A Simulation-Based Method to Estimating Economic Models with Privacy-Protected Data*

Jung Sakong† Alexander K. Zentefis‡

August 5, 2024

Abstract

Differential privacy algorithms typically distort data in ways that bias estimates from standard econometric methods. We describe a simulation-based econometric method that addresses this issue. The approach adapts the Method of Simulated Moments (MSM) for large datasets and models with high-dimensional fixed effects, when traditional MSM is computationally infeasible. We discuss the approach's application to estimating a gravity model of consumer visits using privacy-protected mobile device data. The methodological framework is flexible and applicable to a wide range of settings where economic models are estimated using privacy-preserved data.

JEL classification: C15, C20
Keywords: differential privacy, simulation methods

*We give special thanks to Bo Honoré and Luojia Hu for suggesting parts of the econometric method we describe in this paper and for their very helpful feedback. We are grateful to Jonathan Dingel, Paul Goldsmith-Pinkham, Jungbin Hwang, Chuck Manski, Yuhei Miyauchi, Piyush Panigrahi; and participants at various seminars and conferences for their very helpful comments. The views expressed in this paper are those of the authors and do not reflect those of the Federal Reserve Bank of Chicago or the Federal Reserve System. All errors are our own.

†Federal Reserve Bank of Chicago; 230 La Salle St, Chicago, IL 60604 (email: jung.sakong@chi.frb.org)
‡Yale School of Management; 165 Whitney Ave, New Haven, CT 06510 (email: alexander.zentefis@yale.edu)

1 Introduction

Differential privacy protects individual privacy in datasets while preserving data utility for analysis (Dwork 2006). Its algorithms increasingly mask datasets used in economic research, including the 2020 Census tables and American Community Survey microdata (Ruggles, Fitch, Magnuson and Schroeder 2019), financial transactions (Karger and Rajan 2020), and health records (Allen et al. 2020). To obscure individual identities, differential privacy algorithms typically add noise and truncate or censor certain observations. However, these distortions introduce non-classical measurement error, which impairs traditional econometric methods, like ordinary least squares (OLS), from providing consistent estimates of economic model parameters.

This paper describes an econometric method that uses simulations to address this issue. The approach adapts Daniel McFadden's Method of Simulated Moments (MSM) for large datasets and high-dimensional fixed effects, when traditional MSM would be computationally impractical (McFadden 1989). Broadly speaking, MSM estimates the parameters of an economic model by closely matching moments calculated from observed data with moments calculated from data that is simulated from the model. The parameters of the model iteratively update until the differences between observed and simulated moments are sufficiently small.

In choosing MSM, the approach's key insight is to simulate data from the economic model and then apply the same differential privacy algorithm to the simulated data that the data provider used to privacy-protect the real-world data. The parameters of the model update until the computed moments from the simulated data (after being made "privacy-protected" per the algorithm) closely match the computed moments from the privacy-protected real-world data.

Implementing MSM for economic models with relatively small datasets or few fixed effects is straightforward (Adda and Cooper 2003). However, traditional MSM becomes computationally impractical when applied to very large datasets—potentially requiring multiple simulations of billions of observations—or to models with high-dimensional fixed effects, which could reach the hundreds of thousands or more. These scenarios are common in modern applied econometric work that uses individual-level data, which is particularly vulnerable to differential privacy.

The approach addresses these computational challenges by using stratified sampling and iterating the high-dimensional fixed effects until convergence in a Gauss-Seidel-style method (Guimaraes and Portugal 2010). The remaining steps involve simulating observations from the

sampled data, applying the differential privacy algorithm, and selecting the model's remaining parameters that minimize the weighted sum of squared errors between the simulated and observed data moments.

