RUTGERS
School of Arts and Sciences
Department of Statistics

RUTGERS UNIVERSITY

ETHICAL STATISTICAL LEARNING

Effection of Swapping on the Multinomial Example

Author: Qiru Pan, Yunhao Li, Yicong Li
Supervisor: Linjun Zhang

Dec. 2023

Introduction to Data Swapping

The swapping algorithm is about a set of variables into two distinct groups: swapping variables (V_{Swap}) and holding variables (V_{Hold}). Each record is independently chosen based on a predetermined probability defined by a swap rate parameter, p. Once selected, the V_{Swap} values of these records undergo a random shuffle while the V_{Hold} values remain unchanged.

Sometimes, the swapping process is limited to records that have identical values in a subset of the holding variables. This subset, which may be empty, is referred to as the matching variables, denoted as V_{Match}.

The swapping algorithm is $(c_{swap}, d^u_{HamS}, \epsilon_D)$ differential private with privacy loss budget:

$$\epsilon_D = \begin{cases} \ln(b+1) - \ln o & \text{if } 0 < p \leq 0.5 \\ \max\{\ln o, \ln(b+1) - \ln o\} & \text{if } 0.5 < p < 1 \end{cases}$$

with $o = p/(1-p)$ and b is (roughly) the largest stratum size. c_{swap} is a swapping invariant and d^u_{HamS} is the Hamming distance on unordered datasets.

Data Swapping on Multinomila Example

A. Generating Model

The data pertains the responses to two questions, both of which are binary variables. These answers form a n * 2 table. We employ a random variable $y = (y_1, y_2, y_3, y_4)$ with multinomial distribution to simulate the 4 types of different answer combinations ((0,0), (0,1), (1,0), (1,1)) with probability θ_1, θ_2, θ_3 and θ_4.

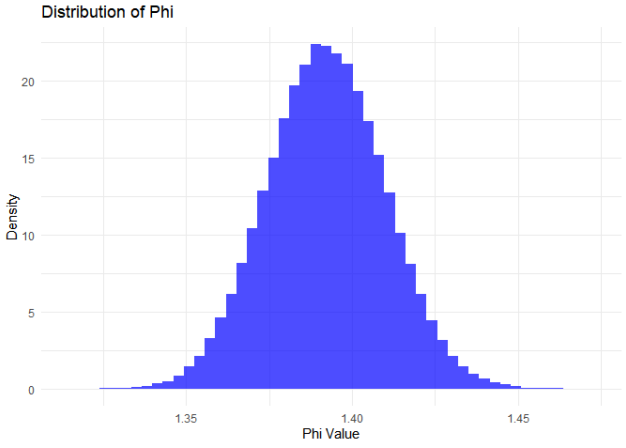
Our primary concern is that by applying the swap algorithm that satisfies Differential Privacy (DP) on the n * 2 data, without altering any marginal totals, would it impact the statistical measure ϕ, defined as:

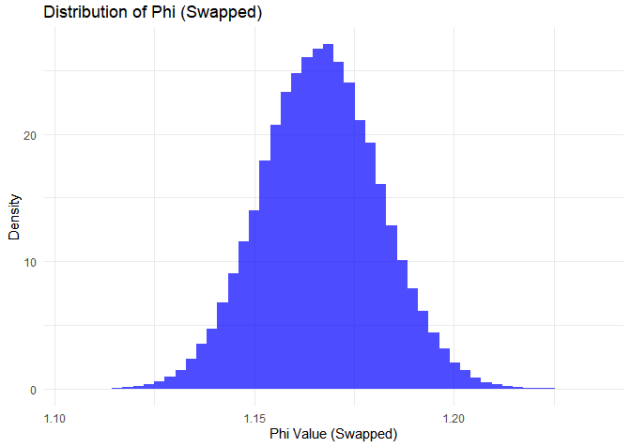
$$\phi = \frac{\theta_1 * \theta_4}{\theta_2 * \theta_3}$$

The ground truth we use to generate data is $\theta^* = (0.27, 0.23, 0.23, .0.27)$.

