Replication Package README: "Identification of Semiparametric Panel Multinomial Choice Models with Infinite-Dimensional Fixed Effects" *

Wayne Yuan Gao† Ming Li‡

January 19, 2026

This replication package accompanies Gao and Li, "Identification of Semiparametric Panel Multinomial Choice Models with Infinite-Dimensional Fixed Effects."

1 Package Contents and Execution Instructions

This replication package contains all code needed to reproduce the paper's tables and figures. The package is organized into two top-level folders: `/simulation` (simulation results) and `/empirical` (empirical results). Confidential proprietary empirical data cannot be redistributed and are therefore not included in the package.

Master Scripts

Task	**Run**	**Outputs Location**
Simulation results (Tables 1–2, 6–8; Figures 2–3)	In MATLAB: run `run_simulation.m` from `/simulation`	`/simulation/result_simulation`
Empirical results (Tables 3–5)	In MATLAB: run `run_empirical.m` from `/empirical`	`/empirical/result_empirical`

Expected outputs are listed in Section 1–2.

*The authors thank Wenli Lyu for excellent research assistance, including the preparation and documentation of this replication package.

†Department of Economics, University of Pennsylvania. Email: waynegao@upenn.edu.

‡Department of Economics and Risk Management Institute, National University of Singapore. Email: mli@nus.edu.sg.

Note: Simulation results are generated entirely from code (no external downloads beyond files included in the package).

Run a Single Table/Figure

To generate only selected outputs (without running the master scripts), use the subfolders below. Run the script(s) that start with `run_` in the corresponding `src_*` folder; outputs are written to the local `result_*` folder and copied to the main result folder when applicable.

Folder	Generates
`/simulation/table_1_2`	Tables 1–2
`/simulation/table_6`	Table 6
`/simulation/table_7`	Table 7
`/simulation/table_8`	Table 8
`/simulation/figure_2_3`	Figures 2–3
`/empirical/table_3`	Table 3
`/empirical/table_4`	Table 4
`/empirical/table_5`	Table 5

Parallel Processing Options (MATLAB)

Some programs support parallel execution. To set the number of CPU cores (example for `run_empirical`):

- In MATLAB command window: `run_empirical(N)` where `N` is the desired core count.
- Example: `run_empirical(12)` for 12 cores.
- Check available cores: `feature('numcores')` or `maxNumCompThreads`.

Helper functions are located in the corresponding `src_*` folders and are called by the main programs above.

2 Computational Requirements

Hardware Used

MacBook Pro equipped with an Apple M2 Pro chip, featuring a 12-core CPU (8 performance cores and 4 efficiency cores) and 32 GB of RAM. No GPU acceleration was used.

Software Requirements

- MATLAB Version: R2024b (Update 5)
- MATLAB toolboxes: Curve Fitting Toolbox, Optimization Toolbox, Parallel Computing Toolbox, Statistics and Machine Learning Toolbox (versions corresponding to MATLAB R2024b)
- R Version: R 4.4.2 (used for auxiliary scripts)
- Operating System: macOS 14.7.3 (23H417)

3 List of Tables and Figures

Please refer to Table 1 and Table 2 below for where the outputs produced by the code are saved.

Table 1: Simulation Results

Item	Code	Output Folder	Output File(s)
Table 1	`run_simulation.m`	`/simulation/result_simulation`	`table1.csv`
Table 2	`run_simulation.m`	`/simulation/result_simulation`	`table2_upper_part.csv`, `table2_bottom_part.csv`
Table 6	`run_simulation.m`	`/simulation/result_simulation`	`table6.csv`
Table 7	`run_simulation.m`	`/simulation/result_simulation`	`table7.csv`
Table 8	`run_simulation.m`	`/simulation/result_simulation`	`table8.csv`
Figure 2	`run_simulation.m`	`/simulation/result_simulation`	`figure2.png`
Figure 3	`run_simulation.m`	`/simulation/result_simulation`	`figure3.png`

Table 2: Empirical Results

Table	Code	Output Folder	Output File
Table 3	`run_empirical.m`	`/empirical/result_empirical/`	`table3.csv`
Table 4	`run_empirical.m`	`/empirical/result_empirical/`	`table4.csv`
Table 5	`run_empirical.m`	`/empirical/result_empirical/`	`table5.csv`

Empirical input data. The empirical results are generated using the intermediate data files `Emp_D3_V2_data.mat` and `SSS2018.mat`. These files are derived from proprietary Nielsen RMS data and cannot be redistributed due to licensing restrictions.

