

· 基础理论与方法 ·

代谢综合征组分的正交因子分析

侯旭宏 贾伟平 包玉倩 陆俊茜 吴元民 顾惠琳 左玉华 姜素英 项坤三

【导读】 结合代谢综合征(MS)组分的正交因子分析实例,介绍正交因子分析模型的原理及其在医学研究中的应用。使用 1998-2001 年上海社区人群 MS 患病率的现况研究资料,选取 35~65 岁资料完整的 1877 例女性作为研究对象,使用 SPSS 统计软件对 MS 相关组分进行正交因子分析,并计算因子得分。因子分析提取的前 6 个互不相关的公共因子可以基本反映原始指标 86% 的信息。通过正交旋转后 6 个公共因子的实际意义很清晰,依次主要反映肥胖、血压、血糖、胰岛素、TG 和 HDL-C 指标的信息。根据给出的因子得分矩阵计算个体各因子得分及总得分。结果表明,MS 的同一组分(除血脂外)内的变量呈高度相关性,但不同组分间虽有统计学关联,但非高度相关。正交因子分析的价值在于探查有高度相关关系的变量群,进而为探究其共同的潜在病理生理机制提供线索。

【关键词】 代谢综合征; 正交因子分析

Orthogonal factor analysis on metabolic syndrome HOU Xu-hong*, JIA Wei-ping, BAO Yu-qian, LU Jun-xi, WU Yuan-min, GU Hui-lin, ZUO Yu-hua, JIANG Su-ying, XIANG Kun-san. Shanghai Clinical Center for Diabetes, Shanghai Diabetes Institute, Department of Endocrinology and Metabolism, Shanghai Jiaotong University Affiliated Sixth People's Hospital, Shanghai 200233, China
Corresponding author: JIA Wei-ping, Email: wpjia@yahoo.com

【Introduction】 To elucidate the principal of orthogonal factor analysis, using an example of factor analysis of metabolic syndrome. The basic structures and the fundamental concepts of orthogonal factor analysis were introduced and data involving 1877 women aged of 35-65 years, selected from a cross-sectional study, which was conducted in 1998-2001 in Shanghai, were included in this study. Factor analysis was carried out using principle components analysis with Varimax orthogonal rotation of the components of the metabolic syndrome. The different components of the metabolic syndrome were not linked closely with the other components and loaded on the six different factors, which mainly reflected by the variables of obesity, blood pressure, plasma glucose, plasma insulin, triglycerides and HDL-cholesterol respectively. Six major factors of the metabolic syndrome were uncorrelated with each other and explained 86% of the variance in the original data. The factor score and total factor score for the individual could be obtained according to the component score coefficient matrix. Although the components of the metabolic syndrome were related statistically, the finding of six factors suggested that the components of the metabolic syndrome did not show high degrees of intercorrelation. As a linear method of data reduction, the mode reduced a large set of measured intercorrelation variables into a smaller set of uncorrelated factors, which explained the majority of the variance in the original variables. Factor analysis was well suited for revealing underlying patterns or structure among variables showing high degrees of intercorrelation.

【Key words】 Metabolic syndrome; Orthogonal factor analysis

正交因子分析(orthogonal factor analysis)是用较少的彼此互不相关的新随机变量来代替原来较多的高度相关的原始随机变量,同时这些新随机变量

能尽可能多地反映原始变量的信息(方差意义下的信息量)。近来因子分析在医学研究,如蛋白组学和代谢组学的研究中已被广泛应用来寻找病因标志物。本文旨在介绍因子分析的基本原理及其应用。在实例分析中,使用上海社区人群代谢综合征(MS)的横断面调查资料,以有因子分析实际意义的人群实测数据为例,对 MS 各组分进行因子分析。

基本原理^[1-3]

1. 正交因子模型的定义: 设观测值 $X =$

基金项目:上海市科委重大项目基金资助(04DE19501)

