cencora
PharmaLex

A Comparative Analysis of Bayesian Sampling Methodologies for Design Space Identification in Quality by Design (QbD) Approach

DENIZ AKINC ABDULHAYOGLU

JEAN-FRANÇOIS MICHIELS

NCS - 2024

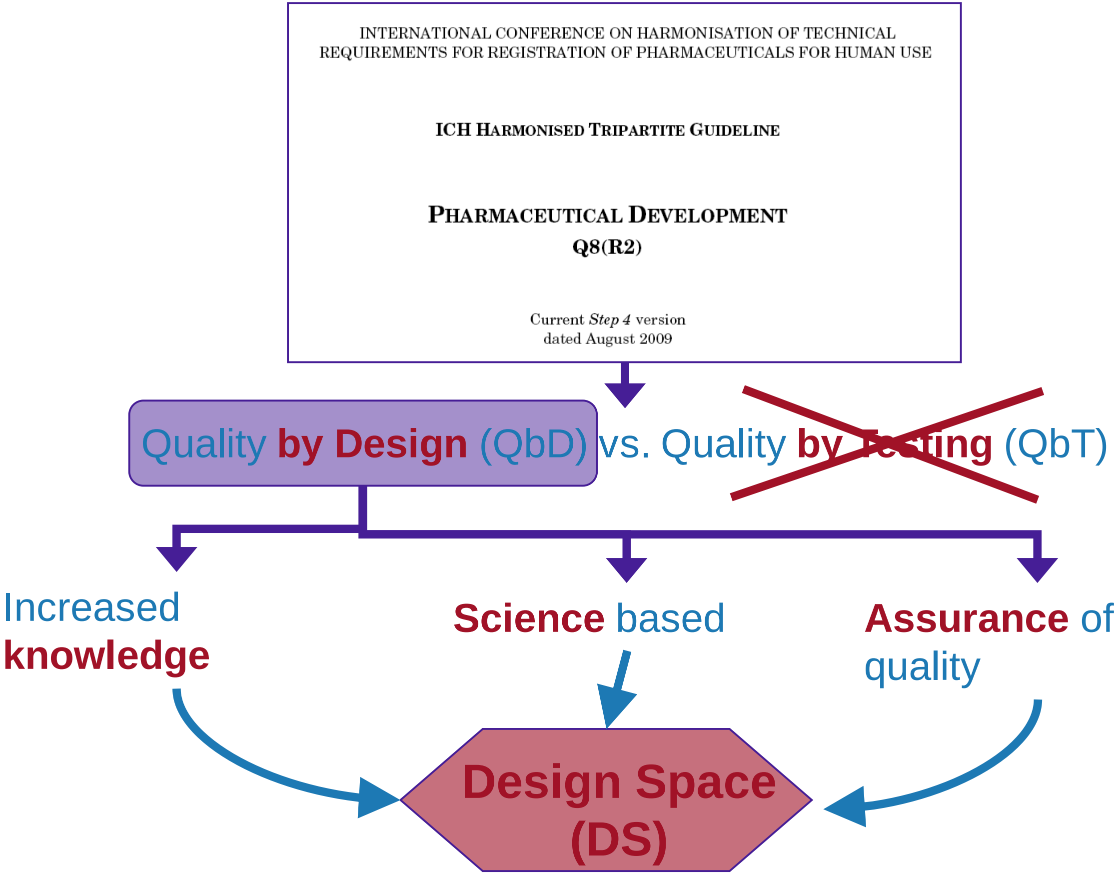
Wiesbaden

September 25-27, 2024

Disclaimer

This presentation is intended to communicate PharmaLex's capabilities which are backed by the author's expertise. However, [PharmaLex GmbH] and its parent, Cencora, Inc., strongly encourage readers to review the references provided with this presentation and all available information related to the topics mentioned herein and to rely on their own experience and expertise in making decisions related thereto as the presentation may contain certain marketing statements and does not constitute legal advice.

