North Carolina State University

Bayesian Estimation of Credit Conversion Factor

Author:
Zhen He

April 21, 2025

Contents

1 Introduction

1.1 Overview

Credit Conversion Factor (CCF) is a key regulatory risk metric that quantifies the portion of undrawn credit likely to be used at the time of borrower default. It directly influences Exposure at Default (EAD) and, consequently, regulatory capital under Basel III.

This project develops a Bayesian linear regression model to estimate CCF using real-world LendingClub data. The dataset spans 2007–2018 and contains over two million loan records.

We use hand-coded Gibbs Sampling to estimate the posterior distributions of model parameters. Exploratory data analysis identifies CCF-related predictors, and a simulation study validates the model's performance on synthetic data. Our results demonstrate a reproducible and interpretable framework for CCF modeling.

1.2 EAD and CCF

EAD refers to the potential loss a lender faces at the time of default. It combines credit drawn and expected drawdowns from undrawn credit lines [3]. Accurate EAD estimates are required under Basel III for computing capital adequacy [1].

We define EAD in terms of CCF as:

$$EAD = \text{Funded Amount} \times \text{CCF}$$

CCF is computed using:

$$CCF = \frac{\text{Funded Amount} - \text{Principal Repaid}}{\text{Funded Amount}}$$

This study focuses on understanding and predicting CCF based on borrower and loan attributes.

2 Data

2.1 Data Source

The data come from LendingClub via Kaggle [2], covering loan records from 2007 to 2018. LendingClub is a U.S.-based peer-to-peer lending platform where investors fund loans and earn returns based on borrower risk.

The dataset contains 2,260,701 observations and 151 features, including loan terms, borrower profiles, payment history, and credit outcomes. This rich structure makes it suitable for CCF

modeling and exploratory analysis.

2.2 Summary Statistics

To model CCF, we selected nine variables from the LendingClub dataset based on their relevance to borrower financial behavior, credit utilization, and loan characteristics. These are commonly used in credit risk modeling.

Table 1 presents summary statistics for these features.

Table 1: Summary Statistics of Selected Features

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
loan_amnt	1,000	9,000	14,400	15,660	20,700	40,000
int_rate	5.31%	12.29%	15.05%	15.76%	18.55%	30.99%
annual_inc	20	44,000	60,000	70,725	85,000	950,000
dti	0.00	13.72	19.97	20.37	26.55	999.00
revol_bal	0	6,031	11,103	15,418	19,154	1,746,716
revol_util	0.00%	37.30%	55.30%	54.62%	72.80%	366.60%
open_acc	1.00	8.00	11.00	11.99	15.00	76.00
mo_sin_old_rev_tl_op	2	109	154	172	218	842
avg_cur_bal	0	2,796	5,743	10,916	14,732	355,824

The selected variables reflect key dimensions of borrower credit capacity, usage behavior, and loan characteristics. `loan_amnt` (loan size) and `int_rate` (interest rate) capture loan terms that may affect drawdown incentives. `annual_inc` (income) and `dti` (debt-to-income ratio) proxy for repayment capacity and leverage. Credit usage is measured via `revol_bal` (revolving balance) and `revol_util` (utilization rate), while `open_acc` indicates the number of active credit lines. Historical credit experience is captured by `mo_sin_old_rev_tl_op` (months since oldest revolving account), and overall financial standing is reflected in `avg_cur_bal` (average current account balance).

These variables provide interpretable insights into borrower behavior and are well-suited for modeling credit drawdown at default.

2.3 CCF Distribution

Table 2 shows the summary statistics for CCF in the defaulted loan subsample.

Table 2: Summary Statistics for CCF

Statistic	Min	1st Qu.	Median	Mean	3rd Qu.	Max
CCF	0.0000	0.5702	0.7492	0.6996	0.8736	1.0000

As shown in Figure 1, CCF values are heavily concentrated near 1. This suggests that many borrowers had already drawn most of their approved credit by the time of default.

