

基于计算机试验的均匀设计

徐洪泉 沈世镒

(南开大学数学系)

摘 要

本文在计算机试验的基础上,提出了最小相关准则和最小距离离差准则,并将信息论中的 Hamming 距离和 Lee 距离引入到计算机试验中,证明了均匀设计在 Hamming 距离下的最优性和部分好格子点均匀设计在 Lee 距离下的最优性. 基于偏差的考虑,给出了一类新的好格子点均匀设计和一个学习算法,利用这个学习算法,给出了基于 Lee 距离的最小距离离差准则的均匀设计表的构造方法. 通过与已有的好格子点均匀设计和循环拉丁方均匀设计作比较,证明了文中的均匀设计在距离和偏差意义下有更好的均匀性.

关键词 均匀设计, 偏差, 计算机试验, D-最优准则, 最小相关准则, 最小距离离差准则, Hamming 距离, Lee 距离

分类号 (中图)O 212. 6; (1991M R); 62K05, 62K99

§1 引 言

均匀设计是由方开泰和王元于1978年提出的,基本方法是近似分析中的数论方法,以偏差为均匀性准则^[1,2]. 均匀设计所用的统计模型是回归分析,一般采用二阶模型. 文献[3]给出了均匀设计的一个综合报告.

近年来,国外对计算机试验有许多的讨论^[4,5],计算机试验中常用的设计是 Latin Hypercube Design (LHD),而 LHD 实质上是一种均匀设计. 文献[6]给出了计算机试验的一个综合报告.

本文1996年7月7日收到

本文获国家自然科学基金(19671048)资助

本文首先讨论了计算机试验模型和最优准则 在D-准则的基础上,提出了最小相关准则和最小距离离差准则,并给出了判别最优设计的一个引理 通过将信息论中的Hamming距离和Lee距离引入到计算机试验中,证明了均匀设计在Hamming距离下的最优性和部分好格子点均匀设计在Lee距离下的最优性 出于偏差的考虑,我们给出了一类新的好格子点均匀设计和一个学习算法 利用这个学习算法,给出了基于Lee距离的最小距离离差准则的均匀设计表的构造方法 通过与已有的好格子点均匀设计和循环拉丁方均匀设计作比较,证明了我们的均匀设计在距离和偏差意义下有更好的均匀性 最后,我们给出了对均匀设计的一些看法

§2 计算机试验模型和最优准则

设 T 是 d 维空间中的一个点集, T 可看作整个试验区域,称 S 是试验次数为 n 的一个设计,如果 S 是 T 中含有 n 个点的子集,记为 $S = \{s_1, s_2, \dots, s_n\}$.

设 Y 是定义在 T 上的一个平稳Gauss过程,具有均值0,协方差

$$\text{Cov}(Y(t), Y(u)) = \sigma^2 \cdot R(t, u), \quad t, u \in T, \quad (1)$$

其中 σ^2 是方差, $R(t, u)$ 是点 t 和点 u 的相关系数 令设计 S 的相关矩阵为

$$R_s = (R(s_i, s_j)). \quad (2)$$

我们只讨论非退化情形,即设 R_s 是正定矩阵

定义1 D-最优准则^[7](又称为熵准则^[8]或信息准则¹⁾定义为

$$\max_S |R_s| = \max_S |R(s_i, s_j)| \quad (3)$$

按主对角线将行列式展开,在 $R(s_i, s_j) \rightarrow 0 (i \neq j)$ 时得到

$$|R(s_i, s_j)| \approx 1 - \sum_{i < j} R^2(s_i, s_j). \quad (4)$$

于是,我们得到作为近似D-最优的最小相关准则 其直观意义是,使设计点之间的平均相关性最小

定义2 最小相关准则(称为C-准则)定义为

$$\min_S \sum_{i \neq j} R^2(s_i, s_j). \quad (5)$$

设 d 是 $T \times T$ 上的一个广义距离^[7]. 当试验者对相关函数 R 无任何先验知识时,一般取 R 为距离的非负单减函数,即取

$$R(t, u) = R(d(t, u)) \geq 0, \quad (6)$$

且 $R = R(d)$ 是距离 d 的单减函数

定义3 设 S 是一个设计,令

$$B_u = B_u(S) = \# \{(s, s') : d(s, s') = u, s \neq s', s, s' \in S\}, u = 0, 1, \dots \quad (7)$$