To obtain the posterior distribution of the generated data, the assumption here is that the prior distribution of θ is a Dirichlet distribution. Given that the likelihood is a multinomial distribution and the multinomial distribution is conjugate to the Dirichlet distribution, we can derive the posterior distribution of θ, which will also be a Dirichlet distribution:

$$\begin{aligned} \text{Prior} &: \theta \sim \text{Dir}(1,1,1,1) \\ \text{Likelihood} &: y|\theta \sim \text{Mult}(\text{n}, \theta) \\ \text{Posterior} &: \theta|y \sim \text{Dir}(1+y_1, 1+y_2, 1+y_3, 1+y_4) \end{aligned}$$



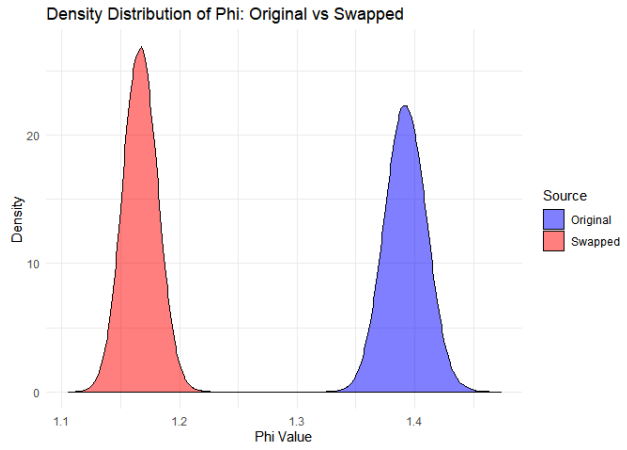


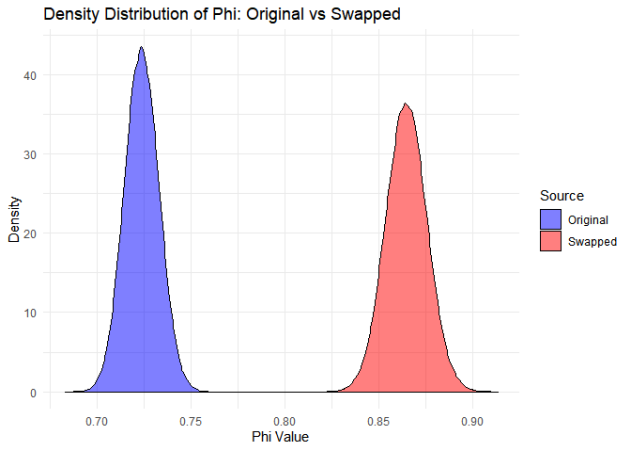
The top left graph represents the distribution of the statistical measure ϕ. The peak of this distribution aligns with the ground truth value calculated as $(0.27 * 0.27)/(0.23 * 0.23) = 1.378$.

We apply the swapping algorithm to the dataset. In the n * 2 table, we take the first column as V_{Hold} and the second column as V_{Swap}. We do not consider V_{Match} in this context. When using the derangement function to perform swapping, such a setting satisfies differential privacy when certain assumptions are met. We set the swap rate $p = 0.5$ which means 50% of the data will be swapped.

It's important to note that this does not mean that 50% of the data has been perturbed, as not all records that have been swapped were altered. We can consider a swap in two scenarios. If both records have the same V_{Hold} or V_{Swap} values, then the value of ϕ remains unchanged. The case where V_{Swap} values are the same is trivial. When V_{Hold} value are the same, for instance, swapping (0,1) with (0,0), the outcome would be (0,0) and (0,1). Consequently, the corresponding values of y_1 and y_2 do not change. In the other scenario, if the V_{Hold} and V_{Swap} values of both records are different, then the value of ϕ will definitely change. For example, swapping (0,0) with (1,1) results in (0,1) and (1,0), and $y = (y_1, y_2, y_3, y_4)$ would correspondingly become $y = (y_1 - 1, y_2 + 1, y_3 + 1, y_4 - 1)$. Conversely, when (0,1) and (1,0) are swapped, the result is (0,0) and (1,1), so that y becomes $(y_1 + 1, y_2 - 1, y_3 - 1, y_4 + 1)$.

