Optimal Time Sampling Strategy in Pharmaceutical Reactions for the Estimation of Accurate DRSM Models

Yachao Dong[1], Christos Georgakis[1], Jason Mustakis[2], Ke Wang[2], Joel Hawkins[2], Jonathan P. McMullen[3] and Kevin Stone[3]

[1]Tufts University, System Research Institute

[2]Pfizer Worldwide R&D

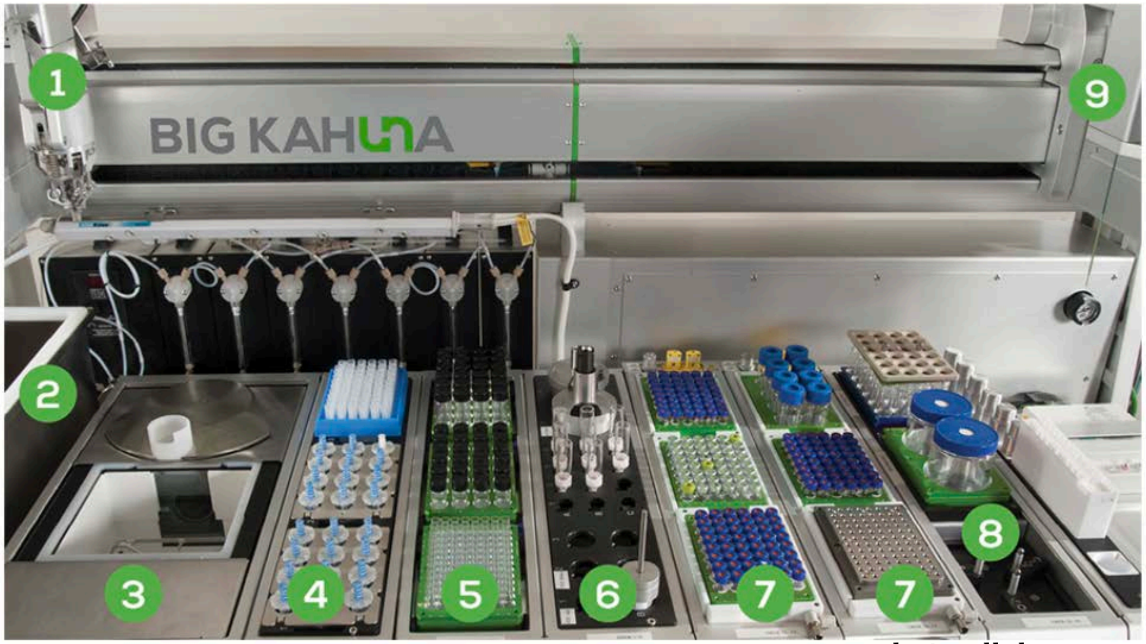
[3] Merck & Co., Inc., Process Research and Development

Nov. 12, 2019, AIChE Annual Meeting, Orlando, FL

Outline

1. Background & Methodology
   - ❑DRSM model
   - ❑Existing time strategies
2. Newly Proposed Strategy
   - ❑Reduction of correlation
   - ❑Equidistant in θ
3. Numerical Results, Comparing
   - ❑Diagonal Dominance
   - ❑Uncertainty Volume

Plethora of Robotic Devices



www.unchainedlabs.com

Big Data of Dynamic Response ➡ Models

Response Surface Methodology (RSM)

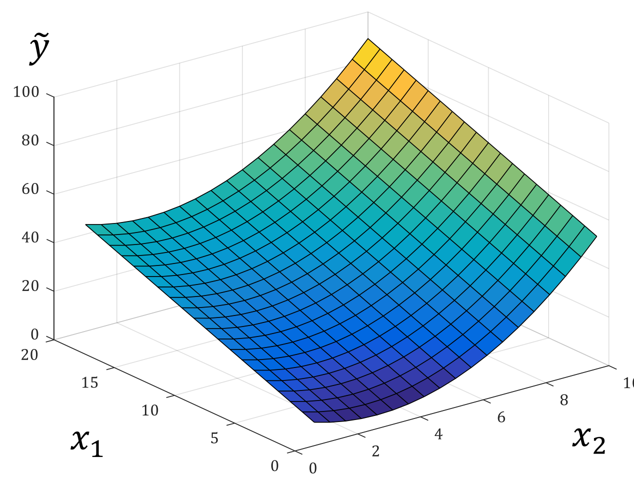
Quadratic form:

$$\tilde{y} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \beta_{ij} x_i x_j + \sum_{i=1}^{n} \beta_{ii} {x_i}^2$$

- $\tilde{y}$: Modeled output
- x_i: Factors of DoE
- β: Coefficient in the model

