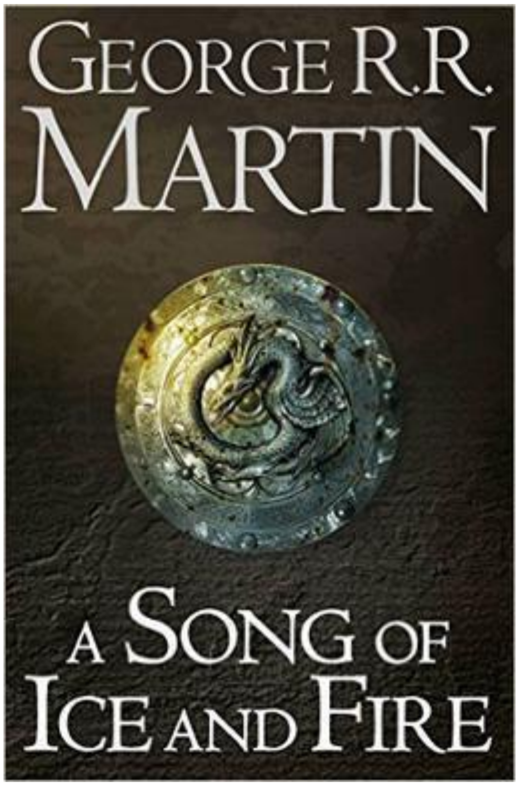
Bayesian Prediction on Game of Thrones Book Series

Predicting Number of Chapters Narrated from the Point of View of each Character



Group Members	**Contribution**
Sneha Shrungarpawar - A20438093	33.33%
Chirag Bhansali - A20436467	33.33%
Kausar Perveen - A20454157	33.33%

Github Link: https://github.com/chiraghbhansali/got-bayesian-analysis

1. INTRODUCTION

1.1. Overview

Game of Thrones is a very famous fantasy TV show which follows the book series "A Song of Ice and Fire" by George R.R Martin. The TV show, however, got diverted from the books in the last couple of seasons. The main reason being that the last two books of the series are not yet published and they are still in the writing phase. Many of the fans found the ending of the TV show quite disappointing, and are now waiting for the book to be published to read about the ending from the book series perspective.

In the existing five books of *A Song of Ice and Fire*, each chapter has been told from the point of view (POV) of a particular character in the series. For example if we say that X is one of the characters in the point of view characters, a chapter that has been told from the perspective of X will be called POV chapter for X. In total, up until book 5, there are 24 characters who have chapters being written from their point of view. A character must have at least one chapter being written from their point of view to be considered as a POV character.

1.2. Goal

The goal of our project is to predict POV of chapters for the characters in the last two books of the series. The past story of POV characters will affect the predictions, since some major characters were killed off in the last five books (Remembering the victims of Joffery's brutality and Red wedding). In addition to this, some characters might not have a very strong importance in the story and might not have any chapters in the subsequent books being written from their point of view.

1.3. Data

The data (Table 1) is obtained from a french fansite of Game of thrones
http://www.lagardedenuit.com
and consists of a matrix with 24 rows and 5 columns with one column for each of the book and one row for the main characters in the series as shown below.

- The columns are novel titles in their abbreviated form .
- The rows are the main characters in the book series.
- The cells/metric of the table are actually the POV chapters of the character in the corresponding book.

1.3.1 Data in terms of Model Parameters

Following is the parametrization of data according to the models used:

M = Matrix of POV characters and chapters as given by Table 1.
M_{ij} = POV chapters for character i in book j.

character	AGOT	ACOK	ASOS	AFFC	ADWD
Eddard	15	0	0	0	0
Catelyn	11	7	7	0	0
Sansa	6	8	7	3	0
Arya	5	10	13	3	2
Bran	7	7	4	0	3
Jon Snow	9	8	12	0	13
Daenerys	10	5	6	0	10
Tyrion	9	15	11	0	12
Theon	0	6	0	0	7
Davos	0	3	6	0	4
Samwell	0	0	5	5	0
Jaime	0	0	9	7	1
Cersei	0	0	0	10	2
Brienne	0	0	0	8	0
Areo	0	0	0	1	1
Arys	0	0	0	1	0
Arianne	0	0	0	2	0
Asha	0	0	0	1	3
Aeron	0	0	0	2	0
Victarion	0	0	0	2	2
Quentyn	0	0	0	0	4
Jon Connington	0	0	0	0	2
Melisandre	0	0	0	0	1
Barristan	0	0	0	0	4

1.4. Prediction Criteria

There are 3 possible ways of prediction here, Point Prediction, Interval Prediction and Probabilistic Prediction.

