A Comparative Analysis of Design Space Optimization Strategies for identifying High-Volume Hypercubes, including a novel algorithm

Sorry,
I rebranded. 😅

Still same subject.
See next ... !! 😉

sanofi

2024-09-27

Speeding up Design space exploration by method of moments approximation.

Is it feasible ?

Yannick Van Haelst

CMC-Biologics Statistics
Data Sciences
Sanofi Global CMC development

2024-09-27

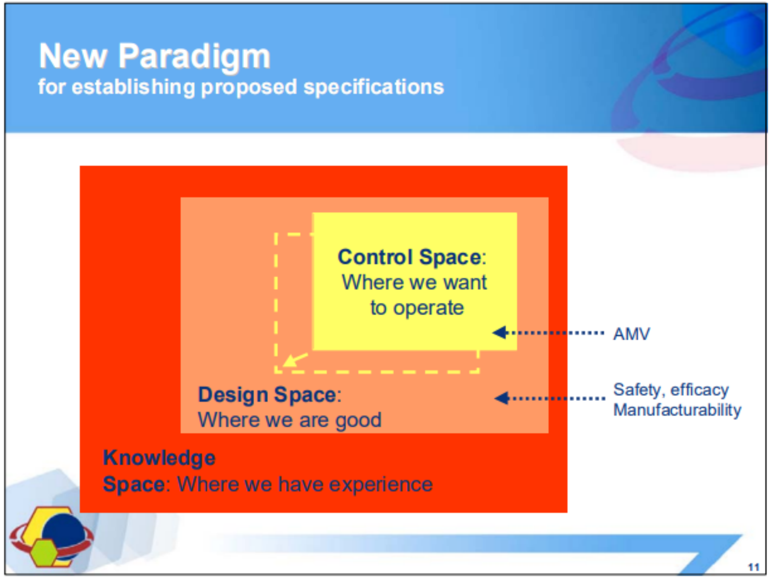
Terminology

Design space(Ds): defined by the multidimensional combination and interaction of input variables (e.g., material attributes) and process parameters that have been demonstrated to provide assurance of quality. (ICH Q8)

Multivariate acceptable ranges(MAR): within multivariate acceptable ranges, any combination of input parameters of a unit operation yields the desired product quality and process performance. (Kunzelmann et al., 2024)→ Hypercube

Edge of failure = hull separating within spec from out of spec. Or a p(within spec) threshold.

Control space = Control Space refers to the specific, defined operating conditions (ranges) within the Design Space where the process is actually controlled during routine production. It represents a narrower subset of the Design Space. (Could be a set of in-process-control limits). (Bhutani et al, 2004)



Excerpt from:
Chen C (2006) Implementation of ICH Q8 and QbD—an FDA perspective. PharmaForum Yokohama, June. https://www.nihs.go.jp/drug/PhForum/Yokohama060609-02.pdf (accessed on 2024-SEP-04)

Current challenges/solutions when exploring Design Space

- When in full control of process input parameters, the problem is easy:

  - Build a model
  - Consider model uncertainty
  - Use statistical inference to find the edge of failure
  - Find a rules set f(inputs, rules) that validate the input settings (the control space).

  → Often simplified to a list of low-high settings, defining a 'hypercube' within the design space. *(like JMP 17.2 Design Space explorer)***

  → Best hypercube (MAR) can be found without the need for a hyper-dimensional grid by means of nested optimization:

  Outer optimization: find largest volume
  $\prod UCL_i - LCL_i * weight \quad (U_{pper}/L_{ower}\, Control\ Limit\ of\ input\ i)$

  for which (Inner optimization):
  `optim(max(p(failure) | in cube) < threshold`

- When process input parameters are variable (i.e. day to day variability, raw material, environmental conditions, …) the problem is hard!!

  Need to integrate out model prediction with respect to routine input variability, ideally proportional to their rate of occurrence

  Current approach is simulation based: **very tedious!§**

  - Classic way:
    - Build a grid in k dimensions.
      (r-points per dimension gives rise to r^k *points)*
    - E(model, inputs)* at each grid point. A.k.a. simulate inputs and perform model prediction n times, then take the average.
    - Delineate the hull or find inscribed hypercube (as before) where p(failure) is lower than a threshold. (and use some ***interpolation technique for course grids***).