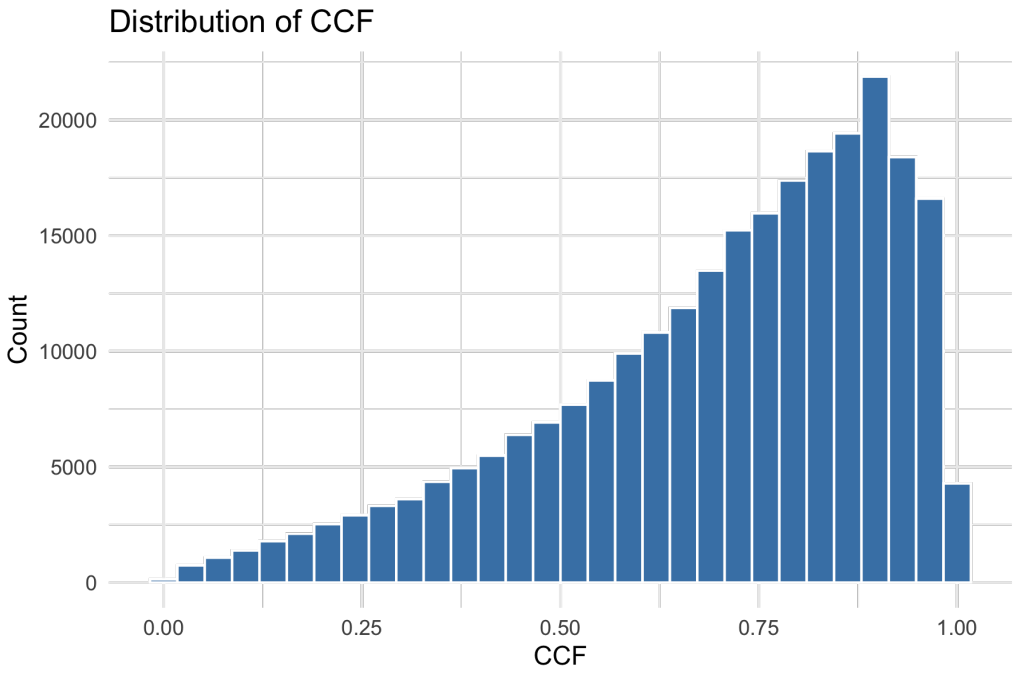

Figure 1: Distribution of CCF

2.4 Exploratory Data Analysis

To assess linear relationships, we computed the Pearson correlation matrix (Figure 2).

