第 8 章　主成分回归与偏最小二乘

李　杰

数据科学学院，浙江财经大学

2019 年 12 月 15 日

内容提要

8.1 主成分回归

8.2 偏最小二乘

8.3 本章小结与评注

☞ 主成分回归的基本思想

- **主成分回归** (principal components regression, PCR)
  - 主成分回归是线性回归模型的一种有偏估计.
- **主成分的思想**(principal components analysis, PCA)
  - 主成分回归是也成为主分量分析, 是一种**降维**的思想, 在**损失少量信息**的前提下, **把多个指标利用正交旋转变换**, 转化为几个综合指标的多元统计分析方法. 通常把生成的综合指标成为**主成分**, 其中每个主成分都是**原始变量的线性组合**, 且各主成分之间**互不相关**.
- **主成分的推导**
  - ✧ 设研究的问题中涉及 p 个指标, 用 $X_1, X_2, \cdots, X_p$ 表示;
  - ✧ p 个指标成随机向量 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$. 设随机向量 $\mathbf{X}$ 的均值为 $\boldsymbol{\mu}$, 方差为 $\boldsymbol{\Sigma}$;

- 对 $\mathbf{X}$ 进行**线性变换**, 形成新的综合变量, 记为 $\mathbf{Y}$, 即

$$\begin{cases} Y_1 = \mu_{11}X_1 + \mu_{12}X_2 + \cdots + \mu_{1p}X_p \\ Y_2 = \mu_{21}X_1 + \mu_{22}X_2 + \cdots + \mu_{2p}X_p \\ \quad \cdots\cdots\cdots\cdots \\ Y_p = \mu_{p1}X_1 + \mu_{p2}X_2 + \cdots + \mu_{pp}X_p \end{cases}$$

- 对线性变换的要求
  ① 记 $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{ip})$, 则 $\boldsymbol{\mu}'\boldsymbol{\mu} = 1$;
  ② Y_i 与 Y_j 不相关 $(i \neq j;\ i, j = 1, 2, \cdots, p)$;
  ③ $\{Y_1, Y_2, \cdots, Y_p\}$ 互不相关, 且 $\mathrm{Var}(Y_1) \geqslant \mathrm{Var}(Y_2) \geqslant \cdots \geqslant \mathrm{Var}(Y_p)$.
- 称 Y_1 为第一个主成分, Y_2 为第二个主成分, 以此类推.

☞ 主成分的基本性质

- **引论**

✧ 设矩阵 $A' = A$, 将 A 的特征根 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 依大小排列, 不妨设 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$, 则对任意向量 $\mathbf{x}$, 有

$$\max_{\mathbf{x} \neq 0} \frac{\mathbf{x}' A \mathbf{x}}{\mathbf{x}' \mathbf{x}} = \lambda_1, \cdots, \min_{\mathbf{x} \neq 0} \frac{\mathbf{x}' A \mathbf{x}}{\mathbf{x}' \mathbf{x}} = \lambda_p$$

- **定理**

✧ 设随机向量 $\mathbf{X} = (X_1, X_2, \cdots, X_p)'$ 的协方差矩阵为 Σ, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ 为 Σ 的特征根, $\gamma_1, \gamma_2, \cdots, \gamma_p$ 为 Σ 的各特征根对应的标准正交向量, 则第 i 个主成分为

$$Y_i = \mu_{i1} X_1 + \mu_{i2} X_2 + \cdots + \mu_{ip} X_p, \quad (i = 1, 2, \cdots, p)$$

此时

$$\mathrm{Var}(Y_i) = \gamma_i' \Sigma \gamma_i = \lambda_i;$$
$$\mathrm{Cov}(Y_i, Y_j) = \gamma_i' \Sigma \gamma_j = 0, \quad i \neq j$$

