BAYESIAN MULTI-OBJECTIVE OPTIMIZATION FOR QUANTITATIVE RISK ASSESSMENT IN MICROBIOLOGY

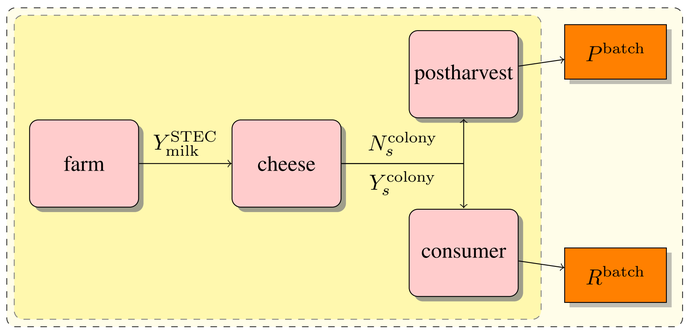
Subhasish Basak[1,2], Julien Bect[2], Laurent Guillier[1], Fanny Tenenhaus-Aziza[3], Janushan Christy[4], Emmanuel Vazquez[2]

[1]Agence nationale de sécurité sanitaire (ANSES), 94700, Maisons-Alfort, France
[2]Université Paris-Saclay, CNRS, CentraleSupélec, L2S, 91190, Gif-sur-yvette, France
[3]Centre national interprofessionnel de l'économie laitière (CNIEL), 75009, Paris, France
[4]Centre technique d'expertise agroalimentaire (ACTALIA), 74800, La-Roche-sur-Foron, France

Context

As a part of the European project ArtiSaneFood, the primary goal of this collaborative work between ANSES, CNIEL and L2S is to establish efficient bio-intervention strategies for cheese producers in France, in order to "economically" reduce the risk of Haemolytic Uremic Syndrome (HUS) caused by Shiga-Toxin producing Escherichia Coli (STEC) present in raw-milk soft cheese. This translates into a multi-objective optimization problem of a stochastic simulator based on a quantitative risk assessment (QRA) model proposed by Perrin et al. (2014), to estimate the Pareto optimal solutions for the process intervention parameters.

QRA simulator



Batch level stochastic simulator

The farm-to-fork continuum for one batch

- Farm module

  STEC and E.coli strains follow same fecal route!

  – A batch of milk is tested with probability $p^{\text{milk}}_{\text{test}}$

  – Farms with E.coli conc. $> l^{\text{sort}}$ are rejected

  STEC conc. $Y^{\text{STEC}}_{\text{milk}}$ in farm milk is computed

- Cheese module

  STEC cells form colonies (clusters) inside cheese

  – No. of colonies (Poisson): N^{colony}_s

  – Size of colonies (LogNormal): Y^{colony}_s for strain s

- Consumer module

  Batch risk is computed using a dose-response model:

$$\Gamma = \sum_s N^{\text{colony}}_s \cdot Y^{\text{colony}}_s \tag{1}$$

$$R^{\text{batch}} = \sum_{\text{age}} g_{\text{age}} \int_\Gamma P[\text{HUS}|\gamma, \text{age}] \cdot \text{p}(\gamma)\text{d}\gamma \tag{2}$$

  averaging over consumer age

- Post-harvest module

  Proportion of rejected batches is estimated using number of test sample n_{sample} given a test rate $p^{\text{cheese}}_{\text{test}}$

$$P^{\text{batch}} = P[\Gamma > 0] = 1 - \exp(-K \cdot n_{\text{sample}})$$

Quantities of Interest (QoI)

Several batches are simulated to estimate

- $R_{\text{avg}} = \mathbb{E}[R^{\text{batch}} \cdot (1 - P^{\text{batch}} \cdot p^{\text{cheese}}_{\text{test}})]$
- $P_{\text{avg}} = \mathbb{E}[P^{\text{batch}} \cdot p^{\text{cheese}}_{\text{test}}]$

QoIs are

- Relative risk: $f_1 = \frac{R_{\text{avg}}}{(1-P_{\text{avg}})\cdot K_1}$, ($K_1$: baseline risk)
- Batch rejection rate: $f_2 = P_{\text{avg}}$

Multi-objective optimization

We consider the multi-objective optimization problem

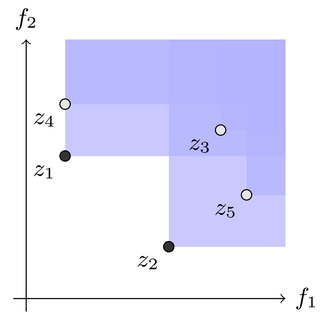
$$\min_{x \in \mathbb{X}} f(x)$$

where, $f = (f_1, f_2, \ldots)$

- The solution set $\mathcal{P}$ contains Pareto optimal points

$$\mathcal{P} = \{x \in \mathbb{X} : \nexists x' \in \mathbb{X}, f(x') \prec f(x)\}$$

  where $f' \prec f \implies f'_i \leq f_i, \forall i$, with at least one of the inequalities being strict



Pareto optimal points z_1 and z_2

- In a stochastic setting, we assume additive noise: for a given $x_i \in \mathbb{X}$, we observe $Z_i = f(x_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \Sigma)$

PALS

Proposed by Zuluaga et al. (2013) and extended for the stochastic case by Barracosa et al. (2021).