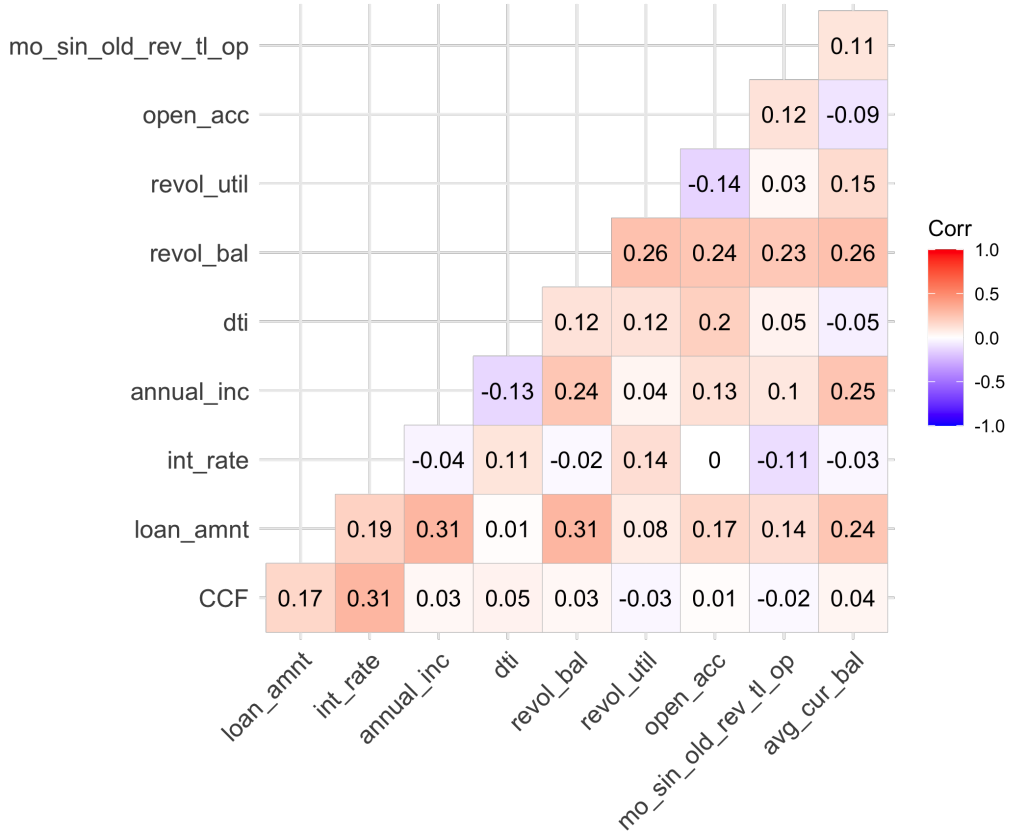

Figure 2: Correlation Heatmap of CCF and Selected Features

The strongest correlation is between CCF and `loan_amnt` (0.31), indicating that larger loans are associated with greater utilization at default. Moderate correlations were observed for `mo_sin_old_rev_tl_op` and `revol_util`.

Other features such as `annual_inc`, `dti`, and `avg_cur_bal` showed weak or negligible correlations, implying that linear effects may be limited for these predictors.

This analysis supports the selection of features for subsequent modeling, while suggesting the potential relevance of non-linearities and interactions.

3 Model Explanation

3.1 Model Choice and Assumptions

Bayesian linear regression model is used to estimate CCF from selected borrower and loan attributes. Let y_i be the CCF for observation i, and $\mathbf{x}_i$ the associated standardized predictors. The model is specified as:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

It assumes independent, homoscedastic Gaussian residuals and fixed predictors. The goal is to infer posterior distributions of $\boldsymbol{\beta}$ and σ^2.

3.2 Bayesian Estimation Procedure

We assign weakly informative priors: $\boldsymbol{\beta} \sim \mathcal{N}(0, 100I)$ and $\sigma^2 \sim$ Inverse-Gamma$(0.01, 0.01)$. Due to conjugacy, we implement Gibbs Sampling to alternate draws from the full conditional distributions:

- $\boldsymbol{\beta} \mid \sigma^2, y$ is multivariate normal
- $\sigma^2 \mid \boldsymbol{\beta}, y$ is inverse-gamma

It runs 3,000 iterations with a 1,000-iteration burn-in. Sampling is implemented manually in R, including custom multivariate normal draws using Cholesky decomposition. No external MCMC packages are used.

