Design and Modeling
in Computer and Physical Experiments

Kai-Tai Fang
BNU - HKBU United International College
and
Chinese Academy of Sciences

Workshop on Uncertainty Analysis and Quantification.

8 - 9 October 2008, Paris

Table of Contents

1 Motivation

In 1978 three big projects in system engineering raised the same type of problems to us. It needs one day calculation in a computer to obtain the output y, a solution of a system of differential equations, from the given input under the true model.

$$y = g(x_1, \cdots, x_s) = g(\boldsymbol{x}).$$

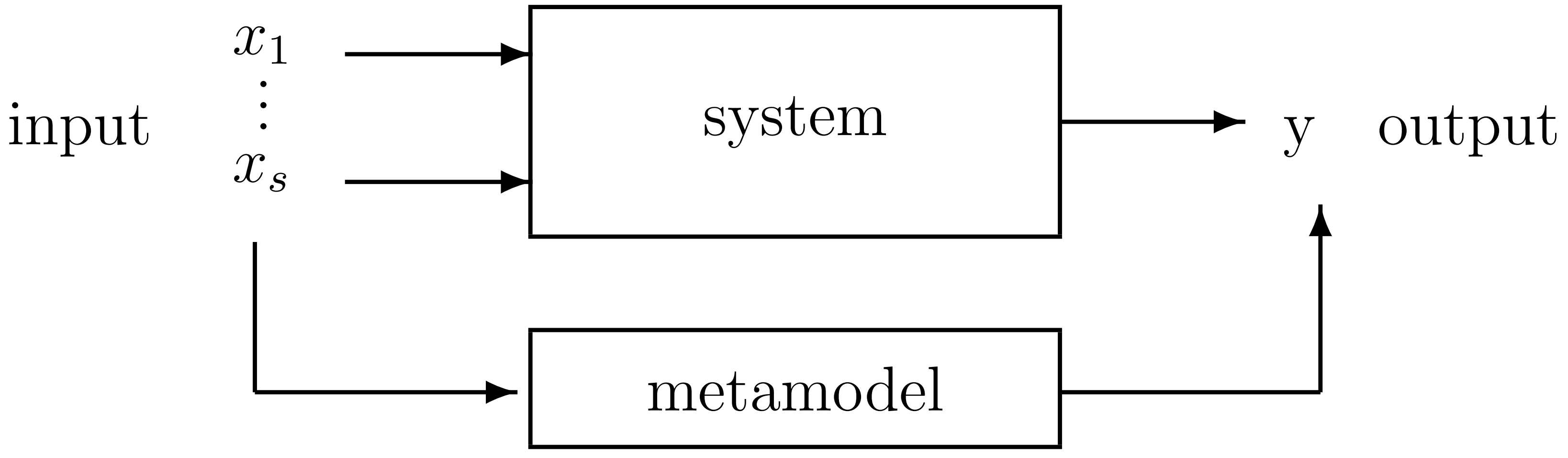
We wish to find a simple and approximate model (metamodel)

$$\hat{y} = \hat{g}(x_1, \cdots, x_s) = \hat{g}(\boldsymbol{x})$$

such that the difference of $|y(\boldsymbol{x}) - \hat{y}(\boldsymbol{x})|$ is small in a certain sense.

We need new concept: design and modeling of computer experiments.

Computer Experiments



Experiments with Model Uncertainty

$$y = g(x_1, \cdots, x_s) + \varepsilon,$$

where g is unknown and ε random error. We want to estimate g based on an experiment.

Both problems need a design that spreads experimental points evenly on the experimental domain.

- Space-filling design
- Uniform design

Both need a number of modeling techniques.

2. Approaches

A. Model

The simplest model is the overall mean model

Suppose that one wants to estimate the overall mean of y

$$\mathrm{Mean}(y) = \int_{C^s} g(\boldsymbol{x}) d\boldsymbol{x}$$

by the sample mean

$$\bar{y}(\mathcal{P}) = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i),$$

where $\mathcal{P} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ is a set of experimental points.

B. Two approaches

- **Latin hypercube sampling**

McKay, Beckaman, and Conover (1979), *Technometrics*

To choose a design $\mathcal{P}$ such that $\bar{y}(\mathcal{P})$ is unbiased estimator of Mean(y) and has a smaller variance than that by a random sampling.

- **Uniform design**

Wang and Fang (Fang (1980), Wang and Fang (1981))

To find a design such that $|\text{Mean}(y) - \bar{y}(\mathcal{P})|$ is as small as possible.

- Overall mean model

Of course, the overall mean model is far not enough in practice, but it is surprising that

- The overall mean model provides a simple way to develop methodology and theory

- The overall mean model results in that LHS/UD have a excellent performance for more comprehensive models.

A comprehensive review can refer to

Fang, K.T. Wang, Y. (1994), *Number-theoretic Methods in Statistics*, Chapman and Hall

Fang, K.T., D.K.J. Lin, P. Winker and Y. Zhang (2000), *Technometrics*

Fang, K.T. and Lin, D.K.J. (2003), one chapter in *Handbook on Statistics 22: Statistics in Industry*, Eds by R. Khattree and C.R. Rao, Elsevier.

Fang, K.T., Li, R. and Sudjianto, A. (2005), *Design and Modeling for Computer Experiments*, Chapman & Hall/CRC Press, London.

Fang, K.T. and Chan, L.Y. (2006), in *Springer Handbook of Enginnering Statistics*, Springer.

Latin Hypercube sampling

Step 1. Independently take s permutations $\pi_j(1), \cdots, \pi_j(n)$ of the integers $1, \cdots, n$ for $j = 1, \cdots, s$, i.e., generate a $LHD(n, s)$;

Step 2. Take ns uniform variates (or called random numbers in Monte Carlo methods) $U_i^j \sim U(0, 1)$, $i = 1, \cdots, n$, $j = 1, \cdots, s$, which are mutually independent. Let $\boldsymbol{x}_k = (x_k^1, \cdots, x_k^s)$, where

$$x_k^j = \frac{\pi_j(k) - U_k^j}{n}, \;\; k = 1, \cdots, n, \;\; j = 1, \cdots, s.$$

Then $D_n = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ is a LHS and is denoted by $LHS(n, s)$.

Example For generating a LHS for $n = 8, s = 2$, in the first step we generate two permutations of $\{1, 2, \cdots, 8\}$ as

(2,5,1,7,4,8,3,6) and (5,8,3,6,1,4,7,2)

to form a $LHD(8, 2)$ that is the matrix on the left below. Then we generate 16=8*2 random numbers to form a 8×2 matrix on the right below:

$$\begin{bmatrix} 2 & 5 \\ 5 & 8 \\ 1 & 3 \\ 7 & 6 \\ 4 & 1 \\ 8 & 4 \\ 3 & 7 \\ 6 & 2 \end{bmatrix}, \quad \begin{bmatrix} 0.9501 & 0.8214 \\ 0.2311 & 0.4447 \\ 0.6068 & 0.6154 \\ 0.4860 & 0.7919 \\ 0.8913 & 0.9218 \\ 0.7621 & 0.7382 \\ 0.4565 & 0.1763 \\ 0.0185 & 0.4057 \end{bmatrix}.$$

Now a LHS is given by

$$\frac{1}{8}\left[\begin{bmatrix} 2 & 5 \\ 5 & 8 \\ 1 & 3 \\ 7 & 6 \\ 4 & 1 \\ 8 & 4 \\ 3 & 7 \\ 6 & 2 \end{bmatrix} - \begin{bmatrix} 0.9501 & 0.8214 \\ 0.2311 & 0.4447 \\ 0.6068 & 0.6154 \\ 0.4860 & 0.7919 \\ 0.8913 & 0.9218 \\ 0.7621 & 0.7382 \\ 0.4565 & 0.1763 \\ 0.0185 & 0.4057 \end{bmatrix}\right] = \begin{bmatrix} 0.1312 & 0.5223 \\ 0.5961 & 0.9444 \\ 0.0491 & 0.2981 \\ 0.8143 & 0.6510 \\ 0.3886 & 0.0098 \\ 0.9047 & 0.4077 \\ 0.3179 & 0.8530 \\ 0.7477 & 0.1993 \end{bmatrix}.$$

