Sampling Plans

Design of Experiments

Sampling plans for linear regression

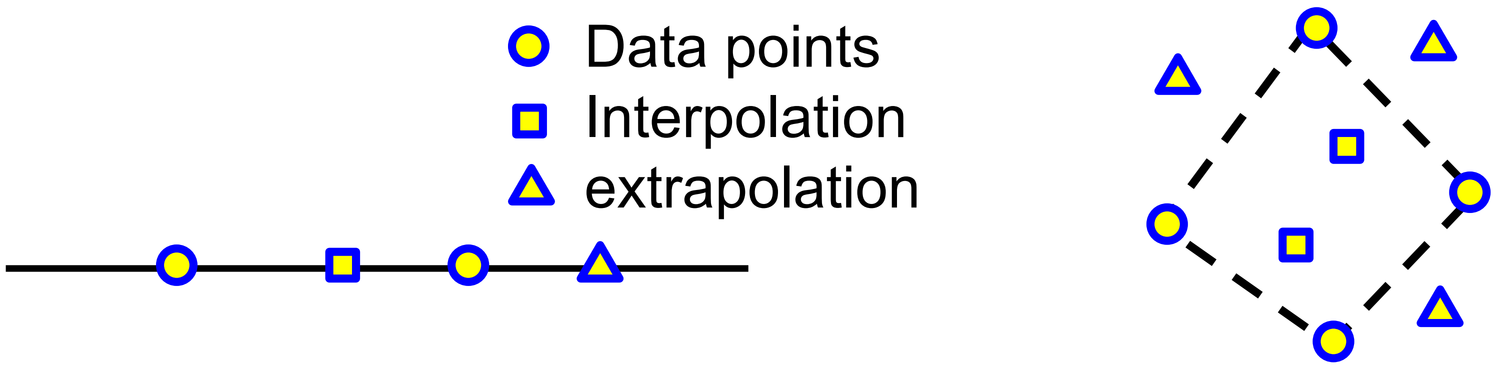
- Given a domain, we can reduce the prediction error by good choice of the sampling points.
- The choice of sampling locations is called "design of experiments" or DOE.
- With a given number of points the best DOE is one that will minimize the prediction variance.
- The simplest DOE is full factorial design where we sample each variable (factor) at a fixed number of values (levels)
  - Ex) with four variables and three levels, we will sample 81 points
  - not practical except for low dimensions

DOE in boxlike domains

- Range: $x_i^l \le x_i \le x_i^u,\ i = 1, \cdots, n$
- Normalize: $\tilde{x}_i = \dfrac{2x_i - x_i^l - x_i^u}{x_i^u - x_i^l} \Rightarrow -1 \le \tilde{x}_i \le 1$

- Full factorial design – include all possible combinations
  - Ex) 2 level, n variables → 2^n DOE
- Fractional factorial design – use a subset of all combinations
  - Ex) n = 10, linear PRS, # of coefficients = 11 $\ll 1024$
  - Fractional factorial design will lose interpolation property

- 1D 2D



- Interpolation: within the convex hull of data points
- n-dimension simplex (n+1 vertices)

$$\mathbf{x} = \sum_{i=1}^{n+1} \alpha_i \mathbf{x}_i, \qquad \sum_{i=1}^{n+1} \alpha_i = 1, \;\; \alpha_i \geq 1$$

  – Convex hull: union of all simplices

Review: linear regression

- Surrogate is linear combination of n_β given shape functions

$$\hat{y} = \sum_{i=1}^{n_\beta} b_i \xi_i(\mathbf{x})$$

- For linear approximation: $\xi_1 = 1 \quad \xi_2 = x$

- Difference (error) between n_y data and surrogate

$$e_j = y_j - \sum_{i=1}^{n_\beta} b_i \xi_i(\mathbf{x}_j), \mathbf{e} = \mathbf{y} - \mathbf{Xb}$$

- Minimize square error: $\mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb})$

- Differentiate to obtain: $\mathbf{X}^T\mathbf{Xb} = \mathbf{X}^T\mathbf{y}$

Model based error for linear regression

- The common assumptions for linear regression
  - The true function is described by the functional form of the surrogate.
  - The data is contaminated with normally distributed error with the same standard deviation at every point.
  - The errors at different points are not correlated.
- Under these assumptions, the noise standard deviation (called standard error) is estimated as

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T \mathbf{e}}{n_y - n_\beta}$$

- $\hat{\sigma}$ is used as estimate of the prediction error.

Prediction variance

- Linear regression model $\hat{y} = \sum_{i=1}^{n_\beta} b_i \xi_i(\mathbf{x})$

- Define $x_i^{(m)} = \xi_i(\mathbf{x})$ then $\hat{y} = \boldsymbol{\xi}^{\mathrm{T}}\mathbf{b}$

- With some algebra

$$V[\hat{y}(\mathbf{x})] = \boldsymbol{\xi}^{\mathrm{T}} \sum\nolimits_b \boldsymbol{\xi} = \sigma^2 \boldsymbol{\xi}^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \boldsymbol{\xi}$$

- Prediction standard error $s_y = \hat{\sigma}\sqrt{\boldsymbol{\xi}^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \boldsymbol{\xi}}$

- Idea: choose sample point to make s_y minimum

Ex) Full factorial design

- $\hat{y}(x_1, x_2) = b_1 + b_2 x_1 + b_3 x_2,\ -1 \le x_1, x_2 \le 1$
- Data: $(-1,-1,y_1), (-1,1,y_2), (1,-1,y_3), (1,1,y_4)$

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} = 4\mathbf{I}, \quad \xi = \begin{Bmatrix} 1 \\ x_1 \\ x_2 \end{Bmatrix}$$

$$\Rightarrow \xi^T(\mathbf{X}^T\mathbf{X})^{-1}\xi = 0.25(1 + x_1^2 + x_2^2)$$

$$\Rightarrow s_y = \frac{\hat{\sigma}}{2}\sqrt{1 + x_1^2 + x_2^2}$$

– Minimum at $(0,0)$, $s_{y,\min} = \frac{\hat{\sigma}}{2}$

– Maximum at $(\pm 1, \pm 1)$, $s_{y,\max} = \frac{\sqrt{3}\hat{\sigma}}{2}$

$$\frac{s_{y,\max}}{s_{y,\min}} = \frac{\frac{\sqrt{3}}{2}\hat{\sigma}}{\frac{1}{2}\hat{\sigma}} = \sqrt{3}$$

