Universitat Autònoma de Barcelona

PhD in Computer Science

Research line: Artificial Intelligence

Research Plan for the PhD Thesis:

Bayesian optimization for calibration of cost-effectiveness simulation models

PhD Student: David Gómez Guillén
PhD Advisor: Mireia Díaz / Josep Lluís Arcos
PhD Tutor: Josep Lluís Arcos
PhD admission date: 29 April 2022

Bellaterra (Cerdanyola del Vallès), 11 December , 2022

Abstract

Cost-effectiveness analyses (CEA) in healthcare are used to evaluate both health outcomes and costs of different strategies (e.g., cancer preventive interventions) to inform decision makers what strategies are more efficient. Many CEA use simulation models that mimic a disease process to evaluate the aggregate effects of thousands or millions of individuals transitioning through different health states. These transitions between states happen because of different events, such as developing a disease, undergoing medical procedures, or getting treatment.

To correctly represent the development of the disease and the different strategies, models include parameters usually obtained from the scientific literature (e.g.: annual probabilities of developing the disease, survival rates or hazard ratios). This set of parameters is often uncertain, and needs be calibrated to fit specific scenarios (e.g.: cancer burden in Spain). That is, the values initially introduced are slightly varied by optimization methods until the desired target is obtained (e.g.: cancer incidence rates by age group). Many classical optimization algorithms such as Nelder-Mead or Simulated Annealing are often used for this purpose, but they can be very time-consuming to find an adequate solution due to their relatively simple heuristic approach, especially when simulation models are large enough and there are many parameters to adjust. Also, they have little flexibility to adapt to the knowledge we may have about the problem simulated.

In this work we explore Bayesian optimization to improve the calibration performance by guiding the process with domain-specific knowledge and exploiting inherent structural properties in the solution space. These improvements include efficient constraint handling, prior specification for parameters, additive model decomposition or GPU-friendly computation, among others.

Keywords: Cost-effectiveness analysis (CEA), simulation modelling, optimization, Bayesian optimization, gaussian processes, artificial intelligence

Contents

1. Introduction

Cost-effectiveness analysis (CEA) in a healthcare context is used to evaluate different medical strategies (e.g., screening, diagnosis, treatment, …) and to inform decision makers which of them are more efficient, from a health and economic perspective. CEA uses simulation models that mimic a disease process to evaluate the aggregate effects of thousands or millions of individuals transitioning through different health states. These transitions between states happen as a result of different events, such as developing a disease, undergoing medical procedures or getting treatment. Simulation models for CEA (from now on CEA models) can be designed and implemented in different ways (decision trees, Markov models, microsimulation models, dynamic transmission models, …), but their outputs always include at least a measure of effectiveness and a measure of cost, though any other computable output of interest can be obtained depending on the interest of the analysis.

These simulated strategies are parameterized by values found from expert sources, such as the scientific literature or expert opinions. These parameters include probabilities (e.g., probability of getting the disease), rates (e.g., recurrences), costs (e.g., treatment cost) or utility values (e.g., utility of the disease health state), among others. The proper specification of these parameters is an important prerequisite before simulating the strategies and being able to draw robust conclusions.

To ensure a reasonable set of parameters, one important procedure is the calibration of the model. By simulating the natural history of the disease, we aim to find the best set of parameters that produces an output as close as possible to a known, expected outcome. This calibration procedure can be seen as an optimization problem that, due to its high dimensionality and other peculiarities, can be a challenging problem for classical methods such as Nelder-Mead or Simulated Annealing, in some cases taking weeks of computational time for a single calibration.

As an illustrative example we can consider a cost-effectiveness analysis of lung cancer [1]. The objective was to measure the effect of different combinations of intensive smoking cessation intervention programmes with cancer screening at different ages, frequencies, and coverages. By simulating these combinations and comparing their outputs it was found, among other results, that the most cost-effective strategy would be to implement intensive smoking cessation interventions at ages 35, 40 and 45, combined with screening every three years between the ages of 55 and 65. However, before reaching valid results the model had to be calibrated several times since sometimes the calibrated parameters were inconsistent with medical evidence (e.g. probability of dying decreasing as the cancer progresses). These multiple calibrations took substantial computational and human effort.

The purpose of this work is to suggest novel methods of calibrating cost-effectiveness models to obtain better models in a more efficient and flexible manner.

2. Research Objectives

The goal of this thesis is the exploration and implementation of new methods to improve the calibration of cost-effectiveness models. The possible improvements can include different aspects, but to achieve the overall objective we need to fulfill these objectives:

- **Exploration and analysis of the solution spaces of cost-effectiveness models**

While exploring optimization methodology we first need to explore the cost-effectiveness models that we are trying to optimize and their behavior in order to identify which structural patterns in their solution space we can exploit. By analyzing different CEA models, we can detect similarities and shared behaviors that we could integrate in the optimization process.