We still use the Dirichlet distribution as the prior for θ, just as in the aforementioned process. The obtained posterior distribution is illustrated in the top right figure. After smoothing both distributions, we place them in one graph, as shown in the bottom left figure, and observe that the distribution function has shifted to the left overall.





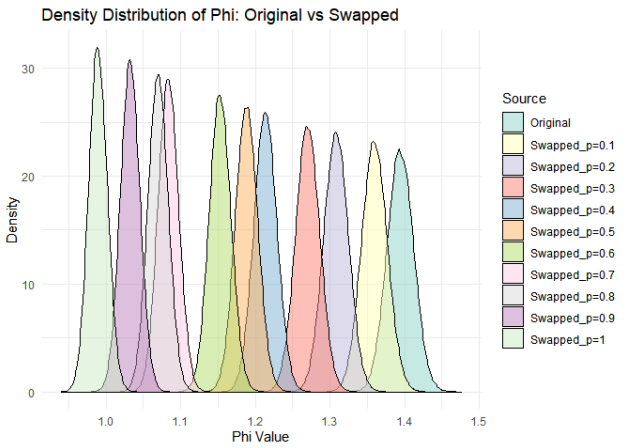
We aim to provide an intuitive explanation for the changes observed in this statistical measure. As can be seen from our previous example, y_1 and y_4, y_2 and y_3 always change simultaneously. In our data generation, since $\theta^* = (0.27, 0.23, 0.23, 0.27)$, y_1 and y_4 are expected to be larger. Due to the random selection of swap targets, there is a higher likelihood of y_1 and y_4 being chosen for swaps. This leads to a decrease in y_1 and y_4. further causing the value of ϕ to decrease. To some extent, the swap algorithm makes y more inclined towards a uniform distribution.

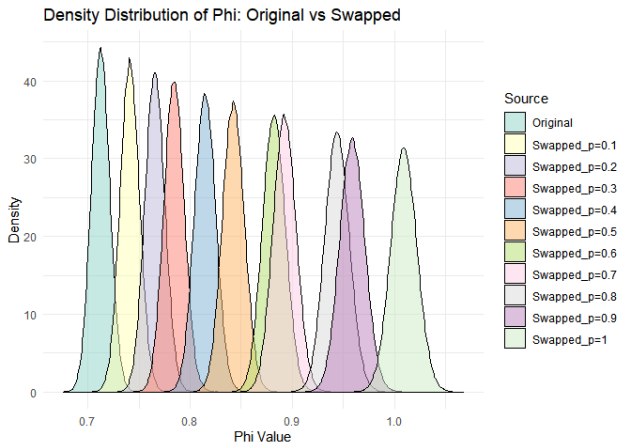
Using similar logic, we conduct experiments with a symmetric $\theta^* = (0.23, 0.27, 0.27, 0.23)$. All other steps remain the same as before. We observe from the top right figure that after the swap, the distribution of ϕ has shifted to the right overall. This is consistent with our intuitive understanding.

Based on this, we can preliminarily conclude that the swapping algorithm does change the value of the statistic, and if we use private data to perform statistical inference directly, we will end up with an incorrect conclusion.

