

# Supplement B - Joint Modeling of Playing Time and Purchase Propensity in Massively Multiplayer Online Role-Playing Games using Crossed Random Effects

BY TRAMBAK BANERJEE ([trambak@ku.edu](mailto:trambak@ku.edu)), PENG LIU ([pliu2@scu.edu](mailto:pliu2@scu.edu)),  
GOURAB MUKHERJEE([gmukherj@marshall.usc.edu](mailto:gmukherj@marshall.usc.edu)),  
SHANTANU DUTTA ([shantanu@marshall.usc.edu](mailto:shantanu@marshall.usc.edu)), AND HAI CHE ([chehai@ucr.edu](mailto:chehai@ucr.edu))

## Code

We provide the full set of analyses code in this repository that can be used to replicate all the results in the main paper. Specific instructions are provided below on how to replicate the tables and figures in the main paper and Supplement A. Additionally, the code relies on specific software and hardware requirements that have been described below in detail.

## Reproducibility workflow

All figures and tables in the paper are reproducible except for figures 1, 2 and 4 in the main paper and figure 6 and table 2 in Supplement A. These figures are developed in Power Point and do not require any numerical inputs. Table 2 is the data dictionary. Below, we provide the steps that must be followed to reproduce the different tables and figures in the paper.

## Supporting software and hardware requirements

The primary software used is R ( $\geq 4.3.0$ ), and the following R packages and their dependencies must be installed for any reproducibility analysis: Rfast ( $\geq 2.0.1$ ), msos ( $\geq 1.2.0$ ), mvtnorm ( $\geq 1.1.1$ ), CVXR ( $\geq 1.0.8$ ), igraph ( $\geq 1.2.6$ ), [ggb \( \$\geq 0.10.0\$ \)](#), GLMMadaptive ( $\geq 0.7.15$ ), lme4 ( $\geq 1.1.26$ ), foreach ( $\geq 1.5.1$ ), doParallel ( $\geq 1.0.16$ ), ggplot2 ( $\geq 3.3.2$ ), gridExtra ( $\geq 2.3$ ), glmLasso ( $\geq 1.5.1$ ), rpqI ( $\geq 0.8$ ), readr ( $\geq 1.4.0$ ), tidyverse ( $\geq 1.3.0$ ), lattice ( $\geq 0.20.41$ ), viridisLite ( $\geq 0.3.0$ ), reshape2 ( $\geq 1.4.4$ ), scales ( $\geq 1.1.1$ ), Rcpp ( $\geq 1.0.5$ ), RcppArmadillo ( $\geq 0.10.1.2.0$ ), RcppProgress ( $\geq 0.4.2$ ).

The analyses presented in the paper are based on the following Hardware specifications: Windows 10, 64 bit, with 128GB RAM on an Intel Xeon Gold 6230 CPU. At a minimum, the

authors recommend access to 64GB RAM and 10 CPU cores for enabling multi-core parallelization.

## Workflow

We provide a sequence of steps for reproducing the different tables and figures in the paper. Expected Run Time for each of these steps will be denoted by ERT.

### Figures 3, 5 in the main paper and Figures 7, 8 in Supplement A

1. Run the script `motivatingfigures.R` in **'processing'** folder to reproduce figures 3 and 5 in the main paper, and figures 7 and 8 in Supplement A. Please make sure that the R working directory is set to `(your folder structure)/ Supplement B/data`. [ERT < 5 minutes]

### Table 3 in Supplement A

1. Run the script `datasummary.R` in **'processing'** folder to reproduce table 3 in Supplement A. The script writes out two CSV files that hold the summary statistics in table 3. Before running this script, please make sure that the working directory is set to `<your folder structure>/Supplement B/data`. [ERT < 5 minutes]

### Table 1 in the main paper

This is the table that presents the selected fixed / composite effects and the coefficient estimates under the submodels Login Indicator, Duration of Play and Purchase Propensity. The following steps when executed in the order described below reproduce table 1.

1. Run `dataprocessing.R` in **'processing'** folder to output `out.RData` which is a list that holds the processed training and prediction data. Before running this script, please make sure that the working directory is set to `<your folder structure>/Supplement B/data` and the output directory is set in line 207. [ERT < 10 minutes]
2. Run `crejm_estimation.R` in **'selection'** folder. This script writes out a list of initial estimates `init.est.RData`. Please make sure that:
  - i. On line 8 - the directory for sourcing the R scripts in the folder `spcov` is set.
  - ii. On line 11 - the appropriate directory for sourcing the R scripts in the folder `library` is set.
  - iii. On line 14 - the working directory is set.