4 Data Dictionary

This section provides a comprehensive reference for all variables used in the simulation and empirical analysis. All variables correspond directly to the mathematical notation in the paper.

4.1 Key Dimension Parameters

Symbol	Description
N	Number of individuals/consumers (simulation) or DMAs (empirical)
D	Number of covariate dimensions
J	Number of choice alternatives (including outside option)
T	Number of time periods
T_0	Number of time period pairs = $T(T-1)/2$
B	Number of Monte Carlo replications (simulation only)

Table 3: Dimension parameters used throughout the code

4.2 Simulation Variables (Tables 1–2, 6–8, Figures 2–3)

Table 4: Simulation Data Dictionary

Variable	Dimension	Description
beta0	$D \times 1$	True parameter vector β_0. Example: [2; 1; 1; 1] for $D = 4$.
theta0	$(D-1) \times 1$	Angular representation of beta0 in spherical coordinates.
core	scalar	Number of CPU cores for parallel processing.
X_B	$N \times D \times J \times T \times B$	Observable product characteristics for all simulations.
y_B	$N \times J \times T \times B$	Binary choice indicators. y_B(i,j,t,b)=1 if individual i chose alternative j in period t.

Continued on next page

Table 4 – (Continued)

Variable	**Dimension**	**Description**
EyA_B	$N \times J \times T \times B$	True conditional choice probabilities $E[y_{ijt}\|A_i]$.
Z_i	$N \times 1$	Latent variable inducing correlation between X and A. Distribution: $N(0,1)$ (baseline) or $U[-\sqrt{3}, \sqrt{3}]$ (Table 6).
A_scale	$N \times 1$	Scaling component of fixed effects. Distribution: $U[2, 2.5]$.
A_location	$N \times J$	Location component of fixed effects. Alternative-specific intercepts.
Xbeta0	$N \times J \times T$	Linear index $\delta_{ijt} = X'_{ijt}\beta_0$.
AXbeta0	$N \times J \times T$	Scaled utility index $A_i(X'_{ijt}\beta_0 + A_{ij})$.
eps	$N \times J \times T$	Type-I extreme value errors ϵ_{ijt}.
T0	scalar	Number of time period pairs: $T(T-1)/2$.
dX_B	$N \times D \times J \times 2T_0 \times B$	Covariate differences $X_{ijt} - X_{ijs}$ (and negatives).
dEy_B	$N \times J \times 2T_0 \times B$	Estimated moment differences $\widehat{E}[y_{ijt} - y_{ijs}\|X]$.
Xflat	$(N \times T_0) \times (2DJ)$	Flattened covariate matrix for sieve estimation.
X_j_sieve	$(N \times T_0) \times K$	Sieve basis including interactions and quadratic terms.
Edy	$(N \times T_0 \times J) \times 1$	LASSO predictions of $E[y_{ijt} - y_{ijs}\|X]$.
theta_B	$3 \times (D-1) \times B$	Estimated bounds in angular coordinates: [lower; mid; upper].
beta_B	$3 \times D \times B$	Estimated bounds in Cartesian coordinates: [lower; mid; upper].
theta_gr	varies	Grid of angular parameters θ for evaluation.

Continued on next page

Table 4 – (Continued)

