Polya-Gamma Distribution and Data Augmentation

Toni Luhdo

March 4, 2025

Introduction

- Logistic regression is a fundamental model for classification problems.
- Bayesian inference requires efficient sampling methods.
- The Polya-Gamma distribution enables an elegant solution to the Gibbs sampling strategy.

Polya-Gamma Distribution

Definition: A random variable ω is called Polya-Gamma distributed with parameters b and c if it has the following density function:

$$f(x \mid b, c) = \{\cosh^b(c/2)\} \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n+b)(2n+b)}{\Gamma(n+1)\sqrt{2\pi x^3}} e^{-\frac{(2n+b)^2}{8x} - \frac{c^2}{2}x} \quad (1)$$

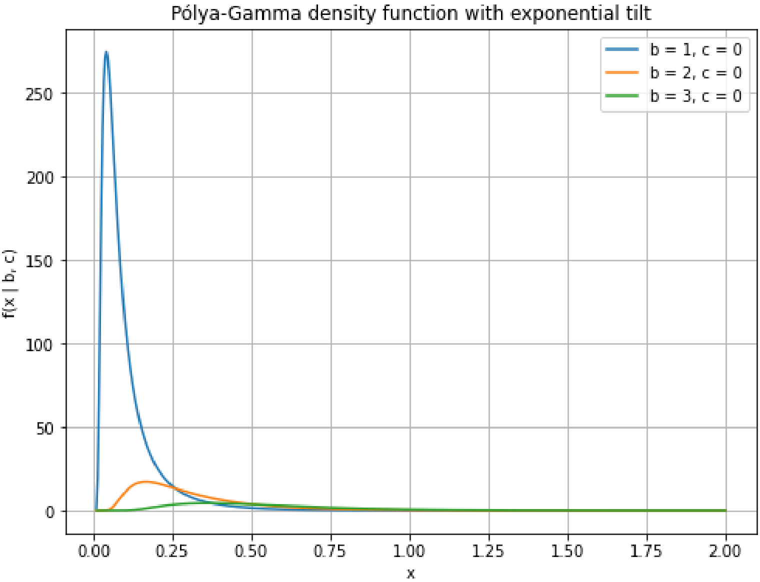
for $x > 0$.

Remark If c $=$ 0 we get the so called standard Polya-Gamma density:

$$f(x \mid b, 0) = \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n+b)(2n+b)}{\Gamma(n+1)\sqrt{2\pi x^3}} e^{-\frac{(2n+b)^2}{8x}} \quad (2)$$

Polya-Gamma Distribution

- $b \in \mathbb{R}^+, \quad b > 0$ Controls the shape of the distribution (shape parameter).
- $c \in \mathbb{R}$ Influences the scaling/tilt (scale parameter).
- $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt, \quad \text{for } x > 0$


Pólya-Gamma density function with exponential tilt
b = 1, c = 0
b = 2, c = 0
b = 3, c = 0
f(x | b, c)
0
50
100
150
200
250
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
x

Laplace Transformation Properties

A different representation of the polya gamma distribution is given by

$$p(\omega|b,c) = \frac{\exp(-c^2\omega/2)p(\omega|b,0)}{E_{PG(b,0)}[\exp(-c^2\omega/2)]} \tag{3}$$

The Laplace transform of the Polya-Gamma distribution is given by:

$$E[\exp(-\omega t)] = \frac{\cosh^b(c/2)}{\cosh^b(\sqrt{\frac{c^2/2+t}{2}})} \tag{4}$$

- Useful for analytical computations and simulations.
- Basis for MCMC methodology in Bayesian inference.

Alternative Representation of the Polya-Gamma Distribution

- The distribution can be represented as an infinite sum of gamma-distributed variables:

$$\omega \sim \sum_{k=1}^{\infty} \frac{G_k}{(k-1/2)^2 + c^2/(4\pi^2)}, \quad G_k \sim \text{Gamma}(b, 1) \quad (5)$$

- This representation highlights the relationship to the gamma distribution and offers alternative computational approaches.

Properties of the Polya-Gamma Class

- Closed under convolution: If $\omega_1 \sim PG(b_1, z)$ and $\omega_2 \sim PG(b_2, z)$ are independent, then:

$$\omega_1 + \omega_2 \sim PG(b_1 + b_2, z) \tag{6}$$

- Expectation:

$$E(\omega) = \frac{b}{2c} \tanh(c/2) \tag{7}$$

Bayesian Inference for Logistic Models

The likelihood function of a logistic model is:

$$P(y_i = 1 | x_i, \beta) = \frac{1}{1 + e^{-x_i^T \beta}} \tag{8}$$

- Introducing a latent Polya-Gamma parameter ω_i enables a Gibbs sampling method.
- The posterior distribution is normal for β and Polya-Gamma for ω.