Ex) Fractional factorial design

- $\hat{y}(x_1, x_2) = b_1 + b_2 x_1 + b_3 x_2,\ -1 \le x_1, x_2 \le 1$
- Data: $(-1,-1,y_1), (-1,1,y_2), (1,-1,y_3)$

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}, \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{bmatrix}, \quad (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{4}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

$\Rightarrow$ $$\xi^T(\mathbf{X}^T\mathbf{X})^{-1}\xi = \frac{1}{2}(1 + x_1 + x_2 + x_1^2 + x_2^2 + x_1 x_2)$$

$\Rightarrow$ $$s_y = \frac{\hat{\sigma}}{\sqrt{2}}\sqrt{1 + x_1 + x_2 + x_1^2 + x_2^2 + x_1 x_2}$$

– At $(0,0)$: $s_y = \frac{\hat{\sigma}}{\sqrt{2}}$, $(-1,-1), (-1,1), (1,-1)$: $s_y = \hat{\sigma}$, $(1,1)$: $s_y = \sqrt{3}\hat{\sigma}$

– Minimum at $(-\frac{1}{3}, -\frac{1}{3})$: $s_{y,\min} = \frac{\hat{\sigma}}{3}$

$$\frac{\partial s_y}{\partial x_1} = \frac{\partial s_y}{\partial x_2} = 0$$

$$\frac{s_{y,\max}}{s_{y,\min}} = \frac{\sqrt{3}\hat{\sigma}}{\frac{1}{3}\hat{\sigma}} = 3\sqrt{3}$$

Prediction variance for full factorial design

- Recall that prediction standard error

$$s_y = \hat{\sigma}\sqrt{\boldsymbol{\xi}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\xi}}$$

- For full factorial design, the domain is normally a box.
- Cheapest full factorial design: two levels (not good for quadratic polynomials).
- For a linear polynomial standard error is then

$$s_y = \frac{\hat{\sigma}}{2^{n/2}}\sqrt{1 + x_1^2 + x_2^2 + \ldots + x_n^2}$$

- Maximum error at vertices

$$s_y = \hat{\sigma}\sqrt{\frac{n+1}{2^n}}$$

- What does the ratio in the square root represent?

DOEs for linear PRS

- Two levels (2 samples in each design variable)
- Orthogonal designs: $[\mathbf{X}^T\mathbf{X}]$ is diagonal → min variance
- Full factorial design: orthogonal design
- Saturated design: n+1 designs for n variables (simplex)
- How to construct orthogonal designs for saturated design?
  – Perfect simplex: the distance b/w all pts are the same

Ex) DOEs for linear PRS

- 3 points linear PRS $\hat{y} = b_1 + b_2 x_1 + b_3 x_2$
- Select $(\sqrt{\tfrac{3}{2}}, -\sqrt{\tfrac{1}{2}}), (-\sqrt{\tfrac{3}{2}}, -\sqrt{\tfrac{1}{2}}), (0, \sqrt{2})$

$$\mathbf{X} = \begin{bmatrix} 1 & \sqrt{\frac{3}{2}} & -\sqrt{\frac{1}{2}} \\ 1 & -\sqrt{\frac{3}{2}} & -\sqrt{\frac{1}{2}} \\ 1 & 0 & \sqrt{2} \end{bmatrix}, \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix}, \quad \xi^T(\mathbf{X}^T\mathbf{X})^{-1}\xi = \frac{1}{3}(1 + x_1^2 + x_2^2)$$

- At (0,0), $s_y = \frac{1}{\sqrt{3}}\hat{\sigma}$
- At $(\pm 1, \pm 1)$, $s_y = \hat{\sigma}$

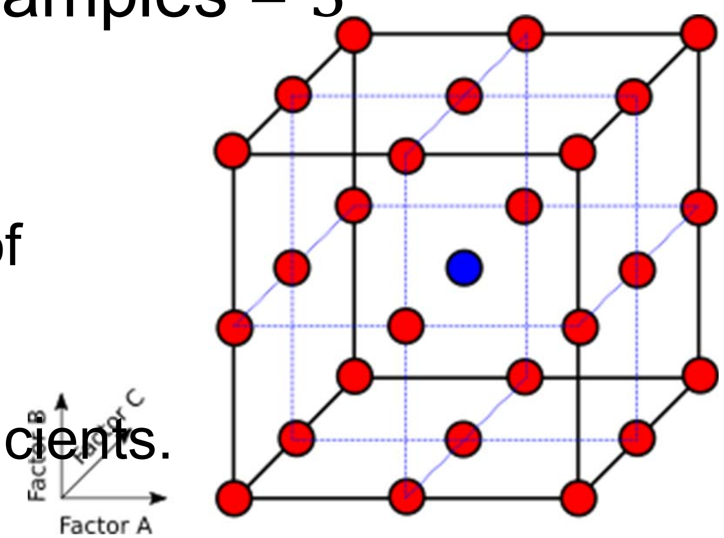
$$\frac{s_{y,\max}}{s_{y,\min}} = \sqrt{3}$$

Improved!

- But, will increase "bias error" (modeling error)

DOE for Quadratic PRS

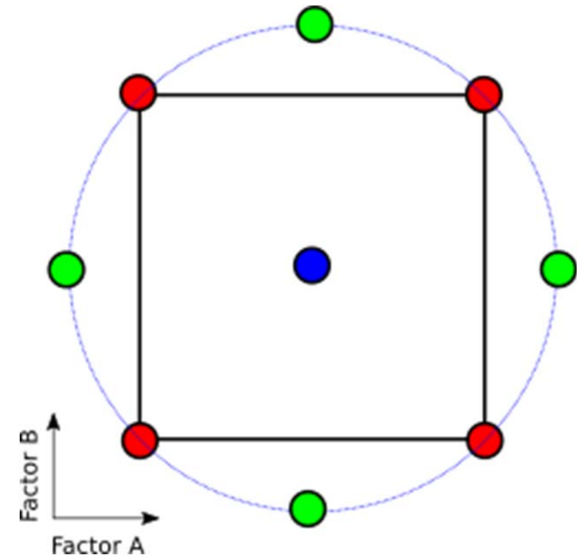
- A quadratic polynomial has $(n+1)(n+2)/2$ coefficients, so we need at least that many points.
- Need at least three different values of each variable.
- Three-level full factorial design, # of samples = 3^n
  - Impractical for n>5
  - Also unreasonable ratio between number of points and number of coefficients
  - Ex) n=8 → $3^8 = 6561$ samples for 45 coefficients.