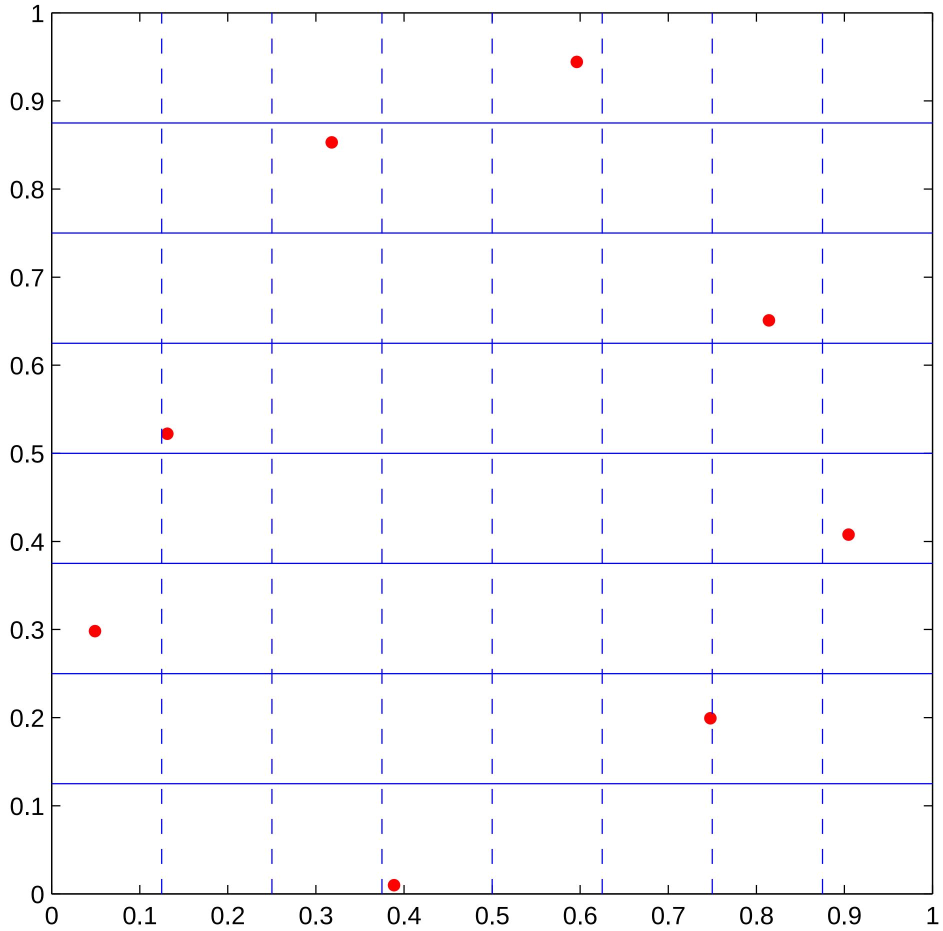


Figure 1: A LHS with eight runs

Advantages of LHS

- ♠ Easy to generate
- ♠ Can work for high dimensional cases

Disadvantages of LHS

- ♠ some designs have a poor quality
- ♠ there is a large space to reduce variance of $\bar{y}(\mathcal{P})$
- ♠ the convergence rate is slow, $O_p(\frac{1}{\sqrt{n}})$

Modified versions of LHS

- centered (midpoint) Latin hypercube sampling
- randomized orthogonal array
- symmetric and orthogonal column Latin hypercubes
- optimal Latin hypercube designs in the following criteria:
  - mean squared error
  - entropy
  - minimax and maximin distance
  - ϕ_p-criterion

Uniform Designs

The uniform design put experimental points uniformly scattered on the experimental domain.

The Koksma-Hlawka inequality provides an upper bound of

$$\text{diff-mean} = |E(y) - \bar{y}(y)| \leq V(g)D(\mathcal{P}),$$

where $D(\mathcal{P})$ is the star discrepancy of the design $\mathcal{P}$, not depending on g, and $V(g)$ is the total variation of the function g in the sense of Hardy and Krause.

The K-H inequality still holds when the star discrepancy is replaced by the centered L_2-, wrap-around L_2- or discrete discrepancy that will be introduced late.

- For construction of uniform designs it is not tractable to find a set of n points, $\mathcal{P} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subset C^s$, such that it has the minimum discrepancy.

- A *U-type design* with n runs and s factors each having respective $q_1, \cdots, q_s$ levels is an $n \times s$ matrix such that the q_j levels in the jth column appear equally often and is denoted by $U(n, q_1 \times \cdots \times q_s)$. When some q_j's are equal, we denote it by $U(n, q^{r_1} \times \cdots \times q_m^{r_m})$ where integers $r_1 + \cdots + r_m = s$. If all the q_j's are equal, denoted by $U(n, q^s)$, it is said to be symmetrical, otherwise asymmetrical.

- Denote by $\mathcal{U}_n(q_1 \times \cdots \times q_s)$ all the $U(n, q_1 \times \cdots \times q_s)$. We want to find one with minimum discrepancy, a UD, denoted by $U_n(q_1 \times \cdots \times q_s)$.

Table 1: $U_{12}(12^4)$

No	1	2	3	4
1	1	10	4	7
2	2	5	11	3
3	3	1	7	9
4	4	6	1	5
5	5	11	10	11
6	6	9	8	1
7	7	4	5	12
8	8	2	3	2
9	9	7	12	8
10	10	12	6	4
11	11	8	2	10
12	12	3	9	6

Table 2: $U_6(3^2 \times 2)$

No	1	2	3
1	1	1	1
2	2	1	2
3	3	2	1
4	1	2	2
5	2	3	1
6	3	3	2

The following remarks are useful for the use of uniform design tables:

Remark 1 Obviously, each number of levels, $q_j (j = 1, \ldots, s)$, should be a divisor of n from the definition of the U-type design.

Remark 2 For a given (n, q, s), the corresponding uniform design is not unique.

A number of UD tables can be found on the UD-website at

http://www.math.hkbu.edu.hk/UniformDesign

Common aspects between LHS and UD

- ♡ space-filling designs for
  - ♯ computer experiments
  - ♯ physical experiments with model uncertainty
- ♡ originally motivated by the overall mean model
- ♡ robust against model specification
- ♡ construction of the design is based on U-type designs

3. An Example

A chemical experiment in pharmaceutics is conducted in order to find the best setup to increase the yield.

Four factors and response are

the amount of formaldehyde (x_1),

the reaction temperature (x_2),

the reaction time (x_3),

the amount of potassium carbolic acid (x_4)

the yield (y)

are under consideration. The experimental domain is chosen to be $\mathcal{X} = [1.0, 5.4] \times [5, 60] \times [1.0, 6.5] \times [15, 70]$ and each factor takes 12 levels in this domain.

A. Design

Table 3: Factors and Levels

Factor	Unit	Level
x_1, the amount of formaldehyde	mol	1.0, 1.4, 1.8, 2.2, 2.6, 3.0, 3.4, 3.8, 4.2, 4.6, 5.0, 5.4
x_2, the reaction temperature	hour	5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60
x_3, the reaction time	hour	1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5
x_4, the amount of potassium carbolic acid	ml	15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70

Table 4: $U_{12}(12^4)$ and related design

No. of runs	1	2	3	4	x_1	x_2	x_3	x_4	y
5	1	10	4	7	1.0	50	2.5	45	0.0795
6	2	5	11	3	1.4	25	6.0	25	0.0118
10	3	1	7	9	1.8	5	4.0	55	0.0109
7	4	6	1	5	2.2	30	1.0	35	0.0991
11	5	11	10	11	2.6	55	5.5	65	0.1266
9	6	9	8	1	3.0	45	4.5	15	0.0717
8	7	4	5	12	3.4	20	3.0	70	0.1319
3	8	2	3	2	3.8	10	2.0	20	0.0900
2	9	7	12	8	4.2	35	6.5	50	0.1739
4	10	12	6	4	4.6	60	3.5	30	0.1176
1	11	8	2	10	5.0	40	1.5	60	**0.1836**
12	12	3	9	6	5.4	15	5.0	40	0.1424

B. Modeling

For experiments with more factors it is popular to consider a set of basis of functions, $\{B_0(\boldsymbol{x}), B_1(\boldsymbol{x}), \ldots\}$ and a maximal model of interest,

$$\hat{g}(\boldsymbol{x}) = \beta_1 B_1(\boldsymbol{x}) + \beta_2 B_2(\boldsymbol{x}) + \ldots, +\varepsilon. \tag{1}$$

Then by techniques for variable selection one of submodel of (1) will be used as a metamdodel. In this example, we consider only linear and quadratic regression models. At first let us try the first-order regression model of the form

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

as it is simple. Based on the data we obtained it results in

$$\hat{y} = -0.0533 + 0.0281x_1 + 0.0010x_2 - 0.0035x_3 + 0.0011x_4. \qquad (2)$$

Its ANOVA table is shown in Table on the next page

Table 5: ANOVA table for model (2)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	4	0.0274	0.0069	8.2973	0.0086
Error	7	0.0058	0.0008		
C Total	11	0.0332			

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
X1	1	0.0180	0.0180	21.8021	0.0023
X2	1	0.0033	0.0033	3.9496	0.0872
X3	1	0.0004	0.0004	0.5150	0.4962
X4	1	0.0046	0.0046	5.6248	0.0495

From the ANOVA table, we find that the model (2) involves an insignificant term 'x_3' with p-value 0.4962. We have to remove this term from the model. By the backward elimination techniques in regression analysis, the resulting model turns out to be

$$\hat{y} = 0.0107 + 0.0289x_1$$

with $R^2 = 57.68\%$ and $s^2 = 0.0014$. This model is not consistent with experience of the experimenter as there are three factors not to be involved in the model.Therefore, a more flexible second-order quadratic regression of the form

$$E(y) = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i \leq j} \beta_{ij} x_i x_j. \tag{3}$$

is considered.

