心 理 学 报 2004,36(2):186~194
Acta Psychologica Sinica

结构方程模型检验:拟合指数与卡方准则*

温忠麟[1,2] 侯杰泰[1] 马什赫伯特[3]

([1] 香港中文大学教育学院,香港)([2] 华南师范大学教科院,广州 510631)

([3] 西悉尼大学教育学院,悉尼,澳大利亚)

摘 要 讨论了 Hu 和 Bentler(1998,1999)推荐的检验结构方程模型的 7 个拟合指数准则,对这 7 个指数的历史、特点和表现做了比较详细的述评。指出了他们基于这 7 个指数的单指数准则和 2 - 指数准则的不足之处。提出了超低显著性水平下的卡方准则,并部分重复他们的模拟例子,将卡方准则与这 7 个指数准则比较,结果说明新的卡方准则优于其中的 6 个,与另一个相当。最后简要说明了应当如何检视拟合指数进行模型检验和模型比较。

关键词 结构方程,模型检验,拟合指数,临界值,卡方检验。

分类号 B841.2

近年来,结构方程(structural equation)分析(包括验证性因子分析)在我国的心理、教育、社会、管理和传播等研究领域已经逐步有了一些应用。面对结构方程分析软件(如流行的 LISREL、EQS、AMOS)输出结果,如何检视诸多的拟合优度统计量(也称为拟合指数,以下简称指数)以检验或选择模型,是应用工作者很感兴趣的问题。本文研究的问题可以简单地归结为:第一,应当根据哪些指数来检验模型?第二,多大的指数值才算是一个"好"的模型?所谓指数,是反映模型与样本数据吻合程度的统计量,所以第一个问题就是用什么统计量来检验所拟合的模型。第二个问题类似于通常的假设检验中统计量(如 t 检验中的统计量)的临界值(以下称为界值)如何确定。实际上,这两个问题都是结构方程分析中很重要、且尚未很好解决的问题。

1 问题的背景

拟合指数的研究在结构方程分析的历史上受到了相当的重视,不少文献与指数的研究有关,而且其中部分文献被高频率引用(如文献[1~5]等)。从 1973 年 Tucker 和 Lewis[1] 提出的第一个指数 TLI 至 1996 年 Marsh 和 Balla[6] 提出的 NTLI,文献上正式发表、有名字的指数有 40 多个。指数研究的历史上可谓争论不断,争论的焦点就是上面提到的两个问题:一是哪些指数比较好?这个问题,从不同的角度分析会有不同的结果,这是指数家族庞大的重要原因。二是指数的界值(cutoff value)取多大合适?很长一段时间比较公认的标准是,相对指数在 0.9 或以上,拟合的模型可以接受[3];RMSEA 小于 0.05 表示模型拟合得好,在 0.05—0.08 之间表示模型基本可以接受[7,8]。新近的结果是,Hu 和 Bentler[9,10] 经过文献分析和模拟研究,对 ML 估计(极大似然估计,他们报告的模拟结果和本文的模拟结果都是基于 ML 估计)和 GLS 估计(广义最小二乘估计),推荐联合使用 SRMR 和以下指数中的一个:TLI、BL89、RNI(或 CFI)、Gamma Hat、Mc 和 RMSEA 来检验模型。他们建议的界值是 TLI、BL89、RNI(或 CFI)和 Gamma Hat 为 0.95,Mc 为 0.9,SRMR 为 0.08,RMSEA 为 0.06。由于作者之一 Bentler(EQS 的作者)的名气和载文刊物《Psychological Methods》[9](1996 年从《Psychological Bulletin》分出来的美国心理协会最主要的心理方法和统计期刊)的影响都很大,他们这种比较苛刻的新准则很可能成为检视指数的普遍标准(参见文献[11])。本文对他们的新准则进行评论和质疑。首先我们根据前人的研究,给出一个好指数应有的特征,对上述指数的表现和特征做出评价;然后就 Hu 和 Bentler[10] 模拟研究中,根据他们的新准则得到的第一类(拒真)和第二类(受伪)错误率的总和,来说明他们的新准则有何不妥;接着我们重复他们的部分模拟,结论是如果

收稿日期:2003 - 02 - 22

* 本研究得到全国教育科学"十五"规划教育部重点课题(DBA010169)以及香港中文大学和华南师范大学心理应用研究中心(教育部文科基地)资助。

通讯作者:温忠麟,Email: wenzl@scnu.edu.cn

选择适当的显著性水平,传统的 χ^2 检验与他们的 Mc 准则相当,比其他 6 个新准则都好。最后,我们简要说明了在检验和评价所拟合的模型时应当如何检视指数。为行文方便,将一个指数连同一个界值称为一个指数准则。

2　指数分类

将众多的指数按其功能分类,有助于对指数的理解和合理使用。1988 年 Marsh 等人最早提出了指数分类的雏形[12],有几个指数是受到他们分类的启发而提出的。后来出现了许多不同的分类方法(如文献[9,10,13~15]等,中文如[16]),分类的依据各有侧重。Marsh, Hau 和 Grayson[17] 将指数分成三大类:一类是绝对指数(absolute index 或 stand-alone index);一类是相对指数(relative index),也称为增值指数(incremental index)或比较指数(comparative index);还有一类是省俭指数(也称为简约指数,parsimony index)。绝对指数又可(有重叠地)分成:直接基于拟合函数的指数、拟合优度指数、基于离中参数的指数、近似误差指数和信息指数;相对指数又可分成有模型复杂性校正和无模型复杂性校正两种。为让读者熟悉 Hu 和 Bentler 建议使用的 7 个指数,表 1 列出了这些指数的计算公式、所属类型,表注中有它们的英文名称。