1.4.1. Point Prediction: Suppose it is believed that a certain character X will have 9 POV chapters in book 6. If we denote X6 as the number of POV chapters for character X in book 6, then X6 = 9 would tell us that the character X will have 9 POV chapters in book 6, which is nothing but a point prediction.

1.4.2. Interval Prediction: On the other hand, suppose it is believed that a certain character X will have 9 POV chapters in book 6. It is quite plausible that X might have 8 or 10 POV chapters instead. Instead of the point prediction X6 = 9, it would be better to give a range of likely values for X6. This is an interval prediction. For example, to say that the interval [7, 9] is an 80% credible interval for X6 is to say that there is an 80% probability that 7 ≤ X6 ≤ 9.

1.4.3. Probabilistic Prediction: Even more refined than an interval of likely values is a probability distribution over all possible values of the number of POV chapters. If we use P() to denote probability, we could say P(X6=9) = 0.5, P(X6=10) = P(X6=8) = 0.2, P(X6=7) = 0.1, and P(X6=other values) = 0. This probability distribution completely describes our belief about X6.

In this project, we will be doing probabilistic prediction and our aim would be to give probability distribution for each of the POV characters.

2. THE MODEL

2.1 Description of model

Let the number of POV chapters for character i in book t by X_{it}, $t \in \{1,2,3,4,5,6,7\}$. Let t_0 and t_1 are times such that the character is 'on-stage' between t_0 and t_1 and 'off-stage' at other times. For example, the character might be killed off at time t_1, that is $X_{it}=0$ for $t<t_0$ and $t>t_1$. For $t_0 \le t \le t_1$ we assume that POV chapters follow a Poisson distribution with parameter λ_i.

2.2 Assumptions

2.2.1. It is inconvenient to have t_0 and t_1 as model parameters, so instead we assume that there are $\tau_i \ge 0$ and β_i such that

$$X_{it} \sim \begin{cases} \mathrm{Pois}(\lambda_i) & \text{if } |t-\beta_i| < \tau_i \\ 0 & \text{otherwise.} \end{cases}$$

This is the same as putting $t_0 = \beta_i - \tau_i$ and $t_1 = \beta_i + \tau_i$ in Section 2.1. Note that we allow τ_i and β_i to take real values, even though t is constrained to be an integer.

2.2.2. It is undesirable for the τ_i, β_i and λ_i for each character to be independent of the other characters as this would give a model with 3N parameters where N is the number of characters. It is unlikely that good predictions could be made from a model with too many parameters. To cut down the number of parameters, we assume that λ_i, β_i and τ_i are random effects which means that they are samples from some underlying probability distribution. One motivation behind this assumption is that the parameters for different characters are assumed to have something in common. For example, there might be a typical value for τ_i which reflects how long the average character is likely to last in A Song of Ice and Fire.

2.2.3. The $\log(\lambda_i)$ are assumed to be normally distributed and β_i and τ_i are also assumed to be normally distributed. However, if there are no constraints on the values of β_i and τ_i, the model becomes difficult to fit, because for example there would be no difference in data generated by $\beta_i = 3$, $\tau_i = 3$ and $\beta_i = 3000$, $\tau_i = 3000$ for a particular character i, regardless of the value λ_i. Because this makes the inference problematic, we assume that the β and τ distributions are truncated in the interval [0, 7].

