

· 讲 座 ·

医学统计学的基本方法(八)

李占魁 潘子昂 王石麟

八、主成分分析

在对各种问题的研究中,为了尽可能详细地了解它的性质和性能,往往通过许多侧面进行调查和测量。例如,在流行病学研究中,调查某人群的尺骨骨矿含量;调查某人群的癌症死亡发生率等。为了弄清某地区人群骨质疏松症或癌亡率高高低的原因,必须了解大群体中每个人(被研究的客体)的衣、食、住、行等各个方面。要对观察客体的衣食住行情况进行表述,必须借助于许多参数,例如吸烟、饮酒、身高、体重等等。每项参数的定量表达,被称为指标,例如每日吸烟量为50支,身高为170cm等。这样,每个客体都要尽可能多的参数,例如 m 项,去描述。 n 个客体的群体,将形成一个 $n \times m$ 元的数据矩阵。

从信息获取角度,尽可能多的观测项目可以避免遗漏。但是,从数理统计角度看,这些项目之间不一定没有关系,甚至会有某种函数关系。例如,某人身高与其体重一定有关;某人总吸烟量与其吸烟开始年龄一定有关等。相关越强烈的两个项目,所能提供的总信息比一个项目所提供的信息增加越少。如果两个项目之间具有函数关系,相关系数为1,提供的信息量与原来一个项目提供的信息量相同。

主成分分析方法可以将许多观察项目综合成新的彼此不相关的参数,每个新参数都是原来初始参数的线性组合。这样就保障不丢失原来的信息。同时,新的综合参数可以提供给我们全面的,排除主观因素和片面性干扰的结论。例如,一个人尺骨骨矿含量多少的原因,不能只根据某一项指标,例如初潮年龄来判断,而是由逐

个多项目综合贡献造成的。所以说,主成分分析为医学和生物学领域对复杂事物的研究,特别是流行病学研究,提供了非常重要的手段。

1 主成分分析原理

假设有 n 个待研究的客体,每个客体的状态用 m 个参数去描写,如 $x_1, x_2, \dots, x_m$ 。我们的目的是寻求新的参数,如 $y_1, y_2, \dots, y_m$,其中每个参数 y_j 都是原来参数 x_j ($j=1, 2, \dots, m$)的线性组合,且彼此线性无关。写成数学方程式的形式便有如下结果:

$$y_i = \sum_{j=1}^m a_{ij} x_j \quad (1)$$

考虑到 (y_i) 彼此线性关的要求,式(1)系数必须满足下列条件:

$$\sum_{i=1}^m a_{ij} a_{ik} = \begin{cases} 0 & j \neq k \\ 1 & j = k \end{cases} \quad (2)$$

这里, $i=j=k=1, 2, 3, \dots, m$ 。

为了行文方便,将各数组向量写成矩阵形式,式(1)和(2)分别写成:

$$\begin{aligned} Y &= AX \\ AA' &= I \end{aligned} \quad (3)$$

(4)

式中, A' 为 A 的转置矩阵; I 为单位矩阵,即 I 的对角元素为1,其他元素为0^[1]。

我们将用 C_{kl} 表示向量 x_k 和 x_l 间的协方差,即

$$C_{kl} = \frac{1}{n} \sum_{i=1}^n (x_{ki} - \bar{x}_k)(x_{li} - \bar{x}_l) \quad (5)$$

写成协方差矩阵形式为:

$$C = \frac{1}{n} XX' \quad (6)$$

Y 的协方差矩阵将是

$$E(YY') = \{AX(AX)'\} = ACA' = A \quad (7)$$

式中, A 为 n 对角矩阵,即非对角元素为0,对角元素为 $\lambda_1, \lambda_2, \dots, \lambda_m$ 。每个 λ_j 表示 y_j ($j=1, 2,$

…… m)的方差。

用 A' 左乘(7)式,便有下列结果:

$$A'ACA' = A' \Lambda = CA' \quad (8)$$

(8)式结果是由于 $A'A=I$, 而 I 为单位矩阵。(8)式为本征方程。求解方程(8),即为寻找满足条件的 $\lambda_j(j=1,2,\dots,m)$,即本征值。为此,将(8)式矩阵相乘形式展开成 $m \times m$ 个方程。为了求得 A_0 的非零解,要求:

$$\begin{vmatrix} C_{11}-\lambda_1 & C_{12} & \dots & C_{1m} \\ C_{21} & C_{22}-\lambda_1 & \dots & C_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & \dots & \dots & C_{mm}-\lambda_1 \end{vmatrix} = 0 \quad (9)$$

