心理学探新 2022,Vol.42,No.3,261 -268
PSYCHOLOGICAL EXPLORATION

基于正则化的探索性中介分析:原理与应用*

邓雅婷 张沥今 潘俊豪
(中山大学心理学系,广州 510006)

摘 要:探索性中介分析被定义为从变量集合中筛选潜在中介变量的方法,该方法能在缺乏理论基础的情况下帮助研究者从数据中挖掘潜在中介机制,提供模型构建上的指导。本文介绍了一种基于正则化的探索性中介分析方法 XMed(exploratory mediation analysis via regularization)。相比于传统探索性中介分析方法,XMed 具有检验力更高、所需样本量更小、能高效地处理高维数据等优点,在认知神经科学、临床心理学等心理学领域有较大的应用潜力。本文主要介绍 XMed 的原理和实现过程,并通过实例分析展示该方法的应用。

关键词:中介效应;探索性中介分析;正则化;Lasso
中图分类号:B841.2 **文献标识码**:A **文章编号**:1003 -5184(2022)03 -0261 -08

变量间的影响机制是心理学领域的一个重要研究主题,而中介分析因其能够剖析自变量和因变量间的作用机制而在社会科学中得到广泛应用。Rucker 等(Rucker et al.,2011)的统计表明,在 2005 年到 2009 年间发表在社会心理学知名期刊 Journal of Personality and Social Psychology(JPSP)和 Personality and Social Psychology Bulletin(PSPB)上的文章中,分别有 59% 和 65% 的文章使用中介模型进行分析。此外,在谷歌学术上检索发现,2010 年至 2019 年间 PSPB 和 JPSP 上分别约有 68% 和 63% 的文章包含至少一个与"mediation"相关的关键词。在管理与组织心理学、消费者行为学等领域,也有相当多实证研究使用了中介分析(Pieters,2017;Wood et al.,2008)。

当前的中介效应分析以验证性中介分析为主流,但验证性分析的方法并不适用于部分情形。验证性分析需要研究者在建模前有较强的先验理论,而后基于理论建立假设模型并进行检验。然而,研究者可能会遇到现有的理论无法指导假设中介模型构建的情况,例如在计算网络心理学(computational cyberpsychology;朱廷劭,2016)等研究领域中,由于缺乏成熟的理论基础,研究者很难从大量网络行为变量中挑选中介变量用于构建中介模型,因此也很难通过验证模型来进一步探索研究问题,类似情况在新兴研究领域更为常见。此外,随着信息技术的发展,功能性磁共振成像(functional magnetic resonance imaging,fMRI)、网络数据抓取、基因测序等技术在心理学研究中的应用逐渐增多,这些技术允许研究者在有限的样本上获取大量的数据特征(van Kesteren & Oberski,2019)。然而在认知心理学、临床心理学等领域,由于数据收集需要付出较大的经济和时间成本,因此也常常出现样本有限而变量数较多的数据集。对这类样本 - 变量数比例较低的数据,一方面,如果不加筛选地将所有变量纳入模型会使模型过于复杂而难以解释,降低模型的泛化能力,还会导致参数估计上的困难;另一方面,研究者也很难仅凭理论就从数据集中毫无遗漏地筛选出重要的变量进行建模。

针对上述验证性中介分析的困境,探索性中介分析的方法能够灵活地进行处理。探索性中介分析指的是从变量集合中筛选出潜在中介变量,并筛除无中介效应的噪声变量的一系列方法(Serang et al.,2017)。它们能够帮助研究者洞悉数据背后的信息,挖掘潜在的中介关系,数据驱动地从数据集中选出具有统计学效应的变量,为后续研究提供理论和实证基础。过去也有研究者提出用于探索性中介分析的方法,但这些方法基于假设检验的结果进行变量选择,可能存在过拟合、效率低等问题(MacKinnon,2008;Serang et al.,2017)。对此,有研究者将机器学习中的正则化方法与结构方程模型(structural equation modeling,SEM)框架下的多重中介模型相结合,提出了一种新的基于正则化的探索性中介分析(exploratory mediation analysis via regularization,XMed;Serang et al.,2017)。其原理是使用正则化方法对中介变量的前后路径进行惩罚,将较小的、不重要的系数压缩至零来筛除中介效应较小的变量,从而达到检测潜在中介变量的目的(Serang et al.,2017)。

XMed 以正则化(regularization)方法作为核心,这一核心使其在处理过拟合(overfitting)问题上有良好表现。过拟合指的是模型在当前样本拟合良

* **基金项目**:国家自然科学基金项目(31871128),广东省自然科学基金项目(2022A1515010367)。
通讯作者:潘俊豪,E - mail:panjunh@ mail.sysu.edu.cn。