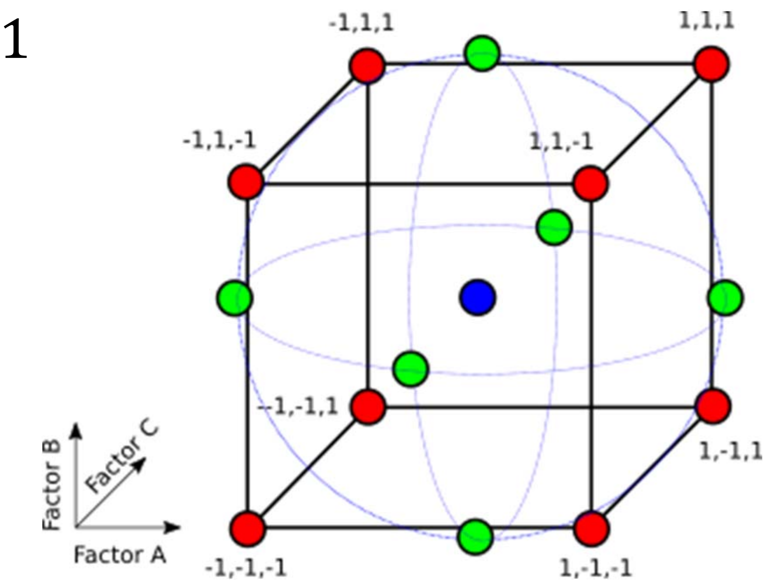


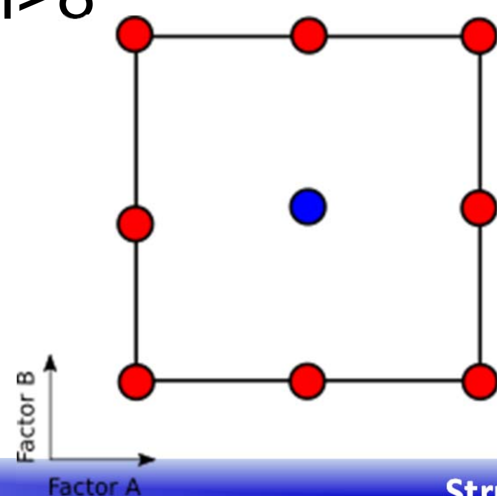
- Rule of thumb is that you want twice as many points as coefficients

Central Composite Design

- Includes 2n vertices, 2n face points plus n_c repetitions of central point
  - # of samples = $2^n + 2n + n_c$
- Can choose $\alpha = 2^{n/4}$ so as to
  - achieve spherical design, achieve rotatability (prediction variance is spherical)
  - Stay in box (face centered) FCCCD, $\alpha = 1$
- Still impractical for n>8

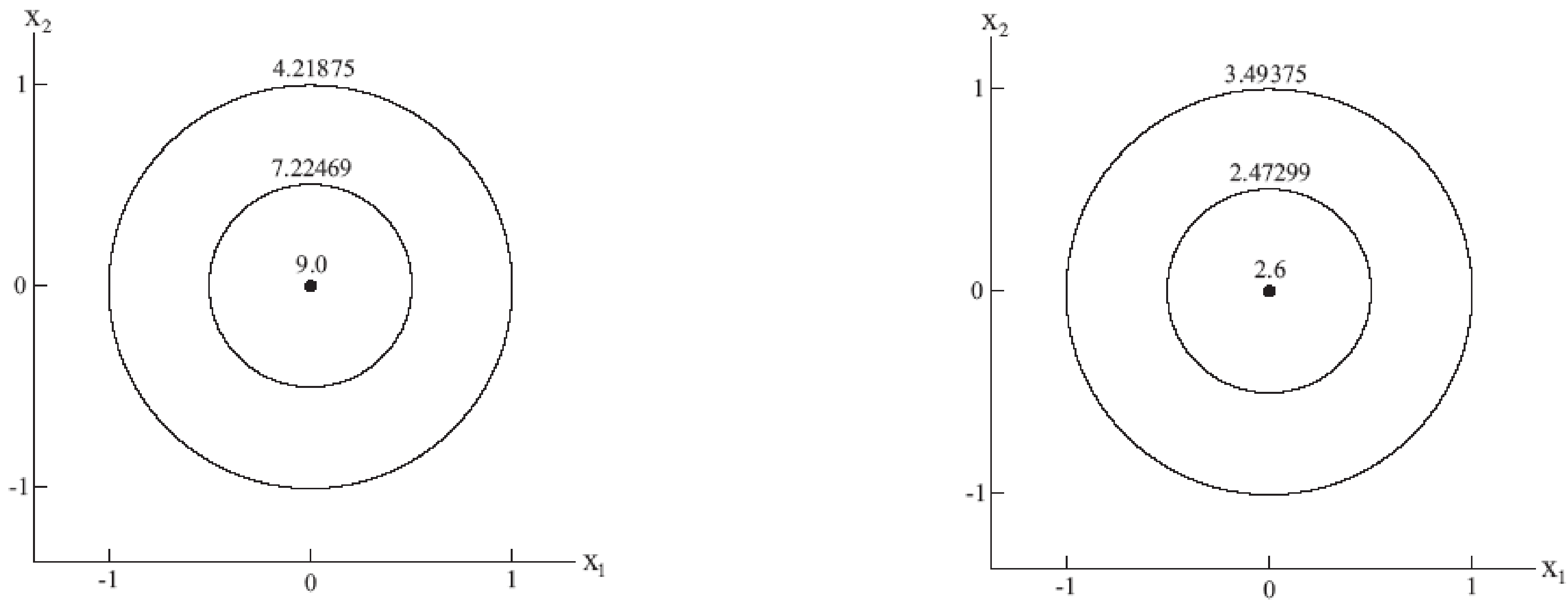






Repeated observations at origin

- Unlike linear PRS, prediction variance is high at origin.
  - Repetition at origin decreases variance and improves stability (uniformity).
  - Repetition also gives an independent measure of magnitude of noise.



Contours of prediction variance for spherical CCD design (no repetition)

With five repetitions, the max prediction variance is significantly reduced and greatly improve the uniformity.