指数虽多,但除了 RMR、SRMR、GFI、AGFI 和 PGFI 外,其他都是统计量 χ^2(以下用 CHI 表示)的函数。绝对指数(如 GFI, SRMR, RMSEA)衡量了所考虑的理论模型(theory model)与样本数据的拟合程度,它只基于理论模型本身,不与别的模型比较。相对指数则将理论模型与虚模型(null model)比较,都可以写成这样的形式:$[f(CHI_N) - f(CHI_T)]/f(CHI_N)$,其中 $f(CHI)$ 是 CHI 的函数,$f(CHI_N)$ 和 $f(CHI_T)$ 分别是拟合虚模型和理论模型得到的 $f(CHI)$。它衡量了相对于虚模型的 $f(CHI)$ 而言,所检验的理论模型的 $f(CHI)$ 减少的比率。最简单的情形是 $f(x) = x$,此时 $f(CHI_N) = CHI_N$,$f(CHI_T) = CHI_T$,对应的相对指数是 NFI[6]。省俭指数是前两类指数派生出来的一类指数,某个指数对应的省俭指数是用省俭比 df_T/df_N 乘以该指数(其中分子和分母分别表示拟合理论模型和虚模型对应的自由度),目的是惩罚复杂模型(即自由度少的模型)。

表 1　Hu 和 Bentler 建议使用的 7 个指数

指数名称及定义	LISREL 缩写	分类
$TLI = (CHI_N/df_N - CHI_T/df_T)/(CHI_N/df_N - 1)$	NNFI	相对、校正
$BL89 = (CHI_N - CHI_T)/(CHI_N - df_T)$	IFI	相对、校正
$CFI = 1 - \max(CHI_T - df_T, 0)/\max(CHI_T - df_T, CHI_N - df_N, 0)$	CFI	相对
$Gamma\ Hat = p/[2(CHI_T - df_T)/(N-1) + p]$	无此指数	绝对、基于离中参数
$Mc = \exp\{-1/2[(CHI_T - df_T)/(N-1)]\}$	无此指数	绝对、基于离中参数
$SRMR = \mathrm{sqrt}\{2\sum_i\sum_j(s_{ij} - \hat{\sigma}_{ij})^2/[p(p+1)]\}$	Standardized RMR	绝对、近似误差
$RMSEA = \mathrm{sqrt}\{\max[(CHI_T - df_T)/(N-1), 0]/df_T\}$	RMSEA	绝对、近似误差、校正

注:N 是样本容量。p 是观测变量个数。$CHI = \chi^2$,它等于拟合函数的极小值的 $(N-1)$ 倍。df 表示自由度。CHI_N 和 df_N 分别表示拟合虚模型(即所有观测变量互不相关的模型,与其他模型相比,它的 CHI 和 df 都比较大)得到的 CHI 和自由度。CHI_T 和 df_T 分别表示拟合待检验的理论模型得到的 CHI 和自由度。max 是最大值函数,exp 是指数函数,sqrt 是平方根函数。s_{ij} 是样本相关系数;$\hat{\sigma}_{ij}$ 是由模型再生的相关系数的估计。TLI(Tucker-Lewis Index) 亦称为 NNFI(Non-Normed Fit Index,非范拟合指数)。BL89 是 Bollen 在 1989 年定义的指数,亦称 IFI(Incremental Fit Index,增值拟合指数)。CFI(Comparative Fit Index,比较拟合指数)是 RNI(Relative Non-centrality Index,相对离中指数)的规范形式(即大于 1 时就取为 1,小于 0 时就取为 0)。Mc (Measure of centrality) 和 Gamma Hat 都不是 LISREL 的指数,但 Mc 是 EQS 的指数。SRMR(Standardized Root Mean square Residual)是标准化残差均方根。RMSEA(Root Mean Square Error of Approximation)是近似误差均方根。

3　一个好的指数应当具有的特征

一个好的指数应当具有如下特征:与样本容量 N 无关;惩罚复杂的模型;对误设模型敏感。

3.1　不受样本容量的系统影响

一个理想的指数,应当与 N 无关或者关系不大,即不受样本容量的系统影响。许多研究者都注意到这一点(如[5,9,12,13,18])。道理并不复杂,因为一个指数如果会随样本容量而系统变化,那么由样本计算的指数是总体指数的有偏估计。一般说来,用它检验模型时,不同的 N 往往会有不同的结果,而且基于总体和基于样本的结果会出入很大。所以,一个指数如果会随样本容量而系统变化,那么就难于判断指数值多大,模型才能接受。Marsh 等

人[12]使用 Monte Carlo 模拟方法,对多种真模型(true model)及误设模型(misspecified model),检查了 31 个指数(其中几个指数当时还没有名字)受样本容量影响的情况。他们发现 TLI 是当时惟一被广泛使用的与样本容量相对无关的指数。McDonald 和 Ho[14]重申,如果一个指数相对地不受取样偏差的影响,可以用来检验模型[5]。

不过也有人指出[19],对于信息指数如 AIC,与 N 有关是合适的。这类指数的功能与其他指数不同,由于本文不涉及这类指数,不多介绍。

3.2 惩罚复杂的模型

一个理想的指数,应当惩罚复杂模型。对一组确定的变量及其关系,要估计的自由参数越多(即自由度越少),模型越复杂。在模型选择时,即使增加的自由参数本来是多余的,也会让人觉得模型拟合在改进。Steiger 和 Lind [20]注意到,使用那些不对复杂模型惩罚的指数,不可避免地导致选择最复杂的模型。例如,模型越复杂,CHI 会越小。在比较嵌套的两个模型时,LISREL 的作者 Jöreskog 和 Sörbom [20]提出的做法是,用两个模型的 CHI 之差作为新的 CHI,新的自由度是两个模型的自由度之差,如果卡方检验不显著,就认为新增的参数是多余的。这一做法后来一直被用于嵌套模型的比较与选择。Bozdogan[21]认为,模型选择需要研究者在过分拟合(参数过多)与不足拟合(参数过少)之间达至合适的平衡。按 McDonald 和 Marsh [5]说法就是,在拟合优度与模型省俭之间找到平衡点。

