This README document provides the basic information on the creation of the final dataset and the replication of empirical and simulation results for the paper titled "Universal Inference for Incomplete Model" by Hiroaki Kaido and Yi Zhang.

If there are any questions about replication of the data set, please contact the author at yzhangjnu@outlook.com

Section 1: replication of the data "rawdatafinal_2010"

The data is cleaned using Stata 17 and store as CSV format. The file "dataprocess040925.do" produces the dataset "rawdatafinal_2010" featured in the paper.

The final data set "rawdatafinal_2010" is created based on the Capital IQ Pro Platform （See S&P Capital IQ Pro | S&P Global (spglobal.com) for more information），Center for Responsive Politics (CRP) (See Lobbying Data Summary • OpenSecrets for more information) and the Bureau of Economic Analysis (See https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas).

Computing Requirement

Our analysis requires STATA version 17 or later with additional packages. One of the required package is "openall" please run "scc install openall" in your Stata to install this package.

Project directory structure

Your project directory should include the following subfolders:

- raw_data: this folder stores all cleaned raw data sets used in the project
- processed_data: This folder stores datasets generated during intermediate steps of the analysis (it should be empty before running the .do file).
- Lobbydata: This folder contains raw datasets required to replicate the bank list file "lobby_id0819" and the lobbying data file "lobby0819.xlsx".
- Distance_data: this folder stores all raw data set used for replicating the file "distance.xlsx".

Next, we describe the contents of raw_data, Lobbydata and Distance_data folders and explain each of them.

Contents of the data folder

Since the processed_data folder is empty any do-file is exexuted, the only folders we need to look at are raw_data, Lobbydata and distance_data folders. The raw_data folder contains the following files:

- bank0720: Raw data from Capital IQ Pro used to generate financial and demographic covariates. (from year 2007 to year 2020)
- enforcement0519: Raw data from Capital IQ Pro used to generate enforcement action outcome variables. (from year 2005 to year 2019)
- lobby_id0819: Cleaned raw data containing unique lobbying banks, including their names and SNL_INSTI_KEY, verified manually using Capital IQ Pro (three columns).
- lobby0819: Cleaned raw data on lobbying information collected from the Center for Responsive Politics (CRP).
- pcpi_growth0812: Raw data on county-level per capita personal income growth from the Bureau of Economic Analysis (BEA), covering the years 2008 to 2012.
- distance: Raw data on the distance from each bank to Washington, D.C.(zipcode:200001), calculated based on their ZIP codes.

To replicate rawdatafinal_2010, please ensure that the directory is set correctly and run the Stata code file dataprocess040925.do.

Please remember to update all relevant directory and base path settings at the beginning of the file and maintain the original folder structure and names as provided in the downloaded zip file. Additionally, you need to create a processed_data folder to store all intermediate results to ensure the code runs successfully.

The following variables are included in rawdatafinal_2010:

- REG_SNL_INSTN_KEY: the SNL database institutional ID key
- Year: year of observation (in this file, they are all 2010)
- CompanyName: name of the bank
- CompanyType: commercial bank or bank holding company
- PrimaryRegulator: OCC, FDIC or FED
- YearEstablished: the established year of the bank
- ParentName: the parent bank name of current bank
- ParentInstitutionKey: the parent bank SNL institutional ID key
- CountyandState: the county and State information of the bank
- ZipCode: the zip code of the bank
- State: the state of the bank
- Initial_Scale: the initial asset size of the bank

- lobby_status: if the bank has at least hired a lobbying firm or filed a lobbying report with the OCC, FDIC, or Fed at least once during the year 2010. Otherwise, it is zero.
- Severe: The outcome variable is defined as a binary indicator that equals 1 if a severe enforcement action was initiated by a regulator (OCC, FDIC or Fed) for the bank I during the year 2010.
- capital_adequacy: (Tier 1 capital divided by risk-weighted assets) *100
- asset_quality: The negative of loan and lease allowance scaled by the total loans
- management_quality: the negative of the number of enforcement actions against personnel and individuals during the year 2010.
- earning: the ratio of net interest income to earning assets
- liquidity: (The ratio of cash to deposits)*100
- sensitive_to_market_risk: The absolute value of the net short-term liabilities divided by the total earning assets
- deposit_to_asset_ratio: (The ratio of total deposits to total assets)*100
- leverage: The debt-to-equity ratio
- total_core_deposit: The deposits made in the bank's natural demographic market
- size: directly measured by total assets of the bank
- age: age (in years) of the bank (given year - establised year)
- growth: real per-capita personal income growth at county-level from 2009 to 2010.
- distance: The distance (in kilometers) between the headquarters of the bank'i and Washington DC.
- Initial_market_size: The initial (in 1998) bank's total assets relative to its within-state peers' total assets.