称 $\{B_u : u = 0, 1, \dots\}$ 为 S 的距离分布 称设计 S_1 比设计 S_2 有较小的距离离差(distance aberration),如果存在 r ,使

$$B_v(S_1) = B_v(S_2), 0 \leq v < r; \text{ 且 } B_r(S_1) < B_r(S_2). \quad (8)$$

¹⁾沈世镒,徐洪泉(1996). 从信息论看试验设计.

称设计 S 具有最小距离离差 (Minimum Distance Aberration, MDA), 如果不存在比 S 有更小距离离差的设计

上述最小距离离差是 $\max_{\min}$ 距离设计^[8] 与 $\min$ aberration^[9] 概念的推广. 在 $R(d)$ 单减的假设下, 一般可以认为 MDA 设计有较小的相关性. 我们有以下的引理:

引理1 设 $R(d) \geq 0$ 是距离 d 的单减凸函数, S_0 是 n 个点的设计, 满足

(i) S_0 的平均距离是所有 n 个点的设计中最大的;

(ii) S_0 是等距的 (任何两点间的距离都相同).

则 S_0 是 D-最优的, 也是 C-最优的

至此, 我们对计算机试验进行了一般性的讨论. 在以下部分, 我们利用信息论中的 Hamming 距离和 Lee 距离对均匀设计作一些探讨.

§3 均匀设计的优良性

设 $S = \{s_1, s_2, \dots, s_n\}$ 是 d 维空间中的一个设计. 为了方便, 将 S 写成 $n \times d$ 的矩阵

$$S = \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix}, \quad s_i = (s_{i1}, s_{i2}, \dots, s_{id}), \quad i = 1, 2, \dots, n \quad (9)$$

如果 s_{ij} 只在 $\{1, 2, \dots, n\}$ 中取值, 则称 S 是一个 n 水平, d 因素, n 次试验的设计, 记为 $S_n(n^d)$.

称设计 S 是均匀设计, 如果矩阵 S 的每一列是 $1, 2, \dots, n$ 的一个排列^[10].

定义4 点 s_i, s_j 之间的 Hamming 距离和 Lee 距离分别定义为

$$d_H(s_i, s_j) = \sum_{k=1}^d w_H(s_{ik} - s_{jk}), \quad (10)$$

$$d_L(s_i, s_j) = \sum_{k=1}^d w_L(s_{ik} - s_{jk}), \quad (11)$$

其中,

$$w_H(x) = \begin{cases} 0, & \text{如果 } x = 0, \\ 1, & \text{如果 } x \neq 0, \end{cases} \quad (12)$$

$$w_L(x) = \begin{cases} |x|, & \text{如果 } |x| \leq n/2, \\ n - |x|, & \text{如果 } n/2 < |x| \leq n, \end{cases} \quad (13)$$

分别为 Hamming 重量和 Lee 重量

在通信理论中, Hamming 距离适合于正交调制 (orthogonal modulation), 而 Lee 距离适合于相位调制 (phase modulation)^[11]. Hamming 距离可用于正交设计的讨论. Lee 距离可看作是环面上的距离. 方开泰和王元提出的均匀设计是以数论方法为基础的. 在此我们用 Lee 距离来讨论均匀设计是适宜的. 例如, 华罗庚和王元 (1978)^[12] 提到, 将单位立方体看成

² 沈世镒, 徐洪泉 (1996). 从信息论看试验设计.

