典型相关分析在均匀设计数据分析中的应用

郭秀花　鲍卫华　罗艳侠　周诗国

（北京军医学院数学教研室）

1.引言

均匀设计（Uniform design）[1, 2]，自 1978 年由中国的著名数学家方开泰和王元提出来之后有了飞速的发展，在国内外各行各业的应用越来越广泛。当各因素的水平数较多时，为了将试验点均匀地散布在试验范围内，探讨因素不同水平组合下响应（因变量）测量值的较佳结果，用均匀设计表来安排实验时，试验数据的结果分析要视因变量的情况而定。当只有 1 个因变量时，若为连续变量的取值，一般采用多元线性回归分析；若为分类变量取值 0、1（或等级资料），可采用 Logistic 回归分析[3]。当有多个因变量时，目前大都多次运用多元线性回归分析，可以得到每个因变量与全部自变量的关系式，但很难找到使每个因变量都达到较佳结果的试验组合。本文利用典型相关分析，探讨全部自变量与全部因变量之间的整体关联性。

2.方法简介

典型相关[4]分析是研究一组指标与另一组指标间关系的统计方法，它将两组指标当作一个整体进行分析，寻找其相应的典型变量，因而将两组指标间的相关信息更加充分地挖掘出来。在许多实际问题中，需要研究两组变量之间的相关性。例如，研究原料的主要质量指标(X_1、… 、X_p)与其相应产品的主要质量指标(Y_1、… 、Y_q)之间的相关性；研究居民的营养状况的一组指标(X_1、… 、X_p)与其健康状况的另一组指标(Y_1、… 、Y_q)之间的相关性等等。当 p=q=1 时，就是两个变量之间的简单相关分析问题；当 P>1、q=1 时，就是一个因变量与多个自变量之间的多元相关分析问题；当 p、q 均大于 1 时，就是研究两组多变量之间的相关性，即为典型相关分析。

利用主成分的思想，找一组系数 $A=(a_1、… 、a_p)'$及 $B=(b_1、… 、b_q)'$，使得新变量

$V_1=a_1X_1+\cdots+a_pX_p\equiv A'x$ 与 $W_1=b_1Y_1+\cdots+b_qY_q\equiv B'Y$ 之间有最大可能的相关系数，称(V_1，W_1)为第 1 对典型相关变量，它们之间相关系数 $r_{(V1, W1)}$简记为 r_1；同理，可求得第 2、第 3、 …、 第 K(K 小于等于 p，q 中较小者)对典型相关变量以及与之对应的相关系数 r_2、… 、 r_k。各对典型相关变量所包括的相关信息互不交叉，且满足：$1\geq r_1\geq r_2\geq\cdots\geq r_k\geq 0$；

$r_{(Vi, Vj)}=0$， $r_{(Wi, Wj)}=0$， $r_{(Vi, Wj)}=0$ （$i\neq j$）；各 V_i和 W_i的均值都为 0，方差都为 1。

求出典型变量对和典型相关系数后，把具有统计学意义的典型相关系数所对应的典型变量对保留下来，并给予合理的解释，是进行典型相关分析好坏的关键。

3.实例应用

资料选自《均匀设计论文选》第 2 集“ 均匀设计在清净型硼酸盐添加剂研制中

的应用 ”燕山石化公司炼油厂研究所的.訾立钧[5]：

6 个自变量名称及取值范围：

原料 A（X1）：210-400

原料 B（X2）：50-240 工业品

溶剂（X3）：500-1400

条件 1（X4）：45-90；条件 2（X5）：100-157；条件 3（X6）：30-220。

在用均匀设计表安排试验后，分别测试产品的 4 个指标，即 4 个因变量：产品的收率（%）Y1；产品中硼元素的百分含量（%）Y2；产品中钙元素的百分含量（%）Y3；产品中的碱值 Y4，试验方案及结果见表 1。

表 1` 研制超碱性硼酸盐清净剂均匀设计实施方案及结果

序号	原料 A	原料 B	条件 1	条件 2	条件 3	条件 4	收率	硼含量	钙含量	碱值
	X1	X2	X3	X4	X5	X6	Y1	Y2	Y3	Y4
1	210	80	1400	55	145	210	77.66	3.9	10.50	202.88
2	220	120	1400	65	130	190	67.06	5.3	10.34	184.09
3	230	160	1300	80	115	170	45.51	5.1	8.57	145.92
4	240	200	1300	90	100	150	44.55	2.8	5.29	26.52
5	250	240	1200	50	148	130	54.49	6.4	7.35	114.42
6	260	70	1200	65	133	110	82.42	3.4	10.67	232.48
7	270	110	1100	75	118	90	60.31	4.1	9.92	186.44
8	280	150	1100	90	103	70	63.02	4.1	9.35	162.88
9	290	190	1000	50	151	50	67.29	5.3	8.96	149.46
10	300	230	1000	60	136	30	56.42	5.5	9.67	181.86
11	310	60	900	75	121	220	90.54	2.3	10.11	233.69
12	320	100	900	85	106	200	81.31	3.3	9.76	191.09
13	330	140	800	45	154	180	76.06	4.5	9.50	190.04
14	340	180	800	60	139	160	69.23	5.0	9.26	183.65
15	350	220	700	70	124	140	60.61	4.4	10.17	136.80
16	360	50	700	85	105	120	93.66	1.5	10.14	234.30
17	370	90	600	45	157	100	80.43	3.0	9.87	223.60
18	380	130	600	55	142	80	76.27	3.0	11.34	204.46
19	390	170	500	70	127	60	73.57	3.8	10.44	227.20
20	400	210	500	80	112	40	65.66	3.7	10.06	181.16

