Credit: © F. Cormon/Auvergne-Rhône-Alpes Tourisme

MASCOT-NUM 2022
Clermont Ferrand, France
7-9 June 2022

Bayesian multi-objective optimization for quantitative risk assessment in microbiology

S. Basak
Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail (ANSES)
Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes (L2S)

Supervisor(s): J. Bect (L2S), L. Guillier (ANSES), F. Tenenhaus-Aziza (CNIEL) and E. Vazquez (L2S)

PhD expected duration: Jan. 2021 - Dec. 2023

Address: L2S, CentraleSupélec, 3 Rue Joliot-Curie, 91190 Gif-sur-Yvette, France

E-mail: subhasish.basak@centralesupelec.fr

Abstract:

In the field of microbiology for the food industry, quantitative risk assessment (QRA) is the scientific discipline that analyzes the links between the practices of producers and consumers and the contamination risk for food-borne diseases. In the context of the ArtiSaneFood project (see Biography), we concentrate our analysis on the work by [2], where the authors propose a stochastic quantitative microbial risk assessment model to assess the risk of Haemolytic Uremic Syndrome (HUS) associated with the five Main Pathogenic Stereotypes of Shiga-toxin producing Escherichia coli (MPS-STEC) in raw-milk soft cheeses. This model is implemented as a stochastic simulator in R that simulates several batches to produce one estimate of the overall risk. The simulator is computationally expensive. In order to control and optimize the risk along with other outputs of the simulator, we thus propose to use Bayesian optimization algorithms. These optimization techniques are useful when the objective functions are expensive to evaluate and their gradients with respect to the design variables are either unknown or expensive as well. These algorithms rely on Bayesian approaches to obtain predictions of the objective functions. Most of the time, they consists of Gaussian process (GP) interpolation or regression. Taking into consideration the stochastic nature of the simulator and the multi-objective framework, for the estimation of Pareto-optimal solutions, we propose to use an extension of the Pareto Active Learning (PAL) algorithm, originally proposed by [4] and later extended for the stochastic setting (PALS) by [1].

The stochastic simulator comprises several hierarchical levels that represent each step in the farm-to-fork continuum. At the batch level, the simulator is composed of the following modules: a farm module followed by a preharvest module (a.k.a. milk sorting module), a cheese production module, a consumer module and a postharvest module (a.k.a. sampling module). Given the inputs of the modules, one particular simulated batch replicates the production process of a single batch of cheese with milk coming from a given set of farms with specific hygiene standards, corresponding to input parameters. The milk is tested for possible rejection in the preharvest milk sorting stage and then processed through the cheese module, which models the evolution of the bacteria during different stages of cheese production, namely storage, molding, draining, salting and ripening. The first three stages correspond to the growth phase where the simulator makes use of ordinary differential equations to model the growth of the bacteria and during the last two stages there is a decline in the concentration.

The risk corresponding to a batch is assessed using a dose-response model which takes into account the consumption behavior of people from different age groups. In the postharvest step the produced batch of cheese is inspected for STEC contamination, by testing small portions of cheese. Several batches are simulated to obtain estimates of the expected value R^{HUS} of the risk and the expected proportion of rejected (destroyed) batches P^{reject}. As a first sub-problem the idea is to minimize these outputs with respect to design variables of the postharvest step, namely the number of test samples (n^{sample}) and the mass of the samples (m^{sample}).

We consider the problem of multi-objective simulation optimization with two objective functions taking values in $[0, 1]$ with two design variables defined on a search domain $\mathbb{X} \subset \mathbb{R}^2$. In such a framework,

the goal is to identify the optimal solutions that represent the best possible trade-offs among the two objectives, defined using the Pareto domination rule ($\prec$). For the objective functions f_1 and f_2, the Pareto set containing all such optimal solutions is given by $\mathcal{P} = \{x \in \mathbb{X} : \nexists x_1 \in \mathbb{X}, f(x_1) \prec f(x)\}$, where $f = (f_1, f_2)$, with at least one inequality being strict. In the stochastic setting, instead of observing the objectives f_i themselves, we observe evaluations $Z_i = f_i + \epsilon_i$, with homoscedastic noise ϵ_i. The first step is to model each objective using an independent Gaussian process [3] prior ξ_i that makes it possible to easily obtain the posterior distributions of the models conditional on the observations Z_i. The next step is to obtain a good approximation $\widehat{\mathcal{P}}_n$ of $\mathcal{P}$, using a sequence $\mathbb{X}_n = \{X_1, X_2, \ldots, X_n\}$ of evaluation points and the posterior distributions of the Gaussian processes ξ_i updated with each iteration. The PALS algorithm does this by classifying the points of the input set $\mathbb{X}$ into three sets at iteration n: a set $P_n \subset \mathbb{X}$ of points that are deemed Pareto-optimal, a set $N_n \subset \mathbb{X}$ of points that are rejected as Pareto-optimal, and a set $U_n \subset \mathbb{X} \setminus (P_n \cup N_n)$ of points that are still "uncertain". This is done by associating with each point $x \in \mathbb{X}$ a region $R_n(x) \subset \mathbb{R}^2_+$ in the objective space, which quantifies the associated uncertainty based on the posterior distribution of ξ_i conditioned with respect to observations until that particular iteration. For a particular evaluation point $x \in \mathbb{X}$, the region $R_n(x)$ is constructed as a hyper-rectangle $\{z \in \mathbb{R}^2_+ : \mu_n(x) - \beta^{1/2}\sigma_n(x) \prec z \prec \mu_n(x) + \beta^{1/2}\sigma_n(x)\}$ with the vector of posterior mean $\mu_n(x)$, posterior variance $\sigma^2_n(x)$ of the Gaussian process prior ξ_i and a scaling parameter β. Each point x is then classified in P_n, U_n or N_n by comparing the position of the regions to each other according to the Pareto rule. Finally given the budget constraint on total number of evaluations, the algorithm selects a new evaluation point X_{n+1} from $U_n \cup P_n$ by maximizing the uncertainty, that is measured with respect to the diameter of $R_n(x)$.

The PALS algorithm as proposed by [1] is easy to implement and inexpensive compared to other Bayesian optimization algorithms that are based on computationally-intensive criteria. This work presents the application of the PALS algorithm to our particular stochastic multi-objective simulator.

Joint work with J. Bect, L. Guillier, F. Tenenhaus-Aziza and E. Vazquez.

References

[1] 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.

[2] 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.

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

[4] 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.

Short biography – I am a second-year doctoral student at the Laboratoire des Signaux et Systèmes (L2S), Université Paris-Saclay and at Agence nationale de sécurité sanitaire (ANSES). Before my PhD, I completed my master in Data Science from Chennai Mathematical Institute (CMI) in India, and I did my bachelor in Statistics from St. Xavier's College, Kolkata, India. My thesis is a part of the European project ArtiSaneFood which aims at controlling food-borne pathogens in artisanal fermented foods of meat and dairy origin produced in the mediterranean region. France is participating to this project through a collaboration between ANSES, CNIEL (Centre National Interprofessionnel de l'Economie Laitière) and L2S. The main goal is to establish methodological recommendations to cheese producers in France, in order to reduce the risk of Haemolytic Uremic Syndrome (HUS) arising out of the consumption of raw-milk soft cheese.