环面更合适(p. 115), 对于非周期函数, 进行周期化可以提高计算精度(p. 185). 又如, 张润楚和王兆军(1996)^[13]提出的均匀设计抽样实际上也将单位立方体看成了环面

利用信息论中的结果^[11], 我们得到:

引理2 在Hamm ing 距离和Lee 距离下, 均匀设计的平均距离是最大的

显然, 均匀设计在Hamm ing 距离下是等距的, 由引理1和引理2得:

定理1 设 $R(d) = R(d_H) \geq 0$ 是Hamm ing 距离 d_H 的单减凸函数, 则均匀设计是D-最优和C-最优的

定理1证明了均匀设计在Hamm ing 距离下是一致最优的 利用Hamm ing 距离无法区分均匀设计, 以下考虑Lee 距离

设 p 是奇素数, 由生成向量 $\pm 1, \pm 2, \dots, \pm \frac{p-1}{2} \pmod{p}$ 可得到好格子点均匀设计 $UG_p\left(p^{\frac{p-1}{2}}\right)$ [定义见下一节], 由生成向量 $1, 2, \dots, p-1$ 可得到好格子点均匀设计 $UG_p(p^{p-1})$. 易证上述两种均匀设计在Lee 距离下是等距的, 由引理1和引理2得:

定理2 设 $R(d) = R(d_L) \geq 0$ 是Lee 距离 d_L 的单减凸函数, 则上述两种好格子点均匀设计是D-最优和C-最优的

§4 均匀设计的构造

4.1 一类新的好格子点均匀设计表

文献[14]给出了四种均匀设计表的构造方法, 我们主要采用好格子点均匀设计(UDGLP).

设 $h_1, h_2, \dots, h_m$ 是所有比 n 小, 且与 n 互素的自然数(其中 $m = \varphi(n)$ 是 n 的Euler数). $n \times m$ 矩阵 $U_{n,m} = (u_{ij})$ 定义为

$$u_{ij} \equiv i \cdot h_j \pmod{n}, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m; \quad (14)$$

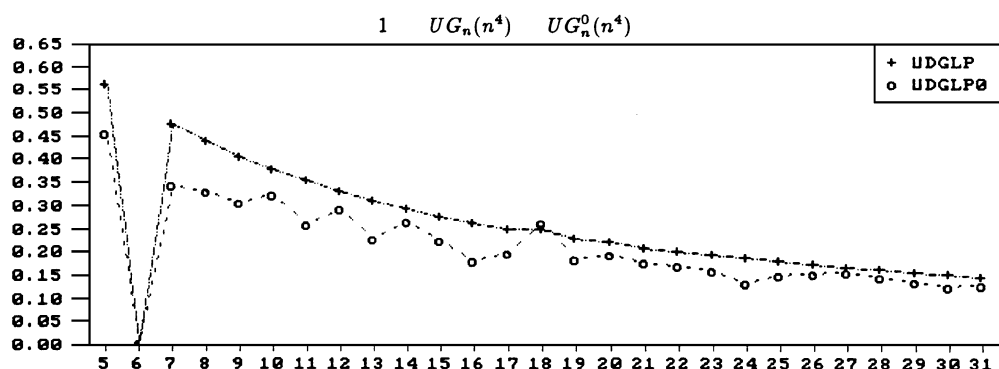
其中的 u_{ij} 在 $1, 2, \dots, n$ 中取值 对于给定的 d , 从 $U_{n,m}$ 表中挑选 d 列, 偏差最小的设计记为 $UG_n(n^d)$. 注意到上述定义的矩阵 $U_{n,m}$ 的最后一行的元素都是 n , 去掉最后一行, 得到 $(n-1) \times m$ 矩阵 $U_{n-1,m}^*$. 对于给定的 d , 从 $U_{n-1,m}^*$ 表中挑选 d 列, 偏差最小的设计记为 $UG_{(n-1)}^*((n-1)^d)$, 称为星型UDGLP^[14].

注意到偏差是依赖于原点的, 为此, 我们提出一类新的UDGLP. 我们将矩阵 $U_{n,m}^0 = (u_{ij})$ 定义为

$$u_{ij} \equiv i \cdot h_j + 1 \pmod{n}, \quad i = 1, 2, \dots, n; j = 1, 2, \dots, m; \quad (15)$$

其中的 u_{ij} 在 $1, 2, \dots, n$ 中取值 对于给定的 d , 从 $U_{n,m}^0$ 表中挑选 d 列, 偏差最小的设计记为 $UG_n^0(n^d)$. 我们称之为零型UDGLP.