B. Effect with Different p

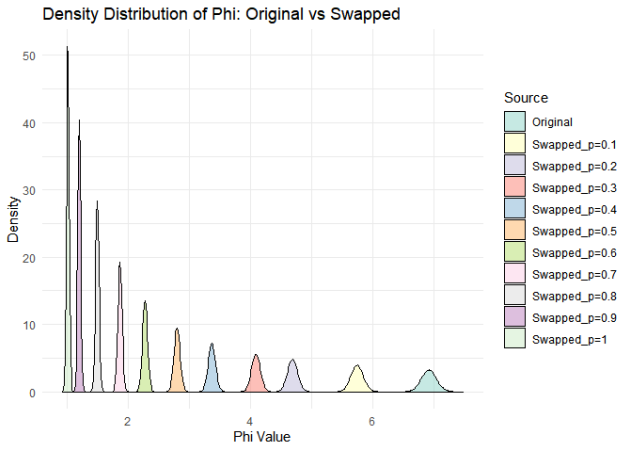
We still use $\theta^* = (0.27, 0.23, 0.23, 0.27)$ and the symmetric $\theta^* = (0.23, 0.27, 0.27, 0.23)$ to generate data. The result of the distribution of ϕ under different swap rate is as follow:





From the left-hand figure, it can be observed that as the swap rate increases, the peak of the ϕ distribution gradually becomes higher, indicating that the data is increasingly concentrated. At the same time, the overall distribution function moves closer to 1, which also suggests an increase in data uncertainty. The right-hand figure presents the exact opposite trend. The peak of the ϕ distribution gradually becomes lower while the overall distribution function moves closer to 1 from another direction. Combining the observations from both graphs, it appears that, in general, the larger the value of ϕ, the higher the peak of its distribution.

We can show it with a more extreme example with the ground truth of $\theta^* = (0.7, 0.1, 0.1, 0.1)$



As expected, the swapping algorithm makes extreme changes more elusive, and as described above, a ϕ far away from 1 will obtain a more dispersed distribution by sampling from the posterior distribution model. As the swap rate increases, the peak of the ϕ distribution gradually rises, and the shape of the probability density estimate becomes more concentrated and sharp.

Therefore, we can draw the following conclusion from the above figures: As p increases, the number of rows selected for swapping increases, so the internal structure (information) of the data itself is destroyed more. When $p = 1$, the structure is completely destroyed. At this time, regardless of the initial setting of θ, the mean value of ϕ is near 1.

C. Conclusions

The efficacy of the swapping algorithm in manipulating data is significantly influenced by the parameterization of θ^*. Variations in θ^* settings yield distinct alterations in the ϕ distribution, indicating a dependency on the chosen parameters.

There is an observable trend where the swapping algorithm appears to gravitate the ϕ value towards unity, a phenomenon potentially attributable to an increase in data homogeneity resulting from the swapping process.

Regarding Data Privacy: Analyzed from a data privacy enhancement perspective, the swapping methodology demonstrates promise in augmenting data anonymity and introducing stochastic elements. This enhancement in randomness serves to disrupt and possibly obfuscate inherent structures or patterns within the original dataset.

Concerning Data Structure Integrity: Conversely, when the objective encompasses the preservation of intrinsic data structures or patterns, alternative strategies may warrant consideration. The current observations suggest that the swapping mechanism may inadvertently distort or diminish original data characteristics.

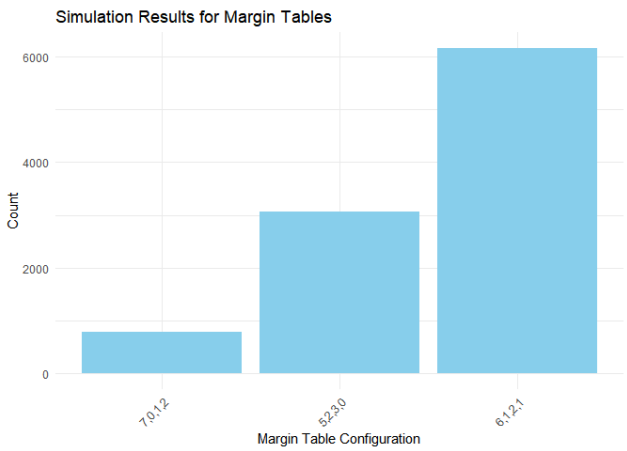
Bayesian Analysis with Swapped Data

We hope to utilize a Bayesian analysis-based algorithm, enabling us to infer and obtain true information under the condition that only the swap algorithm and private data are known.

