多组变量典型相关准则的数值解法

研究生：林雪珍

专业：计算数学

指导老师：刘新国

报告内容

1. 选题依据及背景
2. 典型相关分析的研究状况及已有结果
3. 研究内容及方法

选题依据及背景

典型相关分析（canonical correlation analysis）

- 最早提出：1936年，Hotelling 研究2组变量
- 基本思想：寻找每组变量的线性组合，使得两组的线性组合之间具有最大的相关系数
- 推广：多组变量, Steel(1951)、Horst(1961)、Van de Geer(1984)等
- 数学表现形式：具有约束的最优化问题
- 应用领域：生物学、心理学、计量经济、工业生产及信息技术等

典型相关的研究状况及已有结果

两组变量的典型相关分析:

记两组随机变量$X_1 = (x_{11}, x_{12}, ..., x_{1n_1})$和$X_2 = (x_{21}, x_{22}, ..., x_{2n_2})$ 且它们的方差矩阵及对应的协方差矩阵分别记作Σ_{11}, Σ_{22} 和 Σ_{12}. 如果存在$a_1 \in \mathbb{R}^{n_1}$和$b_1 \in \mathbb{R}^{n_2}$是下述最优化问题的解

$$\rho_1 = \max_{a_1^T \Sigma_{11} a_1 = 1, b_1^T \Sigma_{22} b_1 = 1} \rho(a_1^T X_1, b_1^T X_2),$$

则称$a_1^T X_1, b_1^T X_2$是X_1和X_2 的第一对典型相关变量，ρ_1称为第一个典型相关系数。

典型相关的研究状况及已有结果

两组变量的典型相关分析:

如果存在 $a_k \in \mathbb{R}^{n_1}$ 和 $b_k \in \mathbb{R}^{n_2}$ 使得

(1) $a_k^T X_1, b_k^T X_2$和前面$k-1$ 对典型变量都不相关;

(2) $a_k^T \Sigma_{11} a_k = 1, b_k^T \Sigma_{22} b_k = 1$;

(3) $a_k^T X_1$ 与 $b_k^T X_2$ 的相关系数最大

成立, 则称 $a_k^T X_1, b_k^T X_2$ 是 X_1 和 X_2 的第 k 对典型相关变量，它们之间的相关系数称为第 k 个典型相关系数。

典型相关的研究状况及已有结果

两组变量的典型相关分析:

前人的结论：假设这里的 $\Sigma_{ii}(i=1,2)$ 均可逆，则 a_k 是矩阵$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ 的对应于第 k 大的特征值的特征向量, b_k 是矩阵 $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ 的对应于第 k 大的特征值的特征向量,对应的特征值是典型相关系数的平方。

典型相关的研究状况及已有结果

多组变量的典型相关准则:

假设有n个随机变量$\xi_1, \ldots, \xi_n$，分成m组，形成样本矩阵$X = [X_1, \ldots, X_m], X_j \in \mathbb{R}^{N\times n_j}, \sum_{j=1}^m = n$，$N$为抽样次数. 可设$X$的列已中心化(减去了均值)，从而得样本协方差阵

$$A = X^T X = (A_{ij})_{m\times m} \qquad A_{ii} \in \mathbb{R}^{n_i\times n_i}$$

设所有A_{ii}都可逆，从而形成联合相关矩阵$R = (R_{ij})_{m\times m}, R_{ij} = A_{ii}^{-\frac{1}{2}} A_{ij} A_{jj}^{-\frac{1}{2}}$

典型相关的研究状况及已有结果

Kettenring引入的准则(1971):

假设$\xi_1, \ldots, \xi_n$的分组为$y_1, \ldots, y_m$，考虑y_i的线性组合：$z_i = t_i^T y_i, t_i \in \mathbb{R}^{n_i}$，则$z_1, \ldots, z_m$的协防差阵(由样本形成的)为$A(t) = (t_i^T A_{ij} t_j)_{m\times m}$, 相关矩阵为$R(t) = (t_i^T R_{ij} t_j)_{m\times m}$。他考虑用$R(t)$的适当度量来定义$z_1, \ldots, z_m$的相关性。因此，他引入了5种准则。

典型相关的研究状况及已有结果

Kettenring引入的准则(1971):

$$\text{SUMCOR}: \quad \max_t \quad e^T R(t) e, \quad \text{s.t.} \quad t \in \Sigma_m$$

$$\text{MAXVAR}: \quad \max_t \quad \lambda_{max}(R(t)), \quad \text{s.t.} \quad t \in \Sigma_m$$

$$\text{MINVAR}: \quad \min_t \quad \lambda_{min}(R(t)), \quad \text{s.t.} \quad t \in \Sigma_m$$

$$\text{GENVAR}: \quad \min_t \quad \det(R(t)), \quad \text{s.t.} \quad t \in \Sigma_m$$

$$\text{SSQCOR}: \quad \max_t \quad \|R(t)\|_F^2, \quad \text{s.t.} \quad t \in \Sigma_m$$

其中$\Sigma_m = \{t = \begin{pmatrix} t_1 \\ \vdots \\ t_m \end{pmatrix} \in \mathbb{R}^{n_i}, \|t_i\|_2 = 1\}, e \equiv \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

