STAT 222 Homework #4 **Mary Silva**
UCSC 12 June 2020

1. Assuming the truncated location normal DP mixture model:

$$\begin{aligned} y_i|\boldsymbol{Z}, L_i, \phi &\overset{ind}{\sim} N(y_i|Z_{L_i}, \phi^{-1}) && i = 1, ..., n \\ L_i|\boldsymbol{V} &\overset{iid}{\sim} \prod_{\ell=1}^{N} (p_\ell(\boldsymbol{V}))^1(L_i = \ell) && i = 1, ..., n \\ Z_\ell &\overset{iid}{\sim} N(Z_\ell|m, s^2) && \ell = 1, ..., N \\ V_\ell &\overset{iid}{\sim} Beta(V_\ell|1, \alpha && \ell = 1, ..., N-1 \\ \phi &\sim Gamma(\phi|a_\phi, b_\phi) \end{aligned}$$

where the variance of the normal mixture kernel is ϕ^{-1}, $Gamma(a, b)$ denotes the gamma distribution with mean a/b and

$$p_1(\boldsymbol{V}) = V_1 \qquad p_\ell(\boldsymbol{V}) = V_\ell \prod_{r=1}^{\ell-1}(1 - V_r), \ell = 2, ..., N-1 \qquad p_N(\boldsymbol{V}) = \prod_{r=1}^{N-1}(1 - V_r)$$

The parameters of the DP centering distribution (m, s^2), and the DP precision parameter, α, are fixed, such that the full parameter vector is $\boldsymbol{\theta} = (\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{L}, \phi)$, with $\boldsymbol{Z} = (Z_1, ..., Z_N), \boldsymbol{V} = (V_1, ..., V_{N-1})$, and $\boldsymbol{L} = (L_1, ..., L_n)$.

We implement a mean-field variational method, using the variational approximation:

$$q_\eta(\boldsymbol{\theta}) = q_\beta(\phi) \prod_{\ell=1}^{N-1} q_{\gamma_\ell}(V_\ell) \prod_{\ell=1}^{N} q_{\xi_\ell}(Z_\ell) \prod_{i=1}^{n} q_{\boldsymbol{\pi}_i}(L_i)$$

where $q_\beta(\phi) = Gamma(\phi|\beta_1, \beta_2)$, $q_{\gamma_\ell} = Beta(\gamma_{\ell 1}, \gamma_{\ell 2})$, $q_{\xi_\ell} = N(\xi_{\ell 1}, \xi_{\ell 2})$, and $q_{\boldsymbol{\pi}_i} = \prod_{l=1}^{N} \pi_{i\ell}^{1(L_i=\ell)}$.

2. We consider the data on the incidence of faults in the manufacturing of rolls of fabric, the first 5 rows are in table 1.

length	faults
551	6
651	4
832	17
375	9
715	14
868	8

Table 1: Data: incidence of faults in fabric.

a) We first consider a Poisson regression model, where y_i are assumed to arise independently, given parameters $\theta > 0$ and $\beta \in \mathcal{R}$, from Poisson distributions with means $E(y_i|\beta, \theta) = \theta \exp(\beta x_i)$, such that $\log(\theta)$ is the intercept and β is the slope of a linear regression function under a logarithmic transformation of the Poisson means. The Bayesian model is completed with priors on β and θ:

$$\begin{aligned} y_i|x_i, \theta, \beta &\sim Poisson(\theta \exp(\beta x_i)) \\ \theta &\sim Gamma(\theta_a, \theta_b) \\ \beta &\sim Normal(\mu, \tau^2) \end{aligned}$$

For this problem, we set $\mu = 0$ and $\tau^2 = 1$, which are non-informative given that the βx_i are exponentiated. We also set $\theta_a = 8$ and $\theta_b = 1$ in order for the posterior mean of the y_i to coincide with the mean of the data. The joint posterior is given as:

$$\prod_{i=1}^{n} \left[f(y_i|x_i, \theta, \beta)\right] \pi(\theta)\pi(\beta) \propto \prod_{i=1}^{n} \left[(\theta \exp \beta x_i)^{yi} \exp\left\{-\theta \exp\left(\beta x_i\right)\right\} \frac{1}{y_i!}\right]$$
$$\times \frac{\theta_b^{\theta_a}}{\Gamma(\theta_a)} \theta^{\theta_a - 1} \exp\left\{-\theta_b \theta\right\} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{\frac{1}{2\tau^2}\left(\beta - \mu\right)^2\right\}$$

The full conditional for $\theta|\cdot$ is a Gamma distribution, but for β, the full conditional is not known. Therefore, we use a tuned Metropolis-Hastings algorithm with multivariate Normal joint proposal distribution on θ and β. Figures 1 and 2, we see that there are no problems with convergence. Furthermore, the acceptance ratio is approximately 0.245.