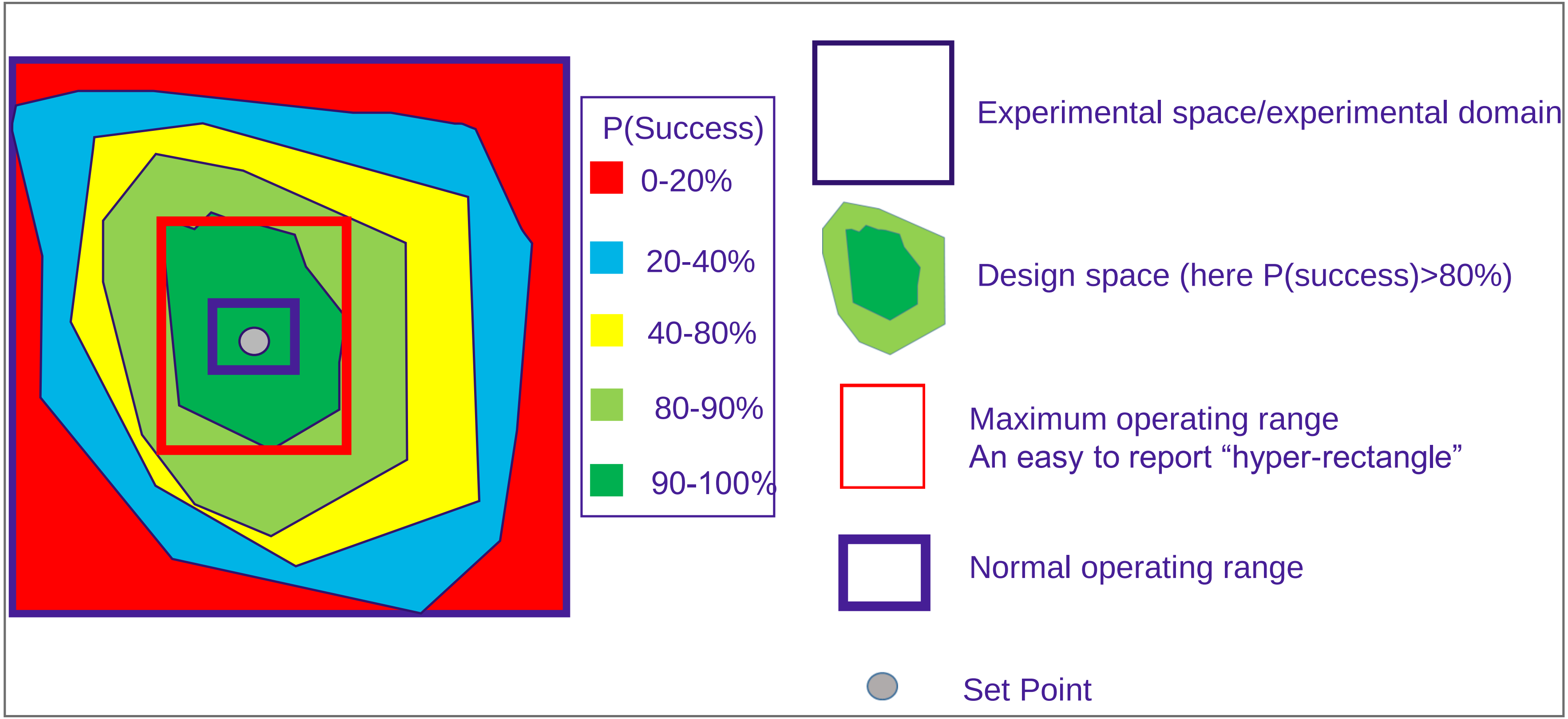
Design Space Identification in QbD approach



Design Space Identification in QbD approach

Data → Fit Bayesian Model → Generate a large grid using a space filling design algorithm (~1000 -5000 points) → Using posterior distributions and for each row of the grid, compute the probability of success to meet specifications → Plot results using a pair plot → Visual Identification of the Design Space

Design Space - 2 factors example



Design Space Identification in QbD approach

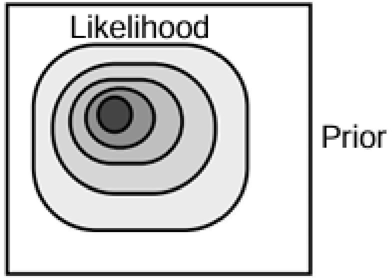
Data → Fit Bayesian Model → Generate a large grid using a space filling design algorithm (~1000 -5000 points) → Using posterior distributions and for each row of the grid, compute the probability of success to meet specifications → Plot results using a pair plot → Visual Identification of the Design Space

