University of Milan

Faculty of Political, Economic and Social Sciences

A Bayesian Approach to Aggregate Insurance Claim Modeling

Final Project in the Subject Bayesian Analysis

Julia Maria Wdowinska (43288A)
Edoardo Zanone (33927A)

Data Science for Economics
II Year
Master's Degree

April 25, 2025

Contents

1 Introduction

2

The first objective of this project was to replicate the analysis conducted by Dudley. The dataset used comprises insurance claim amounts exceeding 1.5 million over a period of five years from an automobile insurance portfolio. The data, originally presented in Rytgaard (1990), is shown in Table 1.

Table 1: Insurance Claim Amounts Exceeding 1.5 Million (Data from Rytgaard, 1990)

Year	**Claim Amounts (in millions)**				
1	2.495	2.120	2.095	1.700	1.650
2	1.985	1.810	1.625	–	–
3	3.215	2.105	1.765	1.715	–
4	–	–	–	–	–
5	19.180	1.915	1.790	1.755	–

The threshold of 1.5 million corresponds to the retention level of an excess-of-loss insurance policy[1].

To model this dataset within a Bayesian framework, assumptions about the distributions of both the number of claims in year t (N_t) and the amount of the i-th claim in year t ($Y_{i,t}$) were necessary. Claims were assumed to occur randomly and independently at a constant rate over time, so N_t was modeled using a Poisson distribution. A Pareto distribution was chosen for $Y_{i,t}$, as a heavy-tailed loss distribution was needed to account for the fact that individual claim amounts are positive and may include large outliers. That is,

$$N_t \sim \text{Poisson}(\theta), \quad 0 < \theta < \infty,$$
$$Y_{i,t} \sim \text{Pareto}(\alpha, \beta), \quad \alpha > 0, \quad 0 < \beta < y.$$

The Pareto(α, β) distribution with support $[\beta, \infty)$ was particularly suitable in this context, as we were modeling claim amounts exceeding a certain threshold.

In addition, the following assumptions were made:

- N_t are independently and identically distributed (i.i.d.) across t,
- $Y_{i,t}$ are i.i.d. across both i and t,
- N_t and $Y_{i,t}$ are independent for all i and t.

Under these assumptions, the aggregate claim amount in year t, denoted by

$$S_t = Y_{1,t} + Y_{2,t} + \cdots + Y_{N_t,t},$$

follows a compound Poisson distribution, since it represents the sum of independent Pareto-distributed random variables. [This is wrong?]

Next, prior distributions for the parameters α, β, and θ were specified. Due to limited prior knowledge about their true values—beyond the assumption that they are strictly positive—vague Gamma priors were chosen:

$$\alpha \sim \text{Gamma}(1, 0.0001), \quad \beta \sim \text{Gamma}(1, 0.0001), \quad \theta \sim \text{Gamma}(1, 0.0001),$$

with the constraint $0 < \beta < \min\{y_{i,t}\}$ to ensure validity of the Pareto distribution. Each of these Gamma priors has a variance of 10^8, implying minimal prior influence so that most of the information about the parameters is derived from the dataset. Additionally, the Gamma distribution is conjugate to both the Poisson and Pareto likelihoods, facilitating analytical tractability in Bayesian inference.

[1] To manage risk exposure, insurers frequently employ reinsurance strategies, which help reduce their financial liability on large claims. Under such arrangements, if a claim amount y exceeds a predetermined threshold d (the retention), the insurer is responsible only for paying up to d, while any excess $y - d$ is covered by the reinsurer.

Finally, the posterior distributions were derived. First, the joint posterior distribution of (α, β) was obtained via Bayes' theorem[2]:

$$\begin{aligned}
\pi(\alpha, \beta \mid \mathbf{y}) &\propto \pi(\alpha) \cdot \pi(\beta) \cdot f(\mathbf{y} \mid \alpha, \beta) \\
&\propto 0.0001 \cdot \exp(-0.0001\alpha) \cdot 0.0001 \cdot \exp(-0.0001\beta) \cdot \prod_{i=1}^{n} \frac{\alpha\beta^{\alpha}}{y_i^{\alpha+1}} \\
&\propto \exp(-0.0001\alpha) \cdot \exp(-0.0001\beta) \cdot \alpha^n \cdot \beta^{n\alpha} \left(\prod_{i=1}^{n} y_i \right)^{-(\alpha+1)} \\
&\propto \alpha^n \cdot \exp(-0.0001\alpha) \cdot \left(\prod_{i=1}^{n} y_i \right)^{-\alpha} \cdot \beta^{n\alpha} \cdot \exp(-0.0001\beta) \\
&\propto \alpha^n \cdot \exp\left(- \left(0.0001 + \sum_{i=1}^{n} \ln(y_i) \right) \alpha \right) \cdot \beta^{n\alpha} \cdot \exp(-0.0001\beta)
\end{aligned}$$