An appealing feature of the simulation-based approach is its flexibility. The researcher needs only to (i) specify the model's probability distribution from which to simulate data and (ii) know the differential privacy algorithm that the data provider used. From our experience, data providers are generally open to sharing this information because differential-privacy algorithms are standardized and revealing them does not compromise privacy. Moreover, as with standard MSM, potential misspecification of the data distribution does not affect the consistency of the model estimates (McFadden 1989; Pakes and Pollard 1989).

After outlining the simulation-based procedure, we discuss its application to estimating a gravity model of consumer flows to bank branches using privacy-protected mobile device geolocation data. Details of this application, and a fuller explanation of the simulation-based approach, are provided in Sakong and Zentefis (2024), which originally proposed the approach. The simulation-based approach estimates a gravity coefficient ranging from -1.45 to -1.26, consistent with estimates for consumer flows to nonfinancial establishments (Agarwal, Jensen and Monte 2018). In contrast, traditional gravity model estimation methods—namely OLS and Poisson pseudo-maximum-likelihood (PPML) from Silva and Tenreyro (2006)—yield estimates about an order of magnitude smaller than the MSM estimates. This comparison highlights the downward bias introduced by differential privacy in traditional econometric approaches and underscores the need for alternative procedures like the simulation-based one.

Overall, we outline an econometric method to estimate economic models via simulation on data protected by differential privacy, and we discuss the method's application to a setting that uses privacy-protected mobile device data. We hope the method proves useful to researchers estimating economic models on privacy-protected data across various applications.

2 A Simulation-Based Method

The method aims to uncover the parameters of the "true" distribution of data implied by an economic model from the privacy-protected, distorted data. Let Y^* denote the true data, and suppose it obeys a probability distribution that accords with an economic model. Specifically, let

$$Y^* \sim F(X;\theta), \tag{1}$$

where F is the probability distribution from the model, X is a vector of observable covariates that are undisturbed by differential privacy, and θ is a vector of model parameters. For example, Y^* might represent individual-level demands for various products, X are anonymized consumer characteristics, and θ are demand elasticities. Let Y denote the observed data used in the estimation, subject to the differential privacy algorithm.

The approach adapts the Method of Simulated Moments (MSM) for large datasets and models with high-dimensional fixed effects. In short, the method (i) simulates the "true" data Y^*, from a presumed data-generating process (DGP) according to Eq. (1), (ii) manipulates the simulated data according to the steps of the differential privacy algorithm, and then (iii) selects model parameters to minimize the distance between selected moments of the manipulated simulated data and the privacy-protected data. The following outlines the method's steps.[1]

(1) Specify the data-generating process. The researcher first specifies the DGP F from which to simulate the true observations Y^*. Whether the chosen DGP is discrete or continuous coincides with the nature of the underlying data. For instance, visitor counts to grocery stores are whole numbers, and so a discrete distribution like Poisson or Negative Binomial would be appropriate. By contrast, dollar spending is real-valued, and so a continuous distribution like the log-normal would be suitable in that case. While it is unlikely that the true data perfectly obey the researcher's selected DGP, potential misspecification of the simulating distribution does not interfere with the consistency of the estimates, as with standard MSM (McFadden 1989; Pakes and Pollard 1989).

(2) Sample the data. The econometric method applies to settings with datasets so large that simulating all observations of Y^* is computationally impractical. Such settings often arise in microeconometric datasets where a single simulation can reach billions of observations, like with high-frequency panel data. Rather than simulating all observations of Y^*, the method includes only stratified sampled ones. If the privacy-preserved dataset shows a non-missing observation of Y, it indicates a genuine observation was drawn from the true distribution that avoided truncation, and this observation is sampled in the simulation with probability 1. For example, if the privacy-protected data shows a positive number of visitors from a Census block group to a particular grocery store, this block group $\times$ store pair is sampled with probability 1. Conversely, if Y is missing, it means either no true data was generated or the observation was

[1] For textbook treatments of MSM, see Adda and Cooper (2003), Davidson and MacKinnon (2004), and Evans (2018).

excluded by the differential privacy methods. The researcher samples from this *alternative* set of missing observations. For instance, no visitors might be in the data from the same block group but to different grocery stores, and the researcher would sample from this alternative set of missing block × grocery store pairs. The sampling probability is a choice of the researcher, denoted p. Probability weights are applied to rebalance the stratified sampled data, and the weights are the reciprocal of the sampling likelihood, following standard practice.