Background

$$\underbrace{\Pr(\boldsymbol{\Theta}|\mathbf{D},\mathrm{M})}^{\text{Posterior}} = \frac{\overbrace{\Pr(\mathbf{D}|\boldsymbol{\Theta},\mathrm{M})}^{\text{Likelihood}}\ \overbrace{\Pr(\boldsymbol{\Theta}|\mathrm{M})}^{\text{Prior}}}{\underbrace{\Pr(\mathbf{D}|\mathrm{M})}_{\text{Evidence}}}$$

Bayes' Theorem

Motivation: Sampling Posterior

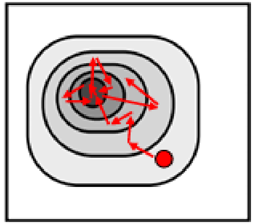


Sampling directly from the likelihood $\mathcal{L}(\Theta)$ is **hard**.

MCMC: Solving a Hard Problem once.
(Markov Chain Monte Carlo)

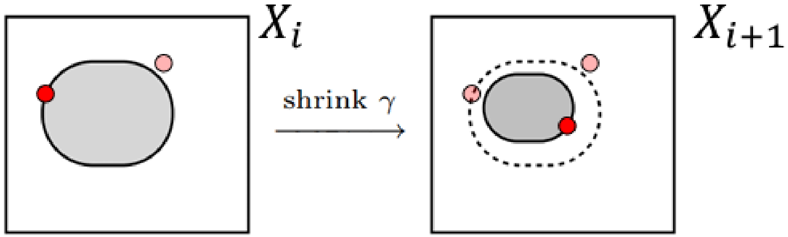
Source: Speagle, Joshua S. (2021)

Motivation: Sampling Posterior



MCMC: Solving a Hard Problem once.
vs
Nested Sampling: Solving an Easier Problem many times.

Sampling $\boxed{\text{uniformly within}}$ bound $\mathcal{L}(\boldsymbol{\Theta}) > \lambda$ is **easier**.



Source: Speagle, Joshua S. (2021)

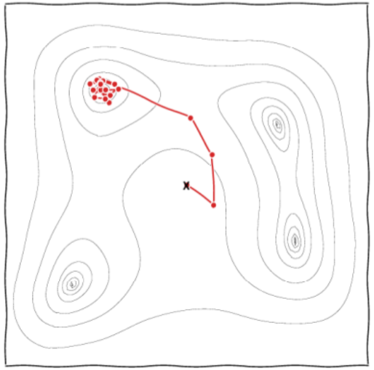
Comparison of Nested Sampling vs. MCMC Sampling

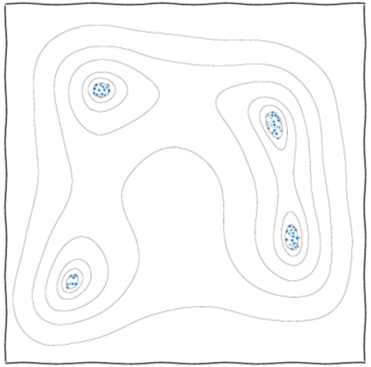
MCMC

- Single “walker”
- Explores posterior
- Fast, if proposal matrix is tuned

Nested Sampling

- Ensemble of “live points”
- Scans from prior to peak of likelihood
- Might be slower, no tuning required





Source: Handley, Will (2023)

Advantages and disadvantages of Nested Sampling

Advantages

- Does not need gradients
- Ensemble sampler
- Multimodal exploration
- Can characterize complex uncertainties in real-time
- Can allocate samples much more efficiently in some cases
- Very parallelizable
- Possesses well-motivated stopping criteria (Skilling 2006; Speagle 2020)
- Can help perform model selection

Disadvantages

- Can not use gradients as “naturally” as HMC.
- Implementations require a prior transform.
- Slow (but steady) and runtime sensitive to size of prior.
- Sampling is more involved.

