中级应用统计学

主成分和因子分析

张晨峰

华东理工大学商学院

2016年10月12日

3 主成分和因子分析

主要内容

- 主成分分析
- 因子分析

3.1 主成分分析

简介

主成分分析（principal components analysis，pca）是由Hotelling于1933年首先提出的。主成分分析是利用**降维**的思想，在损失很小信息的前提下把多个指标转化为几个综合指标的多元统计方法。通常把转化生成的综合指标称之为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间互不相关。

应用

主成分分析更多的是一种达到目的的方法，而非目的本身，这是因为主成分分析频繁地用作许多实证分析的中间步骤。

3.1 主成分分析

总体主成分

设随机向量$X' = [X_1, X_2, ..., X_p]$，其协方差矩阵$\Sigma$，其特征值$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$。 考虑线性组合

$$\begin{cases} Y_1 = \alpha_1' X = \alpha_{11} X_1 + \alpha_{12} X_2 + ... + \alpha_{1p} X_p \\ Y_2 = \alpha_2' X = \alpha_{21} X_1 + \alpha_{22} X_2 + ... + \alpha_{2p} X_p \\ ... \\ Y_p = \alpha_p' X = \alpha_{p1} X_1 + \alpha_{p2} X_2 + ... + \alpha_{pp} X_p \end{cases} \tag{1}$$

3.1 主成分分析

总体主成分

我们可得

$$Var(Y_i) = \alpha_i' \Sigma \alpha_i$$

$$Cov(Y_i, Y_k) = \alpha_i' \Sigma \alpha_k$$

第一主成分是最大方差的线性组合，即它使$Var(Y_i) = \alpha_i' \Sigma \alpha_i$最大化。显然主成分会因为任何$\alpha_1$乘以某个常数而增大，为消除这种不确定性，需要施加约束条件$\alpha_i' \alpha_i = 1$。

3.1 主成分分析

总体主成分

因此我们定义：

- 第一主成分=线性组合$\alpha_1'X$，在$\alpha_1'\alpha_1 = 1$时，它使$Var(\alpha_1'X)$最大；
- 第二主成分=线性组合$\alpha_2'X$，在$\alpha_2'\alpha_2 = 1$和$Cov(\alpha_1'X, \alpha_2'X) = 0$时，它使$Var(\alpha_2'X)$最大；
- 第i个主成分=线性组合$\alpha_i'X$，在$\alpha_i'\alpha_i = 1$和$Cov(\alpha_i'X, \alpha_k'X) = 0$时，它使$Var(\alpha_i'X)$最大；

3.1 主成分分析

定理 (定理 3.1)

若A和B都是对称阵且$B > 0$，则有

$$max \quad \frac{x'Ax}{x'Bx} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = min \quad \frac{x'Ax}{x'Bx}$$

其中$\lambda_1, ..., \lambda_p$为$B^{-1}A$的特征值，使$\frac{x'Ax}{x'Bx}$最大（最小）的向量$B^{-1}A$是对应最大（最小）特征值的特征向量。如果$x'Bx = 1$，我们可得

$$max \quad x'Ax = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = min \quad x'Ax$$

3.1 主成分分析

结论 3.2

设随机向量$X' = [X_1, X_2, ..., X_p]$，其协方差矩阵$\Sigma$，它有特征值-特征向量对$(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$，其中$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$。则第$i$个主成分由

$$Y_i = e_i'X = e_{i1}X_1 + e_{i2}X_2 + ... + e_{ip}X_p$$

给出，此时

$$Var(Y_i) = e_i'\Sigma e_i = \lambda_i$$

$$Cov(Y_i, Y_k) = e_i'\Sigma e_k = 0$$

3.1 主成分分析

推论 3.3

设随机向量$X' = [X_1, X_2, ..., X_p]$，其协方差矩阵$\Sigma$，它有特征值-特征向量对$(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_p, e_p)$，其中$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$。设$Y_1 = e_1'X$，$Y_2 = e_2'X$，...，$Y_p = e_p'X$是主成分，则

$$\sigma_{11} + \sigma_{22} + ... + \sigma_{pp} = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + ... + \lambda_p = \sum_{i=1}^{p} Var(Y_i)$$

