Probabilistic Graphical Models Case study: Hierarchical Bayesian Model

Khaled Fouda

March 2021

Abstract

Here we go over the process of building a hierarchical Bayesian model to model the future number of retweets for a given tweet. We begin by giving a summary on the important methods and theories used in our study. In section two we get introduced to the data and choose appropriate likelihood distribution. In section three we build the graphical model while defining the conjugate priors and computing the posterior's parameters. In the last section we implement the model in R-language and sample from the conditional posterior distributions. This study is accompanied with the code files for section 4 and the exploratory data analysis done in section 2.

Contents

1 Background

1.1 Bayesian Probabilities

In Bayesian statistics of one response variable we assign two distribution. The likelihood $p(y|\theta)$ where we have only a subset of y and we need to model it so we can predict future values of it. We choose it based on either the shape of the frequency (looks like normal or exponential) or based on the nature of the data (Bernoulli for binary data).

The prior $p(\theta)$ can either represent our belief of how the data is distributed (ex: expect the parameter to be normally distributed with mean 3 and sd 1) or it can have no information at all -noninformative- and the model has to learn it all by itself.

We are interested in two other distributions. The first one is the posterior

$$p(\theta|y) = \frac{p(\theta, y)}{\sum_\theta p(\theta, y)} \propto p(\theta, y) = p(\theta)\, p(y|\theta)$$

In order to find an accurate (tight) distribution of θ we will sample this posterior distribution. However the denominator is quite difficult to compute so we will depend on the algorithm that lets us sample the posterior from $p(\theta)\, p(y|\theta)$.
We are also interested in predicting future values of y and hence we need a marginal distribution free of any other parameters and hence

$$p(y) = \frac{p(\theta, y)}{p(\theta|y)} = \sum_\theta p(\theta, y)$$

1.1.1 Conjugate Priors

Given the distribution of the likelihood, if the the prior and posterior has the same distribution with different parameters then we say that the prior is a natural conjugate prior.
This is very useful because if we know that the posterior has a common distribution then we can easily sample from it.

Moreover, if the likelihood belong to the exponential family then there exist a natural conjugate prior associated with it. Below is a table with some examples of likelihood functions and their conjugate priors that we will use to define our priors later. Note that these priors are not unique.

The likelihood of an exponential family has the following representation

$$p(y|\theta) = \Pi_{i=1}^{n} p(y_i|\theta) = \Pi f(y_i)\; g(\theta)^n\; exp(\phi(\theta) \sum u(y_i)) \propto g(\theta)^n\; exp(\phi(\theta) \sum u(y_i))$$

So if we choose our prior as

$$p(\theta) \propto g(\theta)^m\; exp(v\; \phi(\theta))$$

Then the posterior can be represented as

$$p(\theta|y) \propto g(\theta)^{n+m}\; exp(v\; \phi(\theta) \sum u(y_i))$$

Note that $\sum u(y_i)$ is a sufficient statistic for θ.

Likelihood	Model parameters	Prior/Posterior	prior param	posterior param
Binomial	p(probability)	Beta	α,β	$\alpha + \sum y_i$,$\beta + nN - \sum x_i$
Normal	μ, σ^2	Normal, Inverse Gamma	μ_p,σ_b^2,α,β	* check reference
Poisson	λ	Gamma	α,β	$\alpha + \sum x_i$, $\beta + n$

1.1.2 Noninformative prior distributions

When prior distributions have no population basis, they can be difficult to construct, and there has long been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution. The rationale for using noninformative prior distributions is to let the data speak for themselves, so that inferences are unaffected by information external to the current data.
One way of choosing the noninformative priors is by Jeffreys' in-variance principle which states that the noninformative prior is proportional to fisher's information.:

$$p(\theta) \propto I(Y) = -E\,\frac{d^2}{d\theta^2} log(p(y|\theta))$$

For example, if our likelihood has a a binomial distribution $y \sim Bin(n, \theta)$ then we can compute fisher's information to be $\theta^{-.5}\ (1-\theta)^{-.5}$ and we can conclude that the prior $\theta \sim Beta(.5, .5)$

1.1.3 References and extra readings

1. **A Compendium of Conjugate Priors** Deriving conjugate priors for many common distributions.s

2. **Bayesian Data Analysis, by Andrew Gelman** Chapter 2: Single-parameter models.

1.2 Stationary Markov Chains

A stationary MC doesn't change over time. That is $p(x_{i+1}|x_i) = p(x_{k+1}|x_k)$

Definition A stationary distribution of a Markov chain is a distribution π such that :

$$\pi(y) = \sum_x p(y|x)\pi(x)$$

Definition A Markov chain is said to be reversible if there exists a probability distribution π such that

$$p(x|y)\pi(y) = p(y|x)\pi(x)$$

Result If a MC is reversible then it is stationary since,

$$\sum_x p(x|y)\pi(y) = \sum_x p(y|x)\pi(x) = \pi(y)$$

1.3 MCMC Sampling

Simulating a Markov chain until stationarity is achieved then simulating this stationary Markov chain to estimate θ.

1.3.1 Gibbs Sampling

Now assume we have two parameters θ, α with joint posterior distribution $p(\theta, \alpha|y)$ and that this distribution is known but complicated. However, if the two conditional posteriors are easy to sample from then we can use them instead.
Hence the algorithm is as follows:

1. Initialize θ and α to some values.
2. For each time step t:
    - Sample $\theta_t \sim p(\theta|\alpha, y)$
    - Sample $\alpha_t \sim p(\alpha|\theta, y)$

- The sampling order can be sequential like we did above or be random.

- This can be generalized to the case of p parameters.
- If the conditional posteriors are hard to sample from then we cannot use this method.

How to compute the conditional posteriors? Luckily we don't need to sum or integrate. Because we are dealing with proportionality not exact distribution we can safely assume that the conditional posterior of θ is all the elements of the joint posterior that contains θ. All other parameters as well as y are considered constant and can be discarded.

1.3.2 The Metropolis-Hastings Algorithm

Gibbs sampling is a special case of the MH algorithm. Now assume that the conditional posteriors are also complicated.