- **性质 1:** $\mathbf{Y}$ 的协方差矩阵对角阵 Λ, 其对角线元素为 $\lambda_1, \lambda_2, \cdots, \lambda_p$.
- **性质 2:** 记 $\Sigma = (\sigma_{ij})_{p\times p}$, 有 $\sum\limits_{i=1}^{p} \lambda_i = \sum\limits_{i=1}^{p} \sigma_{ii}$.
- **累计贡献率:** 称 $\alpha_k = \dfrac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$ $(k = 1, 2, \cdots, p)$ 为第 k 个主成分 Y_k 的**方差贡献率**. 称 $\dfrac{\sum\limits_{i=1}^{m} \lambda_i}{\sum\limits_{i=1}^{p} \lambda_i}$ 为主成分 $Y_1, Y_2, \cdots, Y_m$ 的**累计贡献率**.
- **性质 3:** $\rho(Y_k, X_i) = \dfrac{\mu_{ki}\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}$ $(k = 1, 2, \cdots, p)$.
- **因子载荷量:** 称第 k 个主成分 Y_k 与原始变量 X_i 的相关系数 $\rho(Y_k, X_i)$ 为**因子载荷量**. 因子载荷量的绝对值大小刻画了该主成分的主要意义及其成因.
- **性质 4:** $\sum\limits_{i=1}^{p} \rho^2(Y_k, X_i)\sigma_{ii} = \lambda_k$.
- **性质 5:** $\sum\limits_{i=1}^{p} \rho^2(Y_k, X_i) = \dfrac{1}{\sigma_{ii}} \sum\limits_{i=1}^{p} \lambda_k \mu_{ki}^2 = 1$.

- **R 代码**

```
france <- read.csv("D:\\documents\\MyDoc\\JobInZufe\\Courseware
                               \\应用回归分析\\数据\\france.csv")
france.prin <- princomp(~GNP+saving+consume,data=france, cor=T)
                                # 主成分分析
summary(france.prin)            # 显示主成分分析结果

pre <- predict(france.prin) # 计算主成分得分
loadings(france.prin)       # 显示载荷矩阵
screeplot(france.prin,type="lines") # 绘制碎石图
comp1 <- pre[,1]            # 提取第一和第二主成分
comp2 <- pre[,2]
import_scale <- scale(france$import,center=T,scale=T)
import_scale <- as.data.frame(import_scale)

france.prin.lm <- lm(import_scale$V1~comp1+comp2-1) # 主成分回归
summary(france.prin.lm)
```

导出主成分回归方程

$$\widehat{import\_scale} = 0.658comp1 - 0.182comp2$$

又因为

$$\begin{aligned} comp1 &= 0.706GNP \qquad\qquad\qquad + 0.707comsume \\ comp2 &= \qquad\qquad -0.999saving \end{aligned}$$

还原变量后的回归方程

$$\widehat{import\_scale} = 0.465GNP + 0.182saving + 0.465consume$$

☞ 偏最小二乘的基本原理

- **问题:** 在实际问题中, 如果用来估计线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

的样本观测 $(y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$ $(i = 1, 2, \cdots, n)$ 数量 $n < p$ 时, 模型如何估计?

- **主成分回归的缺点:** 主成分分析将数据降维, 在损失一定信息量的前提下, 对几个主要成分做主成分回归. 但主成分回归中确定主成分的时候没有考虑与因变量的相关系.
- **偏最小二乘的原理:**
  ① 将因变量, 所有的自变量数据中心化 (之后提到的因变量、自变量都是中心化后的数据);
  ② 做因变量关于每个自变量的一元回归, 得到

$$\hat{y}(x_i) = \frac{\mathbf{x}_i'\mathbf{y}}{\mathbf{x}_i'\mathbf{x}_i} x_i, \quad (i = 1, 2, \cdots, p)$$

其中 $\mathbf{x}_i = (x_{1i}, \cdots, x_{ni})'$, x_i 表示第 i 个变量.

③ 构造自变量的加权线性组合

$$\sum_{i=1}^{p} \omega_i \frac{\mathbf{x}_i'\mathbf{y}}{\mathbf{x}_i'\mathbf{x}_i} x_i$$

ω_i 表示权重，选择有很多，最简单的为 $\omega_i = \mathbf{x}_i'\mathbf{x}_i$，则加权线性组合为

$$t_1 = \sum_{i=1}^{p} \left(\mathbf{x}_i'\mathbf{y}\right) x_i$$

记 t_1 的观测值为

$$\mathbf{t}_1 = \sum_{i=1}^{p} \left(\mathbf{x}_i'\mathbf{y}\right) \mathbf{x}_i.$$

④ 做 $\mathbf{y}$ 对 t_1 的回归，得到

$$\hat{y}(t_1) = \frac{\mathbf{t}_1'\mathbf{y}}{\mathbf{t}_1'\mathbf{t}_1} t_1$$

其拟合值为

$$\hat{\mathbf{y}}(\mathbf{t}_1) = \frac{\mathbf{t}_1'\mathbf{y}}{\mathbf{t}_1'\mathbf{t}_1} \mathbf{t}_1,$$