同理,对 $\lambda_2, \lambda_3, \dots, \lambda_m$,可以得到完全类似的方程及 A_0 非零解的条件方程。这样,所求的 $\lambda_j(j=1,2,\dots,m)$ 为矩阵方程

$$|C - \lambda I| = 0 \quad (10)$$

的 m 的个根,他们被称为协方差矩阵 C 的本征值。相应的,系数矩阵 A 的各个分量 a_j 是特征向量的分量。

至此,主成分分析的目的已经完成,(1)式中综合参量 y_1 的具体表达式已经得到。

2 主成分分析的应用

主成分分析有许多特点和优点,如下述:

2.1 本征值(λ_j)是归一化的,即有 $\sum_{j=1}^m \lambda_j = 1$ 的特性。因此,不但可以知道主成分的大小次序,还可以知道各个主成分对总信息量的贡献的百分率。

2.2 主成分(y_j)是正交的。因此,代表主成分的各个分量也是彼此正交的。于是,在多维相空间中,各主成分所代表的相空间分量向量,形成一个完备的正交集。正由于这一点,保障新的综合参量(y_j)保留了初始数据的所有信息不会丢失。

2.3 主成分分析的一系列矩阵变换和运算,使第一主成分方差最大,因此主成分分析可以完成信息的压缩和浓集作用。这样,将多维相空间降维可以得到实施。研究前几个主成分的特性,便可以大概知道整个群体的特征,为科学研究的复杂性大大得到简化。

由于主成分分析有这些优点而得到广泛应

用。以下将就骨质疏松研究中尺骨骨矿含量调查的数据分析的具体例子,说明该方法的操作步骤和注意事项。

2.4 主成分分析方法的步骤:假设, $x_j(j=1,2,\dots,m)$ 代表尺骨骨矿含量调查的各个项目,例如,性别,年龄,初潮年龄等等。我们称每个观测项目为描述该客体的参数向量。这样, n 个客体构成的群体样本,提供给我们一个数据矩阵 $X = (x_{ij})$ 。通常,脚标的第一个表示矩阵的行序,脚标的第二个表示矩阵的列序。在该具体流行病学研究的例子中, $i=1,2,\dots,n$,表示 n 个客体的序号,通常是个很大的数。 n 表示矩阵的行序,书写论文时非常方便;相似于脚标 i 的用意,我们用脚标 j 表示调查项目的序号,通常不会是过大的数,为横向书写提供了方便。如果给出 x_{ij} 的形式,则表示数据矩阵 (x_{ij}) 的一个元素,例如 $x_{ij}=12$ 岁,则表示第 i 个客体的第 j 个参数的值。如果给出的是向量 x_j ,则意指第 j 个观测项目; $\bar{x}_j$ 意义是,将第 j 个观测参数对所有客体求平均值。主成分分析是要寻求新的综合向量 $y_j(j=1,2,\dots,m)$,目的如前述。 y_j 脚标含意同 x_j ,也顺便明示读者。主成分分析步骤如下:
①初始数据的标准化。我们知道,主成分分析是基于多元统计中的方差分析发展而来的,每个主成分 $y_j(j=1,2,\dots,m)$ 的方差基于本征值 λ_j 。然而, y_j 是 x_j 的线性组合, y_j 的方差的大小,直接与 x_j 的尺度有关系,某一个变量的尺度改变,例如 170 是某人身高,以厘米为单位,1.70 也是该人身高,则是以米为尺度,终将影响到方差 λ_j 。如此等等。若这些尺度间没有可比性,便使 y_j 的方差也失去可比性。这一点,已经成为主成分分析的一个严重问题。为了解决这个困难,必须将初始数据矩阵 (x_{ij}) 标准化。让我们考虑下列变换。

$$U_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{xj}} \quad (11)$$

这里, $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ 为参数 x_j 的样本平均值:

$$S_{xj} = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{(n-1)}}$$

为参数 x_j 的标准偏差

这样,新的数据矩阵称为标准化了的。

②求解本征方程。有了标准化数据阵 (U_i) ,以代替初始数据 (x_{ij}) ;相应的, x_k 和 x_l 的协方差 C_{kl} 将变成 U_k 和 U_l 的相关系数,协方差矩阵 C 将变成相关系数矩阵 R ,方程(11)将变成

$$|R - \lambda I| = 0 \quad (12)$$

式中, λ 及 I 意义同前。但本征方程(12)的解变成归一化,即

$$\lambda_1 + \lambda_2 + \dots + \lambda_i + \dots + \lambda_m = 1 \quad (13)$$

③将本征值系列 λ_i 及相应的主成分 y_i 由大到小排序,以得到有序化的主成分变量。

综上所述,主成分分析是把 m 个随机变量 $x_j (j=1, 2, \dots, m)$ 的总方差 t, v 分解成 $p (p \leq m)$ 个不相关的随机变量 $y_i (i=1, 2, \dots, p)$ 的方差之和 $\lambda_1 + \lambda_2 + \dots + \lambda_p$,使第一主成分的方差最大。比例 $\lambda_1 / \sum_{i=1}^p \lambda_i$ 表明了第一主成分的方差在总方差中的比值,称为第一主成分贡献率。这个值越大,表明 y_1 这个向量综合初始参数向量的能力越强。以下主成分的意义与第一主成分相同^[2]。了解到这一点,就可以明白为什么主成分 y_i 的次序是依本征值 λ_i 的大小顺序排列;就可以明白,主成分分析要求将用不同尺度标称的参数指标的数据标准化的道理。

2.5 主成分分析用于群体聚类结果的显示:上一讲我们曾指出,聚类分析作为多元分析的一个重要内容已得到广泛应用^[3]。聚类分析的结果,可用贝叶斯判据加以分析。然而,预先用眼睛观察而后进行数学处理,为研究聚类结果的生物学意义,提供了许多方便^[4]。

大家知道,人的眼睛至多可以识别三维几何空间内的目标。对于几何空间的概念,人们早已熟悉。借助几何空间的概念,我们引入相空间的概念。

通常,我们用坐标 (x, y) 来描述一目标在

平面 $x-O-y$ 上的位置;用坐标 (x, y, z) 来描述一目标在立体空间中的位置。如果这些坐标 x, y, z 不是几何坐标的意义,而换成描述一个人生活特点的参数指标,例如身高、体重、吸烟量等等,所谓目标的几何位置的概念,将换成一个人的尺骨骨矿含量状态的概念;几何空间的概念换成了相空间的概念。

在概念替换之前,必须讲明替换的条件。条件就是正交要求。在欧几里德几何空间中,向量 x, y, z 彼此正交,即彼此线性无关。而描述骨矿含量的参数之间并非彼此线性无关,因而不能构成多维相空间。主成分分析所得到的一系列主成分 (y_i) 正好满足正交要求而又保留了初始数据的信息^[5]。这是主成分分析对流行病学研究的第一个贡献。其次,要解决人眼观察问题,还必须将多维空间降为三维以下,又要尽可能多地保留群体在多维空间中的结构特点。主成分分析过程中的一系列的矩阵变换和运算解决了这个困难,最大主成分的方差最大,且依次排列。好的质量控制下的初始数据可以作到前几个主成分,例如前3个,集中了总方差的大部分,例如80%以上,从而反映群体的基本特征。

主成分分析还可以用于生物学和医学研究乃至临床观察等许多方面。例如,因子分析就是基于主成分分析进行的。感兴趣的读者可参阅有关文献^[2]。

参 考 文 献

- 1 M.肯德尔(英)著.多元分析.北京:科学出版社,1983,16~25.
- 2 张尧庭,方开泰著.多元统计分析引论.北京:科学出版社,1983,322~328.
- 3 潘子昂,王石麟,李占魁.医学统计学的基本方法(七).中国骨质疏松杂志,1997,3(1):71.
- 4 李占魁,刘秀华,吴元沧.29名小剂量受照人员辐射效应聚类分析.中华放射医学与防护杂志,1989,8(2):97.
- 5 McGill JR and Kowalski SR. Intrinsic dimensionality of smell. Anal Chem, 1977, 49:596.