We need to define a transition function $Q(\theta^{new}|\theta^{old})$ Then at each step t:

1. Sample $\theta^{new} \sim Q(\theta^{new}|\theta^{old})$
2. Accept the new sample with probability $\frac{p(\theta^{new})\,Q(\theta^{new}|\theta^{old})}{p(\theta^{old})\,Q(\theta^{old}|\theta^{new})}$
3. otherwise, set $\theta^{new} = \theta^{old}$

Remember that our goal is to reach a stationary MC process. For a reversible process and stationary process the acceptance probability above is 1.

Moreover, In the basic Metropolis algorithm, we assume that the proposal distribution is symmetric, that is $Q(\theta^{new}|\theta^{old}) = Q(\theta^{old}|\theta^{new})$ and the acceptance probability is reduced to $p = \frac{p(\theta^{new})}{p(\theta^{old})}$

- A common choice for the sampler is one that provides random small steps.

  - For example, if our prior is $\theta \sim Normal()$ we can use $Q(\theta^{new}|\theta^{old}) = N(\theta^{old}, \epsilon)$ where ϵ is chosen to be very small to guarantee having small steps.

- If we have a mix of easy-to-sample conditional posteriors and hard-to-sample ones, we can use Gibbs for the easy ones and M-H for the hard ones inside the sample loop. This approach is called Metropolis-within-Gibbs.

- As with Gibbs sampling, subsequent samples are highly correlated. We can solve this by counting only every k^{th} accepted sample to reduce the dependency.

- How do we know we have reached stationarity?

1.3.3 Convergence

After running an MCMC we need to verify that we have indeed reached a stationary process.

Gelman-Rubin Approach aims to diagnose convergence. Lets assume we run $m/2$ Markov chains each with $4n$ iterations and is randomly initialized. Discard the first half and split the resulting $2n$ samples into two sub-chains. now we can say we have m markov chains with n samples each.

The portion we discarded constitutes the burn-in period where the chains are assumed to be in their transient phase. We now hope that the mxn samples are from the desired stationary distributions. To verify that we are going to compare the between-chain variance with the within-chain variance.

Let $\theta_{i,j}$ be the sample i from the chain j. The between- and within-chain variances, B and W, computed as:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\theta}_{.j} - \bar{\theta}_{..})^2$$

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2 \text{ where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{ij} - \bar{\theta}_{.j})^2$$

Notice The similarities between them and the MS_{ERROR} and MS_{EFFECT} in the ANOVA effect model.

We can now estimate $Var(\theta|y)$ as a weighted average of W and B with :

$$V\hat{a}r(\theta|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

This estimation is unbiased if we reach stationarity. Moreover, W is an underestimate of $Var(\theta|y)$ but it should approach it when $n \to \infty$. We can therefore monitor convergence through

$$\hat{R} = \sqrt{\frac{V\hat{a}r(\theta|y)}{W}}$$

Values of $\hat{R} \leq 1.1$ are acceptable but the closer to 1 the better. We then monitor $\hat{R}$ for each of the parameter we estimate.

1.3.4 References and extra readings

- **MCMC and bayesian Modeling by Martin Haugh** These lecture notes provide an introduction to Bayesian modeling and MCMC algorithms including the Metropolis-Hastings and Gibbs Sampling algorithms.

- **Pattern recognition and machine learning by C.Bishop** Chapter 11: Sampling Method.

2 Model likelihood

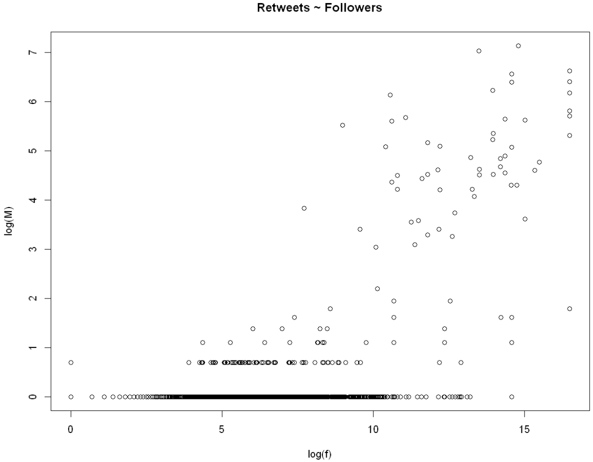
Our data consists of 52 root tweets and recursively all the retweets to them. We have in total 12248 data rows. The relevant elements of each row are:

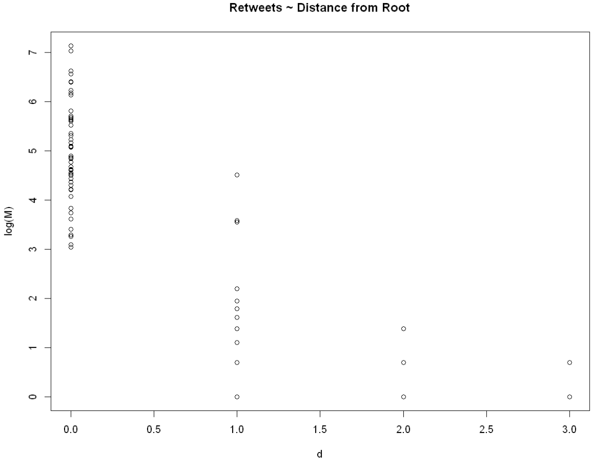
- **case:** 1 to 52 to keep track to which root the re-tweeter belongs.
- **FollowerCount:** The number of followers for the re-tweeter. Denote it f_j^x where x is the case number and j is the re-tweeter's order in that case.
- **DistanceFromRoot:** $= 0$ if it's the root (ie: the tweet's owner). $= 1$ if the user retweeted it directly from the root. > 1 if it was retweeted from a previous retweet. Denote it d_j^x
- **TimeDiff:** The time in minutes between the tweet's origin and the user's retweet of it. I took log(x+1) to control the outliers. Denote it S_j^x
- **Retweets:** The number of retweets received for that user j. For the root user, this is the total number of retweets he received for the original tweet M_0^x. For the re-tweeters, it's the number of retweets they received on their retweet (skewed near 0). Denote it M_j^x

We are interested in modeling the total number of retweets the root user receives M_0^x and the possibility of being able to predict it before observing the full data.

