More Uniformly Distributed Sampling does not Necessarily Lead to More Accurate Models for Computer Experiments

ABSTRACT

It is a general feeling that more uniformly distributed sampling should be better than less uniformly distributed sampling for prediction in computer experiments. A study was conducted to compare four designs for computer experiments, based on simulation tests and statistical analyses ANOVA. Maximin Latin hypercube design (LHMm) nearly always generates more uniformly distributed sampling in two dimensional cases (2D) than does random sampling, Latin hypercube design (LHD), or Minimized centered L_2 discrepancy Latin hypercube design (LHCL2). But mostly there is no significant difference among the means of the prediction errors resulted from employing LHMm and the other designs. Occasionally, even the opposite was seen. The test results indicate that more uniformly distributed sampling does not necessarily lead to more accurate models in 2D cases.

Key words: sampling, experiment design, uniformity, uniform distribution, computer experiments, metamodeling

1. Introduction

It is generally agreed that, by intuition, sampling for computer experiments should be uniformly distributed in the design region (e.g. Koehler and Owen, 1996; Santner, et al, 2003, p124, p126). The more uniformly distributed the sampling is, the better the experimental design is. However, different or even opposite results were observed in a study to compare eighteen designs for computer experiments, based on simulation tests and statistical analyses ANOVA. The design types include random sampling, Latin hypercube design (LHD), Maximin Latin hypercube design (LHMm) based on the Φ_p criterion, Minimized centered L_2 discrepancy Latin hypercube design (LHCL2), etc. The results show that more often there is no significant difference between the error means resulted from applying the different designs for sampling. Where there are significant differences, LHD and LHCL2 often outperform LHMm. However, LHMm nearly always generates more or much more uniformly distributed samples in two-dimensional designs than LHD or LHCL2 does. For more details of the study, please see Liu[1] (2004).

The results seem against the "general feeling" that better designs should have more uniformly distributed sampling. Since uniformity has been taken as a fundamental issue for experimental design in computer experiments and thus has significant impact on research and application, another study with more tests was conducted for further investigation and is described below.

2. Test scheme

There are two groups of testing. In the first group, the four designs mentioned above are employed and compared simultaneously. Twenty test functions are approximated and there are five levels of the sample size. The second group focuses on one function, one level of sample size, and pair-wised comparison at a time. Only 2D cases are tested in both groups.

In the first group, there are two sub-groups: comparing RRMSE and comparing Max-rel-error. In each sub-group, for the same sample size, the four designs are employed for sampling as four different "treatments." For each sampling, a test function is evaluated at every point. A Kriging model is built to fit these sampled points and the corresponding responses, as the

approximation model. To test the prediction accuracy, 10000 validation points are sampled by LHD within the domain [-30,30; -30,30] that is specified for some of the test functions used in this study in the literature and that is used for all the functions. At each point, the values of the test function and of the approximation model are evaluated and compared to find the difference. From all the validation points, a relative error called relative root mean square error (RRMSE) is found by the following formula:

$$RRMSE = \sqrt{\frac{1}{10000}\sum_{i=1}^{10000}\left(\frac{Ft(i)-Fa(i)}{|Ft(i)|+\varepsilon}\right)^2}$$

Ft: function response; Fa: approximation response; ε: 10^{-4}: to guard against possible Ft = 0.

And, the maximum relative error is found by the following formula:

$$Max-rel-error = Max\left(\frac{|Ft(i)-Fa(i)|}{|Ft(i)|+\varepsilon}\right), \qquad i = 1,2,...,10000$$

These two results, RRMSE and Max-rel-error are the two "observations" for approximating the function using the sampling and Kriging model. Nineteen more RRMSE observations and nineteen more Max-rel-error observations are generated by the same procedure for approximating nineteen more test functions. This process is repeated for the other three samplings or "treatments." In all, there are twenty observations for each sampling for comparing RRMSE and Max-rel-error respectively. One-way ANOVA is used to compare the designs by comparing the means of RRMSE in the first sub-group or Max-rel-error in the second sub-group. The whole process outlined above is repeated for another sample size.

The sample sizes (m) tested are 8, 10, 12, 14, 16 for meaningful comparison of the distribution uniformity for which there seems no mathematical definition widely used in the literature. With too few or too more points, it may be difficult to tell if the distributions are uniform by visual inspection. The relative error is employed since the criterion may be more meaningful for prediction and is needed as the "observations" used in comparison with ANOVA. Since 20 different functions are used, using root mean square error may inflate the within group variance to make it difficult to detect small differences in the error means resulted from the different "treatments."

In the second group of testing, many pair-wised comparison between LHMm and one of the other three designs were conducted for one test function only. After sampling, the Kriging model is built and popular root mean square error (RMSE) is found based on 10,000 validation points. The same process is repeated four more times for the same sample size and the same design type. Then, the whole process is repeated for the other design. Then ANOVA is done, each treatment having five observations of RMSE. Then, the whole process is repeated for another function.

The details of the designs, the twenty test functions, and the Kriging models used are described below. To optimize the criterion Φ_p for LHMm and the criterion CL2 for LHCL2, an optimizer using genetic algorithms (GA) was proposed and employed (Liu, 2004).

The following lists the four design types compared.

1) Design type 1: Random sampling (Rd)
2) Design type 2: Latin hypercube design (LHD)

The j^{th} component of the i^{th} sampled point is

$$\mathbf{X}_{ij} = \frac{\pi_{ij} - U_{ij}}{m}$$

where $\boldsymbol{\pi}_{ij}$ is the j[th] element of the i[th] independent uniform random permutations of the integers 1 through m (m is the number of samples), and $\boldsymbol{U}_{ij}$ is the j[th] element of the i[th] independent U [0, 1] (uniform distribution between 0 and 1) random variables independent of the $\boldsymbol{\pi}_{ij}$ (Santner, et al, 2003).