2.3. Hierarchical Model

2.3.1.

$$X_{it} \sim \begin{cases} \mathrm{Pois}(\lambda_i) & \text{if } |t-\beta_i| < \tau_i \\ 0 & \text{otherwise.} \end{cases}$$

for $1 \le i \le N$, and $t \in \{1,2,3,4,5,6,7\}$ with

$\log(\lambda_i) \sim N(\mu_\lambda,\sigma_{\lambda 2})$

$\tau_i \sim N(\mu_\tau,\sigma_{\tau 2})$ truncated to [0,7]

$\beta_i \sim N(\mu_\beta,\sigma_{\beta 2})$ truncated to [0,7]

where σ_λ, σ_τ, $\sigma_\beta > 0$ and μ_λ, μ_τ, $\mu_\beta \in R$. For fixed i, the X_{it} are assumed to be conditionally independent given λ_i, τ_i and β_i. For fixed t and $i \neq j$, the X_{it} and X_{jt} are assumed to be conditionally independent given the values of λ_i, τ_i, β_i and λ_j, τ_j and β_j.

2.3.2. To be explicit, let the data be:

$$(x_{it})_{1\le i\le N}^{1\le t\le d}$$

For $1 \le i \le N$, define L_i as

$$L_i = L_i((x_{it})_{t=1}^{d}, \lambda_i, \tau_i, \beta_i)$$

$$L_i = \prod_{t:x_{it}\neq 0} \frac{e^{-\lambda_i}\lambda_i^{x_{it}}}{x_{it}!} \prod_{t:x_{it}=0} (e^{-\lambda_i}\delta_{|t-\beta_i|<\tau_i} + \delta_{|t-\beta_i|\ge\tau_i}).$$

Then the likelihood is proportional to

$$\int \prod_{i=1}^{N} L_i \frac{1}{\sigma_\lambda} e^{-\frac{(\log(\lambda_i)-\mu_\lambda)^2}{2\sigma_\lambda^2}} \frac{1}{\sigma_\tau} e^{-\frac{(\tau_i-\mu_\tau)^2}{2\sigma_\tau^2}} \frac{1}{\sigma_\beta} e^{-\frac{(\beta_i-\mu_\beta)^2}{2\sigma_\beta^2}} \delta_{0\le\tau_i\le 7}\delta_{0\le\beta_i\le 7}$$

where the integral is over 3N dimensions λ_i, τ_i and $\beta_i \ge 0$ and the symbol δ_p stands for 1 if p is true and 0 if p is false.

2.3.3. A model like this is often called a hierarchical model. The μ_λ, σ_λ, μ_τ, σ_τ, μ_β and σ_β are called hyperparameters to distinguish them from the individual λ_i, τ_i and β_i.

2.4. Method of inference

2.4.1. The model is fitted using Bayesian inference with non-informative $N(0,1000^2)$ priors on the location parameters μ_λ, μ_τ, μ_β and inverse gamma (0.001,0.001) priors on the scale parameters σ_λ, σ_τ, σ_β. Because intractable-looking integrals appear in the likelihood (1), the model is fitted using Gibbs sampling. For the λ_i and β_i, samples are drawn from the marginal distribution using a histogram approximation.

2.4.2. At each iteration of the algorithm, a value of (μ_λ, σ_λ, μ_τ, σ_τ, μ_β, σ_β) is sampled using the theory of the normal distribution and then, for each character i, the values of λ_i, τ_i and β_i are sampled in that order. Then predictions for $X_{i,d+1}$ and $X_{i,d+2}$ are sampled using the definition of Model 2.5. After all iterations are complete, a burn-in is discarded and the output is thinned to make the resulting samples as uncorrelated as possible.

2.4.3. The output of the algorithm is a collection of samples of μ_λ, σ_λ, μ_τ, σ_τ, μ_β ,σ_β) and predictions $X_{i,d+1}$ and $X_{i,d+2}$ for the values of i=1 to i=n where n = ((N − burn-in) / thin).

3. RESULTS

3.1. Problems with Model

Initially, book 4 and book 5 were supposed to be one single book but it was later split into two books. This resulted in a lot of zeros for different subsets of characters in the POV table. The introduction of zeros increased the skewness of data which in turn led to the hierarchical model not fitting the data well.

