# Gibbs Sampling with Data Augmentation for Bayesian Analysis of Binary and Polychotomous Response Data

STAT 230 Final Report

*Damon Bayer & Corey Katz*

*12/16/2019*

## Contents



## 1 Abstract

This project focuses on implementations and applications of several of the algorithms presented in Albert and Chib (1993), including data-augmented Gibbs sampling for probit regression on binary response data, multinomial probit regression on ordinal response data, and t-link regression on binary response data. We compare these method to the equivalent MLE estimates, as well as a more modern MCMC method, No-U-Turn Samplers, as implemented in Stan. Broadly, these methods all resulted in similar point estimates for regression parameters. The data-augmented Gibbs sampling performed significantly faster than the Stan models, but had some issues with mixing in the posterior distributions. Further research has been done to apply the methods here to regression methods on binary and polychotomous response data with other link functions. The authors hope to apply some of the concepts developed here in an ongoing research project to model pitching change decisions in baseball.

## 2 Introduction

Markov Chain Monte Carlo (MCMC) methods revolutionize Bayesian inference, making it possible to sample from complex posterior distributions. As computing power continues to evolve, so do sampling methods, as well as the complexity of models scientists try to fit. While advances have been made to MCMC, sampling

from non-conjugate posterior distributions is still a difficult and time-consuming task. Gibbs sampling is one of the most common forms of MCMC and one of the earliest attempts to alleviate these struggles of sampling from intractable posteriors was the development of data-augmentation Gibbs Sampling for regression models.

Albert and Chib (1993) was the first paper to introduce using data augmentation to fit Bayesian probit regression models to binary and polychotomous data. It builds upon the foundation of Gibbs sampling in Gelfand and Smith (1990) and Tanner and Wong (1987). Adding latent variables has been used in many other areas of statistics and mathematics to aid in computation. By introducing a latent variable to the probit regression model of a binary response variable, they are able to simplify the Gibbs sampling process and make sampling from the posterior distribution easier. This paper laid the groundwork for using data-augmentation Gibbs Sampling for Bayesian inference by enabling the computation of exact posterior distributions of regression coefficients for binary and polychotomous response.

Compared to traditional maximum likelihood estimation (MLE), Albert and Chib's approach is advantageous in the case of small data, where MLE can be biased, as well as in the case of regression models with complicated likelihoods, such as in the multivariate probit case. Additionally, this method enables the modeling of the marginal distribution of residuals, which are on a continuous scale and can be more helpful for outlier detection than frequentist residuals, which only take on two values.

This paper focuses on implementations and applications of several of the algorithms presented in Albert and Chib (1993), including the aforementioned data-augmented Gibbs sampling for probit regression on binary response data, multinomial probit regression on ordinal response data, and t-link regression on binary response data. We compare these method to the equivalent MLE estimates (Generalized Linear Models), as well as a more modern MCMC method, No-U-Turn Samplers, as implemented in Stan (Carpenter et al. 2017). By comparing these three methods, we can better understand the importance of the introduction of data augmentation to Bayesian inference.

# 3 Methodology

In this section, we detail each method for data-augmented Gibbs sampling, including pseudocode for each algorithm.

## 3.1 Probit Model

The first model that data augmentation can be used to find posterior distributions is the probit model. The following is a standard Bayesian probit model with non-informative priors on the regression coefficients:

$$y_i|\pi_i, \vec{\beta}, \vec{x}_i \sim Bernoulli(\pi_i) \quad \text{for} \quad i = 0, ..., n \tag{1}$$

$$\pi_i = \Phi(x_i^T \beta) \tag{2}$$

$$\beta_j \sim N(0, 100), \quad \text{for } j = 0, ..., p \tag{3}$$

Where $\Phi$ is the cumulative density function of the normal distribution.

The idea of data augmentation is to introduce a continuous latent variable derived from a binary response in order to make sampling from the posterior distributions of the probit model coefficients easier. By introducing independent latent variables $Z_1, ...Z_n$, we can now find the joint posterior of $\vec{\beta}$ and $\vec{Z}$ given Y. We can then marginalize over the posterior distribution of $\vec{Z}|\vec{Y}$ and thus we have the conditional posterior of $\beta|\vec{Z}, \vec{Y}$. This method is further simplified if we assume (as we did above) non-informative priors on the regression coefficients, $\vec{\beta}$ (Albert and Chib 1993).

With this approach we are able to create a Gibbs sampler that only needs to sample from truncated normal distributions and multivariate normal distribution, which are extremely easy with today's computing power. Based on the fully conditional posterior of $\beta|\vec{Z},\vec{Y}$, Albert and Chib equated this method of probit regression on binary Y to doing linear regression on the latent variable Z. This will be evident in the algorithm as we use the `lm` function to find the mean of the posterior distribution of $\beta|\vec{Z},\vec{Y}$ (Albert and Chib 1993).

#### 3.1.1 Algorithm:

**Input :** $\vec{Y}$,**X**
**Output:** Posterior Samples of Regression Coefficients, $\vec{\beta}$

Set Number of Samples (Total ($N_s$) and Burn-in) Initialize $\vec{\beta}^{(0)}$
Set $\Sigma = (X^TX)^{-1}$ n = Number of Observations
**for** *k = 1 to* $N_s$ **do**
  **for** *i = 1 to n* **do**
    **if** $y_i$ *= 1* **then**
      Sample $z_i^{(k)}$ from trunc$\mathcal{N}(x_i^T\beta^{(k-1)}$,1,0,$\infty$)
    **else**
      Sample $z_i^{(k)}$ from trunc$\mathcal{N}(x_i^T\beta^{(k-1)}$,1,$-\infty$,0)
    **end**
  **end**
  Regress $\vec{Z}$ onto **X** to find $\vec{\beta}_Z$
  Sample $\beta^{(k)}|\vec{Z}$ from $\mathcal{N}(\hat{\vec{\beta}}_Z^{(k)}, \mathbf{\Sigma})$
**end**

**Algorithm 1:** Probit Regression Using Gibbs Sampler with Data Augmentation

### 3.2 T-link to Approximate Logistic Regression

In this section we discuss an extension of the Gibbs sampler discussed for the probit model. The purpose of this model is to use data augmentation with t-distributions to approximate the logistic regression model for a binary response. The simple Bayesian Logistic Regression model with non-informative priors is as follows:

$$y_i|\pi_i,\vec{\beta},\vec{x}_i \sim Bernoulli(\pi_i) \quad \text{for} \quad i = 0,...,n \tag{4}$$

$$\text{logit}(\pi_i) = x_i^T\beta \tag{5}$$

$$\beta_j \sim N(0,100), \quad \text{for } j = 0,...,p \tag{6}$$

Where $\text{logit}(p) = \frac{p}{1-p}$.

Rather than using the normal cdf (as in the probit model) we use the t-distribution cdf as our link function. Replacing $\Phi$ with the cdf of the t($\nu$) in (2), we achieve the Bayesian t-link model. By generalizing the model, we can now choose a degrees of freedom, $\nu$, where t($\nu$) better fits out model and thus a link function that approximates other well-known link functions. For example, many practitioners prefer logistic regression because of the interpretability of the coefficients as odds ratios. The flexibility of the t-link function allows us to draw from a posterior distribution of $\vec{\beta}$ that is approximately the posterior distribution if we had chosen to use the logit link function. Note that if we set the degrees of freedom equal to infinity, we would revert back to the probit model (Albert and Chib 1993).

A t-distribution with a degrees of freedom of 8 is a fairly close approximation of the logistic regression model, once a correction factor of 0.634 is taken into account (Albert and Chib 1993).

To implement this link function, we introduce a second set of latent variables, $\lambda_i$. This variable is introduced to simulate the heavier tales of the t-distribution compared to the normal distribution. Our implementation is slightly different from Albert and Chib's, due to some issues we encountered with their exact implementation. Albert and Chib describe a procedure for evaluating different degrees of freedom in the t-link in order to choose the best fit to the data, but we were unable to successfully implement it for this project. Because of this, we choose to focus only on 8 degrees of freedom to enable comparisons to logistic regression.

#### 3.2.1 Algorithm:

**Input :** $\vec{Y}$,$\mathbf{X}$
**Output:** Posterior Samples of Regression Coefficients, $\vec{\beta}$

Set Number of Samples (Total ($N_s$) and Burn-in)
n = Number of Observations
Initialize $\vec{\beta}^{(0)}$ and set $\lambda = \vec{1}_n$
$\nu = 8$
**for** *k = 1 to* $N_s$ **do**
  **for** *i = 1 to n* **do**
    **if** $y_i$ *= 1* **then**
      Sample $z_i^{(k)}$ from trunc$\mathcal{N}(x_i^T\beta^{(k-1)},\lambda_i^{-1},0,\infty)$
    **else**
      Sample $z_i^{(k)}$ from trunc$\mathcal{N}(x_i^T\beta^{(k-1)},\lambda_i^{-1},-\infty,0)$
    **end**
  **end**
  W = diag($\vec{\lambda}$)
  Set $\Sigma = (X^TWX)^{-1}$
  $\vec{\beta}_Z = \Sigma X^TWZ$
  Sample $\beta^{(k)}|\vec{Z},\vec{\lambda},\vec{Y}\nu$ from $\mathcal{N}(\hat{\vec{\beta}}_Z^{(k)},\Sigma)$
  Sample $\lambda_i$ from $\Gamma\left(\frac{\nu+1}{2},\frac{\nu+(Z_i-x_i^T\beta)^2}{2}\right)$
**end**

**Algorithm 2:** Tobit Regression Using Gibbs Sampler with Data Augmentation ($\nu = 8$)

### 3.3 Multinomial

Data augmentation Gibbs samplers can also be applied to multinomial response variables. Although there are approaches for both ordered and unordered categories, we will focus on the case of ordered categories. Once again, using a probit model will simplify the conditional posterior distributions. Before discussing the algorithm in depth, we present the ordered multinomial probit (Aitchison and Silvey 1957; Gurland, Lee, and Dahm 1960; McCullagh 1980) model for which we will use Gibbs Sampling to fit.

$$Y_i|\vec{x}_i,\beta \sim Multinomial(p_{ij}) \tag{7}$$

for $i = 1,...,n$ and $j = 1,...,J-1$.

$$p_{ij} = \Phi(\gamma_j - \vec{x}_i^T\vec{\beta}) \tag{8}$$

$$\vec{\gamma}, \vec{\beta} \sim \pi(\vec{\beta}, \vec{\gamma}) \tag{9}$$

where $\pi(\vec{\beta}, \vec{\gamma})$ is the prior on the regression coefficients. We will assume to be non-informative as is commonly done in analysis.