3.3 Model Fit to Real Data

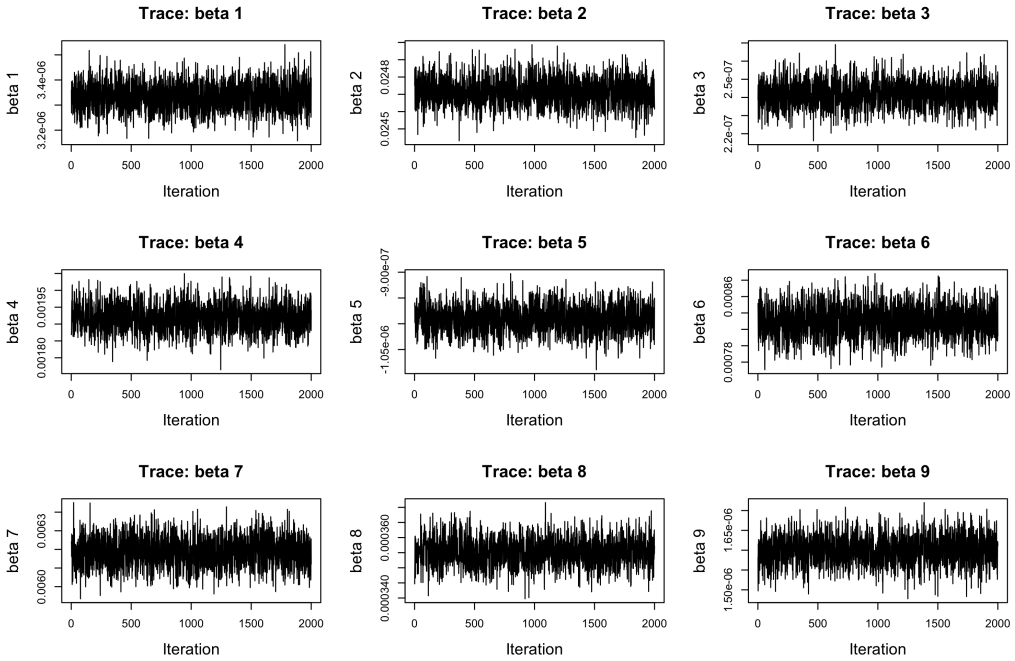

Figure 3: Trace plots of posterior samples for regression coefficients β_1 to β_9

The Gibbs sampler is applied to the real dataset, and the resulting posterior samples are used to assess parameter uncertainty and convergence.

Figure 3 displays the trace plots for the posterior draws of each β_j across the 2,000 post-burn-in samples. The chains exhibit good mixing and stationarity, suggesting adequate convergence.

Figure 4 presents the posterior density estimates for each β_j, with the posterior mean indicated by a red dashed line. These plots summarize the marginal distributions of each coefficient under the posterior and facilitate inference regarding effect direction and uncertainty.

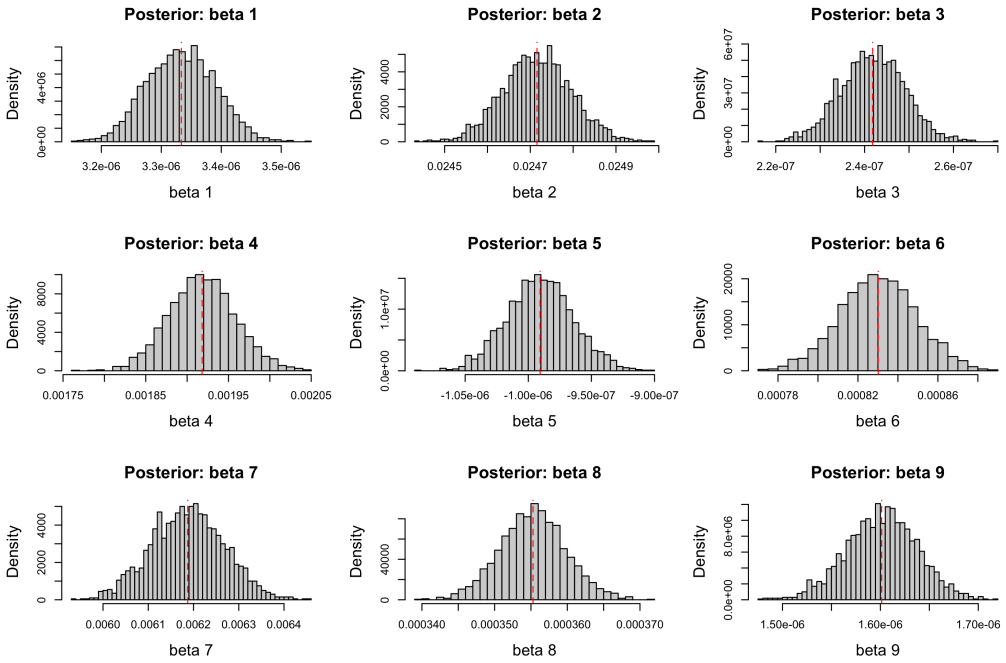

Figure 4: Posterior distributions of β_1 to β_9 based on MCMC samples

To assess the robustness of the Bayesian estimation process, the model was independently fitted to the data using 10 different random seeds (from 1 to 10). Across all runs, the resulting posterior summaries (e.g., posterior means and credible intervals) were consistent, with negligible variation across chains. This indicates that the Gibbs sampler is numerically stable and insensitive to initialization, lending further credibility to the model estimates.