3) Design type 3: Maximin Latin hypercube design (LHMm)
Criterion: Φ_p criterion (slightly modified in form from that of Morris and Mitchell, 1995)

$$\Phi_p = \left[\sum_{i=1}^{m} \sum_{j=i+1}^{n} d_{ij}^{-p} \right]^{1/p}$$

n: number of design variables (2D: n =2)
m: the number of sampled points
d_{ij}: the Euclidean distance between points i and j
p: parameter. The tests show that p values do not impact prediction results in most of the 2D cases, p = 1 being used (Liu[1], 2004)

4) Design type 4: Minimized CL2 Latin hypercube design (LHCL2)
Criterion: Centered L_2 discrepancy CL_2 (Hickernell, 1998)

$$[CL_2(P_m)]^2 = (\tfrac{13}{12})^2 - \tfrac{2}{n} \sum_{k=1}^{m} \prod_{j=1}^{n} [1 + \tfrac{1}{2} | x_{kj} - 0.5 | - \tfrac{1}{2} | x_{kj} - 0.5 |^2]$$

$$+ \tfrac{1}{n^2} \sum_{k=1}^{m} \sum_{j=1}^{m} \prod_{i=1}^{n} [1 + \tfrac{1}{2} | x_{ki} - 0.5 | + \tfrac{1}{2} | x_{ji} - 0.5 | - \tfrac{1}{2} | x_{ki} - x_{ji} |]$$

m: number of samples
n: number of dimensions
P_m: a set of m points

Most of the test functions used are popular functions for testing global optimization methodologies. Some of them are very similar to those "real-world responses" reported in the literature on engineering applications of metamodeling (e.g. Simpson, et al, 2001). The details of the functions are given in Appendix. The approximation models used are Kriging models (Sacks, et al ,1989; Welch, et al, 1992). They are more capable to approximate highly nonlinear functions than many other models e.g. low order polynomials. For the regression part, a constant was used. Gaussian stochastic processes were used as the correlation models. For each dimension, each parameter θ was found through optimization (Nielsen, et al).

3. Results of simulation tests and ANOVA

3.1 Comparing four designs simultaneously

There are two test sub-groups for comparing the designs: Sub-Group 1 based on RRMSE and Sub-Group 2 based on Max-rel-error. For validation, only one sampling group of 10000 points was generated which was used by all the tests respectively in the two sub-groups. The sampling plots and the ANOVA plots together with the p-values are shown in Figures 1 to 10.

For one-way ANOVA, the explanation of the box plots follows. The centerline is the median; the bottom and top lines of the box are the quartiles; the end lines of the whiskers are the ends of the data range. The outliers are those marked as "+". Not all the outliers are shown.

Figure 1. Comparing four designs based on comparing RRMSE, m=8 (m: sample size)

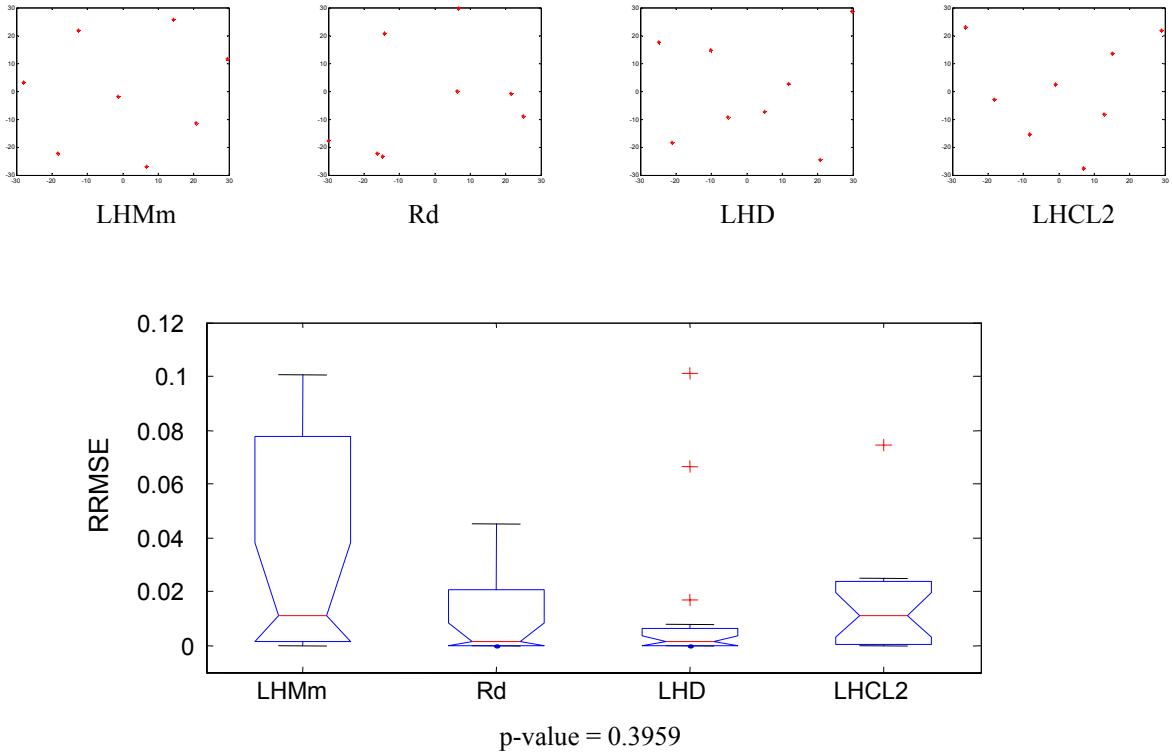


Figure 2. Comparing four designs based on comparing RRMSE, m=10 (m: sample size)