♡ Here the number of unknown parameters is greater than the number of runs and this model is inestimable.

♡ However, this model provide a base and some submodel may fit the purpose of the experiment well. Model (3) is a maximal model of interest.

♡ The remaining study is going to find a submodel of (3) as a metamodel. With MAXR, a technique of selection of variables, we find a good subset model to be

$$\begin{aligned} \hat{y} &= 0.0446+0.0029x_2-0.0260x_3+0.0071x_1x_3 \\ &\quad +0.000036x_2x_4-0.000054x_2^2 \end{aligned} \tag{4}$$

with $R^2 = 97.43\%$ and $s^2 = 0.0001$. The corresponding ANOVA table is shown in Table on the next page.

Table 6: ANOVA table for model (4)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	5	0.0323	0.0065	45.5461	0.0001
Error	6	0.0009	0.0001		
C Total	11	0.0332			

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
X2	1	0.0014	0.0014	10.1949	0.0188
X3	1	0.0125	0.0125	88.0883	0.0001
X1X3	1	0.0193	0.0193	135.5636	0.0001
X2X2	1	0.0024	0.0024	16.8276	0.0063
X2X4	1	0.0062	0.0062	43.6923	0.0006

In the literature, the centered second-order centered quadratic regression model of the form

$$E(y) \quad = \quad \beta_0 + \sum_{i=1}^{4} \beta_i (x_i - \bar{x}_i) + \sum_{i \leq j} \beta_{ij} (x_i - \bar{x}_i)(x_j - \bar{x}_j), \qquad (5)$$

is also suggested for a maximal model of interest, where $\bar{x}_i$ is the sample mean of x_i. In this data set, $\bar{x}_1 = 3.2, \bar{x}_2 = 32.5, \bar{x}_3 = 3.75$ and $\bar{x}_4 = 42.5$. Once again, by using model selection techniques, the final model is

$$\begin{aligned} \hat{y} \quad = \quad & 0.128 + 0.028(x_1 - 3.2) + 0.00094(x_2 - 32.5) + 0.00114(x_4 - 42.5) \\ & + 0.00058(x_3 - 3.75)(x_4 - 42.5) - 0.000082(x_2 - 32.5)^2 \qquad (6) \end{aligned}$$

with $R^2 = 97.05\%$ and $s^2 = 0.0002$. The corresponding ANOVA table is given in Table on the next page.

Table 7: ANOVA table for model (6)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
Model	5	0.0322	0.0064	39.5466	0.0002
Error	6	0.0010	0.0002		
C Total	11	0.0332			

Type III Tests					
Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
X1	1	0.0180	0.0180	110.7654	0.0001
X2	1	0.0031	0.0031	19.2816	0.0046
X4	1	0.0046	0.0046	28.5141	0.0018
X3X4	1	0.0023	0.0023	13.8349	0.0099
X2X2	1	0.0047	0.0047	29.0842	0.0017

C. Maximizing the Response

♡ The best result among the 12 responses is $y_1 = 18.36\%$ at $x_1 = 5.0, x_2 = 40$, $x_3 = 1.5$ and $x_4 = 60$. This can be served as a benchmark.

♡ Maximize y with respect to $x_i, i = 1, \ldots, 4$ under models (4) or (6) respectively over the domain $\mathcal{X}$, that is to find $x_i^*, i = 1, \ldots, 4$ such that

$$\hat{y}(x_1^*, x_2^*, x_3^*, x_4^*) = \max_{\mathcal{X}} \hat{y}(x_1, x_2, x_3, x_4),$$

where $\hat{y}(x_1, x_2, x_3, x_4)$ is given by (4) or (6) respectively.

♡ By some optimization algorithm, it is easily found that

♯ under model (4), $x_1^* = 5.4, x_2^* = 50.2, x_3^* = 1$, $x_4^* = 70$ and the corresponding response $\hat{y}(5.4, 50.2, 1, 70) = 19.3\%$ is the maximum;

♯ under model (6), $x_1^* = 5.4, x_2^* = 43.9, x_3^* = 6.5$, $x_4^* = 70$ and the corresponding response $\hat{y}(5.4, 43.9, 6.5, 70) = 26.5\%$ is the maximum.

♡ We need some additional experiment to judge which model is closed to the real one.

D. Some Modeling Techniques

♯ The basis function method

$$\hat{g}(\boldsymbol{x}) = \beta_1 B_1(\boldsymbol{x}) + \beta_2 B_2(\boldsymbol{x}) + \ldots,$$

where $B_j(\boldsymbol{x})$ can be polynomials, Legendre polynomials, smoothing spline, regression spline, wavelets, Fourier basis, etc.

♯ Kriging method (an interpolation method)

proposed by a South African geologist, D.G. Krige in his Master thesis in 1951.

The Kriging approach was systematically introduced to model computer experiments by Mitchell, Sacks, Welch and Wynn (1989).

♯ Bayesian Approach

See Currin, Mitchell, Morris and Ylvisaker (1991)

♯ Neural Network

♯ Radial basis functions

♯ Local Polynomial Regression

The reader can find the details in Fang, Li and Sudjianto (2005).

The following methods are useful in modeling or/and selection of variables

- ♡ principal component regression
- ♡ partial regression
- ♡ ridge regression
- ♡ robust regression
- ♡ project pursuit regression
- ♡ regression tree
- ♡ LASSO (least absolute shrinkage selection operator)
- ♡ SCAD (smoothly clipped absolute deviation)

4. New Development of The Uniform Design

Difficulty and complexity

† The uniformity is a geometrical criterion, it needs some justification in statistical sense

† Initially, the uniform design theory is based on the quasi-Monte Carlo methods. The useful tool is the number theory. Most statisticians are lack of knowledge of the number theory.

† Construction of uniform design is a NP hard problem. It needs some powerful algorithms in optimization.

† It needs several strong working teams.

4.1. Characteristics of The Uniform Design

From the statistical point of view we want the uniform design is optimal in a certain sense.

† Wiens, D.P.(1991), *Stat. & Prob. Letters*, concern with designs for approximately linear regression models and show that the **uniform design measure** (uniform design for short) is maximin in the sense of maximizing the minimum bias in the regression estimate of σ^2 and is also minimax in the sense of minimizing the maximum bias in the regression estimate of σ^2.

† Xie, M.Y. and Fang, K.T.(2000), *JSPI*, pointed out the uniform measure is admissible and minimax under the model

$$y = g(x_1, \cdots, x_s) + \epsilon,$$

where g is unknown.

† Hickernell, F.J.(1999), *Stat. & Prob. Letters*, considered robust regression models

$$y = g(\boldsymbol{x}) + \varepsilon = \mu + h(\boldsymbol{x}) + \varepsilon,$$

where the function $g(\boldsymbol{x})$ is decomposed into the overall mean value of $g(\boldsymbol{x})$ and mis-specification $h(\boldsymbol{x})$. He proposed two models

‡ average mean-square-error model

‡ maximum mean-square-error model

With a certain condition he proved that the uniform design is optimal under these models.

† Hickernell and Yue (1999), *Statistica Sinica*, considered an approximate model and split the sum squares into variance part and model bias part. They investigated the importance of the variance part and bias part and showed that the variance-minimizing designs can yield substantial bias, whereas bias-minimizing designs are rather efficient. Moreover, bias-minimizing designs tend to spread the points evenly over the domain. This gives another justification of the uniform design.