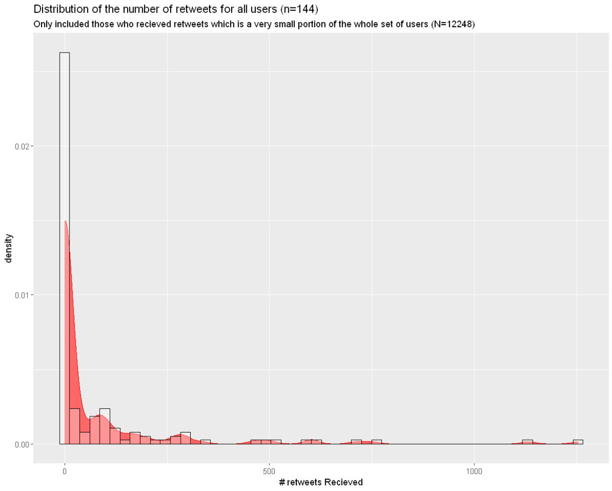
We assume the the number of retweets M_j^x to be related to the number of followers f_j^x and the distance from the root d_j^x. By fitting an ANOVA model we confirm our assumption since the p-values for the model $log(M_j^x) \sim log(f_j^x) + d_j^x$ are all very small.

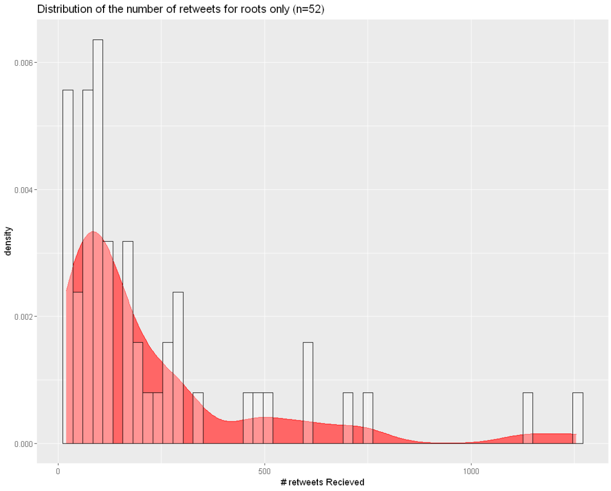
The plots below illustrate the linear relations. The more follower a user get, the more retweets they receive and the further from the root the less likely they will receive retweets.





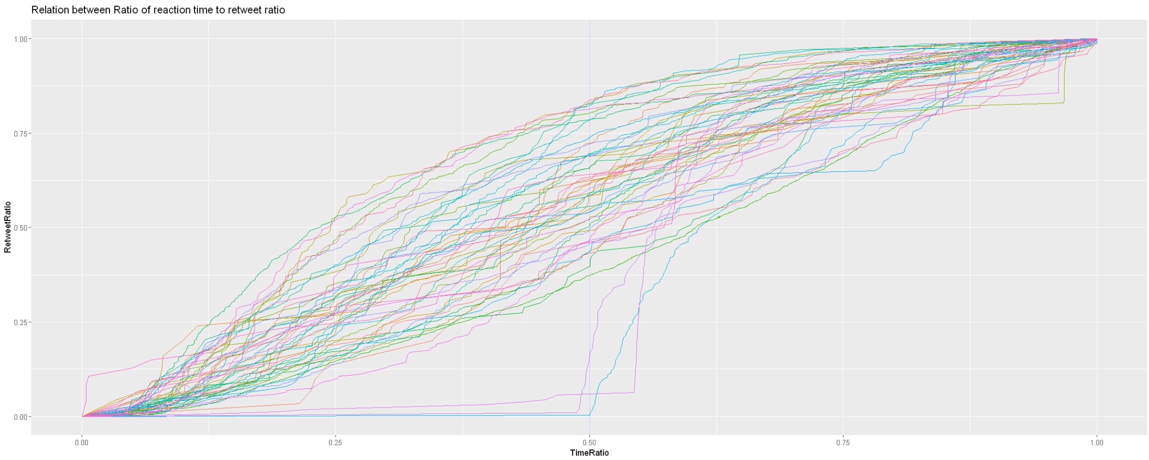
Below is the distribution of M_j^x. For the first plot, all users except those who didn't receive any retweets are shown. In the second, only root users are considered. We notice that the data is highly skewed.





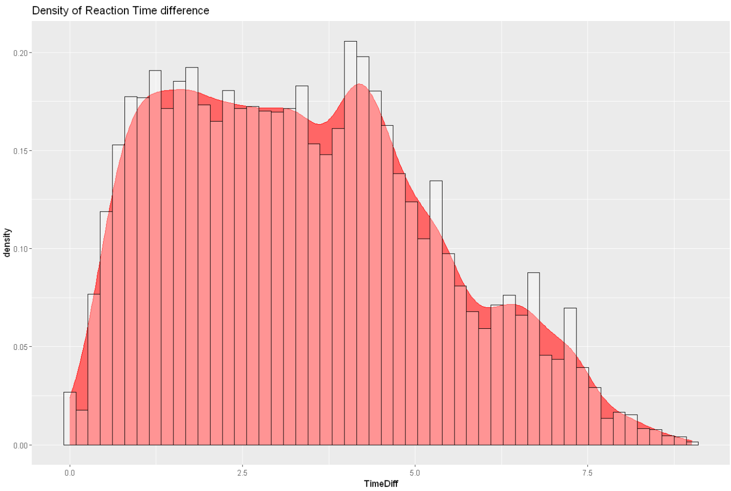
We will assume that the followers are all equally likely to retweet. If we have a number b_j^x representing the probability that any of the followers f_j^x will retweet then we can model $M_j^x \sim Binomial(f_j^x, b_j^x)$. We may also assume that b_j^x is related to both f_j^x and d_j^x. More on modeling them in the next section.

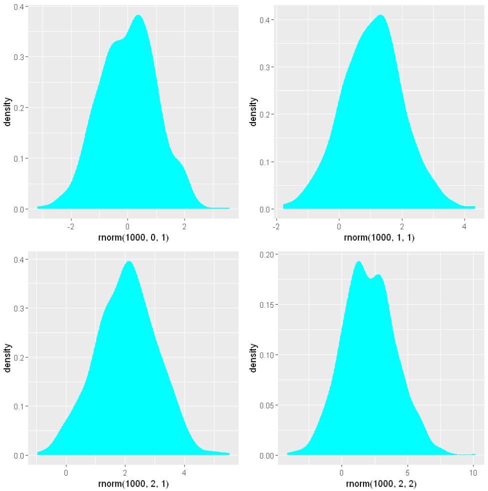
Now we consider another variable of interest, The reaction time. To understand how it affects the total number of retweets, lets analyze the following plot where we plot the time-ratio and retweet-ratio for every root user (in colors) to notice changes in the growth of retweets overtime.



