The Identity Fragmentation Bias: Online Appendix

Tesary Lin* Sanjog Misra†

Simulation A: Scenarios of Identity Fragmentation Bias

To demonstrate sources of estimation bias caused by identity fragmentation, we provide a simulation module where readers can specify various fragmentation and consumer behavior parameters and explore their impacts on the estimates. We also provide numerical examples that map to scenarios introduced in the paper. Below, we introduce the simulation setup and the results corresponding to each scenario.

The Data Generating Process

Following the setup in Section 3.1, we let consumer purchases change linearly with advertising exposure. We use the following values in all the examples below (users are free to choose different values):

- Number of consumers: $N = 1000$;
- Number of devices: $J = 2$;
- Number of covariates: $K = 1$;
- True model parameters: $\alpha = 5, \beta = 1$.

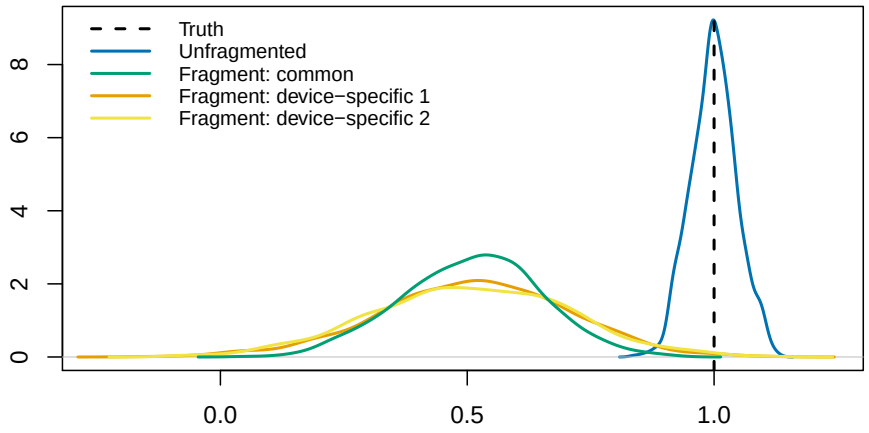
Ad exposures on each device take discrete values (0/1), and we allow correlated ad exposure across devices within a consumer. We change the distribution of device-level covariates and the device usage inclination to illustrate different scenarios. For each scenario, we repeat the simulation 1000 times to get the Monte Carlo distribution for full-data and fragmented estimates. To compare estimates, we plot the distribution of the point estimates from common-effect and device-specific effect estimates, and put them alongside the estimates when using the unfragmented data. A flatter distribution means the point estimates are less stable.

*Boston University Questrom School of Business; tesary@bu.edu;
†The University of Chicago Booth School of Business; sanjog.misra@chicagobooth.edu.

Scenario 1: Symmetric and Independent Exposures

In the best-case scenario, the *SIE* condition holds, so that $E[\widehat{\beta}]$ is attenuated to β/J. To satisfy *SIE*, we let $X_1, X_2 \sim B(N, 0.5)$, $X_1 \perp X_2$, and $\lambda_x = 0.5$. Figure 1 shows that both the common effect and device-specific effect models have the estimates centered at 0.5. Because fragmented data are less reliable, the realized point estimates using fragmented data are less stable, reflected by a more spread-out distribution than the point estimates using unfragmented data.