** E(.) = expectation function =* $\iiint_{-\infty}^{+\infty} model \mid random\ inputs$

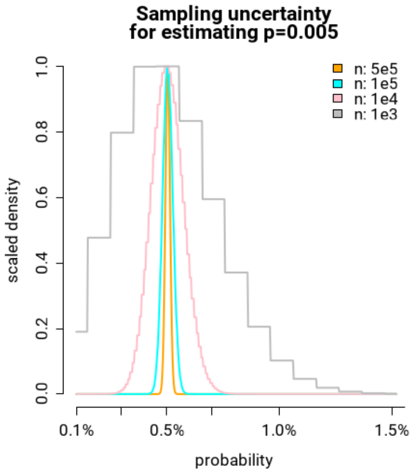
*** For JMP approach, see Lancaster L.(2023)*
§ For calculation time examples, see Taillefer V. & Nasir O. (2020)

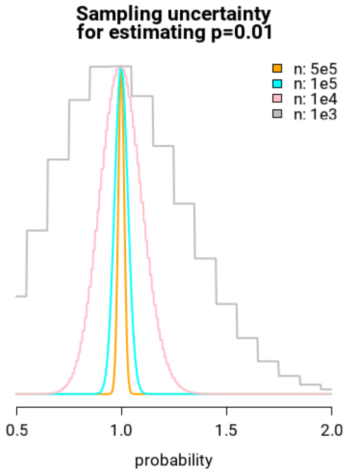
Current challenges & solutions: the double curse

When process input parameters are variable ***(stochastic of nature)***

Curse 1: sampling the tails of a distribution is simulation-expensive.

- In a 'quality by design' setting the edge of failure will be defined with very small risks rates.
  I.e. p(out of spec) < 1%, 0.1% , 0.27% (ideally for 6σ)
- Binomial theorem shows high sampling rates are required to have sufficient precision on those small p-values.





generating n=100'000 (1E5) samples to capture sufficient certainty around 0.05% risk is not a luxury

Workaround 1: **adaptive sampling:** no need to sample expensively everywhere inside the knowledge space. Can be risk-based using binomial confidence intervals as function of current n and $E(p)$:

`stop if` $P(E(p_{failure}), n_{current} < threshold) > \beta$

β= confidence level
Alt. naming: α $(= 1 - \beta)$ reliability risk

Workaround 2: sample a prediction/confidence/tolerance interval and put confidence level on the simulated intervals. This is not the same as the joint distribution! The idea is to take like 95% of the prediction intervals when simulating inputs (sampling for 5% instead of 0.5% on the joint is less expensive). *Like in MODDE 13*

Current challenges & solutions

➢ <u>When process input parameters are variable</u> ***(stochastic of nature)***

Problem is **2 x cursed:**

Curse 2: curse of dimensionality. (Note: also problematic when input factors are fixed but estimates at the points are less expensive)

Example:

	n factors						
grid size	2	3	4	5	6	7	8
16	256	4096	65536	1048576	16777216	2.68E+08	4.29E+09
32	1024	32768	1048576	33554432	1.07E+09	3.44E+10	1.1E+12
64	4096	262144	16777216	1.07E+09	6.87E+10	4.4E+12	2.81E+14
Central Composite Design*	13	19	29	47	81	147	277

= Supported by Modde 13

<u>Known workarounds</u>

- Use space filling design on a 'number of points' budget *(! Mind: budget might be too small for a good estimate)*
- Rejection sampling like in MCMC, focalizing on the design space or edge of failure hull. See *Kusomo et al., 2020* combining rejection sampling for sampling points (curse 2) with a nested adaptive sampling at the point (curse 1).
- Define meta-model, then seek an optimal experimental plan to fit the model on the samples and substitute tedious further simulation by the meta-model for Ds exploration. See *Oberleitner et al., 2024* using a 2nd order response surface 'meta'-model (RSM) on a central-composite design (CCD) *.

Our question:

Full grid

or

Sparse grid (CCD)

(adaptive) Monte-Carlo simulation

(per grid point)

Build Meta model

Define / plot Design space contours

Perform nested optimization for Hyper-cube (MAR)

Usually both are required!

➔ *Hypercube for control limits.*
➔ *Contour plots to explore and validate*

'bypass' simulation and metamodel by a method of moments closed form approximation for the Joint Distribution

Is it possible ?

Method of moments approximation, assumptions

Restricted to:

$Z_{\mathrm{n}} = [1, x_1, x_2, \ldots, x_k, x_1x_2, x_1x_3, \ldots, x_ix_j, \ldots, x_1^2, x_2^2, \ldots, x_k^2]$ (n terms)

$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{(k+1)} x_1 x_2 + \cdots + \beta_{(k+0.5(k(k-1)))} x_{\mathrm{k}} x_{\mathrm{k}-1} + \beta_{(k+0.5(k(k-1)+1)} x_1^2, + \cdots + \beta_{(k+0.5(k(k-1)+k)} x_k^2$

$x_1 \ \ldots x_k \sim N(\mu_i, sigma_i^2)$ are independent random normal

- Interaction and Quadratic are small compared to main effects $(\beta_0 \ldots \beta_k)$ > $\beta_{interaction}, \beta_{quadratics}$
- There are sufficient main terms in the model and their coefficients are the major contributors
- By central limit theorem the joint distribution should approximate a normal distribution, even when the distribution of individually summed terms are not.

