2025, P8116, Group Projects

Project 1: Multivariate Non-Normal Distributions and Correlated Data

In many real-world scenarios, data deviate from the classic assumptions of normality and independence among variables. In this project, you will a simulation study to Generating non-normal, correlated data (e.g., multivariate t-distribution or skewed distributions). Investigating and illustrating what happens if we mis-specify correlation or distributional assumptions in a regression. Find a statistical inference method that can accommodate both non-normality and correlation, then demonstrate its validity through simulations.

In your simulation study, you need to complete the following tasks:

1.**Plan and specify distributions and correlation**

- Select at least one multivariate distribution either with heavier tail or skewness (e.g. multivariate t, skewed-normal. . .)
- Decide on a target correlation structure (e.g. block-diagonal, compound sysmetric or any user-defined correlation matrix) (Extra credit) Use a copula-based transformation (e.g. Gaussian, Clayton or Gumbel) to induce a chosen correlation structure among variables

2.**Implement your data generation method**

- Write an R function to generate data from your chosen distribution
- Evaluate and illustrate whether the empirical distribution of your generated data align with the your target distribution and correlation structure

3.**Design a simulation study 1**

- Fit a regression model to the simulated data while assuming that data are i.i.d and normally distributed
- Conduct a simulation study that varies sample sizes, correlation levels, or degree of non-normality to assess bias in parameter estimates, type I error in hypothesis testing and coverage in conference intervals.
- Demonstrate how inference changes as the severity of assumption deviation increases (more skew, heavier tail, stronger correlation)

4.**Design a simulation study 2**

- Identify a statistical inference tool that could accommodate the correlation and non-normality ((e.g., a robust regression approach, generalized estimating equations (GEE) for correlated data, or a Bayesian method with appropriate priors).
- Use simulation that is similar to Simulation 1 to this inference tool provides valid results for non-normal, correlated data, and compare these results to the standard inference tool in study 1.

5.**Write a report in R-markdown to summarize your simulation study and discuss practical implications in real-world analyses. Please make sure to include the scripts or functions used for data generation and analysis. Your scripts and reports should be reproducible.**

Project 2: Distribution Generation with Advanced Rejection Sampling Methods

In class, you learned a simple acceptance-rejection algorithm (similar to importance sampling), where you sample from a convenient distribution and accept or reject each sample based on the ratio of the target PDF to the convenience PDF. While straightforward, this method can become inefficient when the target distribution is high-dimensional, highly skewed, or multimodal—and the convenience distribution is a poor match.

This project extends your knowledge by introducing two advanced sampling methods—Adaptive Rejection Sampling (ARS) and Slice Sampling—and asks you to compare these with the basic acceptance-rejection approach from class. You will:

1.Implement and evaluate ARS and slice sampling.

2.Compare them to the basic acceptance-rejection method (taught in class).

3.Identify when ARS and slice sampling might be better choices.

Brief Introductions to the Advanced Algorithms

1.Adaptive Rejection Sampling (ARS)

- *Key Idea*: Constructs piecewise linear bounds around the log of the PDF, refining these bounds after evaluating the density at new points. This adaptation often yields higher acceptance rates than basic rejection sampling for log-concave or nearly log-concave distributions.
- *Reference*: Gilks, W. R. & Wild, P. (1992). "Adaptive Rejection Sampling for Gibbs Sampling." Applied Statistics, 41(2), 337–348.

2.Slice Sampling

- *Key Idea*: Samples from the "slice" beneath the curve of the PDF by introducing an auxiliary variable. This method can handle distributions that aren't strictly log-concave and often adapts naturally to local shape changes, therefore handles multimodal or non-log-concave distributions more gracefully.
- *Reference*: Neal, R. M. (2003). "Slice sampling." Annals of Statistics, 31(3), 705–767.

In your simulation study, you need to complete the following tasks:

1.Select a univariate or bivariate target distribution that is NOT trivial to sample from inverse quantile function. You may incorporate features like skewness, heavy tails and multiple modes to highlight the limitation of the basic acceptance-rejection algorithm

2.Implement the three methods to generate the same number of sample (about 5,000-10,000), and compare

- Acceptance rate,
- Average number of iterations per sample point
- Total computing time.
- Accuracy, i.e. how well the empirical distribution match with the true target distribution. Use effective visualization tools to illustrate the difference across the tree methods
- (Extra credit) Parallelize your algorithm into multiple CPU core, and report the CPU time reduction by using parallel computing.

3.Write a report in R-markdown to summarize your simulation study and discuss practical implications in real-world analyses. Please make sure to include the scripts or functions used for data generation and analysis. Your scripts and reports should be reproducible.

Project 3: Hierarchical Logistic Model for mulit-center clinical trial

In a multi-center clinical trial, outcomes are measured across multiple clinics or hospital sites, each with its own unique characteristics. Such data often include

- **Patient-Level Covariates**: Each patient i has a continuous risk factor X_i (e.g., a biomarker value or disease severity score).
- **Clinic-Level Random Effects**: Each clinic j has a random intercept b_j that captures site-specific factors (patient population differences, local procedures, etc.).
- The outcome of interest is **probability of an adverse event** for patient i in clinic j, modeld by

  $$p_{i,j} = logit^{-1}(\alpha + b_j + \beta X_i),$$

  where b_j is the clinic random effect that follows some distribution $f(b)$ that could be heavy-tailed or skewed. The patient-level covariate X_i might also be non-normal — say a Gamma or lognormal distribution to capture substantial skew.
- **Statistics of interest:** We are interested in estimating the overall (population-level) probability of an adverse event, marginalizing over both X (across all patients) and the random effect distribution b (across all clinics):

$$P(\text{Adverse Event}) = \int\int logit^{-1}(\alpha + b + \beta x) f(b) f(x) db dx$$

Note: If the random-effects distribution $f(b)$ and the risk-factor distribution $f_X(x)$ are highly skewed, heavy-tailed, or multimodal, then: Simple Monte Carlo integration may require very large sample sizes to explore all relevant regions (especially tails). When the integral is multi-dimensional (b and x) , and support of x and b can vary widely. It would be desirable to consider variance reduction techniques. You have learned techniques in the class, Control Variates and Importance Sampling.

In your simulation study, you need to complete the following tasks:

1. Specify distributions for b and x. Both b and X should be non-normal.
2. Sample (b_j, x_i) from $f(b)$ and $f(X), and estimate P(Adverse Event) using the simple Monte-Carlo integration.
3. Design a control variate to reduce the variance of estimation
4. Further implementing important sampling strategies to focus on high-probability regions.
5. Compare bias, variance, and CPU time for simple MC, control variates, and importance sampling.
6. (Extra credit) Use cumulative convergence plots to illustrate how quickly each method converges to the "true" value.
7. Write a report in R-markdown to summarize your simulation study and discuss practical implications in real-world analyses. Please make sure to include the scripts or functions used for data generation and analysis. Your scripts and reports should be reproducible.