作者单位:200233 上海市糖尿病临床医学中心 上海市糖尿病研究所 上海交通大学附属第六人民医院(侯旭宏、贾伟平、包玉倩、陆俊茜、项坤三);上海市华阳街道社区卫生服务中心(吴元民、顾惠琳);上海市曹杨街道社区卫生服务中心(左玉华、姜素英)

通讯作者:贾伟平, Email: wpjia@yahoo.com

$(X_1, \dots, X_p)'$ 是 p 个代表研究指标的随机向量。均值 $E(X) = [E(X_1), \dots, E(X_p)]' = \mu = (\mu_1, \dots, \mu_p)'$, 协方差阵 $D(X) = \Sigma = (\sigma_{ij})_{p \times p}$ 。为简化研究指标引入公共因子随机向量 $F = (F_1, \dots, F_m)'$ ($m < p$), 公共因子的均值 $E(F) = (0, \dots, 0)'$, 其协方差阵 $D(F) = I_m$ (m 阶单位矩阵, 表示 F 的各分量方差为 1, 且互不相关, 即正交)。又引入特殊因子 $\epsilon = (\epsilon_1, \dots, \epsilon_p)'$ 与 F 互不相关, 且 $E(\epsilon) = 0$, 同时特殊因子 ϵ_i 间也互不相关。

若 $X = (X_1, \dots, X_p)'$ 满足以下的模型:

$$X = \mu + BF + \epsilon \tag{1}$$

则称模型为正交因子模型。其中 F_1, \dots, F_m 称为 X 的公共因子, 一般对 X 的每一分量都有作用。随机向量 $\epsilon = (\epsilon_1, \dots, \epsilon_p)'$, $\epsilon_1, \dots, \epsilon_p$ 称为 X 的特殊因子, 只对 X_i 起作用。可见因子分析是将原始随机变量表示为公共因子和特殊因子的线性组合。

$B = (b_{ij})_{p \times m}$ 是待估系数矩阵, 称为因子载荷矩阵, 是 SPSS 因子分析输出的重要结果。 b_{ij} ($i = 1, \dots, p; j = 1, \dots, m$) 称为第 i 个变量在第 j 个因子上的载荷 (因子载荷)。因为有 $B = COV(X, F)$, 可见 B 中元素 b_{ij} 刻画了随机变量 X_i 与公共因子 F_j 之间的相关性。

当变量 X_i 是标准化变量 [即 $E(X_i) = 0, Var(X_i) = 1$] 时 (SPSS 在进行因子分析前会先对原始变量进行标化), 原指标 X_i 与新指标 F_j 的相关系数 $\rho(X_i, F_j)$:

$$\rho(X_i, F_j) = \frac{COV(X_i, F_j)}{\sqrt{Var(X_i)}\sqrt{Var(F_j)}} = COV(X_i, F_j) = b_{ij} \tag{2}$$

2. 正交因子模型求解: 即对因子载荷阵 B 的求解。

设 $X = (X_1, \dots, X_p)'$ 是 p 维随机向量, 其协方差阵 $D(X) = \Sigma = (\sigma_{ij})_{p \times p}$ 。 Σ 的特征值为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, 与特征值相应的单位正交特征向量分别为列向量 l_1, l_2, \dots, l_p 。当使用主成分法求解正交因子模型时, 对协方差阵 Σ 进行了近似谱分解。当最后 $p - m$ 个特征值较小时, 因子模型中 B 的初始近似解为 $(\sqrt{\lambda_1} l_1, \dots, \sqrt{\lambda_m} l_m)_{p \times m}$, 称为因子模型的主成分分解。通常选取 $|b_{ij}| > 0.30$ 的系数来建立因子模型。

当求得因子载荷阵 B 后, 可以求得: ①第 j 个

公共因子 F_j 对 X 的所有分量 X_1, \dots, X_p 的方差贡献 $s_j = \sum_{i=1}^p Var(b_{ij}F_j) = \sum_{i=1}^p b_{ij}^2 = \lambda_j$ ($j = 1, 2, \dots, m$)。 s_j 是衡量第 j 个公共因子相对重要性的指标, 据此可以找出最有影响的公共因子。②变量 X_i 共同度 h_i^2 (communalities), 又称为公因子方差, $h_i^2 = \sum_{j=1}^m b_{ij}^2$ ($i = 1, 2, \dots, p$), 与 s_j 的统计意义不同, 它反映了变量 X_i 对公共因子 F_1, \dots, F_m 的共同依赖程度。

