

Bayesian Mixture Models for Complex High-Dimensional Count Data in Phage Display Experiments

Yuan Ji¹, Guosheng Yin¹, Kam-Wah Tsui⁴, Mikhail G. Kolonin², Jessica Sun²,
Wadih Arap^{2,3}, Renata Pasqualini^{2,3}, Kim-Anh Do¹

¹Department of Biostatistics and Applied Mathematics, ²Department of Genitourinary Medical Oncology, ³Department of Cancer Biology, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

⁴Department of Statistics, University of Wisconsin – Madison, Madison WI 53706, USA

Email: *yuanji@mdanderson.org*.

ABSTRACT: Phage display is a biological process used to screen random peptide libraries for ligands that bind to a target of interest with high affinity. Based on a count data set from an innovative multi-stage phage display experiment, we propose a class of Bayesian mixture models to cluster peptide counts into three groups that exhibit different display patterns across stages. Among the three groups, the investigators are particularly interested in the one with an ascending display pattern in the counts, which implies that the peptides are likely to bind to the target with strong affinity. We apply a Bayesian false discovery rate approach to identify the peptides with the strongest affinity within the group. A list of peptides is obtained, among which important ones with meaningful functions are further validated by biologists. To examine the performance of the Bayesian model, we conduct a simulation study and obtain desirable results.

KEY WORDS: Bayesian inference; Gibbs sampler; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Peptide.

1 Introduction

This paper introduces a class of Bayesian mixture models for phage experiments studying peptides. A peptide is a sequence of amino acids linked by peptide bonds between a carbon atom of

one and a nitrogen atom of the next. The length of a peptide is its number of amino acids. For example, a tri-peptide has three amino acids. There are 20 different amino acids, which leads to 4200 distinct tri-peptides (not counting mirror images, e.g., ABC is the same as CBA). For quadri-peptides, the number of distinct types is approximately 20 times larger.

Phage display is a technology commonly used to screen random peptide libraries for specific peptides that bind to a target of interest. It is a useful tool in mapping vasculature, drug discovery, and stem cell research (Arap et al., 2002; Clackson and Lowman, 2004; Nowakowski et al., 2004; Pasqualini and Ruoslahti, 1996). The phage display biopanning (Kolonin et al., 2006) is a novel technique to improve the efficiency in the screening process. It consists of three steps: exposure of a peptide-phage library to a target, removal of unbound phages, and recovery and propagation of bound phages by bacterial infection. The completion of all three steps results in a single stage of phage display biopanning. Kolonin et al.(2006) found that a single biopanning stage of a phage library does not sufficiently enrich high-affinity organ-homing peptides, and hence they propose performing successive stages of panning (usually three to four) to enrich peptides that bind to the targets, thus increasing the chance of identifying them.

Kolonin et al.(2006) applied three stages of phage display biopanning to identify homing peptides for multiple tissues in mice. At each of the three stages of phage display, the total number of tri-peptides i from a small pool is obtained by randomly sampling from the phage library harvested from the mouse organ j , $i = 1, \dots, 4200$, $j = 1, \dots, 6$. The six organs of interest are brain, bowel, kidney, muscle, pancreas and uterus. Therefore, at a given stage of phage display, a vector of 4200×6 integers is produced. Let $m = 6$ and $n = 4200$ represent the total numbers of organs and tri-peptides, respectively. Denote $\mathbf{N}_{ij} = (N_{ij0}, N_{ij1}, N_{ij2})$, where N_{ijk} is the count of tri-peptide i from organ j at stage k , $k = 0, 1, 2$. Consequently, the count data can be represented by the vector $\mathbf{N} = (\mathbf{N}_{ij}, i = 1, \dots, n; j = 1, \dots, m)$.

In Kolonin et al.(2006), a Bayesian beta-binomial model was used to found extended peptides shared among those isolated from a given organ. Alternatively, one can consider all the counts