好,但该良好拟合表现不能泛化到同一总体的其它样本上的现象,这一现象会导致研究结果的可重复性降低(Babyak,2004)。在传统的一般最小二乘(ordinary least square,OLS)回归中,过拟合问题表现为参数估计在小到中等的样本量下容易受到噪声的影响、预测变量的效应被高估、一类错误率增加等(Babyak,2004;Cohen et al.,2003)。如果研究者的目的在于预测变量的选择,则使用传统的变量选择方法(如向前、向后回归和逐步回归)会让过拟合问题被进一步放大,数据中无关变量被选入模型,使模型变得更为复杂且具有误导性(Derksen & Keselman,1992)。而正则化方法通过在损失函数中引入惩罚项就能较好地解决传统方法中的过拟合问题,减小参数对噪声的敏感性,并得到更稳定、可泛化性更强的参数估计。在众多正则化方法中,Lasso(least absolute shrinkage and selection operator;Tibshirani,1996)是最为常用的方法之一,它在生物医学等领域得到了广泛应用,但是在行为科学领域尚未得到普遍重视(McNeish,2015;张沥今 等,2020)。近年来,有研究者将正则化与结构方程模型相结合,提出了正则化结构方程模型(regularized structural equation modeling,RegSEM;Jacobucci et al.,2016)的新分析框架(Jacobucci et al.,2016)。而 XMed 就是将这一框架运用于中介模型而提出的方法,因此也同时继承了正则化方法在处理过拟合问题上的优势。

此外,相较于传统探索性中介方法,XMed 也有其独特优势。传统的探索性中介方法是建立在假设检验的基础上的,其具体做法有两种:第一种是将每个待选择变量作为单独的中介变量,分别建立简单中介模型并进行检验(需要对值进行校正),其中具有显著中介效应的变量将被识别为潜在中介变量(MacKinnon,2008);第二种则是将所有变量作为并行的中介变量纳入模型(即构建多重中介模型),使用 SEM 的框架进行检验,此方法也要矫正 p 值并最终选择中介效应显著的变量(Serang et al.,2017;van Kesteren & Oberski,2019)。然而,两者都可能产生过拟合的问题,尤其前者在变量数多时效率低,且会由于模型误设而额外引入偏差(因没有控制其它中介变量的影响而产生估计偏差;Babyak,2004;Serang et al.,2017;van Kesteren & Oberski,2019),而后者在变量数多于样本数时会导致参数估计收敛上的问题(van Kesteren & Oberski,2019)。相比于传统方法,XMed 能弥补上述的过拟合、效率低的问题,能通过数据降维(即通过变量选择来减少特征数)简化模型。并且模拟研究也表明,XMed 具有更高的敏感性,即在潜在中介变量的选择上有更高的检验力,并且在相同敏感性的前提下,XMed 所需的样本量更少(Serang et al.,2017)。

综上所述,XMed 在心理学、行为学等社会科学领域有广阔的应用前景。首先,在既有理论基础薄弱、信息缺乏的情况下(尤其是在面对新研究领域时),研究者很难在主流的验证性分析框架下对变量间的作用机制进行具体、清晰的理论和假设,而此时 XMed 可以从数据中挖掘关于中介作用机制的信息,为研究者提供理论构建上的指导。根据 Locke(2007)的观点,心理学、管理学等社会科学领域应以归纳(induction)的过程进行研究,即基于对数据的观察和整合来提出概念和理论,例如 Beck(1993)基于他对来访者的观察提出了抑郁的认知理论和干预方法。而 XMed 获取的中介机制信息也可作为归纳推理的证据,为新理论的提出及其进一步验证提供实证基础。其次,过去心理学研究者主要关注的是对因果机制的解释,但是却忽视了理论对未来行为的预测能力,这样的观念在一定程度上导致了心理学研究的低可重复性,而正则化方法能提供更具泛化性的结果,可改善研究结果的可重复性,在心理学可重复性危机的背景下能发挥重要作用(Yarkoni & Westfall,2017;张沥今 等,2020)。最后,随着新兴技术在心理学研究中的推广应用,研究者能更容易地获取大量的数据特征,但是受到经济和人力各方面成本的限制,样本量往往是有限的,对于这种小样本数据,普通的 SEM 建模分析(如基于极大似然估计的多重中介模型分析)可能会遇到参数估计不收敛的问题(Rosseel,2020)。相比之下,XMed 能通过数据降维使模型建立和参数估计变得可行,因此 XMed 是有效处理小样本困境的方法之一(Serang et al.,2017;van Kesteren & Oberski,2019)。

基于上述理由,本文对 XMed 的具体原理和实现步骤进行介绍,并展示相应的案例分析,旨在推广 XMed 在社会科学领域研究中的应用。

1 多重中介效应模型

在中介效应模型中,研究者的是关注自变量对因变量的作用机制,即自变量如何通过中介变量进而对因变量产生影响。一个包含单一中介变量的简单中介模型(simple mediation model)如图 1 所示,其中图 1A 描述的了 X 对 Y 总效应(total effect)c;在图 1B 中,a 表示 X 对 M 的效应,b 表示控制 X 的影响后 M 对 Y 的效应;总效应则被分解为直接效应(direct effect)c'与间接效应(indirect effect)ab,因此总效应与中介效应、直接效应的关系可描述为 $c - c' = ab$;e_1,e_2,e_3 为回归残差(MacKinnon,2008;MacKinnon et al.,1995)。

