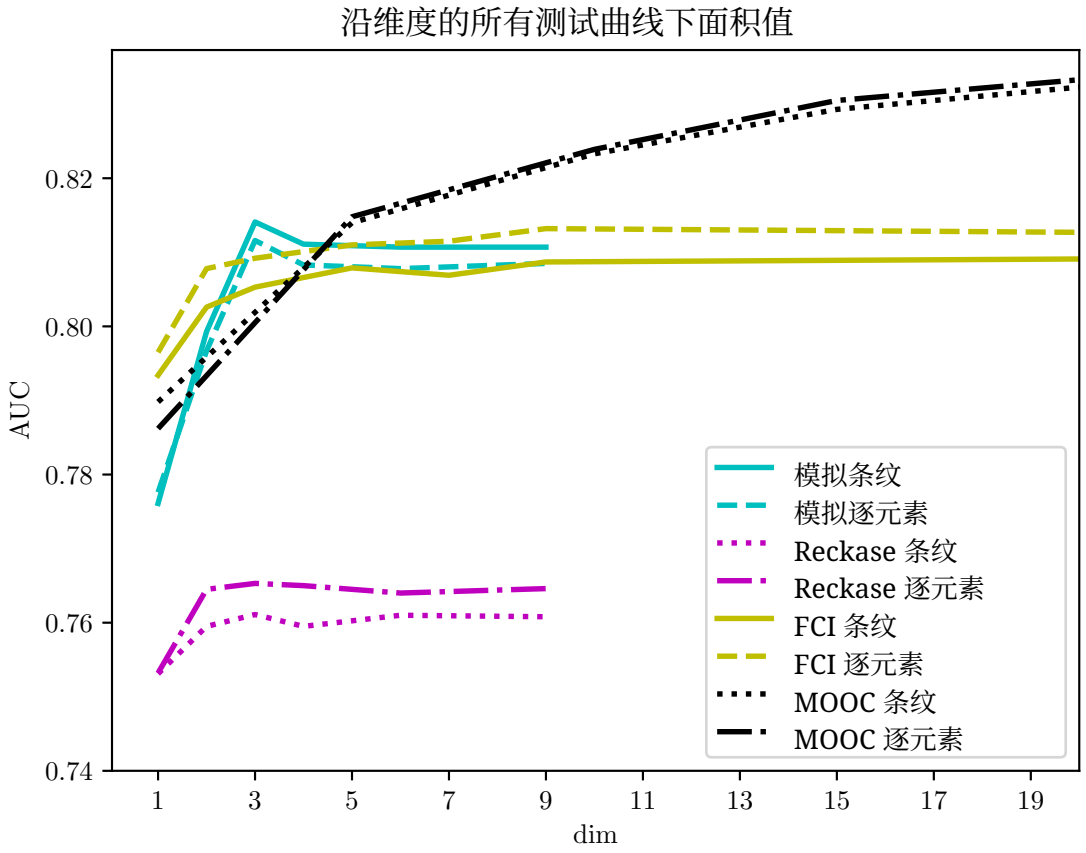
我们在图2中绘制了所有数据集探索性结果中曲线下面积统计量的变化。对于模拟多重单维数据集，性能在真实维度 $d = 3$处达到明显峰值。此后，随着维度增加，性能略有下降并趋于平稳。对于相关因子和交叉载荷题项的模拟响应（标记为“reckase”的曲线），我们观察到整体准确率较低。模型的曲线下面积逐渐增加至 $d = 3$ ，随后趋于平稳，期间仅有小幅波动。力概念测验的性能趋势十分相似，曲线下面积在超过 $d = 9$后逐渐趋平并保持稳定。大规模开放在线课程数据集的结果似乎在 $d = 20$ 处仍在上升，但已有收益递减的迹象。

性能与运行时间

对于模拟数据集，我们在表13中纳入了针对真实数据集的约束联合最大似然估计（CJMLE）基准比较。对于所有考察的模型， $d = 5$。

我们发现，与基准方法相比，正则化能更好地重建缺失数据：未正则化模型在训练集上的似然高于CJMLE，但在测试集上并非如此——换句话说，它发生了过拟合。由于超参数 λ 是通过交叉验证选定的，正则化模型在测试集上的似然高于CJMLE。与CJMLE相比，性能提升在高维大规模开放在线课程数据集上尤为明显；在该数据集上，我们的非正则化方法在 ~ 5 分钟内收敛至更高的似然