Transformation of the Likelihood with Polya-Gamma

The likelihood function of a logistic model is originally given by:

$$P(y_i = 1|x_i, \beta) = \frac{1}{1 + e^{-x_i^T \beta}} \tag{9}$$

Using Polya-Gamma data augmentation, we utilize the identity:

$$\frac{e^{y_i x_i^T \beta}}{1 + e^{x_i^T \beta}} = \int_0^\infty e^{-\omega_i (x_i^T \beta)^2/2} p(\omega_i|1, 0) d\omega_i \tag{10}$$

This allows the likelihood to be rewritten in a form that results in a conditional normal distribution for β, enabling Gibbs sampling.

Gibbs Sampler for the Logistic Model

1. Sample $\omega_i|\beta \sim PG(n_i, x_i^T\beta)$ for all i.
2. Sample $\beta|z,\omega \sim N(m_\omega, V_\omega)$, where:

$$m_\omega = V_\omega X^T\kappa, \quad V_\omega = (X^T\Omega X + B^{-1})^{-1} \tag{11}$$

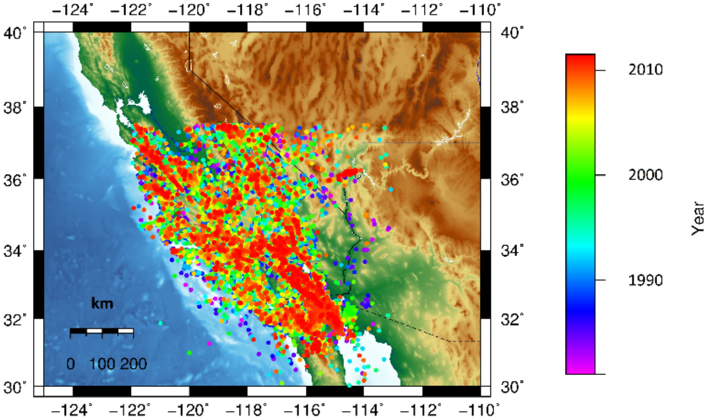
with $\kappa = (y_1 - n_1/2, ..., y_N - n_N/2)$, and Ω being the diagonal matrix of ω_i.
B is the prior covariance matrix for the coefficients β.

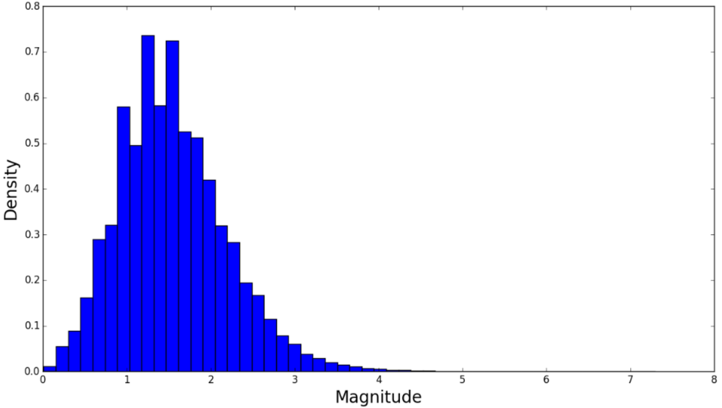
The Gutenberg-Richter Law for Earthquakes

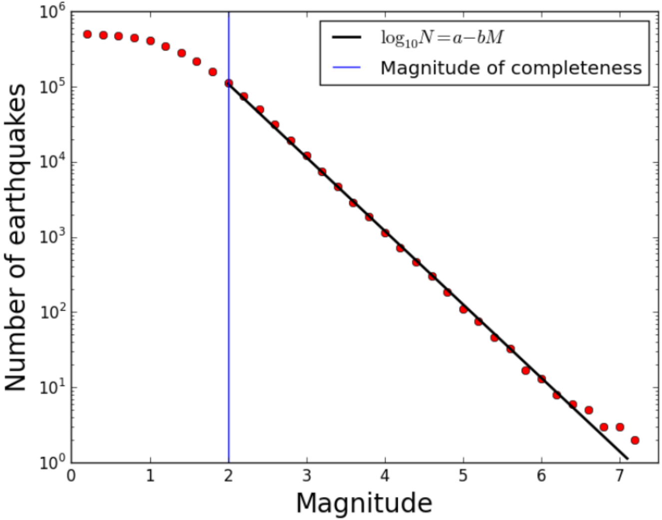
- An empirical law describing the relationship between earthquake magnitude and frequency.
- Proposed by Charles Francis Richter and Beno Gutenberg (1956).
- Gutenberg-Richter Law:

$$\log_{10}(N) = a - bM$$

- N is the cumulative annual number of earthquakes with a magnitude greater than M.
- a represents the total seismicity rate of the observed region.
- b expresses the relative proportion of small to large earthquakes.


−124˚ −122˚ −120˚ −118˚ −116˚ −114˚ −112˚ −110˚
40˚
38˚
36˚
34˚
32˚
30˚
km
0 100 200
2010
2000
1990
Year


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Density
0
1
2
3
4
5
6
7
8
Magnitude



10^6
10^5
10^4
10^3
10^2
10^1
10^0
0
1
2
3
4
5
6
7
$\log_{10} N = a - bM$
Magnitude of completeness
Number of earthquakes
Magnitude

Estimating b with Polya-Gamma Data Augmentation

- The Bayesian estimation of the parameter b can be facilitated using Polya-Gamma data augmentation.
- By transforming the Gutenberg-Richter equation into a logistic model:

$$P(y_i = 1 | m_i, b) = \frac{1}{1 + e^{-m_i b}} \tag{12}$$

  the Polya-Gamma method can be used to develop a Gibbs sampling strategy for determining b.
- Advantages of this method:
  - It enables a robust estimation of b, even with small or noisy data.
  - The Bayesian approach allows for direct quantification of uncertainties in the estimation.
  - It provides a natural way to incorporate additional prior information about b.

Why Use Polya-Gamma for Estimation?

- Modeling earthquake frequency as a Poisson process:

$$N(M) \sim \text{Poisson}(\lambda(M))$$

- Log transformation leads to a linear model:

$$\log N(M) = a - bM$$

- The Polya-Gamma method enables Gibbs sampling for estimation.

Summary

- The Polya-Gamma distribution significantly simplifies Bayesian inference for logistic models.
- The Gibbs sampling strategy allows for efficient numerical computations.
- The method is well-suited for large-scale data analysis and machine learning.

References

Polson, N. G., Scott, J. G., Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504), 1339-1349.

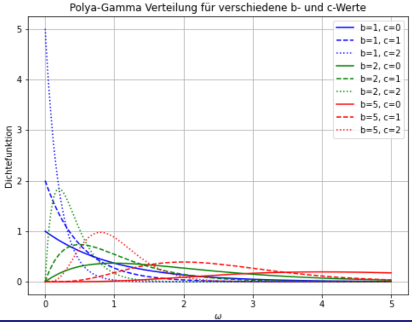
Thanks for your Attention!

Appendix: Meaning of the Parameters b and c in the Polya-Gamma Distribution

- **Parameter b (Shape Parameter):**
  - Determines the scaling of the distribution.
  - Higher values make the distribution narrower and more concentrated.
  - Often derived from the number of trials in the model.
- **Parameter c (Scale Parameter):**
  - Controls the exponential damping.
  - Higher values of c shift the distribution to the left.
  - In logistic models, $c = x^T \beta$.

Appendix: Effects of b and c on the Distribution

- If $b = 1$ and $c = 0$: The distribution resembles a Gamma distribution.
- If c increases: The distribution shifts to the left.
- If b increases: The distribution becomes narrower and more concentrated.



Appendix: Why is $p(\omega_i|1,0)$ in the Likelihood Transformation?

- In logistic regression, the likelihood is given by:

$$P(y_i = 1|x_i, \beta) = \frac{1}{1 + e^{-x_i^T \beta}} \tag{13}$$

- Using the Polya-Gamma transformation, this is rewritten as:

$$P(y_i = 1|x_i, \beta) = \int_0^\infty e^{-\omega_i (x_i^T \beta)^2/2} p(\omega_i|1,0) d\omega_i \tag{14}$$

- **Why $PG(1,0)$?**
  - $b = 1$ because it is a binary logistic regression (Bernoulli model).
  - $c = 0$ because no exponential damping is required.