包含多个中介变量的模型则被称为多重中介模型(multiple mediation model)。多重中介模型包括链式中介和并行中介两种类型:在链式中介中,中介变量之间相互联系,构成一条影响链;而在并行中介模型中,中介变量则呈互相平行的关系(柳士顺,凌

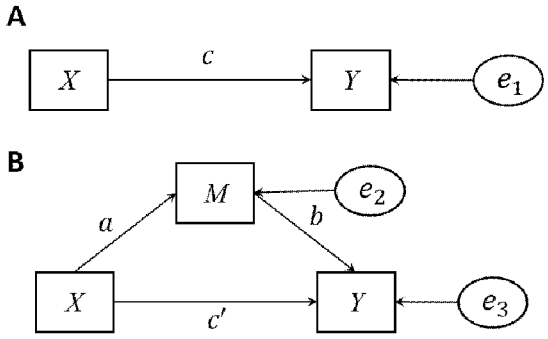

图 1　简单中介模型(温忠麟 等,2005)

文轩,2009)。由于本文所介绍的探索性中介分析方法主要针对并行中介模型的构建,因此对多重中介模型的描述重点关注并行中介的情形。一个包含 k 个中介变量的并行中介模型可用图 2 描述,对应的回归方程如下(Hayes,2017):

$$Y = i + cX + e \quad (1)$$

$$M_j = i_j + a_j X + e_j \quad (2)$$

$$Y = i' + c' X + \sum_{j=1}^{k} b_j M_j + e' \quad (3)$$

其中 $j = 1 \cdots k$,i,而 i',i_j 为截距,e,e',e_j 为回归残差。与简单中介模型类似,c,c' 分别是 X 对 Y 的总效应和直接效应,中介变量 M_j 前后路径系数的乘积 $a_j b_j$ 则是它所对应的特定中介效应(specific indirect effect),并且:

$$c = c' + \sum_{j=1}^{k} a_j b_j \quad (4)$$

多重中介模型相比于简单中介模型更具优势,且在心理学研究中被广泛使用。多重中介模型纳入了多个中介变量,能控制其它中介变量的影响,得到更准确的参数估计,而简单中介效应会因为无法控制其它变量的影响而导致有偏估计,甚至会使中介效应被遮掩(即因为没有考虑其它中介变量的影响而中介效应无法显现的现象;Guyon & Elisseeff, 2003;Judd & Kenny,1981)。此外,多重中介模型能估计各个中介变量的特定中介效应及其效应总和,并允许多个特定中介效应的进行大小比较。由于在心理学、行为科学等社会科学领域中,变量间的作用机制也通常存在多条途径,研究者常常需要考虑多方面因素的复杂影响,以更好地理解变量间的关系,做出更有效的解释和预测,因此多重中介模型的应用也越来越普遍(MacKinnon,2008;方杰 等,2014)。

2　基于正则化的探索性中介分析

2.1　原理

XMed 采用正则化方法进行变量筛选的思想源于 Lasso 回归。Lasso 最早由 Tibshirani(1996) 提出,主要在数据挖掘、机器学习等领域用于线性回归模型的预测变量筛选,以达到简化模型,提高模型可解释性和泛化能力的目的。Lasso 回归的参数估计可通

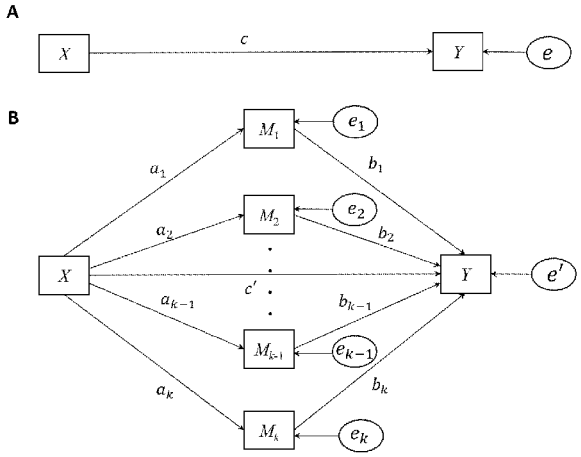

图 2　多重中介模型

过最小化损失函数,即式(5) 获得。

$$\sum_{l=1}^{n} (Y_l - \beta_0 - \sum_{p=1}^{P} \beta_p X_{lp})^2 + \lambda \sum_{p=1}^{P} |\beta_p| \quad (5)$$

其中,P 为多元回归模型中的预测变量数目,β_p 为预测变量 X_{lp} 对应的回归系数。Lasso 回归在传统一般最小二乘(ordinary least square regression, OLS) 回归的基础上加入了惩罚项,即式(5) 中加号后面的项。通过引入该惩罚项,Lasso 回归能够将较小的回归系数压缩至 0,从而达到变量选择的目的。