The main goal of using gradients is to improve dimensionality scaling / reliability

Examples on Statistical Inference

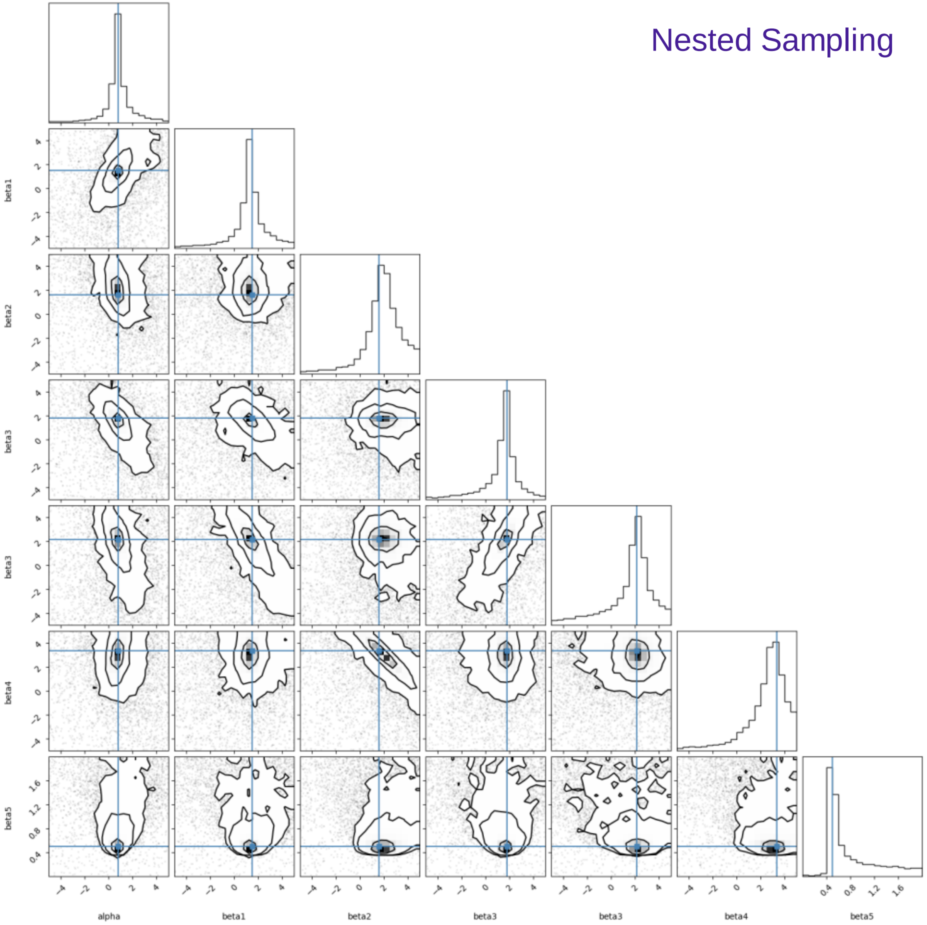
alpha	beta_1	beta_2	beta_3	beta_4	beta_5
0.811829	1.513266	1.590718	1.791521	2.132137	3.339008

$$Y_i = \alpha - \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 - \beta_4 X_1 X_3 + \beta_5 X_2 X_2 + \varepsilon_i$$

Computation Duration with dynesty <2.5min

Computation Duration with PyMC3 <9min

Nested Sampling



Example on Design Space Computation

For a simple case study with two design variables, d := (d1; d2) and a single CQA, s.