FIGURE 2.所有数据集的探索性结果



而 Chen 等人（2019a）的 R/C++ 实现（可在 mirtjml 软件包中找到）需要 ~30分钟。所提出的算法速度很快，部分得益于提前停止；此外，mirtjml 会在应用缺失数据掩码之前，将作答模式误差的整个矩阵保存在内存中，而我们的实现则受益于稀疏计算。我们采用批处理，在每次迭代所用内存与停止前的迭代次数之间进行权衡。尽管如此，通过进一步调整CJMLE实现中的容差和约束参数，仍有可能改善收敛时间和交叉验证误差。

使用辅助数据进行验证

在探索性因子分析的标准方法中，人们会考察载荷相似的项目，以识别或定义因子本身。分析人员通常会事先对这种结构有所预期，而这种结构一般是低维的。然而，在大规模开放在线课程数据集中，研究表明，与规模较小的模型相比，20 维潜在空间解的矩阵补全性能有所提升。应如何理解这种高维表示？要考察一个有 20 列、数百行的载荷矩阵，将极具挑战性且十分耗时。本节提出一种可替代且可自动化的方法，用以展示高维模型在教育情境中的效用，即便各个因子本身无法得到清晰解释。与因子分析方法不同，我们的方法受到协同过滤在推荐系统应用中的启发。对于大规模开放在线课程数据，该方法还利用辅助数据——无模型观测——来“验证”载荷相似的项目之间的实质性内容关系。

具体而言，我们利用大规模开放在线课程开卷期中考试期间发生的网页导航事件，确定最常被查阅的先前作业题目。学生查阅这些题目（题目附有解答）这一行为，表明这些题目与考试内容有实质关联。随后，这组热门题目构成一个参照集。另一方面，高维模型可用于生成一组与期中考试中占主导地位的因子重叠度较高的项目。最后，我们对这两组项目进行比较。如果因子重叠法找回了许多学生实际查阅过的相同项目（即热门项目），我们便可以设想将这些信息用于推荐系统。也就是说，例如在备考时，可以向学生推荐这些作业题目。现在，我们来详细介绍辅助数据，以及如何确定用于比较的项目集。

要识别热门参照项目，需要解析每位学生的网页导航日志。该大规模开放在线课程的期中考试包含五道题目，共有 26 个分别计分的题项。为照顾国际时区差异，学生有 24 小时完成考试。在这场考试中，学生可以查阅任何课程资源，包括讲课视频或教材页面。按导航事件的频次计算，数量最多的一类资源实际上是课程前半部分的作业题。一次导航事件由来源页面、目标页面和时间戳组成。对于每位学生，只统计发生在其期中考试“时间窗口”内（从首次访问到最终提交）的事件。总体而言，这些导航事件可视为由课程资源节点构成的网络。从资源 A 指向资源 B 的边具有方向性，其权重由学生在期中考试期间从一个资源导航至另一个资源的次数决定。如果同一名学生多次经过这条路径，则这些次数都会累加到边的权重中。

为了减少转移网络中的噪声，我们将分析范围限制在同时参加期中考试和期末考试的7518名学生。这些学生可能认真学习了这门大规模开放在线课程，而数万名其他学生则只是偶尔使用。即使如此，网络中仍有超过50万次导航事件，涉及1400种独特资源，包括课程讲课视频和在线教材。将该网络筛选为（a）课程评估题项，以及（b）仅保留权重最高的边后，得到17个项目；课程前半部分共包含102个项目。这就是高人气题项集合。我们注意到，边的权重迅速下降，其分布与帕累托分布相似。我们的方法使用了频率截断值；如果采用较低的截断值，参照集就会保留更多项目。我们的比例截断值（所有转移次数的0.05%）旨在选出一个便于处理的集合，规模大致相当于项目总数的前五分之一。