Figure 1: Slope estimates comparison: attenuation bias



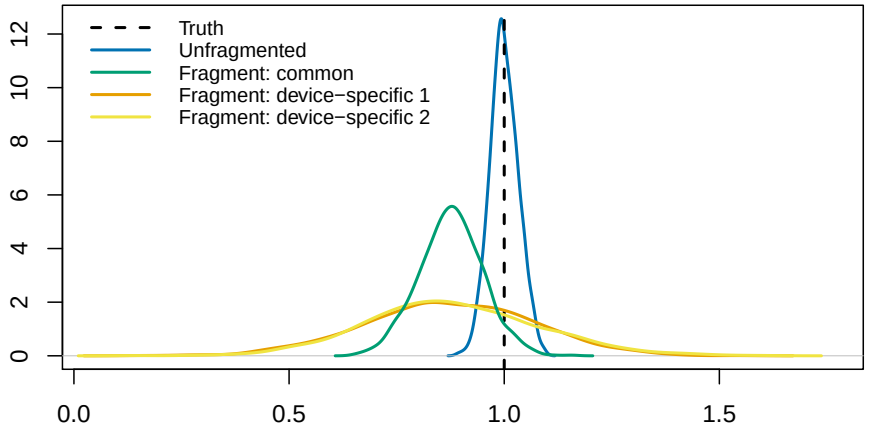
Scenario 2: Omitted Variable Bias

Exposure fragmentation induces an omitted variable bias when a consumer's exposures to ads are correlated across devices. Such correlation is likely when advertisers target different devices that show similar "interest profiles" inferred from behavioral data (e.g., browsing history). In the second specification, we set $corr(X_1, X_2) = 0.75$ while leaving the marginal distributions of X_1, X_2 and λ_x the same as before. Compared to Scenario 1, the fragmented estimates in Scenario 2 take on higher values, with $E[\widehat{\beta}] = 0.875$ for both common effect and device-specific effect models (see Figure 2).

Scenario 3: Activity Bias

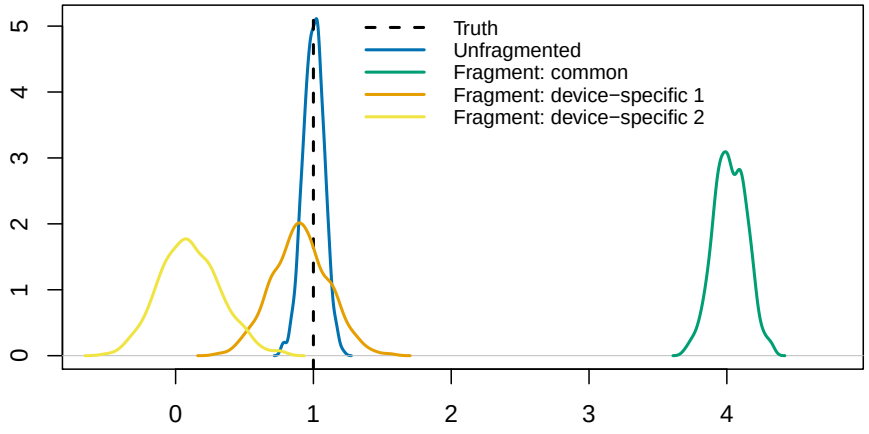
The device-level activity bias is the first consumer behavioral pattern that can induce spurious covariance. To simulate activity bias, we set $X_1 \sim B(N, 0.9)$, $X_2 \sim B(N, 0.1)$, and $\lambda_x = 0.9$, representing a situation where consumers predominantly use one device and thus both see ads and buy things more often on this device. Figure 3 shows that activity bias leads to a substantial upward distortion to the common-effect estimate, with the mean point estimate four times as large

Figure 2: Slope estimates comparison: omitted variable bias



as the true value (see the green line). When the researcher is able to separate different device types and estimate models separately, the resulting estimates do not suffer from such upward distortion. Nevertheless, the attenuation bias remains, thus $E[\beta_1] = 0.9 \cdot \beta$ for device 1 and $E[\beta_2] = 0.1 \cdot \beta$ for device 2 (see the orange and yellow lines).

Figure 3: Slope estimates comparison: activity bias

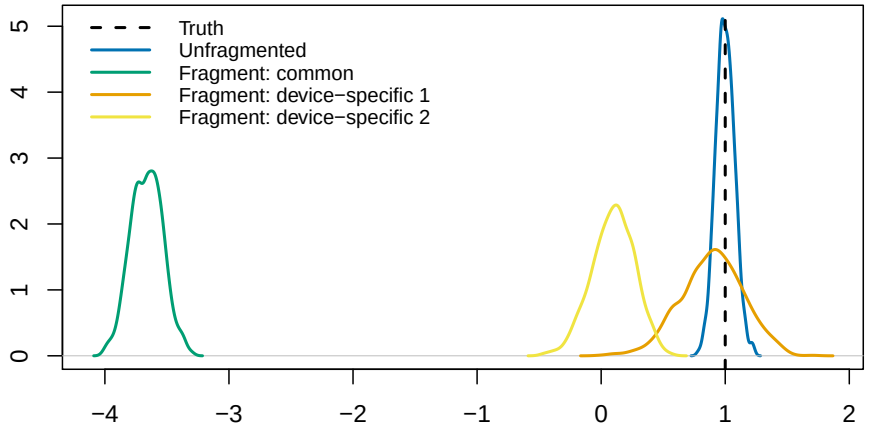


Alternatively, consumers may get exposed to ads mostly on their phones where ad blocking is harder, while completing purchases mostly on their computer. To represent this scenario, let $X_1 \sim B(N, 0.1)$, $X_2 \sim B(N, 0.9)$, and $\lambda_x = 0.9$. The common-effect estimate of ad effect now becomes negative even though the true effect is positive (see Figure 4).