我们对上述基于偏差的三类UDGLP作了比较, 得到结论: $UG_n^0(n^d)$ 一般比 $UG_n(n^d)$ 有更小的偏差, 而 $UG_n^0(n^d)$ 与 $UG_n^*(n^d)$ 之间没有一般的优劣关系 例如, 图1给出了当 $5 \leq n \leq 31$ 且 $d=4$ 时 $UG_n(n^d)$ 与 $UG_n^0(n^d)$ 偏差的比较, 其中UDGLP、UDGLP0分别表示 $UG_n(n^d)$ 和 $UG_n^0(n^d)$ 设计(注: 偏差为0.0表示相应的设计不存在, 以下同). 所以, 我们主要采用 $UG_n^0(n^d)$ 和 $UG_n^*(n^d)$.

图1 $UG_n(n^4)$ 与 $UG_n^0(n^4)$ 偏差的比较

4.2 一个学习算法

为了得到偏差较小的均匀设计, 我们采用了一个学习算法(称为学习算法B). 其基本方法是通过某些列作平移. 学习算法B可概述为以下几步:

1. 对给定的均匀设计 $U_0 = U_n(n^d)$, 计算其偏差, 记为 d_0 .
2. 根据某一规则A, 得到对每一列可能作的平移, 记为 $a = (a_1, a_2, \dots, a_d)$, 其中 $a_i = 1, 0, -1$.
3. 选定某一满足 $a_i \neq 0$ 的列 i , 将 a_i 加到这一列的所有数上(mod n), 得到一个新的设计, 记为 U_i , 并计算其偏差, 记为 d_i .

4. 如果 $d_i < d_0$, 则用 U_i 代替 U_0 , d_i 代替 d_0 , 回到步骤1. 否则, 选择下一个满足 $a_i \neq 0$ 的 i , 回到步骤3. 在没有可选的 i 时, 停止学习, 并输出结果.

学习算法B的效果, 可参见后面的图2和图3. 步骤2中的规则A可参见我们的程序.

4.3 基于最小距离离差准则的均匀设计表的构造

这里, 我们给出基于Lee距离的MDA的均匀设计表的构造方法(称为Lee-MDA构造法). 设 $U_{n,m}$ 是任一均匀设计表, 对给定的 d ,

1. 在Lee距离下, 从所有 $\binom{m}{d}$ 个设计中选择具有MDA的设计, 记为 $U_{A_n}(n^d)$.
2. 对于 $U_{A_n}(n^d)$, 采用学习算法B, 得到 $U_{B_n}(n^d)$.

利用上述Lee-MDA构造法, 我们给出了 $5 \leq n \leq 31, 2 \leq d \leq 6$ 时的两类好格子点(零型和星型)均匀设计表(为了区别, 称为基于Lee-MDA的UDGLP). 我们的程序和均匀设计表可以通过 Email: shensy@SUN.nankai.edu.cn 得到.

4.4 均匀设计表的比较

我们对 $5 \leq n \leq 31, 2 \leq d \leq 4$ 时基于偏差和Lee-MDA构造法的零型和星型好格子点均匀设计的偏差作了比较. 结论是, $d=2$ 时, 两种方法是相当的. $d=3, 4$ 时, 由Lee-MDA构造法得到的UDGLP比基于偏差的UDGLP有更小的偏差, 而与经过学习的UDGLP相当. 例如, 图2、图3分别给出了 $d=3, 4$ 时零型UDGLP偏差的比较. 其中, UDGLP0是偏差最小的零型UDGLP, 而UDGLP0B是对UDGLP0采用学习算法B得到的. ULGLP0B是由Lee-