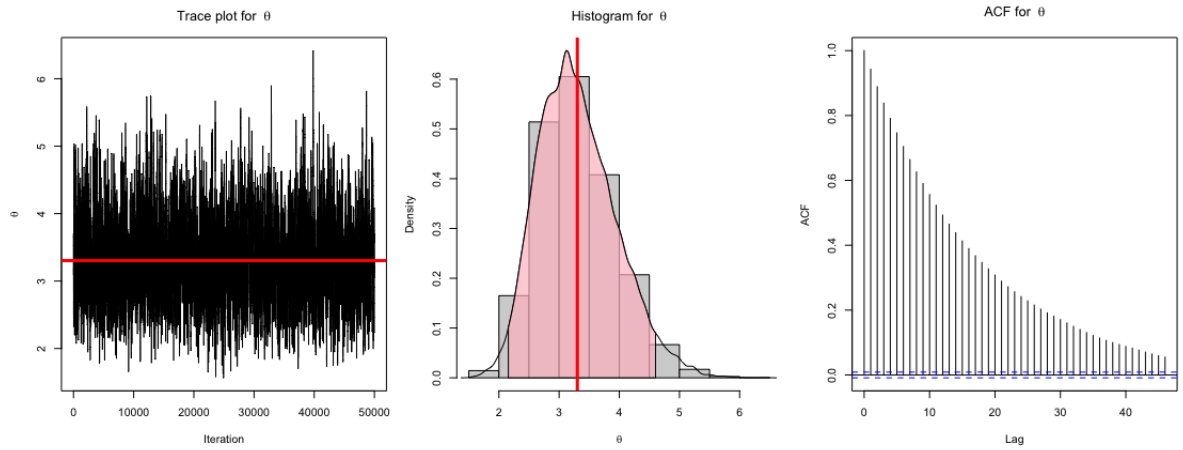

Figure 1: **Left:** Trace, **Middle:** Histogram with 95% credible intervals shaded in pink, Right: ACF for θ.

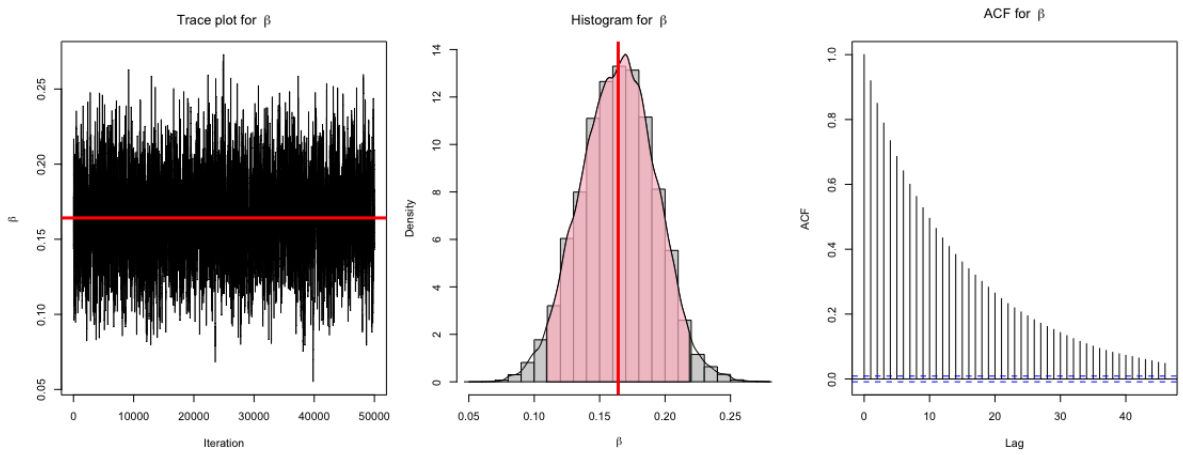

Figure 2: **Left:** Trace, **Middle:** Histogram with 95% credible intervals shaded in pink, Right: ACF for β.

The parameters are summarized in table 2 below. And the posterior predictions are plotted in figure 3.

Parameter	2.5%	50%	97.5%
θ	2.22	3.24	4.71
β	0.11	0.16	0.22

Table 2: Summary for parameters θ and β.

