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]

L2S - Journée des doctorants, 15 Sept, 2022

[1]Agence Nationale Sécurité Sanitaire Alimentaire Nationale (ANSES), Maisons-Alfort, France
[2]Univ. Paris-Saclay, CNRS, CentraleSupeléc, L2S, Gif-sur-yvette, France
[3]Centre National Interprofessionnel de l'Economie Laitière (CNIEL), Paris, France
[4]Centre technique d'expertise agroalimentaire (ACTALIA), La-Roche-sur-Foron, France

Motivation & application

- Project ArtisaneFood: Control food-borne pathogens in artisanal fermented foods for Mediterranean countries
- ArtiSaneFood France
  - Product: Camembert de Normandie (fromage au lait cru)
  - Pathogen: Shiga Toxin producing Escherichia coli (STEC)
  - Disease: Haemolytic Uremic Syndrome (HUS)
- Microbiological Quantitative Risk Assessment (QRA)
- Study impact of intervention steps
  - Preharvest intervention - Farm milk is tested
  - Postharvest intervention - Cheese batches are tested

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.

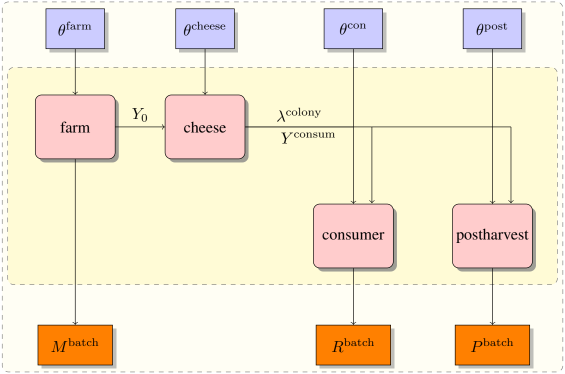
Motivation & application

- Goal: Make methodological recommendations to French cheese producers
- Find optimal values of process intervention parameters
  - f^{sort}: Frequency of milk testing (days)
  - l^{sort}: Milk test threshold (CFU/ml)
  - p^{test}: Proportion of cheese batches tested
  - n^{sample}: Number of cheese samples tested
- Objectives to minimize
  - R^{HUS}: Relative risk of HUS
  - C: Cost of intervention

Contents

1 Quantitative Risk Assessment

- QRA simulator based on model proposed by Perrin et al. (2014)



- Models the farm-to-fork continuum
  - Farm module: Computes STEC concentration in farm milk
  - Cheese module: Evolution of STEC is modelled using ODEs
  - Consumer module: Computes risk averaging over consumer behaviour
- Outputs: risk of HUS (R^{batch}), milk loss (M^{batch}) and proportion of cheese batches rejected (P^{batch})
- Quantities of interest (QoI):

$$R^{\text{HUS}} = \mathbb{E}[R^{\text{batch}} \cdot (1 - P^{\text{batch}} \cdot p^{\text{test}})] / (\mathbb{E}[P^{\text{batch}} \cdot p^{\text{test}}] \cdot K) \quad (1)$$

$$\begin{aligned} C = \mathbb{E}[M^{\text{batch}} \cdot c^{\text{milk}} + P^{\text{batch}} \cdot p^{\text{test}} \cdot c^{\text{cheese}} + \\ c^{\text{milk}}_{\text{test}} / f^{\text{sort}} + n^{\text{sample}} \cdot c^{\text{cheese}}_{\text{test}} \cdot p^{\text{test}}] \end{aligned} \quad (2)$$

c^{milk}, c^{cheese}, $c^{\text{milk}}_{\text{test}}$ and $c^{\text{cheese}}_{\text{test}}$ denotes costs of intervention steps

K is baseline risk (no interventions)

2 Multiobjective optimization

- We consider a biobjective optimization problem of $f = (R^{\mathrm{HUS}}, C)$

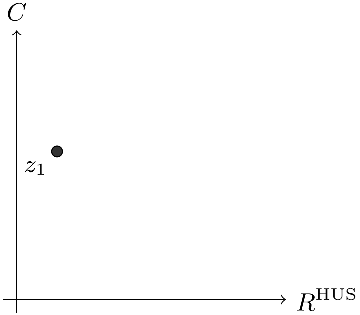
$$\min_{x \in \mathbb{X} \subset \mathbb{R}^4} f(x) \tag{3}$$

- Stochastic: We observe with additive noise $Z(x) = f(x) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \Sigma)$
- Conflicting objectives: There is no unique optimal solution
- The solution set consists of Pareto optimal points

$$\mathcal{P} = \{x \in \mathbb{X} : \nexists x' \in \mathbb{X}, Z(x') \prec Z(x)\} \tag{4}$$

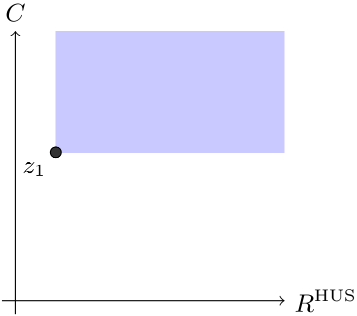
- Where $Z(x') \prec Z(x) \implies Z_i(x') \leq Z_i(x), \forall i$, with at least one strict inequality