如果模型有自由参数去估计本来为零的参数,数值会降低的指数(这里指越大越好的指数)就是恰当地惩罚了模型复杂性。评价一个指数是否惩罚了复杂模型,可以通过分析指数公式的代数特性或者通过数据模拟进行(参见文献[22,23])。有些指数(如 BL89)虽然没有惩罚复杂模型,但考虑到了模型复杂性并做了校正,算是对复杂模型的一种索偿。

3.3 对误设模型的敏感性

一个理想的指数,用同一个总体的不同样本拟合同一个模型时,波动应当小。但用同一个样本拟合误设模型与拟合真模型相比,指数值应当有明显的区别,即对误设模型的敏感性要大。在实际问题中,真模型是不知道的,因而也就不知道所假设的理论模型是真模型还是误设模型。所以要考察指数是否对误设模型敏感,只好借助模拟研究。在模拟研究中,真模型是已知的,如果真模型中的某个参数是零,在理论模型中却自由估计(参数过多);或者,如果真模型中的某个参数不是零,在理论模型中却固定为零(参数过少),都是拟合了误设模型。

Marsh 和他的合作者多次指出[6,12,22,23],应当将指数是否能区分正确模型和各种不同程度的误设模型作为评价指数好坏的一个标准。Hu 和 Bentler[9,10]注意到,以往的研究中参数过少的误设模型中各种指数的相对敏感性被忽略。他们认为,如果拟合了错误的模型,而指数没有敏感地反映,这样的指数就不要用。他们根据前人的研究结果,挑选出 15 个较好的指数进行比较[9],推荐了其中的 7 个对参数过少的误设模型比较敏感的指数(见表 1)。

4 Hu 和 Bentler 建议的 7 个指数述评

4.1 指数 TLI

TLI [1](即 NNFI)是最早出现的相对指数(1973 年)。由于 TLI 可以超出 0 – 1 范围,让人有一种把握不住高低的感觉。历史上对它的误解和不满导致多个指数的产生。1980 年,Bentler 和 Bonett [3]提出两个相对指数:NNFI(即 TLI)和 NFI。NFI 的取值范围是 0 ~ 1,其中 NFI = 1 对应于最好的拟合,NFI = 0 对应于最差的拟合。Bentler 和 Bonett 在同一篇文章中提出的那个多少有点随意的 0.9 准则——指数超过 0.9 认为模型可以接受——后来被广泛应用,使相对指数受到欢迎。1986 年,Bollen[24]从 TLI 的代数形式推测它受样本容量的系统影响,因而提出了 BL86(即 RFI)并以为它不受 N 的影响。后来,他进一步界定了受 N 影响的含义。大量的实际数据研究表明 RFI 受 N 的影响,而 TLI 反而不受影响。1990 年,McDonald 和 Marsh[5],还有 Bentler[4],通过对 TLI 的公式进行代数变形发现,TLI 惩罚复杂模型。1994 年 Marsh 和 Balla[22]的数据模拟再次证实了这一点。对 TLI 的主要批评在于它的样本波动性较大,特别是当虚模型能很好地拟合样本数据的时候。许多研究者(如文献[4,13,25])都在数据模拟中发现 TLI 的样本波动性大这个问题,认为使用时需要慎重。为解决这个问题,1996 年 Marsh 等人[6]提出取值范围为 0 – 1 的 NTLI(规范的 TLI),但似乎没有多大的回应,至今还没有在主要的结构方程分析软件中出现。

4.2 指数 BL89 (即 IFI)

BL89 (即 IFI)是 1989 年 Bollen[26]为了纠正他早前提出的指数 BL86(即 RFI)的缺点(即系统地依

赖于样本容量)而提出的,由于它的公式有一个明显的校正以惩罚自由度小的模型,加上它不依赖于N,不少人都推荐使用[4,9,10,25]。但也有不同的看法[5,6],实际上,1988 年 Marsh 等人[12]就考虑过这个指数并批评过它,由于当时他们用的是不同的名字,被后来的研究者忽视了,结果是用 Bollen 的名字命名为 BL89[26]。Marsh[12]等人将 BL89 的公式变形,试图说明它并没有惩罚到复杂模型,所以认为公式中的校正是不适当的。实际情况是,BL89 中的校正是真的有利于省俭模型(因为分母中减去了自由度,如果两个模型的卡方相同,自由度大的模型对应的 BL89 较大),只是校正力度还不够而已。

4.3　指数 CFI(或 RNI)

1990 年国际著名的心理学刊物《Psychological Bulletin》第 2 期上紧挨着发表了两篇讨论拟合指数的文章[4,5]。前一篇的作者是 Bentler,提出了指数 CFI;后一篇的作者是 McDonald 和 Marsh,提出了指数 RNI。有趣的是两个指数几乎是一样的,不同之处是 RNI 的取值可以在 0 ~ 1 之外,而 CFI 相当于将 RNI 在 0 ~ 1 以外的值进行截取,取值范围变成 0 – 1。两篇文章的投稿时间和接受稿件的时间只相差了 3、4 个月。据说两篇稿件的审稿人相同,他们推荐两篇文章应当一起发表。后来软件上用的是 CFI,所以 RNI 比较少人知道。然而,这两篇文章的引用率都很高。

Bentler[4]用模拟的方法考察了 NFI、TLI、BL89(即 IFI)、RNI 和 CFI,结果发现只有 NFI 会受到样本容量的明显影响。对于基于真模型的小样本(N = 50),CFI 的 SD(标准差)比其他几个指数的都小,而 TLI 的则比较大,所以他比较推崇 CFI。Marsh 等人[6]则更喜欢 RNI,它和 CFI 一样有许多优点,如不受样本容量的系统影响,真模型的 RNI 的均值为 1,能够敏感地反映误设模型的变化。在模型检验和比较方面,RNI 与 CFI 是完全相同的(记得他们在 0 – 1 范围的值是相同的)。但在数据模拟方面,RNI 比 CFI 好[27],因为对于真模型,RNI 的期望值是 1,所以有半数左右的值会超过 1,对应的 CFI 都等于 1,结果 CFI 的期望值小于 1,估计偏低。