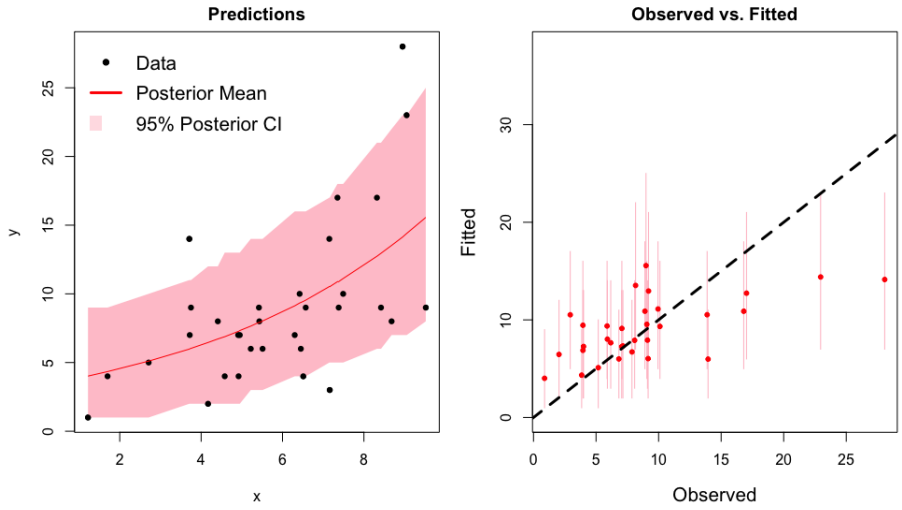

Figure 3: **Left:** Model predictions for each x_i with the 95% posterior credible interval shaded in pink. The red line represents the mean of the predictive means. **Right:** Plot of the observed versus fitted values. The pink bars indicate the 95% predictive regions. A good model would have the dotted lined contained within the bars, which indicates observed equaling the fitted values.

b) Next, we consider an extension in a hierarchical fashion to allow for over-dispersion relative to the Poisson response distribution. In particular, the response distribution can be extended to a negative Binomial under the following hierarchical structure such that the mean of the Gamma distribution for the θ_i is μ and the variance is μ^2/ζ. Under this hierarchical model, $E(y_i|\beta,\mu,\zeta) = \mu\exp(\beta x_i)$ and $V(y_i|\beta,\mu,\zeta) > \mu\exp(\beta x_i)$, thus achieving over-dispersion relative to the Poisson regression model. In this case, the Bayesian model is completed with priors on β, μ, and ζ.

$$
\begin{aligned}
y_i|\theta_i,\beta &\overset{ind}{\sim} Poisson(y_i|\theta_i\exp(\beta x_i)), && i = 1,...n \\
\theta_i|\mu,\zeta &\overset{iid}{\sim} Gamma(\theta_i|\zeta,\zeta\mu^{-1}), && i = 1,...,n \\
\zeta &\sim Gamma(a_\zeta, b_\zeta) \\
\mu &\sim Gamma(a_\mu, b_\mu) \\
\beta &\sim Normal(\mu,\tau^2)
\end{aligned}
$$

In a similar fashion, we apply a Metropolis-Hastings algorithm with a multivariate Normal proposal joint proposal on θ, β, μ, and ζ. We let $\zeta \sim Ga(1, 1/4)$ and $\mu \sim Ga(1, 1/4)$, which leads to a prior mean and prior variance for each θ_i to be approximately 4 and 16, respectively. Also, we use the same hyper parameters for β as part a (i.e. $\beta \sim N(0, 1)$).