  Estimated by regression

Not Time-Resolved

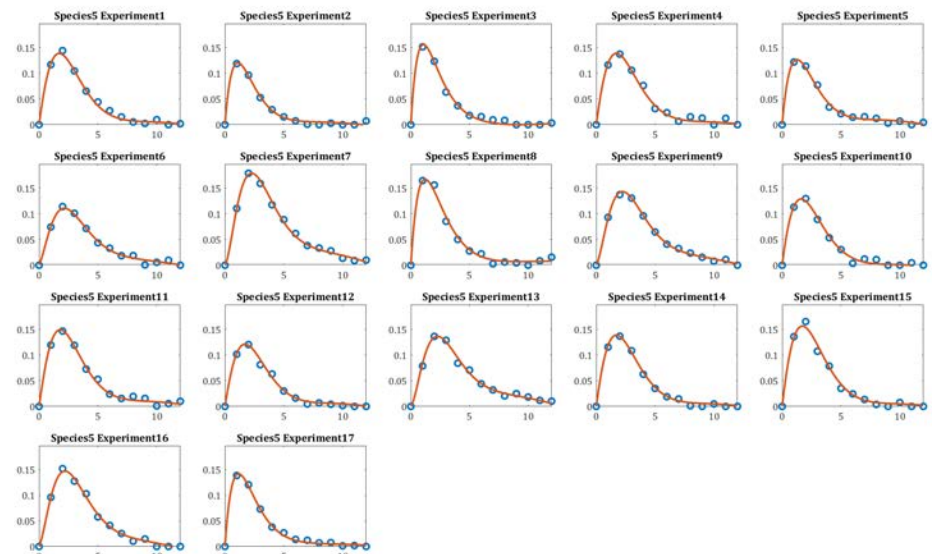


Dynamic Response Surface Methodology (DRSM)

Time as an independent variable:

$$\tilde{y}(t) = \beta_0(t) + \sum_{i=1}^{n} \beta_i(t) x_i + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \beta_{ij}(t) x_i x_j + \sum_{i=1}^{n} \beta_{ii}(t) {x_i}^2$$

One DRSM Model
Each species



Klebanov, Georgakis. *Ind. Eng. Chem. Res.* **2016,** *55* (14), 4022-4034.
Wang, Georgakis. *Ind. Eng. Chem. Res.* **2017,** *56* (38), 10770-10782.
Dong, Georgakis, Mustakis, Hawkins, Lu, Wang, McMullen, Grosser, Stone. *Ind. Eng. Chem. Res.* **2019, ,** *58* (30), 13611-13621.

DRSM Formulation

Latest DRSM Model:

$$\tilde{y}(\theta) = \beta_0(\theta) + \sum_{i=1}^{n} \beta_i(\theta) x_i + \sum_{i=1}^{n} \sum_{j=1}^{i} \beta_{ij}(\theta) x_i x_j$$

Parametrization with Shifted Legendre Polynomials

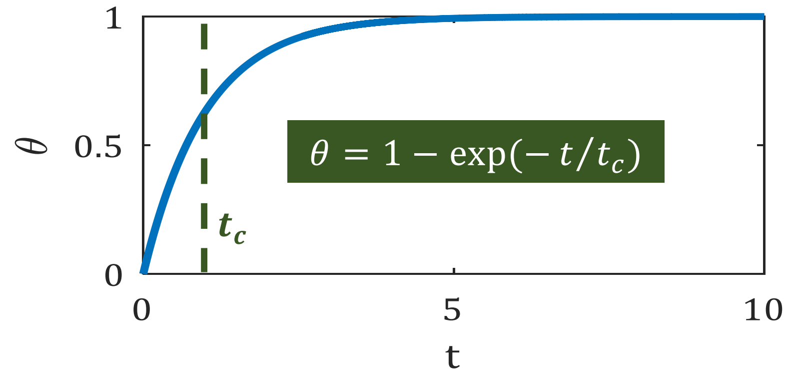
$$\beta_i(\theta) = \sum_{r=0}^{R} \gamma_{i,r} P_r(\theta) \,, \forall i = 0,1, \ldots , n, \ldots$$

$$P_0(\theta) = 1 \quad P_1(\theta) = -1 + 2\theta \quad P_2(\theta) = 1 - 6\theta + 6\theta^2$$

Has been successfully used for:

- ❑Stoichiometry Identification
- ❑Process Optimization

Variable: $t \rightarrow \theta$



Parameters of Model

- Global: R & t_c → BIC criterion
- Local: $\gamma_{i,r}$ → Lasso regression

Add Knowledge-driven Constraints

Existing Time Strategies

- Strategy S1: equidistant in time
  - t=[1, 2, …, 9, 10]
- Strategy S2: first double, later on stable
  - t=[0.06, 0.12, 0.25, 0.5, 1, 2, 4, 6, 8, 10]
- Other strategy in literature*:
  - Sampling time leading to even distribution along concentration
  - Assumes monotonic concentration change

Question to answer:
What is the best strategy for an accurate model?