† Hickernell and Liu (2002), *Biometrika*, They pointed out "When fitting a linear regression model to data, aliasing can adversely affect the estimates of the model coefficients and the decision of whether or not a term is significant. Optimal experimental designs give efficient estimators assuming that the true form of the model is known, while robust experimental designs guard against inaccurate estimates caused by model misspecfication. Although it is rare for a single design to be both maximally **efficient and robust**, it is shown here that uniform designs limit the effects of aliasing to yield reasonable efficiency and robustness together."

4.2. Measure of Uniformity

The star L_p-discrepancy has been widely used in number-theoretic methods and is defined by

$$D_p(\mathcal{P})^p = \int_{C^s} \left| \frac{N(\mathcal{P}, [\mathbf{0}, \boldsymbol{x}))}{n} - \mathrm{Vol}([\mathbf{0}, \boldsymbol{x})) \right|^p d\boldsymbol{x}$$

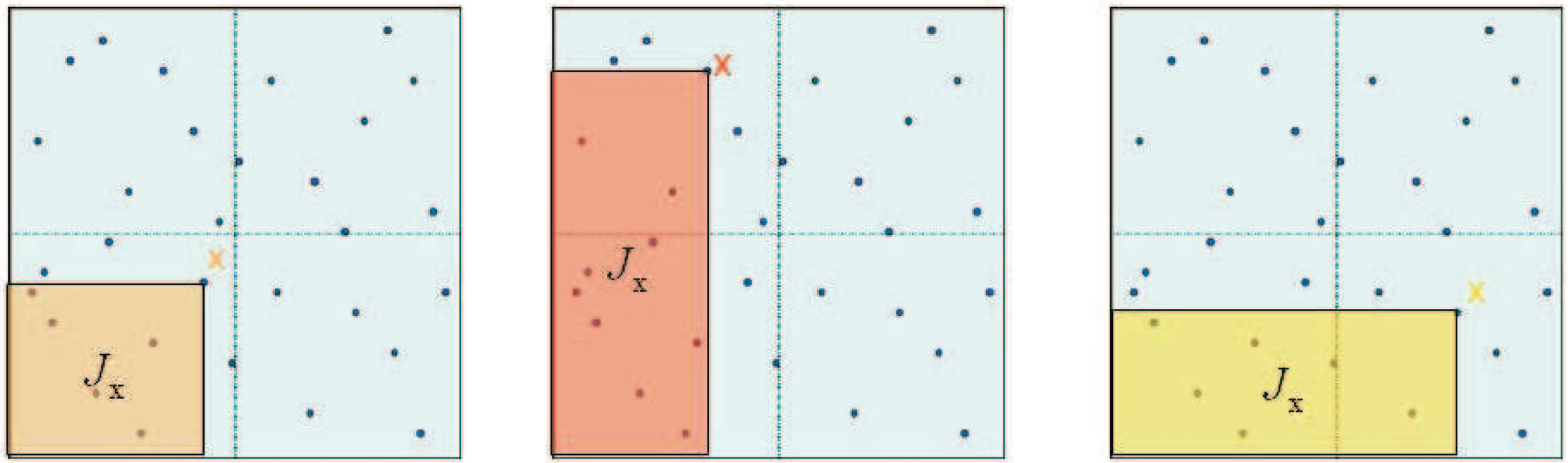
that is the most popular measure from quasi-Monte Carlo methods, where

$[\mathbf{0}, \boldsymbol{x}) = [0, x_1) \times \cdots \times [0, x_s)$,

$N(\mathcal{P}, [\mathbf{0}, \boldsymbol{x}))$ the number of points of $\mathcal{P}$ falling in $[\mathbf{0}, \boldsymbol{x})$,

$\mathrm{Vol}(A)$ the volume of A.

The Discrepancy at $[0, x]$



Ratio = 5/30
= 0.167
Volume = 0.1615
Discrepancy=.0055

Ratio = 9/30
= 0.3
Volume = 0.2875
Discrepancy=.0125

Ratio = 6/30
= 0.2
Volume = 0.2438
Discrepancy=.0438

- The L_∞-discrepancy $D(\mathcal{P})$ is called star discrepancy that can be expressed by

$$D(\mathcal{P}) = \max_{\boldsymbol{x} \in C^s} |F_n(\boldsymbol{x}) - F(\boldsymbol{x})|,$$

where $F_n(\boldsymbol{x})$ is the empirical distribution of $\mathcal{P}$ and $F(\boldsymbol{x})$ is the uniform distribution on C^s.

- The star discrepancy just is the Kolmogorov-Smirnov statistic in goodness-of-fit test.

- The star discrepancy is not easy to compute.

- The star discrepancy is not sensitive enough for construction of uniform designs.

- The L_2-discrepancy is much easier to calculate numerically

$$\begin{aligned}(D_2(\mathcal{P}))^2 &= 3^{-s} - \frac{2^{1-s}}{n}\sum_{k=1}^{n}\prod_{l=1}^{s}(1-x_{kl}^2) \\ &+\frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{s}[1-\max(x_{ki},x_{ji})].\end{aligned}$$

- It ignores differences

$$\left|\frac{N(\mathcal{P},[\mathbf{0},\boldsymbol{x}))}{n} - \mathrm{Vol}([\mathbf{0},\boldsymbol{x}))\right|^2$$

in any low-dimensional subspace.

- star L_p-discrepancy is not invariant under the coordinates rotation

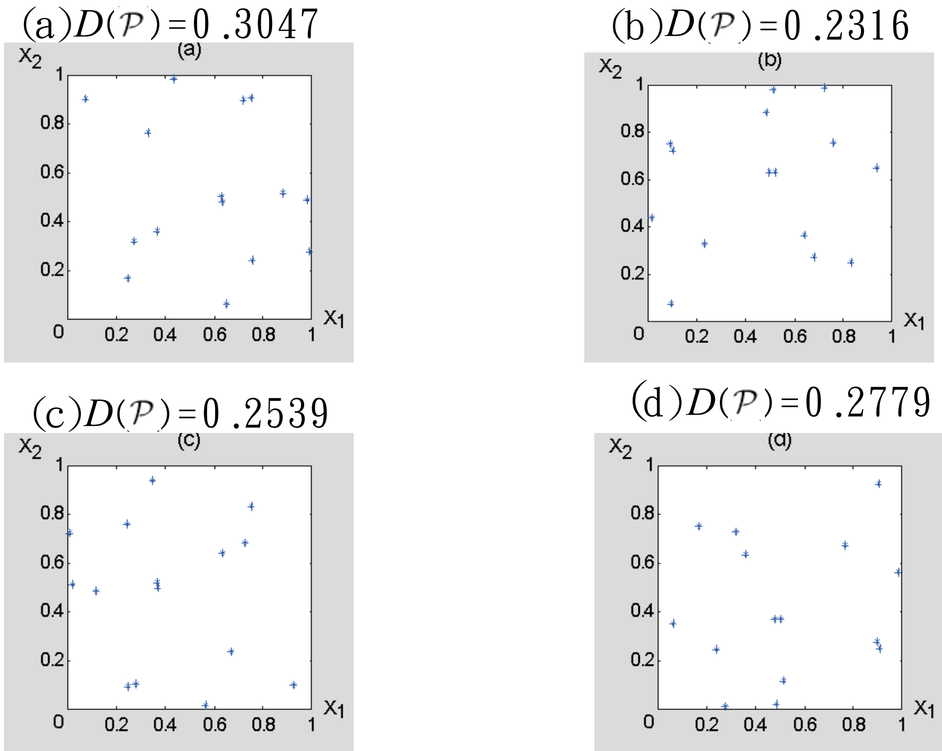


Figure 2: Star discrepancy values in rotations

- We need some new measures of uniformity satisfying:
  - [C1] Invariant under permuting factors and/or runs
  - [C2] It is invariant under the coordinates rotation
  - [C3] It can measure not only uniformity of $\mathcal{P}$ over C^s, but also projection uniformity of $\mathcal{P}$ over C^u, where u is a non-empty subset of $\{1, \cdots, s\}$.
  - [C4] There have some geometric meaning
  - [C5] Easy to compute
  - [C6] Satisfy Koksma-Hlawka-like inequality
  - [C7] It consists with other criteria in experimental designs

A new modified L_2-discrepancy proposed by Hickernell (1998) is defined by

$$(D_2(\mathcal{P}))^p = \sum_{u \neq \emptyset} \int_{C^u} \left| \frac{N(\mathcal{P}_u, J_{\boldsymbol{x}_u})}{n} - \mathrm{Vol}(J_{\boldsymbol{x}_u}) \right|^p d\boldsymbol{x}_u,$$

where the summation takes over all possible marginal dimension of R^s,

- u: a non-empty subset of the set of coordinate indices

  $S = \{1, \cdots, s\}$;
- $|u|$: the cardinality of u;
- C^u: the $|u|$-dimensional unit cube involving the coordinates in u;
- $\mathcal{P}_u$: the projection of $\mathcal{P}$ on C^u;

- $\boldsymbol{x}_u$: the projection of $\boldsymbol{x}$ into C^u;
- $J_{\boldsymbol{x}}$: a rectangle uniquely determined by $\boldsymbol{x}$ and is chosen with some geometric consideration; and
- $J_{\boldsymbol{x}_u}$: the projection of $J_{\boldsymbol{x}}$ on C^u.