- **Exploration and development of novel optimization methods**

The final goal is finding a method that allows a cost-effectiveness analyst to calibrate a model as efficiently as possible. Although the objective should be focused on CEA models, the application for more general problems can be considered as well.

- **Tests and comparison of optimization methods**

The results of the methods explored in the previous objective must be assessed and compared to the conventional methods used for the same purpose. Different aspects should be evaluated, such as execution time, number of function evaluations, or interpretability, possibly among others. As mentioned before, it could be interesting to evaluate the performance of this method on other problems beyond cost-effectiveness analysis.

3. State of the art

Current calibration procedures in CEA models are often reliant on manual trial-and-error or simple methods such as Nelder-Mead or Controlled Random Search (CRS) [2]. Since the parameters of these models tend to have complex relationships between them, it can be difficult to find a set of values that make sense from a domain point of view.

These classical methods use simple heuristics to find global optima, but other alternatives provide a more sophisticated and flexible approach. Sequential Model-Based Optimization (SMBO) is a state-of-the-art methodology used to optimize expensive functions, with success in areas such as hyperparameter tuning in machine learning [3]. As the name implies, SMBO uses a surrogate model to guide the optimization process using principled inference steps to be more efficient and minimize the number of evaluations of the target function.

A popular choice for a surrogate model is a Gaussian Process [4]. These non-parametric regression models allow great flexibility by specifying a kernel function that determines their expressiveness [5], and they can be used to take advantage of the properties of the kind of functions that we want to optimize.

First, there are diverse ways in the literature to add constraints in Gaussian processes [6][7], usually by modelling the constraints in a separate model and integrating this knowledge in the overall inference step. Secondly, usual complications arising in optimization problems due to high dimensionality can be mitigated using techniques such as additive kernels [8][9], input prior specification [10], dimensionality reduction [11] or matrix factorization [12], among others [13].

As mentioned before, SMBO is generally applied to expensive functions due to the method's significant overhead while doing the inference step in each iteration. To reduce the impact of this overhead and enable the reasonable use of this method in less expensive functions other techniques have been developed: batch learning [14][15], parallelization [16] or GPU-friendly approaches [17][18], for example.

Among all these techniques, the focus will initially be in the implementation of Gaussian Processes with Orthogonal Additive Kernels suitable to our problem, with the integration of constraint support via auxiliary surrogate models in the inference process. In the first place, as Duvenaud et al show, additive kernels can be much more efficient in finding patterns in high-dimensional problems such as the one we are considering in this project. Also, as mentioned in the introduction, our problems are often highly constrained, and these constraints can become a problem for traditional methods, used mostly to discard invalid solutions. The SMBO method, however, can be greatly improved by leveraging them and guiding the process toward a good and valid solution.

4. Research methodology and work plan

4.1 Research methodology

- **Literature review**

First, a broad review will be carried out from scientific literature in mathematics, statistics, and computer science regarding optimization of black box functions. This review will focus on methods and strategies that fit the kind of CEA models we need to optimize, as well as their particularities.

- **Analysis**

Before doing extensive development and testing, preliminary analysis of the CEA models will be performed to check if the researched methods are a good fit for our problem. This analysis can also lead to better search terms for a new literature review.

- **Design and development**

After the review and analysis of the CEA models, we plan to adapt the methods found for our particular problem and implement a solution that can generate results.

- **Evaluation**

Once we have results, we can evaluate them and compare them with different alternatives or other classical methods.

- **Dissemination**

Finally, this work is planned to be published in artificial intelligence (technical contributions) and healthcare indexed journals (solution contributions), depending on the relevance of the results for each field. Seminars and conferences will be used to further disseminate the results of this work. A preliminary list of potential ideas for publications include:

1. Comparison between different calibrated values and its impact on the cost-effectiveness result
2. SMBO method for high-dimensional and highly constrained cost-effectiveness models
3. Application of previous method in a cost-effectiveness analysis situation

4.2 Work plan

This project will be developed in incremental sprints of four months. Each sprint will begin with a literature review followed by an analysis phase to check if the strategies found are helpful for our models. After that, the strategy of interest will be implemented in the design and development phase. Lastly, its results will be added to a compilation of results for evaluation. In all these stages the details will be documented with the intention of assisting the process of writing in the last stage of dissemination.

The expected duration of the project is four years. The plan for each year is as follows:

TASK	**MONTH**												
Development of calibration using Bayesian optimization	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												
Implementation of constraints in Bayesian Optimization	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												
Test and validation of calibration on	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												

simplified lung model	3rd												
	4th												
Test and validation of calibration on endometrium model	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												
Test and validation of calibration on external model	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												
CSIC & IDIBELL seminar presentations	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												
Conference submissions and presentations	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												
Manuscript submissions and publications	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												
Thesis writing and defense	YEAR	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
	1st												
	2nd												
	3rd												
	4th												

4.2.1 Detailed work plan for the current academic course

For the current academic course, the sprints will focus on different points of those mentioned in the previous section. In the first three months the focus will be on finding a promising main optimization strategy to focus on. Once a method is found, the next six months are planned to implement, run, and compare results iteratively to build a prototype of an algorithm using two CEA models as benchmarks: first a simplified lung model and, later on, a more complex endometrium model. A poster presentation will be made to present the preliminary development made in the first months to showcase the project in very general terms.