Optimal point selection

- Useful for irregular domain or reduced # of samples
- Select the best n_y points out of large candidates that minimize variance or modeling error
- Variance-based optimal design
  - D-optimality: minimum coefficient variance
  - A-optimality: minimum sum of individual variance

$$\text{Minimize} \quad \sum_{i=1}^{n_\beta} \left| M_{ii}^{-1} \right|$$

  - G-optimality: minimum prediction variance

$$\text{Minimize} \quad v(\mathbf{x}) = \frac{n_y V[\hat{y}]}{\sigma^2} = n_y \xi^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \xi$$

D-optimal design

- Maximizes the determinant of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ to reduce the volume of uncertainties about the coefficients.
- Recall $\Sigma_b = \hat{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} = \hat{\sigma}^2 \dfrac{1}{|\mathbf{X}^{\mathrm{T}}\mathbf{X}|}\mathrm{adj}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) \Rightarrow \Sigma_b \sim \dfrac{1}{|\mathbf{X}^{\mathrm{T}}\mathbf{X}|}$
- Moment matrix

$$\mathbf{M} = \frac{1}{n_y}\mathbf{X}^{\mathrm{T}}\mathbf{X}, \qquad |\mathbf{M}| = \frac{1}{(n_y)^{n_\beta}}|\mathbf{X}^{\mathrm{T}}\mathbf{X}|$$

- Coefficient of variation $\sim \dfrac{1}{|\mathbf{M}|}$
- Maximize $|\mathbf{M}|$ → combinatorial optimization problem
- Finding D-optimal design in higher dimensions is a difficult optimization problem often solved heuristically

Ex) D-optimal design

- For $\hat{y} = b_1 x_1 + b_2 x_2$ with 2 data points (0,0) and (1,0), find the optimum third data point (p,q) in the unit square.
- We have

$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ p & q \end{bmatrix} \quad \mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1+p^2 & pq \\ pq & q^2 \end{bmatrix} \quad \det(\mathbf{X}^T\mathbf{X}) = q^2$$

- arg max $|\mathbf{X}^T\mathbf{X}|$ → $q = 1$
- So that the third point is (p,1), for any value of p
- D-optimal design permits any number of points
- Leads to asymmetrical designs
- Time consuming and not converge well for large n_y

Ex) A-optimal design

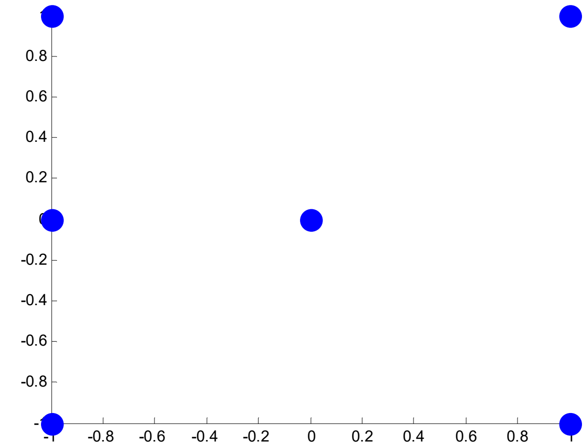
- Repeat the previous example with A-optimality design

$$\mathbf{M}^{-1} = \frac{9}{q^2}\begin{bmatrix} q^2 & -pq \\ -pq & 1+p^2 \end{bmatrix}, \quad \sum_{i=1}^{2} \left| M_{ii}^{-1} \right| = 9\left(1 + \frac{1+p^2}{q^2}\right)$$

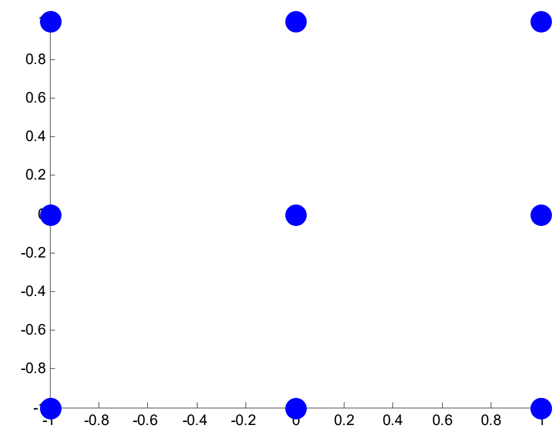
- Minimize $\sum \left| M_{ii}^{-1} \right|$ at $p^2 = 0$ and $q^2 = 1$
- → $(p, q) = (0,1)$

Ex) D-optimality DOE Matlab

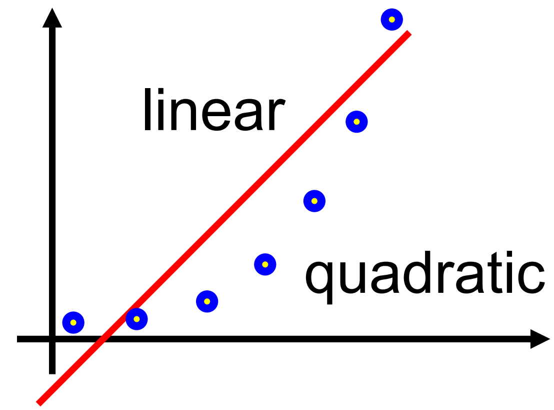
```
%With 6 points:
ny=6;nbeta=6;
[dce,x]=cordexch(2,ny,'quadratic');
>> dce'
    1     1    -1    -1     0     1
   -1     1     1    -1    -1     0
scatter(dce(:,1),dce(:,2),200,'filled')
>> det(x'*x)/ny^nbeta → ans = 0.0055
```




```
%With 12 points:
ny=12;
[dce,x]=cordexch(2,ny,'quadratic');
>> dce'
   -1     1    -1     0     1     0     1    -1
       1     0    -1     1
    1    -1    -1    -1     1     1    -1    -1
       0     0     0     1
scatter(dce(:,1),dce(:,2),200,'filled')
>> det(x'*x)/ny^nbeta → ans =0.0102
```



Minimum bias design



- Bias error: modeling error
- 2nd design moment $M_{\mu} = \frac{1}{V}\int_V \xi\xi^{\mathrm{T}} dV$
  - Approximated 2nd moment $\hat{M}_{\mu} = \frac{1}{n_y}\sum_{k=1}^{n_y} \xi(\mathbf{x}_k)\xi(\mathbf{x}_k)^{\mathrm{T}}$
- Minimum bias design
  - Find sample locations that make the approximated moment to be the same as design moments

$$\hat{M}_{\mu} = M_{\mu}$$

- Min variance design: pts toward boundary, may not low bias
- Min bias design: pts toward centroid, lower variance

Ex) Minimum bias design

- Select 4 pts for $y = x_1^2 + x_2^2$ using linear PRS $\boldsymbol{\xi} = \{1, x_1, x_2\}^{\mathrm{T}}$

$$\xi\xi^T = \begin{bmatrix} 1 & x_1 & x_2 \\ x_1 & x_1^2 & x_1x_2 \\ x_2 & x_1x_2 & x_2^2 \end{bmatrix} \qquad M_\mu = \frac{1}{V}\int_V \xi\xi^{\mathrm{T}} dV = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{bmatrix}$$