CFI(RNI)的不足之处是没有惩罚复杂模型,因为它们没有对模型复杂性作校正。

4.4　指数 Mc 和 Gamma Hat

这两个指数都是 Dk 的函数,Dk = (CHI – df)/(N – 1) 是对离中指数 NCP = CHI – df 的重新标度,以消除 N 的系统影响。1989 年 McDonald[28]提出的指数 Mc,理论取值落在 0 ~ 1 之间(但受抽样误差的影响有可能大于 1),基本上不受 N 的影响,能较好地区分真模型和误设模型[22]。同一年 Steiger[29]提出指数 Gamma Hat,他证明了 Gamma Hat 是 GFI 的一致估计(这样,Gamma Hat 可以作为 GFI 的校正,故亦称 Gamma Hat 为 AGFI*),不受 N 的影响,但没有惩罚复杂模型。

4.5　指数 SRMR

SRMR 早在 1981 年就有了[20],和 RMR 的区别只在于前者是用相关矩阵的结果,而后者是用协方差矩阵的结果。SRMR 取值范围是 0 ~ 1,但 RMR 的上限不是 1。1994 年 Marsh 等人[22]发现 RMR 对误设模型敏感,但受 N 的系统影响,建议不要使用它。1998 年,Hu 和 Bentler[9]的研究说明 SRMR 对误设模型敏感,但认为它受 N 的影响不严重,不仅将它选入 7 个指数之中,而且在两指数准则中,将它作为必选的一个。但从他们的模拟结果中可以发现,SRMR 的表现与真模型密切相关,对于他们所指的"简单"模型,SRMR 对误设模型非常敏感,受 N 的影响很小。但对他们所指的"复杂"模型,受 N 的影响很大,而对误设模型的敏感性降低了很多。

4.6　指数 RMSEA

RMSEA 是 1980 年 Steiger 和 Lind[2]提出的,被广泛使用至今。将离中指数 NCP 重新标度变为 Dk,规范 Dk(即把 Dk 的负值变为 0)变成 PDF,它除于自由度后的平方根就是 RMSEA,虽然 Dk 不受 N 的系统影响,但 RMSEA 受 N 的影响,好在影响不大。与 RMR 相比,RMSEA 受 N 的影响较小,对参数过少的误设模型还稍微敏感一些[22]。RMSEA 是比较理想的指数。Steiger[30]认为,RMSEA 低于 0.1 表示好的拟合;低于 0.05 表示非常好的拟合;低于 0.01 表示非常出色的拟合,这种情形应用上几乎碰不到。

5　Hu 和 Bentler 指数新准则的缺点

Hu 和 Bentler[10]在处理指数的界值问题时强调了推断统计上的一个准则,即合适的界值应当使检验结果的第一类错误率和第二类错误率之和最小。为此,他们比较了他们早先推荐的 7 个指数[9](见表 1),待选的界值包括传统的数值和他们的新准则中的数值。在他们的这两篇文章中,考虑的模型相同,都是两个验证性因子分析(CFA),一个是"简单"模型——3 个相关的因子,各有 5 个指标(indicator),共 15 个负荷。另一个是"复杂"模型——与

“简单”模型的不同之处在于除了 15 个负荷外,增加了 3 个跨因子的负荷,具体说就是 LX(1,3)、LX(4,2)、LX(9,3)不等于零。这样,“简单”模型共有 33 个非零参数(或自由参数),而“复杂”模型共有 36 个非零参数。对于每个模型类型,都考虑了真模型和两个参数过少的误设模型。对于“简单”模型,两个误设模型分别少了 1 个或 2 个因子之间的相关系数(即将其中的 1 个或 2 个相关系数固定为零)。对于“复杂”模型,两个误设模型分别少了 1 个或 2 个跨因子负荷。共设计了 7 种不同的分布类型(包括 3 个稳健性分布和 4 个非稳健性分布)和 6 种样本容量 N(150, 250, 500, 1000, 2500, 5000),模拟重复数是 200。

在他们更加苛刻的界值标准中,建议的界值是 TLI、BL89、RNI(或 CFI)、Gamma Hat 为 0.95,Mc 为 0.9, SRMR 为 0.08,RMSEA 为 0.06。从推断统计中我们知道,一个理想的检验准则应当是,不论 N 是大是小,接受真模型的概率基本上不变;而拒绝假模型的概率随 N 的增加而增加。换句话说,第一类错误率对不同的 N 基本上一致,而第二类错误率随 N 的增加而减少。然而,对于单个指数准则(即只看 7 个指数中的一个),检视一下文献[10]上的表 2,以“简单”模型的 CFI(或 RNI) = 0.95 为界值为例,对真模型的错误拒绝率分别为 25% ($N \leqslant 250$), 3.6% ($N = 500$)和 0.1% ($N \geqslant 1000$)。对上面这三种 N 和误设模型 1(少了因子之间的 1 个相关系数,即 PH(2,1)固定为零),正确拒绝率分别是 47.2%, 30.1%, 和 4.7%;对误设模型 2(少了因子之间的 2 个相关系数,即 PH(2,1)和 PH(3,1)都固定为零),正确拒绝率分别是 59.8%, 45.7%, 和 19.4%。因此,用准则 CFI(或 RNI) = 0.95 的话,第一类错误率与 N 密切相关,而第二类错误率竟然随 N 的增加而迅速增加。对于小 N,正确拒绝率都还不错;但对于大 N,正确拒绝率反而小得可怜。显然,这是不好的检验准则。关键是这个例子不是个别的,所有他们推荐的 7 个指数中,至少有一个误设模型是大 N 的拒绝率小于小 N 的拒绝率。