Hickernell also proposed several definitions of $J(\boldsymbol{x})$ from the geometric consideration:

- symmetric discrepancy (SD)
- centered discrepancy (CD)
- wrap-around discrepancy (WD)
- unanchored discrepancy

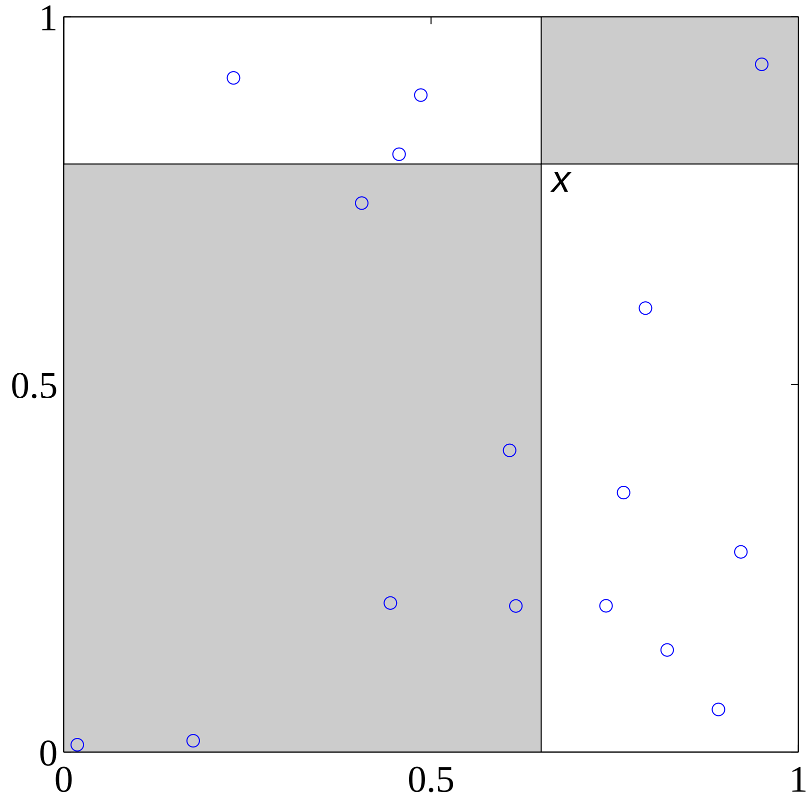


Figure 3: Symmetric discrepancy

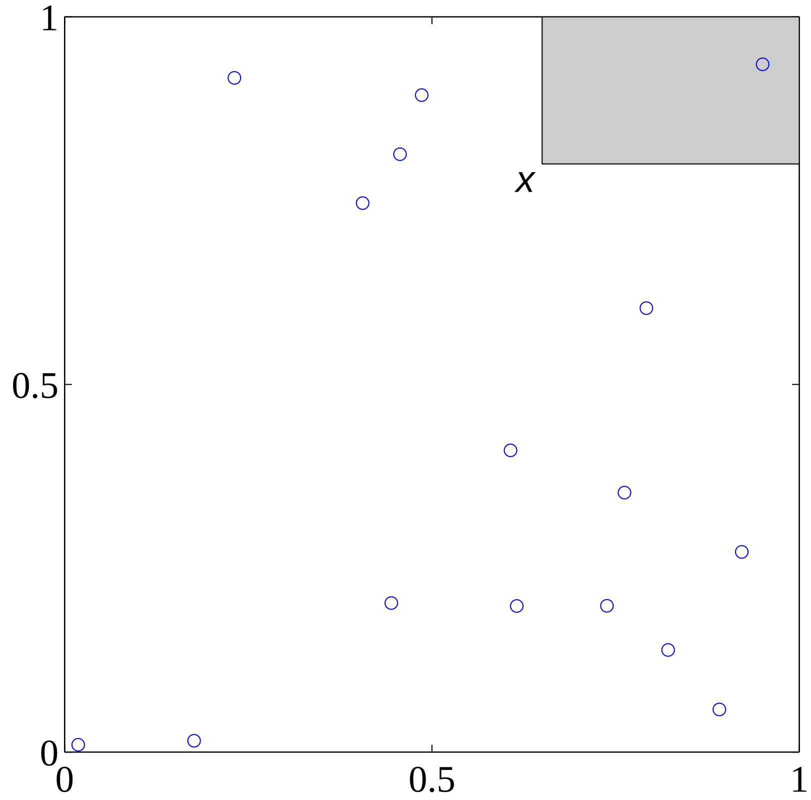


Figure 4: Centered discrepancy

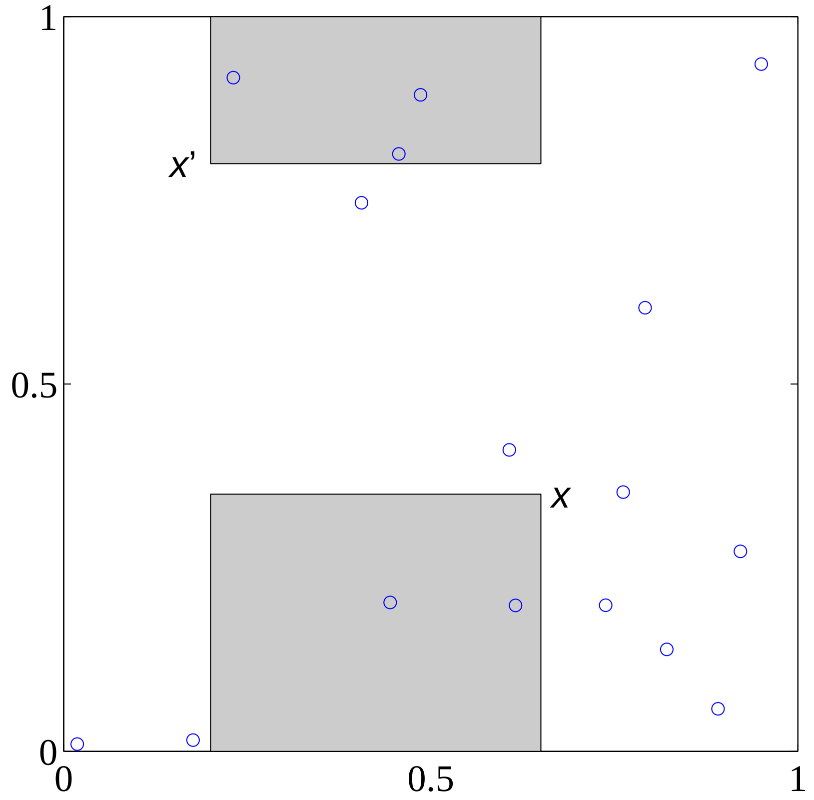


Figure 5: Wrap-around discrepancy

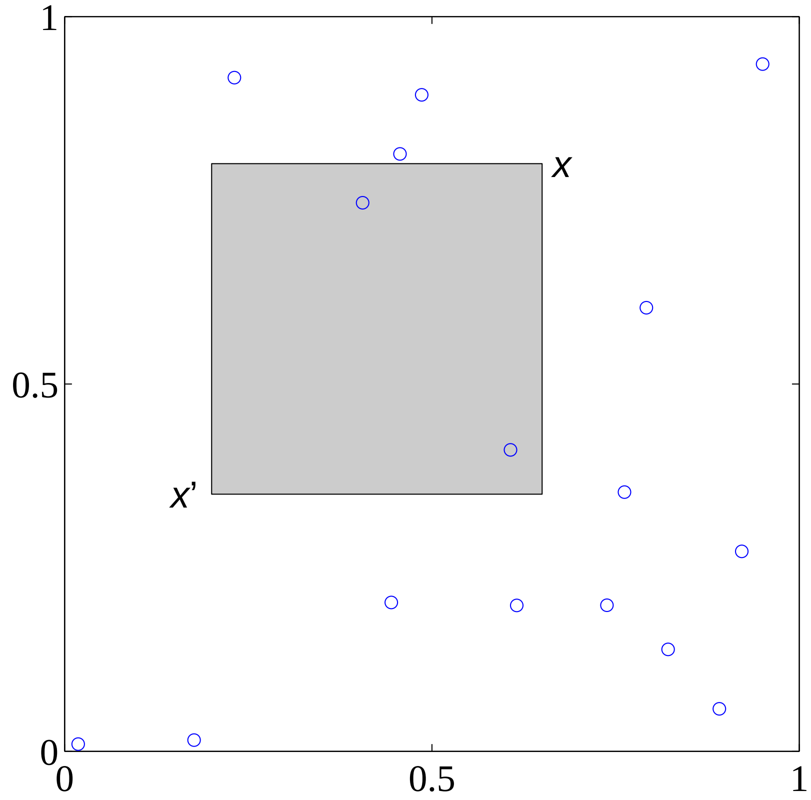


Figure 6: Unanchored discrepancy

A unified definition of discrepancy

Hickernell employed the reproducing kernel of Hilbert space to give a more general definition for the discrepancy.