系数压缩的程度与调整参数(tuning parameter)$\lambda \geqslant 0$ 有关。当 λ 值等于 0 时,Lasso 回归等价于 OLS 回归,即不对系数进行惩罚;而随着 λ 的增大,系数被压缩的程度越大,筛选出的变量则越少。最优 λ 值可通过交叉验证(cross - validation) 选取,也可通过信息准则(information criteria) 如 AIC(Akaike's information criterion) 或 BIC(Bayesian information criterion) 选取(Akaike, 1998;Schwarz,1978;Stone,1974)。如果使用 k - 折交叉验证(k - fold cross validation) 进行参数调整,首先需要确定一组用于比较的 λ 值,而后将数据划分为 k 个子集(其中 1 个子集作为测试集,其余子集作为训练集) 对不同的 λ 值进行测试,最后根据模型表现最佳的 λ 值用于系数惩罚。使用信息准则的方法也需要首先确定一组 λ 值,然后将每一个 λ 值对应的模型与完整的数据集拟合,最后选择最小信息准则的 λ 值作为最优参数值。

虽然 Lasso 回归能进行参数估计,但由于惩罚项会对所有回归系数进行同等程度的压缩,因此除了会将较小的回归系数压缩为 0 以外,对其它回归系数的估计也会有趋于 0 的偏差(Hastie et al., 2009;Tibshirani,1996)。为了得到更准确的估计值,有研究者提出了松弛 Lasso 方法(Relaxed Lasso; Meinshausen,2007)。松弛 Lasso 包括两个阶段:在第一阶段使用最优值筛选变量;第二阶段则是在模型中只纳入第一阶段选出的变量,并使用更小、甚至可以为零的值进行参数估计,以避免参数估计偏差。

近年来,有研究者将 Lasso 回归的正则化思想应用于 SEM 的框架下,提出了 RegSEM(Jacobucci et al. ,2016)。RegSEM 的拟合函数由结构方程模型最大似然估计的拟合函数和正则化惩罚项所组成,表示为式(6):

$$F_{regsem} = \log(|\Sigma|) + tr(C * \Sigma^{-1}) - \log(|C|) - p + \lambda P(\cdot) \quad (6)$$

其中 Σ 为模型协方差矩阵(model - implied covariance matrix),C 为观测数据的样本协方差矩阵(sample covariance matrix),p 是观测变量的数目,λ 则是调整参数。在惩罚项 $\lambda P(\cdot)$ 中,$P(\cdot)$ 代表惩罚函数,不同的正则化方法所对应的惩罚函数形式有所不同,由于本文所介绍的 XMed 采用的是 Lasso 正则化方法,因此 $P(\cdot)$ 是被惩罚系数的绝对值之和。RegSEM 将 SEM 的灵活性和 Lasso 回归的变量选择特性相结合,使正则化的方法能够应用到简单线性回归以外的模型中,Serang 等(2017) 也基于此提出了使用 RegSEM 进行中介变量选择的方法 XMed,具体流程见 2.2。

模拟研究表明,与基于传统中介变量选择思路的方法(例如根据 bootstrap 和 Sobel 检验的结果的显著性进行变量选择) 相比,XMed 在检验力和样本量需求上有更好的表现(Serang et al. ,2017;Serang & Jacobucci,2020)。具体而言,XMed 的敏感性比传统方法要更高,即在检测真实中介变量上有更高的检验力,并且在中介效应较小或中等时表现优异。并且,XMed 所需样本量也比传统方法更小,在达到相同检验力的前提下,基于连续数据的 XMed 所需样本量只需传统方法的一半。尽管 XMed 的高检验力伴随着高一类错误率,但在探索性分析的背景下,高二类错误率造成的后果比高一类错误率更严重。具体而言,如果具有中介效应的变量被遗漏(即犯二类错误),这些变量将很难在后续分析中找回;而如果无中介效应或中介效应很小的变量被保留下来(即犯一类错误),这些变量仍能够在后续根据理论或数据分析结果进行筛除。所以与高二类错误率、低一类错误率的传统方法相比,XMed 显然更适用于探索性分析的背景,这也是 XMed 相对于传统方法的优势之一。

目前已有部分实证研究使用 XMed 进行变量筛选。例如有学者从人际交往、风险活动参与、心理病理学症状、暴露于自杀行为相关的一系列变量中,寻找童年虐待经历和自杀企图之间的关系的中介变量(Ammerman et al. ,2018);也有研究者尝试从与家庭环境和父母教养相关的变量中,寻找中介情绪/行为问题和酗酒问题之间的关系的中介变量(Serang & Jacobucci,2020)。虽然这一新近提出的方法在实证研究中尚未得到广泛应用,但它能有效处理在心理学中较为常见的样本 - 变量数比例低的数据集,具有较大的应用前景,因此下文将具体介绍其实现步骤,并使用实例分析进行展示,以推广 XMed 的应用。

2.2 实现步骤

XMed 的一般实现步骤包括前置工作,以及变量选择和参数估计两个主要阶段。其中前置工作需要将数据标准化以保证 XMed 对各系数的压缩程度一致,然后设定研究者感兴趣的多重中介模型,此模型的中介变量即所有待选择的变量。在此之后则进入变量选择和参数估计的两阶段步骤。

XMed 的第一阶段通过压缩中介变量的前后路径来实现中介变量的选择。在第一阶段,使用 RegSEM 拟合多重中介模型并对所有的 a、b 参数进行惩罚,即为拟合函数加上由这些参数构成的惩罚项,从而压缩中介变量的前后路径系数,其中惩罚项的调整参数 λ 的值可通过交叉验证或信息准则进行选择。如果中介变量 M_j 所对应的系数 a_j 和 b_j 中有至少一个被压缩至 0(即 $a_jb_j = 0$),则该中介变量被删除,其它未被压缩至 0 的中介变量则会被保留作为潜在中介变量。