表 2 按 2 – 指数准则得到的两类错误率(%)之和

2 – 指数准则	稳健性分布						非稳健性分布					
SRMR = 0.08 和下列指数之一	150	250	500	1000	2500	5000	150	250	500	1000	2500	5000
TLI = 0.95	24.9	18.0	12.8	2.8	0.0	0.0	59.3	32.8	9.0	1.1	0.2	0.0
BL89 = 0.95	48.5	52.2	56.8	58.3	77.0	71.8	45.9	25.8	0.0	17.5	29.1	34.5
CFI = 0.95	47.0	50.8	56.0	57.7	66.7	71.7	46.5	25.8	15.7	17.1	28.9	34.1
G_Hat = 0.95	37.3	38.8	39.3	34.2	31.8	21.7	54.1	29.3	0.0	8.5	9.1	8.1
Mc = 0.9	7.8	2.0	0.2	0.0	0.0	0.0	79.6	62.6	26.1	2.8	0.5	0.0
RMSEA = 0.06	18.8	10.8	4.0	0.7	0.0	0.0	69.3	44.3	10.5	1.0	0.0	0.0

注:表中数据是“复杂”真模型和误设模型 1 按 2 – 指数准则得到的两类错误率(%)之和,根据[10]附录中的表 1 ~ 表 12 整理而成。所谓稳健性分布,简单地说是指 ML 估计在这样的分布下与正态分布的差别不大。第二行中的数字表示样本容量 N。G_Hat 表示 Gamma Hat。

下面再看 Hu 和 Bentler[10] 提出的 2 – 指数准则。所谓 2 – 指数准则,就是同时检验 SRMR 和其余 6 种指数中的一个,界值和单个指数的相同。例如,当 CFI < 0.95 且 SRMR > 0.08 时,认为模型拟合得不好(即拒绝所拟合的理论模型),否则就认为模型可以接受。为简便计,下面将这个 2 – 指数准则记为 CFI&SRMR。其他的 2 – 指数准则类似。表 2 列出了 Hu 和 Bantler 模拟例子中,“复杂”真模型和误设模型 1 按 2 – 指数准则得到的两类错误率之和。从表 2 可以看出,对于误设模型 1(少了一个跨因子负荷,即 LX(1,3)固定为零),两类错误率之和(%)在不同的指数准则和不同的 N 中变化很大,最高的为 79.6%,最低的为 0%。现仅以他们所论的稳健性分布结果为例做进一步的分析。图 1 是表 2 中稳健分布下 6 个 2 – 指数准则的比较。其中 N 为横坐标,两类错误率之和为纵坐标。对于 BL89&SRMR 和 CFI&SRMR,大 N (N = 2500 和 5000)的两类错误率之和大于中 N 的(N = 500 和 1000),后者又大于小 N 的(N = 150 和 250)。说明这两个 2 – 指数准则都不好。虽然 G_Hat&SRMR 的表现比上述两个 2 – 指数准则好得多,但大 N 对应的两类错误率之和仍然相当高(在非稳健分布条件下尤甚)。剩下的三个 2 – 指数准则(TLI&SRMR, Mc&SRMR 和 RMSEA&SRMR),看来很不错,对于大 N,两类错误率之和趋于零。不过,这三个准则对于小 N 都不好。连 Hu 和 Bentler[10] 他们自己都说,

TLI&SRMR、Mc&SRMR 和 RMSEA & SRMR 在小样本时往往过高地拒绝真模型(即第一类错误率大)。这种现象对于稳健分布尚不明显,但对于非稳健分布就非常明显了。如果将表 2 中非稳健分布下的结果画出来,会发现当 $N=150$ 和 $N=250$ 时,这三个 2 – 指数准则的两类错误率之和都比前述的三个大。

上述分析表明,Hu 和 Bentler 推荐的 6 个 2 – 指数准则都不理想。下面我们将重复他们的部分模拟,选择适当的显著性水平,用人们熟知的卡方检验做成新的卡方准则,并与他们的 7 个单指数准则比较。

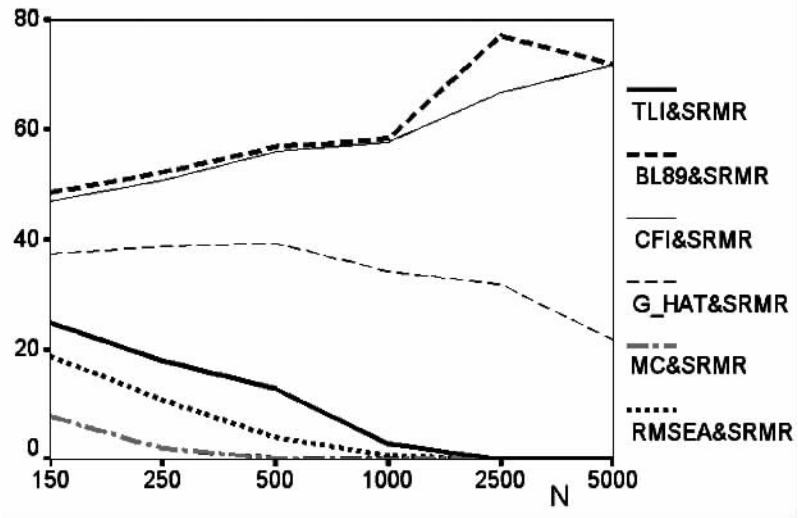

图 1　稳健性分布下不同 2 – 指数准则的比较(横轴为 N,纵轴为两类错误率之和)