Important notes

- Calculation will be exact in the 1st and 2nd moment even when assumptions do not hold
- Deviation from approximation is by missing solution for 3rd and 4th moment of the joint distribution. I.e. treated as if zero like in a Normal distribution.
- Deviation from the approximation can be checked -> take a corner point, simulate and check distributional properties.

1st and 2nd moments for the approximation

Response Surface Model (RSM)

Define:

$x_1, x_2, \dots, x_k \sim \mathcal{N}(\mu_i, \sigma_i^2)$ (independent normal random variables)

RSM terms :

$Z_n = [1, x_1, x_2, \dots, x_k, x_1x_2, x_1x_3, \dots, x_ix_j, \dots, x_1^2, x_2^2, \dots, x_k^2]$ (n terms)

Expectation:

$\hat{Z}_n = [1, \mu_1, \dots, \mu_k, \mu_1\mu_2, \mu_1\mu_3, \dots, \mu_i\mu_j, \dots, \mu_1^2 + \sigma_1^2, \dots, \mu_k^2 + \sigma_k^2]$

Variance $\boldsymbol{\Sigma}_{n\times n} = \mathbf{Var}(\hat{\boldsymbol{Z}})$:

$\mathrm{Var}\,(1) = 0$

$\mathrm{Var}\,(x_i) = \sigma_i^2$

$\mathrm{Var}\,(x_i^2) = 2\sigma_i^4 + 4\mu_i^2\sigma_i^2$

$\mathrm{Var}\,(x_ix_j) = \mu_i^2\sigma_j^2 + \mu_j^2\sigma_i^2 + \sigma_i^2\sigma_j^2$

$\mathrm{Cov}\,(x_i, x_i^2) = 2\mu_i\sigma_i^2$

$\mathrm{Cov}\,(x_ix_j, x_ix_k) = \mu_i\sigma_j^2 + \mu_i\sigma_k^2$

$\mathrm{Cov}\,(x_ix_j, x_kx_l) = 0$ (distinct indices)

Predictor function

Define:

Model coefficients :

$\beta \sim \mathcal{N}(\beta, \mathrm{RMSE}^2(X'X)^{-1})$ $\qquad \frac{RMSE^2}{sigma^2} \sim \frac{\chi^2(dfe)}{dfe}$

Predictor function :

$\mathbb{E}(y) = \hat{Z}'\beta$

Variance:

Model error $\mathrm{Var}(y) = \beta'\Sigma\beta + RMSE\left(\mathrm{tr}\left((X'X)^{-1}\Sigma\right) + \hat{Z}'(X'X)^{-1}\hat{Z}\right)$

Prediction error $\mathrm{Var}(y) = \beta'\Sigma\beta + RMSE^2\left(1 + \mathrm{tr}\left((X'X)^{-1}\Sigma\right) + \hat{Z}'(X'X)^{-1}\hat{Z}\right)$

Approx. deg. freedom:

$\beta'\Sigma\beta$ has df $= \infty$ (under approximation of $\hat{z} \sim \mathrm{MVN}(\hat{Z}_n, \Sigma)$)

Using Welsh-Sattherthwaite

$$df_{approx} = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{\infty} + \frac{V_2^2}{dfe}} = \frac{(V_1 + V_2)^2}{\frac{V_2^2}{dfe}}$$

Where:

$V_1 = \beta'\Sigma\beta$ and $V_2 = RMSE^2\left(1 + \mathrm{tr}\left((X'X)^{-1}\Sigma\right) + \hat{Z}'(X'X)^{-1}\hat{Z}\right)$

Testcases

- Currently only tested on 2 cases.
  - Small number of factors (3)
  - Relevant quadratic and / or interaction terms
  - Reasonable factor input variability.

  - Testcase 1: Viable cell Density optimization on 3 factors
  - Testcase 2: Formulation optimization for viscosity on 3 factors

Test case 1, Viable Cell Density optimization (1/3)

