Estimating Ideology and Influence in Social Media Networks: Opinions on Genome Editing in Livestock

Estimation Using `R` and `RStan` Code

Joseph Navelski[a], Syed Badruddoza[b], Jill McCluskey[c], and Francis Pascual[d]

[a]School of Economic Sciences, Washington State University, Pullman, WA 99164, USA; [b]Department of Agricultural and Applied Economics, Texas Tech University, Lubbock, TX 79409, USA; [c]School of Economic Sciences, Washington State University, Pullman, WA 99164, USA; [d]Department of Mathematics and Statistics, Washington State University, Pullman, WA 99164, USA

All supplementary materials can be found at the corresponding author's GitHub (https://github.com/JNavelski). They are available to all users.

1. Introduction

This supplementary material describes how readers can replicate our methods presented in our article. For a more detailed explanation of the methods applied in our paper, please also refer to [1]. We describe the specific equipment and software we used in the analysis because we expect computation times and numerical results may vary across computer equipment and software versions of `R` and `RStan` [2,3]. For more details, refer to the Reproducibility Chapter on the `RStan` website.

The machine that we used to perform computations and analysis in this supplement was:

- 2018 MacBook Pro - macOS: Monterey Version 12.4, Processor: 2.7 GHz Quad-Core Intel Core i7, SSD: 1T, Memory: 16 GB, and Graphics: Intel Iris Plus Graphics 655 1536 MB.

To replicate our results, the reader needs the following software:

- `R` - Version 4.2.2 (2022-10-31)
- RStudio - Version 2022.07.2, "Spotted Wakerobin" Release (e7373ef8, 2022-09-06) for macOS
- `RStan` - Version 2.26.1 on macOS. For detailed instructions on how to download and install `RStan`, visit these links Website and GitHub [2].

Files necessary to replicate out analysis are available from this link. Download the

CONTACT Address: 2125 E Norris St., Unit 1, Philadelphia, PA 19125, E-mail: joseph.navelski@wsu.edu.

file *GEL_Ideology_and_Influence_Tweet.zip* which contains the following files:

- *GEL_Ideology_and_Influence_Tweet_RScript.R* - This text file is an `R` script that contains the commands to perform analysis as in our paper. Other files in the zipped file serve as inputs to the `R` commands. At the top of this file, the `R` packages necessary for analysis are listed and sourced in. We provide detailed descriptions of the steps to aid the reader.

- *StanModel-FixedLambda-Ident.R* - This text file contains the `RStan` code that defines the statistical model and the inputs for performing Bayesian (MC-MC) estimation. It defines the Jeffreys (noninformative) priors on model parameters and implements the identification approach proposed in the paper. Weadvise against modifying this file unless the reader understands `Stan` code syntax and knows how to write in `Stan`.

- *Bayes_Tweet_Functions.R* - This file presents of additional functions that create results that supplement the main analysis. It includes plotting functions for visualizing the data and plotting the densities of the priors assigned to model parameters.

- *sim_y_mat.rdata* - This file contains the data analyzed in this demonstration. It is simulated from the model fitted to the original Twitter dataset in the main paper. We generated the data to maintain the proprietariness of the original data, but allow the readers to replicate our methods and obtain similar results. The data set is a matrix with n rows representing followers, m columns representing the principals, and elements 1 when the follower follows the principal and 0 or . otherwise.

- *GEL_expert_profile_info.csv* - This file has the demographic information for the Anti-GEL and Pro-GEL principals. Included are their names, Twitter handles, account descriptions, their assumed position on GEL, and whether or not their accounts are Twitter verified.

2. Step-by-Step Instructions and Sample Output for Sanity Checks

The steps to perform a Bayesian analysis on the simulated Twitter data found in *sim_y_mat.rdata* are outlined below.

- Create a folder named Demo somewhere convenient on your PC or laptop. We assume that the complete path to this folder is "C:/Users/MyAcct/Demo". In this directory, unzip the files in *GEL_Ideology_and_Influence_Tweet.zip*. If a folder named "GEL_Ideology_and_Influence_Tweet" is created, move all the files in the latter folder one level up (i.e., out of this folder and onto "C:/Users/MyAcct/Demo").
- Start the RStudio program. Click on **File** on the menu bar (top left below the RStudio logo) and click on **Open File ...**. In the new window, locate the folder "C:/Users/MyAcct/Demo" and select (double-click on) the file *GEL_Ideology_and_Influence_Tweet_RScript.R*. Select **Open**.
- The `R` script file lists all the commands to generate similar results as in the main paper. We recommend that the codes be run in batches. Detailed comments are

included in the script file to indicate how the commands can be run in steps. To execute a section of commands in the script, simply highlight the pertinent lines and click on the **Run** button found on the top-right of the script editor window.