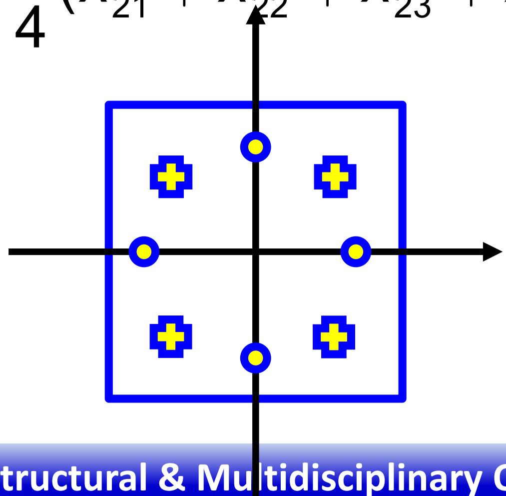
$$\hat{M}_\mu = \frac{1}{n_y}\sum_{k=1}^{n_y} \xi(\mathbf{x}_k)\xi(\mathbf{x}_k)^{\mathrm{T}} \Rightarrow$$

$$\hat{M}_{11} = \frac{1}{4}(1+1+1+1) = 1$$

$$\hat{M}_{22} = \frac{1}{4}(x_{11}^2 + x_{12}^2 + x_{13}^2 + x_{14}^2) = \frac{1}{3}$$

$$\hat{M}_{33} = \frac{1}{4}(x_{21}^2 + x_{22}^2 + x_{23}^2 + x_{24}^2) = \frac{1}{3}$$

- Pic 4 points with symmetry
  - Case 1: $(\pm a, 0), (0, \pm a)$
  - Case 2: $(\pm b, \pm b)$



Ex) Minimum bias design *cont.*

- Case 1: $\hat{M}_{22} = \hat{M}_{33} = \dfrac{1}{4}(a^2 + a^2) = \dfrac{1}{3} \quad \Rightarrow \quad a = \sqrt{\tfrac{2}{3}} = 0.8165$

$$\therefore (\pm 0.8165, 0), (0, \pm 0.8165)$$

$$\mathbf{X} = \begin{bmatrix} 1 & -a & 0 \\ 1 & a & 0 \\ 1 & 0 & -a \\ 1 & 0 & a \end{bmatrix}, \quad \mathbf{y} = \begin{Bmatrix} a^2 \\ a^2 \\ a^2 \\ a^2 \end{Bmatrix}, \quad \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{Bmatrix} 4a^2 \\ 0 \\ 0 \end{Bmatrix}$$

$$\hat{y}(x_1, x_2) = b_1 + b_2 x_1 + b_3 x_2 = a^2 = \frac{2}{3} \qquad \text{Constant!}$$

- Case 2: $\hat{M}_{22} = \hat{M}_{33} = \dfrac{1}{4}(b^2 + b^2 + b^2 + b^2) = \dfrac{1}{3} \quad \Rightarrow \quad b = \sqrt{\tfrac{1}{3}} = 0.5774$

$$\therefore (\pm 0.5774, \pm 0.5774) \quad \Rightarrow \quad \hat{y}(x_1, x_2) = \frac{2}{3}$$

- Fitting with full factorial design: $\hat{y}(x_1, x_2) = 2$
  – Min bias design is better than full factorial design!

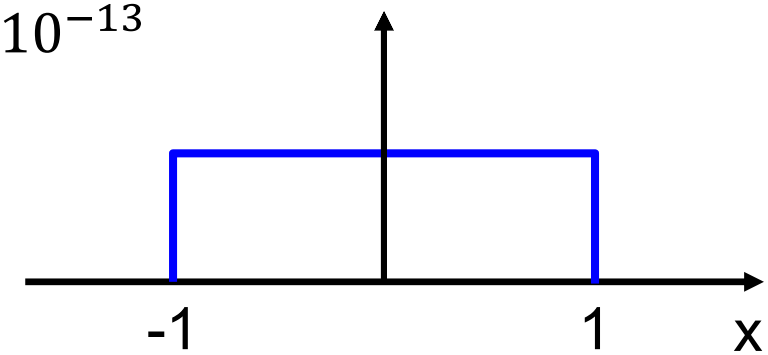
Space-Filling DOEs

- DOE for noisy data tend to place points on the boundary
- When the error in the surrogate is due to unknown functional form, space filling designs are more popular.
  - use values of variables inside range instead of at boundaries
- Space-filling DOE is appropriate only for low-dimension
  - For 10 dimensional space, need 1024 points to have one per orthant.
- Space-filling DOEs
  - Monte Carlo sampling (MCS)
  - Latin hypercube sampling (LHS) uses as many levels as points
  - Orthogonal arrays (OA)

Monte Carlo sampling (MCS)