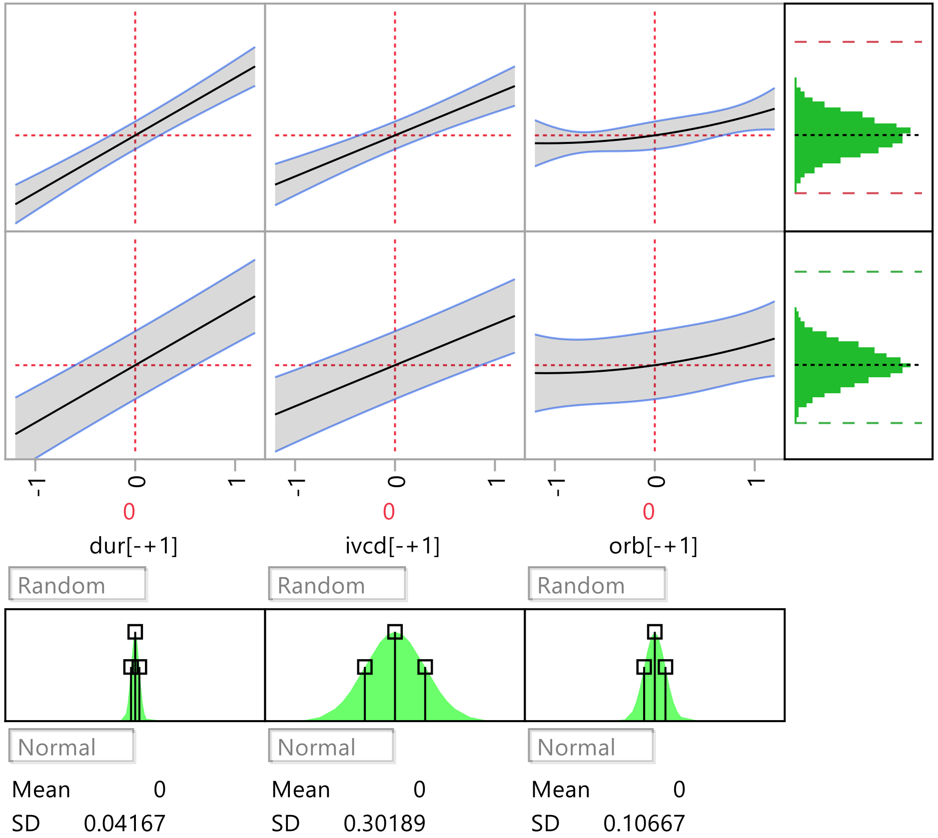
+ Factor distribution
+ Model Error

~ **(Confidence Error)**

+ Factor distribution
+ Model Error
+ Residual Noise

~ **(Prediction Error)**

Joint Distribution

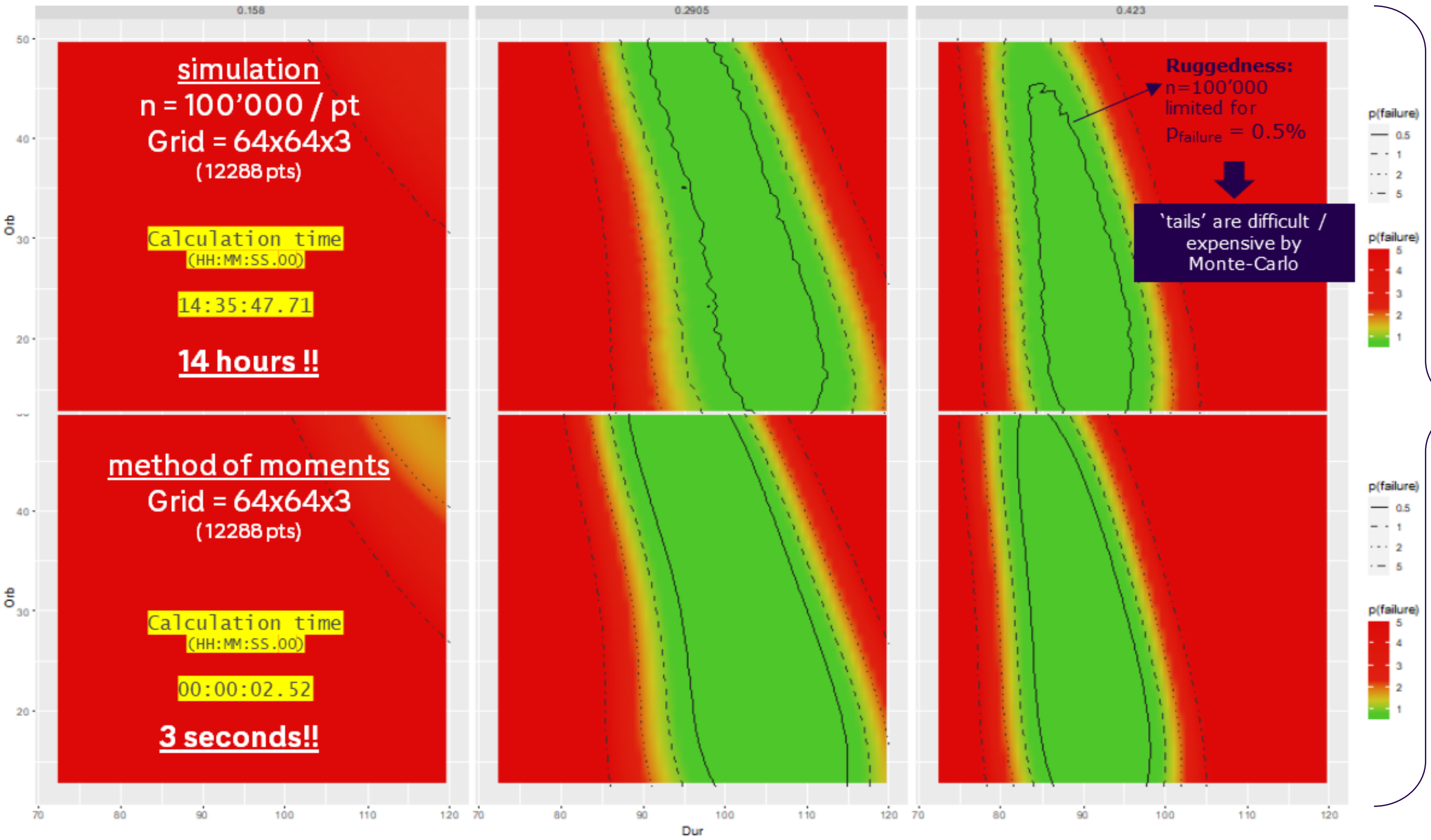


Model:
icept
dur
orb
ivcd

ivcd*dur
orb*dur
orb*ivcd

orb*orb

Test case 1, Viable Cell Density optimization (2/3)