Solution: Data Smoothing

The solution to this problem was found by replacing the Matrix M with a new Matrix such that the new matrix contains the average of Book 4 and Book 5 using the following method.

$$M'_{i4} = c_4 \frac{M_{i4} + M_{i5}}{c_4 + c_5}$$
$$M'_{i5} = c_5 \frac{M_{i4} + M_{i5}}{c_4 + c_5}$$

where c_4 and c_5 represents the total number of chapters in book 4 and book 5 respectively. Data smoothing was done initially, before adding the data to the model.

3.2. Posterior Predictive Distributions

The specifications of model fitting are as follows:

Sampler = Gibbs Sampler
Population size = 100000
Sample size = 1000
Data = M' (The data after applying smoothing)

The sampler was run multiple times to check the stability of results and after scrutinizing the stability, one of the results is shown in the report.

3.2.1. Result of Posterior Predictive Distributions

The results of posterior predictive distribution for 1000 samples for each character is given below. The table shows the distribution of posterior predictive distribution in numerical form and the figures shows the distribution of posterior predictive distribution in graphical form.

3.2.1.1. Posterior Predictive Distribution Numerical Results

Table 2 shows the posterior predictive distributions of POV chapters for Book 6 (Winds of Winter) on the left and for Book 7 on the right. The results of Book 7, as seen, are not satisfactory and one of the reasons is that the POV of characters for Book 7 depends on their POV scores in Book 6, which is unknown as of now.

Posterior Samples for Winds of Winter (Table 2 left)

Character	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Eddard	884	33	27	21	15	8	5	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Catelyn	914	17	14	13	12	10	7	5	3	2	1	1	0	0	0	0	0	0	0	0	0	0
Sansa	161	64	97	127	134	121	97	69	46	27	17	9	5	3	3	2	2	2	2	2	2	2
Arya	143	28	43	68	93	109	115	104	86	66	47	31	21	12	8	5	4	3	2	2	2	2
Bran	176	88	127	144	135	111	81	50	30	18	10	5	3	2	2	2	2	2	2	2	2	2
Jon Snow	136	17	20	34	51	73	89	99	100	92	78	61	45	32	23	15	10	6	4	3	3	2
Daenerys	146	33	54	82	107	119	114	99	77	56	37	23	14	9	6	4	3	2	2	2	2	2
Tyrion	133	13	16	22	36	53	71	87	94	94	87	74	60	47	34	23	17	11	8	6	4	3
Theon	328	155	166	137	93	56	31	16	8	4	2	1	1	0	0	0	0	0	0	0	0	0
Davos	320	150	165	137	97	57	32	18	9	5	3	2	1	1	1	0	1	0	0	0	0	0
Samwell	427	164	151	109	69	37	20	11	5	3	1	1	0	0	0	0	0	0	0	0	0	0
Jaime	412	84	105	109	95	71	50	31	18	10	5	3	1	1	0	0	0	0	0	0	0	0
Cersei	576	122	106	82	51	29	15	8	3	2	1	0	0	0	0	0	0	0	0	0	0	0
Brienne	559	164	119	75	42	20	10	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0
Areo	726	178	58	21	8	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arys	793	147	36	12	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arianne	728	177	58	20	8	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Asha	614	206	100	44	19	8	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Aeron	728	176	58	20	8	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Victarion	615	206	99	43	18	8	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Quentyn	612	205	100	45	19	9	4	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Jon C	726	179	59	20	8	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Melisandre	795	143	37	12	5	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Barristan	615	206	100	42	19	8	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Table 2 (Right) Posterior Samples for Book 7