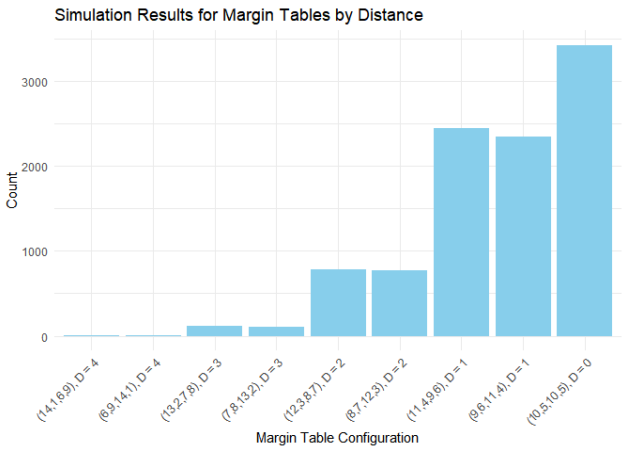
D. Simple examples

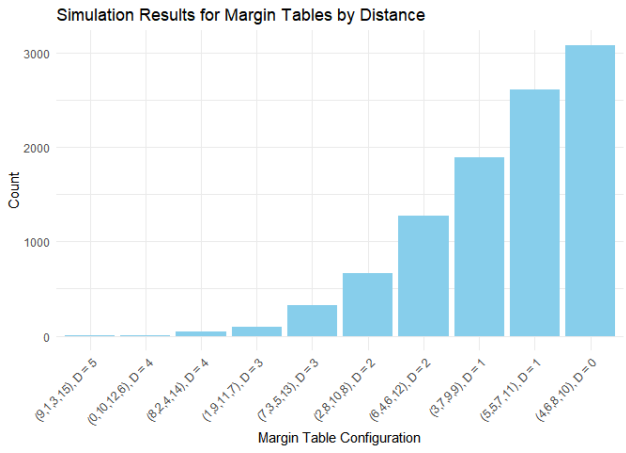
We want to start with a smaller and simpler example. Given a 2x10 marginal table corresponding to micro-data, we try to determine all potential marginal tables (also known as the support) that can be obtained through swap algorithms. Assuming the observed values are $y = (6, 1, 2, 1)$, after one effective swap, the result could be either $(7, 0, 1, 2)$ or $(5, 2, 3, 0)$. We conduct an swap simulation with N=100000 and try to use brute force to obtain the Monte Carlo estimate of $s|y$ when $p = 0.5$ where s is the true distribution before swapping.

The result is as below:



We proceed with a slightly more complex example using Monte Carlo simulation, setting $y = (10, 5, 10, 5)$. Here, we define a distance D between two random variables, which represents the minimum number of swaps needed to transform one y into another. For instance, the distance D between $(10, 5, 10, 5)$ and $(9, 6, 11, 4)$ is 1. From the figure, we can clearly see that as the distance increases, the frequency gradually decreases.





The results exhibit a rather unique shape because our distribution is symmetric in some sense. To add more generality, let's try another more random $y = (4, 6, 8, 10)$. The outcome is the upper right figure.