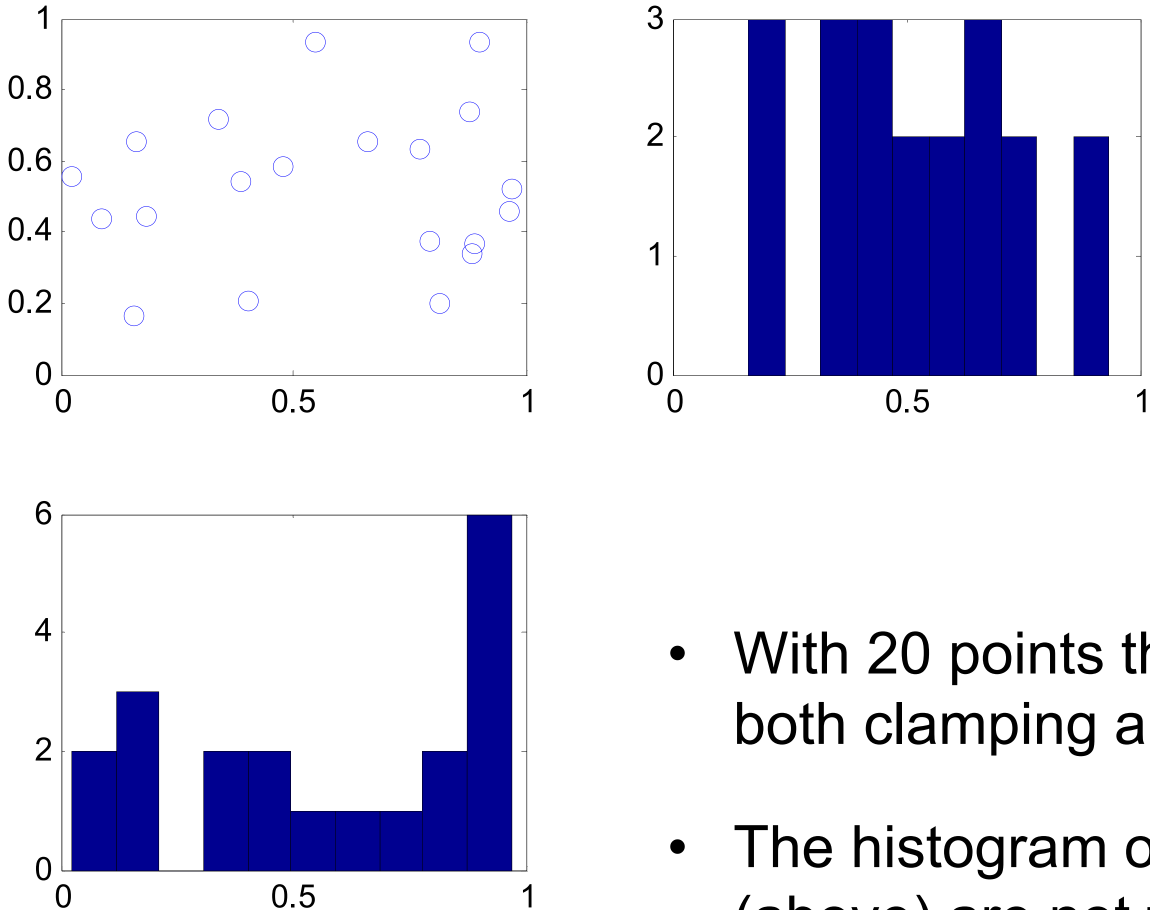
- Regular, grid-like DOE runs the risk of deceptively accurate fit, so randomness appeals.
- Given a region in design space, we can assign a uniform distribution to the region and generate sample points
  – MSC ~ Sampling according to probability distribution
- It is likely, though, that some regions will be poorly sampled
- In 5-dimensional space, with 32 sample points, what is the chance that all orthants will be occupied?

$$\frac{31}{32} \times \frac{30}{32} \times \cdots \frac{1}{32} = 1.8 \times 10^{-13}$$



Ex) MCS

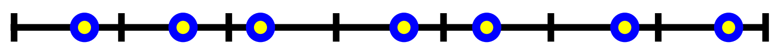
- General 20 samples of $(x_1, x_2) \in (0,1)$ and plot marginal histogram



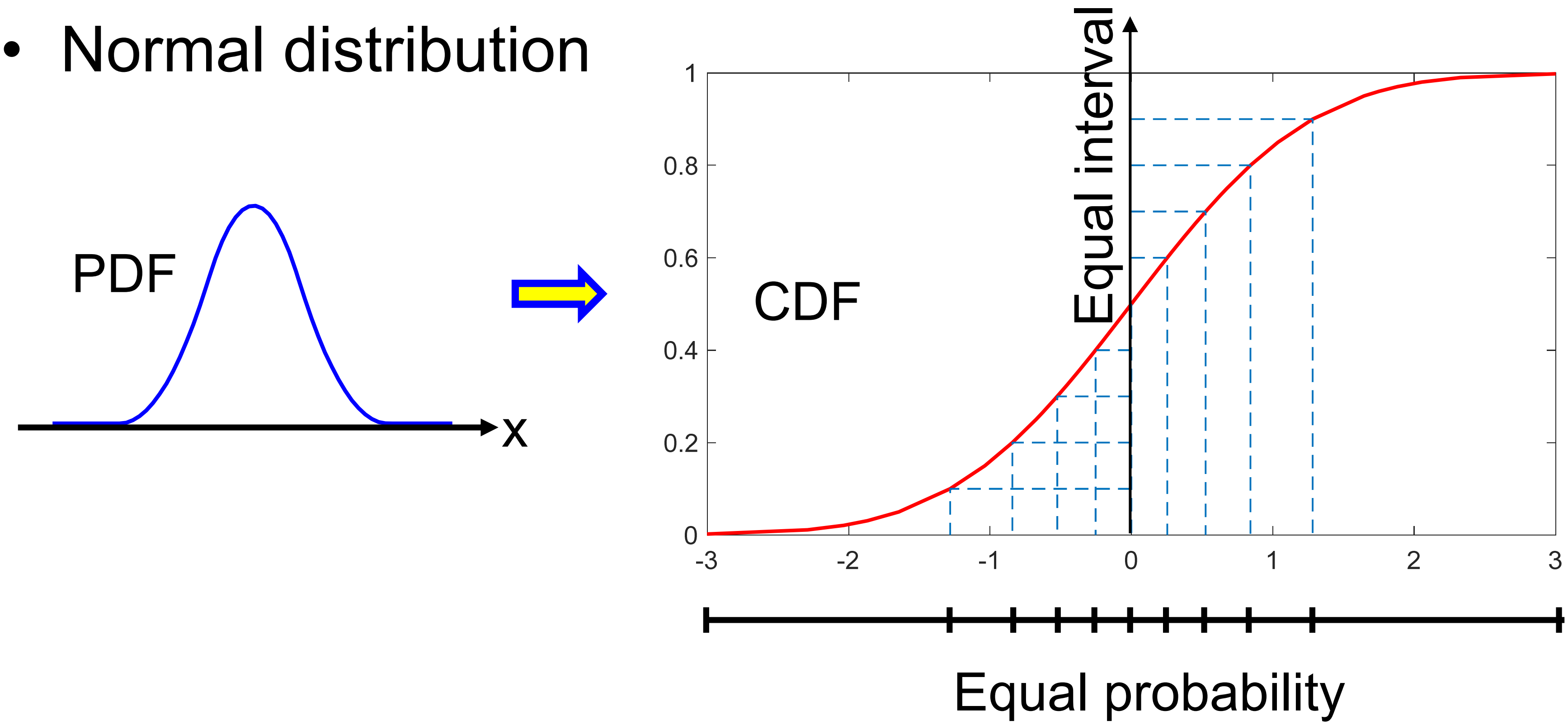
- With 20 points there is evidence of both clamping and holes
- The histogram of x_1 (left) and x_2 (above) are not that good either.