使用完整的课程数据集拟合高维惩罚型多维项目反应理论模型，其中包括期中考试和期末考试题目，以及作业题目。需要注意的是，尽管使用了整门课程的数据来拟合多维

TABLE 14.大规模开放在线课程期中前评估题目交叉列表。受欢迎程度指项目在期中考试中被引用的频率。重叠度指以整场期中考试为基准对因子载荷进行阈值比较。

	高受欢迎程度	低受欢迎程度
高重叠度	11	1
低重叠度	6	84

模型，但约有一半的课程评估题项只有在期中考试之后才可用。由于多个期中题目被编排在同一页面上，因此无法在期中题目层面区分导航事件；我们将期中题目合并起来，以识别最相关的因子。具体而言，先以 1.3 为阈值，将 26 个期中题目的载荷（即区分度）离散化，然后求和。每个因子对应的题目计数介于 0 至 18 之间，我们保留了计数高于 5 这一阈值的八个因子。我们利用这八个因子从其余评估题目池中筛选题目。对于每个因子，均使用相同的区分度阈值来选择题目。总计筛选出 41 个非期中题目。不过，这些题目中只有 12 个来自期中考试前的 102 道作业题目，其余 29 个都在期中考试之后才出现。因此，在期中考试考生可参考的题目中，有 12 道被发现至少在一个期中考试中的主导因子上具有较高重合度。通过此方法发现，期中考试前的 102 道题目中有 90 道与因子的重合度较低。（与之前一样，采用更低的截断值会保留更多因子，从而筛选出更多题目。）

两组项目的比较总结见表14。我们发现，12个高重叠题目中有11个与17个高人气题项的参照集相匹配。如果从推荐的角度将这些视为真正例，那么其余数量可描述为1个假正例、6个假负例和84个真负例。表14中的重叠度和人气标准显然相关。因此，结果确实表明，整个课程响应矩阵的共同因子载荷

有助于识别一组与学生在考试中最常用作参照的项目相似的项目。我们认为，这两种方法得出的结果相互印证，表明这个高维预测模型确实“有用”。

我们在此提出的方法，仅旨在检验多维因子结果在解释意义方面是否合理。我们承认，为了使比较易于开展，该方法在权重和载荷方面采用了一些任意设定的截断值。当然，若不针对这些截断值进行敏感性分析，就无法声称结果具有稳健性；但这超出了我们目前的研究范围。

总结与讨论

我们提出了一种受协同过滤启发、用于拟合多维预测项目反应模型的新型计算方法。我们通过 L_2 正则化项对联合最大似然进行惩罚，以平衡项目参数和个体参数的数量。我们介绍了两种交叉验证估计方法——分层算法和逐元素算法；二者性能相当，运行时间也相近。我们表明，即使观测到的作答数据十分稀疏，该方法仍能有效且高效地估计高维模型的参数。在基准比较中，这些算法表现良好，尤其是在数据集规模较大和/或较为稀疏时。本文主要关注点估计。展望未来，我们指出，该方法还可借助变分推断扩展至贝叶斯框架。

在评估预测性多维模型时，我们提供了两类不同的结果。首先，我们推导出了一些关于常用准确度测量指标预期性能的新分析结果。在比较维度逐渐增加的预测模型时，我们报告了常见的评估度量指标，例如准确率和均方根误差，以及古德曼-克鲁斯卡尔 λ——一种适用于衡量分类变量之间预测关联的指标。其次，在应用于

大规模开放在线课程（MOOC）数据时，我们展示了一个 20 维预测模型从推荐系统视角来看可能具有实用价值。我们注意到，因子结构是根据个体-项目交互的正误评分推导得出的，而热门项目的参照集则基于信息寻求的无模型观测，与正确性完全无关。因此，在 MOOC 案例中观察到的趋同结果表明，预测模型因子载荷所识别出的项目具有实质性的共性。

在心理测量学文献中，20维因子模型并不常见，但在采用协同过滤的机器学习方法中却未必如此。例如，对于教育心理学家而言，作答数据的因子结构通常被假定为简单结构，或被设计成简单结构。在探索性研究中，只有当因子具有可解释性时，才会将其保留在模型中。协同过滤通常用于不同目的，例如应用于为海量稀疏数据集设计的推荐系统。针对这些应用中估计过程的复杂性，人们已取得多项计算方法进展。我们认为，其中一些“技巧”也能惠及心理测量学应用，尤其是使用来自学习环境而非标准化测验或量表的作答数据时。