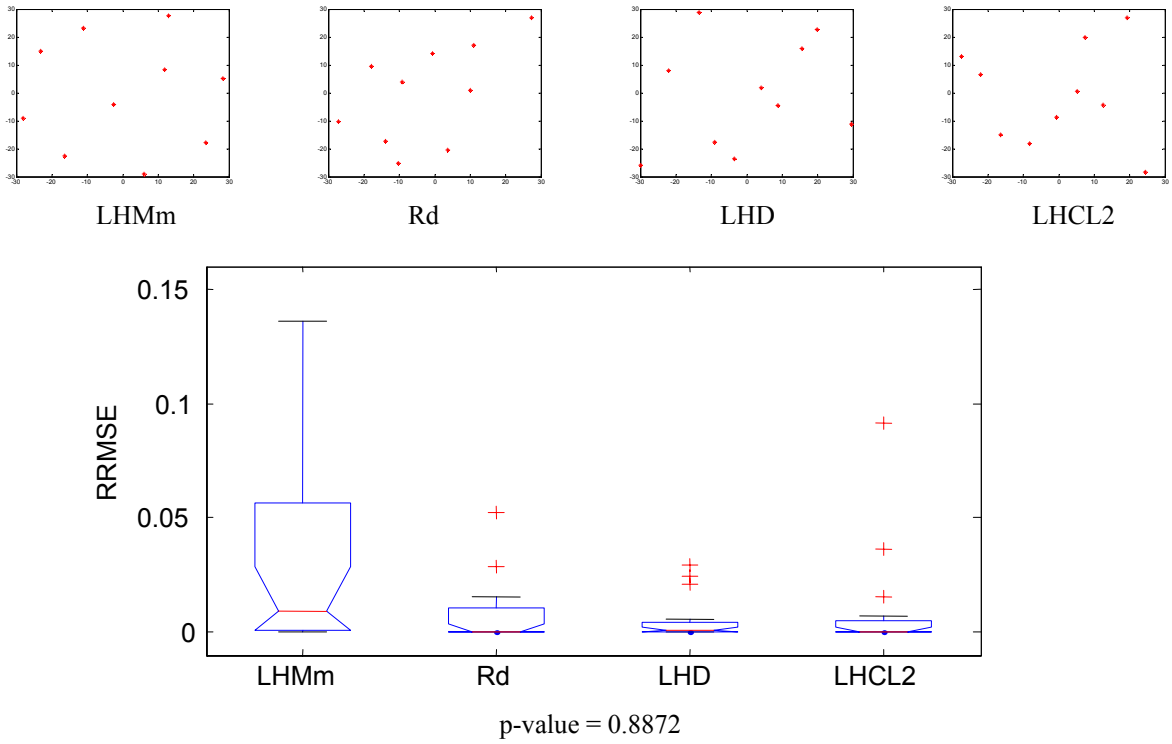
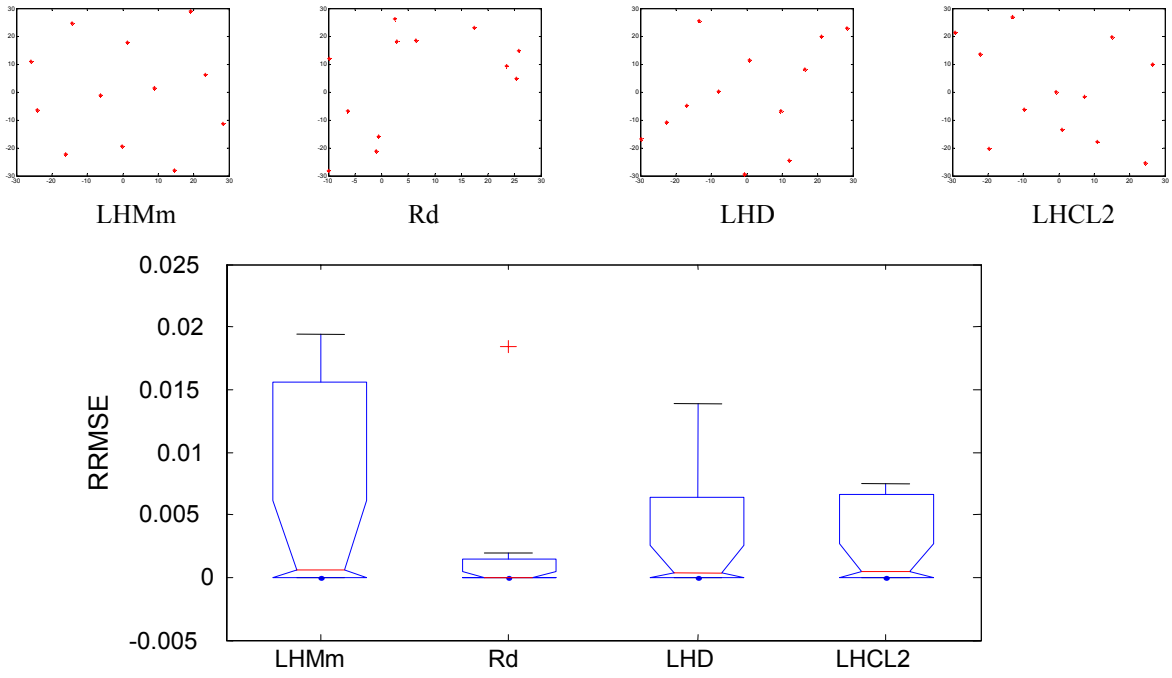
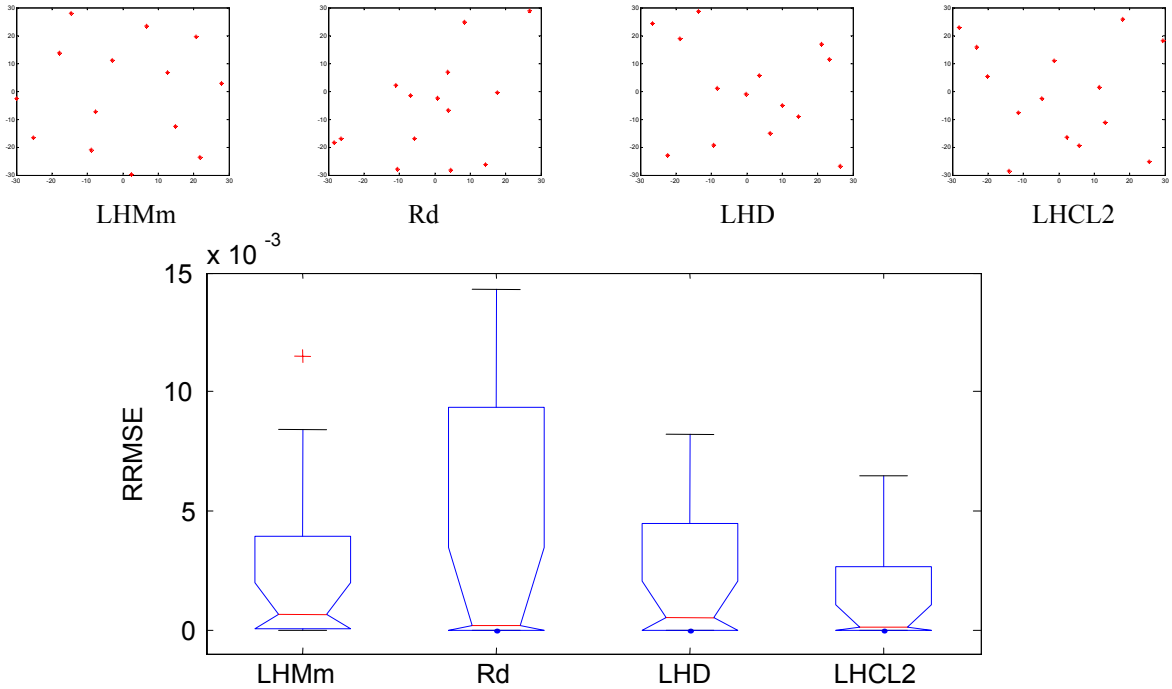