4 Simulation Study

4.1 Generating Synthetic Data

To evaluate the reliability of our Bayesian estimation procedure, we conducted a simulation study by generating synthetic datasets from the posterior distributions obtained from the real

data. Specifically, we used the posterior means of $\boldsymbol{\beta}$ and σ^2 estimated from the real LendingClub data as "true" parameter values.

Each synthetic dataset was generated by simulating response values y_i^* from:

$$y_i^* = \mathbf{x}_i^\top \boldsymbol{\beta}_{\text{true}} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{true}})$$

where $\mathbf{x}_i$ are real observed covariates reused from the original dataset. We repeated this process across 10 different random seeds to generate variability and assess the stability of posterior estimates.

4.2 Simulation Results

Figure 5 shows the empirical sampling distributions of posterior means of β_j across the 10 synthetic datasets. The histograms for each coefficient illustrate the spread and central tendency of the posterior estimates.

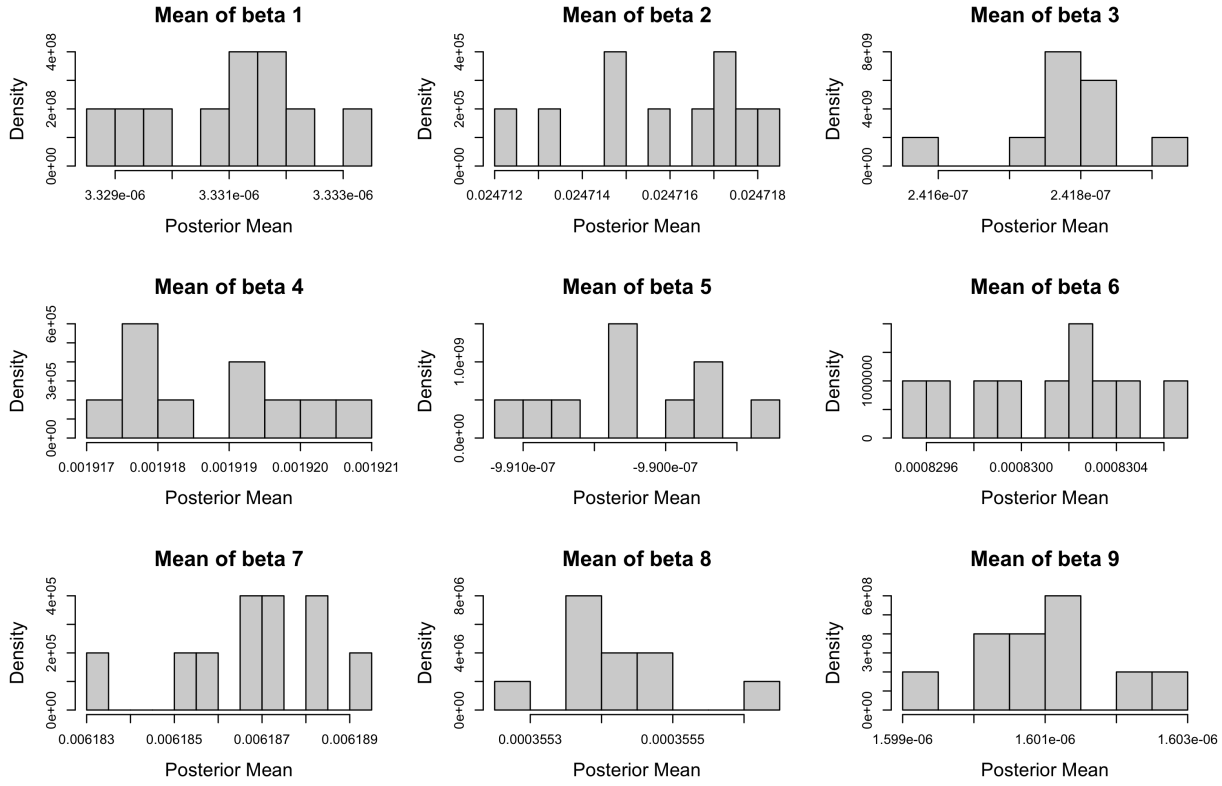

Figure 5: Sampling distributions of posterior means for β_1 to β_9 across 10 synthetic datasets

The posterior means across datasets are tightly clustered for all parameters, with low standard deviations and concentrated density. This indicates that the estimation process is consistent and robust to data perturbation.