*Rothenberg, Boelens, Iron, Westerhuis. *Catalysis Today* **2003,** *81* (3), 359-367.

II. Proposed Strategy

Equidistant in $\boldsymbol{\theta}$

- Reduction of Correlation
- Reduction of Estimator Uncertainty

Key: Matrix B

- 2FI DRSM model:

$$y = \beta_0(\theta) + \sum_{i=1}^{n} \beta_i(\theta) x_i + \sum_{i=1}^{n} \sum_{j=i}^{n} \beta_{ij}(\theta) x_i x_j$$

$$\beta_i(\theta) = \sum_{r=0}^{R} \gamma_{i,r} P_r(\theta)\,, \forall i = 0,1, \ldots, n, \ldots$$

- Important matrix:

$$\boldsymbol{M} = \begin{pmatrix} P_0(\theta_1) & P_1(\theta_1) & \cdots & P_R(\theta_1) \\ P_0(\theta_2) & P_1(\theta_2) & \cdots & P_R(\theta_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_0(\theta_N) & P_1(\theta_N) & \cdots & P_R(\theta_N) \end{pmatrix}$$

- For experiment k:

$$\boldsymbol{X}(k) \equiv [\boldsymbol{I}_{R+1} \quad x_1(k)\boldsymbol{I}_{R+1} \quad \ldots \quad x_{n-1}(k)x_n(k)\boldsymbol{I}_{R+1}]$$

- Fisher information matrix:

$$\boldsymbol{F} = \sum_{i=1}^{K} \boldsymbol{X}^T(i) \boxed{\boldsymbol{M}^T \boldsymbol{M}} \boldsymbol{X}(i)$$

$\boldsymbol{B} = \boldsymbol{M}^T \boldsymbol{M}$

F: related to Variance of parameters

- Sampling time affects Covariance through $\boldsymbol{B}$
- Orthogonal DoE + Diagonal $\boldsymbol{B} \rightarrow$ Parameters γ NOT Correlated

Comparative Indicators

$$\boldsymbol{M} = \begin{pmatrix} P_0(\theta_1) & P_1(\theta_1) & \cdots & P_R(\theta_1) \\ P_0(\theta_2) & P_1(\theta_2) & \cdots & P_R(\theta_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_0(\theta_N) & P_1(\theta_N) & \cdots & P_R(\theta_N) \end{pmatrix} \quad \boldsymbol{B} = \boldsymbol{M^T M}$$

Wishes for $\boldsymbol{B}$:

- Non-diagonal elements as small as possible⇨ γ parameters uncorrelated
  - Indicator 1: Diagonal Dominance. $DD = \max_{i \in 1,\ldots,R+1} \frac{\sum_{j \in 1,\ldots,R+1 \setminus i} |Q_{ij}|}{|Q_{ii}|}$
  - $DD < 1 \Leftrightarrow$ diagonal dominant, prefer smaller DD
- Maximize determinant ⇨ minimize overall estimation variance
  - Indicator 2: Uncertainty Volume, $UV = \frac{1}{\sqrt[R+1]{\det(|\boldsymbol{B}|)}}$
  - Prefer smaller UV

New Approach

Equidistant in theta

- ❑$\theta_1 = \theta_X/N, \theta_2 = 2\theta_X/N, \cdots, \theta_N = \theta_X$
- ❑Back-calculate in time $t_1 \cdots t_N$, and round off

$\theta = 1 - \exp(-t/t_c)$

With this approach:

- ❑$N \rightarrow \infty \Rightarrow \boldsymbol{B}$ is diagonal
- ❑N is finite $\Rightarrow \boldsymbol{B}$ is highly diagonal dominant

Need to estimate time constant t_c, based on at least one experiment

- ❑Through simple DRSM formulation
- ❑For monotonical change, $t_c \approx$ slope of time against $\ln(y)$

Examples of Sampling Time (12 samples)

S1: Equidistant in Time

$[t_1, \cdots, t_{12}]$ = [1, 2, ..., 12]