- iv. On line 16 - the path to `out.RData` (from step 1) is properly set. [ERT ~ 6 hours]
3. Run `crejm_selection.R` in **`selection`** folder. This script writes out a list of selected fixed / composite effect predictors in `selection.RData`. Please make sure that:
  - i. On line 6 - the directory for sourcing the R scripts in the folder `spcov` is set.
  - ii. On line 9 - the appropriate directory for sourcing the R scripts in the folder `library` is set.
  - iii. On line 12 - the working directory is set.
  - iv. On line 14 - the path to `out.RData` (from step 1) is properly set.
  - v. On line 33 - the path to `init.est.RData` (from step 2) is properly set. [ERT ~ 4 hours]
4. Finally, to get the coefficient estimates, run `crejm_postselectionestimation.R` in **`selection`** folder. This script writes out a list `postselection.est.RData` that stores the coefficient estimates, and the estimated covariance matrices of the player and guild specific random effects. Please make sure that:
  - i. On line 9 - the path to `selection.RData` (from step 3) is properly set.
  - ii. On line 13 - the directory for sourcing the R scripts in the folder `spcov` is set.
  - iii. On line 16 - the appropriate directory for sourcing the R scripts in the folder `library` is set.
  - iv. On line 19 - the working directory is set. [ERT ~ 6 hours]

Please be advised that steps 2, 3 and 4 in the sequence above are both memory and time intensive processes.

## Figure 6 in the main paper

1. Figure 6 relies on the output `postselection.est.RData` that is obtained from step (4) above. To reproduce figure 5, run `crejm_randomeffect_network.R` in **`selection`** folder. Please make sure that:
  - i. On line 7 - the working directory is set.
  - ii. On line 8 - the path to `postselection.est.RData` (from step 4 of Table 1) is properly set.
  - iii. On line 38 - the path to the excel file `crejm_network.xlsx` is set. [ERT < 5 minutes]

## Table 2 in the main paper and Table 1, Figures 1 and 2 in Supplement A

Table 2 in the main paper and Table 1 in Supplement A present results related to the predictive performance of CREJM and Benchmarks I, II. Figures 1 and 2 in Supplement A are excel plots that are based on the output from the following steps.

1. Run `g1mm1asso_prediction.R` in **'prediction'** folder to output the False Positive (FP) rates, False Negative (FN) rates, Prediction Errors (PE) and AUC scores for Benchmark I. Please make sure that the working directory and the path to `out.RData` (from step 1 of Table 1) are properly set in lines 5 and 8, respectively, of this script. [ERT < 20 minutes]
2. Run `rpq1_prediction.R` in **'prediction'** folder to output the False Positive (FP) rates, False Negative (FN) rates, Prediction Errors (PE) and AUC scores for Benchmark II. Please make sure that the working directory and the path to `out.RData` (from step 1 of Table 1) are properly set in lines 5 and 8, respectively, of this script. Also, please set the appropriate directory in line 83 for sourcing the R scripts in the folder `library`. [ERT < 30 minutes]
3. Run `crejm_prediction.R` in **'prediction'** folder to output the False Positive (FP) rates, False Negative (FN) rates, Prediction Errors (PE) and AUC scores for CREJM. Please make sure that:
  - i. On lines 8 and 9 - the working directory and the path to `selection.RData` (from step 3 of Table 1) are properly set.
  - ii. On lines 11 and 12 - the appropriate directory for sourcing the R scripts in the folder `library` is set.
  - iii. On line 22 - the appropriate directory for reading `postselection.est.RData` (from step 4 of Table 1) is set.
  - iv. On line 25 - the path to `out.RData` (from step 1 of Table 1) is properly set.

This script uses R packages `foreach` and `doParallel` and expects access to a local cluster of size 10 on line 88. [ERT < 30 minutes]

## Figure 3 in Supplement A

1. Run `crejm_guilrandeffs.R` in **'prediction'** folder to reproduce figure 3 in Supplement A. Please make sure that:
  - i. On lines 8 and 9 - the working directory and the path to `selection.RData` (from step 3 of Table 1) are properly set.
  - ii. On lines 11 and 12 - the appropriate directory for sourcing the R scripts in the folder `library` is set.

- iii. On line 22 - the appropriate directory for reading `postselection.est.RData` (from step 4 of Table 1) is set.
- iv. On line 25 - the path to `out.RData` (from step 1 of Table 1) is properly set.

This script uses R packages `foreach` and `doParallel` and expects access to a local cluster of size 10 on line 83. [ERT < 30 minutes]

## Figures 4 and 5 in Supplement A

Figure 4 presents three heat-maps, one for each of the three responses, that plot the mean predicted correlation over time of all players that are members of guild `k` where `k = 1,..,50`. Please be advised that reproducing figure 4 is both memory and time intensive.

1. Run `crejm_guildcorrelations.R` in ``prediction`` folder to reproduce figures 4 and 5. Please make sure that:
  - i. On lines 8 and 9 - the working directory and the path to `selection.RData` (from step 3 of Table 1) are properly set.
  - ii. On lines 11 and 12 - the appropriate directory for sourcing the R scripts in the folder `library` is set.
  - iii. On line 22 - the appropriate directory for reading `postselection.est.RData` (from step 4 of Table 1) is set.
  - iv. On line 25 - the path to `out.RData` (from step 1 of Table 1) is properly set.

This script uses R packages `foreach` and `doParallel` and expects access to a local cluster of size 20 on lines 87 and 155. [ERT ~ 48 hours]

## Figures 9 and 10 in Supplement A

1. Run `random_effect_structure.R` in ``prediction`` folder to reproduce figures 9 and 10 in Supplement A. Please make sure that:
  - i. On lines 6 and 7 - the working directory and the path to `out.RData` (from step 1 of Table 1) are properly set.