(3) Simulate. The next step is to simulate the true data per Eq. (1) and apply the differential privacy algorithm. The simulation process differs for observations sampled with probability 1 and those sampled with probability p. For the observations sampled with probability 1, the researcher draws random variables from F and applies the differential privacy algorithm. For the observations sampled with probability p, no random variables are drawn. Randomly drawing these high-probability weight observations would amplify noise from the simulation and make the estimation unstable. Instead, the researcher constructs their empirical probability distribution given the parameter estimates. This empirical distribution reflects what would be observed if an infinite number of observations were actually simulated. The empirical distribution to construct will depend on the researcher's chosen DGP and the differential privacy algorithm. Section 3 gives an example based on our discussed application.

(4) Iterate the fixed effects until convergence. High-dimensional fixed effects may enter the economic model's parameter vector θ. Estimating such a high volume of fixed effects using the MSM minimization problem alone would be computationally impractical. Instead, the method uses an iterative routine to identify the fixed effects and lets the moment minimization problem identify the remaining model parameters. The iterative process involves a loop per dimension of fixed effects, and the researcher selects the order of iterations. Here, we illustrate the iterative process with two dimensions of fixed effects common to panel data, where, say, one dimension represents an individual i and the other dimension represents time t. But the processes easily extends to more dimensions. First, the researcher guesses a variety of feasible values for the remaining parameters of θ that the MSM minimization problem will determine. Per parameter guess, the researcher then initializes values for the first set of fixed effects, say those of dimension i. The second set of fixed effects in dimension t are then repeatedly updated until the differences in the average simulated data and average observed data of each time t across all individuals i become sufficiently small. Second, per guess of the remaining parameters and the converged t fixed effects from the first iteration, the second

set of fixed effects in dimension i are then repeatedly updated until the differences in the average simulated data and average observed data of each individual i across all time periods t become sufficiently small. The two sets of fixed effects iterate until both converge. When the number of fixed effects in each dimension is large, this routine produces consistent estimates. The iterative process is similar in spirit to the "zig-zag" algorithm, or Gauss-Seidel method, that is commonly used to identify high-dimensional fixed effects in linear models (Guimaraes and Portugal 2010).

(5) Construct the MSM estimator. Once all dimensions of the fixed effects have converged for each guess of the remaining parameters, the MSM minimization problem then selects the optimal estimate of these parameters. It does so by minimizing the weighted sum of squared errors (expressed in percentage points) between the simulated model moments and data moments. The number and choice of moments are at the researcher's discretion, ideally capturing key parts of the distribution of observed data Y, especially those parts most distorted by privacy protection. The model moments include both simulated draws from data points sampled with probability 1 and components of the empirical distribution representing data points sampled with probability p.

3 Application

To provide an application of the simulation-based approach, we summarize how it was originally used in Sakong and Zentefis (2024). There, we estimate a gravity model of consumer flows to bank branches using the simulation-based econometric approach. Gravity models represent bilateral flows of money, goods, or people from origin locations (e.g., residences, workplaces, home countries) to destination locations (e.g., retail establishments, cities, foreign countries). The gravity model estimation uses privacy-protected geolocation data from mobile devices.