Similar to the binary case, a latent variable $\vec{Z} = (Z_1, ...Z_n)$ is introduced making it possible to simulate from a joint posterior. Further, we can marginalize over $\vec{Z}$ to find the posterior distribution of $\beta|\vec{Z}, \vec{Y}, \vec{\gamma}$, which is the same multivariate normal distribution as the binary case, and the conditional posterior of $\vec{\gamma}|\vec{Z}, \vec{Y}, \vec{\beta}$. This Gibbs sampler requires only sampling from truncated normal distributions, a multivariate distribution, and a uniform distribution (Albert and Chib 1993). The algorithm below presents, in detail, the sampling mechanism of this data augmentation Gibbs sampler.

#### 3.3.1 Algorithm:

**Input :** $\vec{Y}$,**X**
**Output:** Posterior Samples of Regression Coefficients, $\vec{\beta}$ and $\vec{\gamma}$

Set Number of Samples (Total ($N_s$) and Burn-in)
n = Number of Observations
Initialize $\vec{\beta}^{(0)}$ and $\vec{\gamma}^{(0)}$ as the MLE
**for** *k = 1 to* $N_s$ **do**
  Sample $\vec{\gamma}_j$ from $Uniform(max(max(Z_i : Y_i = j), \gamma_{j-1}^{(k-1)}, min(min(Z_i : Y_i = j+1), \gamma_{j+1}^{(k-1)}))$
  **for** *i = 1 to n* **do**
    Sample $z_i^{(k)}|\vec{\beta}, \vec{\gamma}, y_i = j$ from trunc$\mathcal{N}(x_i^T \beta^{(k-1)}$,1,$\gamma_{j-1}, \gamma_j)$
  **end**
  Set $\Sigma = (X^T X)^{-1}$
  Set $\vec{\beta}_Z = \Sigma X^T Z$
  Sample $\beta^{(k)}|\vec{Z}, \vec{\gamma}, \vec{Y}$ from $\mathcal{N}(\hat{\vec{\beta}_Z^{(k)}}, \Sigma)$
**end**

**Algorithm 3:** Ordered Multinomial Probit Regression Using Gibbs Sampler with Data Augmentation

### 3.4 Implementation in Stan (NUTS)

While not the main focus of this paper, we decided to implement the previously described models in a modern MCMC method, a No-U-Turn Sampler (NUTS). NUTS are an extension of Hamiltonian Monte Carlo. We used Stan (through the `Rstan` package in R (Carpenter et al. 2017; Stan Development Team 2019)) to implement this models using NUTS.

## 4 Results

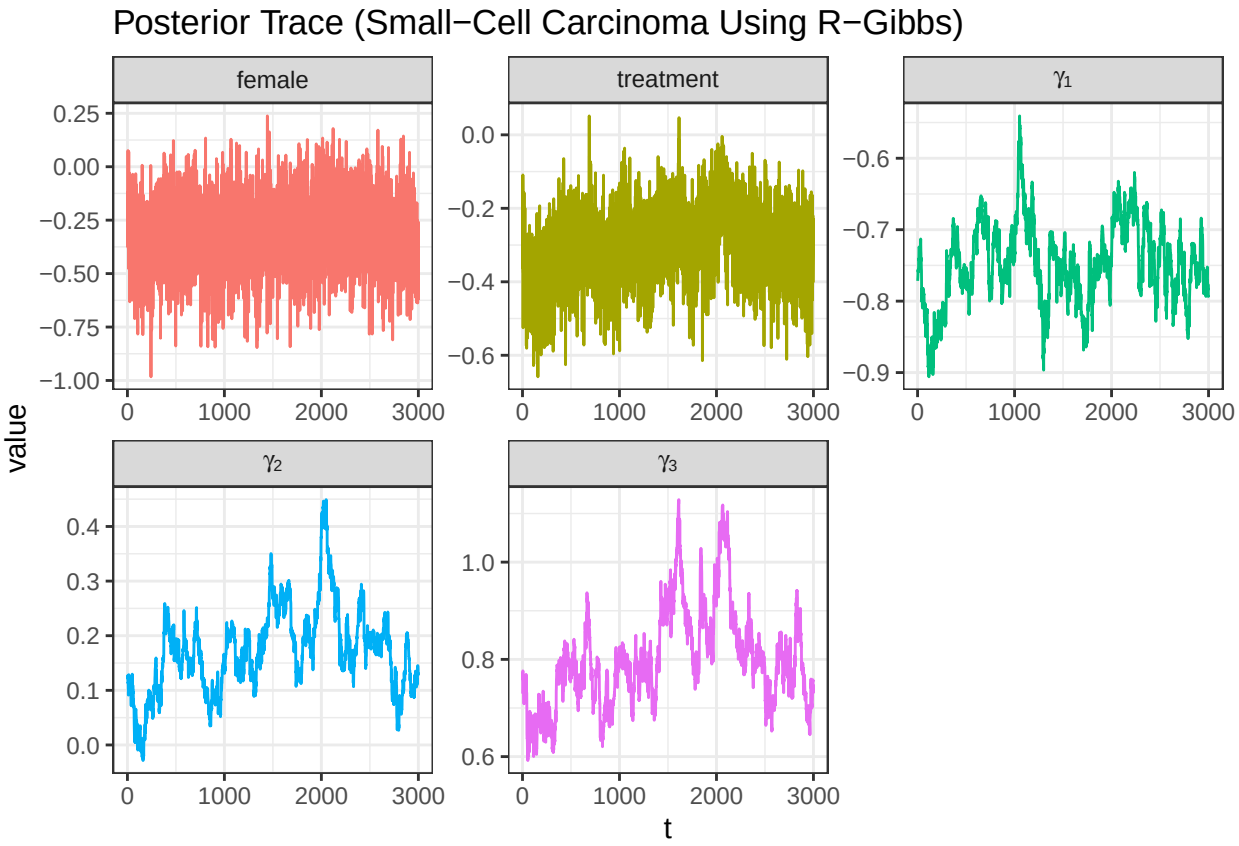
For three data sets, we apply the appropriate previously discussed model as well as standard maximum likelihood estimation and custom models written in Stan (Carpenter et al. 2017) and implemented in the RStan package (Stan Development Team 2019).

## 4.1 Small-Cell Carcinoma

The small-cell carcinoma data comes from our STAT 211 class. Small-cell carcinoma of the lung is an aggressive cancer that can be treated with chemotherapy. Patients with small-cell carcinoma were randomly assigned to one of two therapy options. The outcome of the therapy is recorded on an ordinal scale (1: Progressive, 2: No Change, 3: Partial Remission, 4: Complete Remission). We fit multinomial ordered probit regression models with with therapy option and sex as predictors of outcome. The model parameters are estimated with data augmented Gibbs sampling following the algorithm described in Albert and Chib (1993) and implemented in R, a custom model built in Stan, and traditional maximum likelihood estimation as implemented in the `polr` function in the MASS package (Venables and Ripley 2002). Both Bayesian methods are run to generate 4000 posterior samples, with the first 1000 discarded as burn-in samples, leaving 3000 samples for analysis. The Stan model took 1 minute and 13.9 seconds seconds to run, while the R model completed in 2.4 seconds.

Table 1: Posterior sample summaries for small-cell carcinoma data using R-Gibbs

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| female | -0.3177006 | 0.1695306 | -0.6483304 | -0.3154434 | 0.0143534 |
| treatment | -0.3081612 | 0.1039444 | -0.5136893 | -0.3092713 | -0.1059862 |
| gamma[1] | -0.7478856 | 0.0569057 | -0.8576571 | -0.7492453 | -0.6420117 |
| gamma[2] | 0.1726966 | 0.0795173 | 0.0235177 | 0.1732758 | 0.3510630 |
| gamma[3] | 0.8011110 | 0.1027040 | 0.6390310 | 0.7890847 | 1.0615511 |