We can see that the growth in retweets is logistic where in the first portion of time the growth is exponential but after time it slows down and eventually flatten out.
That shows an interesting relation between the two variable and therefore we will be interested in modeling the reaction time as well.

Below on the left we have the distribution of the reaction time (after taking the log) and on right we have the density of several normal distributions. From this we can assume that the reaction time (in log) has a normal distribution.





3 Priors and Posteriors

To summarize our likelihood, we have:

$$P(M_j^x|f_j^x, b_j^x) = Binomial(M_j^x|f_j^x, b_j^x) = \binom{f_j^x}{m_j^x}(b_j^x)^{(m_j^x)}(1-b_j^x)^{(f_j^x - m_j^x)}$$

$$P(S_j^x|\alpha^x, \tau^x) = Normal(S_j^x|\alpha^x, \tau^x) \propto (\tau^x)^{-n/2} exp\{-\frac{(s_j^x - \alpha^x)^2}{2\tau^x}\}$$

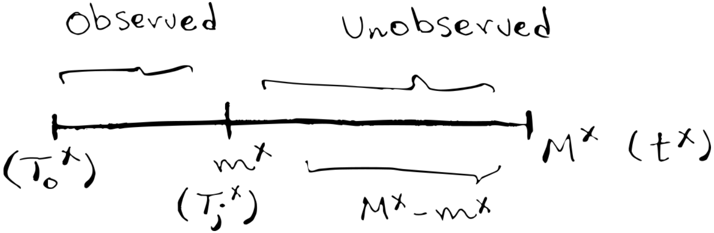
Note here τ^x is the variance for user x I omitted the square to avoid confusion.

Now we write the joint likelihood equation. We first consider the training case where all the retweets are available. Let $S^x = \cup_j S_j^x$ and $M^x = \cup_j M_j^x$ ie, the retweets for user x.

$$P(S^x, M^x|\alpha^x, \tau^x, b^x, m^x) = \Pi_{j=0}^{m^x} Norm(s_j^x|\alpha^x, \tau^x)\ Bin(M_j^x|b_j^x, f_j^x)$$

Now assume tat the data is partially observed. Let m^x be the observed retweets and M^x be the eventual total number of retweets. Therefore we have $M^x - m^x$ unobserved retweets and if the last retweet observed is at time T_j^x and t^x is the time of the last retweet, then the reaction times between T_j^x and t^x are unobserved.

Recall that the reaction time in our data is $S_j^x = log(t_j^x - T_0^x + 1)$ where T_0^x is the time of the first tweet and the added 1 is to avoid having a log of 0.