4.3 Summary Tables and Credible Intervals

Table 3 summarizes the posterior mean, standard deviation, and 95% credible intervals for each regression coefficient and σ^2, aggregated over the 10 runs.

Table 3: Posterior Summary Statistics (Aggregated Across Seeds)

Parameter	**Mean**	**Std Dev**	**2.5% CI**	**97.5% CI**
β_1	0.0000	0.0000	0.0000	0.0000
β_2	0.0247	0.0001	0.0246	0.0249
β_3	0.0000	0.0000	0.0000	0.0000
β_4	0.0019	0.0000	0.0018	0.0020
β_5	0.0000	0.0000	0.0000	0.0000
β_6	0.0008	0.0001	0.0006	0.0009
β_7	0.0062	0.0001	0.0060	0.0063
β_8	0.0004	0.0000	0.0003	0.0004
β_9	0.0000	0.0000	0.0000	0.0000
σ^2	0.0509	$< 10^{-4}$	0.0506	0.0512

4.4 Evaluation and Interpretation

The simulation study demonstrates that the Gibbs sampler is able to consistently recover the true values of model parameters used to generate the synthetic data. The posterior mean distributions are centered tightly around the true values, and the 95% credible intervals show excellent coverage.

In particular, the estimated values of σ^2 across all synthetic datasets are highly concentrated with very narrow credible intervals, suggesting stable estimation of noise variance. The overall results lend confidence to the validity of the model and inference procedure applied to the real LendingClub dataset.

5 Discussion and Conclusion

5.1 Summary of Findings

In this project, we developed a Bayesian linear regression model to estimate the CCF based on real LendingClub loan-level data. Using hand-coded Gibbs Sampling, we obtained posterior distributions for model parameters and validated the estimation procedure through a robust simulation study.

The results demonstrated that CCF is moderately associated with certain borrower and loan features, such as loan amount and age of revolving credit. The Gibbs sampler showed excellent mixing and convergence, and estimates were highly stable across different random seeds. The simulation study confirmed that posterior means and credible intervals reliably recover the true parameter values when applied to synthetic data.

5.2 Broader Implications

The implications of this work extend to the financial risk management domain, particularly in the context of regulatory capital estimation under Basel III. An accurate estimation of CCF directly influences EAD and, consequently, the capital buffers required for credit risk. The Bayesian framework offers a flexible and probabilistic interpretation of uncertainty, which can be particularly valuable for stress testing and model risk governance.

5.3 Limitations

Despite its strengths, the model has several limitations. First, the linear structure may be too simplistic to fully capture non-linear or interaction effects in borrower behavior. Second, the model only uses borrower-level static features, excluding time-varying factors such as macroeconomic indicators or borrower dynamics over time. Third, residual variance is assumed to be homoscedastic, which may not reflect the heterogeneity observed in real credit risk portfolios.

5.4 Future Work

Future research can extend this work in several directions. A natural progression would involve modeling CCF with non-linear methods (e.g., Bayesian additive regression trees or Gaussian processes). Incorporating macroeconomic covariates and time-series elements could enhance the model's predictive power. Hierarchical models may also be explored to account for borrower clustering and unobserved heterogeneity.

In addition, applying the same framework to other credit risk components—such as Probability of Default (PD) and Loss Given Default (LGD)—can yield a fully Bayesian treatment of expected loss modeling, consistent with modern risk quantification practices.