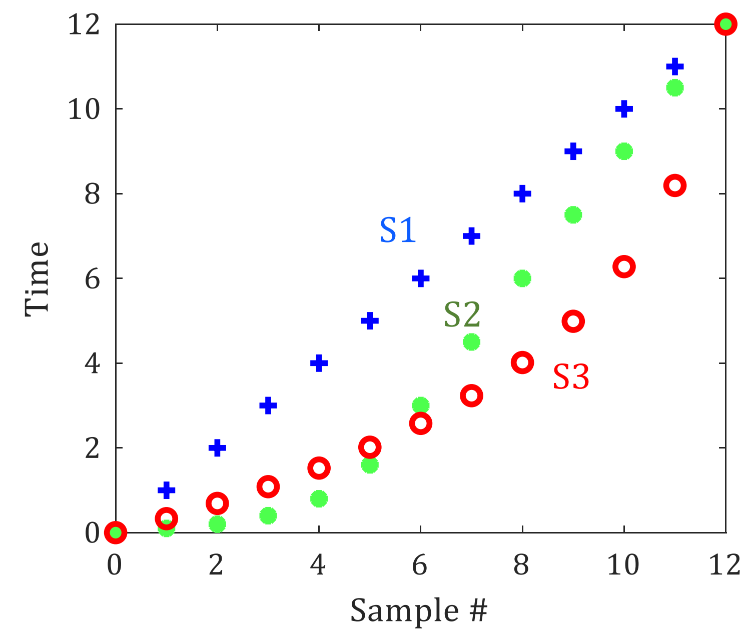
S2: First Double, Later on Stable

$[t_1, \cdots, t_{12}]$=[0.1, 0.2, 0.4, 0.8, 1.6, 3, 4.5, 6, 7.5, 9, 10.5, 12]

S3: Equidistant in Theta

$[t_1, \cdots, t_{12}]$ = [0.3, 0.7 1.1, 1.5, 2.0, 2.6, 3.2, 4.0, 5.0, 6.3, 8.2, 12]

- Assuming $t_c = 4$



III. Numerical Results

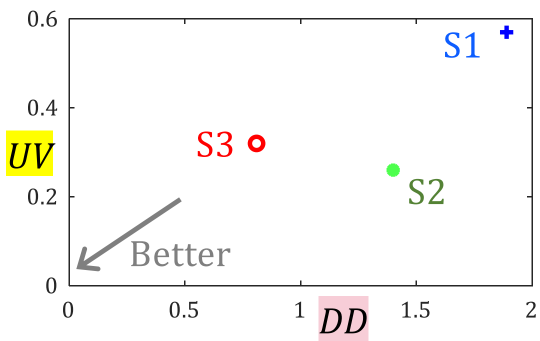
1. **Independent of DoE**
2. **Specific case study based on DoE**

Comparative Indicators (DoE-Independent)

$$M = \begin{pmatrix} P_0(\theta_1) & P_2(\theta_1) & \cdots & P_R(\theta_1) \\ P_0(\theta_2) & P_2(\theta_2) & \cdots & P_R(\theta_2) \\ \vdots & \vdots & \ddots & \vdots \\ P_0(\theta_N) & P_2(\theta_N) & \cdots & P_R(\theta_N) \end{pmatrix} \qquad B = M^T M$$

Indicators

- Diagonal Dominance (DD)
- Uncertainty Volume (UV)
- Prefer strategies of smaller values



Consider the example with R=3 and N=12

Strategy S1: equidistant in time

$$B = \begin{pmatrix} 12 & 5.31 & 1.15 & 0.08 \\ 5.31 & 4.77 & 2.17 & -0.14 \\ 1.15 & 2.17 & 2.16 & 0.78 \\ 0.08 & -0.14 & 0.78 & 1.23 \end{pmatrix}$$

DD=1.89 UV=0.57

Strategy S2: first double, later on stable

$$B = \begin{pmatrix} 12 & 0.56 & 2.96 & -0.91 \\ 0.56 & 5.97 & -0.32 & 1.06 \\ 2.96 & -0.32 & 3.06 & -0.45 \\ -0.91 & 1.06 & -0.45 & 1.73 \end{pmatrix}$$