Scenario 4: Cross-Device Substitution

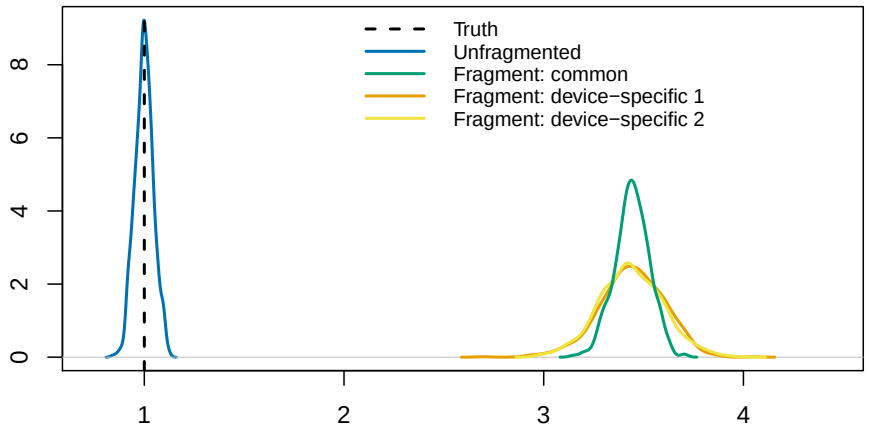
Another behavioral pattern that induces spurious covariance (and one harder to debias) is cross-device substitution. Here, we set $X_1, X_2 \sim B(N, 0.5)$, $X_1 \perp X_2$ as in Scenario 1, but let $\lambda_x =$

Figure 4: Slope estimates comparison: activity bias (reversed)



$(x_1 + 0.01)/(x_1 + x_2 + 0.02)$ so that the device used for purchase depends on which device shows ads more often in that instance. Figure 5 shows that such device substitution creates the same degree of upward bias across model specifications. In particular, device-specific effect models can no longer remove the upward bias.

Figure 5: Slope estimates comparison: device substitution bias

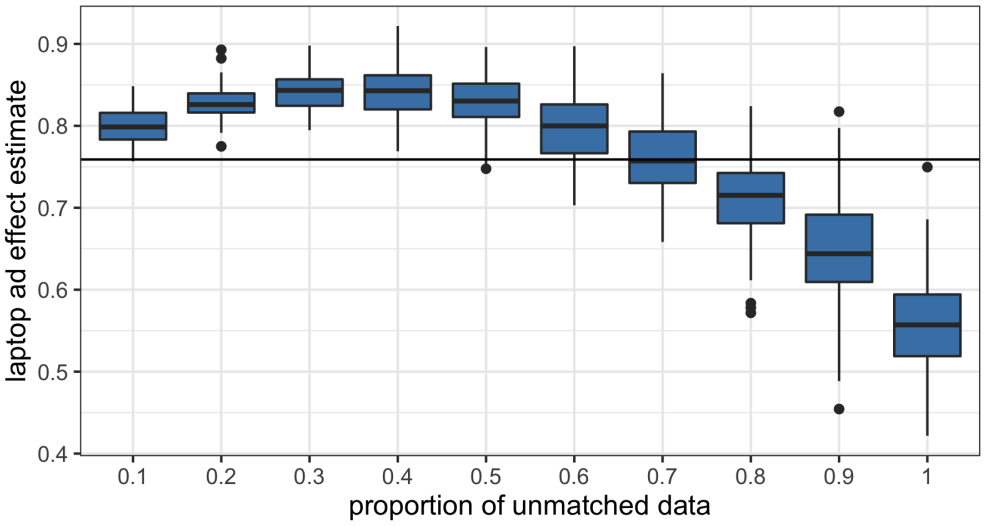


Simulation B: Performance of Partially Matched Estimator

Our second simulation demonstrates how the fragmentation bias changes its sign and magnitude as the proportion of matched records in data changes. To do this, we simulate data with different proportions of matched records. In the true data generating process, consumers have equal propensities of using the mobile or laptop to complete their purchases, and the true ad effect is not device-specific.

Figure 6 compares the estimates using partially matched data (the boxplots) and the true effect (the solid horizontal line). The bias in the estimator does not change monotonically with the proportion of fragmented data. More interestingly, the sign of the bias is positive when most records are matched, but changes to negative when the proportion of fragmented data increases.

Figure 6: Ad effect estimate with partially matched data



Note: The horizontal line represents the true ad effect.