Scotland Vaping Project Report (6.6-6.19)*

Yanpei Cai†

19 June 2025

Simplified ABM from UK Smoking Contagion Model

As per previously discussed, the ABM is simplified to have only four interaction-based parameters, with $\beta_{N,S\to S,S}$ and $\beta_{S,Q\to Q,Q}$ are fixed by following the parameter estimation from Christakis et al ($\beta_{N,S\to S,S} = 0.407185$ and $\beta_{S,Q\to Q,Q} = 0.352138$). The other two parameters, $\beta_{S,N\to Q,N}$ and $\beta_{Q,S\to S,S}$, are calibrated using BOLFI which iteratively minimises the discrepancy between the simulated and observed data to infer their optimal values and posterior distributions.

BOLFI

BOLFI treats parameter calibration as an optimisation problem, where a Gaussian Process (GP) surrogate model approximates the discrepancy between simulations and observations. It sequentially selects new parameter values via Bayesian optimisation (using acquisition functions) that strategically balance exploration of high uncertainty regions with exploitation of low discrepancy areas, iteratively refining the GP surrogate model until convergence to optimal parameters while simultaneously quantifying their posterior uncertainty through the GP's predictive distribution.

Observed Data and Simulated Data

Scotland is witnessing a dramatic surge in vaping in recent years, especially among young adults. As revealed by the Scottish Health Survey (SHeS) smoking tables, the use of e-cigarettes or vaping devices among young adults aged 16 to 24 years has seen a significant increase from 5% to 22% in five years (2019-2023). We used data collected from young adults aged 16 to 24 years and between 2019 and 2023 to represent the alarming circumstances facing us and to avoid the potential influences caused by vital dynamics. Note that data for year 2020 were missing for uncertain reasons, so we omitted this year in both the observed and simulated data. Simulated data were generated by the simplified ABM simulator with different network models.

Priors for $\beta_{N,S\to S,S}$ and $\beta_{S,Q\to Q,Q}$

As we do not have prior knowledge of the priors, we chose $Uniform(0, 1)$ for both $\beta_{N,S\to S,S}$ and $\beta_{S,Q\to Q,Q}$.

†Artificial Intelligence and its Applications Institute, School of Informatics, The University of Edinburgh
Email Address: yanpei.cai@ed.ac.uk

*This is the first-stage report of the project of modelling vaping contagion in Scotland.

Summary Statistics and Discrepancy Measure

Since our data are multidimensional data, to monitor the trend in each column we chose cell-to-cell differences as summary statistics, and calculated the euclidean distance between the simulated and observed data as discrepancy measure. To reduce the effect that high discrepancies have on the GP surrogate model, we took logarithm of the discrepancies for the BOLFI target.

BOLFI Settings

To ensure the performance of the optimisation, we tested different combinations of number of evidence points before and after the optimisation process and acquisition noise variance for $\beta_{N,S\to S,S}$ and $\beta_{S,Q\to Q,Q}$ used in the acquisition function (LCBSC: Lower Confidence Bound Selection Criterion).

Coding Implementation

We further improved our code to a general version so that it can be used for multiple cases by simply changing a few inputs.