DD=1.40 UV=0.26

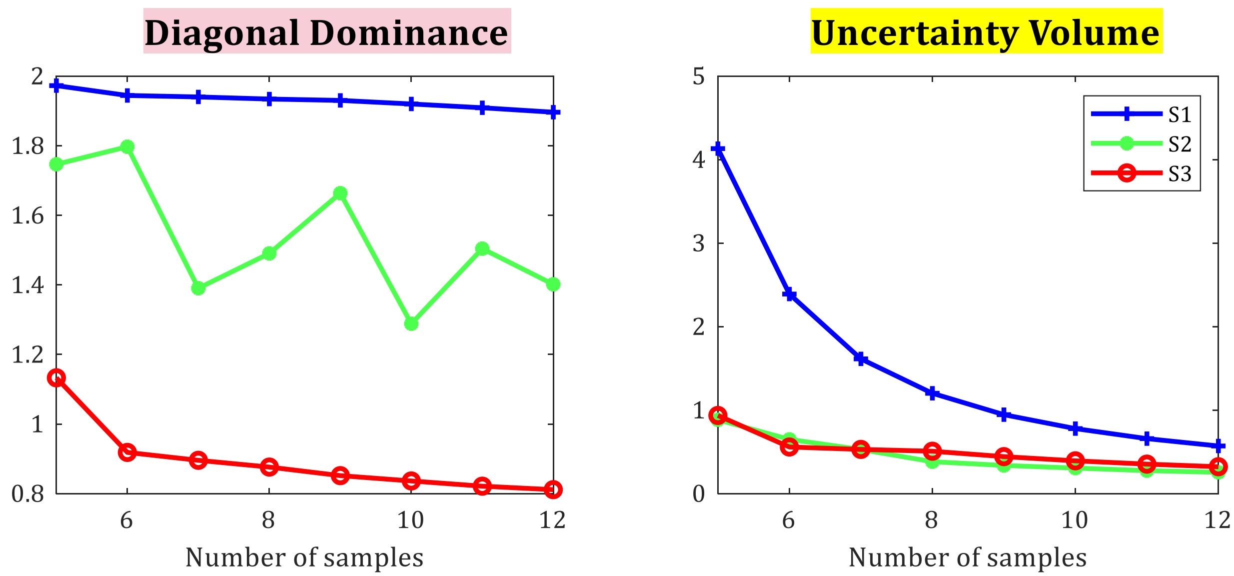
Strategy S3: equidistant in theta

$$B = \begin{pmatrix} 12 & 0.35 & -0.60 & 0.26 \\ 0.35 & 3.59 & 0.29 & -0.57 \\ -0.60 & 0.29 & 1.95 & 0.21 \\ 0.26 & -0.57 & 0.21 & 1.28 \end{pmatrix}$$

DD=0.81 UV=0.32

Varying Number of Samples (R=3)

S1: Equidistant in Time S2: First Double Then Stable S3: Equidistant in Theta



- *DD* **S3** better than **S1** and **S2**
- *UV* **S3** and **S2** are similar, both better than **S1**

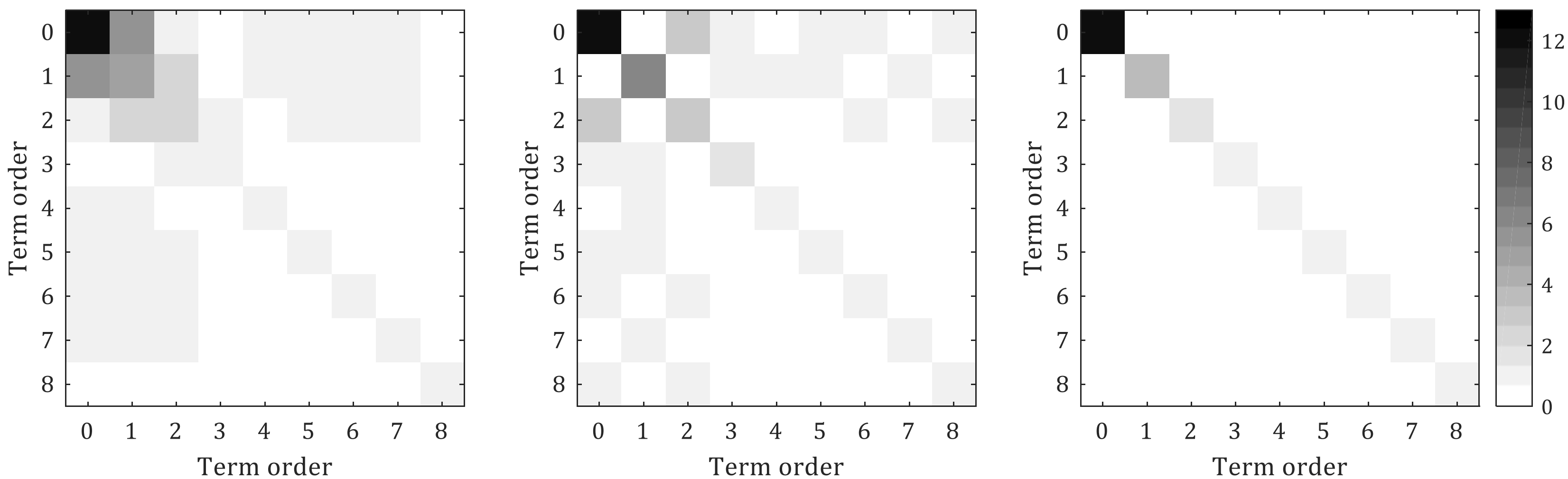
S3: no significant improvement with more than 6 or 7 samples

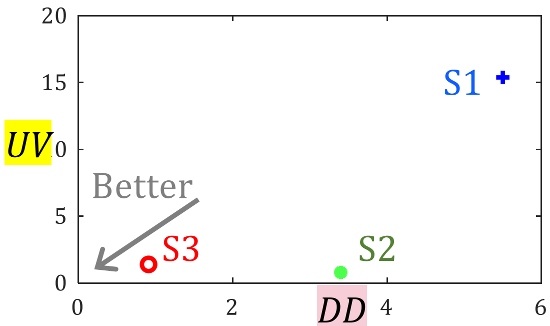
Matrix B of Higher Order (R=8)