As a result, the full conditional posterior distributions of α and β were as follows:

$$\begin{aligned}
\pi(\alpha \mid \beta, \mathbf{y}) &\propto \alpha^n \cdot \exp\left(- \left(0.0001 - n\ln(\beta) + \sum_{i=1}^{n} \ln(y_i) \right) \alpha \right) \\
\pi(\beta \mid \alpha, \mathbf{y}) &\propto \beta^{n\alpha} \cdot \exp(-0.0001\beta)
\end{aligned}$$

which implied that:

$$\begin{aligned}
\alpha \mid \beta, \mathbf{y} &\sim \text{Gamma}\left(n+1, \sum_{i=1}^{n} \ln(y_i) - n\ln(\beta) + 0.0001 \right), \\
\beta \mid \alpha, \mathbf{y} &\sim \text{Gamma}(n\alpha + 1, 0.0001)
\end{aligned}$$

Similarly, the posterior distribution of θ was obtained via Bayes' theorem:

$$\begin{aligned}
\pi(\theta \mid \mathbf{n}) &\propto \pi(\theta) \cdot f(\mathbf{n} \mid \theta) \\
&\propto \exp(-0.0001\theta) \cdot \prod_{t=1}^{T} (\theta^{n_t} \cdot \exp(-\theta)) \\
&\propto \exp(-0.0001\theta) \cdot \theta^{\sum_{t=1}^{T} n_t} \cdot \exp(-5\theta) \\
&\propto \exp(-5.0001\theta) \cdot \theta^{\sum_{t=1}^{T} n_t}
\end{aligned}$$

which implied that:

$$\theta \mid \mathbf{n} \sim \text{Gamma}\left(\sum_{t=1}^{T} n_t + 1, 5.0001 \right)$$

Since all three posterior distributions were standard distributions, the Gibbs sampling method was employed to draw realizations from them. This was implemented using the `JAGS` program, which was called from within `R`. Three Markov chains were run in parallel. The initial values of α, β, and θ were chosen to be well-dispersed and are presented in Table 2.

Table 2: Initial Parameter Values

Chain	α	β	θ
1	1×10^{-5}	1×10^{-5}	1×10^{-5}
2	1×10^{5}	1	1×10^{5}
3	3.076	1.625	3.200

The burn-in period was set to 20,000 iterations. The statistics computed over the results of the subsequent 30,000 iterations are presented in Table 3. A comparison with the statistics reported by Dudley shows a close match, indicating that the model was properly specified and the Gibbs sampler was executed correctly.