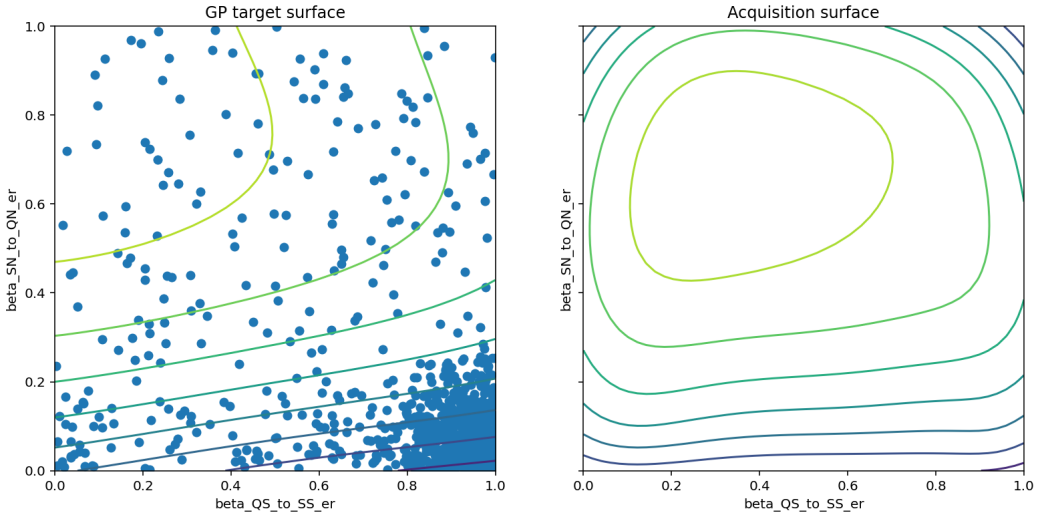
Results

We chose the results from BOLFI tested by a simplified ABM with Erdös-Rényi graph ($n = 1000$ and $p = 0.003$) to show in this section. We first presented the optimal values of $\beta_{N,S\to S,S}$ and $\beta_{S,Q\to Q,Q}$ and the log-discrepancy, and then we visualised the results by various plots. The plots for GP target model and log-discrepancy against parameter value were generated by default methods of BOLFI. We further improved the selection of results by adding a customised contour plot for the predicted mean discrepancies, a customised surfaces plot for the predicted mean discrepancies and their 0.1 and 0.9 quantiles, and two customised contour plots for the posterior's pdf and log-pdf.

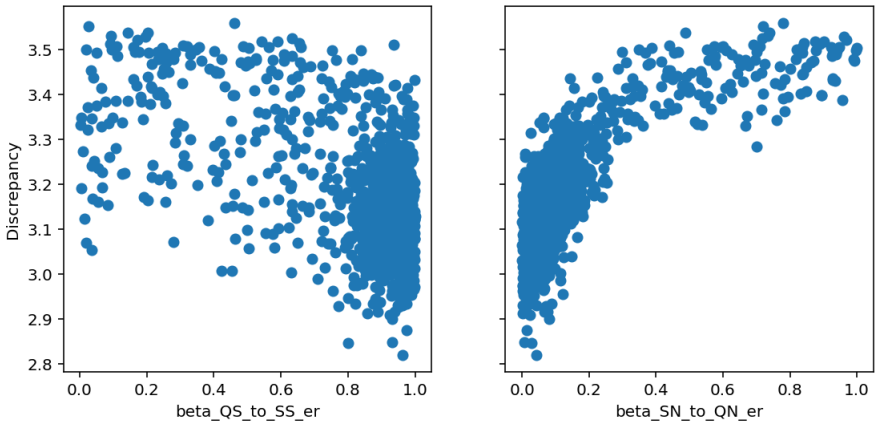
Optimal Parameters and log-discrepancy

```
Progress [==================================================] 100.0% Complete
INFO:elfi.methods.posteriors:Using optimized minimum value (3.0333) of the GP
discrepancy mean function as a threshold
INFO:elfi.methods.posteriors:Using optimized minimum value (3.0333) of the GP
discrepancy mean function as a threshold
Parameters: ['beta_QS_to_SS_er', 'beta_SN_to_QN_er']
Optimal values: [0.96366351 0.04241617]
Optimal log-discrepancy: 2.8204213374411635
Parameters: ['beta_QS_to_SS_er', 'beta_SN_to_QN_er']
Optimal values: [0.963663505701265, 0.04241617441540371]
Optimal log-discrepancy: 2.8204213374411635
```