Strategy S1
Equidistant in Time

Strategy S2:
First double, then stable

Strategy S3:
Equidistant in Theta





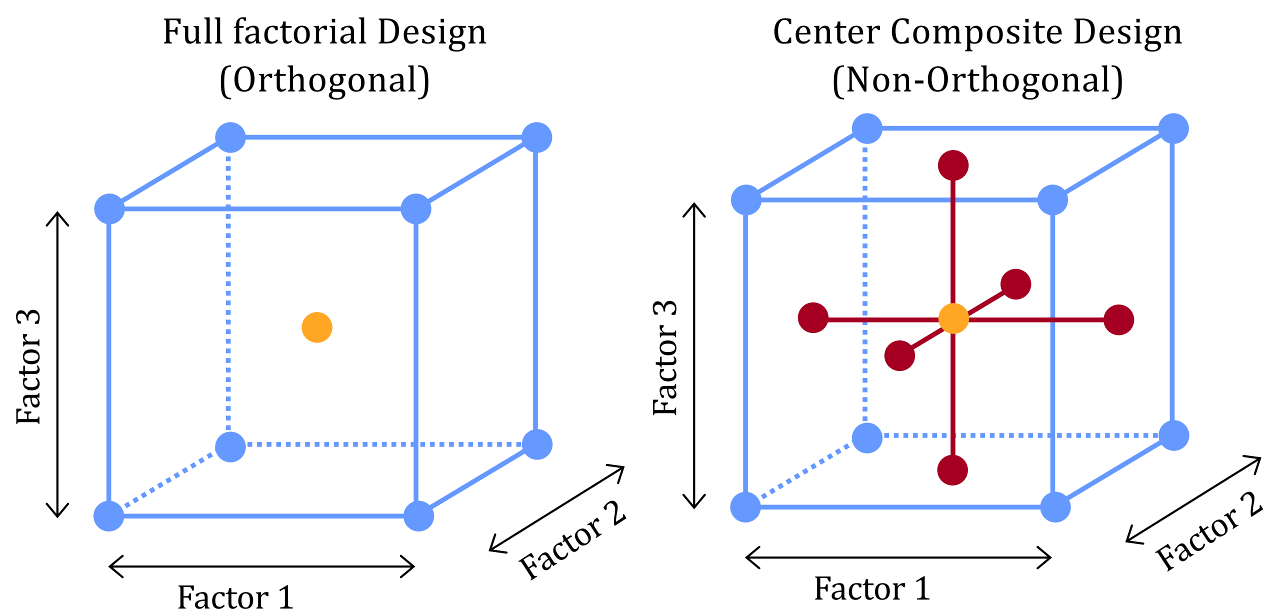
- Diagonal Dominance **S3** is still better
- Uncertainty Volume **S3** is still similar to **S2**, better than **S1**

DoE-Dependent Case Studies

- Simulated case
- 10 species
- 8 linearly independent reactions

- 3-factors
  - Factor 1 temperature 50 ~90 °C
  - Factor 2 concentration of B 0.8 ~1.2 mol/L
  - Factor 3 concentration of D 0~ 2 mol/L