典型相关的研究状况及已有结果

Van de Geer的想法(1984):

$\xi_1, \ldots, \xi_n$的相关性大致来源于两个方面:

“组间” 相关性:$y_1, \ldots, y_m$ (看成整体)的相关性。

“组内” 相关性: y_i 内部(n_i个变量)的相关性。

他认为Kettenring 的分析，强调了组间的相关性，而在一定程度上忽略了组内的相关性。因此他不考虑相关矩阵$R(t)$与R, 而是考虑协方差矩阵$A(t)$与A, 引入了4个准则。

典型相关的研究状况及已有结果

Van de Geer引入的准则(1984):

$$\text{MAXBET}: \quad \max_t \quad e^T A(t) e, \quad \text{s.t.} \quad t \in \Sigma_m$$

$$\text{MAXDIFF}: \quad \max_t \quad e^T (A(t) - D(t)) e, \quad \text{s.t.} \quad t \in \Sigma_m$$

$$\text{MAXNEAR}: \quad \min_t \quad t^T (mD - A) t), \quad \text{s.t.} \quad t \in \Sigma_m$$

$$\text{MAXRAT}: \quad \max_t \quad \frac{t^T A t}{t^T D t}, \quad \text{s.t.} \quad t \in \Sigma_m$$

其中D是A的对角块，即$D = diag(A_{11}, ..., A_{mm}), D(t)$是$A(t)$的对角块即$D(t) = diag(t_1^T A_{11} t_1, ..., t_m^T A_{mm} t_m)$。

典型相关的研究状况及已有结果

Van de Geer引入的准则(1984):

如果 $t = t^{(1)}$ 是上述准则的全局解, $t^{(2)}$ 的求解是在剩余空间里的,即用$X_i(I_{n_i} - t_i t_i^T)$ 代替原来的 X_i. 以上四种准则的解叫做连续 (successive) 形式的.

Van de Geer 还提出了这四个准则的联立 (simultaneous) 形式的解: 令

$$\Omega_m = \{T = \begin{pmatrix} T_1 \\ \vdots \\ T_m \end{pmatrix} \in \mathbb{R}^{n\times r}, T_i^T T_i = I_r\}$$

典型相关的研究状况及已有结果

解为联立(simultaneous)形式的准则:

$$\text{MAXBET}: \quad \max \quad \operatorname{tr}(T^T A T) \quad \text{s.t.} \quad T \in \Omega_m$$

$$\text{MAXDIFF}: \quad \max \quad \operatorname{tr}(T^T A T - T^T D T) \quad \text{s.t.} \quad T \in \Omega_m$$

$$\text{MAXRAT}: \quad \max \quad \frac{\operatorname{tr}(T^T A T)}{\operatorname{tr}(T^T D T)} \quad \text{s.t.} \quad T \in \Omega_m$$

$$\text{MAXNEAR}: \quad \min \quad \operatorname{tr}(m T^T D T - T^T A T) \quad \text{s.t.} \quad T \in \Omega_m$$

典型相关的研究状况及已有结果

一些数值算法:

当$r = 1$ 时，对于MAXBET准则，方法有:

- Horst最早引入的幂法（也称Horst方法）
- Chu和Watterson提出的Gauss-Seidel迭代算法
- 孙继广提出的P-SOR 迭代法
- 秦晓伟提出的对称的P-SOR方法(P-SSOR)

这些算法都不保证得到全局解。

研究内容及方法

几种特殊情形：

- $X_i^T X_i = I$ 或者T_i 是方阵时，这4种准则等价即解都是一样的。
- 当X_i 是列规范正交阵时，这4个准则与SUMCOR准则等价。

猜想这4种准则应该有某种联系。

研究内容及方法

拟研究的问题如下:

- (1)联立形式的解与连续形式的解又有何种联系？它们之间的解有无近似性？
- (2)这四种准则的解之间有无近似性？或者他们之间的解有无联系？
- (3)如何较好的求得全局解？

拟解决的关键科学问题:

- 研究4个准则之间的内在联系
- 发展有效的数值解法

主要文献

(1) Chu M T, Watterson J L. On a multivariate eigenvalue problem, part I: algebraic theory and a power method[J]. Siam Journal on Scientific Computing, 1993, 14(5):1089-1106.

(2) Hotelling, H.: Relations between two sets of variables, Biometrika 28, 321-377 (1936)

(3) Kettenring, J.R.: Canonical analysis of several sets of variales, Biometrika 58, 433-451 (1971)

主要文献

(4) Steel R G D. Minimum Generalized Variance for a set of Linear Functions [J]. Annals of Mathematical Statistics, 1951, 22(3):456-460.

(5) Van De Geer, J.P.: Linear relations among k sets of variables, Psychometrika 49(1), 70-94 (1984)

(6) 秦晓伟，关于解最大相关问题P-SSOR算法的收敛性，计算数学.2011,33:345-356

(7) 孙继广. 多参数特征值问题的一种算法(I)[J]. 计算数学, 1986, 8(2):354-363

致谢

多 谢！