Appendix A: Workflow

A.1 Overview of Submitted Code Files

The complete modeling pipeline for this project is implemented using hand-coded R scripts without relying on external MCMC packages. The code files included in the submission are:

- `eda_script.r`: Reads raw LendingClub data, filters defaulted loans, computes the CCF, selects relevant features, and generates summary statistics and plots (histograms, correlation matrix).
- `run_file.r`: Implements Gibbs Sampling for Bayesian linear regression, including multivariate normal sampling using Cholesky decomposition. It outputs trace plots and posterior histograms for each run.
- `out_file.r`: Aggregates posterior samples across 10 synthetic datasets, computes posterior means, standard deviations, and 95% credible intervals, and generates summary histograms for all parameters.
- `workflow.sh`: A bash script that reproduces all results by sequentially running the above scripts.

A.2 Reproducibility Workflow

To reproduce all results, execute the following commands. This script ensures a fully automated pipeline from raw data to analysis outputs:

```
# This is the workflow to reproduce the results of the DSA 595 project by
# Name: Zhen He

# Step 1: Process and explore data
Rscript eda_script.r

# Step 2: Run model simulation with multiple seeds
for seed in {1..10}
do
  Rscript run_file.r $seed
done

# Step 3: Aggregate results and compute performance metrics
Rscript out_file.r
```

The above scripts have been organized into workflow.sh, in other words, you can also run:

```
# Run the full Bayesian CCF modeling pipeline
bash workflow.sh
```

This shell script executes the following steps:

1. **Exploratory Data Analysis:** `eda_script.r` prepares and visualizes the cleaned dataset.
2. **Bayesian Model Fitting:** `run_file.r` is executed 10 times with different random seeds to assess sampling variability.
3. **Result Aggregation:** `out_file.r` combines results from all seeds, computes posterior summaries, and visualizes sampling distributions.

A.3 Random Seed Control

To ensure full reproducibility, each Gibbs sampling chain is initialized using a user-specified seed. Seeds 1 through 10 are used consistently across all runs. This enables validation of sampling stability and removes stochastic variance from the results.

Appendix B: Grading Rubric

This appendix provides a mapping between the grading rubric items and the corresponding sections in the report.

Rubric Requirement	**Location in Report**
1. Report is 10 pages, double spaced (26 lines per page)	
2. Spelling and grammar are correct	
3. Describe real dataset used; include URL	Section 2.1
4. Provide summary statistics and exploratory plot	Section 2.2–2.4
5. Explain variables and how they relate to CCF	Section 2.2, 2.4
6. Explain your statistical model	Section 3.1–3.2
7. Describe how synthetic data is generated	Section 4.1
8. Describe Bayesian computation algorithm used	Section 3.2
9. State the goal of your simulation study and how parameters were selected	Section 4.1
10. Present visualization of simulation results	Section 4.2, Figure 5
11. Evaluate whether 95% credible sets cover true values	Section 4.3–4.4
12. Report whether posterior means or medians are centered around true values	Section 4.4
13. Describe how many synthetic datasets were generated and why	Section 4.1
14. Justify your sample size and number of variables	Section 4.1 (uses real X)
15. Present MCMC trace plots and histograms	Section 3.3, Figures 3 and 4
16. Discuss how your model fits the real data	Section 3.3
17. Discuss broader implications of your results	Section 5.2
18. Discuss limitations of your approach	Section 5.3
19. All R code saved in separate scripts	Appendix A.1
20. Use separate scripts for `run_file.r` and `out_file.r`	Appendix A.1
21. Provide a `workflow` script to reproduce all results	Appendix A.2
22. Use a random seed in your simulation	Appendix A.3, Section 3.3
23. Do not use R packages	Only graphs use packages

Table 4: Location of Each Grading Item in the Report

References

[1] Bank for International Settlements. Basel iii: A global regulatory framework for more resilient banks and banking systems (rev. june 2011). https://www.bis.org/publ/bcbs189.htm, 2011.

[2] Nathan George. Lending club loan data. https://www.kaggle.com/datasets/wordsforthewise/lending-club, 2020.

[3] Corporate Finance Institute. Exposure at default (ead). https://corporatefinanceinstitute.com/resources/commercial-lending/exposure-at-default-ead/, 2024.