Let $\mathcal{X}$ be the experimental domain that is a measurable set of R^s and $K(\boldsymbol{x}, \boldsymbol{w})$ be a symmetric and positive definite function on $\mathcal{X} \times \mathcal{X}$, i.e.,

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{w}) &= K(\boldsymbol{w}, \boldsymbol{x}), \text{ for any } \boldsymbol{x}, \boldsymbol{w} \in \mathcal{X}, \\ \sum_{i,j=1}^{n} a_i a_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) &> 0, \text{ for any } a_i \in R, \boldsymbol{x}_i \in \mathcal{X}, i = 1, \cdots, n, \end{aligned}$$

Let F be the uniform distribution on $\mathcal{X}$ and $F_n(\boldsymbol{x})$ be the empirical distribution of $\mathcal{P}$. The L_2-discrepancy is defined by

$$D_K^2(\mathcal{P}) = ||F - F_n||_K^2 = \int_{\mathcal{X}\times\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{w}) d(F - F_n)(\boldsymbol{x}) d(F - F_n)(\boldsymbol{w}).$$

It can be expressed

$$\begin{aligned} D_K^2(\mathcal{P}) &= \int_{\mathcal{X}\times\mathcal{X}} K(\boldsymbol{x}, \boldsymbol{w}) dF(\boldsymbol{x}) dF(\boldsymbol{w}) \\ &\quad - \frac{2}{n} \sum_{i=1}^{n} K(\boldsymbol{x}, \boldsymbol{x}_i) dF(\boldsymbol{x}) + \frac{1}{n^2} \sum_{i,j=1}^{n} K(\boldsymbol{x}_i, \boldsymbol{x}_j), \end{aligned}$$

Different kernel K's imply different measures of uniformity. For example, the CD and WD have the respective kernel

$$
\begin{aligned}
K_c(\boldsymbol{x}, \boldsymbol{w}) &= \prod_{i=1}^{s}[1 + \frac{1}{2}|x_i - 0.5| + \frac{1}{2}|w_i - 0.5| - \frac{1}{2}|x_i - w_i|], \\
K_w(\boldsymbol{x}, \boldsymbol{w})) &= \prod_{i=1}^{s}[\frac{2}{3} - |x_i - w_i| + |x_i - w_i|^2],
\end{aligned}
$$

where $\boldsymbol{x} = (x_1, \cdots, x_s)$ and $\boldsymbol{w} = (w_1, \cdots, w_s)$.

All the L_2-discrepancies proposed by Hickernell has an analytic formula for computing. For example,

$$(CD_2(\mathcal{P}))^2 = \left(\frac{13}{12}\right)^s - \frac{2}{n}\sum_{k=1}^{n}\prod_{j=1}^{s}(1+\frac{1}{2}|x_{kj}-0.5|-\frac{1}{2}|x_{kj}-0.5|^2)$$
$$+\frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{s}\left[1+\frac{1}{2}|x_{ki}-0.5|+\frac{1}{2}|x_{ji}-0.5|-\frac{1}{2}|x_{ki}-x_{ji}|\right].$$

$$(WD_2(\mathcal{P}))^2 = \left(\frac{4}{3}\right)^s + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{s}\left[\frac{3}{2}-|x_{ki}-x_{ji}|(1-|x_{ki}-x_{ji}|)\right].$$

Advantages of Hickernell's Approach:

- The new discrepancy can satisfy [C1]∼[C5].
- It is easy to prove each discrepancy satisfying the Kpksma-Hlawka inequality by the Cauchy-Schwartz's inequality.
- It can easily define new discrepancies. For example, discrete discrepancy proposed by Hickernell and Liu (2002) and Fang, Lin and Liu (2003).

Let $\mathcal{X}_j = \{\frac{1}{2q_j}, \frac{3}{2q_j}, \cdots, \frac{2q_j-1}{2q_j}\}$ and $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_s$. Take the kernel

$$K(\boldsymbol{x}, \boldsymbol{w}) = \prod_{j=1}^{s} k_j(x_j, w_j), \text{ for any } \boldsymbol{x}, \boldsymbol{w} \in \mathcal{X}, \tag{7}$$

where

$$k_j(x_j, w_j) = \begin{cases} a & \text{if } x_j = w_j, \\ b & \text{if } x_j \neq w_j, \end{cases} \quad a > b > 0.$$

The discrepancy has the following formula

$$D(\mathcal{P}; K) = \left\{ -\prod_{j=1}^{s} \left[\frac{a + (q_j - 1)b}{q_j} \right] + \frac{1}{n^2} \sum_{i,j=1}^{n} \prod_{k=1}^{s} \left[a^{\delta_{z_{ik} z_{jk}}} b^{1-\delta_{z_{ik} z_{jk}}} \right] \right\}^{1/2},$$

4.3. Construction of Uniform Designs

- For one-factor experiment, the unique UD on [0,1] is

$$\left\{\frac{1}{2n}, \frac{3}{2n}, \cdots, \frac{2n-1}{2n}\right\}$$

with $CD_2^2 = \frac{1}{12n^2}$.

To finding a $U_n(q^s)$ with $s > 2$ is a very difficult problem. There are two approaches:

† theoretic approach

† numerical search

A. Construction of UDs by resolvable balanced incomplete block designs

A balanced incomplete block design (BIBD) with parameters (n, s, m, t, λ), denoted by BIBD(n, s, m, t, λ), is an arrangement of n treatments into s blocks of size t, where $t < n$, such that each treatment appears in m blocks, and every pair of treatments appears in exactly λ blocks.

A BIBD(n, s, m, t, λ) is said to be *resolvable*, denoted by RBIBD(n, s, m, t, λ), if its blocks can be partitioned into m sets of blocks, called *parallel classes*, such that every treatment appears in each parallel class precisely once.

Table 8: An RBIBD$(10, 45, 9, 2, 1)$

	$\boldsymbol{P}_1$	$\boldsymbol{P}_2$	$\boldsymbol{P}_3$	$\boldsymbol{P}_4$	$\boldsymbol{P}_5$	$\boldsymbol{P}_6$	$\boldsymbol{P}_7$	$\boldsymbol{P}_8$	$\boldsymbol{P}_9$
b_1^j	{1,10}	{2,10}	{4,9}	{3,7}	{2,8}	{5,7}	{5,6}	{1,7}	{1,6}
b_2^j	{8,9}	{5,8}	{3,10}	{4,10}	{6,9}	{2,4}	{3,4}	{2,5}	{2,7}
b_3^j	{4,5}	{3,6}	{7,8}	{1,2}	{5,10}	{1,9}	{1,8}	{4,6}	{4,8}
b_4^j	{6,7}	{7,9}	{2,6}	{5,9}	{1,3}	{3,8}	{7,10}	{3,9}	{3,5}
b_5^j	{2,3}	{1,4}	{1,5}	{6,8}	{4,7}	{6,10}	{2,9}	{8,10}	{9,10}

There are 10 runs, 45 blocks of size 2, each treatment appears in 9 blocks, every pair of treatments appears in exactly one block.

Table 9: $U(10; 5^9)$

Row	1	2	3	4	5	6	7	8	9
1	1	5	5	3	4	3	3	1	1
2	5	1	4	3	1	2	5	2	2
3	5	3	2	1	4	4	2	4	4
4	3	5	1	2	5	2	2	3	3
5	3	2	5	4	3	1	1	2	4
6	4	3	4	5	2	5	1	3	1
7	4	4	3	1	5	1	4	1	2
8	2	2	3	5	1	4	3	5	3
9	2	4	1	4	2	3	5	4	5
10	1	1	2	2	3	5	4	5	5

There are many ways to construct RBIBDs.