Variable	**Dimension**	**Description**
b_gr	varies $\times D$	Grid of coefficient vectors on unit sphere.
Q_gr	varies	Criterion function values $Q(\beta)$ at grid points.
Qmin, Qmax	scalar	Minimum and maximum criterion values.
theta_Qmin	varies $\times(D-1)$	Angular parameters achieving minimum Q.
beta_Qmin	varies $\times D$	Cartesian parameters achieving minimum Q.
agg_theta_q	varies $\times 3$	Aggregated search results: $[\theta_1,\ \theta_2,\ Q]$. Used in Table 6.
nlogn	scalar	Tuning parameter for set estimation (Table 6 only). Values: 0, 0.01, 0.1, 1. Controls $\widehat{c}$ in $\widehat{B}_{\widehat{c}}$.
theta_nlogn	varies $\times(D-1)$	Angular parameters where $Q(\theta) \leq Q_{min} + \widehat{c}$ (Table 6).
Mid_bias	$1 \times D$	Mean bias: $\frac{1}{B}\sum_b(\widehat{\beta}_{mid,b} - \beta_0)$.
Ub_MD, Lb_MD	$1 \times D$	Mean deviation of upper/lower bounds.
dUL_mean	$1 \times D$	Mean width of identified set.
Mid_MSE	$1 \times D$	Mean squared error by coordinate.
SD	$1 \times D$	Standard deviation of midpoint estimator.
RMSE	$1 \times D$	Root MSE by coordinate: $\sqrt{MSE_d}$.
Mid_rsMSE	scalar	Root of sum of MSE: $\sqrt{\sum_d MSE_d}$.
Mid_MND	scalar	Mean normed deviation: $\frac{1}{B}\sum_b \|\widehat{\beta}_b - \beta_0\|_2$.
Mid_SAB	scalar	Sum of absolute bias: $\sum_d \|bias_d\|$.
dul_sum	scalar	Sum of mean widths: $\sum_d E[\widehat{\beta}_{upper,d} - \widehat{\beta}_{lower,d}]$.
Ub_MSE, Lb_MSE	$1 \times D$	MSE for upper/lower bounds (Table 6).
Ub_rsMSE, Lb_rsMSE	scalar	Root of sum of MSE for bounds (Table 6).

Continued on next page

Table 4 – (Continued)

Variable	Dimension	Description
Ub_MND, Lb_MND	scalar	Mean normed deviation for bounds (Table 6).
dX	$N \times D \times J \times T_0$	Covariate differences (input to Qfunc).
dE	$N \times J \times T_0$	Estimated moments $\widehat{H}(E[y_{ijt} - y_{ijs}\|X])$.
b	$D \times 1$	Candidate parameter vector.
dE1	$N \times J \times T_0$	Transformed moments: $2\Phi(\max(dE, 0)) - 1$.
indb	$N \times J \times T_0$	Index differences: $X'_{ijt}b - X'_{ijs}b$.
dbm	$N \times J \times T_0$	Moment condition indicators.
opt	scalar	Objective function value: $\sum_{i,j,t} dE1_{ijt} \cdot dbm_{ijt}$.

Naming Conventions:

- Suffix _B: Variable stored across all Monte Carlo replications
- Prefix d: First-difference variable (e.g., dX $= X_t - X_s$)
- Suffix _gr: Grid of values for search algorithm
- Suffix _0 or 0: True parameter value
- Mid, Ub, Lb: Midpoint, upper bound, lower bound of identified set

4.3 Empirical Variables (Tables 3–5)

Table 5: Empirical Data Dictionary

Variable	Dimension	Description
N	scalar	Number of DMAs (Designated Market Areas) = 205.

Continued on next page

Table 5 – (Continued)

Variable	Dimension	Description
J	scalar	Number of products = 5 (3 brands + other + outside option).
T	scalar	Number of weeks = 52 (year 2015).
D	scalar	Number of covariates = 3 (Price, Promo, Price×Promo).
File: Emp_D3_V2_data.mat		
X_B	$N \times 2 \times J \times T$	Covariates. X_B(:,1,:,:) = Price, X_B(:,2,:,:) = Promo.
EyA_B	$N \times J \times T$	DMA-level market shares s_{cjt}.
File: SSS2018.mat		
X	$(N \times J \times T) \times 3$	Flattened covariates. Columns: [Price, Promo, Price×Promo].
Y	$(N \times J \times T) \times 1$	Flattened market shares.
s_{cjt}	$N \times J \times T$	Observed market share for DMA c, brand j, week t.
Price_{cjt}	$N \times J \times T$	Weighted-average unit price (across UPCs).
Promo_{cjt}	$N \times J \times T$	Binary promotion indicator: 1 if feature or display, 0 otherwise.
$\text{Price}_{cjt} \times \text{Promo}_{cjt}$	$N \times J \times T$	Interaction term allowing price sensitivity to vary under promotion.
u_{cjt}	$N \times J \times T$	Measurement error. Unobserved, assumed $E[u_c \mid X_c, A_c] = 0$.
beta_hat	$D \times 1$	RCLM (Random Coefficients Logit Model) estimate via grid search.
M	scalar	Number of random coefficient draws for RCLM (default: 2000).