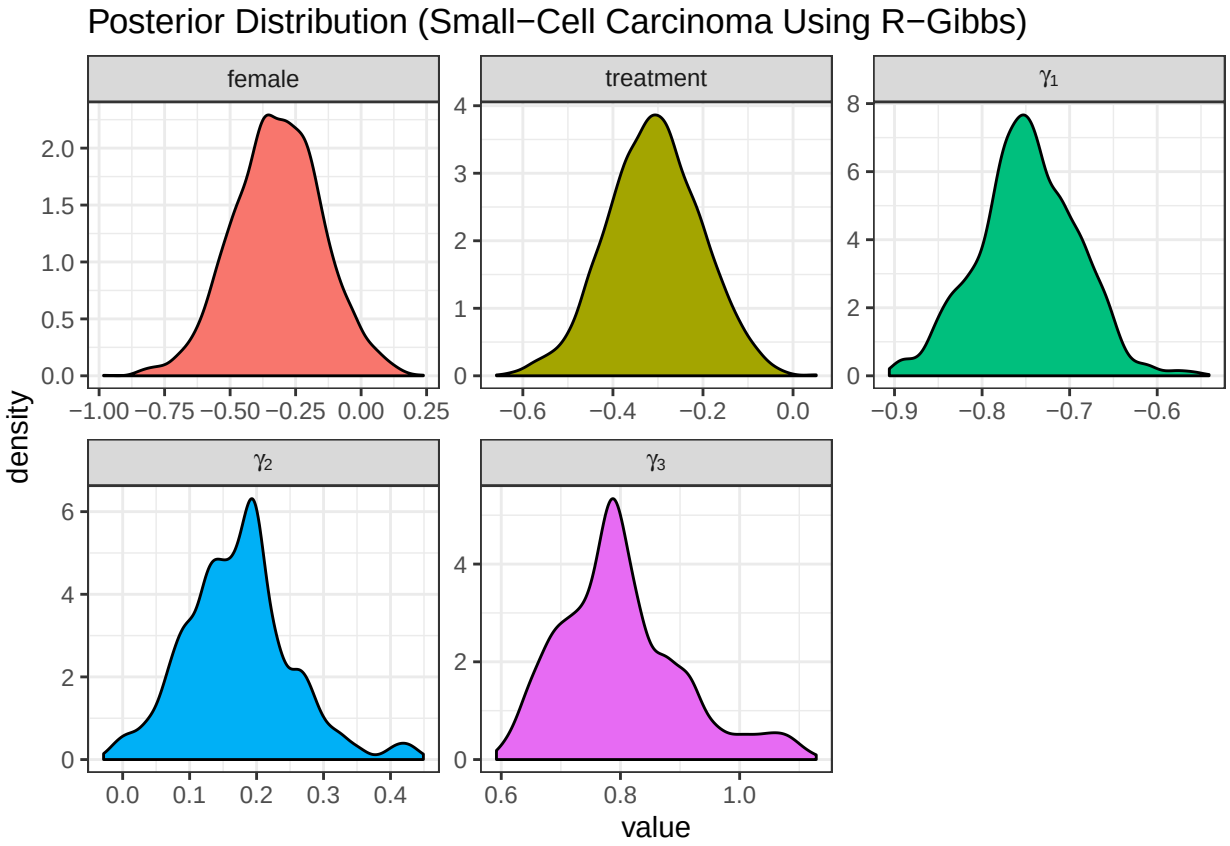
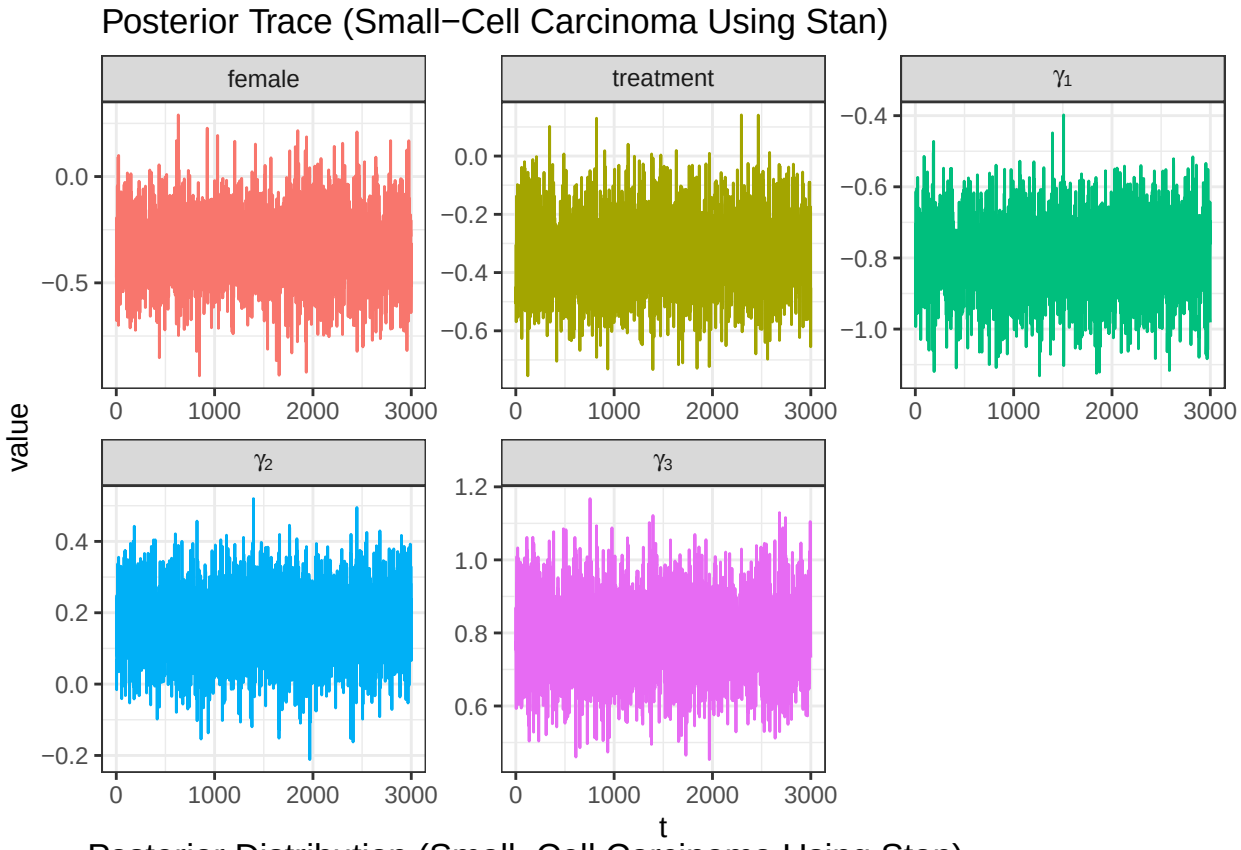



Table 2: Posterior sample summaries for small-cell carcinoma data using Stan

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| female | -0.3383656 | 0.1702371 | -0.6777031 | -0.3369185 | -0.0045383 |
| treatment | -0.3311899 | 0.1282283 | -0.5829356 | -0.3309320 | -0.0730097 |
| gamma[1] | -0.7992210 | 0.1048007 | -1.0079013 | -0.7988986 | -0.5917817 |
| gamma[2] | 0.1684474 | 0.1002620 | -0.0302685 | 0.1692335 | 0.3625867 |
| gamma[3] | 0.7914654 | 0.1058342 | 0.5811806 | 0.7903738 | 1.0012122 |


Posterior Trace (Small−Cell Carcinoma Using Stan)
female
treatment
γ1
γ2
γ3
value
t
0.0
−0.5
−0.2
−0.4
−0.6
−0.8
−1.0
0.4
0.2
1.2
1.0
0.8
0.6
0
1000
2000
3000

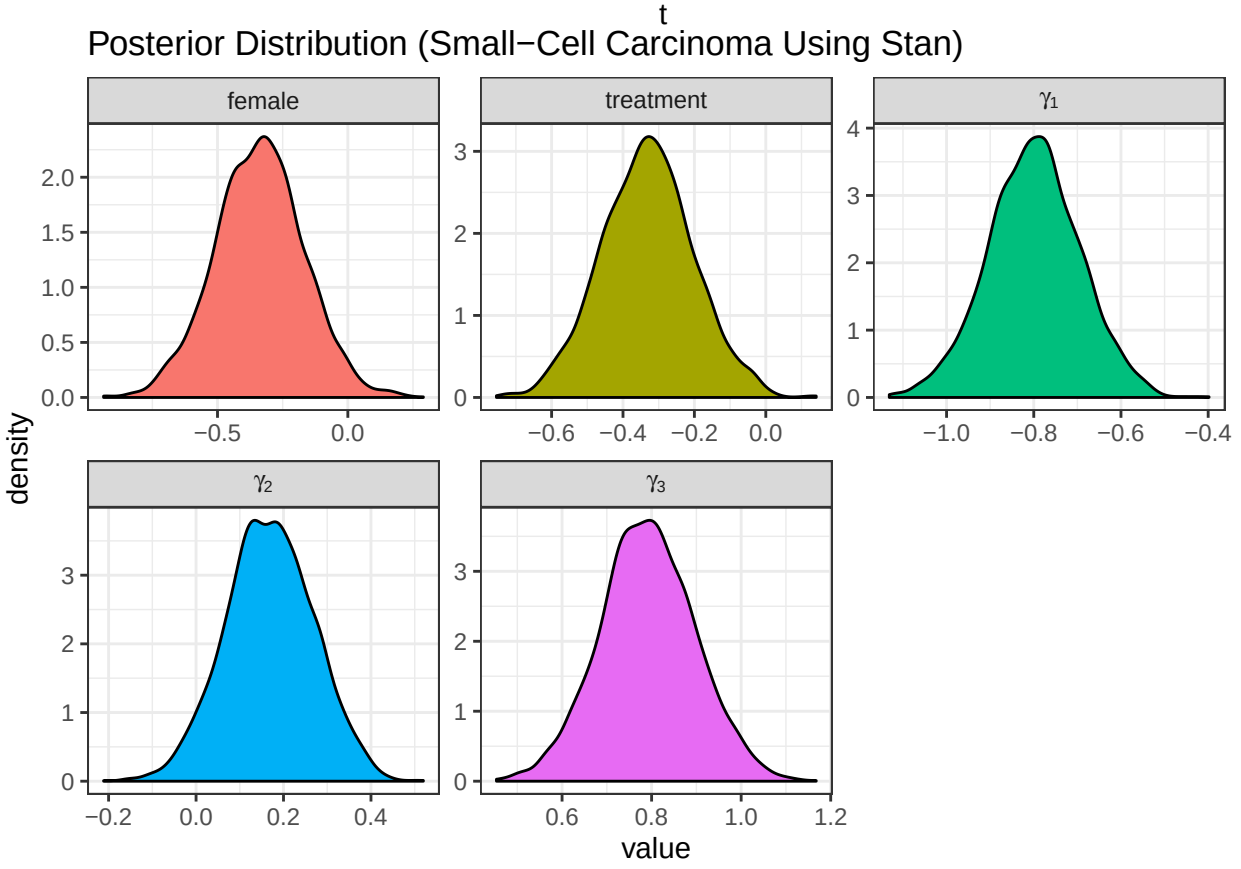

Posterior Distribution (Small−Cell Carcinoma Using Stan)
female
treatment
γ1
γ2
γ3
density
value
2.0
1.5
1.0
0.5
0.0
3
2
1
4
−0.5
−0.6
−0.4
−0.2
−1.0
−0.8
0.2
0.4
0.6
0.8
1.2

Table 3: MLE estimates for small-cell carcinoma data

| Variable | Estimate | Std. Error |
|---|---|---|
| female | -0.3401606 | 0.1749021 |
| treatment | -0.3344764 | 0.1254351 |
| gamma[1] | -0.7994983 | 0.1053802 |
| gamma[2] | 0.1649279 | 0.0988553 |
| gamma[3] | 0.7818721 | 0.1068102 |

For this set of data, the results from the three models are fairly similar. The trace plots for the Gibbs sampler show a large degree of auto-correlation for the *gamma* variables, which is a likely explanation for the difference between the Gibbs and Stan estimates. We also note asymmetric posterior distributions. This issues will be further discussed in Section 5.