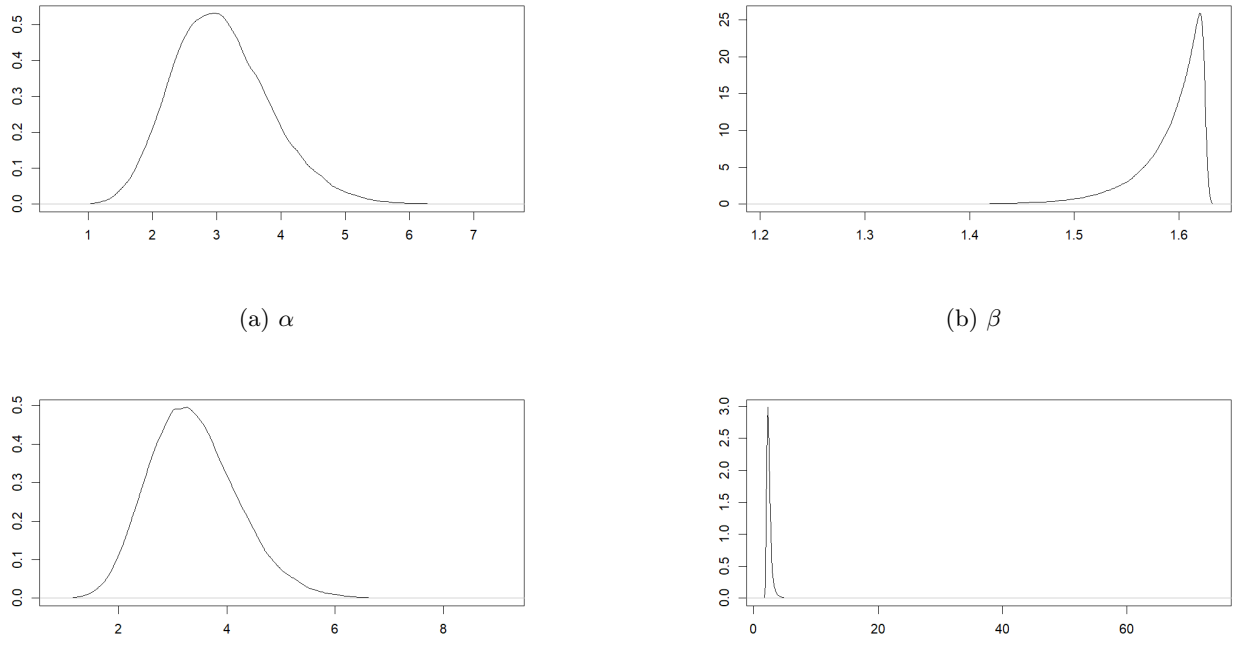
In addition, density plots were generated for each of the parameters and for $E[Y]$, as presented in Figure 1. The resulting densities for the parameters resemble Gamma distributions, with the density of β appropriately truncated at $\min\{y_{i,t}\} = 1.625$. The density plot for $E[Y]$ displays a right-skewed distribution that permits very large values, albeit with very low probability—consistent with expectations.

[2]Here, assuming that α and β are independent, the joint prior $\pi(\alpha, \beta)$ was computed as $\pi(\alpha) \cdot \pi(\beta)$.

Table 3: Posterior Statistics

	Mean	Standard Deviation	95% Bayesian Credible Interval
α	3.079	0.763	(1.771, 4.741)
β	1.592	0.035	(1.498, 1.624)
θ	3.396	0.821	(1.982, 5.192)
$E[Y]$	2.499	0.621	(2.024, 3.621)

Note: $E[Y]$ was calculated for each simulated set of parameters α and β, and from these values, the mean, standard deviation, and 95% Bayesian credible interval were subsequently computed.



(a) α (b) β

(c) θ (d) $E[y]$

Figure 1: Posterior Densities

The posterior means of α and β were used as parameters of the Pareto distribution, and the corresponding cumulative distribution function (CDF) was plotted against the empirical cumulative data ($y_{i,t}$). Similarly, the posterior mean of θ was used as the parameter of the Poisson distribution, and its CDF was plotted against the empirical cumulative data (n_t).

The Pareto(3.079, 1.592) distribution provides a close fit to the empirical data. The Poisson(3.396) distribution also fits the observed frequencies quite well.

Throughout all computations, a burn-in period of 20,000 iterations was applied, with only samples from iterations 20,001 to 50,000 retained for analysis. This approach followed the assumptions made by Dudley. However, it is essential to verify that the chains have indeed converged to a stationary distribution after discarding the initial samples. The first method of assessment involves visual inspection. Figure 4 presents trace plots for all three parameters, showing the sampled values across iterations. These plots indicate good mixing, suggesting that the chains have converged.

The convergence of the chains was also assessed using the Gelman–Rubin diagnostic. The diagnostic was applied to the post-burn-in iterations (20,001–50,000), and the results are summarized in Table 4.

Table 4: Potential Scale Reduction Factors (Gelman–Rubin Diagnostic)

Parameter	Point Estimate	Upper C.I.
α	1.00	1.00
β	1.00	1.00
θ	1.00	1.00
Multivariate PSRF	1.00	

All univariate potential scale reduction factors (PSRFs) have point estimates and upper confidence bounds at 1.00. The multivariate PSRF is also equal to 1.00. These values suggest that the Markov chains have likely