Instructions to reproduce raw datasets

In this section, we discuss how to replicate all data contained in the raw_data folder. Since some of this data is not publicly available, users must download the raw data through an authorized account on the Capital IQ Pro platform. We specifically focus on replicating "bank0720" and "enforcement0519" data files.

For publicly available data, we have provided the original dataset on per capita personal income growth at the county level from 2008 to 2012 in the file pcpi_growth0812. The distance file records the distance from each bank to Washington, D.C., calculated using a simple Python script. Additionally, in the Lobbydata folder, we have included all data downloaded from the CRP website, along with instructions on how to reproduce the lobby_id0819 and lobby0819 data files.

Bank0720

To construct the financial covariates, we list the variables downloaded from the S&P Capital IQ platform, with the corresponding six-digit unique field IDs provided in parentheses for replication purposes. The file bank0720.xlsx contains the following variables (listed in column order; definitions of all variables can be found on the platform):

- Bank basic information: (1) Year (also called record year) (2) Company name (201127); (3) SNL institution key (201128); (4) Federal Reserve ID (201126); (5) Regulatory ID (205116); (6) Company type (205241); (6) Primary regulator (205275); (7) Year established (225998).
- Parent information: (9) Parent name (205174); (10) Parent institution key (205130); (11) Parent state (205177); (12) Parent company type (205276); (13) Parent regulatory ID (205175).
- Address and Geographic: (14) Location (232520); (15) Zip code (205129); (16) County and state (205133); (17) State (205128).
- Risk taking: (18) Return of assets (ROA) (ROAA 205264); (19) GRB Total Equity Capital (215404); (20) Initial Scale (21) Total Asset (215382); (22) Equity/Assets-Ratio (215627); (23) Off Balance Sheet Unused commitment growth (216227); (24) Total Real Estate Loans (216892); (25) Total C\&I Loans (215808); (26) Total Consumer Loans (215813) (Notice not loan ratio, e.g. not 205850); (27) Total Loans and Lease (215825); (28) CI/Loan ratio (215514); (29) Consumer/Loan ratio (215515); (30) Real estate/Loan ratio (215513); (31) Loans 90 days or more past due (PD total loan) (216442); (32) Non-accrual Loans (216515); (33) PD90/loan ratio (215485); (34) Non-accrual/loan ratio (215486).
- Financial Variables: The financial variables are selected based on the CAMELS rating system. The CAMELS rating system assesses the strength of a bank through six categories. These categories include capital adequacy, assets, management capability, earnings, liquidity, and sensitivity. In general, the rating system is on a scale of one to five, with one being the best rating and five being the worst rating. We chose proxy variables for the CAMELS rating. For capital adequacy: (35) Tier 1 risk-based capital ratio (215628); (36) Tier 1 capital (215619); (37) risk-weighted assets (215622). For asset quality: (38) loan and lease allowance ratio (215663). For liquidity (39) Total Deposit (206127) (40) Total Cash (215361).For earning: (41) net interest income (215417); (42) earning assets/total assets (206037). For sensitivity to market risk:(43) Net short-term liabilities (215675). For size: (44) Deposit to total asset ratio (216946); (45) Total core deposit: total core deposit/total deposit (215536); (46) Leverage ratio (215630) (47) total earning asset (48) Regular core deposit.

enforcement0519

To construct the enforcement action outcome variables, we list the variables downloaded from the S&P Capital IQ platform with variable names for replication purposes (Definition of variables can be checked through platform).

- Key Institution
- Institution Name
- Regulatory Agency
- Institution Type
- Current/Historical
- Regactiontype
- Issue Date
- Modification Date
- Termination Date

pcpi_growth0812

For regional economic growth, we downloaded county-level per capita personal income growth data for the years 2008 to 2012 from www.bea.gov/data/. Below, we list the variables retained for empirical analysis:

- GeoFips: the county-level Fips code
- GeoName: the county and state name
- 2008: growth rate from 2007 to 2008 for the county
- 2009: growth rate from 2008 to 2009 for the county
- 2010: growth rate from 2009 to 2010 for the county
- 2011: growth rate from 2010 to 2011 for the county
- 2012: growth rate from 2011 to 2012 for the county

distance

To simplify the replication process, we directly provide a distance file containing the following variables:

- REG_SNL_INSTN_KEY: the SNL database institutional ID key
- ZipCode: the zip code of the bank
- distance: the distance between bank and Washington DC (zip code: 20001) in km.