3. 寻找最优载荷矩阵: 当求得的初始因子载荷矩阵相应列 (例如第 j 列) 的数值较为平均 (未趋向 0 或 1) 时, 说明公共因子 F_j 对原始变量的分类聚集不明确, 也就是说该公共因子的实际意义不清晰。此时可对初始因子载荷阵施行旋转变换, 使其每一列元素的平方向 0 或 1 两极转化, 来提供清晰的因子模型。最常使用的方法为方差最大正交旋转, 即反复右乘各种正交矩阵 Γ 。使得 B 的每一列 (即因子载荷向量) 数值分散 (即其平方趋于 1 或 0), 这样因子载荷阵的样本方差就越大, 相应的因子分析模型就越具有简化结构。即 ΓB 具有更明显的实际意义。

4. 因子得分: 在实际分析中, 常需要用公共因子代替原始变量, 来描述研究对象的特征。因而往往将公共因子表示为变量的线性组合, 即因子的得分函数。

$$F_j = \beta_{1j}X_1 + \beta_{2j}X_2 + \dots + \beta_{pj}X_p \quad (j = 1, \dots, m) \tag{3}$$

用它可计算每个研究对象第 j 个公因子的得分, 估计因子得分的常用方法有回归法和加权最小二乘法。SPSS 可以将计算出的经过标化的因子得分作为新变量存在数据库中以供使用。

按每个公因子得分可对每个研究对象进行排序。也可以计算每个研究对象的因子总得分, 即每个公因子得分乘以该因子的方差贡献率的累计。

$$\text{因子总得分} = \sum_{i=1}^m [F_i \times (\lambda_i / \sum_{i=1}^p \sigma_{ij})] \tag{4}$$

实例分析

MS 是 2 型糖尿病和动脉硬化性心血管疾病的一组危险因素, 包括高血压、糖耐量异常、血脂异常和肥胖。目前可根据下列指标判定 MS: ①肥胖: 腰围 (WC)、腰臀比 (WHR)、BMI; ②血压: SBP、DBP; ③血糖: 空腹血糖 (FPG)、餐后 2 h 血糖 (PG2h); ④胰岛素水平: 空腹胰岛素 (FINS)、餐后 2 h 胰岛素

(INS2h);⑤血脂:HDL-C、TG。

1.对象与方法:采用正交因子分析来探讨 MS 组分间的相关关系。利用上海市糖尿病研究所 1998 年 9 月至 2001 年 11 月对华阳、曹杨社区 15~95 岁人群进行的 MS 横断面调查资料。选取年龄 35~65 岁且相关分析资料完整的 1877 例女性作为研究对象,对 MS 组分的相关变量进行因子分析。

· 使用 SPSS 11.5 软件对资料进行因子分析的主要步骤:

(1)因子分析中随机变量 (X_1, \dots, X_p) (本例中 $p=11$) 的选择:即 analysis→data reduction(资料简化)→factor analysis(因子分析)。打开主对话框,选取 MS 各组分变量 (WC、WHR、BMI、SBP、DBP、FPG、PG2h、FINS、INS2h、HDL-C、TG) 进入 variables 框,作为待分析变量。

[说明] 在实际应用中,为了消除量纲不同带来的不合理影响,常将原始数据进行标准化,用标化后的样本协方差阵 S 作为总体协方差阵 Σ 的估计。SPSS 软件会自动对原始变量先进行标化后再进行因子分析。

(2)因子模型建立的可行性分析:打开 descriptive 子对话框,在 correlation matrix 复选框中,选中 KMO (Kaiser-Meyer-Olkin Measure of Sampling Adequacy) and Bartlett's test of sphericity (检验变量间的相关性),选中 coefficient(给出相关系数阵)和 significance level(给出所有变量相关系数单侧检验的 P 值)。相关系数阵及其显著性水平(表 1)。