Character	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	29	20	21
Eddard	897	28	24	19	13	9	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Catelyn	921	14	13	12	11	9	7	5	3	2	1	1	0	0	0	0	0	0	0	0	0	0
Sansa	321	53	82	102	108	98	77	56	36	22	13	8	4	3	2	2	2	2	1	1	1	1
Arya	304	24	37	56	75	88	91	84	70	53	38	25	17	10	7	4	3	3	2	2	2	1
Bran	331	75	103	116	110	89	62	42	24	14	7	4	3	2	1	2	1	1	1	1	2	1
Jon Snow	297	14	17	27	42	59	72	82	81	74	63	49	37	26	18	12	8	6	4	3	2	2
Daenerys	305	27	45	69	87	95	94	79	62	45	31	19	11	7	5	3	2	2	2	2	2	1
Tyrion	297	12	13	19	30	45	57	69	76	75	70	59	49	37	28	19	13	9	6	4	3	3
Theon	521	111	119	96	66	40	22	11	5	3	1	1	1	0	0	0	0	0	0	0	0	0
Davos	517	108	116	97	67	41	23	12	6	3	2	1	1	1	0	0	0	0	0	0	0	0
Samwell	629	110	99	69	42	24	12	6	3	2	1	0	0	0	0	0	0	0	0	0	0	0
Jaime	627	58	68	70	60	44	30	18	11	6	3	2	1	1	0	0	0	0	0	0	0	0
Cersei	726	86	68	51	31	18	8	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0
Brienne	720	110	76	45	23	12	6	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0
Areo	813	124	39	12	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arys	856	101	26	8	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Arianne	813	125	39	13	5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Asha	747	139	63	28	12	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Aeron	813	123	39	13	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Victarion	750	137	64	26	11	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Quentyn	746	139	66	27	11	5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Jon C	812	124	40	13	5	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Melisandre	857	100	25	8	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Barristan	751	137	64	26	11	5	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 2

3.2.1.2. Visual Representation of Posterior Predictive Distribution

The following figure (Figure 1) shows the visual representation of posterior predictive distribution. From the graphs it can be interpreted that Eddard, Catelyn, Areo, Arys, Melisandre and Jon C have the lowest variances while on the other hand Jon and Tyrion have the highest.

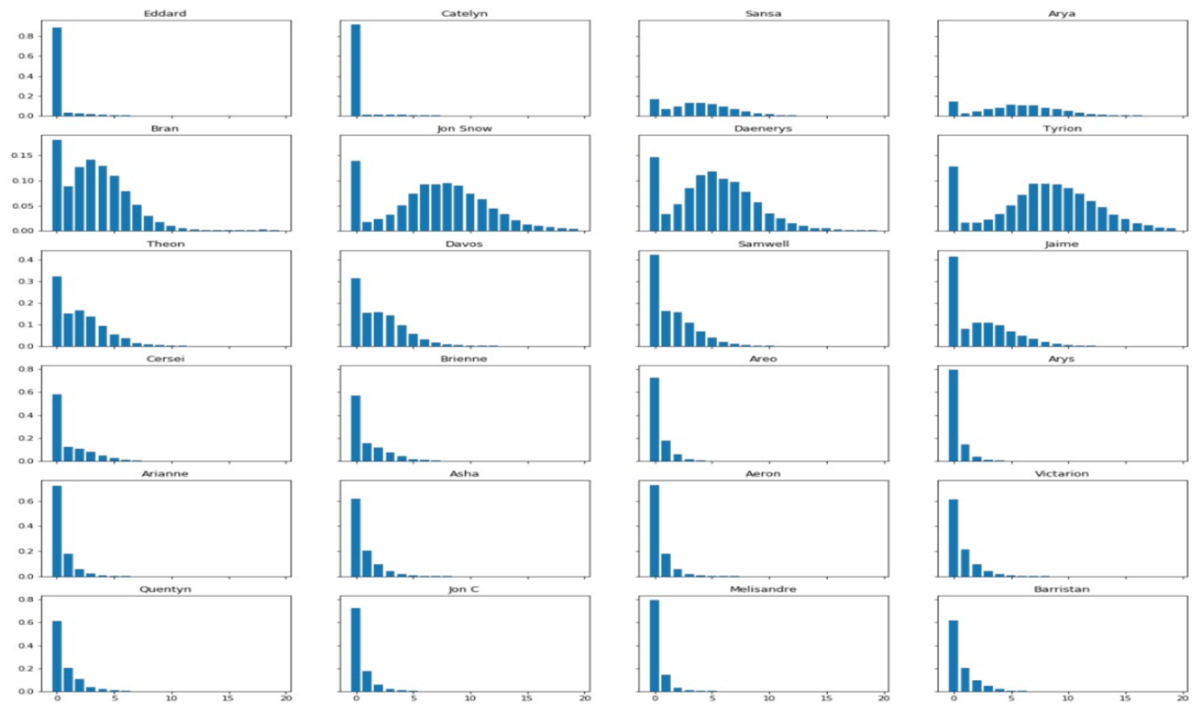


Figure 1: Histograms representing the posterior predictive distributions of characters

3.3. Probabilities for Zero POV Chapters

The first column in Table 2 corresponds to a specific character having zero chapters written from his/her POV. The probabilities are shown in Figure 2. Since Eddard was killed in the first book, we can see that the probability of Eddard having zero chapters in Book 6 and Book 7 is more than 0.8.