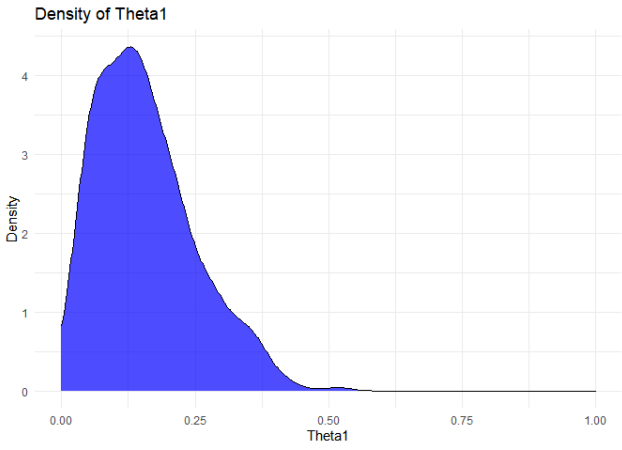
In our previous experiments, we learned how to determine the support for potential marginal tables that can be obtained through a swapping algorithm. Additionally, using Monte Carlo simulations, we estimated the likelihood of obtaining each marginal table given a specific margin and swap rate. Upon further observation, we also discerned the frequency patterns of obtaining marginal tables with varying 'distances' under different swap rates through the swapping algorithm.

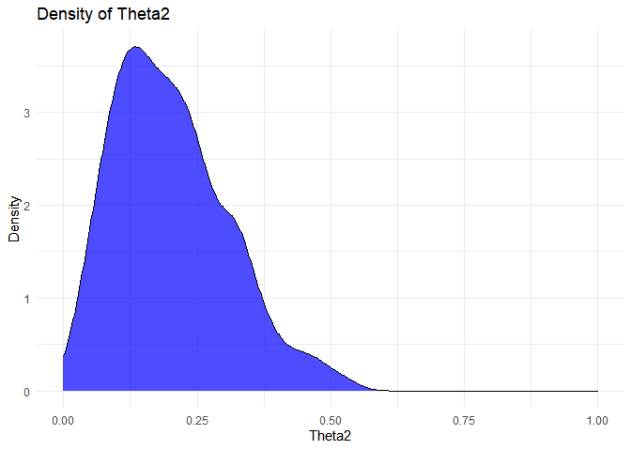
E. Statistical Inference with Approximate Bayesian Computation

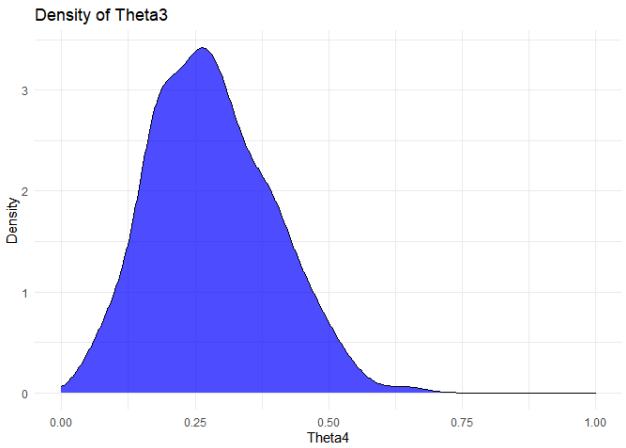
Given an observed result s_{obs}, we aim to estimate the true value of θ through Approximate Bayesian Computation (ABC). Firstly, we assume that the prior for $\theta^{(i)}$ is a Dirichlet distribution: $\theta^{(i)} \sim rmDir(1, 1, 1, 1)$. Then, we sample $y^{(i)}$ using a multinomial distribution: $y^{(i)}|\theta^{(i)} \sim rmMult(n, \theta)$. Following this, we apply the swapping algorithm to $y^{(i)}$ and have a swapped random variable $s^{(i)}$. The specific method of swapping is the same as described earlier. In each iteration, we compare the observed marginal table s_{obs} with the $s^{(i)}$. If $s_{obs} = s^{(i)}$, we retain the current parameter $\theta^{(i)}$, otherwise, we skip the current iteration.