参考文献

Alldredge, J., & Gilb, N.（1976）。岭回归：一份附注书目。《国际统计评论》，*44*(3)，355–360。Andersen, E.（1970）。条件最大似然估计量的渐近性质。《英国皇家统计学会会刊：*B*辑 *(* . . . ，*32*(2)，283–301。Bartholomew, D. J., Knott, M., & Moutsaki, I.（2011）。潜变量模型与因子分析 （第3版）。伦敦：Arnold。Bennett, J., & Lanning, S.（2007）。Netflix 奖。载于*KDD* 杯及研讨会论文集 （第2007卷，第35页）。Bergner, Y.（2017）。学习分析中的测量及其用途。学习分析手册，35–48。Bergner, Y., Colvin, K., & Pritchard, D. E.（2015）。存在缺失或多次作答时，如何根据作业题目估计能力。第五届国际学习分析与知识会议论文集——*LAK '15*。Bergner, Y., Droschler, S., & Kortemeyer, G.（2012）。基于模型的协同过滤对学生作答数据的分析：机器学习项目反应理论。教育数据挖掘。Billsus, D., & Pazzani, M. J.（1998年）。学习协同信息过滤器。载于国际机器学习大会（第98卷，第46–54页）。Birnbaum, A.（1968）。若干潜在特质模型及其在推断应试者能力中的应用。载于心理测验分数的统计理论。Bock, R. D., & Aitkin, M.（1981）。项目参数的边际最大似然估计：期望最大化算法的应用。《心理测量学》，*46*(4)，443–459。Bradley, A. P.（1997）。ROC曲线下面积在评估机器学习算法中的应用。《模式识别》，*30*(7)，1145–1159。Browne, M. W.（2001年6月）。探索性因子分析中解析旋转概述

分析。《多变量行为研究》，36(1)，111–150。Cai, L. (2010). 使用 Metropolis–Hastings Robbins–Monro算法进行高维探索性项目因子分析。《心理测量学》，75，33–57。Cen, H., Koedinger, K., & Junker, B. (2006). 学习因素分析——一种用于认知模型评估与改进的通用方法。载于《国际智能辅导系统会议》(第164–175页)。Chalmers, R. P. (2012). mirt：一个多维项目反应理论软件包，适用于

R

环境。《统计软件杂志》，48(6)，1–29。Chen, Y., Li, X., & Zhang, S. (2019a). 高维探索性项目因子分析的联合极大似然估计。《心理测量学》，84(1)，124–146。Chen, Y., Li, X., & Zhang, S. (2019b, June). 大规模数据的结构化潜因子分析：可识别性、可估计性及其影响。*arXiv:1712.08966 [*统计*]*。Cho, S.-J., & Rabe-Hesketh, S. (2011). 具有交叉随机效应模型的交替插补后验估计。《计算统计与数据分析》，55(1)，12–25。Chrysafiadi, K., & Virvou, M. (2013). 学生建模方法：过去十年的文献综述。《应用专家系统》，40(11)，4715–4729。Desmarais, M. C., & Pu, X. (2005). 一种不含隐节点的贝叶斯学生模型及其与项目反应理论的比较。《国际教育人工智能杂志》。Doan, T.-N., & Sahebi, S. (2019). 用于预测学生表现的基于秩的张量分解。载于第*12*届教育数据国际会议论文集

a

挖掘 （第288–293页）。Fan, J., & Li, R.（2001）。通过非凹惩罚似然进行变量选择及其 Oracle 性质。《美国统计学会会刊》，*96*(456)，1348–1360。

Goodman, L. A., & Kruskal, W. H.（1954）。交叉分类的关联测量。《美国统计学会会刊》，*49*(268)，732–764。Haberman, S.（1977）。指数响应模型中的最大似然估计。《统计年刊》，*5*(5)，815–841。Hastie, T., Tibshirani, R., & Friedman, J.（2009）。《统计学习要素》（第2版）。纽约：Springer。Hestenes, D., Wells, M., & Swackhamer, G.（1992）。力概念测验。《物理教师》，*30*(3)，141–158。Hirose, K., & Yamamoto, M.（2015）。因子分析模型中的非凹惩罚似然稀疏估计。《统计与计算》，*25*(5)，863–875。Holland, P.（1990）。项目反应理论模型的抽样基础。《心理测量学》，*55*(4)，577–601。Hu, B., Zhou, Y., Wang, J., Li, L., & Shen, L.（2009）。项目反应理论在协同过滤中的应用。载于 W. Yu, H. He, & N. Zhang（编），神经网络进展——*ISNN 2009*（第766–773页）。柏林、海德堡：Springer。Jin, S., Moustaki, I., & Yang-Wallentin, F.（2018年9月）。探索性因子分析的近似惩罚最大似然：正交情形。《心理测量学》，*83*(3)，628–649。Kingma, D. P., & Ba, J.（2015）。Adam：一种随机优化方法。载于第*3*届国际学习表征会议（*ICLR 2015*），美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集。取自http://arxiv.org/abs/1412.6980Koren, Y., & Bell, R.（2015）。协同过滤进展。载于《推荐系统手册》（第77–118页）。Springer。Lan, A., Waters, A., Studer, C., & Baraniuk, R.（2013）。用于学习与内容分析的稀疏因子分析。*arXiv* 预印本。取自http://arxiv.org/abs/1303.5685