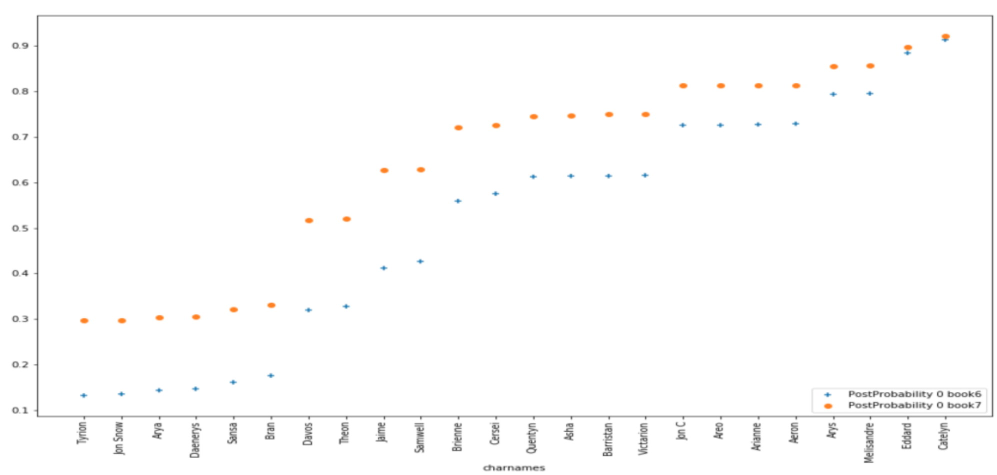


Figure 2: Probability of having zero POV chapters in Book 6 and Book 7

4. Testing and Validation

4.1. Testing Method

The testing method that we used was testing by inference. Data was generated using the model, and this data was used in the chosen inference method to see whether the given inference method could return the parameters that were used to generate the data set.

4.1.1. Testing

To test the method of inference, the model was fitted with the following specifications:

Data sets = 100

$(\mu\lambda,\sigma\lambda,\mu\tau,\sigma\tau,\mu\beta,\sigma\beta) = (1.3,0.75,2,1,4,1.5)$ (Hyperparameters values $\approx$ posterior medians for the hyperparameters obtained from one of the fits of Hierarchical Model)

$$\mu\lambda,\sigma\lambda,\mu\tau \sim N(0,0.12)$$

$$\sigma\tau,\mu\beta,\sigma\beta \sim \exp(\Delta) \text{ where } \Delta \sim N(0,0.012)$$

The Data sets were used to calculate α–credible intervals for each of the location and scale hyperparameters by taking the 100α% of the posterior distributions. This led to 600 credible intervals per α. This is plotted in Figure 3.