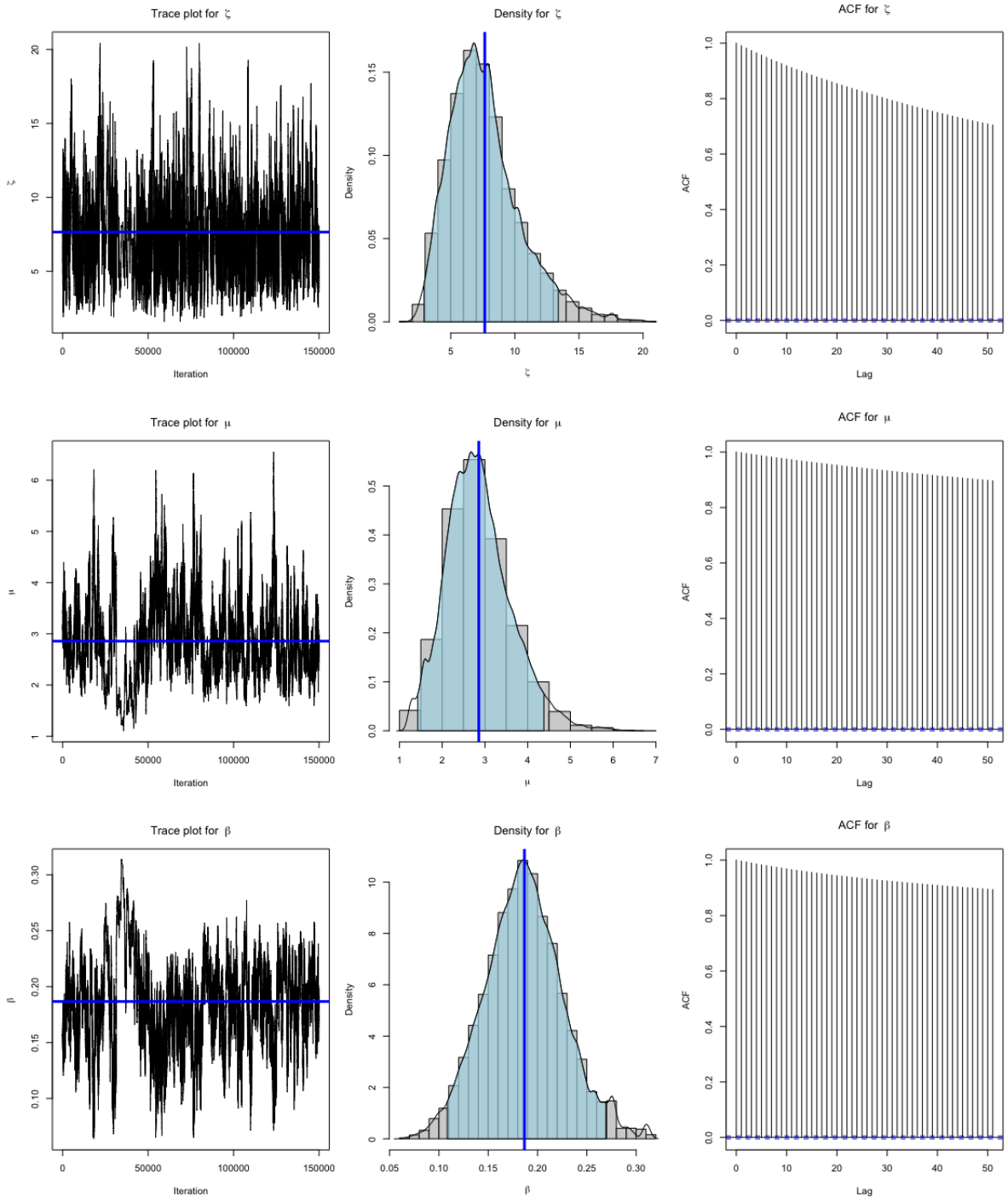

Figure 4: Trace, histogram/density, and ACF plots for ζ (top), μ(center row) and β(bottom).

Figure 4 shows no issues with convergence. The overall acceptance rate was approximately 0.275. We plot the posterior predictions in figure 5. In this case, to make predictions for a new observation y_0 we must also generate a new θ_0 based on the posterior samples for ζ and μ. Furthermore, we do not use the posteriors for each θ_i to make predictions at the x_i. This would imply knowing the random effect for a particular observations. This model assumes the random effect is unknown in order to make a prediction for a new observation.

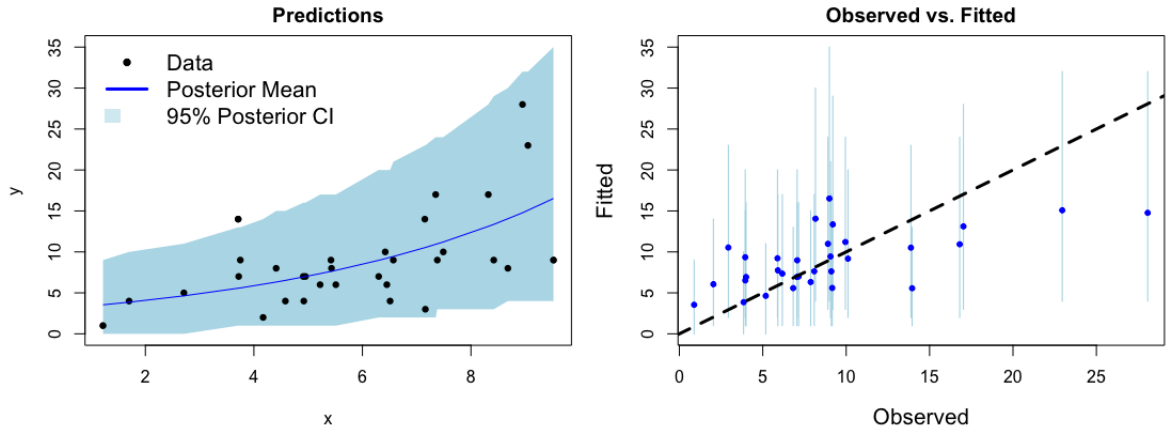

Figure 5: **Left:** Model predictions for each x_i with the 95% posterior credible interval shaded in light blue. The blue line represents the mean of the predictive means. **Right:** Plot of the observed versus fitted values. The blue bars indicate the 95% predictive regions. A good model would have the dotted lined contained within the bars, which indicates observed equaling the fitted values.