Figure 3. Comparing four designs based on comparing RRMSE, m=12 (m: sample size)



p-value = 0.4041

Figure 4. Comparing four designs based on comparing RRMSE, m=14 (m: sample size)



p-vale = 0.7300

Figure 5. Comparing four designs based on comparing RRMSE, m=16 (m: sample size)

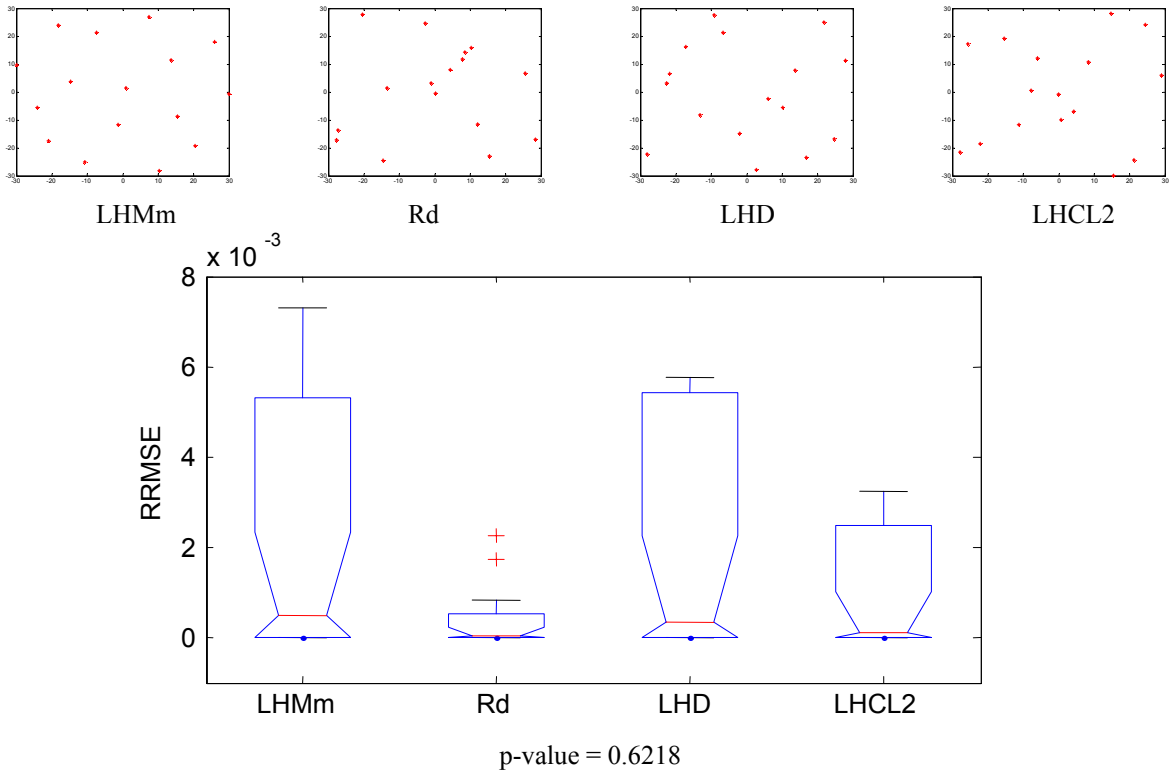


Figure 6. Comparing four designs based on comparing Max-rel-error, m=8 (m: sample size)

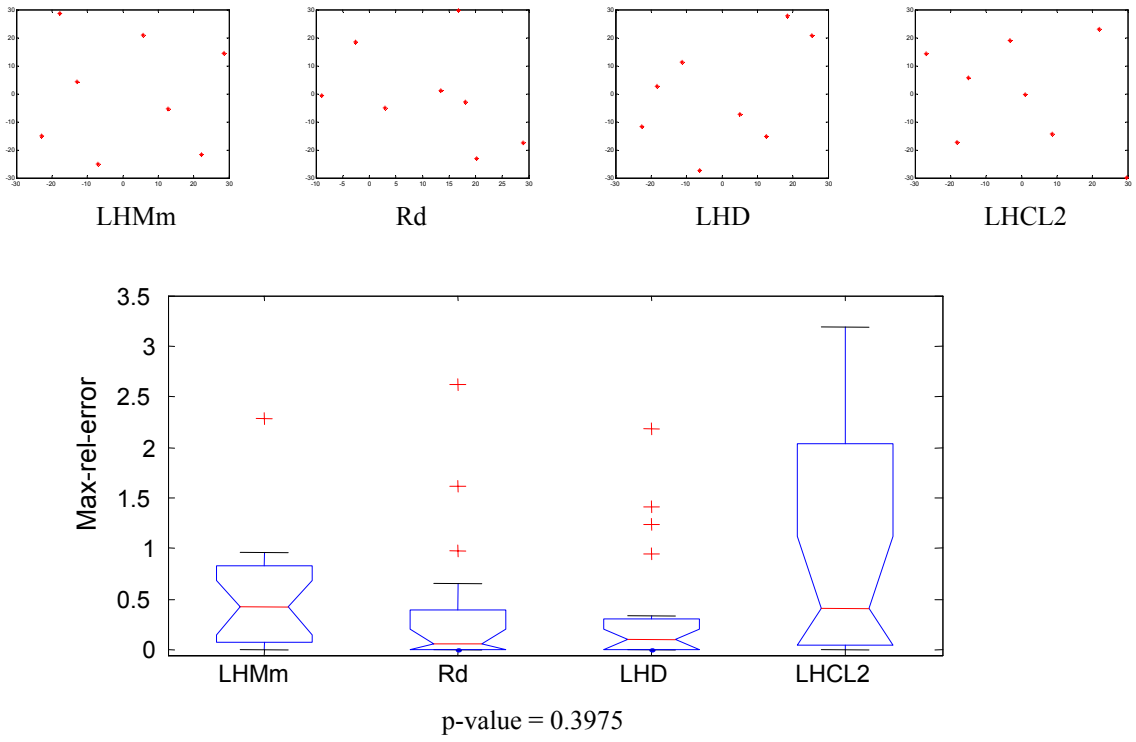


Figure 7. Comparing four designs based on comparing Max-rel-error, m=10 (m: sample size)