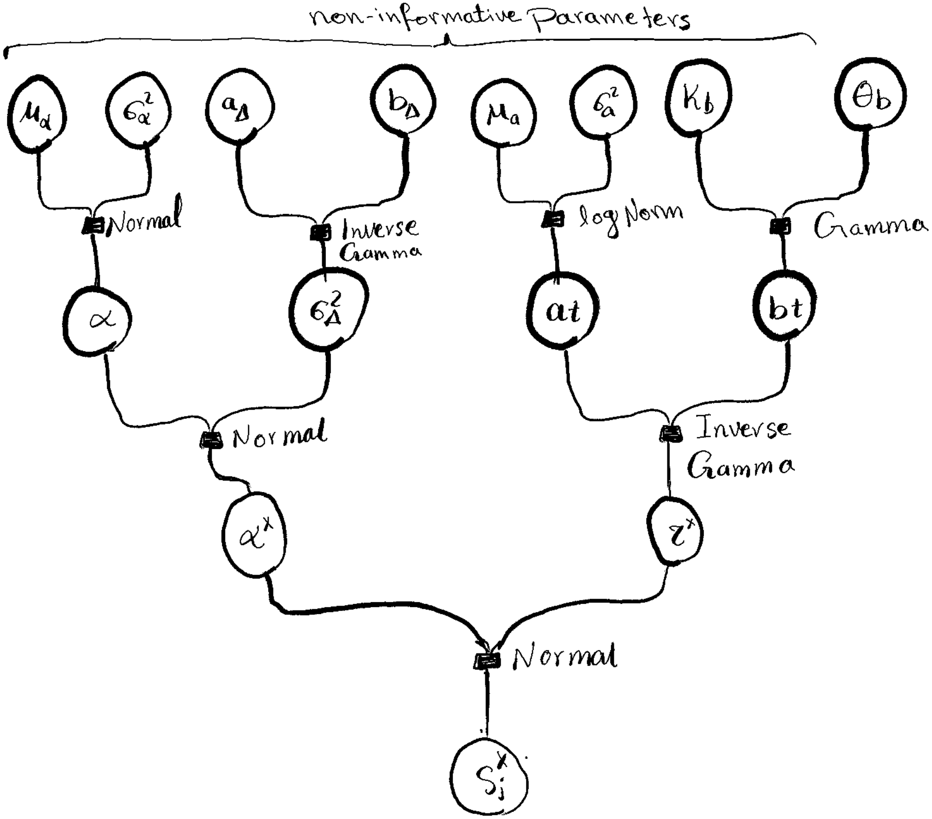


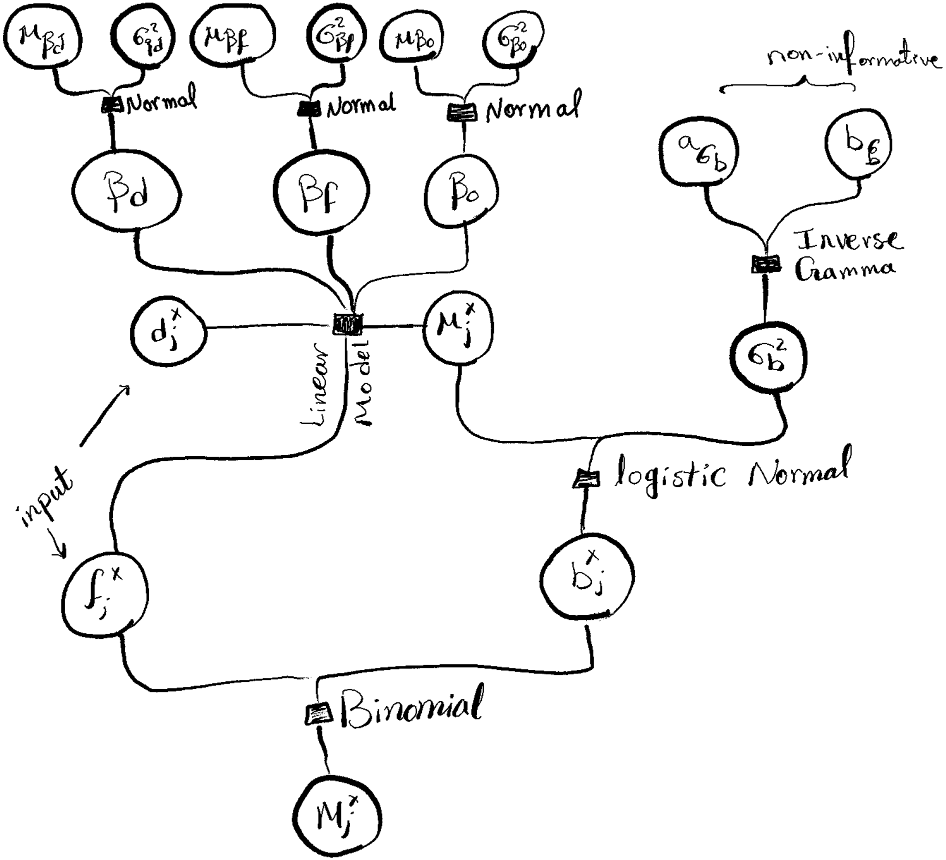
Here we assume that the data is ordered. ie, m_2^x is the second retweet received at time $t_2^x > t_1^x$. We can then use order statistic properties to compute the joint likelihood when the data is partially observed.

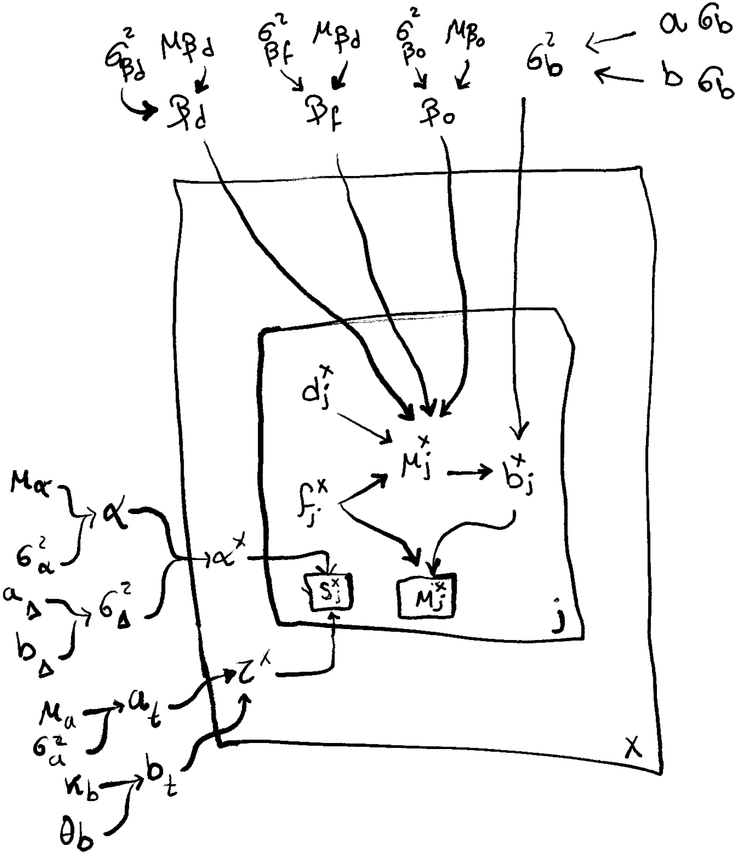
$$P(S_{t^x}^x, m_{t^x}^x|\alpha^x, \tau^x, m^x, M^x) = \prod_{j=0}^{m^x} \binom{M_j^x}{m_j^x} f(s_j^x|\alpha^x, \tau^x)(1 - F(t^x - T_j^x|\alpha^x, \tau^x))^{M_j^x - m_j^x}$$

The combination term is there because we have $M_j^x - m_j^x$ possible unobserved retweets.

The chosen model and priors are illustrated below. We will analyze then in details for the rest of this section.


non-informative Parameters
μ_α
σ_α^2
a_Δ
b_Δ
μ_a
σ_a^2
k_b
θ_b
Normal
Inverse Gamma
log Norm
Gamma
α
σ_Δ^2
at
bt
Normal
Inverse Gamma
α^x
τ^x
Normal
S_j^x


$\mu_{\beta d}$
$\sigma^2_{\beta d}$
$\mu_{\beta f}$
$\sigma^2_{\beta f}$
$\mu_{\beta o}$
$\sigma^2_{\beta o}$
non-informative
Normal
Normal
Normal
a_{σ_b}
b_{σ}
βd
βf
βo
Inverse Gamma
d_j^x
μ_j^x
σ_b^2
Linear Model
input
logistic Normal
f_j^x
b_j^x
Binomial
M_j^x


$\sigma^2_{\beta d}$
$\mu_{\beta d}$
β_d
$\sigma^2_{\beta f}$
$\mu_{\beta d}$
β_f
$\sigma^2_{\beta o}$
$\mu_{\beta o}$
β_o
σ^2_b
$a\ \sigma_b$
$b\ \sigma_b$
d_j^x
μ_j^x
b_j^x
f_j^x
S_j^x
M_j^x
j
X
μ_α
α
σ^2_α
α^x
a_Δ
σ^2_Δ
b_Δ
τ^x
μ_a
a_t
σ^2_a
κ_b
b_t
θ_b

3.1 Conjugacy

We first need to justify some of our priors. We assume that the Normal-IG is a conjugate prior for the normal likelihood and that Gamma is conjugate prior for the IG.