c) Last, we develop a semiparametric DP mixture regression model for the count responses y_i, which includes as limiting cases both of the parametric regression models discussed above:

$$
\begin{aligned}
y_i|x_i, \theta_i, \phi &\overset{ind}{\sim} Poisson(y_i|\theta_i \exp(\beta x_i)), && i = 1, ..., n \\
\theta_i|G &\overset{ind}{\sim} G, && i = 1, ..., n \\
G|\alpha, \zeta, \mu &\sim DP(\alpha, G_0 = Gamma(\zeta, \zeta\mu^{-1})) \\
\alpha &\sim Gamma(a_\alpha, b_\alpha) \\
\beta &\sim Normal(\mu, \tau^2) \\
\mu &\sim Gamma(a_\mu, b_\mu) \\
\zeta &\sim Gamma(a_\zeta, b_\zeta)
\end{aligned}
$$

We use the same priors for β, μ, and ζ. In addition, we let $\alpha \sim Ga(1, 1/2)$.

Expressions for the full conditionals (Mostly copied from hw 2, so please excuse any incorrect parameters). We derive the expressions for the full conditionals of the Pòlya urn based Gibbs sampler, which can be used for posterior simulation from $p(\boldsymbol{\theta}, \alpha, \beta, \mu, \zeta|data)$, where $data = \{y_i : i = 1, ..., n\}$.

A key property for the implementation of the Gibbs sampler is the discreteness of G, which includes a partitioning of the θ_i. From the lecture notes we use the following notation

- n^*: the number of distinct elements(clusters) in the vector $(\theta_1, ..., \theta_n)$.
- $\theta_j^*, j = 1, ..., n^*$: the distinct θ_i
- $\boldsymbol{w} = (w_1, ..., w_n)$ is the vector of configuration indicators, defined by $w_i = j$ if and only if $\theta_i = \theta_j^*, i = 1, ..., n$

- n_j is the size of the j^{th} cluster, i.e. $n_j = |\{i : w_i = j\}|, j = 1, ..., n^*$.

The vectors $(n^*, \boldsymbol{w}, \theta_1^*, ..., \theta_{n^*}^*)$ and $(\theta_1, ..., \theta_n)$ are equivalent.

For each $i = 1, ..., n$, $p(\theta_i | \{\theta_{i'} : i' \neq i\}, \alpha, \mu, \beta, \zeta, \boldsymbol{y})$ is simply a mixture of n^{*-} point masses and the posterior for θ_i based on y_i,

$$\frac{\alpha q_0}{\alpha q_0 + \sum_{j=1}^{n*^-} n_j^- q_j} h(\theta_i | \mu, \tau^2, \phi, y_i) + \sum_{j=1}^{n*^-} \frac{n_j^- q_j}{\alpha q_0 + \sum_{j=1}^{n*^-} n_j^- q_j} \delta_{\theta_j^*}(\theta_i)$$

For the purposes of coding, let $A = \frac{\alpha q_0}{\alpha q_0 + \sum_{j=1}^{n*^-} n_j^- q_j}$ and $B_j = \frac{n_j^- q_j}{\alpha q_0 + \sum_{j=1}^{n*^-} n_j^- q_j}$, and

- $q_j = f(y_i | \theta_j^*, \beta)$
- $q_0 = \int f(y_i | \theta_i, \beta) g_0(\theta | \zeta, \mu) d\theta$
- $h(\theta_i | \mu, \beta, \zeta) \propto f(y_i | \theta_i, \beta) g_0(\theta_i | \zeta, \mu)$
- g_0 is the density of $G_0 = Gamma(\theta_i | \zeta, \zeta\mu^{-1})$
- The superscript "-" denotes all relevant quantities when θ_i is removed from the vector $\boldsymbol{\theta}$, e.g. $n*^-$ is the number of clusters in $\{\theta_{i'} : i' \neq i\}$

Integrating out θ, we obtain the following

$$\begin{aligned}
q_0 &= \int Poi(y_i | \theta_i exp(\beta x_i) Ga(\theta_i | \zeta, \frac{\zeta}{\mu} d\theta) \\
&= \int \frac{1}{y_i!} \left(\theta_i exp\left(\beta x_i\right)\right))^{y_i} \exp\left\{-\theta_i exp(\beta x_i)\right\} \frac{\zeta\mu^{-1}}{\Gamma(\zeta)} \theta_i^{\zeta-1} \exp\left\{-\zeta\mu^{-1}\theta_i\right\} d\theta \\
&= \int \frac{1}{y_i!} \theta_i^{y_i} \left(\exp(\beta x_i)\right)^{y_i} \exp\left\{-\theta_i \exp(\beta x_i) - \theta_i \zeta\mu^{-1}\right\} \frac{(\zeta\mu^{-1})\zeta}{\Gamma(\zeta)} \theta_i^{\zeta-1} d\theta \\
&= \frac{1}{y_i} \left(\exp(\beta x_i)\right)^{y_i} \frac{(\zeta/\mu)^\zeta}{\Gamma(\zeta)} \int \theta_i^{y_i+\zeta-1} \exp\left\{-\theta_i \left(\frac{\zeta}{\mu} + \exp(\beta x_i)\right)\right\} d\theta
\end{aligned}$$