XMed 的第二阶段基于第一阶段变量选择的结果来进行参数估计。由于在第一阶段所有 a、b 参数都受到压缩,因此会引入一定程度偏差。为了避免第一阶段由于参数压缩而造成参数估计偏差,第二阶段采用了与松弛 Lasso 相同的逻辑,使用第一阶段保留的变量作为中介变量构建新的多重中介模型进行拟合,同时将惩罚项中的 λ 值设为 0(即不进行惩罚),以此得到更准确的参数估计。

根据模型中变量类型的不同,XMed 有不同的 R 语言实现方法,分析流程可用图 3 的流程图来表示。目前,XMed 可应用于所有变量为连续数据的情况,或结果变量(outcome variables) 中存在二分变量的情况 (Serang et al. ,2017;Serang & Jacobucci, 2020)。首先,需要完成的前置工作是对标准化数据和通过 lavaan 包(Rosseel,2012) 中的 sem 函数来构建模型,作为后续参数压缩和变量选择的基础。在第一阶段,若变量都为连续变量,需要使用 regsem 包中的 cv_regsem 函数和 multi_optim 函数分别进行 λ 值的调整和 a、b 参数的压缩,根据压缩后的参数估计可判断哪些变量被选择(详见附录与实例分析部分);而当结果变量中存在二分变量时,已有学者编写了该情况下的实现函数 xmed,因此直接调用 regsem 包的 xmed 函数就可以完成第一阶段的全部过程(Serang & Jacobucci,2020)。第二阶段则是使用第一阶段保留下来的变量作为中介变量来进行模型拟合,这一步可以用一般的 SEM 分析工具实现,如 Mplus(Muthén & Muthén,2017),lavaan(Rosseel, 2012) 等。本文将在下一节解结合实例分析展示连续变量情形下 XMed 的 R 语言实现。

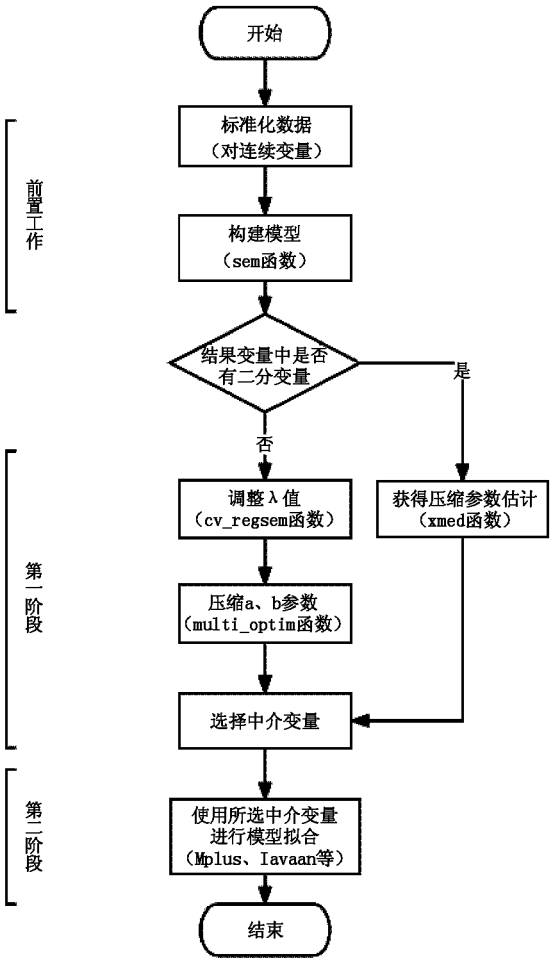

图 3　XMed 代码分析流程图

3　实例分析

本节将通过实例分析向读者展示 XMed 方法在实证研究中的应用和具体分析过程。假设研究者想要探讨亚裔儿童与教师之间的人际互动会如何对儿童自身的学习行为产生影响，但目前没有足够证据和理论基础支持相关模型的构建，此时就可以使用 XMed 数据驱动地探索可能的中介机制。本实例分析使用一份来自早期儿童追踪研究项目（Early Childhood Longitudinal Studies，ECLS；Tourangeau et al.，2015）公开数据集 ECLS - K:2011 来对分析过程和分析结果予以展示和说明。该数据集包含 2010 ~ 2011 年关于美国学校儿童的认知、情绪、学业等方面发展的数据，数据的获取方式包括对儿童各方面能力的测试，对儿童及其父母、老师的访谈和问卷调查等。首先选取 2011 年春季学期亚裔儿童的数据，筛除含缺失数据以及不符合要求的案例后，所得数据总样本量 N = 848（男生 383 人），平均年龄为 72.18 个月（s = 4.37）。以儿童与教师的亲密程度作为自变量，儿童的学习行为作为因变量，儿童的自我控制、人际交往、专注、孤独/悲伤、冲动/过度反应分别作为待选择的潜在中介变量*，本实例分析的目的在于通过 XMed 筛选出具有中介效应的变量，探讨儿童与教师的亲密程度对儿童学习行为可能的作用机制。实例分析所用 R 语言程序参考 Serang 等（2017），详见附录。