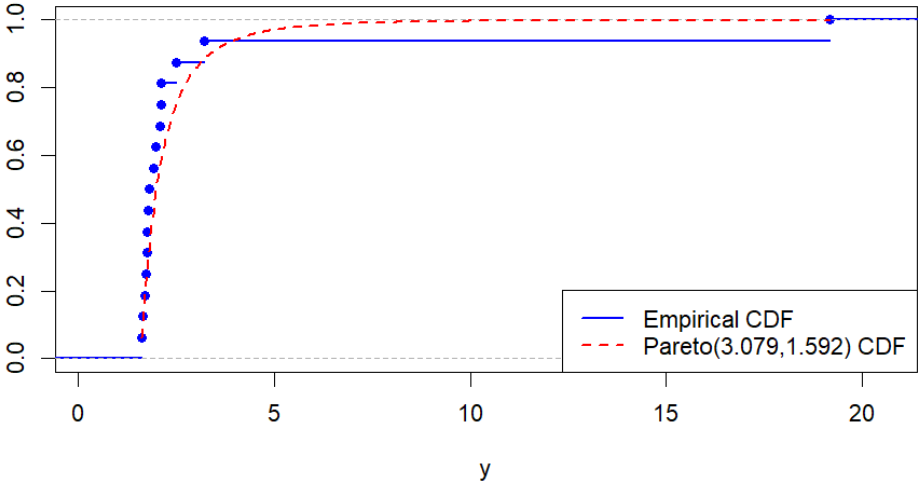


Figure 2: Empirical vs. Fitted Pareto CDF

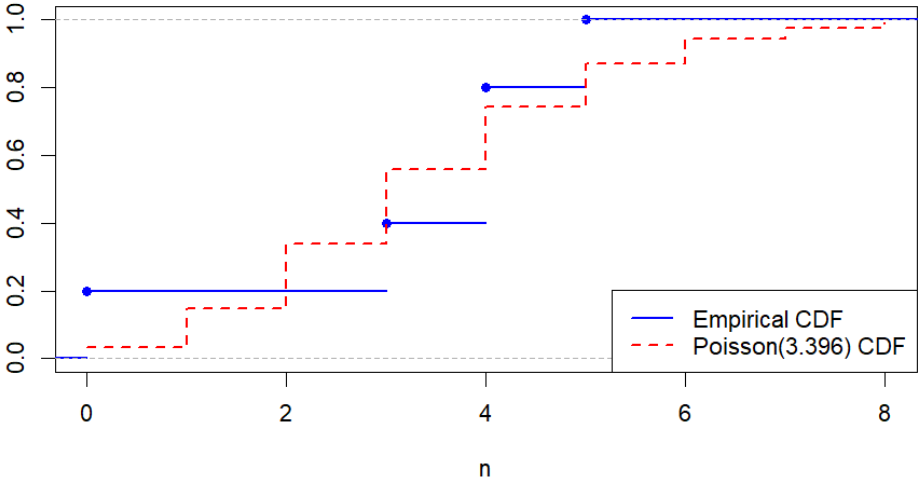


Figure 3: Empirical vs. Fitted Poisson CDF

converged, both for individual parameters and jointly.

Figure 5 shows how the univariate PSRF point estimates evolve with increasing iterations. Throughout all iterations, all estimates remain below 1.1, which is commonly considered an acceptable threshold for convergence. This further confirms that the chains have likely reached a stable distribution.

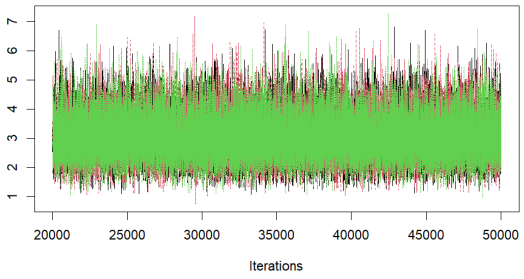
In addition, autocorrelation plots were generated for all three parameters (see Figure 6), and values at lags 1 through 10 are reported in Table 5. Several high autocorrelations were observed, particularly for β, which motivated the use of a thinning interval of 10 iterations, as suggested by Dudley.

Consequently, the chains were rerun with this thinning. Figure 7 presents the corresponding trace plots, and Figure 8 shows the updated autocorrelation plots. The trace plots indicate that the chains have mixed well, and all autocorrelation values at lags 2, 3, and beyond have become negligible.