at a given stage as a two-way contingency table, which has been extensively studied by Agresti (2002), Leonard and Hsu (1999), Raftery (1997), Spiegelhalter and Smith (1982), and Albert (1997), among others. When multiple tables are involved, independence is usually assumed for measurements across different tables. However, due to the sequential nature of the phage display experiment, the counts of the same peptide in the same organ across different stages are apparently correlated. In addition, since the tables are of high dimensions, statistical models for these tables often involve a large number of unknown parameters. One approach of accounting for correlations among measurements and reducing the number of parameters is via Bayesian hierarchical modeling. Specifically, we propose a class of mixture models for analyzing the phage data with the following considerations. Since the recovered phage library at each stage was enriched before it was re-injected into the mouse for the next stage, the tri-peptide counts from the phage library were expected to increase across the multiple stages of the experiment if they were to bind to specific organs with high affinity. Likewise, the counts for the tri-peptides that did not bind would stay unchanged or even decrease, since only a certain number of peptides could bind to an organ. To capture these specific features in the data, we model the three counts of each peptide as three Poisson random variables and assume that the log Poisson means are expressed as a linear function of the stage index k . Therefore, the sign of the slope in the linear function indicates whether the counts of the peptide decrease or increase over stages. For the prior distribution of the slope parameters, we propose a mixture of three normal distributions representing different trends in the display patterns of the peptides. Through posterior inference, we identify organ-specific tri-peptides that exhibit an ascending trend across consecutive stages. We use a Bayesian false discovery rate (FDR) approach (Newton et al., 2004) to select the tri-peptides binding to organs with high affinity. Posterior computation is based on a Markov chain Monte Carlo (MCMC) algorithm combining a Gibbs sampler with a Metropolis-Hastings step (Geman and Geman, 1984; Hastings, 1970; Metropolis et al., 1953).

The remainder of the article is organized as follows. We propose the Bayesian mixture model

in Section 2. We develop an MCMC simulation scheme for obtaining random samples from the posterior distributions in Section 3, and illustrate the application of the proposed models and the FDR procedure using real data in Section 4. We perform a simulation study to further assess the properties of our models in Section 5, and provide some concluding remarks in Section 6.

2 Probability model

2.1 Likelihood

Suppose that, conditional on the parameters (μ_{ij}, β_{ij}) , the count of phage i from organ j at stage k

$$N_{ijk} | \mu_{ij}, \beta_{ij} \sim Poi(\mu_{ij} e^{k\beta_{ij}}), \quad (1)$$

where $Poi(x)$ denotes a Poisson distribution with mean x . Considering model (1) as a Poisson regression on stage index k , we can see that $\log \mu_{ij}$ is the intercept, and β_{ij} is the slope for the covariate k . A positive value of β_{ij} indicates an ascending trend in the three consecutive counts, thus implying a strong association between tri-peptide i and organ j .

Let $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ be vectors representing all the parameters μ_{ij} and β_{ij} , respectively. The likelihood function for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ is given by

$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\beta} | \mathbf{N}) &= \prod_{i=1}^n \prod_{j=1}^m \prod_{k=0}^2 p(N_{ijk} | \mu_{ij}, \beta_{ij}) \\ &= \prod_{i=1}^n \prod_{j=1}^m \frac{e^{-\mu_{ij}(1+e^{\beta_{ij}}+e^{2\beta_{ij}})} \mu_{ij}^{N_{ij0}+N_{ij1}+N_{ij2}} e^{\beta_{ij}(N_{ij1}+2N_{ij2})}}{N_{ij0}! N_{ij1}! N_{ij2}!}. \end{aligned} \quad (2)$$

2.2 Priors

We propose a mixture of three normal distributions as the prior of β_{ij} that corresponds to three different display patterns: an ascending pattern, a descending pattern, and an oscillating pattern around a constant.

Let $\boldsymbol{\lambda}_{ij} = (\lambda_{ij1}, \lambda_{ij2}, \lambda_{ij3})$ and assume

$$\boldsymbol{\lambda}_{ij} \sim \text{Multi}(1; \pi_1, \pi_2, \pi_3),$$

which is a multinomial distribution with one draw from three components with probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$. Conditional on $\boldsymbol{\lambda}_{ij}$, the prior density of β_{ij} is given by

$$p(\beta_{ij} | \boldsymbol{\lambda}_{ij}) = \phi(\beta_{ij} | s_1, \tau_1^2)^{\lambda_{ij1}} \phi(\beta_{ij} | s_2, \tau_2^2)^{\lambda_{ij2}} \phi(\beta_{ij} | s_3, \tau_3^2)^{\lambda_{ij3}},$$

where $\phi(\cdot | s, \tau^2)$ denotes the density of a normal distribution with mean s and variance τ^2 .