Continued on next page

Table 5 – (Continued)

Variable	Dimension	Description
eta_draws	$D \times M$	Random coefficients $\eta \sim N(0, \sigma^2 I_D)$ with $\sigma = 1$.
beta_candidates	$L \times D$	Grid of candidate β vectors on unit sphere (L = grid size).
Q_values	$L \times 1$	Criterion values: $\sum_j (s_j^{pred} - s_j^{obs})^2$ for each candidate.
beta10	$D \times 1$	Cyclic Monotonicity method estimate (normalized).
ols_norm	$D \times 1$	OLS estimate (normalized to unit sphere).
olsfe_norm	$D \times 1$	OLS with fixed effects estimate (normalized).
mlogitfe_norm	$D \times 1$	Multinomial logit with fixed effects estimate (normalized).

5 Raw Data Availability and Processing

Proprietary raw data. The empirical application uses proprietary Nielsen Retail Measurement Services (RMS) scanner data on popcorn sales in the United States for year 2015. The raw Nielsen data cannot be redistributed in this replication package due to licensing restrictions.

How to obtain the proprietary data. Researchers wishing to access the Nielsen RMS scanner data should apply through the Kilts Center for Marketing at the University of Chicago Booth School of Business. Data access information is available at:

`https://www.chicagobooth.edu/research/kilts/research-data/nielseniq`

Researchers affiliated with participating institutions may access the data through their institutional agreement. Independent researchers can request access by contacting the Kilts Center directly. Access typically requires approval of a research proposal and execution of a data use agreement.

Raw-data processing (restricted Nielsen RMS data). The empirical analysis uses Nielsen Retail Measurement Services (RMS) scanner data. Because the raw Nielsen data

and the associated processing code are restricted under the Kilts Center data use agreement, we do not distribute the raw data or the raw-data processing scripts in this replication package. This section documents, at a high level, how the raw data were transformed into the intermediate dataset used to create the MATLAB inputs for estimation.

High-level processing steps. Starting from the Nielsen RMS store-week scanner files and the corresponding product and store crosswalks, we perform the following operations:

1. Restrict the raw RMS transactions to popcorn products (module 1328) in calendar year 2015.
2. Identify the top three brands by market share and aggregate all remaining brands into an "other" category.
3. Merge store identifiers to DMA information and aggregate store-level observations to a DMA-week level panel.
4. Construct key variables used in estimation, including weighted-average prices, promotion indicators (feature $\vee$ display), and market shares.
5. Output a DMA-week-brand panel dataset that serves as the intermediate input for constructing the MATLAB data objects used in estimation.

Confidential data inputs.

- RMS_1328_2015.dta – Weekly store-level scanner data for popcorn (module 1328) in 2015
- products_1328.dta – Product characteristics and brand information
- stores.dta – Store characteristics including DMA identifiers

6 References and Data Citation

Main paper. Gao, Wayne Yuan, and Ming Li (2026). *Identification of Semiparametric Panel Multinomial Choice Models with Infinite-Dimensional Fixed Effects*. Manuscript, University of Pennsylvania and National University of Singapore.

Online appendix. Gao, Wayne Yuan, and Ming Li (2026). *Online Appendix for "Identification of Semiparametric Panel Multinomial Choice Models with Infinite-Dimensional Fixed Effects"*. Available upon publication alongside the main paper.

Data source. The empirical analysis uses proprietary Nielsen Retail Measurement Services (RMS) scanner data provided by NielsenIQ through the Kilts Center for Marketing Data at the University of Chicago Booth School of Business. Access to these data requires a license and can be requested from:

`https://www.chicagobooth.edu/research/kilts/research-data/nielseniq`

Due to confidentiality and licensing restrictions, the empirical data are not included in this replication package.