1	A + B ⇆ C + D
2	C → D + E
3	E → F
4	B + D ⇆ G
5	G → D + H
6	A + F → I
7	2A → J
8	B + J → 2E + I

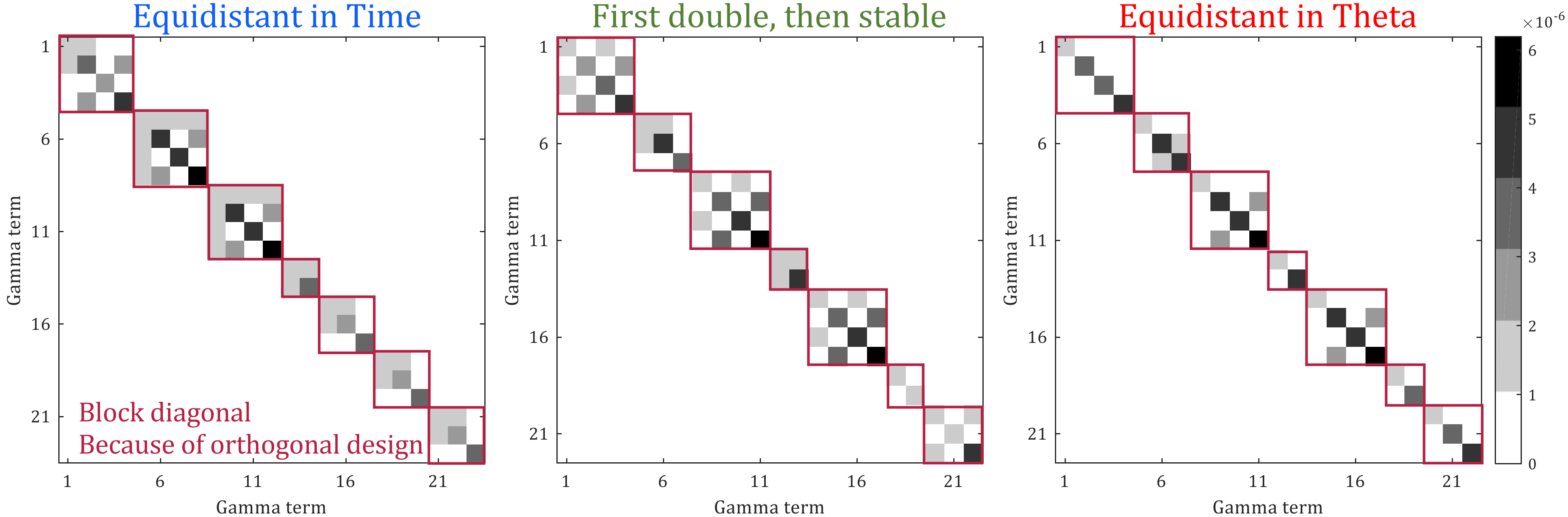


Covariance Matrix (Orthogonal Design)

Strategy S1
Equidistant in Time

Strategy S2:
First double, then stable

Strategy S3:
Equidistant in Theta



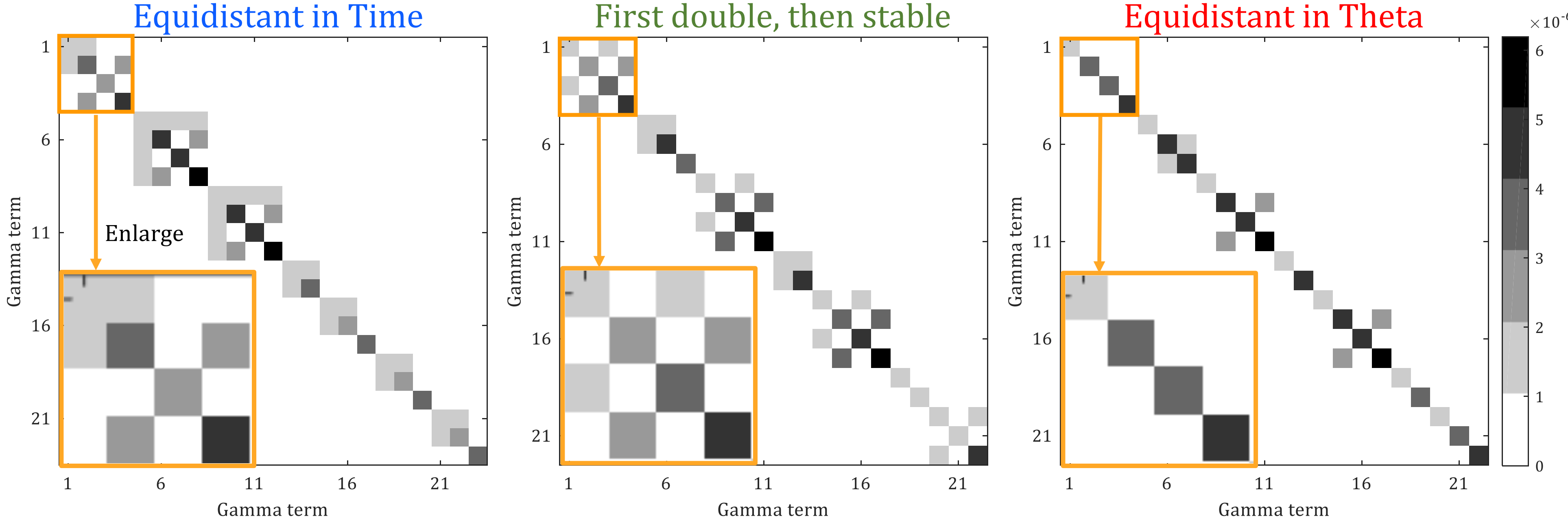
S3 leads to the smallest correlation

Covariance Matrix (Orthogonal Design)