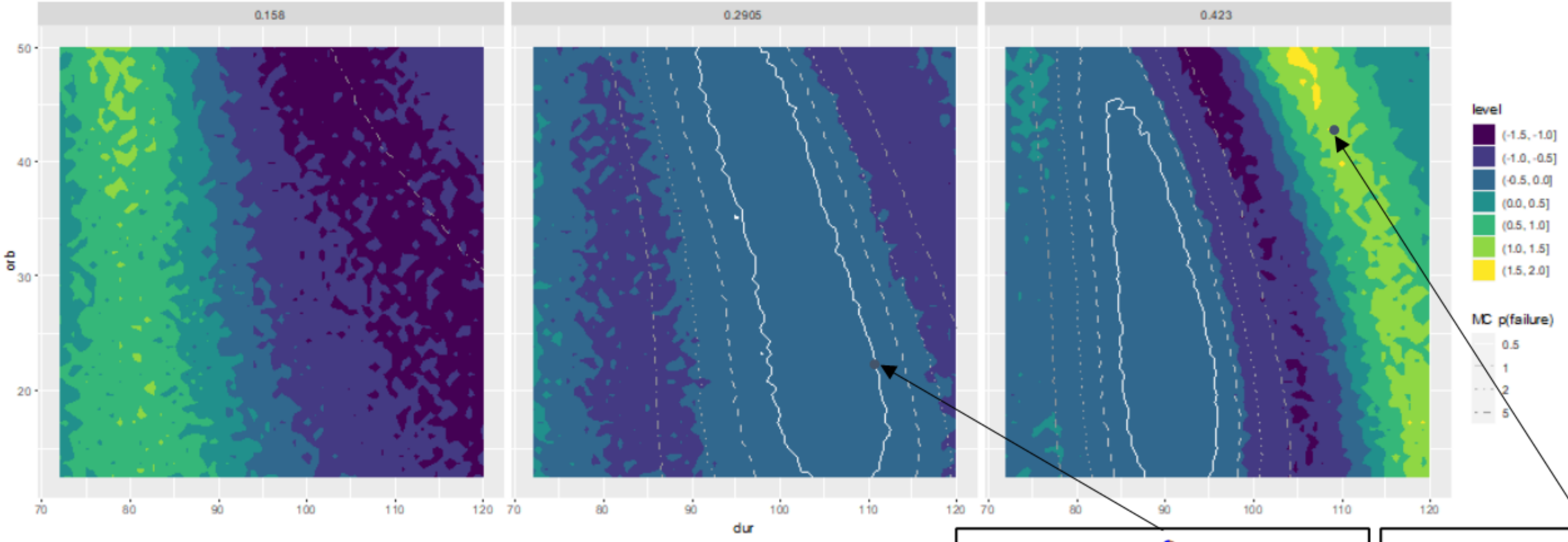


➔ Very comparable results for 'method of moments' compared to simulated reference.

➔ Huge difference in calculation times !

Test case 1, Viable Cell Density optimization (3/3)

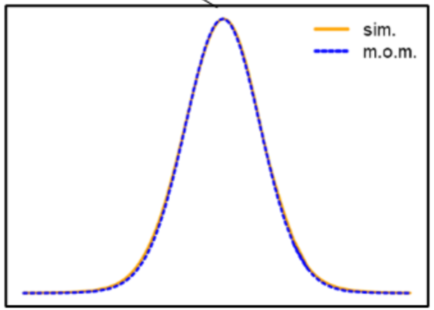
Method of moments compared to simulation reference: P(failure)-value difference