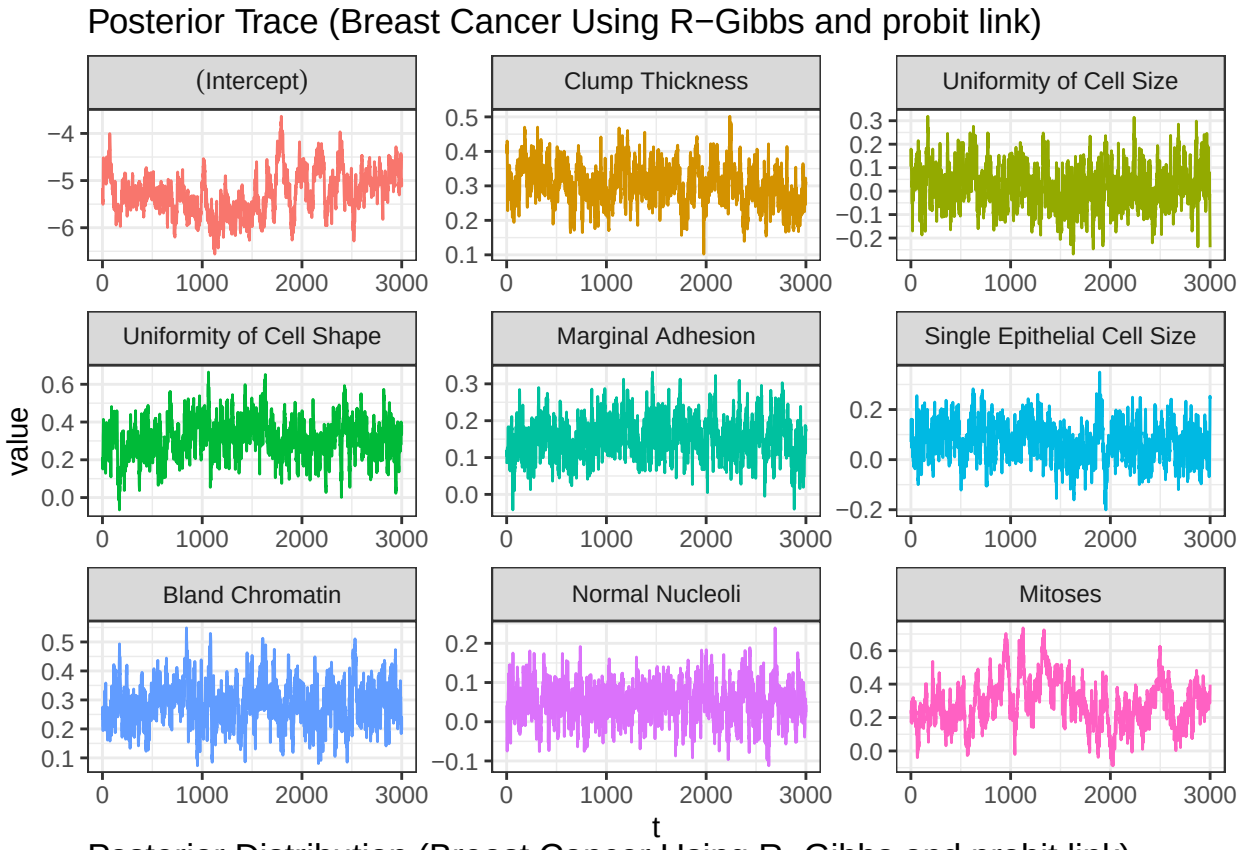
## 4.2 Breast Cancer

### 4.2.1 Probit Regression

We evaluate the probit regression methods with the Wisconsin Breast Cancer dataset, obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg (Wolberg and Mangasarian 1990). The data reports 9 discrete measurements for 699 observations of clumps of breast cancer cells as well as the response variable, indicating whether or not the clump is malignant (1) or benign (0). The variables are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. Only the Bare Nuclei variable included any missing data, so it was not included in this analysis. The model parameters are estimated with data-augmented Gibbs sampling following the algorithm described in Albert and Chib (1993) and implemented in R, a custom model built in Stan, and traditional maximum likelihood estimation as implemented in the `glm` function in R (R Core Team 2019). Both Bayesian methods are run to generate 4000 posterior samples, with the first 1000 discarded as burn-in samples, leaving 3000 samples for analysis. The Stan model took 6 minutes and 10.5 seconds seconds to run, while the R model completed in 7.8 seconds.

Table 4: Posterior sample summaries for breast cancer data using R-Gibbs and probit link

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| (Intercept) | -5.2657995 | 0.4516886 | -6.0869080 | -5.2781142 | -4.3524946 |
| `Clump Thickness` | 0.3030510 | 0.0570216 | 0.1891526 | 0.3017604 | 0.4131211 |
| `Uniformity of Cell Size` | 0.0229394 | 0.0891064 | -0.1458150 | 0.0219254 | 0.2035911 |
| `Uniformity of Cell Shape` | 0.3118907 | 0.1020941 | 0.1119921 | 0.3143989 | 0.5132866 |
| `Marginal Adhesion` | 0.1532512 | 0.0538371 | 0.0546572 | 0.1515937 | 0.2611788 |
| `Single Epithelial Cell Size` | 0.0733985 | 0.0731480 | -0.0708377 | 0.0732059 | 0.2141101 |
| `Bland Chromatin` | 0.2863431 | 0.0725573 | 0.1497924 | 0.2856310 | 0.4414197 |
| `Normal Nucleoli` | 0.0509178 | 0.0490190 | -0.0488792 | 0.0528914 | 0.1421364 |
| Mitoses | 0.2845107 | 0.1376854 | 0.0269171 | 0.2745250 | 0.5957231 |


Posterior Trace (Breast Cancer Using R−Gibbs and probit link)
(Intercept)
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bland Chromatin
Normal Nucleoli
Mitoses
value
t

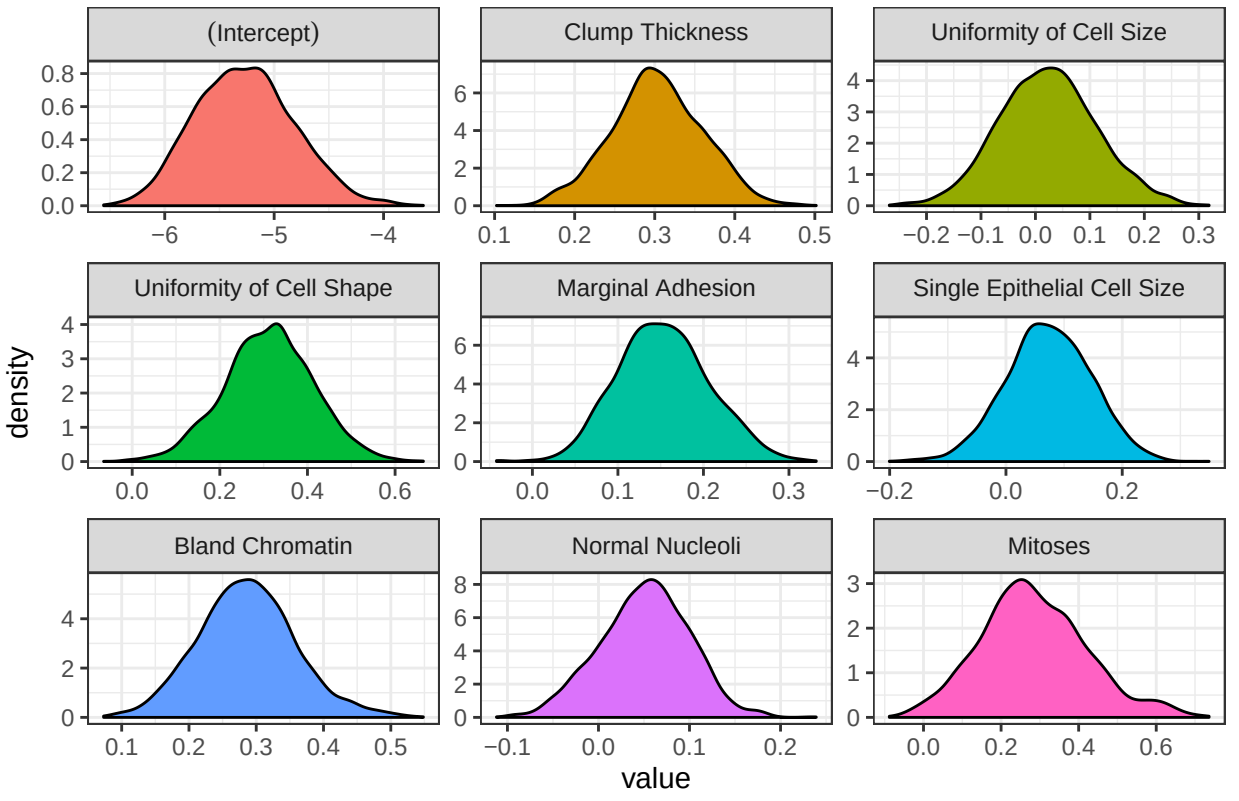
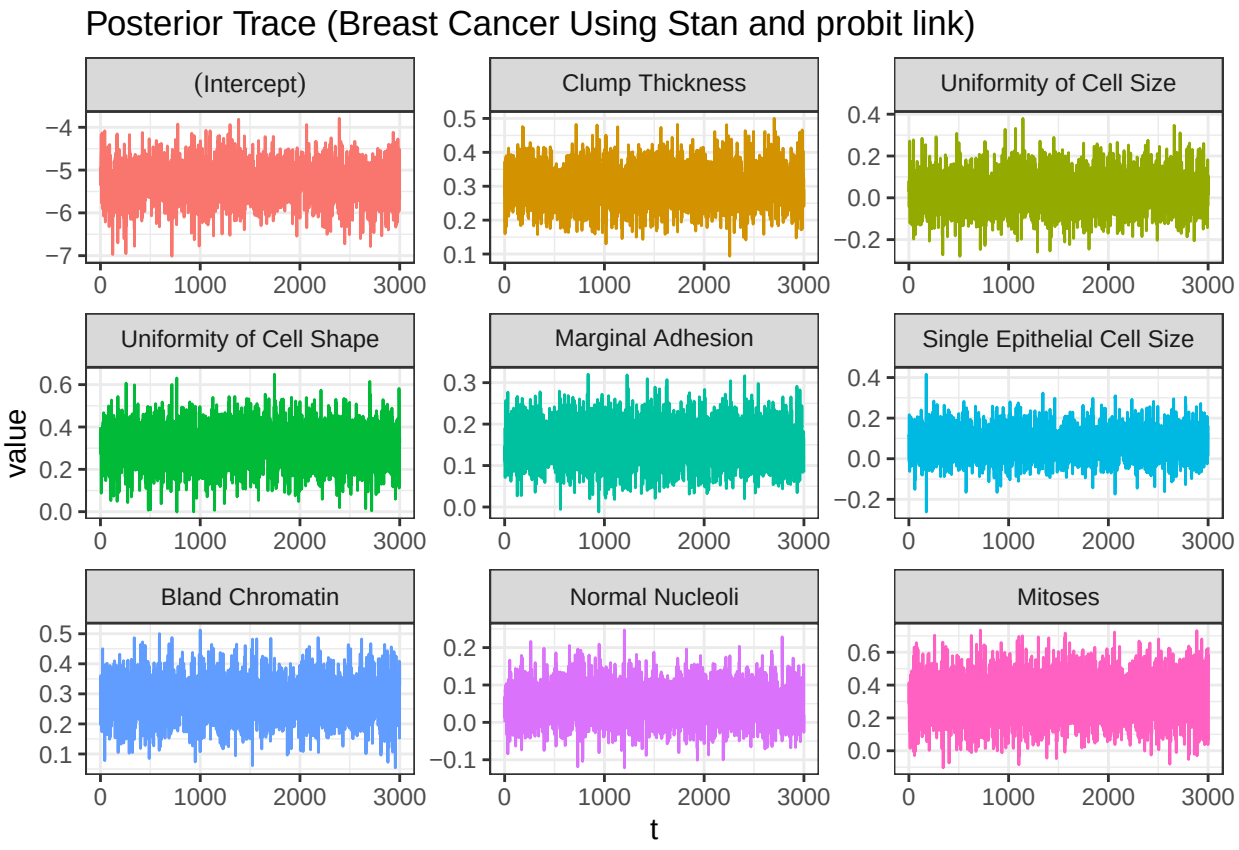