Latin hypercube sampling (LHS)

- Each variable range divided into n_y equal probability intervals. One sample at each interval.
  - Choose one sample randomly within the interval
- Uniform distribution

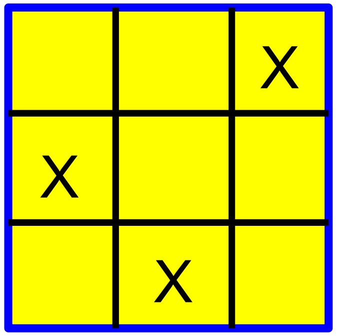


- Normal distribution



LHS procedure

- m designs, n samples (intervals)
- Create $n \times m$ matrix
- Each column: permutation of 1, 2, …, n
- Ex) 2 designs, 3 samples (levels)



$$\begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 2 \end{bmatrix}$$

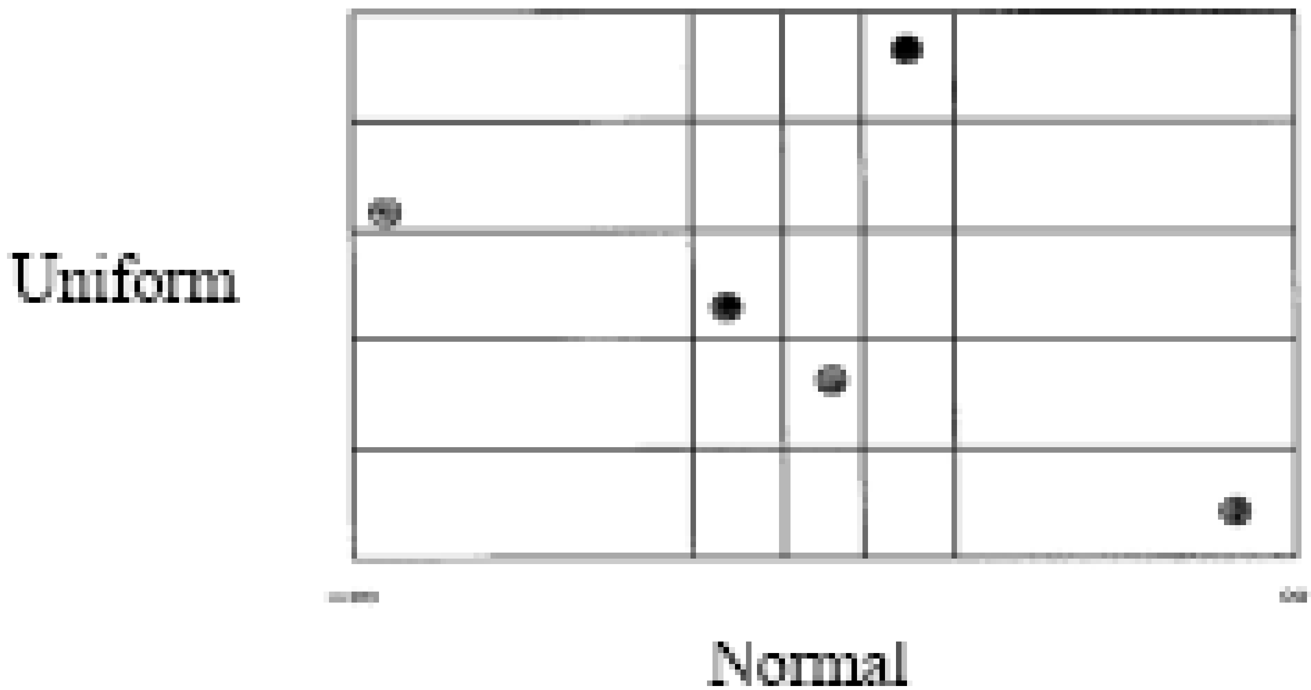
- Choose location randomly within an interval

Ex) LHS

- x_1: normal distribution, x_2: uniform distribution
- Choose 5 sample using LHS

Chosen intervals

x_1	x_2
1	2
2	3
3	4
4	1
5	5



(Courtesy of Wyss and Jorgensen 1998)

Latin Hypercube definition matrix

- For n points with m variables: m by n matrix, with each column a permutation of 1,…,n
- Examples

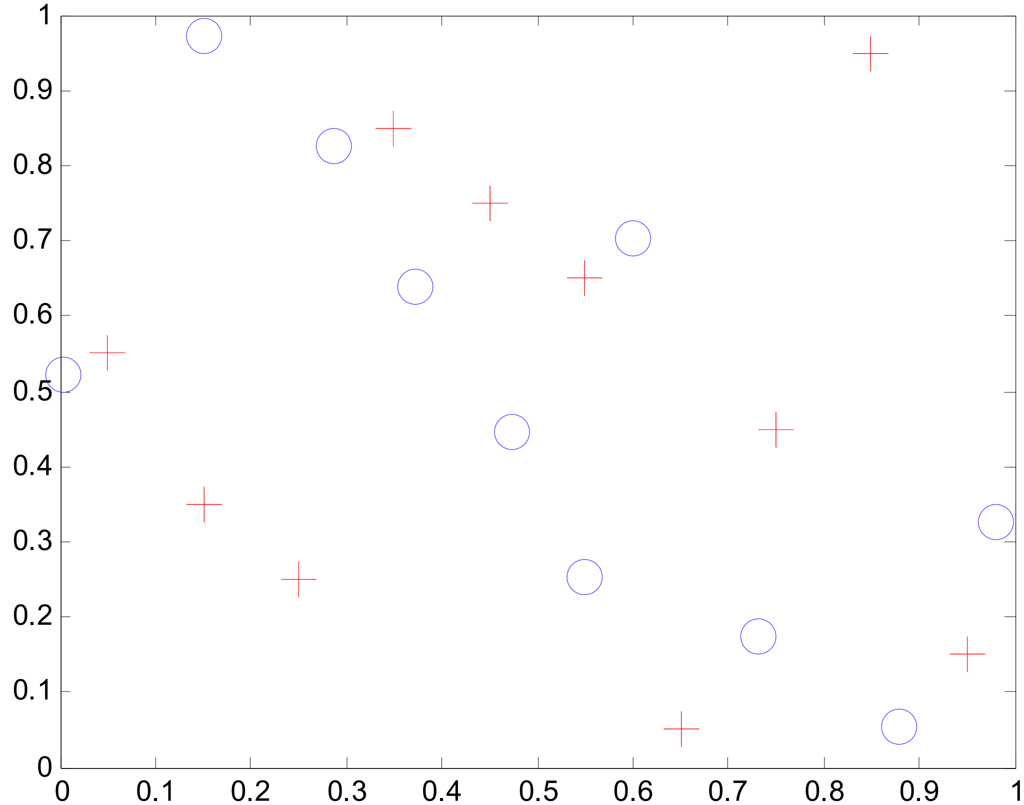
$$\begin{bmatrix} 1 & 3 \\ 2 & 1 \\ 3 & 2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 & 4 \\ 4 & 1 & 2 \\ 3 & 3 & 3 \\ 2 & 4 & 1 \end{bmatrix}$$

- Points are better distributed for each variable, but can still have holes in m-dimensional space.