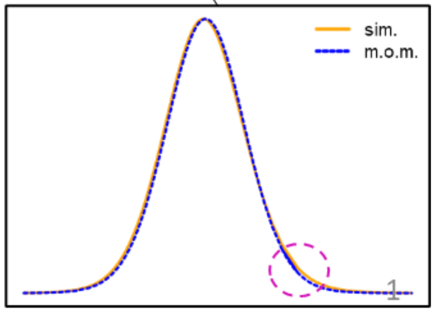


- Differences are not bigger than p(mom) – p(sim) +/- 2%
- In the region of interest (0.5% to 1 % failure risk) it is smaller (-0.5 to 0 %)

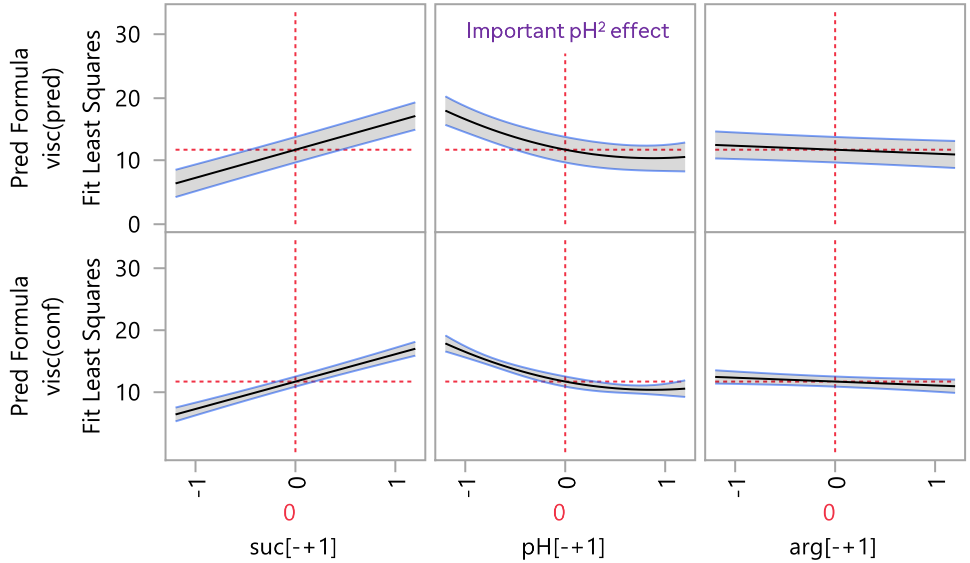
Conclusion: Differences appear acceptable.

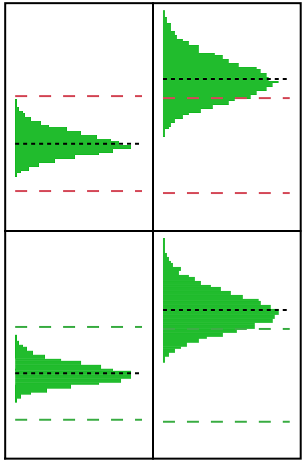
Remark: the difference is close to binomial sampling uncertainty when estimating a prob of 0.5% with n=100’000 simulations.





Test case 2, Viscosity response in a formulation (1/3)





Midpoint
suc 0
pH 0
arg 0

Corner
suc 1
pH -1
arg -1

+ Factor distribution
+ Model Error
+ Residual Noise

~ **(Prediction Error)**

+ Factor distribution
+ Model Error

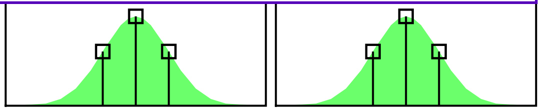
~ **(Confidence Error)**

Joint Distribution

Wide*perturbation: σ = 1/6 (High-Low)