[说明] KMO 是通过比较相关系数和偏向关系系数的大小来评估样本量是否充足,其取值范围 0~1。KMO 取值越小 (<0.6),说明越不适宜做因子分析。一般 KMO 值 >0.8 时效果好, 0.7 时效果尚

可。本例 $P=0.707$ 。说明可进行因子分析。Bartlett's test of sphericity 是对无效假设(相关阵变量间不相关)进行检验。本例 P 为 0.000,说明变量间相关性好,可以建立因子模型。

(3)判断因子分析中需要保留的公因子个数:在 extraction 子对话框的 method 下拉列表中选择 principle components(主成分分析法),在 statistics 复选框中选中 initial solution 来输出协方差阵 Σ 的特征值 (λ_i) , 方差贡献率 $(\lambda_i / \sum_{i=1}^p \sigma_{ij})$, 方差累积贡献率 $(\sum_{i=1}^m \lambda_i / \sum_{i=1}^p \sigma_{ij})$ (表 2), 及变量 X_i 的共同度 (communalities)(该研究结果未列出)。

[说明] 根据 initial solution 输出结果,可以帮助判断因子模型中需要保留的公因子的个数 m 。

选择合适的公共因子个数 m 的方法:①选择主成分分析结果中特征值 $\lambda_i > 1$ 的主成分个数 m ;②将主成分按照特征值从大到小的顺序排列,取使主成分累积贡献率 $\sum_{i=1}^m \lambda_i / \sum_{i=1}^p \lambda_i \geq 85\%$ 的最少主成分个数 m 。

(4)初始(未经旋转的)因子模型的建立:在 extraction 子对话框的 method 下拉列表中选择 principle components(主成分分析法),在 extraction 子对话框的 extract 单选框组,选择公共因子提取标准——number of factors(自定义提取公共因子的数量),本研究自定义提取 6 个公共因子。选中 unrotated factor solution(显示未经旋转变换的因子提取结果),结果见表 3。

[说明] SPSS 中选择公共因子提取标准有① eigenvalues over 1(特征值 >1)或② number of factors(自定义提取公共因子的数量)。本研究如以提取标准①作为选择公共因子提取标准,则可提取 4 个公

表1 MS 组分间的相关系数矩阵

指标	WC	WHR	BMI	SBP	DBP	FPG	PG2h	FINS	INS2h	TG	HDL-C
WC	-	0.83 ^a	0.86 ^a	0.35 ^a	0.33 ^a	0.27 ^a	0.37 ^a	0.29 ^a	0.29 ^a	0.25 ^a	0.16 ^a
WHR		-	0.61 ^a	0.33 ^a	0.28 ^a	0.26 ^a	0.37 ^a	0.26 ^a	0.29 ^a	0.26 ^a	0.15 ^a
BMI			-	0.35 ^a	0.34 ^a	0.27 ^a	0.35 ^a	0.33 ^a	0.30 ^a	0.24 ^a	0.14 ^a
SBP				-	0.80 ^a	0.23 ^a	0.29 ^a	0.24 ^a	0.24 ^a	0.24 ^a	0.06 ^b
DBP					-	0.23 ^a	0.27 ^a	0.24 ^a	0.21 ^a	0.25 ^a	0.06 ^b
FPG						-	0.68 ^a	0.25 ^a	0.17 ^a	0.37 ^a	0.03
PG2h							-	0.22 ^a	0.32 ^a	0.32 ^a	0.09 ^a
FINS								-	0.40 ^a	0.24 ^a	0.07 ^b
INS2h									-	0.23 ^a	0.06 ^b
TG										-	0.10 ^a
HDL-C											-

注:^a $P < 0.001$; ^b $P < 0.05$

表2 各公共因子的特征值及其对总方差的贡献率(%)