从而总方差中属于第k个主成分(被第k个主成分所解释的)比例为

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + ... + \lambda_p}$$

又称为第k个主成分的方差贡献率。

3.1 主成分分析

系数向量

e_{ik}的大小量度第k个变量对第i个主成分的重要程度，而不管其他变量如何。特别地，e_{ik}与Y_i和X_k之间的相关系数成比例。

结论 3.4

设$Y_1 = e_1'X$，$Y_2 = e_2'X$，...，$Y_p = e_p'X$是主成分，则

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

是成分Y_i和变量X_k之间的相关系数，又称为因子负荷量。在实践中，有较大(按绝对值)系数的变量，趋向于有较大的相关，故这两个重要性的测度经常给出相似的结果。

3.1 主成分分析

从相关矩阵得到主成分

主成分也可以从标准化变量

$$\begin{cases} Z_1 = \frac{(X_1-\mu_1)}{\sqrt{\sigma_{11}}} \\ Z_2 = \frac{(X_2-\mu_2)}{\sqrt{\sigma_{22}}} \\ \cdots \\ Z_p = \frac{(X_p-\mu_p)}{\sqrt{\sigma_{pp}}} \end{cases} \tag{2}$$

得到，采用矩阵记号

$$Z = (V^{\frac{1}{2}})^{-1}(X-\mu)$$

3.1 主成分分析

从相关矩阵得到主成分

所以我们有

$$Cov(Z) = \rho$$

Z的主成分可从X的相关矩阵ρ的特征向量得到。

3.1 主成分分析

结论 3.4

有$Cov(Z) = \rho$的标准化变量$Z' = [Z_1, Z_2, ..., Z_p]$的第$i$主成分由

$$Y_i = e_i' Z = e_i' (V^{\frac{1}{2}})^{-1} (X - \mu)$$

给出，而且

$$\sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$$

$$\rho_{Y_i, Z_k} = e_{ik} \sqrt{\lambda_i}$$

需要注意的是，由协方差矩阵和相关系数矩阵导出的主成分是不同的，因此标准化不是无关紧要的。如果测量单位不是同量纲的，那么变量可能应该标准化。

3.1 主成分分析

主成分个数

- 一种能帮助我们确定主成分合适个数的有用的视觉工具，是所谓的崖底碎石图。
- 或者选取主成分个数使得累积贡献率达到某个数值(例如85%)以上

3.1 主成分分析

样本主成分

在实际研究工作中，总体协方差阵Σ与相关阵R通常是未知的，因此需要通过样本数据来估计，即

$$S = \frac{1}{n-1}\sum_{k=1}^{n}(x_{ki} - \overline{x_i})(x_{ki} - \overline{x_i})'$$

$$R = (r_{ij})_{pp}, \quad r_{ij} = \frac{S_{ij}}{\sqrt{S_{ii}S_{jj}}}$$

S为样本协方差矩阵，作为总体协方差阵Σ的无偏估计；R是样本相关矩阵，为总体相关矩阵的估计。

3.2 因子分析

简介

因子分析（factor analysis）是主成分分析的推广。它也是利用降维的思想，由研究原始变量相关矩阵内部的依赖关系出发，把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多变量统计分析方法。因子分析的实质目的是，只要可能，就用几个潜在的但不能观测的随机变量取描述许多变量间的协方差关系，这些随机量叫做因子。

3.2 因子分析

因子分析的基本思想

因子分析的基本思想是根据相关性大小把原始变量分组，使得同组内的变量之间相关性较高，而不同组的变量间的相关性则较低。每组变量代表一个基本结构，并用一个不可观察的综合变量表示，这个基本结构就称为公共因子。对于某一具体问题，原始变量可以分解为两部分之和的形式，一部分是少数几个不可测的所谓公共因子的线性代数，另一部分是与公共因子无关的特殊因子。

3.2 因子分析

正交因子模型

有p个成分的观测随机向量X，有均值μ和协方差矩阵Σ。因子模型要求X是线性依赖于不可观测的称之为公共因子的随机变量$F_1, F_2, ..., F_m$和p个附加的称之为误差或有时也称之为特殊因子的变差源$\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$，具体的，因子分析模型是：

$$\begin{cases} X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + ... + l_{1m}F_m + \varepsilon_1 \\ X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + ... + l_{2m}F_m + \varepsilon_2 \\ ... \\ X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + ... + l_{pm}F_m + \varepsilon_p \end{cases} \tag{3}$$

3.2 因子分析

正交因子模型

矩阵形式$X-\mu=LF+\varepsilon$ 系数l_{ij}为第i个变量在第j个因子上的载荷，故矩阵L是因子载荷阵。我们设定

$$E(F)=0, Cov(F)=I$$

$$E(\varepsilon)=0, Cov(\varepsilon)=\Psi=\begin{bmatrix}\psi_1 & 0 & \dots & 0\\ 0 & \psi_2 & \dots & 0\\ \dots & \dots & \dots & \dots\\ 0 & 0 & \dots & \psi_p\end{bmatrix}$$