For details regarding the calculation, please refer to the Distance_data folder. The raw dataset, bank_zipcode.xlsx, was downloaded from Capital IQ Pro. The distance calculations were performed using a simple Python script with the uszipcode, SearchEngine and mpu.haversine_distance packages. You will need to install the required

packages by runing “pip install mpu” and “pip install SearchEngine” . After installing the packages, execute the script “04distance.py” to get distance.

lobby_id0819

To construct the lobbying outcome variable, we match lobbying records with all bank records based on the parent bank name. The algorithm used in this empirical application consists of three steps:

- First step: Identify and record the corresponding REG_SNL_INSTN_KEY for unique lobbying banks using the Capital IQ Pro platform.
- Second step: Perform manual name-matching in cases where REG_SNL_INSTN_KEY matching is unsuccessful. This involves verifying information using external sources, such as web searches, to complete the name-matching process.
- Third step: combine step 1 and step 2 and generate a unique bank list with REG_SNL_INSTN_KEY

The unique bank list is generated by the “lobby_clean.do” script. Readers can find the “lobby_id0819” unique bank list in the Lobbydata/stata folder. This unique list is then used to match information from Capital IQ Pro and other sources, with all records consolidated into the same file. The following variables are retained in this data file:

- Client: the name of the bank
- Full name: the parent’s name of the bank
- REG_SNL_INSTN_KEY： the SNL database institutional ID key

lobby0819

To construct the lobbying variables, we provide the file with the following variables:

- client: the name of the bank in the lobbying report
- Year: lobbying report year
- Total: total expenditure for a given bank in the given year.
- client_clue: the name of the bank who revealed the lobbying record
- lobbyingfirmhired: the name of the lobbying firm that the bank hired.
- totalamount: the total expenditure for each firm that the bank hired.
- lobbyist: the name of the lobbyist.
- revolvingdoorprofiles: if the lobbyist is a movement from the public sector (such as government) to private sector, then the value is equal to one. Otherwise, it is zero.

- formermemberofcongress: if the lobbyist is a movement from the congress senator to private sector, then the value is equal to one. Otherwise, it is zero.
- lobby_status: if the bank has the lobbying record or hire a lobbyist then it is equal to one. (currently they should all equal to one)
- experience: the lobbying experience in year that the bank has.
- Target: If their lobbying strategic efforts by individuals or organizations to influence specific regulatory decisions on FDIC, OCC, or FED, then it is equal to one. Otherwise, it is zero.

We have shared all data downloaded from the CRP website, which are stored in the folder Lobbydata/rawdata1030. The constructed files can be found in the Lobbydata/stata folder. Please do not modify any subfolders within the Lobbydata/stata folder, as they are required for processing and replicating the files.

Section 2: replication of the Table 1 descriptive statistics and Table 2 confidence region in the empirical illustration.

Computing Requirement

Beginning with this section, the remaining replication code is written in MATLAB 2017 or later versions. No additional packages are required to compute Table 1 and Table 2. The code for Table 2 should be run on the Shared Computing Server.

Table1: Descriptive Statistics

Table 1 can be replicated by running the script replication_table1.m. Please ensure that both the script and the working directory are set up correctly. The summary statistics produced by this script correspond to those presented in Table 1 of the paper.

Table2: Confidence Intervals

Table 2 can be replicated to execute the code on a shared computing server. In this package, we provide both main file and batch file for the replication process.

Replication procedures

- Step 1: create a folder in the same directory called “Results” on the server.

- Step 2: Open **the main file: app_giivasf_scc3.m**: the change the saving results path accordingly in Line 122-126 so that results can be saved in Results directory on the server.
- Step 3: We provide two examples to achieve confidence interval for entire population CP(0) and small asset size with CP(1) (rest of examples can be achieved accordingly).

Example 1: Entire population: CP(0)

- Open **the main file: app_giivasf_scc3.m**: Setup Line 18 to cf_obj= 'ASF' , line 20 to cf_decision=0, and Line 21 cf_gridpoints =200. Then save the file on the server.
- Open **the batch file: run_app_givasf_scc3.sh**: Setup Line 9 to -t 1-200.
- Submit the batch file to the server. Once all jobs are finished, download all files from the"Results" folder on the server.
- Place the get_results.m file in the Results folder. Open the file "get_results.m" and Setup line 14 to the corresponding local directory on your personal computer where all 200 files are stored. Then run the main file get_results.m to obtain confidence interval for case "Entire population: CP(0)".

