Supplementary Material

Adaptive data collection for intra-individual studies affected by adherence

Greta Monacelli[1 2 *], **Lili Zhang**[1 2], **Winfried Schlee**[3], **Berthold Langguth**[3], **Tomás E. Ward**[1 2], **Thomas B. Murphy**[3 4]
[1]Dublin City University
[2]Insight SFI Research Centre for Data Analytics
[3]University of Regensburg
[3]University College Dublin

[*]Corresponding author: Greta, Monacelli; greta.monacelli2@mail.dcu.ie

March 9, 2023

This document contains supplementary material on the implementation of the algorithms presented in **Adaptive data collection for intra-individual studies affected by adherence** by Monacelli et al. (2022).

S.1. Practitioner guidelines for the selection of the starting point

Three are two competing goals which require the attention of the practitioner in the selection of the starting point $S^\star$. On one hand, $S^\star$ has to be large enough so that the method of moments estimation provides a "good enough" estimate of the parameters and does not encounter computational errors. A sufficient large number of observations is necessary to obtain a "good" estimate through the method of moments. Indeed, as the sample size increases, the estimation of the mean squared error decreases, see Figure 1. Notice that if we start collecting the additional task data at time $S^\star$, then smallest sample size considered by Algorithms 1 and 2 is $S^\star - 1$. Hence, the preference for a large $S^\star$.

Moreover, in order to avoid computational errors, the condition

$$0 < \hat{\sigma}_s^2 < \hat{\mu}_s(1 - \hat{\mu}_s). \tag{1}$$

has to be satisfied for all $s \in \{S^\star - 1, \dots, N\}$, where $\hat{\sigma}_s^2$ and $\hat{\mu}_s$ are the method of moments estimates of the variance and expected value of a Beta distributed sub-sample $\mathbf{x}_{1:s}$. This condition is always satisfied if the estimated expected value and variance are replaced by their true values. Assuming that the estimates $\hat{\sigma}_s^2$ and $\hat{\mu}_s$ tend to the true values σ^2 and μ^2 as s tends to infinity — as it is happens for good estimates, then we expect Condition (1) to hold for large data sample — i.e. large values of $S^\star$, see also Figure 1. Nevertheless, because of the structure of both Algorithms 1 and 2, it is necessary to estimate the parameters for small sample sizes. This may result in computational problems both in a simulated and real setting as the likelihood of obtaining a constant — or almost constant — sequence of observations is increased. In those cases, the sample variance is zero — or almost zero, violating the left hand side of Condition (1). This causes the estimates to either be undefined or explode. Alternatively, the right hand side of Condition (1) may not be satisfied, producing negative estimates for the parameters. This problem is non-trivial as the experimental design considered and the real data analysed can contain sequences of all 0 and 1 — or values too close to them. Then, both $\hat{\sigma}_s^2$ and $\hat{\mu}_s$ tend to zero, sometimes resulting in a violation of Condition (1), see also Figure 3 and Section S.3.

we may obtain too small estimates of the variance and of the right hand expression of Condition (1). All these suggest that a greater $S^\star$ is less likely to produce errors in the code.

On the other hand, $S^\star$ needs to be small enough so that the risk of collecting too little data for non-adherent subjects is low. As the adherence of the subjects decreases exponentially (as it is shown in Figure 2, page 9 of the main manuscript), an early starting point increases the chances of obtaining enough data for a larger number of participants. The final selection of $S^\star$ needs to achieve a balance between these two goals. In this work, we set $S^\star = 6$. Plotting the optimal region $A^\star$ can further aid the researcher in the selection of the experimental design and highlight how the choice of the starting point impacts the significance of the following triggers, see Figure 2.