The mobile device data are from SafeGraph. The data are monthly and report the numbers of visitors from home Census block groups to bank branches. The company applies a differential privacy algorithm to these visitor counts to avoid identifying people. First, Safegraph adds Laplace noise to all positive counts of visitors to a branch from each home Census block group of the branch's visitors. Second, they round each of these block group $\times$ branch visitor counts down to the nearest integer. Third, they drop from the data all rounded visitor counts less than 2. Fourth, if a rounded visitor count equals 2 or 3, they raise it to 4. These last two

data adjustments render our sample subject to both truncation from below and censoring from below, leading to non-classical measurement error. Roughly 84% of the observed visitor counts equal 4, which implies a substantial amount of data distortion.[2]

We assume that the true visitor counts, denoted V^*_{ij}, are Poisson distributed with means given by a standard fixed-effects gravity model:

$$V^*_{ij} \sim \text{Pois}\left(\exp\left(\gamma_i + \lambda_j - \beta \log \text{Distance}_{ij}\right)\right). \tag{2}$$

Since Harrigan (1996) first introduced them, fixed-effects gravity models are now ubiquitous to the spatial economics and trade literature (Head and Mayer 2014). The parameter γ_i is a block group fixed effect that captures all characteristics of block group i's residents that contribute to them visiting *any* bank branch (e.g., residents' wealth and income, cash needs, or amount of financial sophistication). The parameter λ_j is a branch fixed effect that captures all characteristics of branch j that make it a destination for residents of *any* block group (e.g., the bank's brand, the courteousness of the branch's staff, or the building's aesthetic appeal). The parameter β is the elasticity of visitor flows with respect to distance, and the term Distance_{ij} is the geographic distance between block group i and branch j. Applying the general notation from the previous section, we estimate the parameters $\theta \equiv \left(\{\gamma_i\}_{\forall i}, \{\lambda_j\}_{\forall j}, \beta\right)$. This amounts to estimating roughly 200,000 block group fixed effects, 50,000 branch fixed effects, and one elasticity.

In the estimation, we apply SafeGraph's differential privacy algorithm to the Poisson draws. We draw with probability 1 all block group $\times$ branch pairs with positive visitor counts and we sample with probability $1/2000$ from the alternative set of pairs with missing visitor counts. Because the Laplace noise is added after the Poisson draw, the empirical distribution we use in the simulation is a truncated and censored Laplace distribution whose mean is the realization of the Poisson draw. For the simulation, we construct 7 components of this empirical distribution: (i)-(ii) the probability that the visitor count equals 0, as well as exceeds 0; (iii)-(iv) the probability that the visitor count equals 4, as well as exceeds 4; (v) the expected visitor count; and (vi)-(vii) the expected natural logarithm of visitor counts, conditional on visitor counts exceeding 0, as well as exceeding 4. Finally, we use 6 unconditional moments in the minimization problem to identify β: (i)-(ii) the fractions of visitor counts equaling 0 and equaling 4; (iii)-(iv) the average log distances, when visitor counts equal 0 and equal 4;

[2]SafeGraph asks all researchers who use the company's data to include the disclaimer: "SafeGraph is a data company that aggregates anonymized location data from numerous applications in order to provide insights about physical places, via the Placekey Community. To enhance privacy, SafeGraph excludes census block group information if fewer than two devices visited an establishment in a month from a given census block group." The documentation to the SafeGraph data is here: SafeGraph Documentation.

and (v)-(vi) the OLS coefficients from regressing log visitor counts onto their associated log distances, when visitor counts are greater than 0 and 4 respectively.

We run the estimation month-by-month of the sample, and the point estimates of the gravity coefficient, β, range from about -1.45 to -1.26, and they are fairly stable month-to-month. Thus, across the country, if a representative branch were located 1% farther away from a representative block group, the number of residents from that block group who travel to that branch would drop by around 1.26–1.45% per month. By comparison, Agarwal et al. (2018) estimate a gravity model of consumer expenditures in nonfinancial sectors. and they find a gravity coefficient of -1.05 for the average out-of-home purchase.