残差为

$$\mathbf{y}^{(1)} = \mathbf{y} - \hat{\mathbf{y}}(\mathbf{t}_1)$$

⑤ 再做每个自变量 x_i 关于 t_1 的回归,

$$\hat{x}_i(t_1) = \frac{\mathbf{t}_1'\mathbf{x}_i}{\mathbf{t}_1'\mathbf{t}_1} t_1, \quad i = 1, 2, \cdots, p$$

相应的拟合值为

$$\hat{\mathbf{x}}_i(\mathbf{t}_1) = \frac{\mathbf{t}_1'\mathbf{x}_i}{\mathbf{t}_1'\mathbf{t}_1}\mathbf{t}_1 \ (i = 1, 2, \cdots, p)$$

残差为

$$\mathbf{x}_i^{(1)} = \mathbf{x}_i - \hat{\mathbf{x}}_i(\mathbf{t}_1) \ (i = 1, 2, \cdots, p).$$

⑥ 将 $\mathbf{y}^{(1)}, \mathbf{x}_1^{(1)}, \cdots, \mathbf{x}_p^{(1)}$ 作为新的因变量和自变量, 重复上述步骤, 逐步求得 $t_1, t_2, \cdots, t_r$, 其中 $r = \text{rank}(X'X)$.

⑦ 最后做 y 对 $t_1, t_2, \cdots, t_r$ 的最小二乘回归. 经过变量间的转换, 最终可得到 y 关于 $x_1, x_2, \cdots,$ x_p 的回归方程.

☞ 偏最小二乘算法 (Wold 算法)

假定数据 (因变量, 所有的自变量) 都已经中心化.

① 初始化: $\mathbf{y} \to \mathbf{y}_0$, $\mathbf{X} \to \mathbf{X}_0$, $\mathbf{0} \to \hat{\mathbf{y}}_0$, $\mathbf{0} \to \hat{\mathbf{X}}_0$;

② 对 $a = 1$ 到 r, 重复进行:

③ $\mathbf{t}_a = \mathbf{X}_{a-1}\mathbf{X}'_{a-1}\mathbf{y}_{a-1}$

④ $\hat{\mathbf{y}}_a = \dfrac{\mathbf{t}_a\mathbf{t}'_a}{\mathbf{t}'_a\mathbf{t}_a}\mathbf{y}_{a-1} + \hat{\mathbf{y}}_{a-1}$

⑤ $\mathbf{y}_a = \mathbf{y}_{a-1} - \dfrac{\mathbf{t}_a\mathbf{t}'_a}{\mathbf{t}'_a\mathbf{t}_a}\mathbf{y}_{a-1}$

⑥ $\hat{\mathbf{X}}_a = \dfrac{\mathbf{t}_a\mathbf{t}'_a}{\mathbf{t}'_a\mathbf{t}_a}\mathbf{X}_{a-1}$

⑦ $\mathbf{X}_a = \mathbf{X}_{a-1} - \hat{\mathbf{X}}_a$

⑧ $\mathbf{X}_a\mathbf{X}'_a$ 中的主对角元素近似等于 0, 循环中止.

☞ 交叉验证法 (cross validation)

- **问题:** Wold 算法如何终止 (即如何得到最优的 a)?
- **交叉验证法 (cross validation)**
  - ✧ 将资料矩阵 $\mathbf{X}$, $\mathbf{y}$ 分组, 并删除其中的第 l 数据, 将去掉第 l 组数据的资料矩阵记为 $\mathbf{X}(-l)$, $\mathbf{y}(-l)$;
  - ✧ 以 $\mathbf{X}(-l)$, $\mathbf{y}(-l)$ 为基础, 用 PLS 方法算出预测方程 $\hat{y}_a$ 的表达式.
  - ✧ 将 $x_{l1}, x_{l2}, \cdots, x_{lk}$ 代入 $\hat{y}_a$, 将其预测值记为 $\hat{y}_{al}(-l)$, 残差值记为 $y_l - \hat{y}_{al}(-l)$.
  - ✧ 记
    $$L(a) = \sum_{l=1}^{n} (y_l - \hat{y}_{al}(-l))^2$$
    $L(a)$ 从整体上反映了第 a 步预测方程的好坏.
  - ✧ 选择使得 $L(a)$ 最小的值 a^*, 即
    $$a^* = \arg \min_{0 \leqslant l \leqslant r} L(a)$$