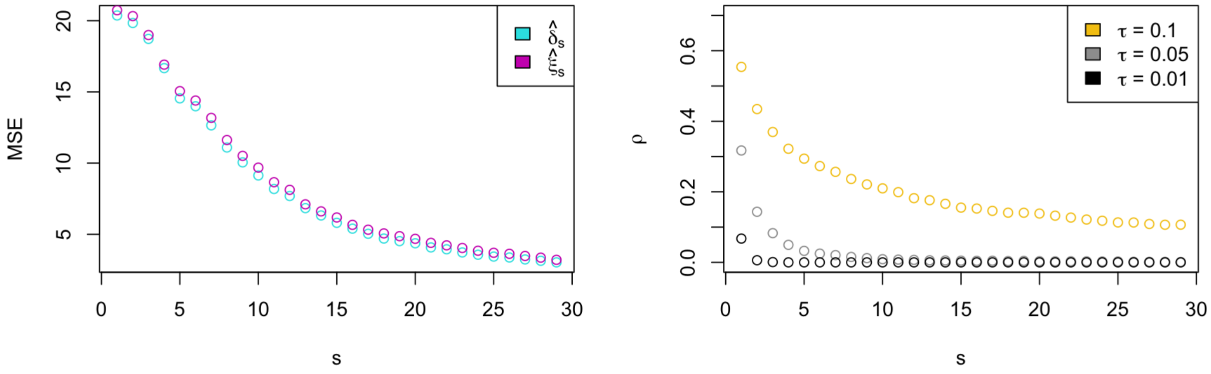

Figure 1: On the left, a simulation of the mean squared error (MSE) for different data sample sizes in the method of moment estimation of the Beta distribution parameters. For each sample size s, we simulated 5000 pair of Beta distribution parameters from a uniform distribution in $[0.5, 10]$. Then, for each pair, s observations were sampled and the method of moments estimates computed. As expected, the MSE of these estimates decreases as the sample size increases. On the right, an analysis of the prevalence of data which violates the left hand side of Condition (1) in the same simulation setting. In particular, ρ is the approximated probability that the variance for a sample of size s from a Beta distribution is smaller than τ. We notice that ρ decreases as the sample size increases.

S.2. Stopping rule

Reducing the variance of the final number of triggers V is a second meaningful objective when computing the optimal significance level $\alpha^\star$: indeed, once the mean has been fixed, this concentrates the values of V in proximity of $\mathrm{E}(V)$. Recall that $\mathrm{Var}(V) = (N - S + 1)\alpha(1 - \alpha)$ and define a second expected utility function $U_2 : \{2, \dots, N\} \times [0, 1] \longrightarrow \mathbb{R}$ as

$$U_2(S, \alpha) = -\mathrm{Var}(V). \tag{2}$$

A contour plot of the continuous extension of the function U_2 into the domain $D = [2, N] \times [0, 1]$ is given in Figure 2.

As the variance is only a secondary goal, we consider the design optimization problem for U_2 constrained to the domain $A^\star$, i.e.

$$(S', \alpha') \in \arg \max_{(S^\star, \alpha^\star) \in A^\star} U_2(S^\star, \alpha^\star). \tag{3}$$

Then,

$$\begin{aligned} U_2(S^\star, \alpha^\star) &= -(N - S^\star + 1)\alpha^\star(1 - \alpha^\star) \\ &= -v(1 - \alpha^\star) \end{aligned}$$

where the second-last step is justified by the constraint $(S^\star, \alpha^\star) \in A^\star$. Thus, the optimal solution for the optimization problem in Equation (3) is $(N - v, 1)$. This design corresponds to triggering the additional task in the last v interactions of the subject with the app. There are two main problems with this approach: first, in real applications, there is no guarantee that the subject will adhere to the study until the end; second, in order to trigger the additional task on significantly high or low values x_t we aim to implement α as small as possible.

In conclusion, reducing the variability of the total number of triggers produced by the algorithm results in a trivial solution for the design optimization problem. Nevertheless, the argument above shows that there is a trade-off between the need for precision — decreasing the variance of the total number of observations — and the need for unlikeliness — keeping the significance level α small. In this paper we assume that both $S^\star$ and, subsequently, $\alpha^\star$ are chosen to be relatively small. This can result in high variance for the final number of triggers. In other words, the algorithm may trigger either too many or too few tasks for certain subjects.

Hence, we suggest to add some simple deterministic rules to either force an additional task if there were not enough or stop their triggering if too many have already been triggered. In this paper, we added a stopping rule which, for each subject, prevented the algorithm to trigger the additional task more than 10 times. In Table 1, we show that, at least in a simulated context, this choice has almost no impact on Algorithms 1 and 2.