Posterior Distribution (Breast Cancer Using R−Gibbs and probit link)
(Intercept)
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bland Chromatin
Normal Nucleoli
Mitoses
density
value

Table 5: Posterior sample summaries for breast cancer data using Stan and probit link

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| (Intercept) | -5.2803740 | 0.4474952 | -6.1788551 | -5.2675580 | -4.4570687 |
| `Clump Thickness` | 0.3020233 | 0.0581248 | 0.1891183 | 0.3010757 | 0.4166279 |
| `Uniformity of Cell Size` | 0.0283373 | 0.0899331 | -0.1404464 | 0.0256264 | 0.2084892 |
| `Uniformity of Cell Shape` | 0.3066612 | 0.0950854 | 0.1178779 | 0.3088803 | 0.4907881 |
| `Marginal Adhesion` | 0.1525317 | 0.0502227 | 0.0576146 | 0.1520253 | 0.2512294 |
| `Single Epithelial Cell Size` | 0.0739874 | 0.0750008 | -0.0656497 | 0.0713772 | 0.2254342 |
| `Bland Chromatin` | 0.2812440 | 0.0686278 | 0.1482095 | 0.2824382 | 0.4200620 |
| `Normal Nucleoli` | 0.0526038 | 0.0495117 | -0.0427017 | 0.0537263 | 0.1511080 |
| Mitoses | 0.3191038 | 0.1369070 | 0.0561501 | 0.3168852 | 0.5875316 |



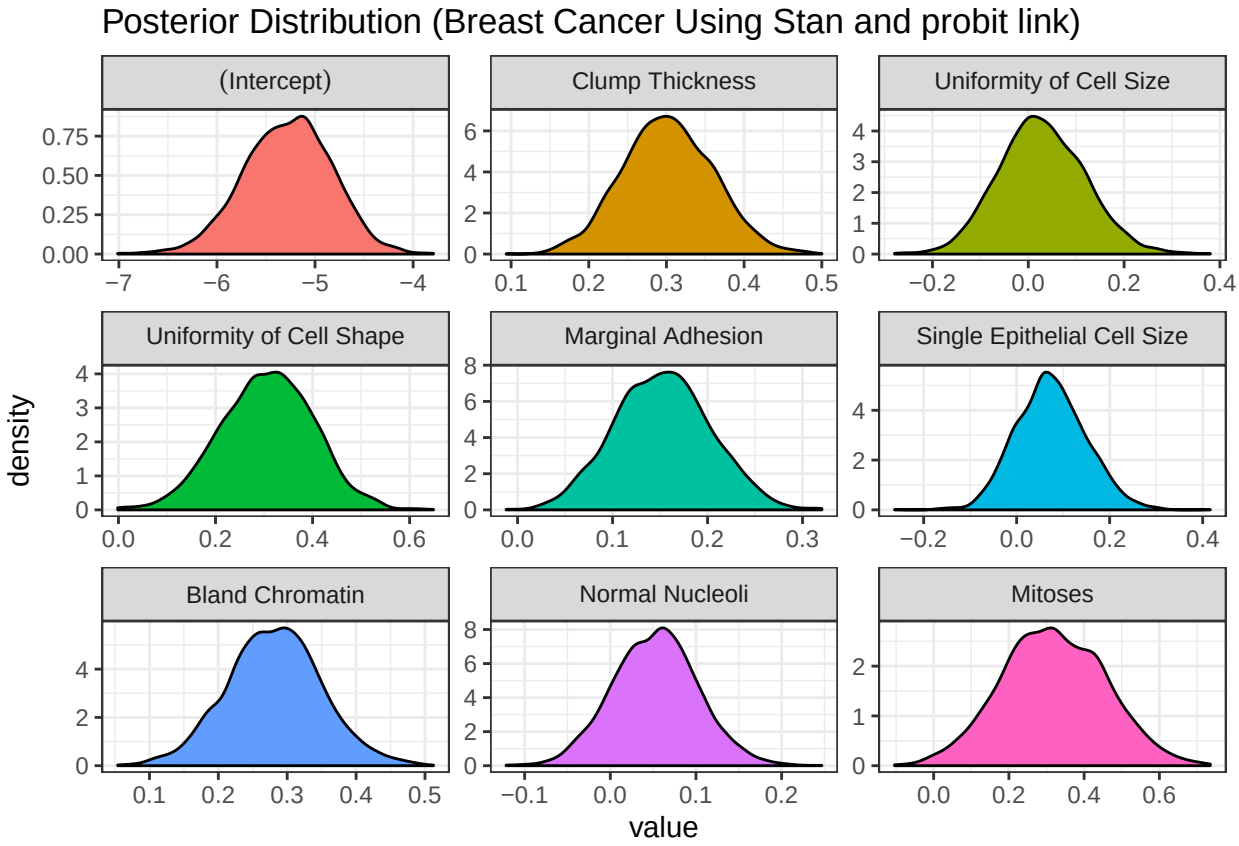



Table 6: MLE estimates for breast cancer data and probit link

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -5.1320418 | 0.4516719 |
| `Clump Thickness` | 0.2897958 | 0.0572147 |
| `Uniformity of Cell Size` | 0.0156388 | 0.0871703 |
| `Uniformity of Cell Shape` | 0.3037402 | 0.0931808 |
| `Marginal Adhesion` | 0.1493079 | 0.0518700 |
| `Single Epithelial Cell Size` | 0.0715684 | 0.0743288 |
| `Bland Chromatin` | 0.2758152 | 0.0715357 |
| `Normal Nucleoli` | 0.0538840 | 0.0509148 |
| Mitoses | 0.3109816 | 0.1430757 |

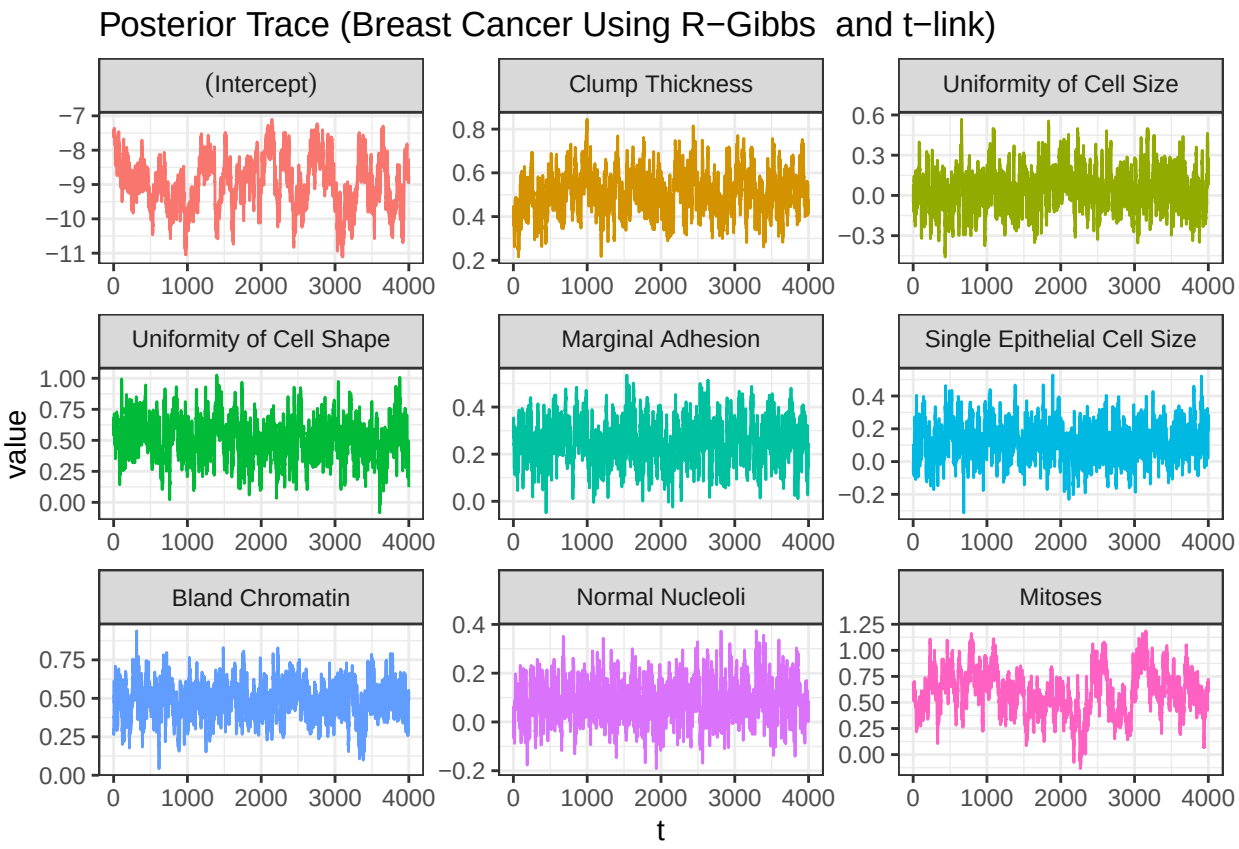
We note that the probit regression model for this data garnered very similar results from all three methods. While the Stan model mixed better than the Gibbs Sampler, the posterior distributions are very similar. Mixing issues are discussed in more detail in Section 5.

#### 4.2.2 T-Link/Logistic Regression

We now run the same analysis using the t-link with eight degrees of freedom, the logistic regression model run in Stan and traditional maximum likelihood estimation as implemented in the `glm` function in R (R Core Team 2019). Each sample from the t-link model is divided by 0.634 to make comparisons to the logistic models. Both Bayesian methods are run to generate 6000 posterior samples, with the first 2000 discarded as burn-in samples, leaving 4000 samples for analysis. The Stan model took 6 minutes and 55.7 seconds seconds to run, while the R model completed in 49.3 seconds.