Theorem

The following $U_n(q^s)$ can be constructed by the Algorithm RBIBD-UD, where the categorical discrepancy is employed as the measure of uniformity:

(a) If n is even, then a uniform $U_n((\frac{n}{2})^{l(n-1)})$ exists, where l is a positive integer.

(b) If $n \equiv 3\ (mod\ 6)$, then a uniform $U_n((\frac{n}{3})^{\frac{n-1}{2}})$ exists.

(c) If $n \equiv 0\ (mod\ 3)$ and $n \neq 6$, then a uniform $U_n((\frac{n}{3})^{n-1})$ exists.

(d) If $n \equiv 4\ (mod\ 12)$, then a uniform $U_n((\frac{n}{4})^{\frac{n-1}{3}})$ exists.

(e) If $n \equiv 0\ (mod\ 4)$, then a uniform $U_n((\frac{n}{4})^{n-1})$ exists.

(f) If $n \equiv 0\,(mod\,6)$, then a uniform $U_n((\frac{n}{6})^{n-1})$ exists except possibly $n \in \{174, 240\}$.

(g) If $n \equiv 0\ (mod\ 6)$, then a uniform $U_n((\frac{n}{6})^{2(n-1)})$ exists.

(h) If $n \equiv 5\ (mod\ 20)$, then a uniform $U_n((\frac{n}{5})^{\frac{n-1}{4}})$ exists except possibly $n \in \{45, 225, 345, 465, 645\}$.

(i) If $n \equiv 5\ (mod\ 10)$ and $n \neq 15$, then a uniform $U_n((\frac{n}{5})^{\frac{n-1}{2}})$ exists except possibly
$n \in \{45, 115, 135, 195, 215, 225, 235, 295, 315, 335, 345, 395\}$.

(j) If $n \equiv 0\ (mod\ 5)$ and $n \neq 10, 15$, then a uniform $U_n((\frac{n}{5})^{n-1})$ exists except possibly $n \in \{70, 90, 135, 160, 190, 195\}$.

A lot of contributions on this direction:

‡ Lu, X and Meng, Y. (2000), *JSPI*

‡ Fang,K.T. Ge, G. and Liu, M.Q. (2002), *CSA Bulletin*

‡ Fang,K.T. Ge, G. and Liu, M.Q. (2002, 2003), *Science in China*

‡ Qin, H. (2002), *Ph.D thesis*

‡ Fang,K.T. Ge, G., Liu, M.Q. and Qin, H. (2004), *Discrete Math.*

‡ Chatterjee, K., Fang, K.T. and Qin, H. (2005), *JSPI*

‡ Fang, K.T., Lu, X., Tang, Y. and Yin, J. (2004), *Discrete Math.*

B. Construction of UDs by Optimization & other techniques

♢ Good lattice point method, Wang Y. and Fang, K.T. (1980, 1981), Fang and Li (1994)

♢ Latin square method, Fang, K.T. Shiu, W.C. and Pan, J.X. (1999), *Statistica Sinica*)

♢ Cutting method, Ma, C.X. and Fang, K.T. (2004), *International Journal of Materials and Product Technology*

♢ Collapsing method, Fang and Qin (2003), *Statist. & Prob. Letters*

To finding a $U_n(q^s)$ with $s > 2$ is a NP hard problem in the sense of computation complexity.

◇ Optimization

- ‡ Local search algorithm
- ‡ simulated annealing algorithm
- ‡ stochastic evolutionary algorithm
- ‡ threshold accepting algorithm
  - ♯ Winker and Fang (1997), *SIAM J. Numer. Anal.*
  - ♯ Fang and Ma (2001), *J. Complexity*
  - ♯ Fang, Ma and Winker (2002), *Math. Computation*
  - ♯ Fang, Lu and Winker (2003), *J. Complexity*

C. Lower bounds

For a given discrepancy D and the domain $\mathcal{U}(n, q^s)$, if we could find a strict lower bound of D, it will be helpful in searching UDs.

$\diamondsuit$ Fang and Mukerjee (2000, *Biometrika*) found a lower bound of CD_2 for $q = 2$

$\diamondsuit$ Fang, Lin and Liu (2003, *Metrika*) found a lower bound of the discrete discrepancy

$\diamondsuit$ Fang, Lu and Winker (2003, *J. Complexity*) found a lower bound of WD_2 for $q = 2, 3$ and of CD_2 for $q = 2$.

$\diamondsuit$ Fang, Tang and Yin (2004), found a lower bound of WD_2 for all the cases

$\diamondsuit$ Fang, Maringer, Tang and Winker (2004), found a lower bound of CD_2 for $q = 3, 4$.

$\diamondsuit$ Zhang, Fang, Li and Sudjianto (2005), *Annals of Statistics*) given a unified approach to construction of lower bounds for various criteria in factorial and supersaturated design. We use majorization theory that can treat all the cases in a unified way.

5. Usefulness of uniformity in experimental designs

5.1. Uniformity and Orthogonality

- Fang and Winker (1998) found that many existing orthogonal designs are also uniform designs under CD_2. They proposed a conjecture that each orthogonal design is also a uniform design under a certain discrepancy.

- Ma, Fang and Lin (2002, *JSPI*) proved that the conjecture is true for full design with odd number of levels or two level design if CD_2 or WD_2 is chosen as the criterion.

5.2. Uniformity and Aberration

- For a factorial design, $\mathcal{P}$, of n runs with s q-level factors, the word-length pattern, $(A_1(\mathcal{P}), \cdots, A_s(\mathcal{P}))$, gives a rich information about confounding situation of $\mathcal{P}$.

- Resolution and the minimum aberration are two popular criteria for comparing fractional designs.

- Fang and Mukerjee (2000, *Biometrika*) found an analytic relationship between CD_2 and word-length pattern for a 2-level factorial design.

- Fang and Ma (2001, *MCQMC*) and Fang, Ma and Mukerjee (2001, *MCQMC*) extended Fang-Mukerjee's result to a 3-level design.

- Ma and Fang (2001, *Metrika*) proposed the generalized word-length pattern and give the connection with CD_2 and WS_2.

- Hickernell and Liu (2002, *Biometrika*) proposed the projection discrepancy pattern or generalized word-length pattern. They indicated that the uniform design limit aliasing.

- Fang and Qin (2004, *Science in China*) proposed the uniformity pattern and related criteria for 2-level designs.

5.3. Uniformity in supersaturated designs

- Whenever the run size of a design is insufficient for estimating all the main effects, the design is called supersaturated.

Lin (1993, 1995) *Technometrics*,

Yamada, Ikebe, Hashiguchi and Niki (1999), *JSPI*

- There are many criteria for comparing supersaturated designs, such as
  - $E(s^2)$ (Booth and Cox, 1962),
  - $\mathrm{Ave}(\chi^2)$ (Yamada and Lin, 1999, *Stat. & Prob. Letters*)
  - $\mathrm{Ave}(f^2)$ (Fang, Lin and Ma, 2000, *JSPI*)

- However, the discrepancy can be used for comparing supersaturated designs,

– Liu and Hickernell (2002, *Statistica Sinica*) proposed a discrete discrepancy and found its lower bound and find link with $E(s^2)$

– Fang, Ge, Liu (2002, *Science in China* & *Calcutta Statistical Association Bulletin*)

– Fang, Lin and Liu (2003, *Metrika*)

– Fang, Ge, Liu and Qin (2004, *Discrete Math.*)

– Qin and Fang (2004, *Metrika*)

– Zhang, Fang, Li and Sudjianto (2005, *The Annals of Stat.*)

5.4. Uniformity and Isomorphism

- $d(n, q, s)$: a factorial design of n runs and s factors each having q levels.

- Two designs are called to be isomorphic if one can be obtained from another by exchanging factors, runs and permuting levels of one or more factors.

- For identifying 2 $d(n, q, s)$'s, a complete search compares $n!(q!)^s s!$ design pairs. This is a NP hard problem.

- Draper and Mitchell (1968), Chen and Lin (1991), Clark and Dean (2001), proposed several methods.

- For two isomorphic designs they have the same CD_2-value and the same distribution of projection CD_2-value.

- Ma, Fang and Lin (2001, *J. Complexity*) proposed a powerful algorithm for detecting non-isomorphic designs.

5.5. Uniformity and Hadamard matrices

- A Hadamard matrix of side n is an $n \times n$ matrix with every entry either 1 or -1, which satisfies that $HH^T = nI$.

- Two Hadamard matrices are called equivalent if one can be obtained from the other by some sequence of row & column permutations & negations.