We provide below graphical and numerical results from running the file *GEL_Ideology_and_Influence_Tweet_RScript.R*. These results serve as a means for readers to confirm that their setups are correct and as a guide for researchers who wish to analyze their data using our proposed methods.

2.1. The Simulated Input Data Based on the Analysis of the Twitter Data

Figure 1 is a heat map of the raw data set *Sim_Y_Mat.rdata* that we simulated. It visualizes all the responses in the input data for analysis. The columns represent the $n = 46$ GEL experts and the rows represent the $m = 3,383$ followers. A black space indicates following or a connection between an expert and a follower and white spaces otherwise. A total of $N = m \times n = 155,618$ p_{ij} values are to be estimated.

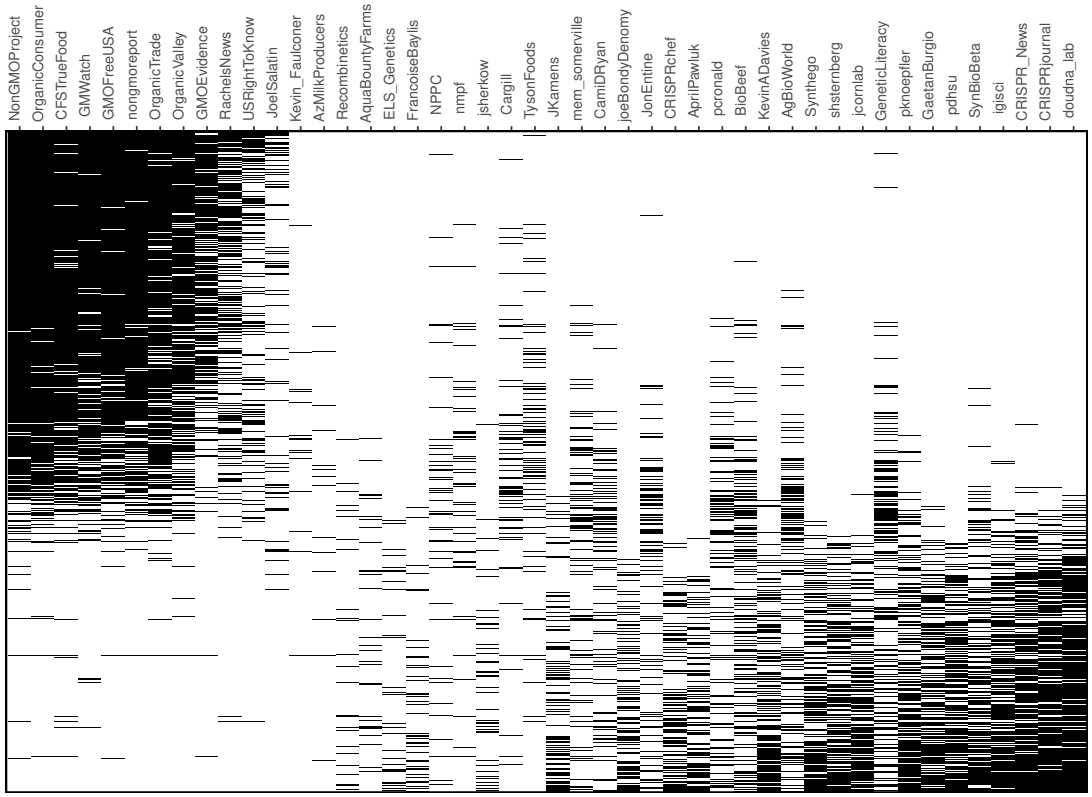


Figure 1. A heat map of the simulated following relationships between $m = 3,383$ followers and $n = 46$ GEL experts on Twitter.