Fixed

0



Normal

Mean	0
SD	0.33

Normal

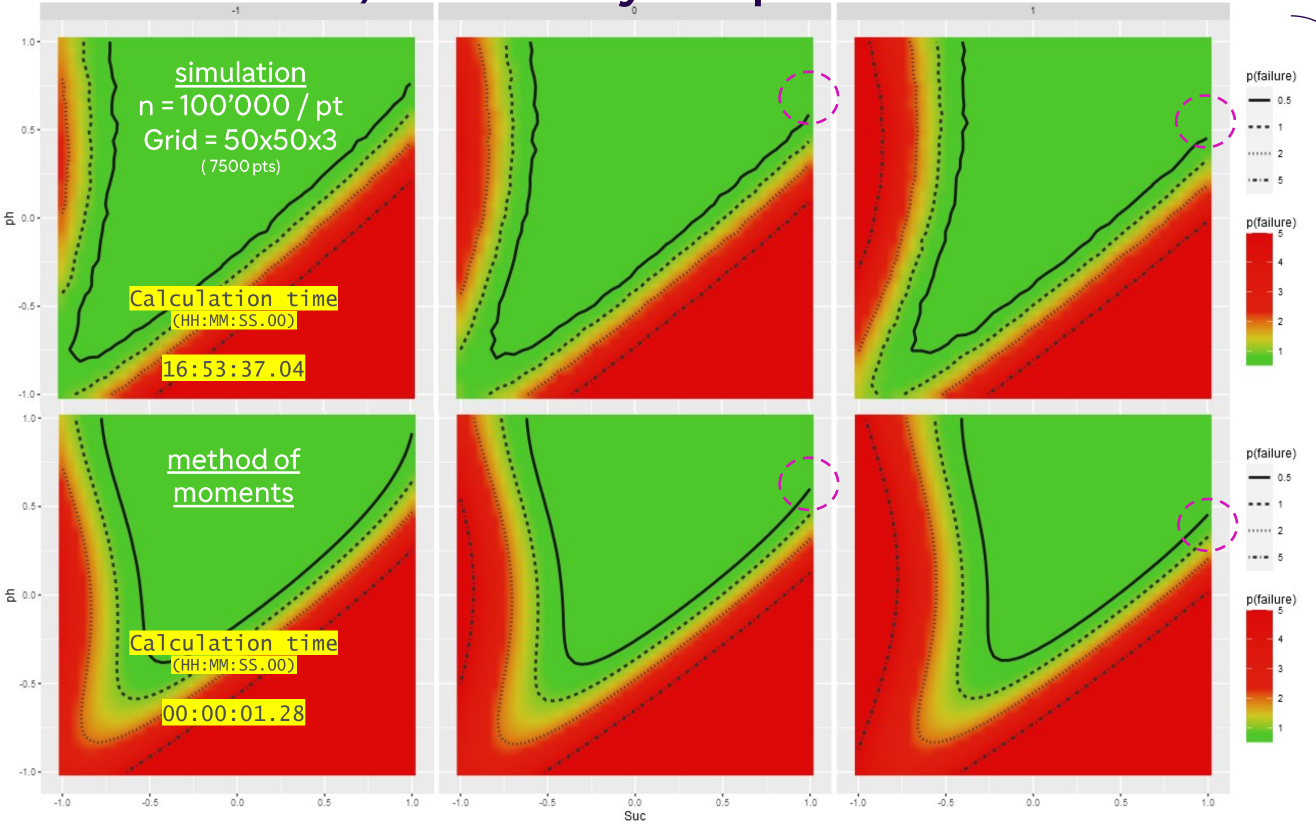
Mean	0
SD	0.33

*** ±3σ covers entire DOE range**

Model:

icept
suc
pH
arg
suc*pH
pH*pH

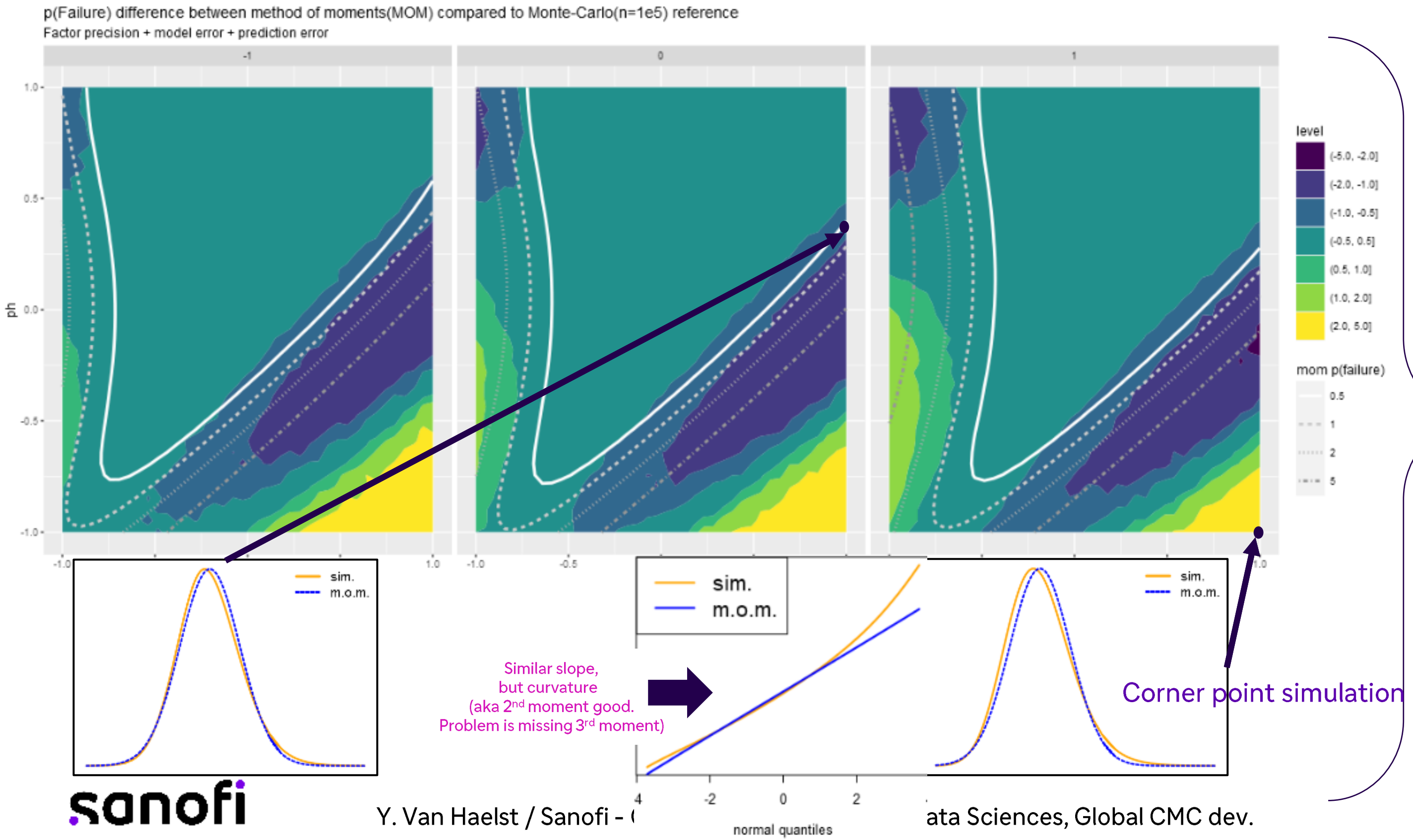
Test case 2, Viscosity response in a formulation (2/3)