3.1.1 Normal-IG for Normal likelihood

We begin to prove the first assumption. Assume that the likelihood $D \sim Normal(\mu, \sigma^2)$ and consider the case when σ^2 is known but μ isn't. (we will treat each parameter individually while fixing the other. Since we are using Gibbs sampling where we sample from each parameter while fixing the others, this should not cause problems.).

$$p(D|\mu) \sim Normal(\mu, \sigma^2) \propto exp\{\frac{-1}{2\sigma^2}\sum_x (x_i - \mu))^2\}$$

We claim that the prior of the following shape is conjugate

$$p(\mu) \propto exp\{\frac{-1}{2\sigma_o^2}(\mu - \mu_o)^2\} \sim Normal(\mu|\mu_o, \sigma_o^2)$$

proof: we now show that the posterior is normally distributed.

$$p(\mu|D) \propto p(\mu)p(D|\mu) \propto exp\{\frac{-1}{2\sigma_o^2}(\mu - \mu_o)^2 + \frac{-1}{2\sigma^2}\sum_x (x_i - \mu))^2\}$$

$$= exp\{\frac{-1}{2\sigma^2}\sum_x(x_i^2 - 2x_i\mu + \mu^2) - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_o + \mu_0^2)\}$$

$$exp\{\frac{-\mu^2}{2}[\frac{1}{\sigma_o^2}+\frac{n}{\sigma^2}]+\mu[\frac{\mu_o}{\sigma_o^2}+\frac{\sum x_i}{\sigma^2}]-[\frac{\mu_o^2}{2\sigma_o^2}+\frac{\sum x_i^2}{2\sigma^2}]\}$$

We want the expression above to be similar to:

$$exp\{\frac{-1}{2\sigma_n^2}(\mu^2-2\mu\mu_n+\mu_n^2)\}$$

For some parameters σ_n^2 and μ_n. We now solve the two equations for the parameters.

$$\frac{-\mu^2}{2\sigma_n^2}=\frac{-\mu^2}{2}[\frac{1}{\sigma_o^2}+\frac{n}{\sigma^2}]$$

Solving this we get for σ_n^2 we get

$$\sigma_n^2=\frac{\sigma^2\sigma_o^2}{n\sigma_o^2+\sigma^2}=(n+\sigma^2\sigma_o^{-2})^{-1}\sigma^2 \tag{1}$$

And by equating the terms for μ we get:

$$\frac{\mu\mu_n}{\sigma_n^2}=\mu[\frac{\mu_o}{\sigma_o^2}+\frac{\sum x_i}{\sigma^2}]$$

then,

$$\mu_n=(n+\sigma^2\sigma_o^{-2})^{-1}\sum x_i \tag{2}$$

And hence the posterior $p(\mu|D)\sim Normal(\mu|\mu_n,\sigma_n^2)$

Moreover, The predictive posterior also has a normal distribution,

$$p(x|D)=\int p(x|\mu)p(\mu|D)\,d\mu=\int N(x|\mu,\sigma^2)\;N(\mu|\mu_n,\sigma^2)\;d\mu=Normal(x|\mu_n,\sigma_n^2+\sigma^2)$$

But for this presentation we will focus on the posterior model more than the predictive posterior.

We now fix μ and show that σ^2 has an IG conjugate. Let $\lambda = \sigma^2$ to avoid confusion about the square. We claim that $\lambda \sim IG(\lambda|\alpha, \beta)$ is conjugate.
proof:

$$p(\lambda) \propto \lambda^{-\alpha-1} exp(-\beta/\lambda)$$

Then the posterior is,

$$p(\lambda|D) \propto p(\lambda)p(D|\lambda) \propto \lambda^{-\alpha-1} exp(-\beta/\lambda)\ \ \lambda^{-n/2}\ exp(\frac{-1}{\lambda}\sum(x_i-\mu)^2)$$

$$= \lambda^{(-(\alpha+n/2)-1)}\ \ exp(\frac{-1}{\lambda}(\frac{1}{2}\sum(x_i-\mu)^2+\beta))$$

From this we see that $\lambda \sim IG(\alpha_n, \beta_n)$ where,

$$\alpha_n = \alpha + n/2 \text{ and } \beta_n = \frac{1}{2}\sum(x_i-\mu)^2 + \beta \tag{3}$$

3.1.2 Gamma for-Inverse Gamma

Now assume we have an Inverse-Gamma likelihood of the shape

$$p(D|\alpha,\beta) = \prod IG(x_i|\alpha,\beta) \propto (\prod x_i)^{-\alpha-1} e^{-\beta\sum\frac{1}{x_i}}\ \beta^{n\alpha}(\frac{1}{\Gamma(\alpha)})^n$$

Assume first that α is fixed. We claim that $\beta \sim Gamma(\beta|\alpha_o|\beta_o)$ is a conjugate prior.
Proof: The posterior can be calculated as,

$$p(\beta|D) \propto p(D|\beta)p(\beta) \propto e^{-\beta\sum\frac{1}{x_i}}\beta^{n\alpha}\beta^{\alpha_o-1}e^{-\beta\beta_o}$$

$$= \beta^{n\alpha+\alpha_o} exp\{-\beta(\beta_o + \sum \frac{1}{x_i})\}$$

Hence we can see that the resulting posterior is

$$Gamma(\beta \,|\, n\alpha + \alpha_o \,,\, \beta_o + \sum \frac{1}{x_i}) \tag{4}$$

Now we fix β and try to find a conjugate prior for α. By expressing the likelihood above as an exponential family - by taking e^{log} - we get,

$$p(D|\alpha) = e^{(-\alpha-1)log(\prod x_i)} \beta^{n\alpha} e^{n(log(1)-log(\Gamma(\alpha)))}$$

$$\propto e^{(-\alpha-1)\sum log(x_i)} (\beta^{\alpha} e^{-log(\Gamma(\alpha))})^n$$

$$= e^{\phi(\alpha)\sum u_i(x_i)} \;\; [g(\alpha)]^n$$

A conjugate prior will have the shape of,

$$\propto e^{\phi(\alpha)v} \;\; [g(\alpha)]^m$$

For some v and n. However, there is no obvious conjugate so we will use an alternative method (Metropolis-Hastings) to sample α.
Nevertheless, we still need to assign a prior distribution for alpha. Since $\sum log(x_i)$ is a sufficient statistic for α, it's reasonable to assume a log-normal prior.