主成分	特征值			初始载荷平方和 $s_j = \sum_{i=1}^6 b_{ij}^2$			旋转后的载荷平方和 $s_j^* = \sum_{i=1}^6 (b_{ij}^*)^2$		
	特征值	方差贡献率	方差累积贡献率	特征值	方差贡献率	方差累积贡献率	特征值	方差贡献率	方差累积贡献率
1	4.1	37.6	37.6	4.1	37.6	37.6	2.5	23.1	23.1
2	1.3	12.2	49.8	1.3	12.2	49.8	1.8	16.4	39.6
3	1.3	11.5	61.4	1.3	11.5	61.4	1.7	15.3	54.9
4	1.0	9.2	70.5	1.0	9.2	70.5	1.4	13.1	67.9
5	1.0	8.8	79.3	1.0	8.8	79.3	1.0	9.1	77.0
6	0.7	6.6	85.9	0.7	6.6	85.9	1.0	8.9	85.9
7	0.6	5.7	91.6	-	-	-	-	-	-
8	0.4	3.5	95.1	-	-	-	-	-	-
9	0.3	2.5	97.6	-	-	-	-	-	-
10	0.2	1.8	99.4	-	-	-	-	-	-
11	0.1	0.6	100.0	-	-	-	-	-	-

注:协方差阵 Σ 的特征值(λ_i),方差贡献率($\lambda_i/\sum_{i=1}^6 \sigma_{ij}$),方差累积贡献率($\sum_{i=1}^m \lambda_i/\sum_{i=1}^6 \sigma_{ij}$)

共因子。为方便说明问题,本研究采用标准②自定义提取 6 个公共因子,此时累积贡献率达 85.9% (表 3)。

表3 初始因子载荷矩阵 $B(b_{ij})$

指标	因子 1	因子 2	因子 3	因子 4	因子 5	因子 6
WC	0.81 ^a	-0.51 ^a	0.01	-0.14	0.12	0.05
WHR	0.75 ^a	-0.44 ^a	0.05	-0.13	0.10	0.07
BMI	0.78 ^a	-0.43 ^a	-0.01	-0.06	0.12	0.02
SBP	0.63 ^a	0.31 ^a	-0.62 ^a	-0.09	-0.04	-0.07
DBP	0.61 ^a	0.34 ^a	-0.64 ^a	-0.10	-0.06	-0.04
FPG	0.56 ^a	0.46 ^a	0.48 ^a	-0.27	0.04	-0.14
PG2h	0.65 ^a	0.34 ^a	0.42 ^a	-0.22	0.03	-0.31 ^a
FINS	0.51 ^a	0.10	0.06	0.64 ^a	0.02	0.01
INS2h	0.51 ^a	0.07	0.10	0.64 ^a	0.06	-0.17
TG	0.49 ^a	0.32 ^a	0.21	0.01	-0.22	0.73 ^a
HDL-C	-0.20	0.25	-0.07	-0.01	0.93 ^a	0.18
方差贡献率(%)	37.60	12.20	11.50	9.20	8.80	6.60

注:公共因子提取的方法为主成分分析法;^a 原始变量在公因子上的载荷量的绝对值 $|b_{ij}| > 0.30$

(5)初始因子模型的旋转:在 rotation 子对话框的 method 复选框中选择 varimax(最大正交旋转)。旋转后的因子模型见表 4。

表4 方差最大正交旋转后的因子载荷矩阵 $B^*(b_{ij}^*)$

指标	因子 1	因子 2	因子 3	因子 4	因子 5	因子 6
WC	0.95 ^a	0.15	0.13	0.13	-0.06	0.06
WHR	0.85 ^a	0.12	0.15	0.11	-0.06	0.10
BMI	0.85 ^a	0.17	0.12	0.19	-0.04	0.05
SBP	0.20	0.91 ^a	0.12	0.12	-0.01	0.06
DBP	0.16	0.92 ^a	0.11	0.11	-0.01	0.09
FPG	0.12	0.09	0.88 ^a	0.08	0.02	0.20
PG2h	0.23	0.13	0.87 ^a	0.17	-0.05	0.03
FINS	0.16	0.12	0.07	0.79 ^a	-0.01	0.16
INS2h	0.16	0.09	0.15	0.82 ^a	-0.02	-0.01
TG	0.13	0.13	0.21	0.13	-0.05	0.94 ^a
HDL-C	-0.10	-0.02	-0.02	-0.03	0.99 ^a	-0.04
方差贡献率(%)	23.10	16.40	15.30	13.10	9.10	8.90