MDA 构造法得到的零型UDGLP. 对于星型UDGLP, 也有类似的图形

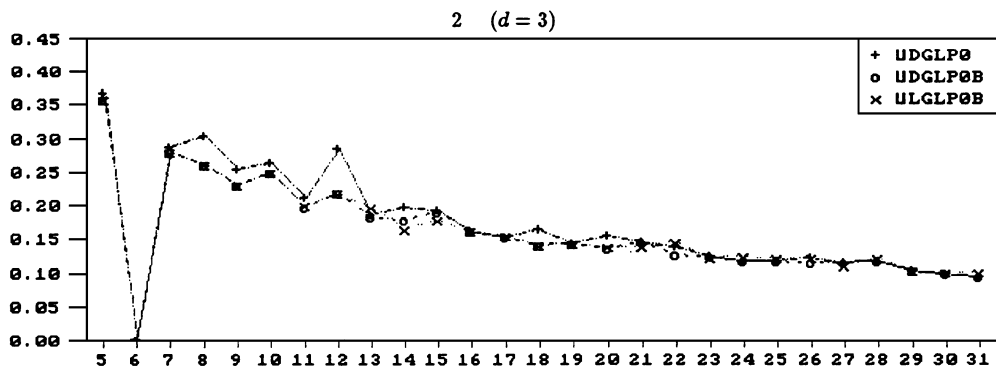


图2 零型好格子点均匀设计偏差的比较

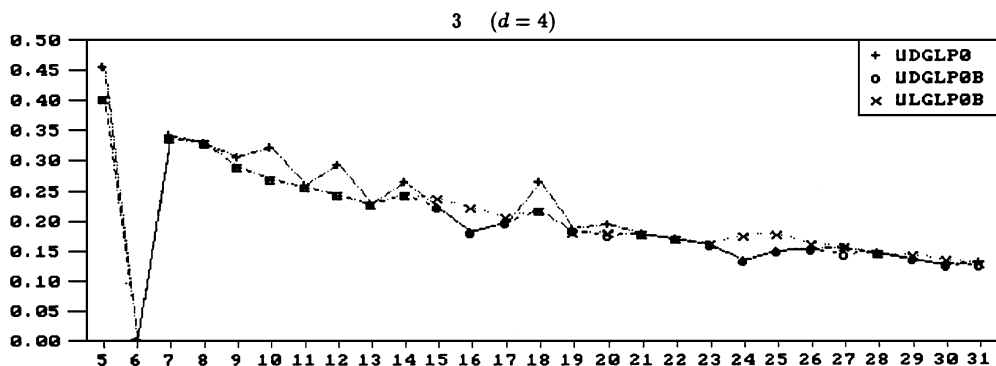


图3 零型好格子点均匀设计偏差的比较

当 $d=2, 3, 4$ 时, 我们与文献[14]中的循环拉丁方均匀设计(UDCLS)也作了比较³, 结果是类似的. 需要指出的是, UDCLS 的偏差并不比零型和星型UDGLP 中的较小偏差更小. 另外, 在MDA 准则下, UDCLS 是不可取的, 因为它们的距离较小.

当 $d=5, 6$ 时, 我们与文献[10]中的表18作了比较^[3]. 结论是, 我们的均匀设计表有较小的偏差. 例如, $n=26, d=5$ 时, 我们给出的四个设计的偏差分别为 0.2146, 0.1833, 0.2486, 0.2238, 都小于表18^[10]中设计的偏差 0.2564. 再如, $n=26, d=6$ 时, 我们给出的四个设计的偏差分别为 0.2391, 0.2340, 0.2547, 0.2155, 都小于表18^[10]中设计的偏差 0.3024.

§5 对均匀设计的几点看法

1. 文献[15]就一阶线性模型讨论了D-准则和A-准则下的均匀设计. 我们认为, 就一阶

³文献[10]表18中的偏差是错误的——作者注

线性模型讨论均匀设计是没有意义的, 因为对于一阶线性模型, 最优设计是二水平的, 而不是均匀设计

2. 正交设计是基于方差分析模型的, 均匀设计比正交设计复杂, 采用计算机试验模型是适宜的 例如, 引理2证明了均匀设计在 Hamming 距离和 Lee 距离下的平均距离是最大的, 而定理1证明了均匀设计在 Hamming 距离下是一致最优的 关于计算机试验的一般模型和数据分析, 可参考有关文献

3. 以偏差为准则的均匀设计的理论依据是近似分析中的数论方法^[1,2]. 在近似分析中主要讨论大样本性质 ($n \rightarrow \infty$), 而均匀设计考虑的是小样本, 两者是有本质区别的 偏差是依赖于原点的, 它没有平移性 例如, 我们通过对 $UG_n(n^d)$ 的平移得到的 $UG_n^0(n^d)$ 有较小的偏差 再如, 我们可以找出一些偏差最小的均匀设计, 但在直观和距离上, 它们并不均匀 因此, 在构造均匀设计时, 考虑距离准则是必要的 基于前面的讨论, 我们推荐 Lee 距离和最小距离离差准则

