Categorical Bayesian Optimization for Indirect Hard Model Selection

Luise F. Kavena, Daniel Jungena, Jan G. Rittiga, Alexander Mitsosa,b,\*

aRWTH Aachen University, Process Systems Engineering (AVT.SVT), Aachen, Germany

bJARA-SOFT, Aachen, Germany

*\*amitsos@alum.mit.edu*

Abstract

Model-based evaluation of Raman spectroscopy data is routinely conducted by use of indirect hard modeling (IHM). To develop a suitable IHM for a given analyte system, multiple decisions regarding the model settings must be made. These decisions require expert knowledge and exhaustive model development to identify the best suitable IHM for the analyte system. Herein, we propose the application of Bayesian optimization (BO) for the selection of most suitable settings for IHM development. We apply the BO algorithm available in the open-source software package BoFire to support the decision-making in the IHM development process. The results indicate that by leveraging BO, the IHM generation can be significantly accelerated and IHM performance can be exploited.

**Keywords**: Bayesian optimization, categorical variables, Raman spectroscopy.

* 1. Introduction

Indirect hard modelling (IHM) constitutes a physically-supported approach that predicts concentrations from spectral data such as near infrared (NIR) or Raman spectroscopy (Alsmeyer et al., 2004). The IHM method represents a multivariate regression technique, enabling concentration predictions of mixtures comprising overlapping component peaks (Alsmeyer et al., 2004, Kriesten et al., 2008). Meanwhile, the method relies on small calibration data sets (Alsmeyer et al., 2004, Echtermeyer et al. 2021). Finding the optimal IHM is a laborious and time-consuming task. Depending on the applied settings and the number of calibration measurements, the model generation can take several minutes. Additionally, the current state-of-the-art software PEAXACT lacks a programmable interface. The generation of an IHMs entails several decisions regarding the model settings. The performance of the IHM, quantified by calculating the coefficient of determination (R2) and root mean squared error of cross-validation (RMSECV) value for each component in the system under consideration, depends highly on the selected settings. Current IHM development often lacks a systematic approach to optimizing the model performance. We propose the strategic development of IHM models using Bayesian optimization (BO) with categorical inputs. BO is an established, efficient, and data-driven optimization method for intricate processes and (costly) experiments. We leverage BO algorithms that can handle categorical input variables to enhance the IHM development. We benchmark the strategic IHM development via BO with a random sampling of the categories and exhaustive sampling, i.e., creating all possible combinations of model settings.

* 1. Methods
     1. Indirect Hard Modeling

Within the development process of IHM evaluation models, several decisions must be made, e.g., regarding the qualitative spectral analysis, pretreatment, and hard model settings. In this work, we consider the choice of type of baseline and standardization for spectral pretreatment and fitting mode as decision variables in the problem setup for IHM development, since they have been shown to highly affect the model performance. As a case study, we develop an IHM for the mixture during the polymerization of *N*-Isopropylacrylamide-based microgels. Hence, the IHM components include the monomer (NIPAM), the polymer (PNIPAM), and the solvent (deionized water).

* + 1. Bayesian Optimization Algorithms

BO algorithms that can deal with categorical input variables include the open-source software packages: BoFire, CoCaBo (Ru et al., 2019), GPyOpt, Gryffin (Häse et al., 2021), Nemo-Bo, and NEXTorch (Wang et al., 2021). These software packages differ in the type of surrogate models and acquisition functions and optimization algorithms incorporated. All software packages can deal with continuous, discrete, and categorical type input variables. For BO-supported indirect hard modeling, we use the BoFire software package and compare the performance to random selection of sampling points. Furthermore, the data is approximated by a random forest surrogate model because random forest models are inherently well suited to handle categorical variables.

* 1. Results

We generated all possible IHM evaluation models with the combinations of settings, resulting in a total of 80 IHMs. In Figure 1, the progression of the objective values (R2 and RMSECV) over the number of experiments is shown when applying the BoFire software package. Reaching a high R2 value and a low RMSECV value for each component of the polymer system in the model with a low number of experiments are desirable. Figure 1 shows the 68 experiments conducted after initialization with 12 experiments. Within the first 20 experiments, only two instances with performance decrease (visually detectable by the two spikes in the progression) occur. Overall, it becomes clear that the algorithm follows a strategic approach, as the last experiments are stepwise resulting in a worse performance. Yet, a suitable stopping criterion needs to be defined for future applications of the BO supported method, as based on the exemplary progression of the objective values depicted in Figure 1 different criteria, e.g., number of experiments resulting in similar outcome or a fixed number of experiments to be conducted, would have yielded a different model performance. In contrast, randomly picking IHM settings to test results in high fluctuations of model performances due to the missing strategy. Therefore, the strategic BO-supported approach is preferable.

Ein Bild, das Text, Reihe, Diagramm, Schrift enthält.

Automatisch generierte Beschreibung

**Figure 1** Results of the IHM development supported by BoFire. Left: R2 objective value for all components in the IHM evaluation model over number of experiments. Right: RMSECV objective value for all components in the IHM evaluation model over number of experiments.

* 1. Conclusions

In summary, the proposed method involving BO for IHM development allows an efficient decision-making process, focusing on the most relevant model settings for analysis. The findings indicate that BO is promising for selecting IHM settings and could be used in future studies for IHM evaluation of more complex systems involving multiple components and with additionally considered IHM settings, e.g., fingerprint region or number of peaks in deriving pure component models. By supporting the decision-making process in IHM development, the most suitable settings resulting in the best possible model performance can be determined with reduced effort.

References

F. Alsmeyer, H.-J. Koß, and W. Marquardt, 2004, Indirect spectral hard modeling for the analysis of reactive and interacting mixtures, Applied Spectroscopy, 58(8):975–985.

BoFire: Bayesian Optimization Framework Intended for Real Experiments, Github: https://github.com/experimental-design/bofire.

A. Echtermeyer, C. Marks, A. Mitsos, and J. Viell, 2021, Inline Raman Spectroscopy and Indirect

Hard Modeling for Concentration Monitoring of Dissociated Acid Species, Applied

spectroscopy, 75(5):506–519.

GPyOpt: A Bayesian Optimization framework in python, 2016. Github: http://github.com/

SheffieldML/GPyOpt.

F. Häse, M. Aldeghi, R. J. Hickman, L. M. Roch, and A. Aspuru-Guzik, 2021, Gryffin: An algorithm for Bayesian optimization of categorical variables informed by expert knowledge, Applied Physics Reviews, 8(3):031406, Github: https://github.com/aspuru-guzik-group/gryffin.

E. Kriesten, F. Alsmeyer, A. Bardow, and W. Marquardt, 2008, Fully automated indirect hard modeling of mixture spectra, Chemometrics and Intelligent Laboratory Systems, 91(2):181-193.

Nomadic Exploratory Multi-objective Optimisation (NEMO), Python Package: https://pypi.org/project/nemo-bo/.

PEAXACT version 5.8, S-PACT GmbH, Aachen, Germany.

B. Ru, A. S. Alvi, V. Nguyen, M. A. Osborne, and S. J. Roberts, 2019, Bayesian optimisation over multiple continuous and categorical inputs, Github: https://github.com/rubinxin/CoCaBO\_code.

Y. Wang, T.-Y. Chen, and D. G. Vlachos, 2021, Nextorch: A design and bayesian optimization toolkit for chemical sciences and engineering, Journal of chemical information and modeling, 61(11):5312–5319, Github: https://github.com/VlachosGroup/nextorch.