2.2. Plots of Jeffreys Prior Densities and Histograms of Initial Values

Figure 2 plots the seven densities of the prior distributions assigned to model parameters and four histograms of initial values used in the Bayesian (MC-MC) estimation. The prior distributions also show the points we used for initial values indicated by

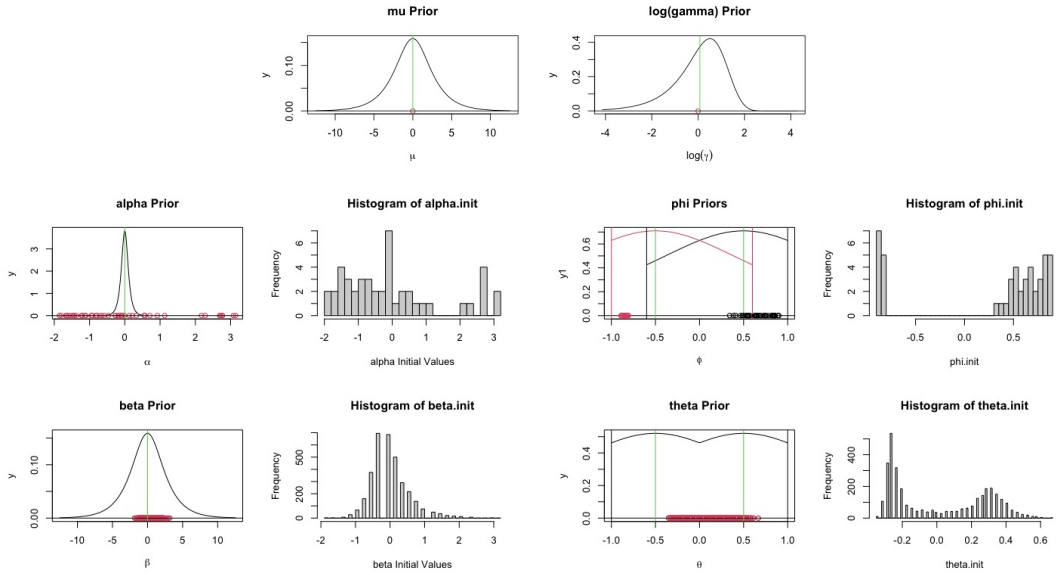


Figure 2. Plots of densities of Jeffreys prior distributions and histograms of initial values.

small red circles. The top row of plots are the prior distributions for μ and γ, and the bottom two rows are the prior distributions and histograms of initial values for α_j, ϕ_j, β_i, and θ_i. These plots are outputs in the estimation code and they should be used as a sanity check when replicating the analysis. Additionally, each prior distribution has its own set of location and scale parameters, which we pre-define in the replication code to get these distributional shapes.

2.3. Graphical and Numerical Results

2.3.1. Bayesian Estimates and Credible Intervals for Expert Popularity and Ideal Points

Figure 3 (a) and (b), plots Bayesian estimates of expert popularity ($\widehat{\alpha}_j$) and expert ideal point ($\widehat{\phi}_j$), respectively, based on MC-MC samples from the posterior distributions. Each parameter estimate comes with a box-plot whose "whiskers" represent the 95% credible interval. These figures resemble the corresponding results in the main paper. For a deeper understanding of the plots, we refer the reader to [1].

2.3.2. Bayesian Estimates of Follower Engagement, Ideal Points, and Other Model Parameters

Figures 4 (a) and (b) are histograms of the Bayesian estimates for the engagement ($\widehat{\beta}_i$) and the ideal point ($\widehat{\theta}_i$), respectively, for all followers. These plots are similar to those in the main paper which also demonstrates their interpretation. Table 1 provides statistics describing the posterior distributons of engagements, ideal points, and the rest of the model parameters. The file *Bayes_Tweet_RScript.R* contains `R` codes for creating a simpler form of this table in LaTeX format.