- Initial design I_0, $n = 0$, $\beta > 0$
- Step I – $\forall x \in \mathbb{X}$, construct confidence rectangles

$$R_n(x) = \{z \in \mathbb{R}^2_+ : R^-_n(x) \prec z \prec R^+_n(x)\}$$

  using posterior means $\mu_{n,j}(x)$ and variances $\sigma^2_{n,j}(x)$

$$R^{\pm}_{n,j}(x) = \mu_{n,j}(x) \pm \beta^{1/2}\sigma_{n,j}(x)$$

  based on I_n (indep. GP models for f_j, $j \in \{1, 2\}$)

- Step II – Classify each $x \in \mathbb{X}$ into one of three classes

  – Deemed Pareto optimal P_n:

$$P_n = \{x \in \mathbb{X} | \nexists x' \in \mathbb{X} \setminus \{x\}, R^-_n(x') \prec R^+_n(x)\}$$

  – Non Pareto optimal N_n:

$$N_n = \{x \in \mathbb{X} | \exists x' \in \mathbb{X} \setminus \{x\}, R^+_n(x') \prec R^-_n(x)\}$$

  – Unclassified $U_n = \mathbb{X} \setminus (P_n \cup N_n)$

- Step III – Sample the next point of evaluation

$$X_{n+1} = \underset{x \in (U_n \cup P_n)}{\arg\max} \, \|R^-_n(x) - R^+_n(x)\|_2$$

$$I_{n+1} = I_n \cup (X_{n+1}, Z_{n+1}),\ Z_{n+1} = f(X_{n+1}) + \varepsilon_{n+1}$$

- Increase n and Repeat I-III until $n \leq n_{\max}$

PALS with quantiles

PALS is not directly suitable for our application:

- f_1 is not the expectation of a simulator output,
- we build a GP model for R_{avg} instead.

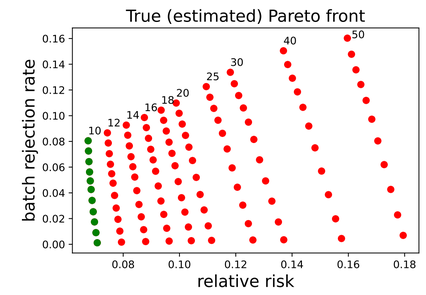
So, we propose using quantiles of QoI to construct confidence rectangles of coverage probability α

$$R^-_{n,j}(x) = q_{n,j}(\alpha/2) \text{ and } R^+_{n,j}(x) = q_{n,j}(1 - \alpha/2)$$

with $q_{n,j}(.)$ being a posterior quantile of $f_j(x)$

Experimental results

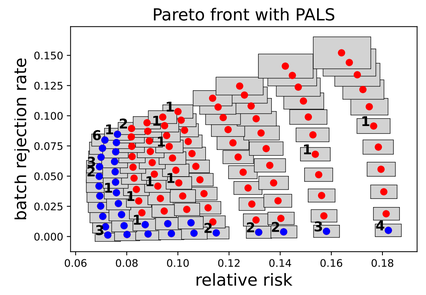
- Minimizing f, given $p^{\text{milk}}_{\text{test}} = 1, p^{\text{cheese}}_{\text{test}} = 0.5$, w.r.t.
  – $n_{\text{sample}} \in \{1, 5, 10, 15, \cdots, 50\}$
  – $l^{\text{sort}} \in \{10, 12, 14, 16, 18, 20, 25, 30, 40, 50\}$
- True Pareto front: estimated using 5000 samples for each of 11×10 input pair
  – A GP regressor is used to smooth the predictions



(stripes corresponding diff. values of l^{sort})

  – Pareto optimal (green) and dominated points (red)

- Pareto front estimated using PALS
  – $|I_0| = 10$, $n_{\max} = 30$, samples/evaluation $= 200$



(integers denoting evaluation counts)

- With PALS using significantly less (40×200) budget, the user can provide the following insights
  – Points with $l^{\text{sort}} > 12$ are dominated (red)
  – $l^{\text{sort}} \leq 12$ remain unclassified (blue)

Future work

- Integrate milk loss as objective and $p^{\text{milk}}_{\text{test}}$, $p^{\text{cheese}}_{\text{test}}$ as design variables
- Take the correlation between outputs into account

References

F. Perrin, F. Tenenhaus-Aziza, V. Michel, S. Miszczycha, N. Bel, and M. Sanaa. Quantitative risk assessment of haemolytic and uremic syndrome linked to O157:H7 and non-O157:H7 shiga-toxin producing escherichia coli strains in raw milk soft cheeses. *Risk Analysis*, 35(1):109–128, 2014.

M. Zuluaga, G. Sergent, A. Krause, and M. Püschel. Active learning for multi-objective optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 462–470. PMLR, 2013.

B. Barracosa, J. Bect, H. Dutrieux Baraffe, J. Morin, J. Fournel, and E. Vazquez. Extension of the Pareto Active Learning method to multi-objective optimization for stochastic simulators. In *SIAM Conference on Computational Science and Engineering (CSE21)*, Mar 2021.

This work is part of the ArtiSaneFood project (grant number : ANR-18-PRIM-0015) which is part of the PRIMA program supported by the European Union.