The integrand is recognized as the kernel for the Gamma distribution.

$$\begin{aligned}
&= \frac{1}{y_i! \Gamma(\zeta)} \left(\exp(\beta x_i)\right)^{y_i} (\zeta/\mu)^\zeta \times \frac{\Gamma(y_i + \zeta)}{\left(\frac{\zeta}{\mu} + \exp(\beta x_i)\right)^{y_i+\zeta}} \\
&= \frac{\Gamma(y_i + \zeta)}{y_i! \Gamma(\zeta)} \frac{\left(\exp(\beta x_i)\right)^{y_i}}{\left(\zeta/\mu + \exp(\beta x_i)\right)^{y_i+\zeta}} (\zeta/\mu)^\zeta \\
&= \frac{(y_i + \zeta - 1)!}{(\zeta - 1)! y_i!} \frac{(\zeta/\mu)^\zeta \left(\exp(\beta x_i)\right)^{y_i}}{\left(\zeta/\mu + \exp(\beta x_i)\right)^{y_i+\zeta}} \\
&= \binom{y_i + \zeta - 1}{y_i} \frac{(\zeta/\mu)^\zeta (\exp(\beta x_i))^{y_i}}{\left(\zeta/\mu + \exp(\beta x_i)\right)^{y_i} \left(\zeta/\mu + \exp(\beta x_i)\right)^\zeta}
\end{aligned}$$

Which reduces to a Negative binomial$\left(\zeta, \frac{\zeta/\mu}{\exp(\beta x_i) + \zeta/\mu}\right)$. Furthermore, we found $h(\theta_i | \mu, \beta, \zeta) \propto Gamma(\zeta, \exp(\beta x_i) + \zeta/\mu)$. The full conditional for α is derived as in homework 3 problem 2. We update β in the MCMC with a normal proposal distribution, and a joint multivariate normal distribution for ζ and μ.

The trace, densities, and ACF plots for ζ, μ, α, and β are shown in figure 6.