6　卡方准则及其与 Hu 和 Bentler 指数准则的比较

为了探索合适的卡方检验并与 Hu 和 Bentler 指数新准则比较,我们部分地重复他们的数据模拟[9,10]。首先,我们只考虑正态分布。无论在理论上还是应用上,正态分布都是最重要的。如果他们的准则在正态分布下都不好,就不值得推广。第二,我们只考虑他们的"复杂"模型(模型和总体参数详见[9])。和他们的设计一样,我们使用真模型和两个参数过少的误设模型。样本容量与他们的设计一样,即 N = 150, 250, 500, 1000, 2500 和 5000,重复模拟 200 次。

6.1　选择卡方检验的显著性水平

统计量 CHI 在模型拟合方面有非常重要的作用。不仅 CHI 本身是一个重要的指数,而且大多数指数都是 CHI 的函数。但当 N 比较大时,CHI 检验被认为不好,因为 CHI 会随 N 的增大而不断增大,结果是任何模型都会被拒绝。注意到 CHI 只是近似服从 χ^2 分布,所以如何选择显著性水平对检验结果很重要。这里我们要突破显著性水平为 0.05 或 0.01 的限制。表 3 是我们的模拟结果——不同显著性水平 α 和不同样本容量 N 的 CHI 检验的两类错误率(%)之和。

从表 3 容易看出,当 $N=150$ 时,误设模型 1(少了一个跨因子负荷,即 LX(1,3) 固定为零)对应的两类错误率之和随 α 的增加而降低,说明第二类错误是主要的,因为第一类错误率总是随 α 的增加而增加。这容易解释,因为误设模型 1 与真模型相当接近,样本小时很难拒绝它,受伪的机会大,即第二类错误率大。然而,对于较大的 N,几乎都可以正确地拒绝误设模型 1,即第二类错误率很小,主要是第一类错误。误设模型 2(少了两个跨因子负荷,即 LX(1,3) 和 LX(4,2) 都固定为零)由于与真模型差别较大,即使是小样本,也基本上可以正确地拒绝它。事实上,对于误设模型 2,除了 $\alpha=0.0001$,其余的 CHI 检验的第二类错误率均为零,这时两类错误率之和几乎就是第一类错误率,随 α 的增加而增加(见表 3 下半部分)。注意到第一类错误率与显著性水平 α 相差很大,说明不能用传统的 $\alpha=0.05$。考虑选择 α 使两类错误率之和尽量小,由表 3 知,除了 $N=150$ 外,取 $\alpha=0.0005$ 是最好的(如果 $N\geqslant 500$,取 $\alpha=0.0001$ 甚至更好)。对于 $N=150$,权衡两个误设模型的结果,取 $\alpha=0.01$ 是比较合适的。称这种分别不同 N 在 CHI 检验中取不同显著性水平的做法为卡方准则。

表 3 CHI 检验的两类错误率(%)之和与 α 和 N 的关系

样本容量 N	150	250	500	1000	2500	5000
误设模型 1(少了 1 个跨因子负荷)						
α = 0.0001	38.0	2.5	0.0	0.0	0.0	0.0
0.0005	24.5	0.5	0.0	0.0	0.0	0.5
0.001	21.5	1.0	0.5	0.0	0.0	1.0
0.005	14.5	2.5	1.5	0.0	0.5	1.5
0.01	12.5	3.0	2.0	2.0	2.0	2.5
0.05	12.5	10.0	9.5	8.0	5.5	4.5
误设模型 2(少了 2 个跨因子负荷)						
α = 0.0001	1.0	0.0	0.0	0.0	0.0	0.0
0.0005	0.0	0.0	0.0	0.0	0.0	0.5
0.001	0.5	0.5	0.5	0.0	0.0	1.0
0.005	2.5	2.0	1.5	0.0	0.5	1.5
0.01	4.0	2.5	2.0	2.0	2.0	2.5
0.05	11.0	10.0	9.5	8.0	5.5	4.5

上述卡方准则虽然也是卡方检验,但与传统的卡方检验是不一样的。传统的卡方检验的临界值只与自由度有关,与 N 无关。但在我们的卡方准则中,临界值与 N 和自由度都有关,这一点与部分指数有异曲同工之处。就以表现比较好的 Mc(见表 1、表 4)为例,取界值为 0.9,当 Mc 大于 0.9 时认为模型拟合得好。由 $Mc = \exp\{-1/2[(CHI_T - df_T)/(N-1)]\}$,通过简单的代数运算可知,$Mc > 0.9$ 与 $CHI_T < df_T - 2(N-1)\ln(0.9)$ 等价。这说明 Mc 准则也是做卡方检验,临界值是 $df_T - 2(N-1)\ln(0.9)$,将 $\ln(0.9) = -0.105$ 代入得 $df_T + 0.21(N-1)$,这个临界值也与 N 和自由度都有关。当 N 很大时,临界值也很大,相应的显著性水平就很小。所以对于大样本,Mc 准则其实也是一种微显著性水平的卡方检验。

现在说明一下当 N 较大时卡方准则(还有 Mc 等准则)的 CHI 检验中可以取这么低的显著性水平的理据。通常的显著性检验(如两总体均值差异的 t 检验),显著性水平 α(即第一类错误率)不能太小,因为第二类错误率 β 会随 α 的减少而增加,结果两类错误率之和 $\alpha + \beta$ 可能随 α 的减少而变大,得不偿失。但在结构方程中,CHI 对误设模型很敏感。像上面的模拟例子,误设模型 1 与真模型其实相差很小(只差了一个跨因子负荷),但模拟结果告诉我们,除了 $N = 150$ 以外,CHI 都变得很大,即使 $\alpha = 0.0005$,第二类错误率仍然几乎为零。只要一个统计量足够敏感,能把与真模型差别很小的误设模型区分开来,意味着第二类错误不容易出现,在这个前提下,α 自然是越小越好。