We compare these MSM estimates to estimates from traditional methods of estimating gravity models, namely ordinary least squares (OLS) and Poisson pseudo-maximum likelihood (PPML), the latter having been introduced by Silva and Tenreyro (2006). Doing so lets us evaluate the magnitude of the bias introduced by the differential privacy. Whereas the gravity coefficient estimates from the MSM range from -1.45 to -1.26, the estimates from OLS range from -0.062 to -0.038, roughly twenty to thirty times smaller in magnitude. When the sample is limited to block group $\times$ branch pairs with greater than 4 visitors, which are unaffected by SafeGraph's truncation and censoring, the OLS estimates rise in magnitude, ranging from -0.33 to -0.27, which is still roughly four to five times smaller in magnitude than the MSM estimates. Computed over all block group $\times$ branch pairs, the PPML estimates register higher magnitudes than the OLS ones, ranging in values from -0.108 to -0.066. But they still are roughly ten to twenty times smaller in magnitude than the MSM estimates.

Overall, the application reveals the downward bias that differential privacy methods may introduce to traditional methods of estimating gravity models, and it stresses the need for the alternative econometric procedure.

4 Conclusion

We outline a simulation-based econometric method for estimating economic models using data protected by differential privacy. The approach adapts the Method of Simulated Moments for large datasets often used in applied microeconometric research. We discuss the method's application to estimating a gravity model using privacy-protected geolocation data from mobile devices. Estimates from the simulation-based approach differ significantly from those obtained using standard gravity model estimation methods, producing values that align more with the literature. Due to its flexibility, we hope the simulation-based method will be useful

to researchers estimating economic models with large datasets affected by differential privacy.

References

Adda, J. and R. Cooper (2003) *Dynamic Economics: Quantitative Methods and Applications*: MIT Press.

Agarwal, S., J. B. Jensen, and F. Monte (2018) "The geography of consumption," May, Working paper. Georgetown University, Washington, D.C.

Allen, J., C. Bavitz, M. Crosas et al. (2020) "The OpenDP White Paper," May, https://projects.iq.harvard.edu/files/opendifferentialprivacy/files/opendp_white_paper_11may2020.pdf, Working Paper.

Davidson, R. and J. G. MacKinnon (2004) *Econometric Theory and Methods*, 5: Oxford University Press New York.

Dwork, C. (2006) "Differential privacy," in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, 1–12, Springer.

Evans, J. W. (2018) "Simulated Method of Moments (SMM) Estimation," *QuantEcon Notes*, https://notes.quantecon.org/submission/5b3db2ceb9eab00015b89f93.

Guimaraes, P. and P. Portugal (2010) "A simple feasible procedure to fit models with high-dimensional fixed effects," *The Stata Journal*, 10 (4), 628–649.

Harrigan, J. (1996) "Openness to trade in manufactures in the OECD," *Journal of International Economics*, 40 (1-2), 23–39.

Head, K. and T. Mayer (2014) "Gravity equations: Workhorse, toolkit, and cookbook," in *Handbook of International Economics*, 4, 131–195.

Karger, E. and A. Rajan (2020) "Heterogeneity in the marginal propensity to consume: Evidence from Covid-19 stimulus payments," Working paper. Federal Reserve Bank of Chicago, Chicago, IL.

McFadden, D. (1989) "A method of simulated moments for estimation of discrete response models without numerical integration," *Econometrica*, 995–1026.

Pakes, A. and D. Pollard (1989) "Simulation and the asymptotics of optimization estimators," *Econometrica: Journal of the Econometric Society*, 1027–1057.

Ruggles, S., C. Fitch, D. Magnuson, and J. Schroeder (2019) "Differential privacy and census data: Implications for social and economic research," in *AEA Papers and Proceedings*, 109, 403–08.

Sakong, J. and A. K. Zentefis (2024) "Bank Branch Access: Evidence from Mobile Device Data," January, Working Paper. Yale University, New Haven, CT.

Silva, J. S. and S. Tenreyro (2006) "The log of gravity," *Review of Economics and Statistics*, 88 (4), 641–658.