在第一阶段，首先使用 lavaan 包构建多重中介模型，该模型需要纳入所有待选择的中介变量，也就是构建以儿童与教师的亲密程度作为自变量，儿童的学习行为作为因变量，儿童的自我控制、人际交往、专注、孤独/悲伤、冲动/过度反应作为中介变量的并行中介模型。构建模型后，研究者需要选择惩罚模型中的哪些参数，regsem 包的 extractMatrics 函数能将模型中的每一条路径赋予一个序号，研究者只需要找出被惩罚参数所对应的序号，作为后面所用函数的输入，例如本研究中的被惩罚参数是模型中的 10 个中介路径前、后系数，在 extractMatrics 的输出中对应的序号为 2 - 11，因此 2 - 11 的序号会作为后面 cv_regsem、multi_optim 函数的输入，告诉程序需要对模型的前后路径系数进行压缩。而后调用 cv_regsem 函数寻找最优值，本研究设置了以 0 为起始值、步长为 0.005 的 100 个 λ 值的集合，并使用 BIC 信息准则从中选择最优值。结果显示当模型的 BIC 值最小，因此最终以 0.07 作为最优参数值。使用 multi_optim 函数进行系数压缩，得到的各路径系数估计结果如表 1。由表 1 可知，孤独/伤心和冲动/过度反应的中介效应被压缩至零，说明这两个变量因为其中介效应较小而被筛除，最终自我控制、人际交往和专注被选为潜在中介变量以进行后续分析。

表 1　第一阶段分析结果

中介变量	a 路径	b 路径	特定中介效应
自我控制	0.27	0.18	0.05
人际交往	0.47	0.29	0.14
专注	0.33	0.54	0.18
孤独/悲伤	0	0	0
冲动/过度反应	0	-0.01	0
总中介效应		0.34	

为了避免第一阶段的参数估计偏差，第二阶段

* 所用变量在 ECLS - K:2011 中的对应变量名分别为：与教师的亲密程度（X2CLSNSS）、儿童的学习行为（X2TCHAPP）、自我控制（X2TCHCON）、人际交往（X2TCHPER）、专注（X2ATTNFS）、孤独/悲伤（X2PRNSAD）、冲动/过度反应（X2PRNIMP）。这些变量在数据集中均为连续变量。

针对潜在中介变量建立无惩罚模型。由于这一阶段等价于最大似然估计中介效应分析，因此 SEM 软件都可用于第二阶段的分析。在此使用第一阶段筛选所得的三个潜在中介变量构建并行中介模型，并使用 lavaan 包进行模型检验和系数估计，估计结果如表 2 所示。结果说明，自我控制、人际交往和专注可能在儿童与教师的亲密程度和儿童的学习行为之间起中介作用，研究者可对此作进一步的理论解释，或据此构建新的理论模型并进行验证。

表 2 第二阶段分析结果

中介变量	*a* 路径	*b* 路径	特定中介效应
自我控制	0.30	0.19	0.06
人际交往	0.49	0.30	0.15
专注	0.36	0.55	0.20
总中介效应		0.37	

4 讨论

随着心理学各个子领域的快速发展，探索性中介分析未来会有更多的用武之地。传统的验证性方法强调用理论指导研究，但随着数据采集技术的发展，理论的发展逐渐跟不上数据体量增加的速度，因此会出现现有理论不足以指导模型构建的情况。随着新领域和交叉学科领域不断地涌现，这一情况会变得愈加常见。而探索性中介能从大数据中提取有价值的信息，指导后续研究的进行和新理论的构建，实现理论和数据的协同发展。

本文介绍了一种新的探索性中介分析方法 XMed，并对其原理和实现过程进行了详细阐述。相比于传统的基于统计显著性筛选变量的方法，XMed 不仅具有检验力更高、所需样本量更少的优点，而且相比于传统方法有更高的容错性，这些优势使其更适用于探索性的情景（Serang et al.，2017）。此外，XMed 继承了 Lasso 正则化机器学习算法的优势，在分析变量数目较多的数据集时具有良好的表现，未来在认知神经科学、网络心理学等领域能发挥其独特作用（张沥今 等，2020）。

然而，XMed 目前仍有一定的局限性和发展空间。首先，XMed 不能提供准确的标准误和区间估计，尽管有研究者认为可以借助 bootstrap 的方法进行估计，但该方法被认为难以获得可靠的估计，到目前为止仍没有一个公认的能为 Lasso 计算标准误和区间估计的较好方法（Casella et al.，2010；Serang & Jacobucci，2020；van Erp et al.，2019）。其次，XMed 本质上是一个探索性的分析方法，可能会带来较高的一类错误率（即模型中可能会纳入较多冗余变量），这需要研究者在后续的验证性分析中对模型进行验证（Lieberman & Cunningham，2009）。在应用研究方面，XMed 适用的建模场景还比较有限，目前 XMed 仅适用于模型中所有变量为连续变量，或结果变量（因变量和中介变量）中存在二分变量的情形。未来研究可以将其应用至更多的数据类型中（如顺序变量），或含潜变量（即无法直接测量，需要通过显变量反映的变量）的模型，并探索其方法表现。最后，尽管 XMed 可用于中介变量的选择，但其本质是对路径的选择，当中介变量前、后路径效应都比较小而中介效应的确存在时，XMed 会无法识别该中介变量，未来研究可考虑对这一方面进行改进，例如 van Kesteren 和 Oberski（2019）提出的坐标中介过滤法（Coordinate - wise Mediation Filter）就能处理这一局限，但这一方法对计算资源的耗费较大。