6.2 卡方准则与 Hu 和 Bentler 指数新准则的比较

表 4 列出了基于卡方准则和 Hu – Bentler 的单指数准则检验的两类错误率之和。从 Hu – Bentler[9] 模拟设计的总体参数不难计算出,对于误设模型 1 和误设模型 2,总体的 SRMR 分别是 0.057 和 0.070,对于大样本($N \geqslant 1000$),对应的 SRMR 与总体的很接近。所以当用 0.08 作为界值时,对于大样本,都不能正确地拒绝误设模型,即第二类错误率是 100%。所以,在正态条件下,SRMR 准则的表现不好。对于 2 – 指数准则,由于一定有 SRMR 检验,所以对于大样本,第二类错误率也是 100%。

表 4 基于卡方准则和 Hu – Bentler 的单指数准则的两类错误率(%)之和

指数	CHI	TLI	BL89	CFI	G_Hat	Mc	SRMR	RMSEA
误设模型 1								
N = 150	12.5	28.5	50.5	47.5	42.5	11.0	81.0	33.5
250	0.5	17.0	52.0	50.0	36.0	1.0	94.5	17.5
500	0.0	9.0	52.5	51.0	32.0	0.0	100.0	4.5
1000	0.0	1.5	61.0	61.0	33.0	0.0	100.0	1.5
2500	0.0	0.0	68.0	67.5	26.5	0.0	100.0	0.0
5000	0.5	0.0	75.0	75.0	20.5	0.0	100.0	0.0
误设模型 2								
N = 150	4.0	0.5	2.5	2.0	1.0	5.0	38.0	0.0
250	0.0	0.0	0.0	0.0	0.0	0.0	59.0	0.0
500	0.0	0.0	0.0	0.0	0.0	0.0	85.5	0.0
1000	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
2500	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
5000	0.5	0.0	0.0	0.0	0.0	0.0	100.0	0.0

注:CHI 检验的显著性水平为 $\alpha = 0.01$ ($N = 150$)或 0.0005 ($N \geqslant 250$)。SRMR 的界值为 0.08,TLI、BL89、CFI 和 G_Hat 的界值为 0.95,Mc 的界值为 0.90,RMSEA 的界值为 0.06。

容易看出,卡方准则得到的两类错误率之和几乎都小于 Hu – Bentler 的 7 个单指数准则(见表 4)。只有 Mc 准则可以与卡方准则相媲美,RMSEA 和 TLI(即 NNFI)准则也比较好,SRMR 最差。所以比较的结果是,至少在正态条件下,卡方准则和 Mc 准则优于 Hu – Bentler 的其他准则。新近版本的 EQS 中有 Mc 指数,但 LISREL 中没有,需要用公式自己计算。要注意的是这里卡方检验时显著性水平与传统的 0.05 或 0.01 不同,当 N 大于 250 时,显著性水平取为 0.0005(当 N 大于 500 时还可以更小)。

7　结论和讨论

上面分析表明,Hu – Bentler 推荐的指数准则存在明显瑕疵,不宜推广(但本文主要是指出他们的新指数准则的不足之处,没有讨论为什么会出现这些问题,有兴趣的读者可参阅[31]),用他们的模型模拟分析,结果表明我们提出的卡方准则和 Mc 准则优于他们的其他 6 个准则,至少在正态情形如此。因此,就本研究结果而言,卡方准则可以列入优秀指数准则之一。更明确一些,卡方准则的显著性水平是:$N \leqslant 150$ 时 $\alpha = 0.01$,$N = 200$ 时 $\alpha = 0.001$,$N = 250$ 时 $\alpha = 0.0005$,$N \geqslant 500$ 时 $\alpha = 0.0001$。但对于 $N \geqslant 1000$ 的大样本,$\alpha = 0.0001$ 还是不够小,即卡方值往往很大而导致拟合得很不错的模型都被拒绝。因此我们建议在 $N < 1000$ 时才使用卡方准则。

有两点需要特别指出。第一,我们质疑 Hu – Bentler 推荐的指数准则,并不是否定相应的指数本身。指数和指数准则是两个不同的概念,一个指数连同一个界值才是一个指数准则。同一个指数,界值不同的指数准则产生的两类错误率之和可能很不一样。我们的模拟结果只是说明,他们提出的苛刻的新界值不合适。根据好指数的三个特征,他们筛选的 7 个指数在现有的诸多指数中是比较好的,其中 NNFI、CFI、RMSEA 和 Mc 比较优秀。第二,我们提出的卡方准则和 Mc 准则在模拟研究中优于其他 6 个指数准则,但卡方准则作为一种新的思路(即直接用 N 去调整界值,而不像许多指数那样用 N 去调整 CHI,然后与固定的界值比较),不少问题有待研究,例如,在非正态情形,卡方准则的表现如何? 对于更一般的真模型和误设模型,它的表现又如何? 当然,对于后面一个问题,所有的指数准则都有待研究。

那么,就目前的认识水平,应当如何检视指数呢? 根据本文的讨论,我们建议使用如下的指数和传统界值:NNFI 和 CFI(界值为 0.9);Mc(界值 0.85);RMSEA(界值为 0.08)。我们的卡方准则值得推荐。只要根据其中多个准则(包括卡方准则)模型是好的拟合,就可以从某些角度认为模型可以接受。当然,其他指数也要参考,不能离界值太远。在比较嵌套的两个模型时,除了做 Jöreskog 和 Sörbom [20] 提出的卡方检验(见 3.2 节),可以参考或报告上述指数的变化情况。在一般的模型比较时,除了报告上述指数外,可以报告 CHI/df,这个值小的模型较好。

在实际应用中,模型检验和模型选择除了检视上述指数外,应当综合考虑问题的背景、参数估计值的意义、模型的可解释性、各备选模型的表现和省俭原则等。研究者的专业知识、智慧和使用结构方程的经验相结合,有望对拟合的模型做出合适的评价或从众多的备选模型中选出最好的模型。

参　考　文　献

1　Tucker L R, Lewis C. The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 1973, 38: 1 ~ 10

2　Steiger J H, Lind J M. Statistically – based tests for the number of common factors. Paper presented at the Psychometrika Society Meeting, IowaCity, May, 1980