Example 2: Asset size with small bank: CP(0)

- Open **the main file: app_giivasf_scc3.m**: Setup Line 18 to cf_obj= 'CASF' , line 20 to cf_decision=0, line 19 to cf_csize='small' and Line 21 cf_gridpoints =200. Then save the file on the server.
- Open **the batch file: run_app_givasf_scc3.sh**: Setup Line 9 to -t 1-200.
- Submit the batch file to the server. Once all jobs are finished, download all files from the"Results" folder on the server.
- Place the get_results.m file in the Results folder. Open the file "get_results.m" and Setup line 14 to the corresponding local directory on your personal computer where all 200 files are stored. Then run the main file get_results.m to obtain confidence interval for case "asset size small banks: CP(0)".

Section 3: replication of the Table 3 -Table 5 size, power and computation time illustration.

Computing Requirement

Starting from this section, the rest of replication code are written in MATLAB 2017 or later versions. We need to install an extra solver for this simulation which is called CVX solver (see https://cvxr.com/cvx/ for more information). The user needs to register on the website

and download the standard license accordingly. To install the CVX solver in your computer please see https://cvxr.com/cvx/doc/install.html. It will generate a csolve.m file for the rest of computing.

Replication of Table 3

Before running the code, create a subfolder named "Results" in the same directory as the main file "main_ex1_scc.m". Open main_ex1_scc.m, change the value in line 6 to 50, 100, and 200, and run the code three times—once for each value. The results for each case will correspond to those presented in Table 3.

Replication of Table 4

The replication of table 4 requires running the code on the Shared Computing Server. Notice that we need to upload main_ex2_scc.m and the corresponding batch file run_main_ex2_scc in the same folder.

Step 1: Upload the batch and main files—run_main_ex2_scc.sh, main_ex2_scc.m, to the shared computing server working directory.

Step 2: Create a folder on the server named Results in the same directory.

Step 3: Open main_ex2_scc.m, set line 10 to hgrid = linspace(0, 2, M), line 11 to n = 50, and line 61 to the corresponding results directory. Save the file.

Step 4: Submit the batch files as jobs to the server. Wait until all jobs are completed.

Step 5: Repeat Steps 3 and Step 4 with n=50,100,200,300.

Step 6: Download all files to your personal computer from the Results folder on the server.

Step 7: In your PC, open the file get_results_main_ex2.m, update lines 29 and 31 to the corresponding local directory where the downloaded files are stored.

Step 8: Run get_results_main_ex2.m to generate table 4.

Replication of Table 5

The replication of table 5 requires running the code on the Shared Computing Server. Notice that we need to upload main_ex2_comptime.m, bcs_ex2_comptime.m, and the corresponding batch file run_main_ex2_comptime, run_bcs_ex2_comptime in the same folder.

Step 1: Upload the 2 batch files and 2 main files to the shared computing server working directory, create a Results folder in the same working directory.

Step 2:

- Open main_ex2_comptime.m, set line 6 to S=1000, line 10 to hgrid = linspace(0, 0.15, M), line 11 to n = 5000, and line 64 to the corresponding Results directory. Save the file.
- Open bcs_ex2_comptime.m, set line 2 to S=100, line 6 to hgrid = linspace(0, 0.15, M), line 7 to n = 5000, and line 107 to the corresponding Results directory. Save the file.

Step 3: Submit the batch files as jobs to the server. Wait until all jobs are completed.

Step 4: Download all files to your personal computer from the Results folder on the server.

Step 5: Open downloaded files and check median of variable "comptime" for main_ex2_comptime results, and median of variable "comptime" for bcs_ex2_comptime results.

Replication of Figure 2

To replicate Figure 2, we provide instructions for the cases with n=7500 and n=5000 for both the cross-fit LR test and the moment-based test. The replication steps are as follows:

Step 1: Upload the batch and main files—run_main_ex2_scc.sh, run_bcs_ex2.sh, main_ex2_scc.m, and bcs_ex2.m—to the shared computing server working directory, create a Results folder in the same working directory.

Step 2:

- Open main_ex2_scc.m, set line 10 to hgrid = linspace(0, 0.15, M), line 11 to n = 7500, and line 61 to the appropriate results directory. Save the file.
- Open bcs_ex2.m, set line 10 to hgrid = linspace(0, 0.15, M), line 11 to n = 7500, and line 105 to the appropriate results directory. Save the file.

Step 3: Submit the batch files as jobs to the server. Wait until all jobs are completed.

Step 4: Repeat Steps 2 and 3 with n=5000.

Step 5: Download all files to your personal computer from the Results folder on the server.

Step 6: In your PC, open the file get_results_main_ex2.m, update lines 29 and 31 to the corresponding local directory where the downloaded files are stored.

Step 7: Run get_results_main_ex2.m to generate Figure 2.