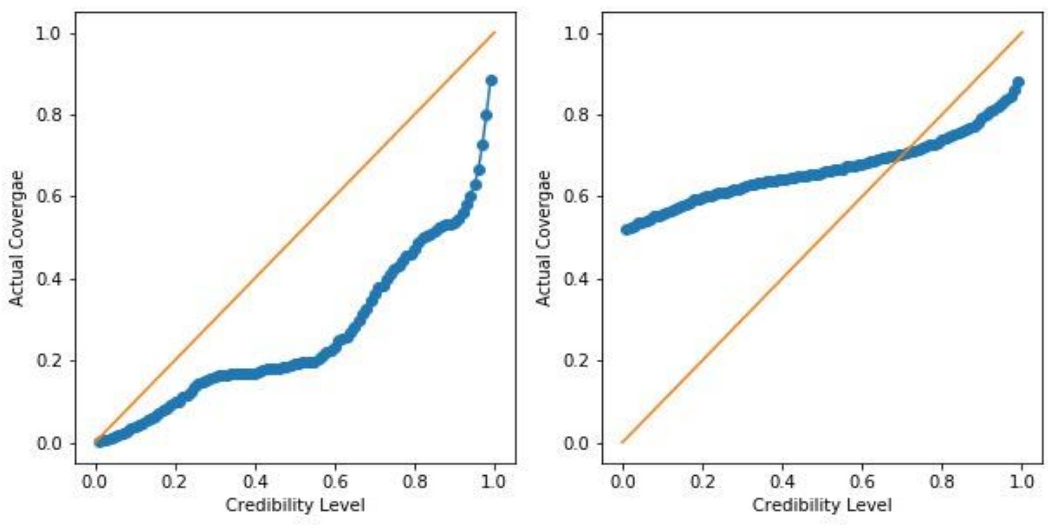


Figure 3

4.1.2. Justification

The procedure for inference testing was carried out for the one-step-ahead predictions for book 6. The result, shown in Figure 3 (right panel) shows that the credible intervals have greater coverage than they should. In figure 3, the reason for the credible intervals to have greater coverage than they should is because POV can only take integer values and so the closed interval $[q_{\alpha/2}, q_{1-\alpha/2}]$ obtained by taking the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution will in general cover more than $100\alpha\%$ of the posterior samples.

4.2. Evaluations

Since evaluating a model requires testing it on unseen data and in this case the unseen data will be Book 6 of ASOF. It is not possible for us to test on Book 6, since the data is not available. The test was then conducted on the existing books to check the accuracy of the model.

4.2.1. Procedure

Books selected = Book 1 and Book 2
POV characters = 9
M = 9 x 2
True values = Third column of matrix M in Table 1

The model was fitted and evaluated using the above parameters. Figure 4 shows the result of fitting the model to Matrix M. The intervals displayed are central 50% (solid lines) and 80% (dotted lines) credible intervals. The coverage is satisfactory but the intervals are much too wide to be of interest.

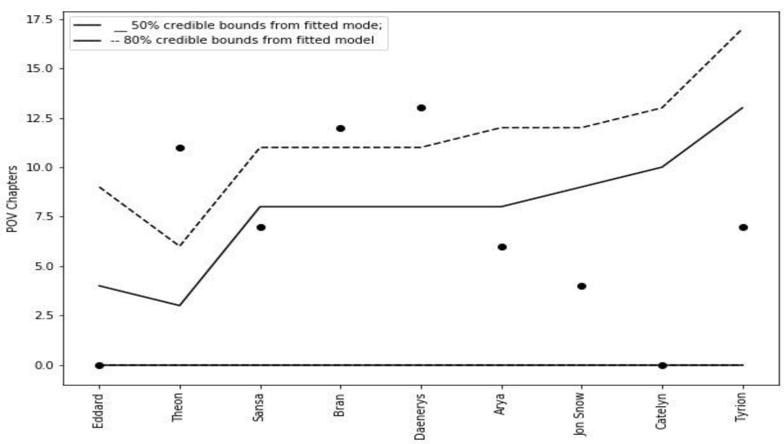


Figure 4

5. ISSUES WITH MODEL

5.1. Model Issues

It is more common for us to have used negative binomial distribution for count data instead of Poisson but using negative binomial distribution would have introduced more complexity in the model and so it was traded in favor of Poisson. This gives us no other strong reason for the choice of Poisson.

5.2. Data Issues

5.2.1. New Data Issues

The model does not take into account the introduction of new characters. The solution can be to find the mode of POV chapters for all the given characters and find the difference with previous data to evaluate how many chapters will be told from POV of new characters but pointing out the number of new POV characters is not possible.

5.2.2. Missing Data for Evaluation

As discussed earlier the lack of data for Book 6 affected the evaluation of the model and the model cannot be validated until the release of Book 6 and subsequently Book 5.

References

[1] A Song of Ice and Fire by George R. R. Martin
[2] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
[3] Heike Trautmann, Detlef Steuer, Olaf Mersmann and Bj ̈orn Bornkamp (2014). truncnorm: Truncated normal distribution. R package version 1.0-7. http://CRAN.R-project.org/package=truncnor