Pareto optimal solutions: the objective space



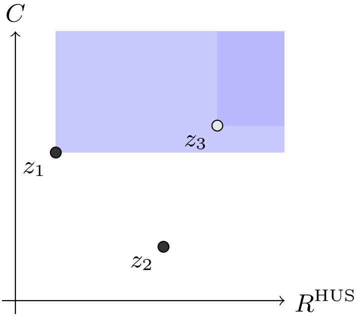
One observation $z_1 = (R_1^{\mathrm{HUS}}, C_1)$

Pareto optimal solutions



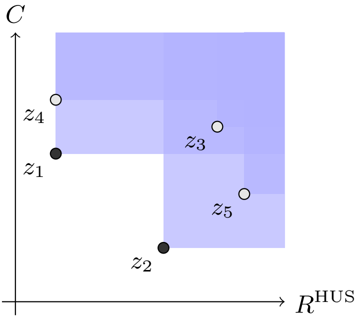
Dominated area by z_1 in objective space

Pareto optimal solutions



z_1 dominates z_3 but not z_2

Pareto optimal solutions



z_3, z_4 & z_5 dominated by z_1 and z_2

Problem formulation

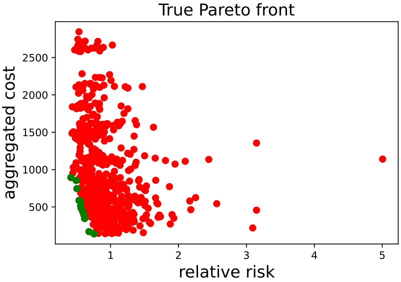
- Optimize the function $f = (R^{\mathrm{HUS}}, C)$
  - Input space $(f^{\mathrm{sort}}, l^{\mathrm{sort}}, p^{\mathrm{test}}, n^{\mathrm{sample}}) \in \mathbb{X} \subset \mathbb{R}^4$

$$\begin{aligned} f^{\mathrm{sort}} &\in \{10, 20, 30, 40, 50\} \\ l^{\mathrm{sort}} &\in \{10, 20, 30, 40, 50\} \\ p^{\mathrm{test}} &\in \{0.1, 0.2, 0.3, 0.4, 0.5\} \\ n^{\mathrm{sample}} &\in \{5, 10, 20, 30, 50\} \end{aligned}$$

  - Objective space $(R^{\mathrm{HUS}}, C) \in \mathbb{R}^2$
- The input space $\mathbb{X}$ is discrete and finite (625 points)
- We want to estimate the Pareto set $\mathcal{P} \subset \mathbb{X}$

Naive solution: Brute force Monte Carlo

- Expensive: Heavy MC evaluated $\forall x \in \mathbb{X}$, takes > 4 days!



- Each point is an estimated average over 5000 iterations
- Pareto optimal (green) and non Pareto optimal (red) points

3 Stochastic Pareto Active Learning (PALS)

- Proposed by Zuluaga et al. (2013) and extended by Barracosa et al. (2021)
- Why PALS ?
  - Easy to implement and inexpensive
  - Does not have computationally-intensive criteria like other Bayesian optimization algorithms (see, e.g., Hernandez-Lobato et al., 2016)
  - Suitable for optimizing expensive and stochastic simulators
- PALS at a glance:
  - Gaussian process (GP) model to construct a inexpensive surrogate
  - Samples cleverly the points in $\mathbb{X}$ to evaluate
  - Classifies points in $\mathbb{X}$ using confidence rectangles

Surrogate modelling: Gaussian process regression

- For each of the QoIs we consider the data generative model

$$Z_j = \xi(x_j) + \varepsilon_j \tag{5}$$

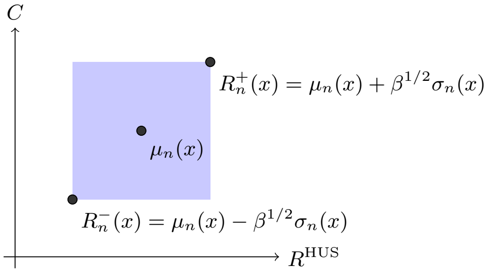
  where

  - $\xi \sim \mathrm{GP}(m, k)$ and $\varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$, independent of ξ
  - mean function $m : \mathbb{R}^4 \to \mathbb{R}$, kernel $k : \mathbb{R}^4 \times \mathbb{R}^4 \to \mathbb{R}$

- The parameters of m, k and noise variance are estimated using the method of maximum likelihood

- Knowing m and k the posterior $\xi | Z_1, Z_2, \ldots, Z_n, m, k$ can be computed by solving a system of linear equations (see, Rasmussen and Williams, 2006)