Table 7: Posterior sample summaries for breast cancer data using R-Gibbs and t-link

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| (Intercept) | -8.9478410 | 0.7545748 | -10.4608408 | -8.9265365 | -7.5847849 |
| `Clump Thickness` | 0.5112914 | 0.0962183 | 0.3326653 | 0.5073785 | 0.7061969 |
| `Uniformity of Cell Size` | 0.0451144 | 0.1530782 | -0.2427769 | 0.0453854 | 0.3636352 |
| `Uniformity of Cell Shape` | 0.5096562 | 0.1626622 | 0.1892346 | 0.5116899 | 0.8363108 |
| `Marginal Adhesion` | 0.2504394 | 0.0897844 | 0.0732464 | 0.2496116 | 0.4260844 |
| `Single Epithelial Cell Size` | 0.1220531 | 0.1107471 | -0.0975137 | 0.1228943 | 0.3379523 |
| `Bland Chromatin` | 0.4868111 | 0.1145345 | 0.2631670 | 0.4866126 | 0.7138250 |
| `Normal Nucleoli` | 0.0879714 | 0.0824531 | -0.0717133 | 0.0847024 | 0.2569476 |
| Mitoses | 0.5954966 | 0.2187600 | 0.1763032 | 0.5951070 | 1.0145938 |



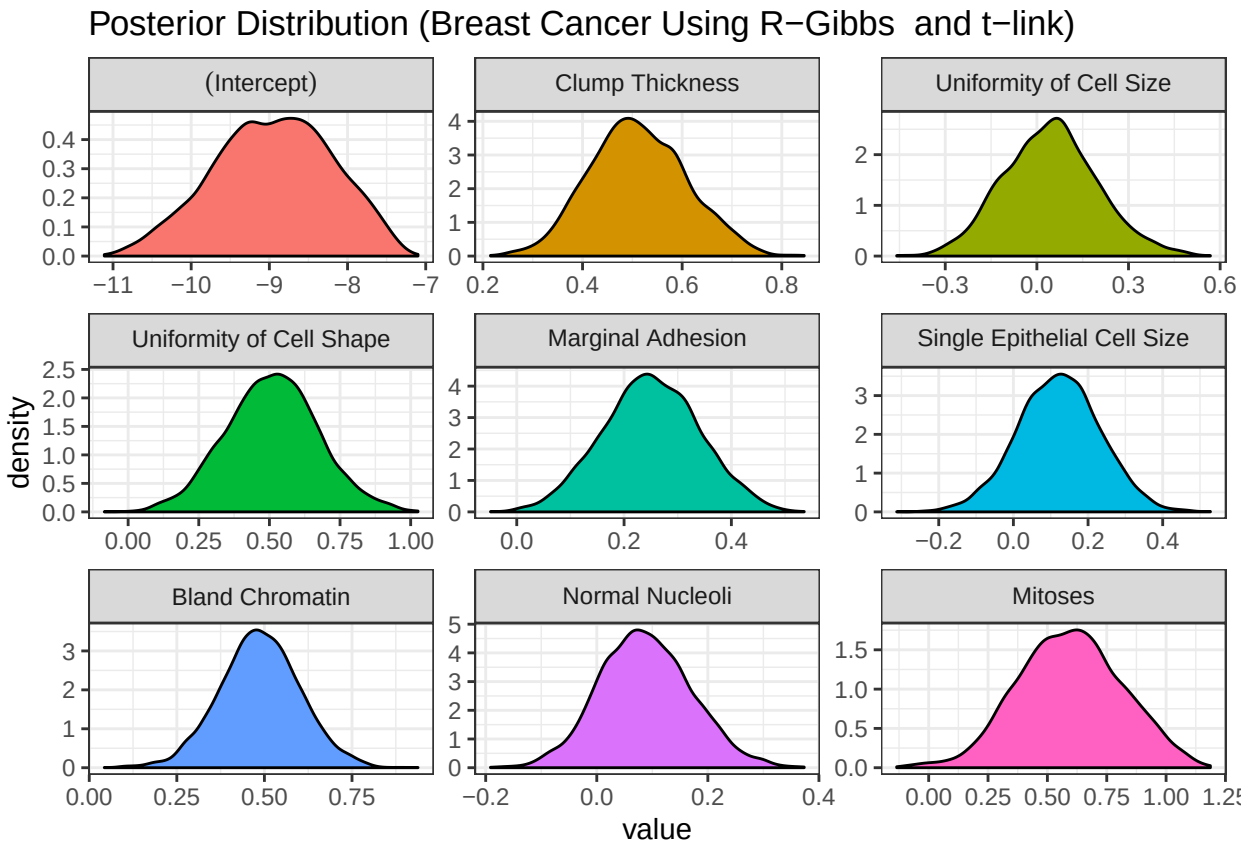
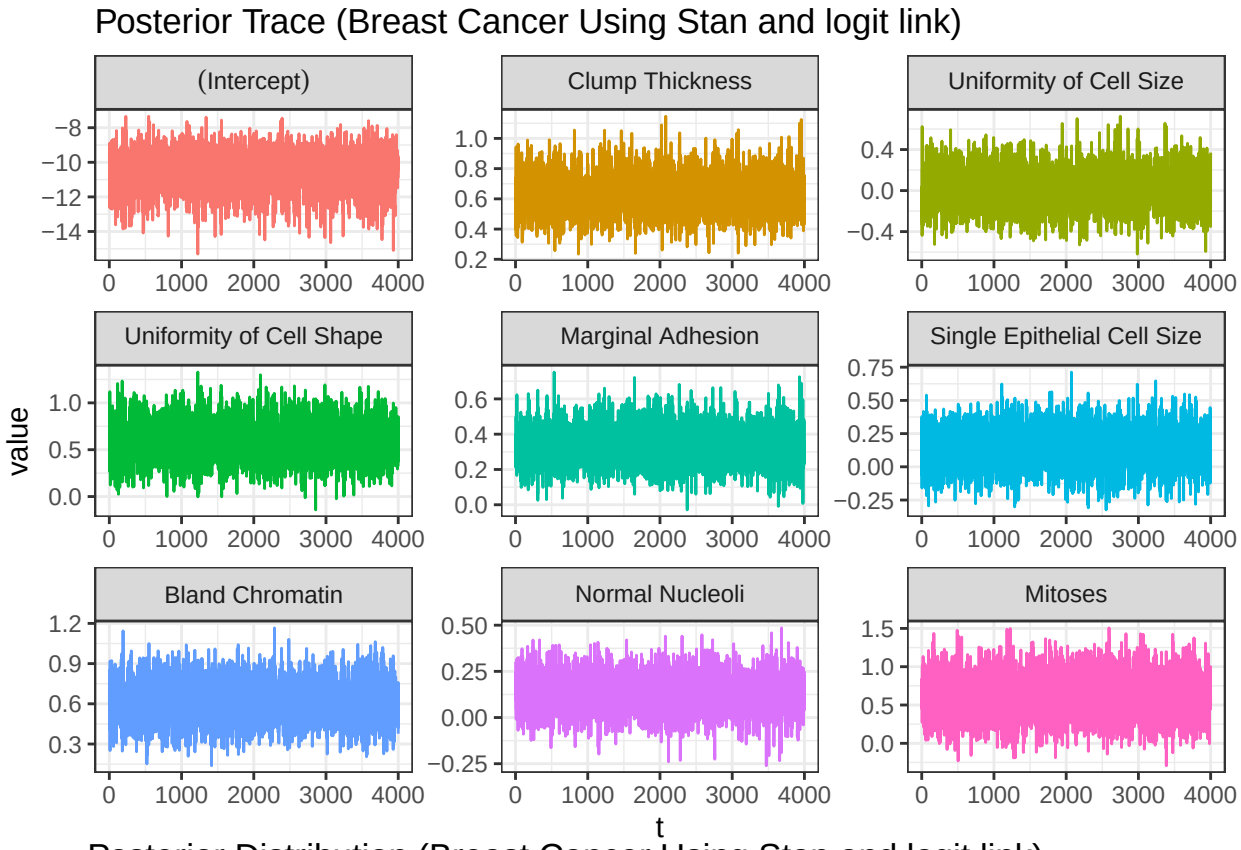



Table 8: Posterior sample summaries for breast cancer data using Stan and logit link

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| (Intercept) | -10.5446308 | 1.1086470 | -13.0068829 | -10.4491418 | -8.5919845 |
| `Clump Thickness` | 0.6262489 | 0.1248959 | 0.3989534 | 0.6201545 | 0.8904694 |
| `Uniformity of Cell Size` | 0.0104245 | 0.1866697 | -0.3463081 | 0.0078354 | 0.3948086 |
| `Uniformity of Cell Shape` | 0.5884213 | 0.2045167 | 0.1923511 | 0.5906412 | 0.9963114 |
| `Marginal Adhesion` | 0.3360092 | 0.1053344 | 0.1356647 | 0.3328388 | 0.5502999 |
| `Single Epithelial Cell Size` | 0.1349902 | 0.1478499 | -0.1585615 | 0.1298016 | 0.4280293 |
| `Bland Chromatin` | 0.6176422 | 0.1484599 | 0.3340841 | 0.6147750 | 0.9164285 |
| `Normal Nucleoli` | 0.1297881 | 0.0998482 | -0.0650426 | 0.1290370 | 0.3265395 |
| Mitoses | 0.6114814 | 0.2811762 | 0.0806540 | 0.6095192 | 1.1638068 |


Posterior Trace (Breast Cancer Using Stan and logit link)
(Intercept)
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bland Chromatin
Normal Nucleoli
Mitoses
value
t