- The Hadamard matrix has payed an important role in construction of experimental designs, code theory and others.

- To identify the equivalence of 2 Hadamard matrices by a complete search is an NP hard problem, when n increases.

- There is a unique equivalence class of Hadamard matrices of each order 1, 2, 4, 8 and 12.

Table 10: Number of equivalent classes

Order	16	20	24	28	32	36
number of classes	5	3	60	487	≥ 66000	≥ 200

- Fang and Ge (2004, *Math. Computation*) proposed an powerful algorithm that can easily detect inequivalent Hadamard matrices based on the sequence of symmetric Hamming distances. They found that there are at least 382 inequivalent classes for Hadamard matrices of order 36.

Comments on The Uniform Design

- Flexibility in design and modeling
- Easy to understand and use
- Good for nonlinear models
- Can be applied on complicated system
- Can be used for several occations:
  - experiments with unknown model
  - computer experiments
  - experiments with mixtures

Comments on The Uniform Design

"Another approach to space-filling design using methods from number theory is briefly described in Exercise 7.7. This approach is reviewed by Fang, Wang and Bentler (1994) and its application in design of experiments discussed in Ch. 5 of Fang and Wang (1994). In the computer science literature the method is often called quasi-Monte Carlo sampling; see Neiderreiter (1992)."

"Another type of space-filling design specifies points in the design space using methods from number theory. The resulting design is called a uniform, or uniformly scattered design."
– D.R. Cox and N. Reid (2000), The Thoery of the Design of Experiments.

Comments on The Uniform Design

"An important class of designs are so-called lattices. These have received considerable attention in number theory under ths heading of low discrepancy sequences. A principal text is Niederreiter (1992) and Fang and Wang (1994) (an their earlier work) make a considerable contribution in applications to statistics, including design."

"Some conclusions are that the lattice designs do surprisingly well and a good integer lattice is robust against changes of criterion." (Concluded from a two-dimensional exercise for comparing Latin hypercube design modified Latin hypercube design and lattice designs)

– R.A. Bates, R.J. Buck, E. Riccomagno and H.P. Wynn (1996), JRSS-B, 58, 77-94 (with discussion).

Comments on The Uniform Design

"If some of the noise factors have more than three levels, the run size of the orthogonal array for the noise factors may be too large. An alternative is to employ a smaller plan with uniformly spread point for the noise factors. These plans include Latin hypercube sampling (Koehler and Owen, 1996) and "uniform" designs based on number-theoretic methods (Fang and Wang, 1994). Since the noise array is chosen to represent the noise variation, uniformity may be considered to be a more important required than orthogonality."

– C.F.Jeff Wu and M. Hamada, p.445, "Experiments planing, analysis, and parameter design optimization".

Applications of UDs

- There are more than 500 hundreds case studies published in more than one hundred journals.
- More than one hundred theoretic research papers have been published in various journals.
- Ford Motor Company has been used UD for car development and "Design for Six Sigma".
- A nationwide society "The Uniform Design Association of China" was established in 1994.

Ford Motor Company

21500 Oakwood Blvd.
POEE - FK143 Zone E-3, MD65
Dearborn, MI 48121

May 2, 2003

To: Prof. Kai-Tai Fang
Department of Mathematics
Hong Kong Baptist University
Kowloon Tong, Hong Kong

Dear Prof. Fang,

As a continuation of your visit to Ford last year, we would like to invite you again to visit Ford Motor Company for one month from June 11 to July 10, 2003.

In the past few years, we have tremendous successes in using Uniform Design for computer experiments. The technique has become a critical enabler for us to execute "Design for Six Sigma" to support new product development, in particular, automotive engine design. Today, computer experiments using uniform design have become standard practices at Ford Motor Company to support early stage of product design before hardware is available.

We would like to share with you our successful real world industrial experiences in applying the methodology that you developed. Additionally, your visit will be very valuable for us to gain more insight about the methodology as well as to learn the latest development in the area. During the visit, we would like to strengthen our research collaboration to accelerate both theoretical development and practical implementation of design and analysis of computer experiments to support complex engineering design such as automotive engine.

Ford Motor Company will provide all necessary travel accommodation during this trip.

Sincerely,

A. Sudjianto

Agus Sudjianto, Ph.D. (e-mail: asudjian@ford.com)
Engineering Manager, Analytical Powertrain
Ford Motor Company

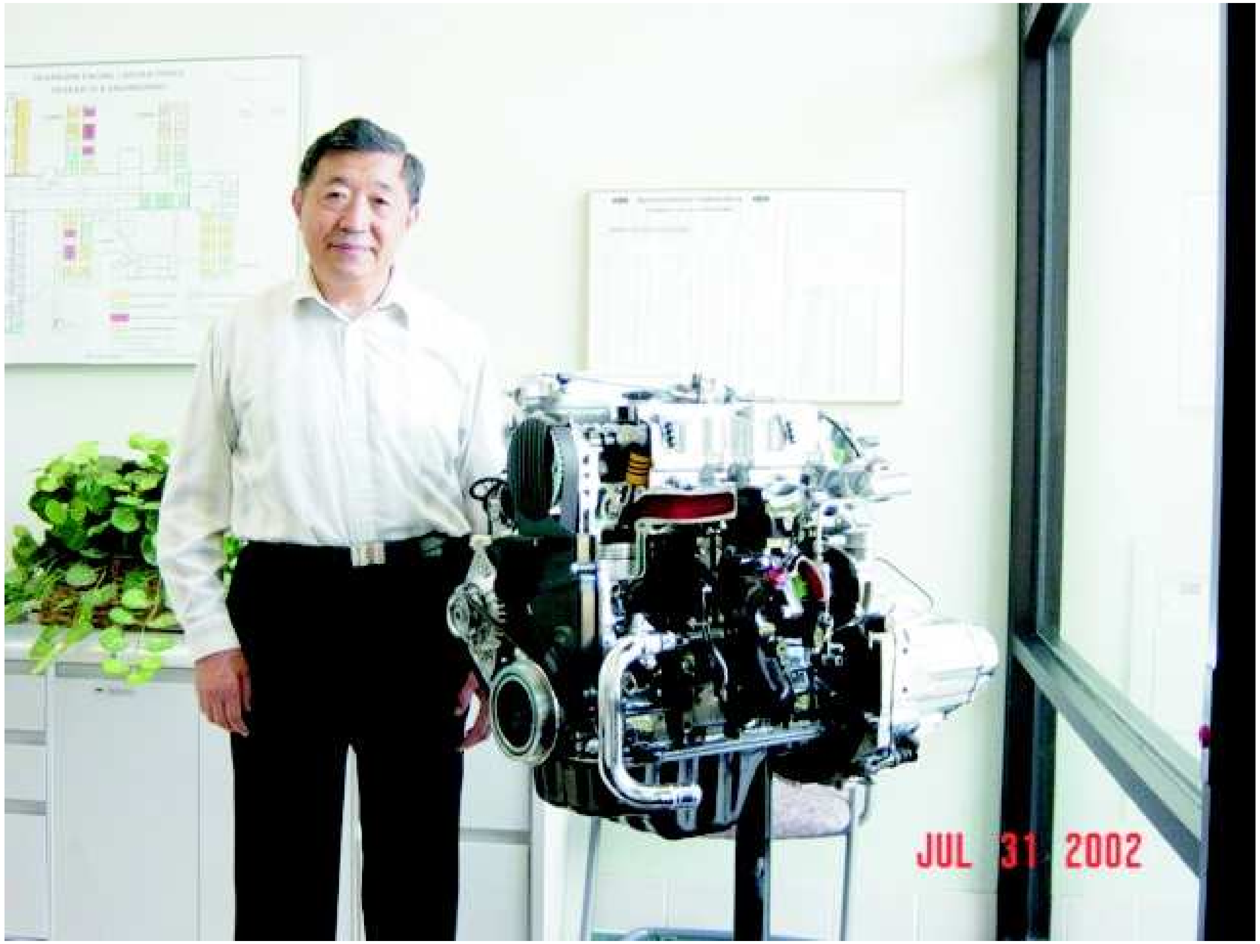


Figure 7: Kai-Tai Fang was in FORD in 2002

Conclusion

- Space filling designs are useful in experiments with model uncertainty

Modeling is a process

- All the models are wrong, but some models are useful
- Modeling is more ART than Science
- Team working

THANK YOU!!

Kai-Tai Fang

Beijing Normal Univ. - Hong Kong Baptist Univ.

United International College

Email: ktfang@uic.edu.hk

Web: www.math.hkbu.edu.hk/∼ktfang