The prior density of β_{ij} is given by

$$p(\beta_{ij}) = \pi_1 \phi(\beta_{ij} | s_1, \tau_1^2) + \pi_2 \phi(\beta_{ij} | s_2, \tau_2^2) + \pi_3 \phi(\beta_{ij} | s_3, \tau_3^2). \quad (3)$$

Furthermore, we propose hierarchical priors for the means, s_1 and s_3 (while fixing $s_2 = 0$), and for the variances, τ_1^2 , τ_2^2 and τ_3^2 . Specifically, we take the densities of s_1 and s_3 to be

$$s_1 \sim N(m_1, \eta_1^2) \quad \text{and} \quad s_3 \sim N(m_3, \eta_3^2),$$

where m_1 and m_3 are fixed positive and negative scalars, respectively. For the variances τ_1^2 , τ_2^2 and τ_3^2 in (3), we assume that their priors are independent and follow the same inverse gamma (*IG*) distribution, i.e.,

$$\tau_l^2 \sim IG(a_\tau, b_\tau), \quad l = 1, 2, 3,$$

where a_τ and b_τ take positive values. We assume a Dirichlet (*Dir*) prior for $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi} \sim \text{Dir}(\pi_{1,0}, \pi_{2,0}, \pi_{3,0}),$$

and $(\pi_{1,0}, \pi_{2,0}, \pi_{3,0})$ are fixed positive scalars. We specify the prior distribution for the baseline count parameter μ_{ij} to be

$$\mu_{ij} \sim \mu_0 G(\alpha, 1/\alpha), \quad \alpha > 0,$$

where $G(a, b)$ denotes a gamma distribution with mean ab . Finally, we let

$$\mu_0 \sim IG(a_{\mu_0}, b_{\mu_0}),$$

where a_{μ_0} and b_{μ_0} are positive scalars.

In Section 4, we will assign appropriate values to these hyperparameters based on prior information provided by the investigators. Note that the priors for s_1 , s_3 , and τ_l^2 ($l = 1, 2, 3$) cannot be improper; otherwise, their posterior distributions would also be improper (McLachlan and Peel, pp. 125-126, 2000; Wasserman, 2000). Due to this result, we carefully check the convergence of MCMC for the set of priors elicited to ensure the validity of our results.

3 Model fitting using MCMC

The joint posterior distribution is log-concave, facilitating the Gibbs sampling based on the full conditional distributions. Closed forms are available for the full conditionals of all the parameters except for β_{ij} , for which we use the Metropolis-Hastings algorithm to update. Specifically, we divide β_{ij} into three groups based on the current values of λ_{ij} . If $\lambda_{ij} = (1, 0, 0)$, β_{ij} follows the first normal distribution in the mixture (3) with density $\phi(\beta_{ij}|s_1, \tau_1^2)$. Similarly, β_{ij} follows the second or the third normal distribution with density $\phi(\beta_{ij}|s_2, \tau_2^2)$ or $\phi(\beta_{ij}|s_3, \tau_3^2)$ when $\lambda_{ij} = (0, 1, 0)$ or $(0, 0, 1)$, respectively. Therefore, at each MCMC iteration, the set of β_{ij} following the same normal distribution is updated simultaneously, using the Metropolis-Hastings algorithm. We outline the MCMC algorithm in which the parameter values are updated based on the corresponding full conditional distributions. The derivation of the full conditionals is a straightforward and thus omitted.

1. Initialize all parameter values.
2. For each pair (i, j) , sample μ_{ij} , from a gamma distribution.
3. Sample μ_0 from an inverse gamma distribution.
4. Sample each of the λ_{ij} from a multinomial distribution with one draw from three components.

5. Based on the sampling values of λ_{ij} , sample β_{ij} as follows:

- Let $\beta^{(1)} = \{\beta_{ij} : \lambda_{ij} = (1, 0, 0)\}$ denote the vector containing all the β_{ij} such that $\lambda_{ij} = (1, 0, 0)$. Therefore, these β_{ij} follow the first normal distribution in the mixture prior with density $\phi(\beta_{ij} | s_1, \tau_1^2)$. Similarly, let $\beta^{(2)} = \{\beta_{ij} : \lambda_{ij} = (0, 1, 0)\}$ and $\beta^{(3)} = \{\beta_{ij} : \lambda_{ij} = (0, 0, 1)\}$.
- The full conditional of each $\beta^{(l)}$, $l = 1, 2, 3$, is the product of the full conditionals of its components. Therefore, the β_{ij} within each $\beta^{(l)}$ are updated simultaneously using the random walk Metropolis-Hastings method (Gilks, Richardson, and Spiegelhalter 1991).

6. Sample s_1 and s_3 from normal distributions.

7. Sample τ_1^2 , τ_2^2 , and τ_3^2 from inverse gamma distributions.

8. Sample π from a Dirichlet distribution.

9. Repeat steps 2-8 until the Markov chain converges.

4 A case study of multi-stage phage display

We implement the Bayesian mixture models and the proposed MCMC algorithm for the phage display data. The fundamental problem is to identify the combination of tri-peptides and organs that have ascending display patterns in their counts, i.e., to identify the tri-peptide i with relatively large positive slopes β_{ij} for organ j .