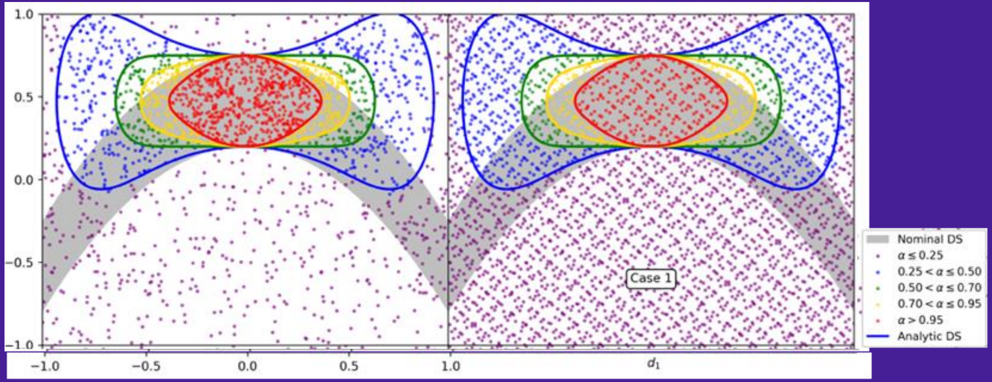
$$s = \theta d_1^2 + d_2$$

with the model parameter θ. The goal is to characterize the probabilistic DS inside the knowledge space K := [-1; 1]^2 imposed by the following CQA limits:

$$0.2 \leq s \leq 0.75$$

$$\mathcal{D}_{\text{nom}} := \{\boldsymbol{d} \in [-1,1]^2 : 0.20 \leq d_1^2 + d_2 \leq 0.75\}$$

$$\theta \sim \mathcal{N}(0,1)$$



$$\mathcal{D}_\alpha := \left\{\boldsymbol{d} \in [-1,1] : \mathbb{P}\left[0.20 \leq \theta d_1^2 + d_2 \leq 0.75 \,\middle|\, \theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta)\right] \geq \alpha\right\}$$
$$= \left\{\boldsymbol{d} \in [-1,1] : \operatorname{erf}\left(\frac{0.75 - \mu_\theta d_1^2 - d_2}{\sqrt{2}\sigma_\theta d_1^2}\right) - \operatorname{erf}\left(\frac{0.20 - \mu_\theta d_1^2 - d_2}{\sqrt{2}\sigma_\theta d_1^2}\right) \geq 2\alpha\right\}$$

Reliability value	Samples drawn by standard Monte Carlo (Algorithm 1)	Samples drawn by nested sampling (Algorithm 2)
Uncertainty scenario: $\theta \sim \mathcal{N}(0,1)$		
$0.95 \leq \alpha$	247	500
$0.70 \leq \alpha < 0.95$	139	231
$0.50 \leq \alpha < 0.70$	172	181
$0.25 \leq \alpha < 0.50$	666	418
$\alpha < 0.25$	2,025	532
Total	3,249	1,862

Source: Kusumo, et al (2019)

Conclusions

To summarize, Nested Sampling is an attractive algorithm framework for Bayesian inference and design space computation because

- it explores the parameter space globally,
- it handles multi-modal distributions and phase transitions well,
- it initializes and terminates at a well-defined point without cumbersome supervision, and
- it leverages efficient strategies from Bayesian parameter estimation for generating replacement proposals during the DS characterization.
- it is effective for larger DS characterization with a handful of process parameters, in the presence of a complex dynamic model and realistic model uncertainty.
- it can even be competitive of MCMC with optimization methods relying on process flexibility concepts, even in low-dimensional DS characterization problems

Further Studies

- High Dimensional experiments:

- Advanced kinetic models based on ordinary differential equations

- Multimodal and mixture models

Thank you!

Jean-François Michiels

Associate Director
Cencora PharmaLex
Data Strategy & Quantitative Sciences

Deniz Akinc

Senior Manager
Cencora PharmaLex
Data Strategy & Quantitative Sciences

References

1. ICH Q8 (R2) Pharmaceutical development - Scientific guideline

2. Skilling, J. (2006). Nested sampling for general Bayesian computation.

3. Kusumo, K. P., Gomoescu, L., Paulen, R., García Muñoz, S., Pantelides, C. C., Shah, N., & Chachuat, B. (2019). Bayesian approach to probabilistic design space characterization: A nested sampling strategy. *Industrial & Engineering Chemistry Research*, *59*(6), 2396-2408.

4. Kusumo, Kennedy P., et al. "Nested Sampling Strategy for Bayesian Design Space Characterization." *Computer Aided Chemical Engineering*. Vol. 48. Elsevier, 2020. 1957-1962.

5. Speagle, J. S. (2020). dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, *493*(3), 3132-3158.