Strategy S1
Equidistant in Time

Strategy S2:
First double, then stable

Strategy S3:
Equidistant in Theta

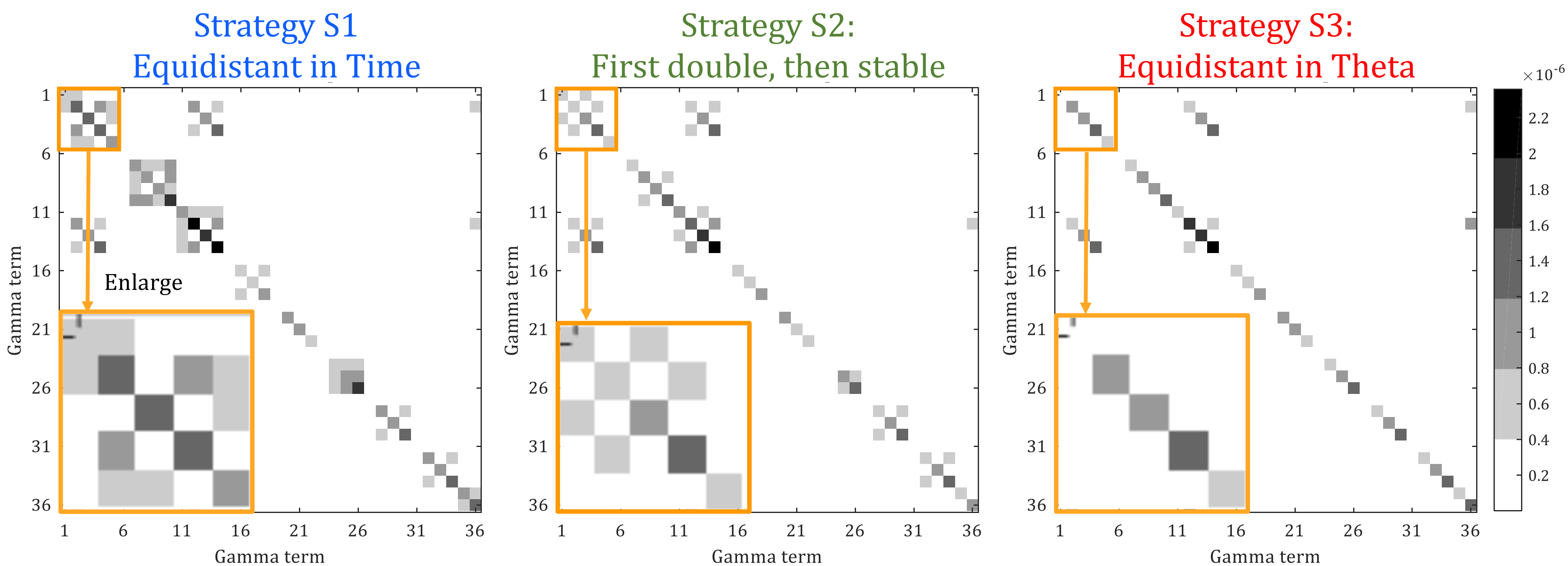


$$y = \beta_0(\theta) + \sum_{i=1}^{n} \beta_i(\theta) x_i + \sum_{i=1}^{n} \sum_{j=i}^{n} \beta_{ij}(\theta) x_i x_j$$

$$\beta_i(\theta) = \sum_{r=0}^{R} \gamma_{i,r} P_r(\theta), \forall i = 0,1, \dots, n, \dots$$

The difference manifests in each block
Representing one beta function

Covariance Matrix (Central Composite Design)



Not Block diagonal:
CCD is not orthogonal

Again, S3 leads to the smallest correlation

Confidence Interval Size

S1: Equidistant in Time S2: First Double Then Stable S3: Equidistant in Theta

Confidence Interval				
Species	Reference (mol/L)	Change in percentage (%)		
	Strategy S1	**Strategy S2**	**Strategy S3**	
	12 samples	12 samples	12 samples	7 samples
1	0.0025	-22	-9	-7
2	0.0036	-1	-5	7
3	0.0023	-4	-12	-12
4	0.0040	-32	-17	-14
5	0.0023	3	-27	4
6	0.0049	-12	-11	13
7	0.0025	-1	-1	8
8	0.0016	5	-4	-16
9	0.0014	0	-16	-8
10	0.0010	-2	-20	4
Average over species	0.0026	**-6**	**-12**	**-2**

- S3: $t_c = 3.4$
  - ❑ Average values of over species
- Confidence interval averaged over:
  - ❑27 experiments, 100 time instants

S3: Smallest Confidence Interval

- ○ 12 Samples: **S3** < **S2** < **S1**
- ○ **S3** (7 samples) ≈ **S1** (12 samples)

Conclusions

- DRSM Model Accurately Predicts Dynamic Response Data
- **We Proposed a New Sampling Strategy**
- **Basic Idea: Equidistant in Theta**
  - ❑Reduces Correlation of Estimated Parameters
  - ❑Reduces Overall Uncertain Volume

Thank you for Your Attention!