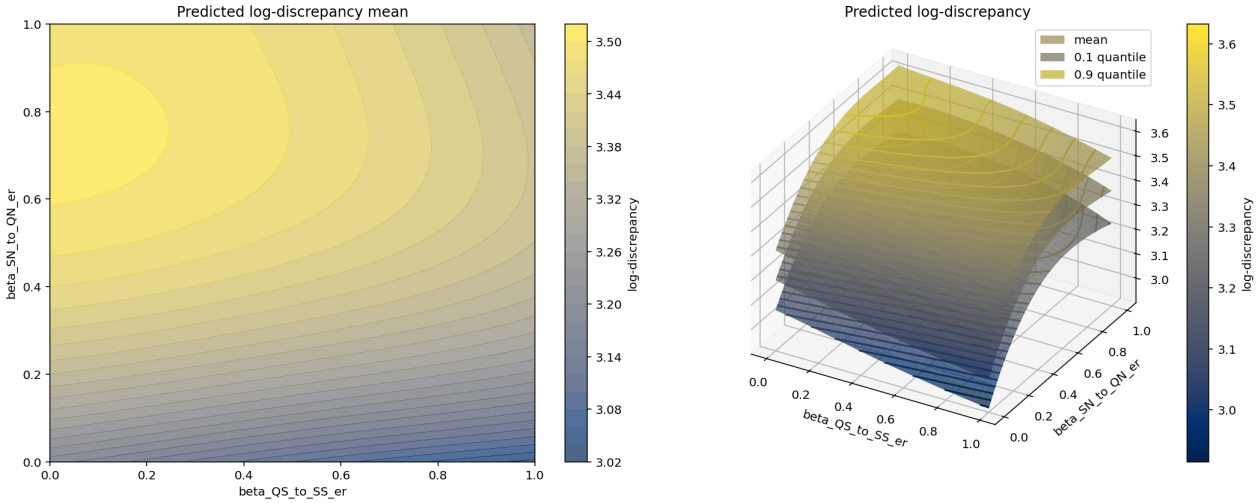
GP Target Model



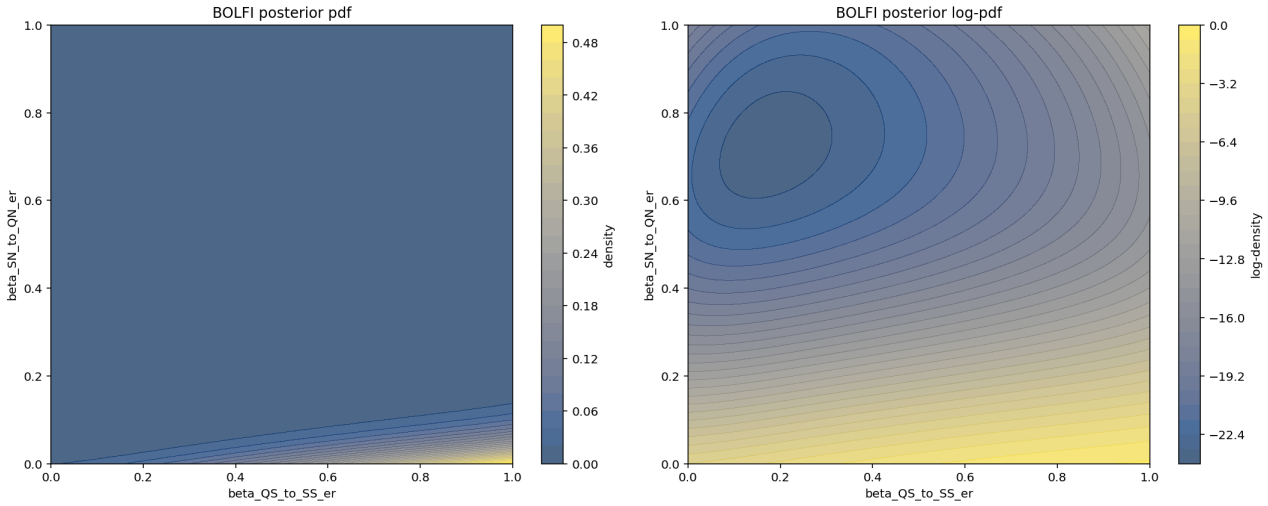
Log-discrepancy against Parameter Value (Default)



Log-discrepancy against Parameter Value (Contour and Surfaces)



BOLFI Posterior (Contours)



Observed Data and Simulated Data with Optimal Parameters

Index	% NEVERSMOKER	% SMOKER	% QUITTER
2019	78	5	17
2021	75	6	19
2022	61	15	24
2023	59	22	19

Index	% NEVERSMOKER	% SMOKER	% QUITTER
2019	78	5	17
2021	74	9.8	16.2
2022	71.7	12.5	15.8
2023	69	15.5	15.5

Discussion

As can be seen from the last section, the discrepancy between the simulated data and observed data is still very large, though we have used the optimal parameter values. We believe that this level of discrepancy is primarily caused by the inappropriate value used for $\beta_{N,S\to S,S}$, as there is a considerable mismatch between the % NEVERSMOKER columns of the observed and simulated data. Since $\beta_{N,S\to S,S}$ is the only parameter that manipulates the state change from a never-smoker to a smoker, this inappropriate value use would also influence the calibration for other parameters.

The observed data might also be improved from another side of view. We could possibly combine the data collected from the youngsters with our current data to represent the trend of e-cigarette or vaping device use within a single group of people over years. This would be particularly useful to monitor the trend over a longer period of time when we also intend to consider the vital dynamics.