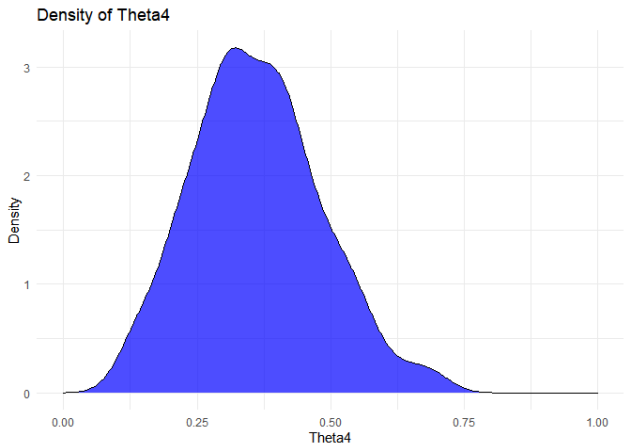
Through multiple iterations, we expect to obtain the distribution of the parameter θ concerning the observed data s_{obs}. This distribution will reflect the uncertainty and range of variation of the parameters given the observed data.

We set the ground truth of $\theta^* = (0.1, 0.2, 0.3, 0.4)$, the result of iterating 100,000 times is as follow:

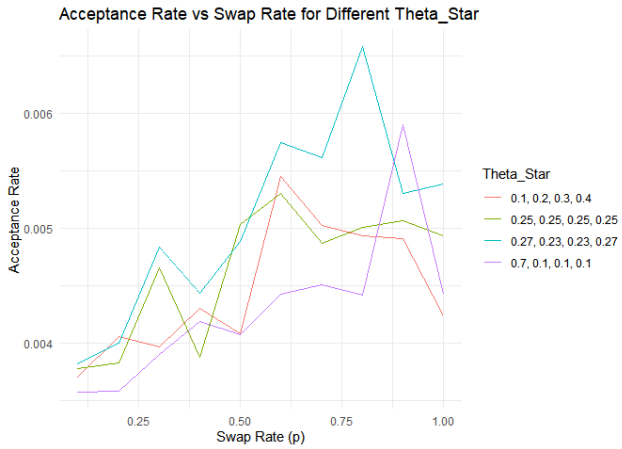








The figures show that the value of θ simulatied form ABC is close to the ground truth θ^*. To further validate the power of the experiment, it is necessary to record the acceptance rate and examine the impact of different settings on this rate. Therefore, in the experiment, we iterate 100,000 times for different swap rates p with different θ^* sets to observe their respective acceptance rates.



From the above figure, it can be observed that regardless of the true value of θ^*, the overall trend of the acceptance rate increases and then decreases with the increase in swap rate. Based on previous discussions, we believe that when p is close to 1, the data structure is completely disrupted, leading to the failure of ABC to function as intended.

Reference

1. Bailie, J., Gong, R., and Meng, X.-L. (2023). "Can Swapping be Differentially Private? A Refreshment Stirred, not Shaken." Retrieved from https://conference.nber.org/conf_papers/f178188.pdf.
2. Dalenius, T. and Reiss, S. P. (1982). "Data-Swapping: A Technique for Disclosure Control". Journal of Statistical Planning and Inference, 6(1), 73–85. doi: 10.1016/0378-3758(82)90058-1.
3. Fienberg, S. and McIntyre, J. (2004). "Data Swapping: Variations on a Theme by Dalenius and Reiss". In Privacy in Statistical Databases. doi: 10.1007/978-3-540-25955-8_2.
4. Gong, R. (2019). "Exact Inference with Approximate Computation for Differentially Private Data via Perturbations." arXiv:1909.12237.
5. Ju, N., Awan, J. A., Gong, R., and Rao, V. A. (2022). "Data Augmentation MCMC for Bayesian Inference from Privatized Data." arXiv:2206.00710.
6. Zhang, L. (2023). "STAT 16:960:690: Ethical Statistical Learning".