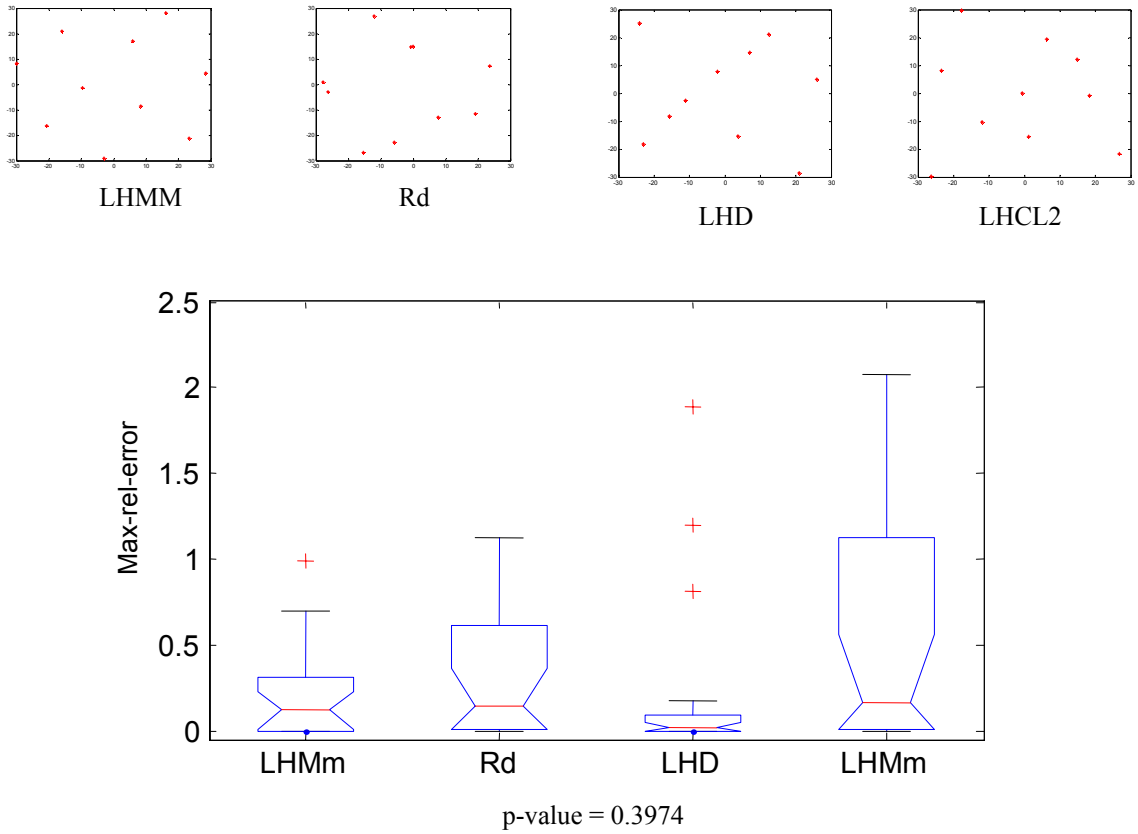


Figure 8. Comparing four designs based on comparing Max-rel-error, m=12 (m: sample size)

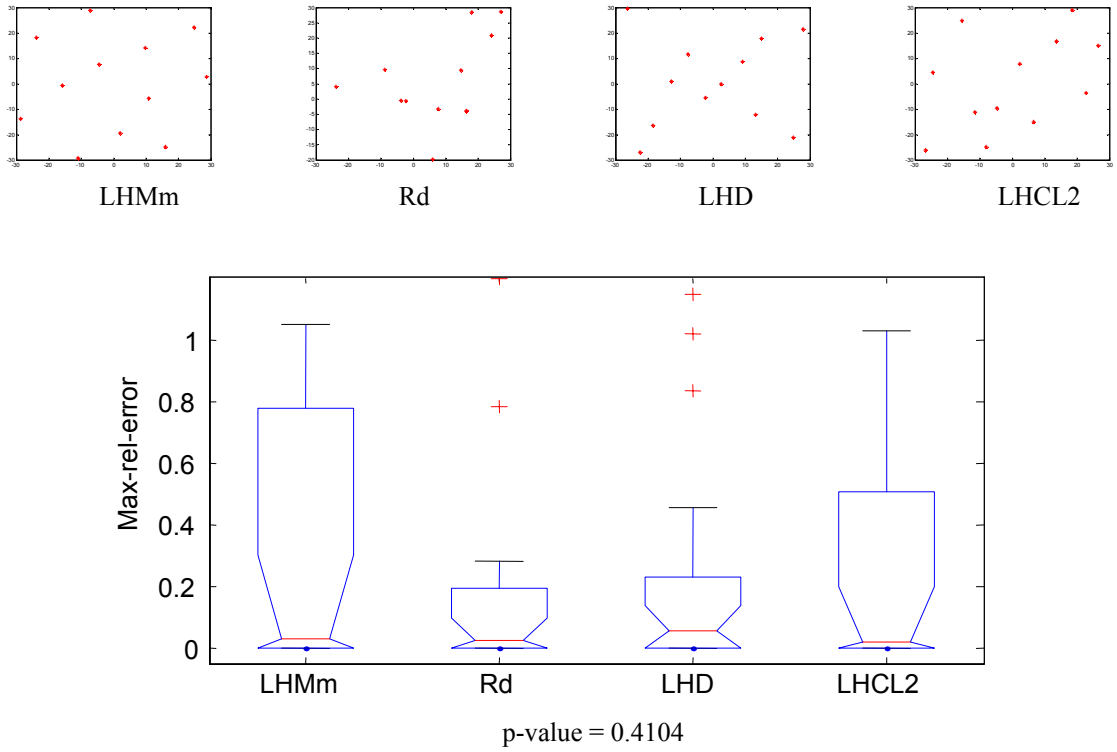


Figure 9. Comparing four designs based on comparing Max-rel-error, m = 14 (m: sample size)