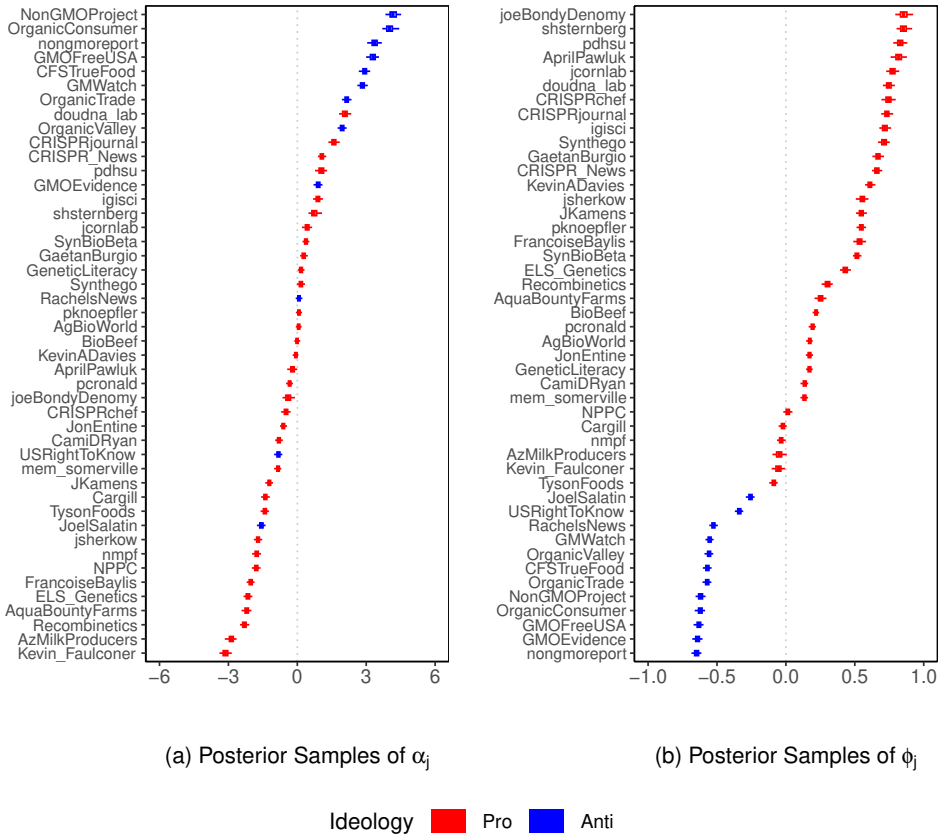
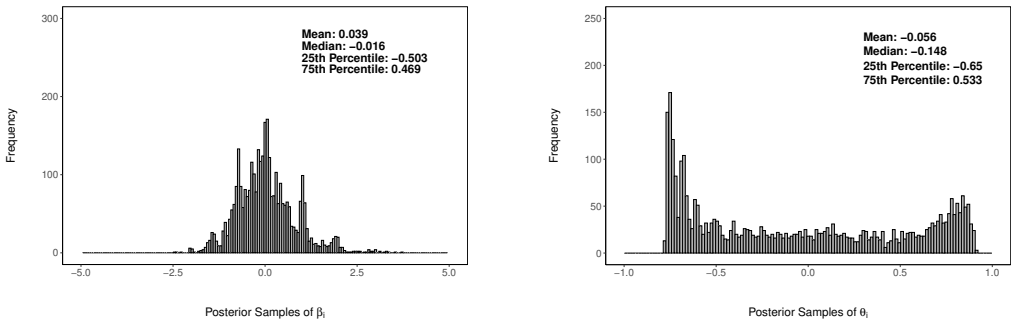


Figure 3. Bayesian estimates and 95% credible intervals for (a) popularity ($\widehat{\alpha}_j$) and (b) ideal point ($\widehat{\phi}_j$) for experts in the simulated Twitter data. Pro-GEL and Anti-GEL experts are colored blue and red, respectively.



(a) Engagement of Follower $\widehat{\beta}_i$.

(b) Ideology for Follower $\widehat{\theta}_i$.

Figure 4. Histograms and summary statistics for the estimates of (a) engagements $\widehat{\beta}_i$ and (b) ideal points $\widehat{\theta}_i$ for the 3,383 followers in the simulated Twitter data.

Table 1. Summary Statistics and 95% Credible Intervals for the Engagement $\widehat{\beta}_i$ and Ideal Point $\widehat{\theta}_i$ Estimates for the 3,383 Followers, and for μ and γ in the Simulated Twitter Data Analysis

Estimator	Mean	Median	SD	Lower 95% CI	Upper 95% CI
$\widehat{\beta}_i$	0.039	-0.016	0.799	-1.412	1.903
$\widehat{\theta}_i$	-0.056	-0.148	0.583	-0.763	0.871
$\widehat{\mu}$	-8.981	-8.965	0.223	-9.427	8.572
$\widehat{\gamma}$	9.615	9.606	0.228	9.158	10.068