3　Bentler P M, Bonett D G. Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 1980, 88: 588 ~ 606

4　Bentler P M. Comparative fit indices in structural models. Psychological Bulletin, 1990, 107: 238 ~ 246

5　McDonald R P, Marsh H W. Choosing a multivariate model: Noncentrality and goodness – of – fit. Psychological Bulletin, 1990, 107: 247 ~ 255

6　Marsh H W, Balla J R, Hau K T. An evaluation of incremental fit indices: A clarification of mathematical and empirical processes. In: Marcoulides G A, Schumacker R E eds. Advanced structural equation modeling techniques. Hillsdale, NJ: Erlbaum, 1996. 315 ~ 353

7　Browne M W, Cudeck R. Alternative ways of assessing model fit. In: Bollen K A, Long J S eds. Testing Structural Equation Models. Newbury Park, CA: Sage, 1993. 136 ~ 162

8　Jöreskog K G, Sörbom D. LISREL 8: Structural equation modeling with the SIMPLIS command language. Chicago: Scientific Software International, 1993

9　Hu L, Bentler P M. Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, 1998, 3: 424 ~ 453

10　Hu L, Bentler P M. Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 1999, 6: 1 ~ 55

11　Byrne, B. M. Structural equation modeling with AMOS: Basic con-

cepts, applications and programming. Mahwah, NJ: Erlbaum,2001
12 Marsh H W, Balla J R, McDonald R P. Goodness – of – fit indices in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 1988, 103: 391 ~ 410
13 Bollen K A, Long J S. Introduction. In: Bollen K A, Long J S eds. Testing Structural Equation Models. Newbury Park, CA: Sage, 1993. 1 ~ 9
14 McDonald R P, Ho M R. Principles and practice in reporting structural equation analyses. Psychological Methods, 2002, 7: 64 ~ 82
15 Mulaik S A, James L R, Alstine J V, Bennett N, Lind S, Stilwell C D. Evaluation of goodness – of – fit indices for structural equation models. Psychological Bulletin, 1989, 105: 430 ~ 445
16 Wang Q, Li J B. Confirmatory Factor Analysis(in Chinese). Hangzhou: Zhejiang University Press, 2002. 117 ~ 135
(王权,李金波. 实证性因素分析. 杭州:浙江大学出版社,2002. 117 ~ 135)
17 Marsh H W, Hau K T, Grayson D. Goodness of fit evaluation in structural equation modeling. In: Maydeu – Olivares A, McCardle J eds: Psychometrics. A Festschrift to Roderick P. McDonald. NJ: Erlbaum. (in press)
18 Anderson J C, Gerbing D W. The effect of sampling error on convergence, improper solutions, and goodness – of – fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 1984, 49: 155 ~ 173
19 Browne M W. Cross – validation methods. Journal of Mathematical Psychology, 2000, 44: 108 ~ 132
20 Jöreskog K G, Sörbom D. LISREL: User's Guide. Chicago: International Education Services,1981
21 Bozdogan H. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 1987, 52: 345 ~ 370
22 Marsh H W, Balla J R. Goodness – of – fit indices in confirmatory factor analysis: The effect of sample size and model complexity. Quality & Quantity, 1994, 28: 185 ~ 217
23 Marsh H W, Hau K T. Assessing goodness of fit: Is parsimony always desirable? Journal of Experimental Education, 1996, 64: 364 ~ 390
24 Bollen K A. Sample size and Bentler and Bonett's nonnormed fit index. Psychometrika, 1986, 51: 375 ~ 377
25 Gerbing D W, Anderson J C. Monte Carlo evaluations of goodness – of – fit indices for structural equation models. In: Bollen K A, Long J S eds. Testing Structural Equation Models. Newbury Park, CA: Sage, 1993. 40 ~ 65
26 Bollen K A. A new incremental fit index for general structural equation models. Sociological Methods and Research, 1989, 17: 303 ~ 316
27 Goffin R D. A comparison of two new indices for the assessment of fit of structural equation models. Multivariate Behavioral Research, 1993, 28: 205 ~ 214
28 McDonald R P. An Index of Goodness – of – fit based on noncentrality. Journal of Classification, 1989, 6: 97 ~ 103
29 Steiger J H. EzPATH: A supplementary module for SYSTAT and SYSGRAPH. Evanston, IL: SYSTAT, 1989
30 Steiger J H. Structure model evaluation and modification: An interval estimation approach. Multivariate Behavioral Research, 1990, 25: 173 ~ 180
31 Marsh H W, Hau K T, Wen, Z. In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralising Hu & Bentler's (1999) findings. Structural Equation Modeling. (in press)

STRUCTURAL EQUATION MODEL TESTING: CUTOFF CRITERIA FOR GOODNESS OF FIT INDICES AND CHI-SQUARE TEST

Wen Zhonglin[1,2], Hau Kit-Tai[1], Herbert W. Marsh[3]
([1] *Faculty of Education, The Chinese University of Hong Kong, Hong Kong, China*)
([2] *Faculty of Education, South China Normal University, Guangzhou* 510631, *China*)
([3] *Faculty of Education, University of Western Sydney, Australia NSW*2560)

Abstract

Recent cutoff criteria for goodness of fit indices in structural equation analysis proposed by Hu and Bentler(1998, 1999) were discussed. After reviewing their recommended 7 indexes, we demonstrated the inappropriateness of their single index or 2-index combinational rules. Subsequently we proposed that a new rule based on Chi-square test at a certain extremely low significant level was more useful for their purposes. Comparing with Hu and Bentler's rules, we showed that our proposed rule is better and more appropriate. Finally, we discussed guidelines on ways to evaluate a fitted model or to compare alternative models.

Key words structural equation, model testing, fit index, cutoff value, Chi-square test.