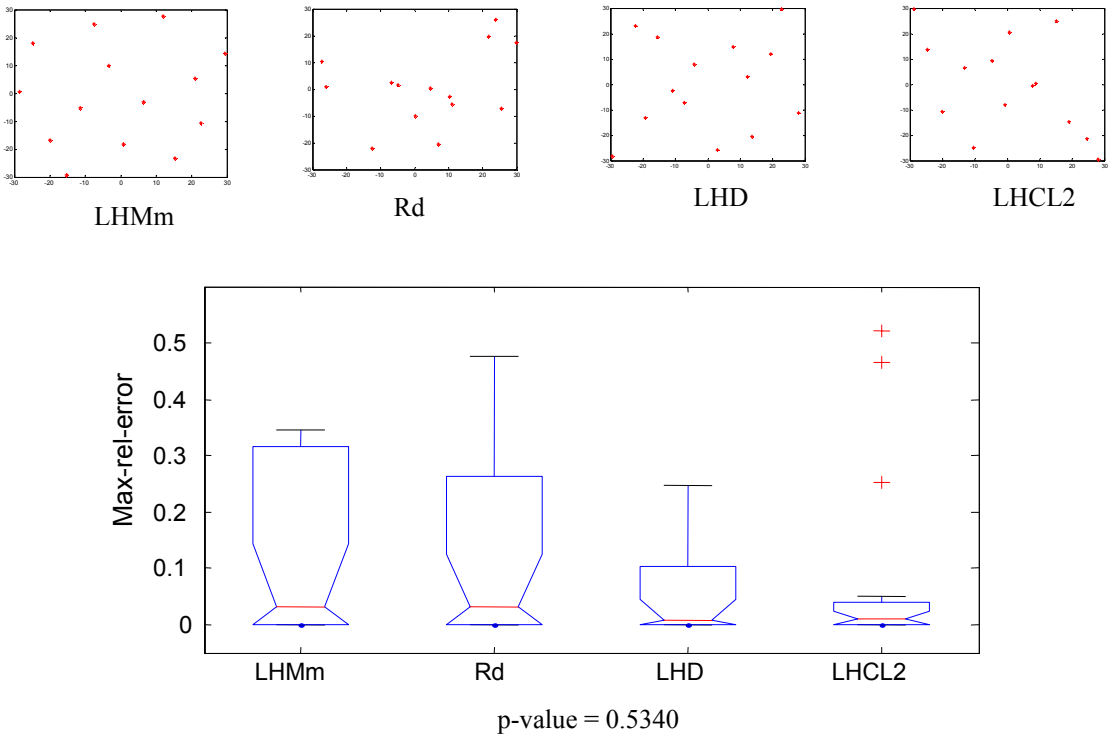
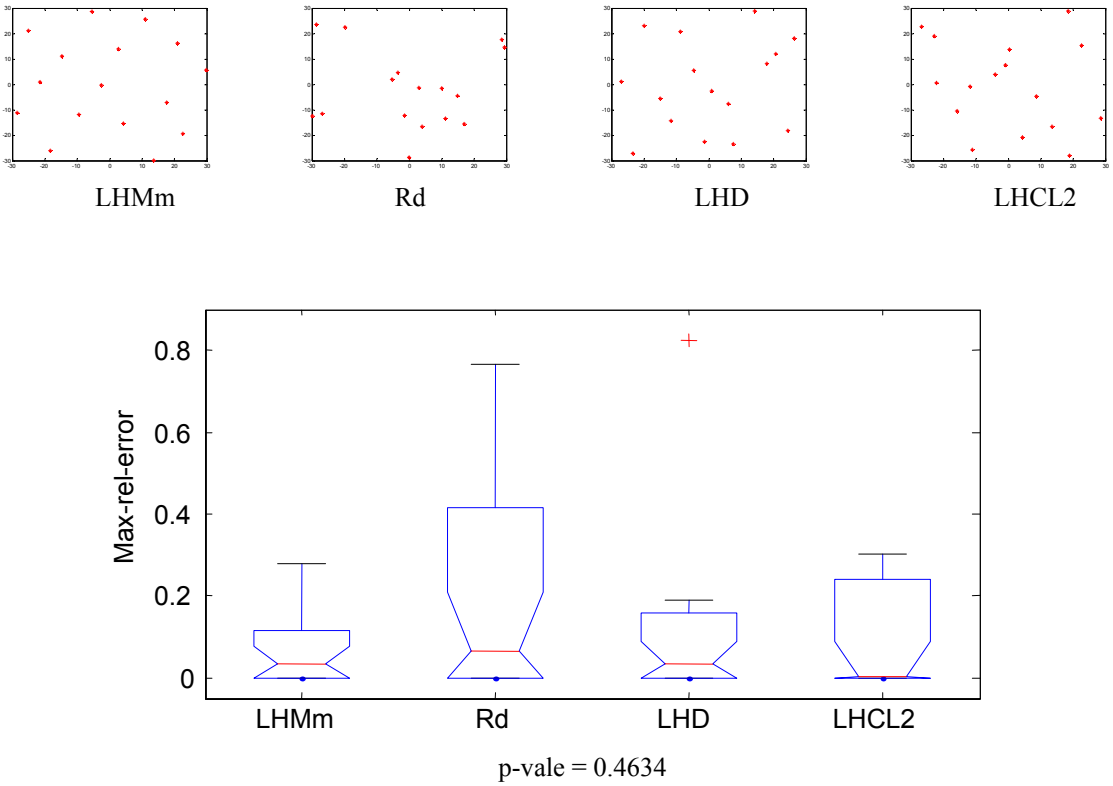


Figure 10. Comparing four designs based on comparing Max-rel-error, m= 16 (m: sample size)