综上，通过结合机器学习中的正则化方法，XMed 实现了探索性的中介变量选择，为研究者们提供了一个新的建模思路。本文建议应用研究者在实证研究中使用 XMed 来发掘数据中蕴藏的更丰富的信息，而方法学研究者可考虑改进 XMed 使其适用于更多应用场景。

参考文献

方杰，温忠麟，张敏强.（2014）. 基于结构方程模型的多重中介效应分析. *心理科学*，*37*（3），735 - 741.

柳士顺，凌文辁.（2009）. 多重中介模型及其应用. *心理科学*，*32*（2），433 - 435 + 407.

温忠麟，侯杰泰，张雷.（2005）. 调节效应与中介效应的比较和应用. *心理学报*，*37*（2），268 - 274.

朱廷劭.（2016）. *大数据时代的心理学研究及应用*. 科学出版社.

张沥今，魏夏琰，陆嘉琦，潘俊豪.（2020）. Lasso 回归：从解释到预测. *心理科学进展*，*28*（10），1777 - 1791.

Akaike，H.（1998）. Information theory and an extension of the maximum likelihood principle. In E. Parzen，K. Tanabe，& G. Kitagawa（Eds.），*Selected papers of hirotugu akaike*（pp. 199 - 213）. Springer New York.

Ammerman，B. A.，Serang，S.，Jacobucci，R.，Burke，T. A.，Alloy，L. B.，& McCloskey，M. S.（2018）. Exploratory analysis of mediators of the relationship between childhood maltreatment and suicidal behavior. *Journal of Adolescence*，*69*，103 - 112.

Babyak，M. A.（2004）. What you see may not be what you get：A brief，nontechnical introduction to overfitting in regression - type models. *Psychosomatic Medicine*，*66*（3），411 - 421.

Beck，A. T.（1993）. *Cognitive therapy of depression：A personal reflection*. Aberdeen，Scotland：Scottish Cultural Press.

Casella，G.，Ghosh，M.，Gill，J.，& Kyung，M.（2010）. Penalized regression，standard errors，and Bayesian lassos. *Bayesian Analysis*，*5*（2），369 - 411.

Cohen，J.，Cohen，P.，West，S. G.，& Aiken，L. S.（2003）. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Derksen，S.，& Keselman，H. J.（1992）. Backward，forward and stepwise automated subset selection algorithms：Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*，*45*（2），265 - 282.

Guyon，I.，& Elisseeff，A.（2003）. An introduction to variable and feature selection. *The Journal of Machine Learning Research*，*3*，1157 - 1182.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer Science & Business Media.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression - based approach*. Guilford Publications.

Jacobucci, R. (2017). *Regsem: Regularized structural equation modeling*. ArXiv: 1703.08489 [Stat]. http://arxiv.org/abs/1703.08489

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555 - 566.

Judd, C. M., & Kenny, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation Review*, *5*(5), 602 - 619.

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: Re - balancing the scale. *Social Cognitive and Affective Neuroscience*, *4*(4), 423 - 428.

Locke, E. A. (2007). The case for inductive theory building. *Journal of Management*, *33*(6), 867 - 890.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Taylor & Francis Group/Lawrence Erlbaum Associates.

MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, *30*(1), 41 - 62.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471 - 484.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*(1), 374 - 393.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide (version 8)*. Muthén & Muthén.

Pieters, R. (2017). Mediation analysis: Inferring causal processes in marketing from experiments. In P. S. Leeflang, J. E. Wieringa, T. H. Bijmolt, & K. H. Pauwels (Eds.), *Advanced methods for modeling markets* (pp. 235 - 263). Springer.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, *40*(3), 879 - 891.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*(2), 1 - 36.

Rosseel, Y. (2020). Small sample solutions for structural equation modeling. In R. Vande Schoot & M. Miocević (Eds.), *Small sample size solutions* (pp. 226 - 238). Routledge.

Rucker, D. D., Preacher, K. J., Tormala, Z. L., & Petty, R. E. (2011). Mediation analysis in social psychology: Current practices and new recommendations. *Social and Personality Psychology Compass*, *5*(6), 359 - 371.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461 - 464.

Serang, S., & Jacobucci, R. (2020). Exploratory mediation analysis of dichotomous outcomes via regularization. *Multivariate Behavioral Research*, *55*(1), 69 - 86.

Serang, S., Jacobucci, R., Brimhall, K. C., & Grimm, K. J. (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 733 - 744.