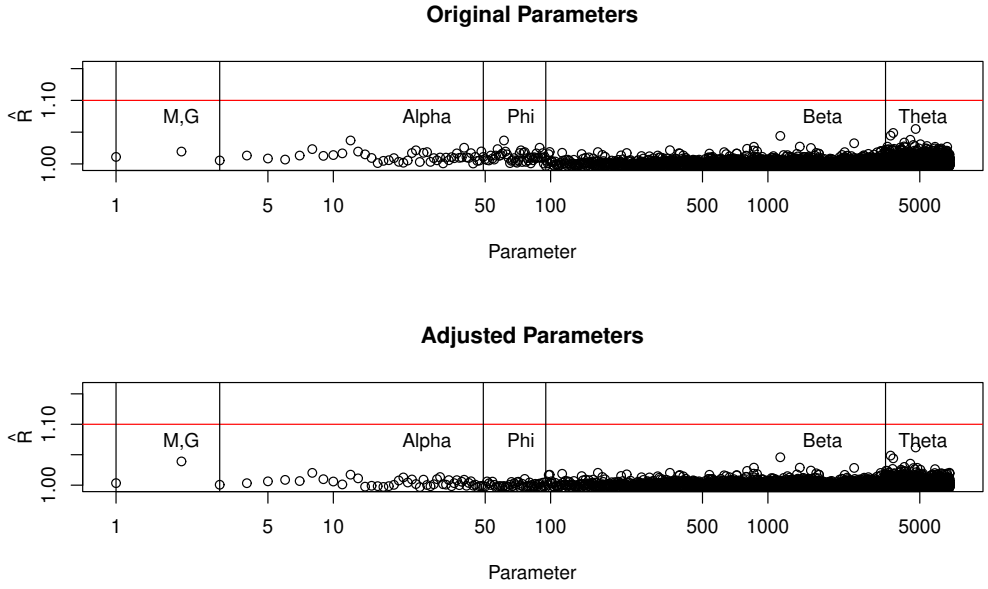


Figure 5. $\widehat{R}$ diagnostic plots for the original parameters (top) and adjusted parameters (bottom) for the simulated Twitter data.

2.4. Diagnostics for Assessing Model Fit

2.4.1. MC-MC Diagnostics

Figure 5 plots the $\widehat{R}$ values for the original (top) and adjusted, i.e., identifiable, (bottom) model parameters from the Bayesian estimation. Observe that the $\widehat{R}$ values are all below the threshold of 1.1 that the red line indicates. This threshold is the standard for MC-MC estimation recommended by [4].

2.4.2. Prediction Diagnostics

Figure 6 gives a heat map of the estimates $\widehat{p}_{ij}$ of $p_{ij} = \Pr(\text{follower } j \text{ follows influencer } i)$. A visual check confirms that the patterns in the $\widehat{p}_{ij}$'s values in Figure 6 follow closely those of the y_{ij}'s values in Figure 1. This suggests an adequate model fit and predictions.

The confusion matrix in Figure 7 (a) is created with an optimal threshold of $\tau_{\widehat{p}} = 0.246$ that is determined by maximizing [5]'s J-statistic. This optimal approach is standard practice that is meant to improve prediction ([6,7]). The overall model accuracy is $0.206 + 0.655 = .861$ and, thus, prediction with this model is 86.1% accurate. Figure 7 (b) is a plot of the receiver operating characteristic (ROC) curve with the