4. 偏差计算公式是由文献[16, 17]给出的, 其计算复杂度是 $O(n^d)$, 只适合于 n 和 d 较小的情形 我们提出的基于距离离差准则和学习算法的构造法, 在时间上也有明显的优势 例如, 我们花了526秒, 得到 $UG_{31}^0(31^4)$, 偏差为0.1305 而我们得到偏差为0.1289的 $UB_{31}^0(31^4)$ 只花了5秒 以上时间是在我们的奔腾机(主频120M)上得到的 如果只采用距离准则, 学习算法和偏差计算是没有必要的

(本文作者通讯地址: 天津市南开大学数学系 邮编 300071)

参 考 文 献

- [1] 方开泰, 均匀设计, 应用数学学报, 3(1980), 363-372
- [2] 王 元, 方开泰, 关于均匀设计与试验设计(数论方法), 科学通报, 26(1981), 65-70
- [3] Fang Kaitai, Hickernell, F. J., The uniform design and its applications, ISI第五十届国际会议邀请报告, 北京, 1995
- [4] Sacks, J., Schiller, S. B. and Welch, W. J., Designs for computer experiments, *Technometrics*, **31** (1989), 41-47.
- [5] Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P., Designs and analysis of computer experiments (with discussion), *Statistical Science*, **4**(1989), 409-435
- [6] 张润楚, 王兆军, 关于计算机试验的设计理论和数据分析, 应用概率统计, **11**(1994), 420-436
- [7] Johnson, M. E., Moore, L. M. and Ylvisaker, D., Minimax and maximin distance designs, *Journal of Statistical Planning and Inference*, **26**(1990), 131-148
- [8] Shewry, M. C. and Wynn, H. P., Maximum entropy sampling, *Journal of Applied Statistics*, **14** (1987), 165-170
- [9] Fries, A. and Hunter, W. G., Minimax aberration 2^{k-p} designs, *Technometrics*, **22**(1980), 601-608
- [10] 方开泰, 均匀设计与均匀设计表, 科学出版社, 北京, 1994
- [11] Berlekamp, E. R., Algebraic Coding Theory, McGraw-Hill, Inc., New York, 1968
- [12] 华罗庚, 王元, 数论在近似分析中的应用, 科学出版社, 北京, 1978

- [13] 张润楚, 王兆军, 均匀设计抽样及其优良性质, 应用概率统计, 13·3(1996), 337-347.
- [14] 方开泰, 构造均匀设计表的若干方法, 中国均匀设计学会第二届学术交流会论文, 香港浸会大学, 1995
- [15] 丁元, 均匀设计优良性初探, 应用概率统计, 2(1986), 153-160
- [16] Bundschuh, P., 朱尧辰, 低维有限点集偏差的精确计算公式(I), 科学通报, 38(1993), 669
- [17] 朱尧辰, 低维有限点集偏差的精确计算公式(III), 科学通报, 39(1994), 1724

THE UNIFORM DESIGN BASED ON COMPUTER EXPERIMENTS

Xu Hongquan Shen Shiyi

(Dept. of Math., Nankai University, Tianjin 300071)

Abstract

Based on computer experiments the concepts of minimum correlation and minimum distance aberration (MDA) are proposed. By applying the Hamming distance and Lee distance of information theory, it is shown that uniform designs are optimal with Hamming distance and some uniform designs based on good lattice point sets (UDGLP) are optimal with Lee distance. A new class of UDGLP and a learning algorithm are suggested for discrepancy. With this algorithm, a construction method for uniform designs is given based on the MDA criterion of Lee distance. Compared with other UDGLP and UDCLS, the new uniform designs have better uniformity under distance and discrepancy.

Key Words Computer Experiments, Discrepancy, D-optimum, Hamming Distance, Lee Distance, Minimum Correlation, Minimum Distance Aberration, Uniform Design.

Subject Classification (CL)O 212.6; (1991M R)62K05, 62K99