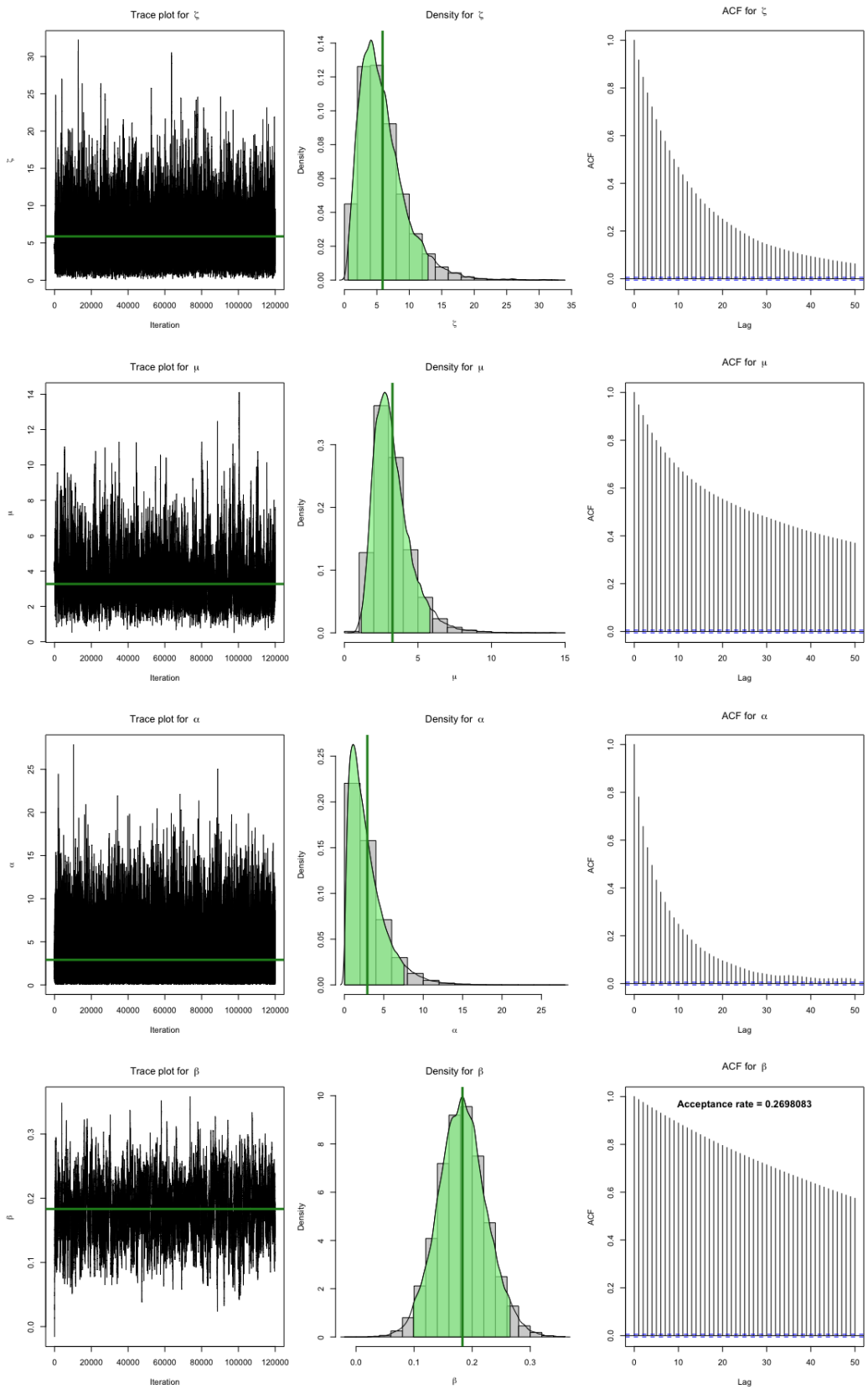

Figure 6: Self-explanatory at this point, sorry. For (ζ, μ), the mean acceptance ratio for the joint proposal is 0.289. The mean acceptance ratio for β is 0.27.

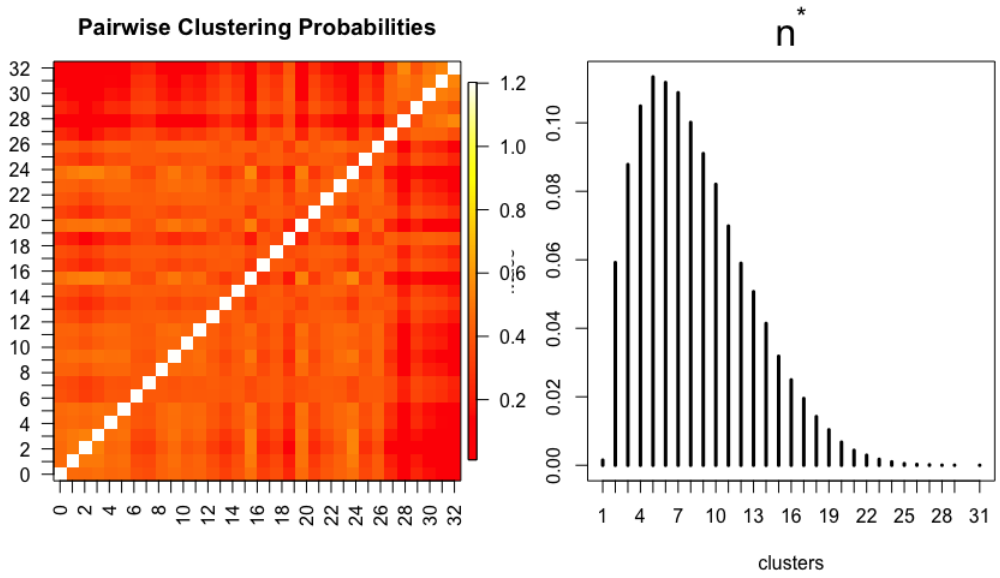

Figure 7: Left: Pairwise Clustering probabilities. Right: (Right) Posterior for n^* with prior $\alpha \sim Gamma(1, 0.5)$.