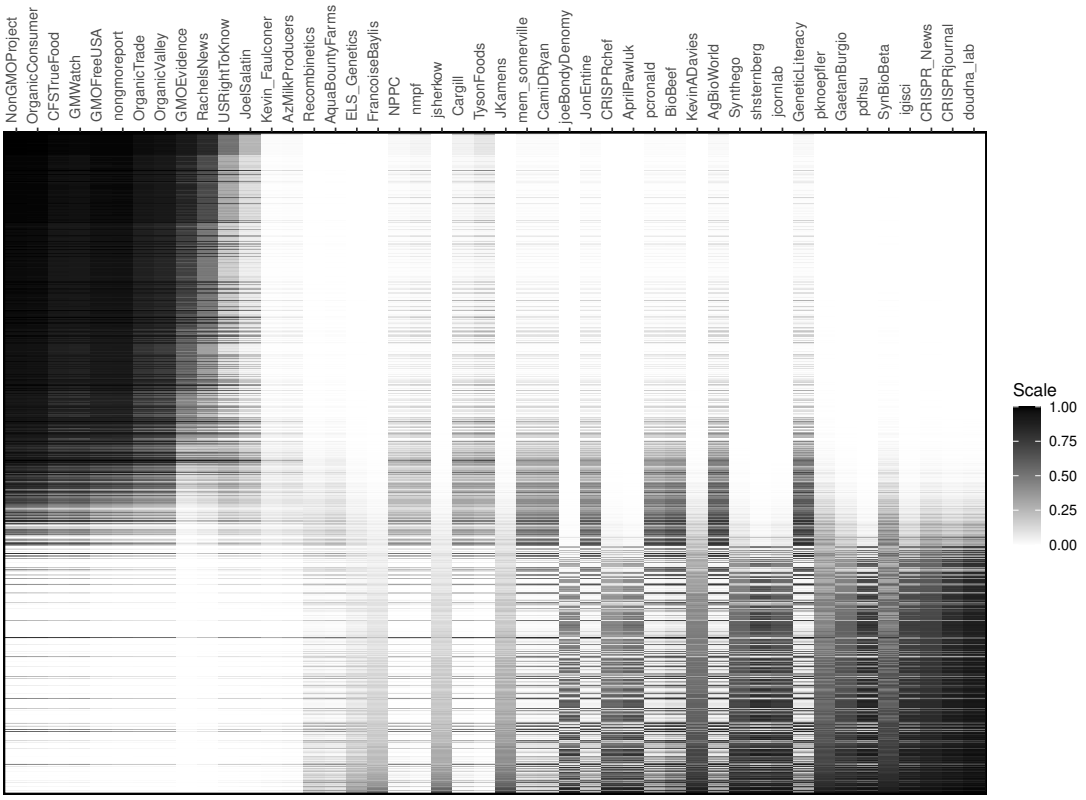
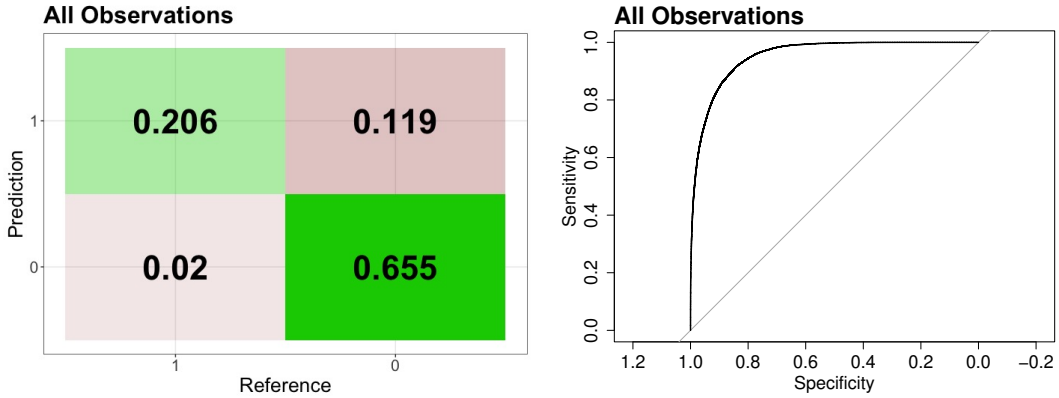


Figure 6. A heat map of Bayesian estimates $\widehat{p}_{ij}$ of the probabilities $p_{ij} = \Pr(\text{follower } i \text{ follows expert } j)$.



(a) Confusion Matrix

(b) ROC Curve

Figure 7. The confusion matrix of classification rates (a) with prediction threshold of $\tau_{\widehat{p}} = 0.246$ for $\widehat{p}_{ij}$ and a plot of the ROC curve (b) for the simulated Twitter data.

desirable shape of being relatively close to a right angle with the vertex at the point $(0, 1)$. Details related to the ROC curve are given in [1].

References

[1] Navelski J, Pascual J. New methods in estimating the latent space network model with an application to the us media industry on twitter. Working Paper. 2023;.

[2] Stan Development Team. `RStan`: the `R` interface to `Stan` ; 2021. Version 2.21.3. Accessed on January 11, 2023; Available from: https://mc-stan.org/.

[3] `R` Development Core Team. `R` project ; 2023. Accessed on February 10, 2023; Available from: http://www.r-project.org.

[4] Gelman A, Carlin JB, Stern HS, et al. Bayesian data analysis. 3rd ed. CRC Press; 2014.

[5] Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–35.

[6] Schisterman EF, Faraggi D, Reiser B, et al. Youden index and the optimal threshold for markers with mass at zero. Statistics in medicine. 2008;27(2):297–315.

[7] Fluss R, Faraggi D, Reiser B. Estimation of the youden index and its associated cutoff point. Biometrical Journal: Journal of Mathematical Methods in Biosciences. 2005;47(4):458–472.