3.2 Joint Posterior

Going back to our diagram. The joint posterior is the product of the likelihoods and priors which is the same as taking the product of the CPD in the graph. In other words, the product of the densities of each node given its parents. Here we are considering the full model where all the data is observed. If we considered the partially observed data then we need to use the joint distribution $p(S^x_{t^x}, m^x_{t^x})$ we computed above.

Let $\Phi = (\alpha, \sigma^2_\Delta, \bigcup_x \alpha^x, a_t, b_t, \bigcup_x \tau^x, \beta_o, \beta_f, \beta_d, \sigma^2_b, \bigcup_x \bigcup_j b^x_j)$ be the list of all the model parameters.

$$p(\Phi|M,S) = p(\alpha)p(\sigma^2_\Delta)\prod_x p(\alpha^x|\alpha,\sigma^2_\Delta)p(a_t)p(b_t)\prod_x p(\tau^x|a_t,b_t)\prod_x\prod_j p(S^x_j|\alpha^x,\tau^x)$$

$$*\, p(\beta_o,\beta_f,\beta_d)p(\sigma^2_b)\prod_x\prod_j p(b^x_j|\mu^x_j,\sigma^2_b)\prod_x\prod_j p(M^x_j|f^x_j,b^x_j)$$

3.3 Conditional Posteriors

It's not easy to sample directly from the joint posterior distribution above. However, using MCMC sampling we focus at one parameter a time and consider everything else to be constant.
For this section we will derive the conditional posterior for each parameter using the two conjugates we proved above and using either the diagram or the joint posterior.
Using the joint posterior above, for parameter ζ we only consider the densities that has ζ in its parameterization.

3.3.1 α^x , τ^x

Using the concept above, we have the posterior to be,

$$p(\alpha^x|\bigcup_j S_j^x) \propto p(\alpha^x|\alpha,\sigma_\Delta^2)\prod_j p(S_j^x|\alpha^x,\tau^x) = \text{ Prior X likelihood}$$

$$p(\tau^x|\bigcup_j S_j^x) \propto p(\tau^x|a_t,b_t)\prod_j p(S_j^x|\alpha^x,\tau^x) = \text{ Prior X likelihood}$$

And since we defined the likelihood to be Gaussian, we define the prior/posterior of α^x and τ^x be Gaussian and Inverse-Gamma as shown earlier. We will use

$$p(\alpha^x|\bigcup_j S_j^x) \propto N(\alpha^x|\alpha,\sigma_\Delta^2)\prod_j N(S_j^x|\alpha^x,\tau^x)$$

$$p(\tau^x|\bigcup_j S_j^x) \propto IG(\tau^x|a_t,b_t)\prod_j N(S_j^x|\alpha^x,\tau^x)$$

And using the parameterization we derived in equations (1)(2)(3) we have that the posterior of α^x is Gaussian with parameters:

$$\mu' = (M^x + \tau^x\sigma_\Delta^{-2})^{-1}\sum_j s_j^x$$

$$\sigma^{2'} = (M^x + \tau^x\sigma_\Delta^{-2})^{-1}\tau^x$$

and τ^x is an IG with parameters:

$$\alpha' = a_t + \frac{M^x}{2}$$

$$\beta' = b_t + .5\sum_{j=1}^{M^x}(s_j^x - \alpha^x)^2$$

Where M^x is the total number of retweets for root tweet x.

3.3.2 α , σ_Δ^2

Again, we do the same exact thing we did.

$$p(\alpha|\alpha^x) \propto p(\alpha|\mu_\alpha, \sigma_\alpha^2)\prod_x p(\alpha^x|\alpha, \sigma_\Delta^2)$$

$$p(\sigma_\Delta^2|\alpha^x) \propto p(\sigma_\Delta^2|a_\Delta, b_\Delta)\prod_x p(\alpha^x|\alpha, \sigma_\Delta^2)$$

Here we have the likelihood to be Gaussian as well. We choose the priors/posteriors to be Normal-IG where the paramters of the priors are noninformative. $N(0, 100^2)$ and $IG(.5, .5)$. Hence,

$$p(\alpha|\alpha^x) \propto N(\alpha|\mu_\alpha, \sigma_\alpha^2)\prod_x N(\alpha^x|\alpha, \sigma_\Delta^2)$$

$$p(\sigma_\Delta^2|\alpha^x) \propto IG(\sigma_\Delta^2|a_\Delta, b_\Delta)\prod_x N(\alpha^x|\alpha, \sigma_\Delta^2)$$

And using the parameterization we derived in equations (1)(2)(3) we have that the posterior of α to be Normal with parameters:

$$\mu' = (X + \sigma_\Delta^2\sigma_\alpha^{-2})^{-1}\sum_x \alpha^x$$

$$\sigma^{2'} = (X + \sigma_\Delta^2\sigma_\alpha^{-2})^{-1}\sigma_\Delta^2$$

and σ^2_Δ is an IG with parameters:

$$\alpha' = a_\Delta + \frac{X}{2}$$

$$\beta' = b_\Delta + .5\sum_{x=1}^{X}(\alpha^x - \alpha)^2$$

Where X is the number of root tweets.

3.3.3 a_t , b_t

As shown in 3.1.2, For an IG likelihood, Gamma is a conjugate prior for the rate parameter. Hence,

$$p(b_t|\tau^x) \propto p(b_t|k_b, \theta_b)\prod_x p(\tau^x|a_t, b_t) =\propto G(b_t|k_b, \theta_b)\prod_x IG(\tau^x|a_t, b_t)$$

We choose the prior's parameters to be noninformative. $k_b = .5, \theta_b = 1/500$. The posterior can be evaluated using formula (4).

$$p(b_t|\tau^x) \propto Gamma(Xa_t + k_b, \theta_b + \sum_x \frac{1}{\tau^x})$$

As for a_t we will give it a noninformative log-normal$(0, 10^2)$ prior and a normal step-wise transition function,

$$Q(a_{t+1}|a_t) \sim Normal(a_t, .2)$$

Where the small variance is to guarantee making only small steps. We can then use M-H method to sample a_t.