$$Cov(\varepsilon, F)=0$$

3.2 因子分析

正交因子模型的协方差结构

- $Cov(X) = LL' + \Psi$，或

$$Var(X_i) = l_{i1}^2 + ... + l_{im}^2 + \psi_i$$

$$Cov(X_i, X_k) = l_{i1}l_{k1} + ... + l_{im}l_{km}$$

- $Cov(X, F) = L$，或

$$Cov(X_i, F_i) = l_{ij}$$

有m个公共因子贡献的第i个变量的方差部分，叫做第i个共性方差，属于特殊因子的部分，常称为独特方差或特殊方差。

3.2 因子分析

估计方法

若Σ明显背离一个对角阵，那就可以准备考虑因子模型了。最流行的参数估计方法有主成分方法和极大似然估计。

主成分方法

令Σ有特征值-特征向量对(λ_i, e_i)，$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$，则

$$\Sigma = LL' + \Psi$$

其中

$$L = [\sqrt{\lambda_1}e_1, \sqrt{\lambda_2}e_2, ..., \sqrt{\lambda_m}e_m], m < p$$

3.2 因子分析

Example 9.3 (Factor analysis of consumer-preference data) In a consumer-preference study, a random sample of customers were asked to rate several attributes of a new product. The responses, on a 7-point semantic differential scale, were tabulated and the attribute correlation matrix constructed. The correlation matrix is presented next:

Attribute (Variable)		1	2	3	4	5
Taste	1	1.00	.02	.96	.42	.01
Good buy for money	2	.02	1.00	.13	.71	.85
Flavor	3	.96	.13	1.00	.50	.11
Suitable for snack	4	.42	.71	.50	1.00	.79
Provides lots of energy	5	.01	.85	.11	.79	1.00

3.2 因子分析

Table 9.1

Variable	Estimated factor loadings $\tilde{\ell}_{ij} = \sqrt{\hat{\lambda}_i}\,\hat{e}_{ij}$ F_1	F_2	Communalities $\tilde{h}_i^2$	Specific variances $\tilde{\psi}_i = 1 - \tilde{h}_i^2$
1. Taste	.56	.82	.98	.02
2. Good buy for money	.78	−.53	.88	.12
3. Flavor	.65	.75	.98	.02
4. Suitable for snack	.94	−.10	.89	.11
5. Provides lots of energy	.80	−.54	.93	.07
Eigenvalues	2.85	1.81		
Cumulative proportion of total (standardized) sample variance	.571	.932		

3.2 因子分析

因子旋转

用一个正交变换从初始载荷得到的所有因子载荷，同样有能力重现协方差矩阵。因子载荷的一个正交变换，称之为因子旋转。

Table 9.5

Variable	Estimated factor loadings F_1	F_2	Communalities $\hat{h}_i^2$
1. Gaelic	.553	.429	.490
2. English	.568	.288	.406
3. History	.392	.450	.356
4. Arithmetic	.740	−.273	.623
5. Algebra	.724	−.211	.569
6. Geometry	.595	−.132	.372

3.2 因子分析

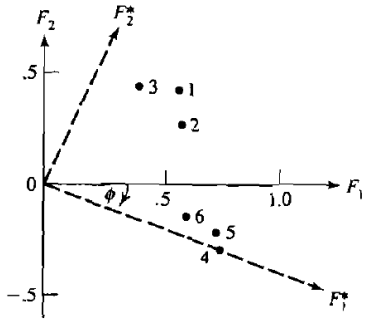

Figure 9.1 Factor rotation for test scores.

Table 9.6

Variable	Estimated rotated factor loadings F_1^*	F_2^*	Communalities $\hat{h}_i^{*2} = \hat{h}_i^2$
1. Gaelic	.369	.594	.490
2. English	.433	.467	.406
3. History	.211	.558	.356
4. Arithmetic	.789	.001	.623
5. Algebra	.752	.054	.568
6. Geometry	.604	.083	.372

3.2 因子分析

因子得分

因子分析中，感兴趣的通常是因子模型的参数。然而，公共因子的估计值，称为因子得分，也是我们需要的。因子得分不是在通常意义下对未知参数的估计，相反的，它们是对不能观测的随机因子向量F_j的值的估计。因子得分的常用方法有加权最小二乘法和回归法。两种因子得分方法都有两个共同的要素：

- 它们把估计的因子载荷和特殊方法当作真值处理
- 它们涉及原始数据的线性变换，可能是作中心化或标准化

3.2 因子分析

主成分分析与因子分析的区别

- 因子分析的目的是要探查能对变量起解释作用的公共因子和特殊因子，主成分分析只是寻找能解释诸多变量绝大部分变异的几组彼此不相关的新变量（主成分）。
- 因子分析把变量表示成各因子的线性组合；主成分分析把主成分表示成各变量的线性组合。
- 主成分分析不需要专门假设，因子分析则需要。
- 主成分分析中，当给定协方差或相关矩阵的特征根唯一时，主成分一般是固定的；而因子分析中因子不是固定的
- 和主成分分析比较，由于因子分析可以使用旋转帮助解释因子，在解释方面更加有优势。