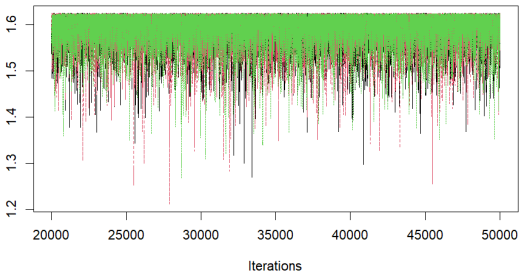
Since the ultimate goal of the analysis was to predict the values of S_t, the posterior predictive distribution was employed. For a variable z, this distribution is defined as:

$$\pi(z \mid \mathbf{y}) = \int_{\Theta} f(z \mid \phi)\, \pi(\phi \mid \mathbf{y})\, d\phi$$

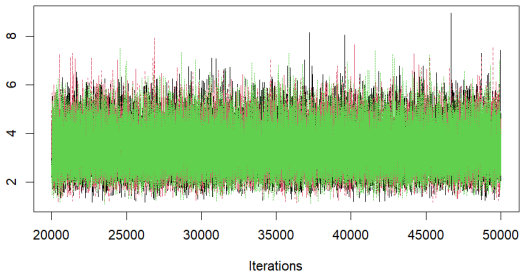
where $\phi = (\alpha, \beta, \theta)$, $\pi(\phi \mid \mathbf{y})$ is the posterior distribution of ϕ, and $f(z \mid \phi)$ is the likelihood of z given ϕ. This



(a) α

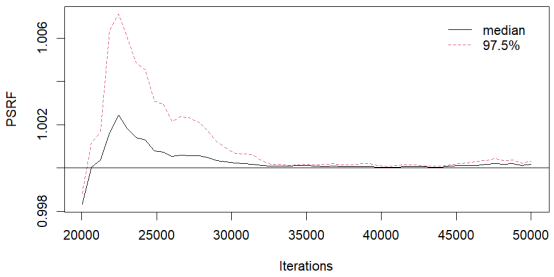


(b) β

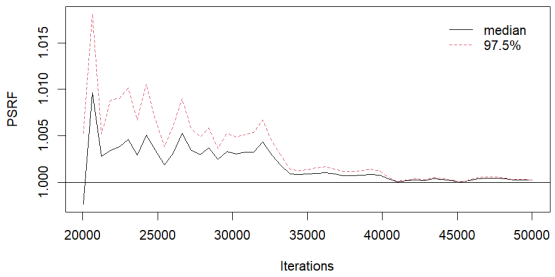


(c) θ

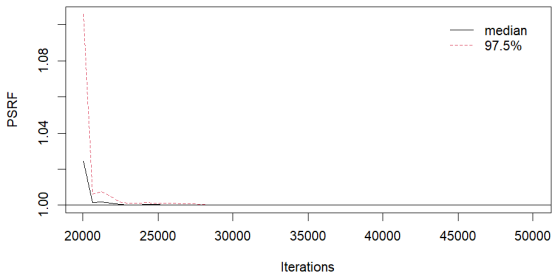
Figure 4: Trace Plots



(a) α



(b) β

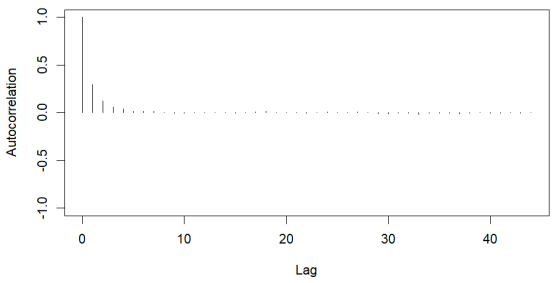


(c) θ

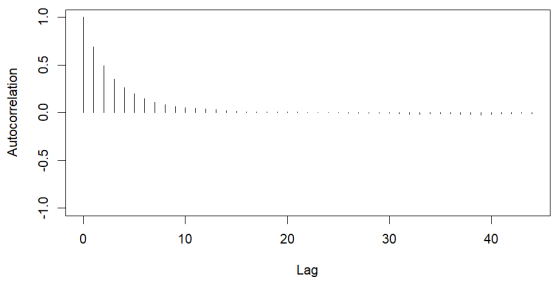
Figure 5: PSRF Values (Gelman–Rubin Diagnostic)

approach accounts for the uncertainty in ϕ by integrating over its possible values, weighted by their posterior probabilities.