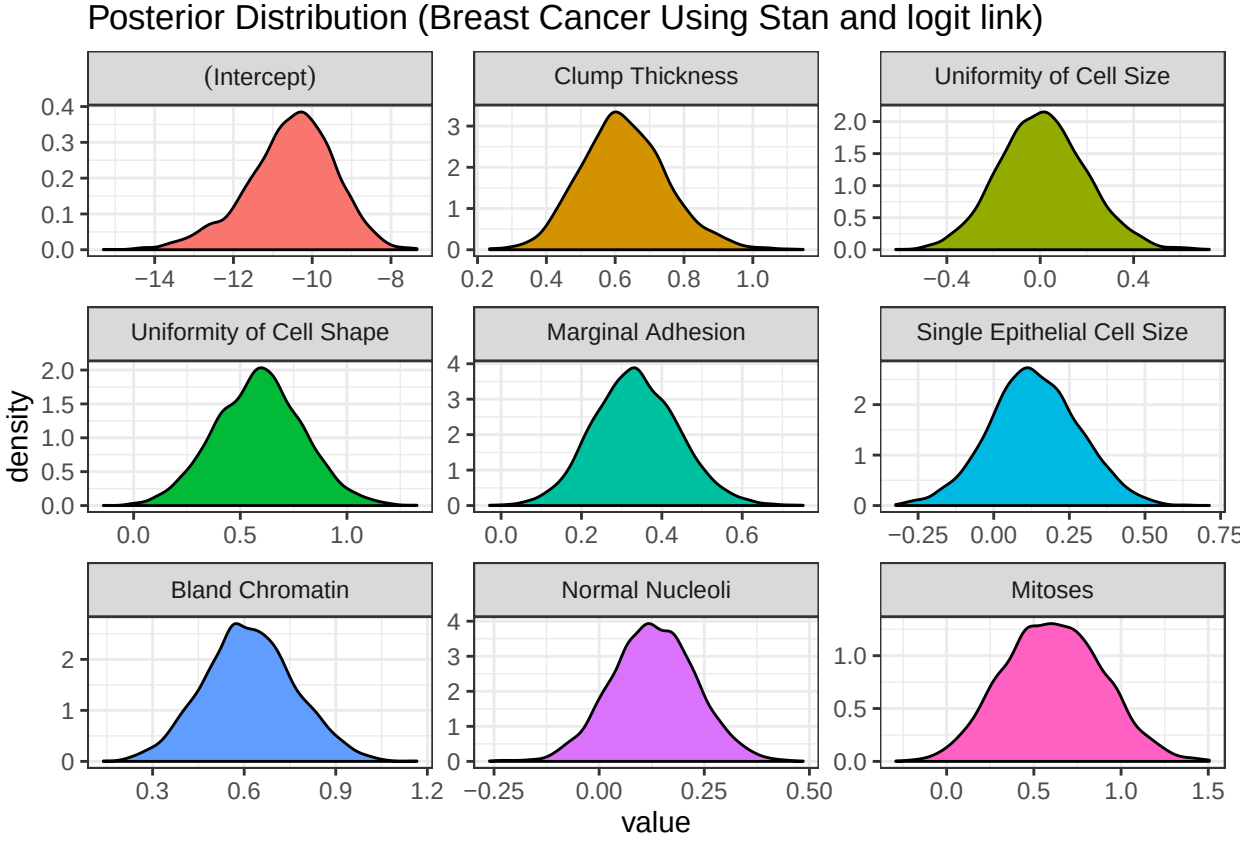

Posterior Distribution (Breast Cancer Using Stan and logit link)
(Intercept)
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bland Chromatin
Normal Nucleoli
Mitoses
density
value

Table 9: MLE estimates for breast cancer data and logit link

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | -9.9456360 | 1.0323642 |
| `Clump Thickness` | 0.5775660 | 0.1190322 |
| `Uniformity of Cell Size` | -0.0115529 | 0.1759137 |
| `Uniformity of Cell Shape` | 0.5679361 | 0.1912659 |
| `Marginal Adhesion` | 0.3136807 | 0.1003632 |
| `Single Epithelial Cell Size` | 0.1305624 | 0.1405972 |
| `Bland Chromatin` | 0.5799514 | 0.1455813 |
| `Normal Nucleoli` | 0.1231927 | 0.0986913 |
| Mitoses | 0.6078538 | 0.3241578 |

We note that the transformed t-link point estimates closely match the Stan model coefficients as well as the maximum likelihood estimates. The Gibbs sample trace plot exhibit strong auto-correlation in some plots, but the posterior distributions look fairly normal. There are no obvious problems with the posterior distributions generated with the Stan model.

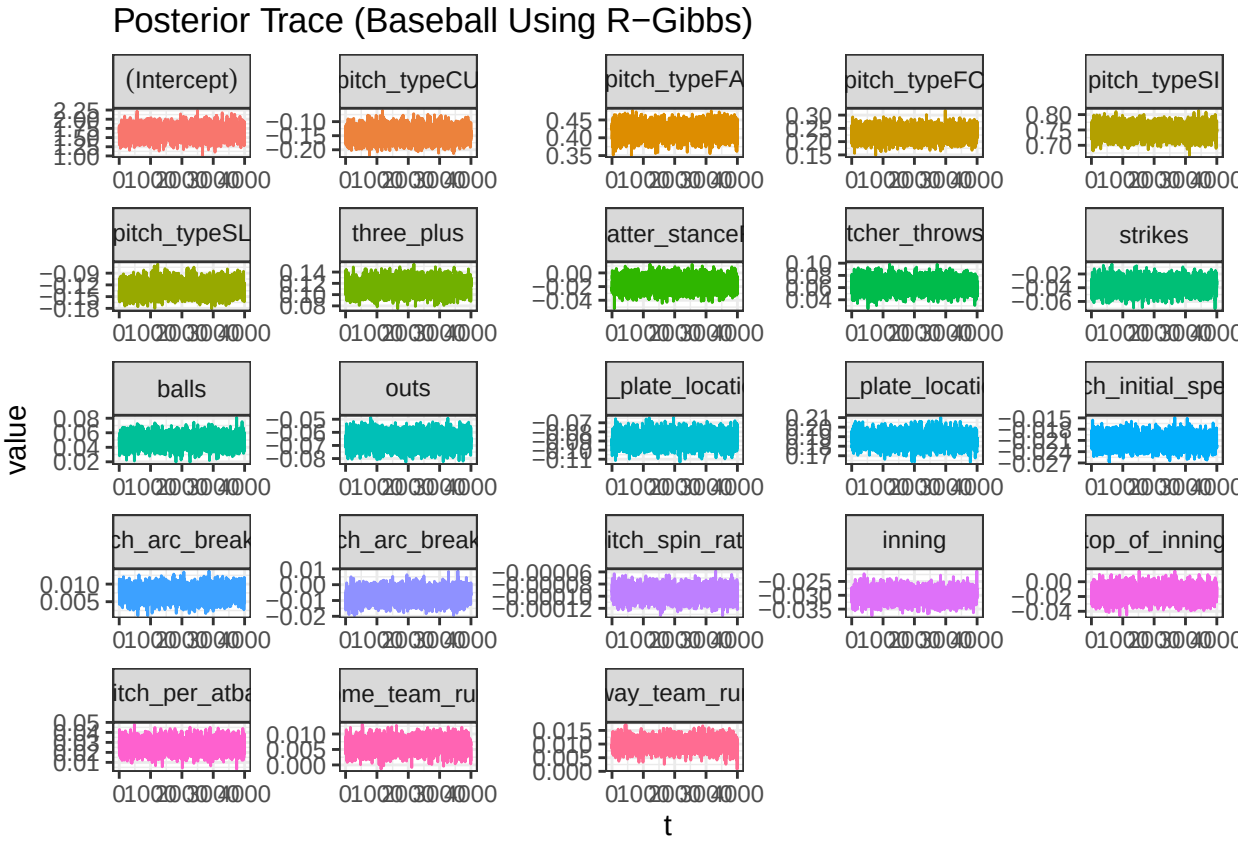
### 4.3 Baseball

We evaluate the performance of data-augmented Gibbs sampling on a larger datasets (more predictors and observations than the breast cancer data), by fitting probit regression models on baseball data. The baseball data comes from Major League Baseball pitch tracking software and was prepared by the Los Angeles Dodgers for the purpose of predicting whether or not a pitch was put into play. There are 100,000 pitches (observations) thrown by Dodger's starting pitchers and 19 features of each pitch, including velocity, spin rate, and location. The binary response indicates whether or not a pitch was put into play (0: not put into play, 1: put into play). Missing data was removed, leaving 99,254 observations with 18 covariates for analysis. The model parameters are estimated with data augmented Gibbs sampling following the algorithm described in Albert and Chib (1993) and implemented in R and traditional maximum likelihood estimation as implemented in the `glm` function in R (R Core Team 2019). Both Bayesian methods are run to generate 5000 posterior samples, with the first 1000 discarded as burn-in samples, leaving 4000 samples for analysis. We also attempted to fit a custom model in Stan but the samples were generated very slowly (fewer than 500 samples per hour). In contrast, the R model generated all 5000 samples in 15 minutes and 41 seconds.

Table 10: Posterior sample summaries for baseball data using R-Gibbs

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| (Intercept) | 1.6419615 | 0.1641598 | 1.3275671 | 1.6432777 | 1.9577907 |
| pitch_typeCU | -0.1427578 | 0.0223181 | -0.1875506 | -0.1427039 | -0.0990722 |
| pitch_typeFA | 0.4187926 | 0.0181987 | 0.3820077 | 0.4188624 | 0.4544004 |
| pitch_typeFC | 0.2290787 | 0.0213407 | 0.1871727 | 0.2289236 | 0.2717159 |
| pitch_typeSI | 0.7452006 | 0.0191260 | 0.7073330 | 0.7452225 | 0.7822887 |
| pitch_typeSL | -0.1282329 | 0.0162129 | -0.1598217 | -0.1282685 | -0.0963311 |
| three_plus | 0.1138644 | 0.0114908 | 0.0918353 | 0.1136950 | 0.1355962 |
| batter_stanceR | -0.0157872 | 0.0088088 | -0.0331719 | -0.0157398 | 0.0011848 |
| pitcher_throwsR | 0.0640037 | 0.0100326 | 0.0445198 | 0.0640910 | 0.0836733 |
| strikes | -0.0369098 | 0.0091613 | -0.0552714 | -0.0368880 | -0.0193004 |
| balls | 0.0487882 | 0.0079058 | 0.0331798 | 0.0489391 | 0.0639353 |
| outs | -0.0660833 | 0.0050520 | -0.0760450 | -0.0661052 | -0.0563236 |
| pitch_plate_location_x | -0.0880404 | 0.0067681 | -0.1014258 | -0.0879550 | -0.0749833 |
| pitch_plate_location_z | 0.1874257 | 0.0063152 | 0.1750398 | 0.1874246 | 0.1995573 |

| Variable | mean | sd | 2.5% | 50% | 97.5% |
|---|---|---|---|---|---|
| pitch_initial_speed | -0.0211755 | 0.0016717 | -0.0244022 | -0.0212001 | -0.0179145 |
| pitch_arc_break_x | 0.0073551 | 0.0017686 | 0.0038227 | 0.0073578 | 0.0107850 |
| pitch_arc_break_z | -0.0068673 | 0.0038864 | -0.0145743 | -0.0068384 | 0.0005285 |
| pitch_spin_rate | -0.0000956 | 0.0000100 | -0.0001149 | -0.0000956 | -0.0000758 |
| inning | -0.0301068 | 0.0020301 | -0.0340363 | -0.0301332 | -0.0260424 |
| top_of_inning | -0.0157989 | 0.0084155 | -0.0318350 | -0.0159557 | 0.0010927 |
| pitch_per_atbat | 0.0263585 | 0.0060101 | 0.0148513 | 0.0263168 | 0.0383177 |
| home_team_runs | 0.0062439 | 0.0020007 | 0.0022450 | 0.0062631 | 0.0100559 |
| away_team_runs | 0.0099700 | 0.0019942 | 0.0061492 | 0.0099993 | 0.0137278 |