3.3.4 b_j^x

b_j^x represents the probability that a follower f_j^x will retweet the tweet x. We give it a logistic Normal prior and transition function. We take the logit to guarantee that it's bounded between 0 and 1. Moreover, its mean is a Bayesian linear regression model of f_j^x and d_j^x. More on that later. We will use M-H to sample b_j^x.

3.3.5 σ_b^2

Since the likelihood is logit-Normal, if we transformed the xs by $logit(x)$ we get the a usual Normal likelihood and we can give σ_b^2 and conjugate prior/posterior Inverse-Gamma().

The prior's parameters are noninformative : $p(\sigma_b^2|a_{\sigma_b}, b_{\sigma_b}) \sim IG(.5, .5)$.

The posteriors parameters are calculated using formula (3) to be:

$$p(\sigma_b^2|\bigcup_x \bigcup_j b_j^x) \sim IG(a', b')$$

where,

$$a' = a_{\sigma_b} + \frac{N}{2}$$

$$b' = b_{\sigma_b} + .5 \sum_x \sum_j (logit(b_j^x) - \mu_j^x)^2$$

Where N is the total number of retweets in the dataset. ie. $\sum_x \sum_j (1)$

3.3.6 μ_j^x, β

We assumed that the mean of b_j^x to be related to the number of followers f_j^x and the distance from the root d_j^x.
Fitting a linear model would be a reasonable choice.

Let $X = \{1, log(f_j^x + 1), log(d_j^x + 1)\}$ and $\beta = \{\beta_o, \beta_f, \beta_d\}$ Then the likelihood can be expressed as

$$logit(b_j^x) \sim Normal(\beta^T X, \sigma_b^2)$$

Let us assume a noninformative Multivariate-Gaussian prior on β with parameters:

$$\mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ and covariance matrix S} = \begin{pmatrix} 100^2 & 0 & 0 \\ 0 & 100^2 & 0 \\ 0 & 0 & 100^2 \end{pmatrix}$$

The log posterior distribution is then,

$$logp(\beta | \bigcup_x \bigcup_j logit(b_j^x)) \propto logp(\beta|\mu, S) + \sum_x \sum_j logp(logit(b_j^x)|\beta)$$
$$\propto -\frac{1}{2}\beta^T S^{-1}\beta - \frac{1}{2\sigma_b^2}\sum_x \sum_j (logit(b_j^x) - \beta^T X)^2$$

Assume X has the shape (3,N), β has the shape (3,1) and $L = \bigcup_x \bigcup_j logit(b_j^x)$ has the shape (N,1) Then the equation above can be vectorized to,

$$= -\frac{1}{2}[\beta^T S^{-1}\beta + \frac{1}{\sigma_b^2}\{\beta^T X^T X\beta - 2L^T X\beta + L^T L\}$$

Matching that with the density of a multivariate Gaussian with Covariance matrix Σ and mean vector μ

$$-\frac{1}{2}(\beta - \mu)^T \Sigma^{-1}(\beta - \mu)$$

We get the following parameters for the posterior multivariate Gaussian of β

$$\text{Precision Matrix: } \Sigma^{-1} = \sigma_b^{-2} X^T X + S^{-1}$$

$$\mu = \sigma_b^{-2} \Sigma X^T L$$

---

4 Implementation

- Please refer to the code files.
I created a loop to sample all the conditional posteriors, one at a time using Gibbs sampling for most of them. Since I implemented it in R, it wasn't possible to define random variables so I defined a function for each posterior, prior and likelihood that acts like a random variable.

Using Gelman-Rubin approach in section 1.3.3 I evaluated the following metrics for the samples. Note that I sampled 15000 samples for each parameter and for 3 Markov chains. I discarded the first half of the samples as discussed earlier and divided the resulting samples in two. I ended up with 6 chains and 3750 each.

Parameter	mean	Estim var	low.bound	upp.bound	R_hat	n_eff	MC SE
Beta 0	0.00257	0.00050	0.00213	0.00300	1.570	10	0.00704
beta f	-0.04932	0.00002	-0.04936	-0.04928	2.829	7	0.00159
beta d	-0.04303	0.00114	-0.04358	-0.04248	1.913	8	0.01175
sigmaS.b	0.00008	0.00000	0.00008	0.00008	1.000	36818	0.00000
alpha	3.23328	0.01265	3.22969	3.23688	1.000	14229	0.00094
sigmaS.delta	0.53555	0.01341	0.53184	0.53926	1.000	16387	0.00090
a.t	8.48976	0.37501	8.47041	8.50912	1.006	522	0.02682
b.t	14.17313	0.58179	14.14907	14.19720	1.006	528	0.03320
alpha.X.1	2.43828	0.13025	2.42673	2.44983	1.000	16449	0.00281
alpha.X.2	2.18274	0.09349	2.17296	2.19253	1.000	23315	0.00200
alpha.X.3	2.45972	0.14180	2.44767	2.47177	1.000	11668	0.00349
tauS.X.1	2.82817	0.55260	2.80438	2.85195	1.000	13827	0.00632
tauS.X.2	2.25119	0.31178	2.23332	2.26906	1.000	21587	0.00380
tauS.X.3	3.67390	0.82412	3.64485	3.70294	1.000	9914	0.00912

$\hat{R}$ Measures the stationarity of our samples. Those with values < 1.1 are considered stationary. We see that we reached stationarity for all of our parameters except for the betas. We may need to reconsider our linear regression assumption.

Next we observe the mean, estimated variance of sampling the posterior, the lower and upper 95% confidence intervals. We confirm that we have small variability and tight intervals which is what we are looking for.

Finally, we look at the effective sample size which is the number of samples needed to get similar results if our samples were independent. The samples from our method are highly correlated and hence we expect small effective sample size. We again observe that our modeling for the beta might need reconsideration. The Monte Carlo Standard Error (MC-SE) is the square root of the accuracy of average of the simulations as an estimate of the posterior mean. We tend to have small MC-SE which is a good thing.

References

[1] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*, Jul 2020.

[2] Martin Haugh. Mcmc and bayesian modeling. 2017.

[3] Daniel Fink. A compendium of conjugate priors. 1997.

[4] Kevin Murphy. Conjugate bayesian analysis of the gaussian distribution. 11 2007.

[5] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[6] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian data analysis*. Texts in Statistical Science Series. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

[7] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. A bayesian approach for predicting the popularity of tweets. *CoRR*, abs/1304.6777, 2013.