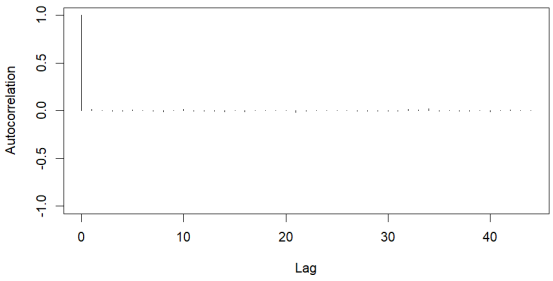
Let $\mathbf{n}$ denote the observed data and N_f represent a future observation. Then the posterior predictive



(a) α



(b) β



(c) θ

Figure 6: Autocorrelation Plots

Table 5: Autocorrelations at Lags 1–10

Lag	α	β	θ
1	0.297	0.705	0.003
2	0.119	0.509	0.005
3	0.061	0.373	−0.005
4	0.036	0.278	−0.004
5	0.023	0.210	0.001
6	0.016	0.161	−0.002
7	0.010	0.126	−0.004
8	0.009	0.102	−0.004
9	0.002	0.081	−0.001
10	0.007	0.065	0.004

distribution for N_f is:

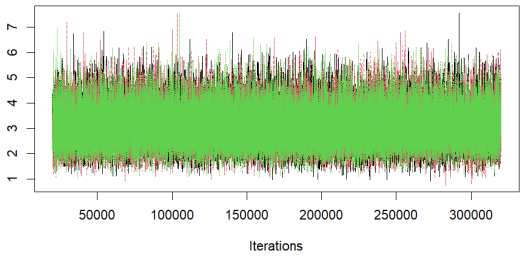
$$\begin{aligned} p(N_f = n \mid \mathbf{n}) &= \int_0^\infty p(N_f = n \mid \theta)\pi(\theta \mid \mathbf{n})\, d\theta \\ &= \mathbb{E}_{\theta|\mathbf{n}}\left[p(N_f = n \mid \theta)\right] \\ &= \mathbb{E}_{\theta|\mathbf{n}}\left[\frac{\theta^n e^{-\theta}}{n!}\right] \end{aligned}$$

Since the integral could not be solved analytically, it was approximated using samples from the posterior distribution obtained via MCMC. Specifically, the expectation was approximated as:

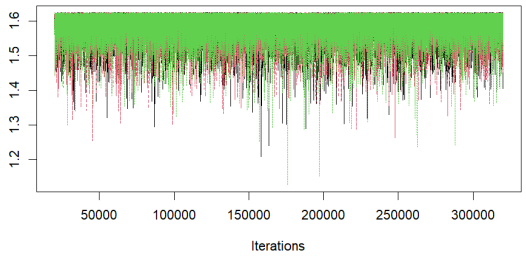
$$p(N_f = n \mid \mathbf{n}) \approx \frac{1}{m}\sum_{i=1}^{m} \frac{(\theta^{(i)})^n e^{-\theta^{(i)}}}{n!} \tag{1}$$

where $\theta^{(i)}$ is the i-th sample from the MCMC chain and m is the number of iterations after burn-in and thinning.

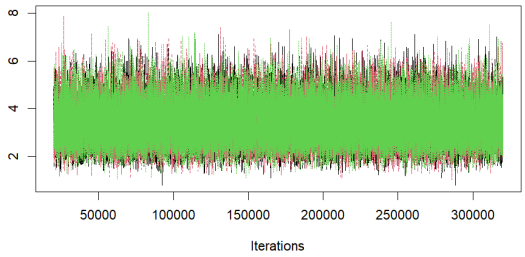
Similarly, let $\mathbf{y}$ denote the observed data, and let Y_f represent a future observation. The posterior predictive cumulative distribution function (CDF) of Y_f is given by:



(a) α

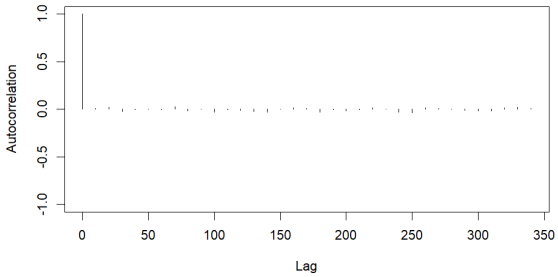


(b) β

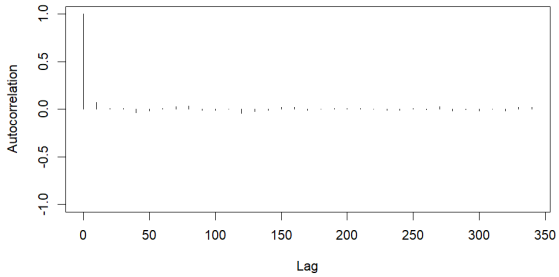


(c) θ

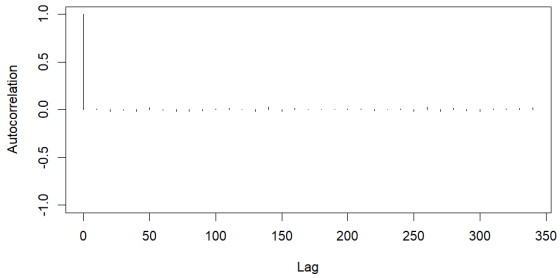
Figure 7: Trace Plots After Thinning



(a) α



(b) β



(c) θ

Figure 8: Autocorrelation Plots After Thinning

$$
\begin{aligned}
p(Y_f \leq y \mid \mathbf{y}) &= \int_{\mathbf{u}} p(Y_f \leq y \mid \mathbf{u})\, \pi(\mathbf{u} \mid \mathbf{y})\, d\mathbf{u} \\
&= \mathbb{E}_{\mathbf{u}|\mathbf{y}} \left[p(Y_f \leq y \mid \alpha, \beta) \right]
\end{aligned}
$$

Again,

$$p(Y_f \leq y \mid \mathbf{y}) \approx \frac{1}{m} \sum_{i=1}^{m} \left(1 - \left(\frac{\beta^{(i)}}{y} \right)^{\alpha^{(i)}} \right) \tag{2}$$

where $\alpha^{(i)}$ and $\beta^{(i)}$ denote the i-th samples from the MCMC chain, and m is the number of post-burn-in, thinned iterations.

Table 6 contains the estimated probabilities of $N_f = n$ for $n = 0, \ldots, 14$.

Table 6: Estimates of $p(N_f = n \mid \mathbf{n})$

n	Probability
0	0.0452
1	0.1279
2	0.1918
3	0.2023
4	0.1685
5	0.1178
6	0.0719
7	0.0393
8	0.0196
9	0.0091
10	0.0039
11	0.0016
12	0.0006
13	0.0002
14	0.0001

These results are consistent with those obtained by Dudley. As Y_f is a continuous variable, the probability density function (PDF) was estimated rather than discrete probabilities. The estimation employed the inverse cumulative distribution function (CDF) method.

A total of 1000 values $U \sim \text{Uniform}(0, 1)$ were generated. For each value, the transformation

$$y^{(i)} = \frac{\beta^{(i)}}{(1 - U)^{1/\alpha^{(i)}}}$$

was applied. For each i, the mean of the resulting values was computed. The resulting values were then used to approximate the PDF of Y_f via kernel density estimation (KDE). The estimated predictive density is illustrated in Figure 9.

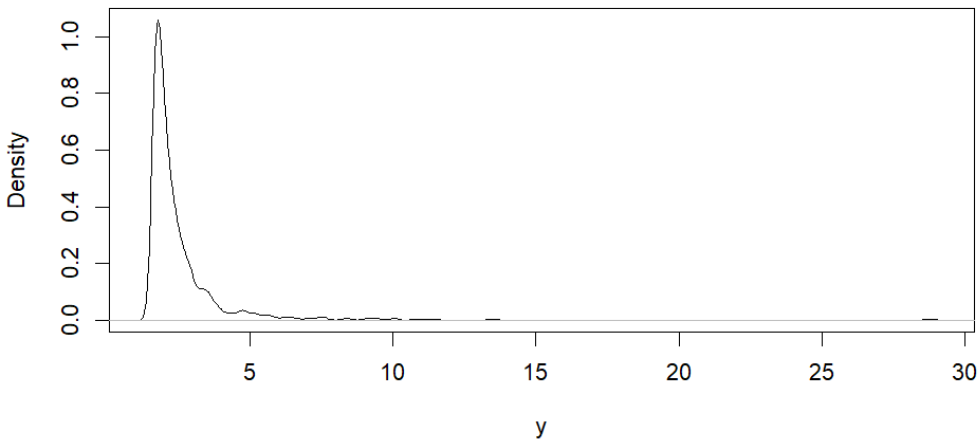


Figure 9: Estimated Predictive PDF of Y_f