注:^a 原始变量在旋转后的公因子上的载荷量的绝对值 $|b_{ij}^*| >$

(6)计算因子得分:打开 scores 子对话框,选择 save as variables(将计算出的经过标化的因子得分作为新变量存在数据库中),同时激活 method 单选框组,选择计算因子得分的方法,本例选 regression(回归法);选择 display factor score coefficient matrix(显示因子得分的系数矩阵);系统会同时给出因子得分的系数矩阵。

2. 结果:

(1)MS 组分间的相关系数见表 1。除 HDL-C 外,各变量间均呈正相关($P < 0.05$),HDL-C 与 FPG 无统计学相关,与其他变量均呈负相关($P < 0.05$)。

(2)各主成分的特征值及其对总方差的贡献率、协方差阵 Σ 的特征值(λ_i),方差贡献率($\lambda_i/\sum_{i=1}^6 \sigma_{ij}$)、方差累积贡献率($\sum_{i=1}^m \lambda_i/\sum_{i=1}^6 \sigma_{ij}$),见表 2。

表 2 第 1 列给出所有主成分(11 个),按照特征值从大到小的顺序排列。即第一主成分的(方差意义下的)信息量大于等于第二主成分的信息量,大于等于第 p 主成分的信息量。第一主成分的特征值为 4.1,它解释了总变异的 37.6%,余类推。因子载荷阵 $B(m=6)$ 旋转前后,6 个公共因子的累积贡献率均为 86%。但旋转前后各公共因子相应特征值所反映原始变量的信息量发生改变。

(3)初始因子载荷矩阵 $B(m=6)$ 见表 3。在因子模型中保留了 6 个公共因子,在表 3 初始因子载荷矩阵 B 中,有 10 个变量在因子 1 上的载荷量的绝对值 > 0.30 ,有 8 个变量在因子 2 上的载荷量的绝对值 > 0.30 ,有 4 个变量在因子 3 上的载荷量的绝对值 > 0.30 。初始因子载荷矩阵 B 中公共因子的实际意义不清晰。

(4) 方差最大正交旋转后的因子载荷矩阵 B^* ($m=6$) 见表 4。此时 6 个公共因子的实际意义很清晰。第 1 个公共因子对应的列(即原变量在第 1 个公共因子的载荷或因子载荷), 除 WC(0.95)、WHR(0.85) 和 BMI(0.85) 较大外, 其余各变量数值的绝对值均 < 0.30 , 这表示可以用 WC、WHR 和 BMI 来解释公共因子 1(或者说公共因子 1 主要反映肥胖指标信息)。原变量在公共因子 2 上载荷量的绝对值 > 0.30 的是 SBP(0.91) 和 DBP(0.92), 在公共因子 3 上载荷量的绝对值 > 0.30 的是 FPG(0.88) 和 PG2h(0.87), 在公共因子 4 上载荷量的绝对值 > 0.30 的是 FINS(0.79) 和 INS2h(0.82), 在公共因子 5、6 上载荷量的绝对值 > 0.30 的分别是 HDL-C 变量(0.99) 和 TG 变量(0.94), 即公共因子 2~6 依次主要反映血压、血糖、胰岛素、HDL-C 和 TG。

(5) 因子得分系数阵见表 5。通过因子得分系数阵就可以将所有公共因子表示为各变量的线性组合。由此可得公共因子 F_1 的得分函数:

$$F_1 = 0.45X_1 + 0.40X_2 + 0.38X_3 - 0.07X_4 - 0.09X_5 - 0.09X_6 - 0.05X_7 - 0.08X_8 - 0.09X_9 + 0.06X_{10} - 0.04X_{11}$$

表5 因子得分系数阵(回归法)