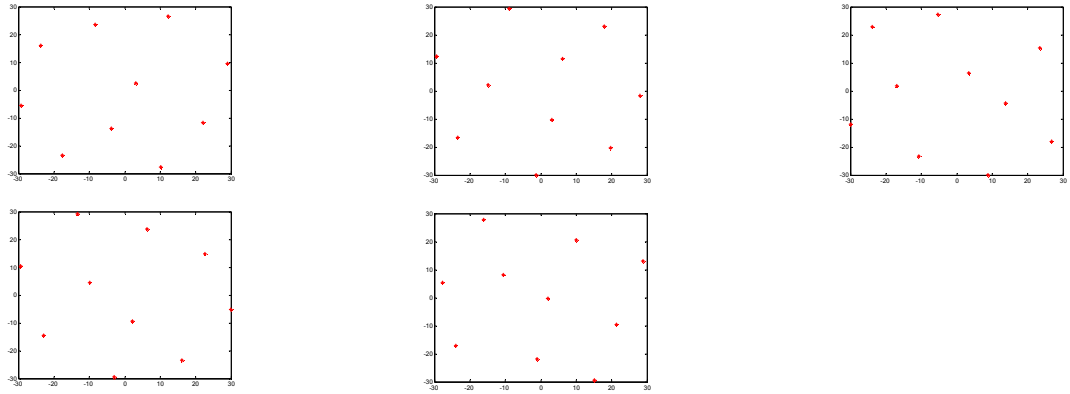


3.2. Pair-wised comparison between LHMm and Random sampling (Rd)

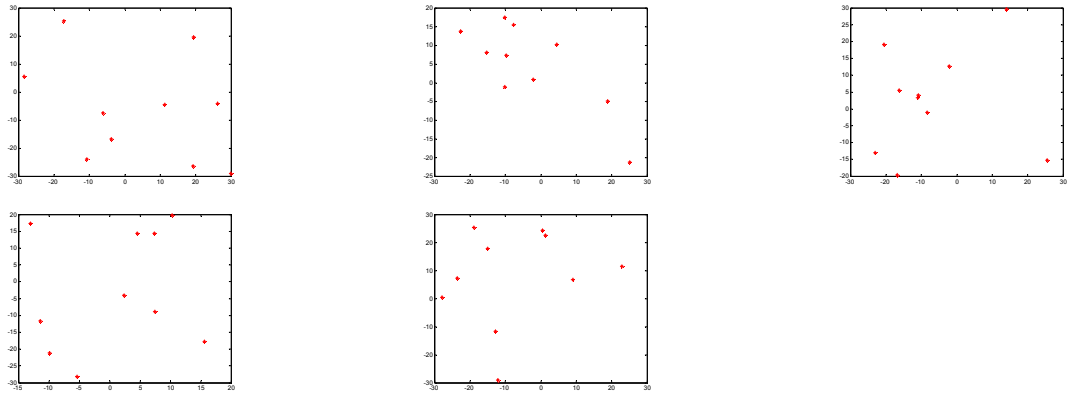
The pair-wised comparison between LHMm and one of the other three designs show that again mostly there is no significant difference in the RMSE means resulted from using LHMm and the other design. To show less uniformly distributed sampling can result in higher prediction accuracy occasionally, one pair-wised comparison between LHMm and random sampling is provided in Figure 11. The sole test function is AC. The sample size is 10.

Figure 11. Comparing LHMm and Rd based on comparing RMSE, m=10 (m: sample size), function AC

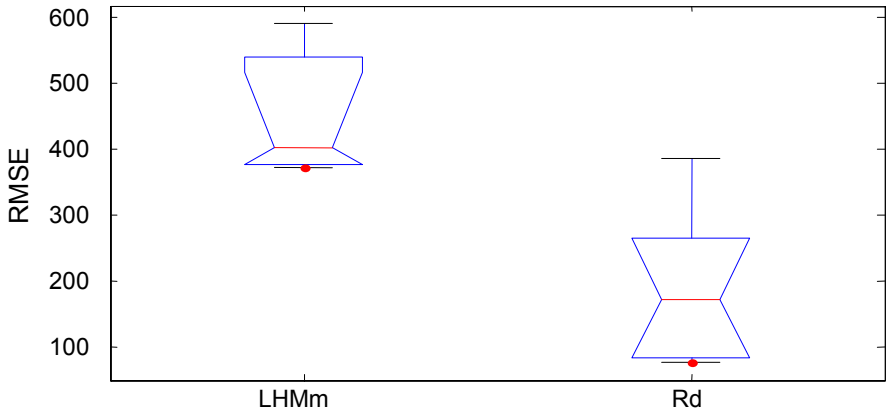
1) LHMm



2) Random sampling (Rd)



3) ANOVA Plot and p-value



P-value = 0.00624.

4. Conclusions and discussion

It has been shown that Maximin Latin hypercube design (LHMm) generates more or much more uniformly distributed sampling than does random sampling, Latin hypercube design (LHD), or Minimized centered L_2 discrepancy Latin hypercube design (LHCL2). However, in most cases, there is no statistically significant difference in the means of RRMSE, Max-rel-error, or RMSE, regardless of the critical p-value selected as 0.01 or 0.05. In some cases, most of the points are along one straight line or cluster in a corner, but still there is no significant difference in the prediction error means resulted from these samplings and LHMm samplings. In Section 3.2, LHMm is significantly worse than random sampling despite more uniformly distributed sampling by LHMm, instead of being better as expected.

Santner, et al (2003) described a sampling along a straight line, resulting in higher accuracy at the line but poor prediction elsewhere. In that case, the maximum error is expected to be higher than that with more uniformly distributed sampling. The extreme cases like that one is not seen or enforced in this study. But for some less severe cases shown, where most points are distributed along one line or cluster in a corner, there seem no significant difference in either RRMSE or Max-rel-error resulted from these "bad" samplings and "better" LHMm samplings.

In terms of the variance or the range of the data (either RRMSE or Max-rel-error), it seems that uniformity does not have consistent impact either. But sample sizes do have the impact. As the sample size becomes larger, the range mostly becomes smaller, which can be seen from the figures (1-10) above (the figures have different axis limits for the errors).