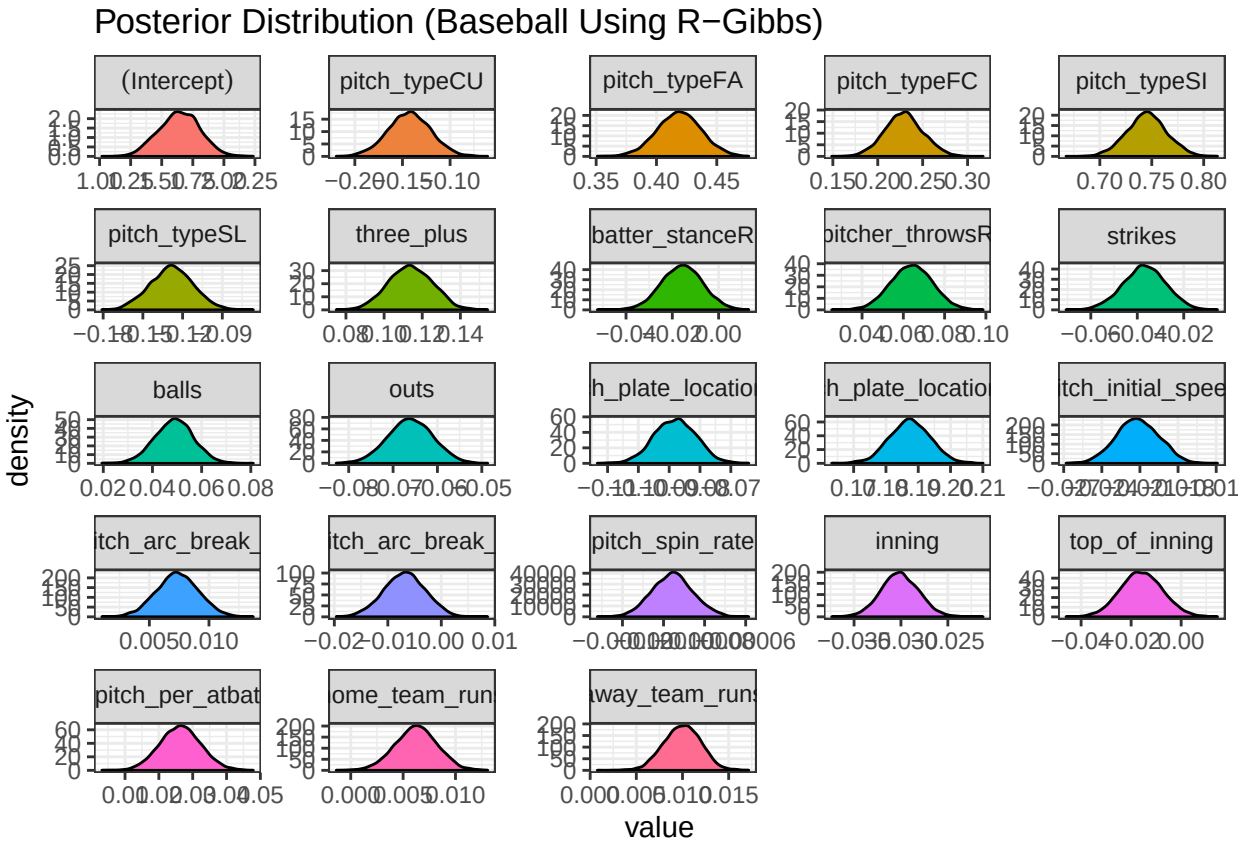



Table 11: MLE estimates for baseball data

| Variable | Estimate | Std. Error |
|---|---|---|
| (Intercept) | 1.6457301 | 0.1643384 |
| pitch_typeCU | -0.1425477 | 0.0226587 |
| pitch_typeFA | 0.4184923 | 0.0185715 |
| pitch_typeFC | 0.2282183 | 0.0219511 |
| pitch_typeSI | 0.7453712 | 0.0188182 |
| pitch_typeSL | -0.1279136 | 0.0167241 |
| three_plus | 0.1136658 | 0.0114023 |
| batter_stanceR | -0.0156687 | 0.0088127 |
| pitcher_throwsR | 0.0637459 | 0.0099670 |
| strikes | -0.0369765 | 0.0102701 |
| balls | 0.0489444 | 0.0088985 |
| outs | -0.0661775 | 0.0050947 |
| pitch_plate_location_x | -0.0881536 | 0.0068380 |
| pitch_plate_location_z | 0.1874354 | 0.0063953 |
| pitch_initial_speed | -0.0212061 | 0.0016916 |
| pitch_arc_break_x | 0.0073317 | 0.0017433 |
| pitch_arc_break_z | -0.0067832 | 0.0038488 |
| pitch_spin_rate | -0.0000956 | 0.0000100 |
| inning | -0.0300979 | 0.0020700 |
| top_of_inning | -0.0160490 | 0.0083282 |
| pitch_per_atbat | 0.0262550 | 0.0070157 |
| home_team_runs | 0.0062091 | 0.0019981 |
| away_team_runs | 0.0100125 | 0.0020411 |

We note that the MLE results and the results from the data-augmented Gibbs Sampler is quite similar when comparing the mean and standard deviation estimates of the two methods. We can see from the trace plots that the Gibbs Sampler mixed well.

# 5 Discussion

The data-augmented Gibbs sampling algorithms applied in this project largely performed well and achieved comparable estimates to the Stan models, while running in significantly less time. The speed improvements were particularly impressive in the case of the baseball data, which enabled us to perform a Bayesian analysis when we would have otherwise not have had sufficient time. This model also mixed quite well, so inference should be trusted with the Gibbs Sampler. The chief shortcoming of these methods is the poor mixing properties we observed with the Small-Cell Carcinoma data and the Breast Cancer data (although not as severe). Simulating much greater samples and using thinning did not lead to any improvement in mixing. Despite the poor mixing of the chains, the posterior means and medians closely aligned with the results observed in the Stan model and obtained with maximum likelihood estimation.

In addition the use of flat, improper priors may be of concern to some practitioners. The data-augmented Gibbs Sampler can accommodate the use of specific priors, but this can make the sampling more difficult. This was not a concern in our analysis, given that we are not subject matter experts in any of the applications we demonstrated here.

Although the models described in Albert and Chib (1993) and demonstrated in this project remain popular to this day (Chakraborty and Khare 2017), much work has been done to expand upon its foundation. Several papers have made efforts to apply the strategy described here to the logit model (Holmes and Held 2006; Frühwirth-Schnatter and Frühwirth 2010; Gramacy and Polson 2012; Polson, Scott, and Windle 2013) as wells as negative binomial regression (Pillow and Scott 2012).

Current research is being done to analyze baseball data as it pertains to pitcher changes and in-game decision making (Kwon, Katz, and Stern 2019). A hierarchical probit model is being used to estimate spline functions of how pitchers "age" while playing in a Major League Baseball game. This model was adapted from Berry, Reese, and Larkey (1999), which looked at players actually aging over their careers, but uses a probit link instead of a logit link function. The probit link is being used in order to use data-augmented Gibbs Sampling for the sampling of the posterior distributions and for parameter estimation on the spline functions.

# 6 References


Aitchison, John, and Samuel D Silvey. 1957. "The Generalization of Probit Analysis to the Case of Multiple Responses." *Biometrika* 44 (1/2). JSTOR: 131–40.

Albert, James H., and Siddhartha Chib. 1993. "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association* 88 (422). Taylor & Francis: 669–79. https://doi.org/10.1080/01621459.1993.10476321.

Berry, Scott M., C. Shane Reese, and Patrick D. Larkey. 1999. "Bridging Different Eras in Sports." *Journal of the American Statistical Association* 94 (447). [American Statistical Association, Taylor & Francis, Ltd.]: 661–76.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software, Articles* 76 (1): 1–32. https://doi.org/10.18637/jss.v076.i01.

Chakraborty, Saptarshi, and Kshitij Khare. 2017. "Convergence Properties of Gibbs Samplers for Bayesian Probit Regression with Proper Priors." *Electron. J. Statist.* 11 (1). The Institute of Mathematical Statistics; the Bernoulli Society: 177–210. https://doi.org/10.1214/16-EJS1219.

Frühwirth-Schnatter, Sylvia, and Rudolf Frühwirth. 2010. "Data Augmentation and Mcmc for Binary and Multinomial Logit Models." In *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, edited by Thomas Kneib and Gerhard Tutz, 111–32. Heidelberg: Physica-Verlag HD. https://doi.org/10.1007/978-3-7908-2413-1_7.

Gelfand, Alan E, and Adrian FM Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85 (410). Taylor & Francis Group: 398–409.

Gramacy, Robert B., and Nicholas G. Polson. 2012. "Simulation-Based Regularized Logistic Regression." *Bayesian Anal.* 7 (3). International Society for Bayesian Analysis: 567–90. https://doi.org/10.1214/12-BA719.

Gurland, John, Ilbok Lee, and Paul A Dahm. 1960. "Polychotomous Quantal Response in Biological Assay." *Biometrics* 16 (3). JSTOR: 382–98.

Holmes, Chris C., and Leonhard Held. 2006. "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression." *Bayesian Anal.* 1 (1). International Society for Bayesian Analysis: 145–68. https://doi.org/10.1214/06-BA105.

Kwon, Jimmy, Corey Katz, and Hal Stern. 2019. "A Statistical Approach to Pitching Change Decisions."

McCullagh, Peter. 1980. "Regression Models for Ordinal Data." *Journal of the Royal Statistical Society. Series B (Methodological)* 42 (2). [Royal Statistical Society, Wiley]: 109–42.

Pillow, Jonathan W., and James Scott. 2012. "Fully Bayesian Inference for Neural Models with Negative-Binomial Spiking." In *Advances in Neural Information Processing Systems 25*, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1898–1906. Curran Associates, Inc. http://papers.nips.cc/paper/4567-fully-bayesian-inference-for-neural-models-with-negative-binomial-spiking.pdf.

Polson, Nicholas G., James G. Scott, and Jesse Windle. 2013. "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables." *Journal of the American Statistical Association* 108 (504). Taylor & Francis: 1339–49. https://doi.org/10.1080/01621459.2013.829001.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stan Development Team. 2019. "RStan: The R Interface to Stan." http://mc-stan.org/.

Tanner, Martin A., and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (398). [American Statistical Association, Taylor & Francis, Ltd.]: 528–40.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. Fourth. New York: Springer. http://www.stats.ox.ac.uk/pub/MASS4.

Wolberg, William H, and Olvi L Mangasarian. 1990. "Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology." *Proceedings of the National Academy of Sciences* 87 (23). National Academy of Sciences: 9193–6. https://doi.org/10.1073/pnas.87.23.9193.

# 7 Appendix

All code for this analysis is available in our GitHub repository: https://github.com/damonbayer/230-Final