Stone, M. (1974). Cross - validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111 - 133.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267 - 288.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Hagedorn, M. C., Daly, P., & Najarian, M. (2015). *Early childhood longitudinal study, kindergarten class of* 2010 - 11 (*ECLS - K*: 2011). *User's manual for the ECLS - K*: 2011 *kindergarten data file and electronic codebook, public version* (NCES 2015 - 074). U. S. Department of Education. National Center for Education Statistics.

vanErp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31 - 50.

vanKesteren, E. - J., & Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 710 - 723.

Wood, R. E., Goodman, J. S., Beckmann, N., & Cook, A. (2008). Mediation testing in management research: A review and proposals. *Organizational Research Methods*, *11*(2), 270 - 295.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100 - 1122.

附录:

实例分析部分 R 语言程序

```
library(lavaan)
library(regsem)

#将数据标准化
data_sca <- data.frame(scale(data6))

#将待选择变量纳入中介模型
model <-
  '#c_prime 直接效应
  X2TCHAPP ~ c_prime * X2CLSNSS
  #a 路径
  X2TCHCON ~ a1 * X2CLSNSS
```

```
    X2TCHPER ~ a2 * X2CLSNSS
    X2ATTNFS ~ a3 * X2CLSNSS
    X2PRNSAD ~ a4 * X2CLSNSS
    X2PRNIMP ~ a5 * X2CLSNSS
    #b 路径
    X2TCHAPP ~ b1 * X2TCHCON + b2 * X2TCHPER
+ b3 * X2ATTNFS + b4 * X2PRNSAD + b5 * X2PRNIMP
    #特定中介效应
    a1b1 := a1 * b1
    a2b2 := a2 * b2
    a3b3 := a3 * b3
    a4b4 := a4 * b4
    a5b5 := a5 * b5
    #总效应
    c := c_prime + a1 * b1 + a2 * b2 + a3 * b3 + a4 *
b4 + a5 * b5'

  #拟合模型
  fit_res <- sem(model, data_sca)
  summary(fit_res)
  #找到需要惩罚的参数序号,本例中为 2 到 11
  extractMatrices(fit_res) $ A

  #第一阶段
  #使用 BIC 信息准则调参
  fit.reg.tune <- cv_regsem(model = fit_res, type = "
lasso", fit.ret = "BIC",
                            pars_pen = c(2:11), n.
lambda = 100,lambda.start = 0,
                            jump = 0.005)#给 pars_
pen 参数传入需要惩罚的参数序号

  #找到最小 BIC 所对应的 lambda 值
  bics <- fit.reg.tune[[2]][,"BIC"] #提取所有 bic 值
  plot(seq(0,0.495,by = 0.005),bics,main = "BIC by
lambda",
       xlab = "lambda",ylab = "BIC") #绘制 bic 变化
图
  min.bic <- min(bics) #找到最小 bic
  lambda <- fit.reg.tune[[2]][which(bics == min.
bic), "lambda"] #确定最优 lambda 值

  #用最优 lambda 值再拟合一次
  fit.reg1 <- multi_optim(fit_res, type = "lasso",pars_
pen = c(2:11),lambda = lambda, gradFun = "ram", opt-
Method = "coord_desc")
  summary(fit.reg1)
  #输出特定中介效应
  fit.reg1[["defined_params"]]

  #第二阶段
  #只纳入第一阶段选择出的 3 个中介变量
  model2 <-
  '#直接效应
  X2TCHAPP ~ c_prime * X2CLSNSS
  #a 路径
  X2TCHCON ~ a1 * X2CLSNSS
  X2TCHPER ~ a2 * X2CLSNSS
  X2ATTNFS ~ a3 * X2CLSNSS
  #b 路径
  X2TCHAPP ~ b1 * X2TCHCON + b2 * X2TCHPER +
b3 * X2ATTNFS
  #中介效应
  a1b1 := a1 * b1
  a2b2 := a2 * b2
  a3b3 := a3 * b3
  #总效应
  c := c_prime + a1 * b1 + a2 * b2 + a3 * b3 '

  fit2 <- sem(model2, data = data_sca)
  summary(fit2)
  fitmeasures(fit2)#输出拟合指标
  inspect(fit2, 'r2')#输出 R 方
```

The Principle and Application of Exploratory Mediation Analysis via Regularization

Deng Yating Zhang Lijin Pan Junhao
(Department of Psychology, Sun Yat-sen University, Guangzhou 510006)

Abstract: Mediation analysis is common in social science. Exploratory mediation analysis is defined as a series of data-driven methods for identifying potential mediators from a set of variables. It offers insights into the potential mediation process from data and provides guidance on model building. This article introduces the approach of exploratory mediation analysis via regularization(XMed). Compared to conventional exploratory mediation analysis approaches, XMed has higher sensitivity and needs less sample size. Moreover, it can handle high-dimensional data efficiently, which endows XMed a great potential for application in fields including cognitive neuroscience and clinical psychology. This article focuses on the principle and implementation of XMed. An empirical analysis is included to demonstrate the application of XMed.

Key words: mediation; exploratory mediation analysis; regularization; Lasso