指标	因子 1	因子 2	因子 3	因子 4	因子 5	因子 6
WC	0.45	-0.07	-0.07	-0.09	0.04	-0.03
WHR	0.40	-0.08	-0.04	-0.09	0.03	0.02
BMI	0.38	-0.04	-0.07	-0.01	0.04	-0.05
SBP	-0.07	0.57	-0.04	-0.05	0.00	-0.07
DBP	-0.09	0.58	-0.05	-0.05	-0.00	-0.03
FPG	-0.09	-0.05	0.61	-0.09	0.03	0.01
PG2h	-0.05	-0.03	0.61	-0.02	-0.04	-0.22
FINS	-0.08	-0.05	-0.11	0.64	0.02	0.06
INS2h	-0.09	-0.06	-0.01	0.68	-0.00	-0.16
TG	-0.04	-0.06	-0.12	-0.06	0.03	1.07
HDL-C	0.06	0.00	-0.00	0.01	1.01	0.03

F_2, \dots, F_6 也可类似得到。据此, 可得到每个个体经过标化的 6 个公共因子得分(作为新变量保存在数据库中)。可以分别按这 6 个公共因子(1~6 依次主要反映肥胖、血压、血糖、胰岛素、HDL-C 和 TG)来对研究对象进行评估, 还可通过计算 MS 的因子总得分来对个体进行评估。

同时系统给出的因子得分系数矩阵(本研究未给出), 表明 6 个公共因子间完全不相关。

讨 论

在不同种族人群中, 使用因子分析探讨 MS 不同组间潜在的共同病理生理联系, 但有的研究对结果不能进行恰当解释。在本项研究人群中, 提取的 6 个公共因子可以基本反映全部指标信息的 86%(方差意义下的)。进行正交旋转后, 公共因子 1~6 依次主要反映肥胖、血压、血糖、胰岛素、HDL-C 和 TG。通过实例分析, 注意到不同变量在公共因子上的聚集始终是根据变量间的相关系数, 即相关性越高, 越可能聚集在 1 个公共因子上。当根据公共因子的不同提取标准来构建因子模型时, 所保留的公共因子的个数可能不同, 原始变量在相应公共因子上的载荷量也可能不同。这也是对相同变量进行因子分析时, 产生不同因子模型的原因之一。其他的原因还有变量在不同人群中相关性的差异^[4,5]。

綜上述, 正交因子分析模型是一种对高度相关的多变量的简化分析方法, 该线性模型用较少的互不相关的新随机变量(公共因子)来代替原来较多的原始随机变量, 同时这些新随机变量又能尽可能多地反映原随机变量的信息(方差意义下的信息量), 即不丢失原始指标的预测价值。在因子模型中 F 是不可观测的, 当 $m > 1$, 旋转会改变 X 的信息量在不同公共因子上的分布。在旋转调整公因子载荷量的同时, 选择更具有实际意义的因子模型, 所以因子模型是对主成分分析的发展。不同变量在公共因子上的聚集始终是根据变量间的相关系数, 也就是说, 对高度相关变量进行因子分析才会有实际意义。所以因子分析的价值在于探查有高度相关关系的变量群, 进而为探究其共同的潜在机制提供线索。

参 考 文 献

- [1] 高惠璇. 应用多元统计分析-因子分析. 北京: 北京大学出版社, 2005: 293-312.
- [2] 丁秀林, 任雪松. 多元统计分析-因子分析. 北京: 中国统计出版社, 1999: 171-182.
- [3] 袁志发, 周静芊. 多元统计分析-因子分析. 北京: 科学出版社, 2002: 207-215.
- [4] Edwards KL, Austin MA, Newman B, et al. Multivariate analysis of the insulin resistance syndrome in women. *Arterioscler Thromb*, 1994, 14: 1940-1945.
- [5] Choi KM, Lee J, Kim KB, et al. Factor analysis of the metabolic syndrome among elderly Koreans—the South-west Seoul Study. *Diabet Med*, 2003, 20: 99-104.

(收稿日期: 2007-11-23)

(本文编辑: 张林东)