Improved LHS

- Since some LHS designs are better than others, it is possible to try many permutations. What criterion to use for choice?

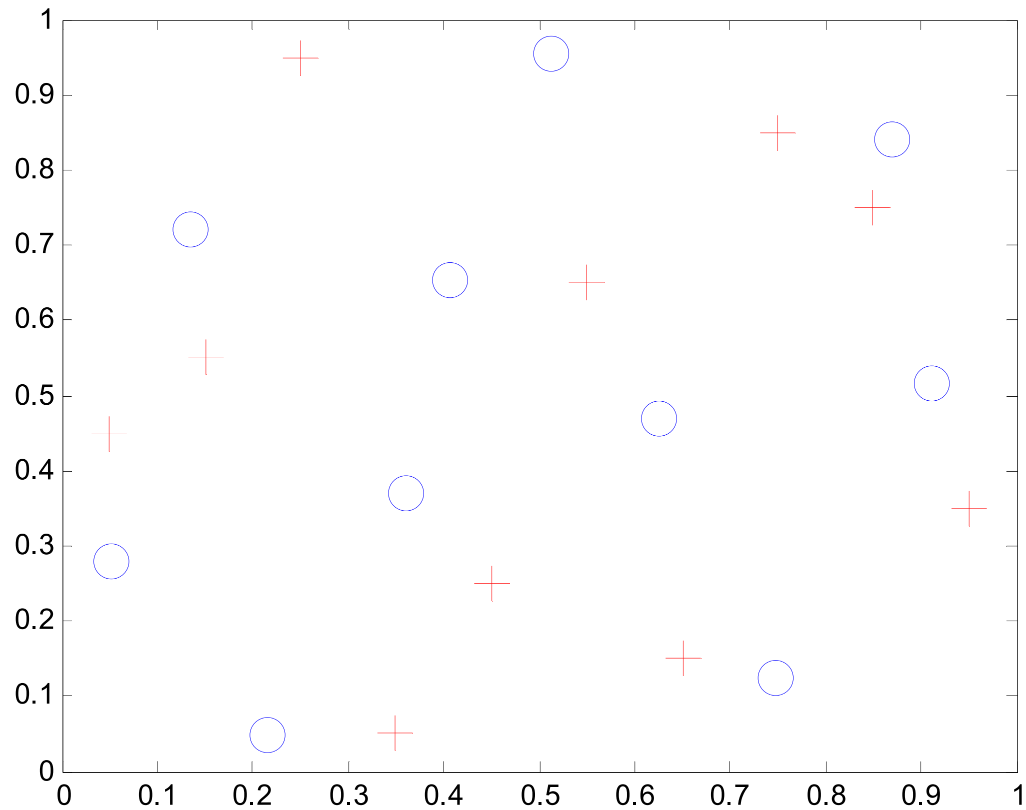
- One popular criterion is minimum distance between points (maximize). Another is correlation between variables (minimize).
  - Matlab lhsdesign uses by default 5 iterations to look for "best" design.
  - The blue circles were obtained with the minimum distance criterion. Correlation coefficient is -0.7.
  - The red crosses were obtained with correlation criterion, the coefficient is -0.055.



More iterations

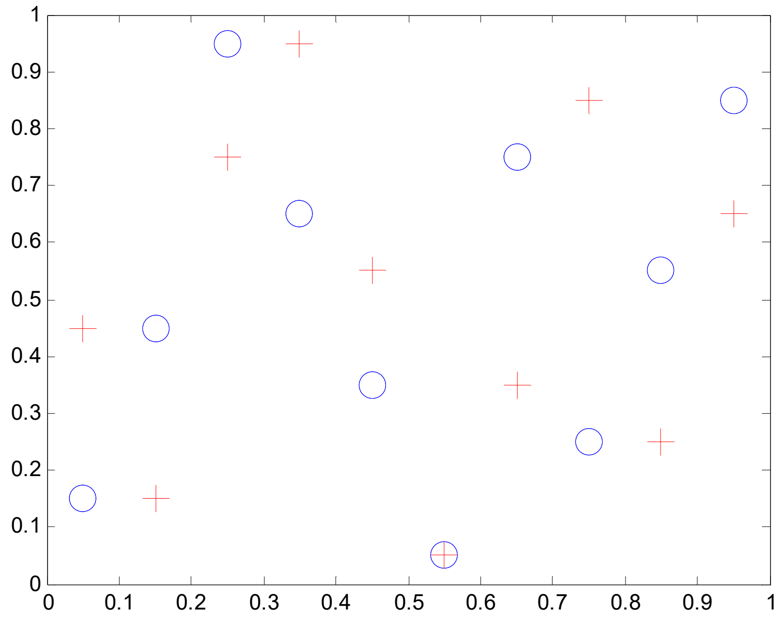
- With 5,000 iterations the two sets of designs improve.
- The blue circles, maximizing minimum distance, still have a correlation coefficient of 0.236 compared to 0.042 for the red crosses.

  - With more iterations, maximizing the minimum distance also reduces the size of the holes better.
  - Note the large holes for the crosses around (0.45,0.75) and around the two left corners.



Reducing randomness further

- We can reduce randomness further by putting the point at the center of the box.
- Typical results are shown in the figure.
- With 10 points, all will be at 0.05, 0.15, 0.25, and so on.



Review of various DOWs

- Questions before DOE
  - Large or small noise
  - Type of surrogate
  - # of variables (n)
  - # of samples (n_y)
  - All the simulations at once or adaptive sampling
- DOE recommendations

1. Low dimension with large noise
   - Full factorial design or CCD (box domain)
   - D-optimal design (irregular domain) + adaptive sampling

Review of various DOWs

2. Low dimension + small noise
   – Minimum bias, LHS, orthogonal arrays (box domain)
   – MCS, optimized distance design (irregular domain)
3. High dimension + large noise
   – Block design, fractional factorial design, CCD (box domain)
   – D-optimal design (irregular domain)
4. High dimension + small noise
   – LHS, optimized distance design