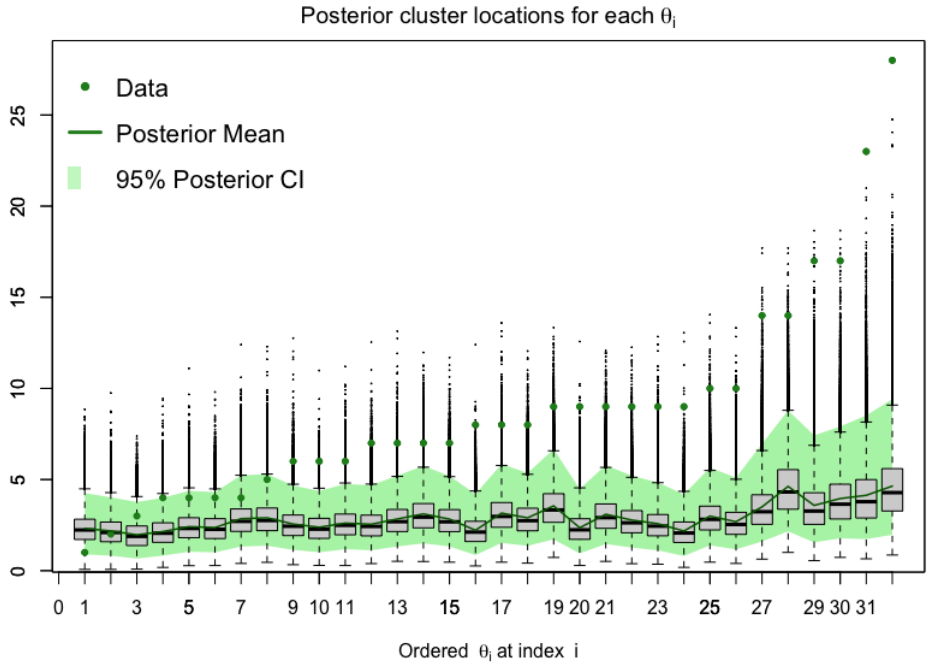

Figure 8: Boxplots of the (ordered) θ_i's with their corresponding 95% posterior credible intervals.

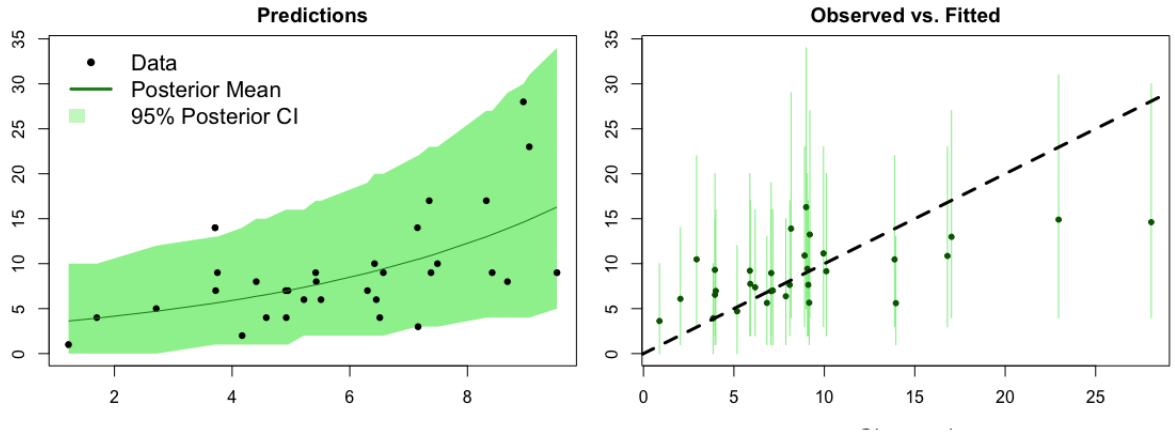

Figure 9: **Left:** Model predictions for each x_i with the 95% posterior credible interval shaded in light green. The green line represents the mean of the predictive means. **Right:** Plot of the observed versus fitted values. The green bars indicate the 95% predictive regions. A good model would have the dotted lined contained within the bars, which indicates observed equaling the fitted values.

The predictions at each x_i are comparable to the hierarchical model's. Both model (b) and model (c) visually appear to be an improvement over model (a) based on Figure 9.

To compare the 3 models, I used the Gelfand and Ghosh criterion, which is defined as follows: Let $\boldsymbol{y}$ denote the observed data. Let $\mu_\ell = E(z_\ell|\boldsymbol{y})$ and $\sigma^2_\ell = Var(z_\ell|\boldsymbol{y})$. Then let $G = \sum_l (\mu_\ell - y_\ell)^2$, and $P = \sum_\ell \sigma^2_\ell$. Then the Gelfand and Ghosh criterion is defined as $D = G + P$, where D seeks to reward goodness of fit penalizing complexity. So the smaller the D, the better the model. The GG criterion is summarized for each model in table 3.

Model	G	P	D = G+P
Model (a)	351.875	97.6825	449.5575
Model (b)	365.9388	150.0788	516.0176
Model (c)	339.5346	145.8948	485.4294

Table 3: Gelfand and Ghosh criterion for each model.

For all 3 models, the goodness of fit term is relatively close. For model (a), there is variance and less penalty, obviously because the model is less complicated. But visually, model (c) did better at capturing the observed data based on the 95% posterior predictive intervals.