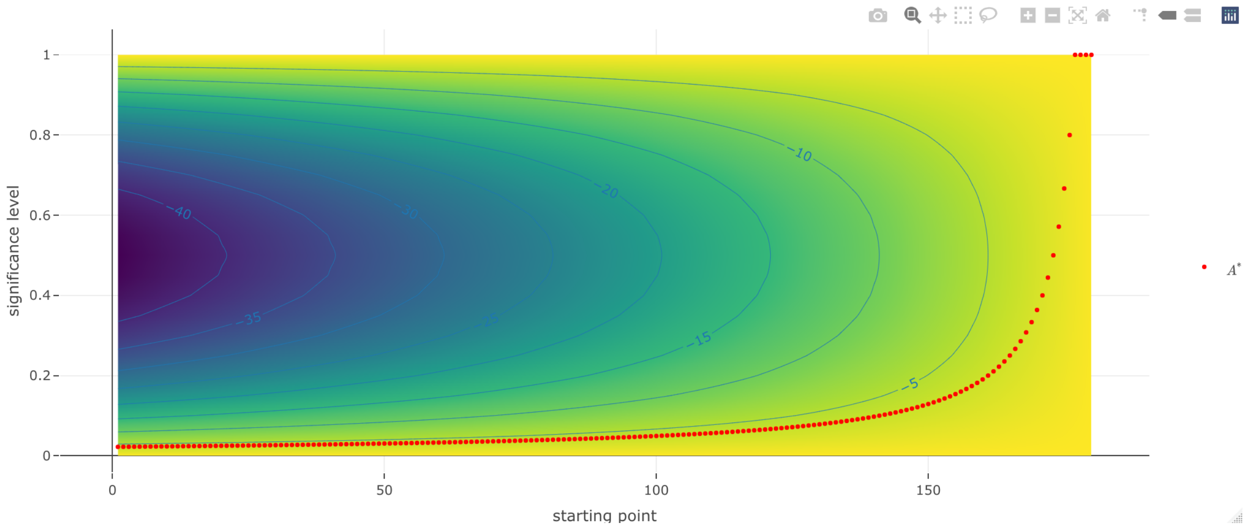

Figure 2: The red dots represent the set $A^\star$ computed for an experiment of length $N = 180$ and assumed $v = 4$. Notice that, as a function of the optimal starting point $S^\star$, the optimal significance level $\alpha^\star$ increases slowly in the first half of the experiment. This suggests that any starting point S in the first half of the experiment should not impact significantly the final performance of the algorithm for completely adherent subjects. The extension of the function U_2 in Equation (2) into the domain $D = [2, N] \times [0, 1]$ is represented by the contour plot.

Table 1: In the burden side of the Table we report the proportion of subjects for which at least one additional task has been prevented by a stopping rule set to 10. In the information loss side of the Table we report the proportion of subjects for which at least a false positive or a false negative was introduced by the same stopping rule.

	Burden			Information Loss		
	static	alg1	alg2	static	alg1	alg2
Sim data 1	0.218	0.002	0	0.219	0.001	0
Sim data 2	0.201	0.073	0	0.201	0.064	0

S.3. Other computational problems

A careful selection of a starting point $S^\star$ following the instructions in Section A.1. partially solves the computational problems of the estimation of the Beta parameters. Nevertheless, we could still occasionally encounter computational errors due to a violation of Condition (1). The probability of this happening converges to zero, but it is always non-zero for a finite samples. Therefore, only when Condition (1) is not satisfied, we added in our code some artificial data to regularize the problem. Since one artificial data point may not be enough to increase the variance of the sample — a subject may report a constant sequence of exactly that value — we decided to add two artificial observations from the set $[0, 1]^2$.

To show how this choice impacts the estimation of the parameters, we conducted a sensitivity analysis. We show the results in Figure 3. We conclude that points placed on the corners of $[0, 1]^2$ and on the diagonal $\{(h, h) \in [0, 1]^2\}$ should be avoided. Points in the corners do not guarantee that the right hand side of Condition (1) is satisfied since we increase the likelihood of obtaining a sequence of only 0 or 1 — or a sequence of values too similar to it. On the other hand, points close to the diagonal results in a higher mean squared error. In this work, we used the artificial data 0.4 and 0.6.

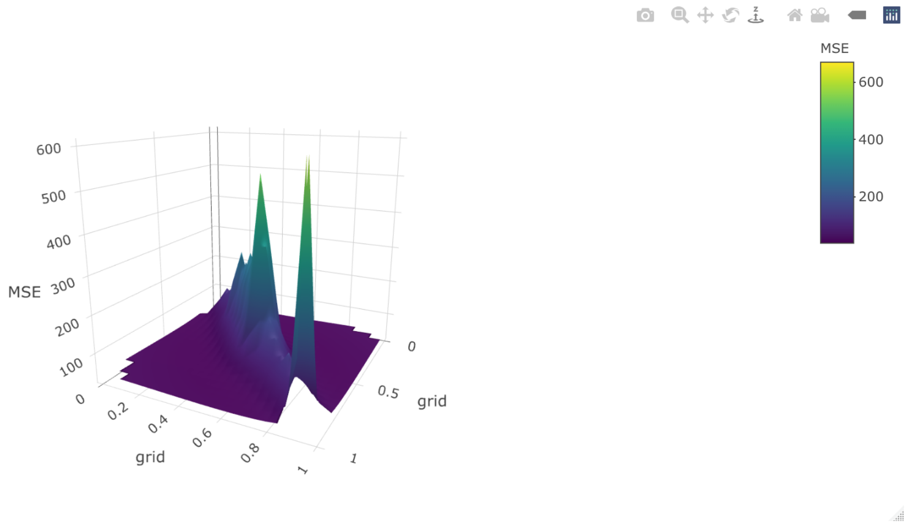

Figure 3: Sensitivity analysis fro the impact of the artificial data on the Beta parameters estimation. The Figure was obtained by considering 1624 simulated samples of length 5 for which Condition (1) is not satisfied. We only considered sample of length 5 as it is the smallest sample size considered in this work, i.e. $S^\star - 1 = 5$, and Condition (1) is more likely to not be satisfied for small sample sizes. The mean squared error of the Beta parameters estimation obtained by adding artificial — taken from a discretization of the set $[0, 1]^2$ — was computed. The figure shows the sum of the MSE for both the first and second parameters.