4 对典型相关系数及其假设检验的主要结果见表 2。

表2　　4对典型相关系数及其假设检验的主要结果

	Canonical Correlation	Eigenvalue	Proportion	Cumulative	Approx F	Pr > F
1	0.975479	19.6439	0.9328	0.9328	3.3477	0.0005
2	0.707759	1.0037	0.0477	0.9805	0.9585	0.5169
3	0.497595	0.3291	0.0156	0.9961	0.5978	0.7702
4	0.275498	0.0821	0.0039	1.0000	0.3559	0.7857

从表2可知，自变量与因变量之间的关联性主要来自第1对典型相关系数。标准化相关系数如下：

Standardized Canonical Coefficients for the 'VAR' Variables

	V1	V2	V3	V4
X1	0.9027	-3.2013	-3.7864	1.8656
X2	-0.8824	0.0086	-0.2370	-0.0126
X3	0.3907	-3.1206	-3.4855	1.5958
X4	0.4925	-0.3394	5.2554	-6.8571
X5	0.4391	-1.3872	5.4047	-6.6354
X6	-0.0738	-0.3945	-0.9303	-0.5176

Standardized Canonical Coefficients for the 'WITH' Variables

	W1	W2	W3	W4
Y1	0.1075	-1.5891	-1.6313	-0.1510
Y2	-0.5742	-1.2350	0.0211	-0.0408
Y3	-0.0159	0.0670	-0.7055	1.9957
Y4	0.5384	0.4408	2.3776	-1.4701

观察 V1 与 X1-X6 之间、W1 与 Y1-Y4 之间标准化典型相关系数的大小和符号，可以得出结论：自变量主要是：原料 A、原料 B，且与原料 A 成正比、与原料 B 成反比；因变量主要是：产品中硼元素的百分含量（%）、产品中的碱值，且与产品中硼元素的百分含量（%）成反比、与产品中的碱值成正比。

4.结语

对于均匀设计模型的数据分析，要找出自变量与因变量彼此之间的内在联系，除了本文提到的典型相关分析之外，还可以采用数学结构模型[6]等方法；均匀设计模型数据的预测分析可以采用人工神经网络（已见报道）[7]、小波分析[8]等方法。

模型应是实际原型的模拟，而不是原型本身。利用各种数学模型探讨某个实际问题，就是要寻找最为符合实际规律的较好的模型，这一点很像“盲人摸象”，很难或者说不可能达到最佳的境界。即使使用了最复杂的数学处理而建立的数学模型，实际上同真正的现实情况还有不少距离。但是，随着数学模型发展的不断完美，会有更多的数学方法渗入到均匀设计模型分析，使得建立的各种数学模型逐步接近实际，为其理论和实践的发展服务。

（本文诚蒙方开泰教授指导，特此致谢！）

参考文献

[1] 方开泰，马长兴着.正交与均匀试验设计.科学出版社，2001

[2] Fang K. T. and Li, J. K. Some new uniform designs, Chinese Science Bulletin, 21, 1921-24, 1994.

[3] 郭秀花，张学中，徐桂永等.Logistic 回归模型介绍及其在均匀设计数据分析

中的作用.均匀设计理论及其应用研讨会论文集.香港浸会大学，1999 年 10 月
[4] 郭秀花，李东，徐桂永等．典型相关在未婚青年生殖健康分析中的应用.中国公共卫生.2001，17（3）：257-258
[5] 訾立钧 均匀设计在清净型硼酸盐添加剂研制中的应用.均匀设计论文选(第 2 集)：156-163
[6] 李青云，孙厚才，陈智勇，等.均匀设计与人工神经网络技术在三峡二期围堰柔性材料配合比优化中的应用研究．均匀设计论文选（第 2 集）：139-147
[7] 孙尚拱编着.医学多变量统计与统计软件（第一版）.北京医科大学出版社，2000
[8] 郭秀花，林济南，曹务春.探讨基于小波分析的季节性时间序列预测模型.数理医药学杂志.2003，16（3）：195-197