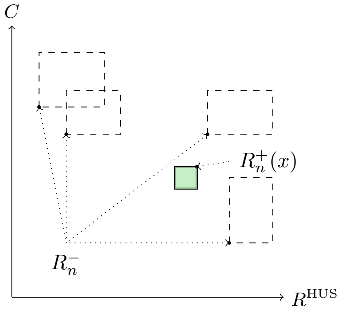
PALS confidence rectangle



Confidence rectangles in PALS

- μ_n and σ_n^2: posterior mean and variance of the GP model
- β: coverage probability, n: number of data points

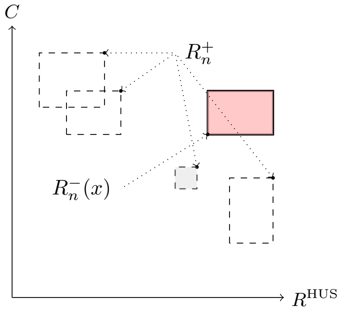
Deemed Pareto optimal



$$P_n = \{x \in \mathbb{X} | \nexists x' \in \mathbb{X} \setminus \{x\}, R_n^-(x') \prec R_n^+(x)\} \tag{6}$$

- The pessimistic (R^+) outcome of the green box is not dominated by the optimistic (R^-) outcome of any other box

Non Pareto optimal



$$N_n = \{x \in \mathbb{X} | \exists x' \in \mathbb{X} \setminus \{x\}, R_n^+(x') \prec R_n^-(x)\} \tag{7}$$

- The optimistic (R^-) outcome of red box is dominated by the pessimistic (R^+) outcome of at least one other box

PALS algorithm

- Classification is done $\forall x \in \mathbb{X}$ at each iteration $n < n_{\max}$
- Each point is classified as one of the following:
  - P_n: Deemed Pareto optimal
  - N_n: Non Pareto optimal
  - $U_n = \mathbb{X} \setminus (P_n \cup N_n)$: Unclassified
- Sample the next point of evaluation

$$X_{n+1} = \underset{x \in P_n \cup U_n}{\arg\max} \, ||R_n^-(x) - R_n^+(x)||_2 \tag{8}$$

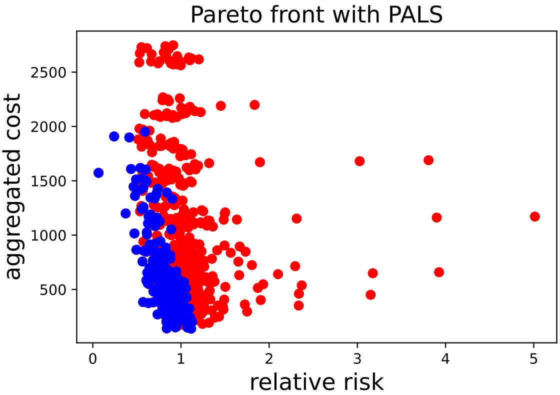
where the uncertainty is maximum

4 PALS with quantiles

- PALS as proposed by Barracosa et al. (2021) is not suitable when:
  - QoI is not an expectation of the simulator outputs
- Several batches are simulated to estimate
  - $R_{\text{avg}} = \mathbb{E}[R^{\text{batch}} \cdot (1 - P^{\text{batch}} \cdot p^{\text{test}})]$
  - $P_{\text{avg}} = \mathbb{E}[P^{\text{batch}} \cdot p^{\text{test}}]$
- QoI is $R^{\text{HUS}} = \frac{R_{\text{avg}}}{(1-P_{\text{avg}}) \cdot K}$
- Basak et al. (2022) propose using quantiles to construct rectangles estimated from the sample paths of GP

Estimated Pareto front with PALS (with quantiles)

Unclassified (blue) and non Pareto optimal (red)



5 Perspectives

- PALS with quantiles is still not able to classify well between
  - P_n: Deemed Pareto optimal
  - U_n: Unclassified
- This is possible due to use of confidence rectangles for classification
  - When the observations are too close in the objective space
- We propose Rectangle-free version of PALS (WIP!)

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, volume 28 of Proceedings of Machine Learning Research, pages 462–470, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

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), Virtual Conference originally scheduled in Fort Worth, Texas, United States, Mar 2021.

Daniel Hernandez-Lobato, Jose Hernandez-Lobato, Amar Shah, and Ryan Adams. Predictive entropy search for multi-objective bayesian optimization. In Maria Florina Balcan and Kilian Q. Weinberger, editors, Proceedings of The 33rd International Conference on Machine Learning, volume 48 of Proceedings of Machine Learning Research, pages 1492–1501, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/hernandez-lobatoa16.html`.

C. E. Rasmussen and C. K. I. Williams. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA, 2006.

Subhasish Basak, Julien Bect, Laurent L. Guillier, Fanny Tenenhaus-Aziza, Janushan Christy, and Emmanuel Vazquez. Bayesian multi-objective optimization for quantitative risk assessment in microbiology. MASCOT-NUM 2022, June 2022. URL `https://hal.archives-ouvertes.fr/hal-03715857`. Poster.

Thank you for your attention!

Questions?