TASK	MONTH											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12
Development of calibration using Bayesian optimization	■	■	■	■	■	■	■	■	■	■	■	■
Implementation of constraints in Bayesian Optimization					■	■	■	■	■	■	■	■
Test and validation of calibration on simplified lung model				■	■	■	■	■	■	■		
Test and validation of calibration on endometrium model											■	■
CSIC & IDIBELL seminar presentations		■	■						■	■		

5. Conclusions

We have outlined the general research plan for the first academic year. Since the scope of this work involves a healthcare perspective, other concerns besides purely computer science criteria have to considered, so unexpected modifications to this plan may occur.

As a summary, the objective of this thesis consists in an efficient way to calibrate CEA models while considering and taking advantage of knowledge from a healthcare domain. This endeavour will prove useful for healthcare professionals in the development of new medical strategies to improve the quality of life of the population in an economically sustainable basis.

6. References

1. Diaz, M., Garcia, M., Vidal, C., Santiago, A., Gnutti, G., Gómez, D., Trapero-Bertran, M., & Fu, M. (2021). Health and economic impact at a population level of both primary and secondary preventive lung cancer interventions: A model-based cost-effectiveness analysis. Lung Cancer, 159, 153–161
2. Moriña, D., & Díaz, M. (2017). Calibration Approach Impact on Health and Cost-Effectiveness Outcomes in a Decision Analytic Framework. *Value in Health*, *20*(9), A407–A408
3. Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems*, *24*.
4. Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.
5. Duvenaud, D. K. (2014). *Automatic Model Construction with Gaussian Processes*. 157.
6. Gardner, J. R., Kusner, M. J., Xu, Z., Weinberger, K. Q., & Cunningham, J. P. (2014). Bayesian optimization with inequality constraints. *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, II-937-II–945.
7. Cartis, C., Massart, E., & Otemissov, A. (n.d.). *Constrained global optimization of functions with low effective dimensionality using multiple random embeddings*. 42.

8. Duvenaud, D. K., Nickisch, H., & Rasmussen, C. (2011). Additive Gaussian Processes. *Advances in Neural Information Processing Systems*, *24*.

9. Lu, X., Boukouvalas, A., & Hensman, J. (2022). Additive Gaussian Processes Revisited. *Proceedings of the 39th International Conference on Machine Learning*, 14358–14383.

10. Souza, A., Nardi, L., Oliveira, L. B., Olukotun, K., Lindauer, M., & Hutter, F. (2021). Bayesian Optimization with a Prior for the Optimum. In N. Oliver, F. Pérez-Cruz, S. Kramer, J. Read, & J. A. Lozano (Eds.), *Machine Learning and Knowledge Discovery in Databases. Research Track* (pp. 265–296). Springer International Publishing.

11. Cartis, C., Massart, E., & Otemissov, A. (2022). Bound-constrained global optimization of functions with low effective dimensionality using multiple random embeddings. *Mathematical Programming*.

12. Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. (2016). Fast Direct Methods for Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 252–265.

13. Binois, M., & Wycoff, N. (2022). A Survey on High-dimensional Gaussian Process Modeling with Application to Bayesian Optimization. *ACM Transactions on Evolutionary Learning and Optimization*, *2*(2), 8:1-8:26.

14. Nguyen, Vu & Rana, Santu & Gupta, Sunil & Li, Cheng & Venkatesh, Svetha. (2016). Budgeted Batch Bayesian Optimization. 1107-1112. 10.1109/ICDM.2016.0144.

15. Azimi, J., Jalali, A., & Fern, X. Z. (2012). Hybrid batch Bayesian optimization. *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 315–322.

16. Daulton, S., Balandat, M., & Bakshy, E. (2021). Parallel Bayesian Optimization of Multiple Noisy Objectives with Expected Hypervolume Improvement. *Neural Information Processing Systems*.

17. Wang, K., Pleiss, G., Gardner, J., Tyree, S., Weinberger, K. Q., & Wilson, A. G. (2019). Exact Gaussian Processes on a Million Data Points. *Advances in Neural Information Processing Systems*, *32*.

18. Munawar, A., Wahib, M., Munetomo, M., & Akama, K. (2009). Theoretical and Empirical Analysis of a GPU Based Parallel Bayesian Optimization Algorithm. *2009 International Conference on Parallel and Distributed Computing, Applications and Technologies*, 457–462.