- Differences between simulation and mom are apparent
- Appears to be shifted to the right on 'suc' scale.

Test case 2, Viscosity response in a formulation (3/3)

Method of moments compared to simulation reference: P(failure)-value difference



- ➔ Differences are present
- ➔ leads to underestimation:
  - p=0.5% in mom underestimates by 0.5-1 %
  - p = 1% in mom underestimates by 1-2%
- ➔ QQ-plot evaluation indicates result of unaccounted skewness.

<u>Conclusion:</u>

- Differences are small but sufficient relevant to further investigate
- Since qq-plot indicates mostly skewness misspecification, elucidating 3rd moment could correct

References

Bhutani, H., Kurmi, M., Singh, S., Beg, S., & Singh, B. (2004). Quality by design (QbD) in analytical sciences: an overview. *Quality Assurance*, *3*, 39-45.

Chen C (2006) Implementation of ICH Q8 and QbD—an FDA perspective. PharmaForum Yokohama, June. https://www.nihs.go.jp/drug/PhForum/Yokohama060609-02.pdf (accessed on 2024-SEP-04)

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). (2022). ICH Q8 (R2): Pharmaceutical development. [Online guideline]. European Medicines Agency. https://www.ema.europa.eu/en/ich-q8-r2-pharmaceutical-development-scientific-guideline (accessed on 2024-SEP-04)

Kunzelmann, M., Thoma, J., Laibacher, S. *et al.* An *in-silico* approach towards multivariate acceptable ranges in biopharmaceutical manufacturing. *AAPS Open* **10**, 7 (2024). https://doi.org/10.1186/s41120-024-00095-y

Kusumo, K. P., Gomoescu, L., Paulen, R., García-Muñoz, S., Pantelides, C. C., Shah, N., & Chachuat, B. (2020). Nested Sampling Strategy for Bayesian Design Space Characterization. In *Computer Aided Chemical Engineering* (Vol. 48, pp. 1957-1962). Elsevier.

Lancaster L(2023) Finding Optimal Operating Regions for Critical Quality Attributes with the Design Space Profiler - (2023-US-30MP-1447) [Online video] https://community.jmp.com/t5/Discovery-Summit-Americas-2023/Finding-Optimal-Operating-Regions-for-Critical-Quality/ta-p/651670 (accessed on 2024-SEP-04)

Oberleitner, T., Zahel, T., & Herwig, C. (2024). Identifying design spaces as linear combinations of parameter ranges for biopharmaceutical control strategies. *Computers & Chemical Engineering*, *183*, 108555.

Taillefer, V., & Nasir, O. (2020). Statistical methodology for the determination of “univariate” and “multivariate” Proven Acceptable Ranges (PAR) with a Design of Experiments. In *Proceedings of the APEX Conference 2020*.

Authors & responsibilities

Cesaraccio, Gaelle (AIXIAL GROUP): providing TestCase 1 + porting TestCase 2 to R and test-running the simulations

Van Haelst, Yannick (Sanofi): literature + mathematical conceptualization + coding R framework + presentation

Caroline Leveder (Sanofi) slides review; Vincent Taillefer (Sanofi) Modde expertise & initial PAR work (APEX 2022)

Is RSM m.o.m. good enough ?

Could we leverage a simplified 3rd / 4th moment function ?

Should we use CCD, and an RSM metamodel on sampled moments ?

We appreciate your input !

sanofi

Thank *you*

sanofi