Lord, F. M. (1980年). 项目反应理论在实际测验问题中的应用. 纽约：Routledge. Lord, F. M., Novick, M., & Birnbaum, A. (1968). 心理测验分数的统计理论。 Addison-Wesly Publishing. Martin, B., Mitrovic, T., Mathan, S., & Koedinger, K. R. (2010). 利用学习曲线评估和改进自适应教育系统。用户建模与用户自适应交互：个性化研究期刊, *21*, 249–283. Palmer, H. (2004). 条件最大似然估计。载于*SAGE*社会科学研究方法百科全书 (pp. 168–169). Sage Publications. Pan, J., Ip, E. H., & Dub´e (2017, December). 验证性因子分析中事后模型修正的另一种方法：贝叶斯套索。《心理学方法》, *22*(4), 687–704. Pan, J., Ip, E. H., & Dub´e (2019, November). 多层异质因子分析及其在生态瞬时评估中的应用。《心理测量学》. Pel´anek, R. (2016). 埃洛等级分系统在自适应教育系统中的应用。《计算机与教育》, *98*, 169–179. Prechelt, L. (1998年). 提前停止——但何时停止？载于《神经网络：实用技巧》(pp. 55–69). Springer. Reckase, M. (2009). 多维项目反应理论. 纽约：Springer. Reye, J. (2004). 基于信念网络的学生建模。《国际教育人工智能杂志》, *14*, 1–33. Sahebi, S., Lin, Y.-R., & Brusilovsky, P. (2016). 用于非结构化领域学生建模和表现预测的张量分解。载于第九届国际教育数据挖掘会议论文集 (pp. 502–506). Seaton, D. T., Bergner, Y., Chuang, I., Mitros, P., & Pritchard, D. E. (2014). 谁在大规模开放在线课程中做什么？《*ACM*通讯》, *57*(4), 58–65. Shi, J., Xu, Y., & Baraniuk, R. (2014). 稀疏双线性逻辑回归。*arXiv*

预印本, 1–27. 获取自 http://arxiv.org/abs/1404.4104Stewart, J., Zabriskie, C., Devore, S., & Stewart, G. (2018). 多维项目反应理论与力概念测验。《物理评论：物理教育》

n

研究， *14*(1), 10137. Su, X., & Khoshgoftaar, T. M. (2009). 一项关于协同过滤技术的调查。《人工智能进展》， *2009*年(第3节), 1–19. Sun, J., Chen, Y., Liu, J., Ying, Z., & Xin, T. (2016). 通过L1正则化进行多维项目反应理论模型的潜变量选择。《心理测量学》, *81*(4), 921–939. Tibshirani, R. (1996). 通过Lasso进行回归收缩与选择。《英国皇家统计学会会刊：*B*辑》, *58*(1), 267–288. Trendafilov, N. T., & Adachi, K. (2015). 稀疏结构与简单结构载荷。《心理测量学》, *80*(3), 776–790. Trendafilov, N. T., Fontanella, S., & Adachi, K. (2017, 9 月). 稀疏探索性因子分析。《心理测量学》 , *82*(3), 778–794. Yao, Y., Rosasco, L., & Caponnetto, A. (2007年). 关于梯度下降学习中的提前停止。《构造性逼近》, *26*(2), 289–315. Zhou, Y., Wilkinson, D., Schreiber, R., & Pan, R. (2008). 面向Netflix Prize的大规模并行协同过滤。载于国际管理算法应用会议 (第337–348页). Zhu, Y., Shen, X., & Ye, C. (2016年1月). 潜因子模型中的个性化预测与稀疏性追求。《美国统计学会会刊》, *111*(513), 241–252.