It has been shown that more is not necessarily better than less uniformly distributed sampling for prediction accuracy in computer experiments in 2D cases. It might be true for higher dimensions also. Thus, maybe it is not appropriate to use uniformity as the sole criterion to compare different experimental designs or as the only goal to be pursued for better designs. Much research effort has been devoted to "better" designs with more uniformly distributed sampling. It seems that better goals than uniformity need to be studied. What is more important than uniformity or "better designs" may be enough samples for prediction accuracy, as is discussed by the author in Liu[1] (2004) and Liu[2] (2005).

Because of very limited resources, the range of this study is very limited. Only Kriging models are used to test twenty functions to compare four designs. Larger range may be needed. Since it is not easy to check visually if the sampling is uniformly distributed for the dimensions higher than two, comparison for uniformity for higher dimensions has not been done. Much more and deeper study is needed beyond this preliminary investigation. Nevertheless, the tests done have revealed that uniformity may not be as important as many thought for prediction accuracy. Since sampling uniformity has been taken as a fundamental issue for computer experiments, this observation may be useful to the users as well as the researchers for their choice or development of the experimental designs.

In comment on the earlier manuscript of this paper, Professor Bill Notz provided more observations and really thoughtful analysis or explanation (private communication):

"The results you have found agree with what we have observed. Our experience is that any design that is reasonably uniform seems to work well. Only very nonuniform designs (for example, designs that take most of their observations on the boundary) seem to perform poorly.

We believe the reason is that interesting features of response surfaces (locations of maxima, minima, regions where the response surface varies greatly) generated by simulations tend to be "uniformly" spread out over the design region. The variation in the location of these interesting features is larger than subtle variations in the uniformity of designs. Thus, only designs that take observations in a very limited portion of the design space perform poorly."

ACKNOWLEDGEMENT

We are grateful to Professor Bill Notz who read the earlier manuscript and provided the detailed comment, explanation, and more observations on sampling uniformity as well as sample size effect.

REFERENCE

Hickernell, F.J. (1998): A generalized discrepancy and quadrature error bound, *Mathematics of Computation*, 67, 299-322.

Koehler, J. R., Owen, A. B. 1996: *Computer experiments*. In Ghosh, S. and Rao, C. R., editors, *Handbook of Statistics*, 13, 261--308. Elsevier Science, New York.

Liu[1], L., 2004: Employing simulation and optimizer to optimize experimental design and structural topology, dissertation, Systems Science Ph.D. Program, Portland State University.

Liu[2], L., 2005: Could enough samples be more important than better designs for computer experiments? 38th Annual Simulation Symposium, Spring Simulation Multiconference, San Diego, April 2-8, 2005.

Morris, M. D. and Mitchell, T. J., 1995, "Exploratory Designs for Computer Experiments," Journal *of Statistical Planning and Inference*, 43, 381-402.

Nielsen, H.B., Lophaven, S.N., and Søndergaard, J.: DACE: A MATLAB Kriging Toolbox, http://www.imm.dtu.dk/~hbn/dace.

Santner, T., Williams, B, and Notz, W. 2003: *Design and Analysis of Computer Experiments*, Springer-Verlag, New York.

Simpson, T. W., Lin, D. K. J. and Chen, W. (2001) "Sampling Strategies for Computer Experiments: Design and Analysis," International Journal of Reliability and Applications, 2:3 (209-240).

Appendix Test functions

1) Function 1(AC): Ackley's path function

$$f(x) = -a * \exp(-b * \sqrt{(1/n)(\sum_{i=1}^{n} x_i^2)} - \exp(\frac{1}{n}\sum_{i=1}^{n}\cos(cx_i)) + a + e$$

a = 20; b = 0.2; c = 2·pi; I =1:n;

2) Function 2 (AX): Axis parallel hyper-ellipsoid function

$$f(x) = \sum_{i=1}^{n} i x_i^2$$

3) Function 3 (DE): De Jong's function 1

$$f(x) = \sum_{i=1}^{n} x_i^2$$

4) Function 4 (RB): Rosenbrock's valley (De Jong's function 2)

$$f(x) = \left[\sum_{i=1}^{n} 100(x_{i+1} - x_i^2)\right]^2 + (1 - x_i)^2$$

5) Function 5 (RY): rotated hyper-ellipsoid function

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{i} x_{ij}^2$$

6) Function 6 (MI): Michalewicz's function

$$f(x) = -\left(\sum_{i=1}^{n}\sin x_i\right) * \left[\frac{\sin(ix_i^2)}{\pi}\right]^{2m}$$

m=10;

7) Function 7 (BR): Branins' rcos function

$$f(x_1, x_2) = a(x_2 - bx_1^2 + cx_1 - d)^2 + e(1-f)\cos x_1 + e$$

a=1, b=5.1/(4·pi^2), c=5/pi, d=6, e=10, f=1/(8·pi);

8) Function 8 (GD): Goldstein-Price's function

$$f(x_1, x_2) = [1 + (x_1 + x_2 + 1)^2 (19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)] *$$
$$[30 + (2x_1 - 3x_2)^2 (18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)]$$

9) Function 9 (SX): Six-hump camel back function

$$f(x_1, x_2) = (4 - 2.1x_1^2 + \frac{x_1^4}{3})x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$$

10) Function 10 (PK): Peaks function

$$f(x, y) = 3(1-x)^2 \exp(-x^2 - (y+1)^2) - 10(\frac{x}{5} - x^3 - y^5)\exp(-x^2 - y^2) -$$
$$\frac{1}{3}\exp(-(x+1)^2 - y^2)$$

11) Froth function

f= -13+x+((5-y)y-2)y-29+x+((y+1)y-14)y;

12) Helix2 function

$$f = 10(\sqrt{x^2 + y^2} - 1)$$

13) Rose function

f = 10(y-x^2)

14) Sing2 function

f = $\sqrt{5}$ (x-y);

15) Sing3 function

f = (x-2y)^2;

16) Sing4 function

f = $\sqrt{10}$ ((x-y)^2) ;

17) Wood1 function

f = 10*(y-x^2);

18) Wood3 function

f = $\sqrt{90}$ (y-x^2);

19) Wood5 function

f= $\sqrt{10}$ (x+y-2);

20) Wood6 function

f= (x-y)/ $\sqrt{10}$