4.1 Bayesian FDR

Following Newton et al.(2004), we use the posterior probability

$$\xi_{ij} = \Pr(\beta_{ij} > 0 | \mathbf{N}),$$

as the primary criterion for selecting tri-peptides. The posterior probabilities are sorted and a “cutoff” value κ_δ is used to threshold probabilities with an FDR value δ .

Consider the selection of tri-peptides under a hypothesis-testing framework where the null hypothesis $H_{ij}^{(0)} : \beta_{ij} \leq 0$ is tested against the alternative hypothesis $H_{ij}^{(1)} : \beta_{ij} > 0$. Note that the posterior probability ξ_{ij} is equivalent to the posterior probability of the alternative, $\Pr(H_{ij}^{(1)} | \mathbf{N})$. We rank the tri-peptides according to increasing values of $(1 - \xi_{ij})$, and select a list S containing tri-peptide and organ pairs with $(1 - \xi_{ij}) \leq \kappa$, for some bound κ . The posterior expected number of false discoveries is then

$$FD(\kappa) = \sum_{i=1}^n \sum_{j=1}^m (1 - \xi_{ij}) I_{(1 - \xi_{ij} \leq \kappa)}. \quad (4)$$

For a target FDR value δ , we can find a data-driven cutoff κ_δ , the largest value among all κ satisfying

$$\frac{FD(\kappa)}{||S||} \leq \delta,$$

where $||S||$ is the size of the list.

Below is the proposed selection procedure based on this FDR approach.

- Obtain posterior samples for all values of β_{ij} . For each combination of i and j , estimate ξ_{ij} using the posterior samples from the MCMC iterations.
- Rank the tri-peptide and organ pairs (i, j) according to increasing values of $(1 - \xi_{ij})$.
- Starting from the highest ranking pair (i, j) with the smallest value of $(1 - \xi_{ij})$, move down to the G th highest ranking pair and stop when $S_G/G > \delta$ is first reached, where S_G is the sum of $(1 - \xi_{ij})$ for all the G pairs.

For various values of the target FDR δ , we obtain different sets of tri-peptides and report these sets to the investigators.

We note that FDR is an approach for adjusting multiplicity. Alternative methods can also be employed, such as directly thresholding posterior probabilities under a rigorous Bayesian

hypothesis testing framework (Müller et al., 2004). When communicating with collaborative investigators, we report FDR to the investigators as an estimate of the error rate for the list of peptides we provide.

4.2 Results

We performed the following prescreening procedure before applying the proposed models to the phage display data. We first computed $C_{ij} = \sum_{k=0}^2 N_{ijk}$, the sum of the observed peptide counts for each tri-peptide and organ pair (i, j) over the three stages. We excluded the tri-peptide and organ pairs with $C_{ij} \leq 3$. The counts of these tri-peptides were too small and the investigators were not interested in them. Nevertheless, we implemented our method for all the tri-peptides and our models performed well. Here, we describe the detailed model-fitting procedure for the 257 peptide and organ pairs satisfying $C_{ij} > 3$.

Wasserman (2000) showed that one cannot use improper “flat” priors for the parameters in the mixture components because they will lead to improper posteriors. The parameter values in our proposed priors are elicited by consulting with the investigators and by using the information in the data. According to the investigators, most of the phage counts are small in the initial stage. Therefore, we center μ_0 , the parameter representing the initial phage counts, at 1. Specifically, we take $a_{\mu_0} = 3$ and $b_{\mu_0} = 0.5$ as parameters for the inverse gamma prior for μ_0 with mean 1 and variance 1. We take $\alpha = 0.1$ to induce a large variance of 10 for μ_{ij} . We set $m_1 = -1$, $m_3 = 1$, $\eta_1^2 = 0.3^2$ and $\eta_3^2 = 0.3^2$. Hence the normal priors of the means s_1 and s_3 are centered around -1 and 1 , respectively. We let $\pi_{1,0} = \pi_{2,0} = \pi_{3,0} = 50$, and let the inverse gamma priors of τ_1^2 and τ_3^2 be centered around 0.2 . The prior for τ_2^2 is set to be an inverse gamma distribution with mean 1 and variance 1. Based on these priors, the induced marginal prior for the key parameter β_{ij} is shown in Figure 1. The induced prior of β_{ij} places most mass between -2 to 2 , and is flat between -1 and 1 . According to the Poisson model (1), since the mean phage

count equals $\mu_{ij}e^{k\beta_{ij}}$, the difference in the mean phage counts between two adjacent stages k and $k + 1$ equals $e^{(k+1)\beta_{ij}} - e^{k\beta_{ij}}$ (assuming $\mu_{ij} = 1$, its prior mean). Therefore, the shape of the prior density implies how large the differences of two mean phage counts could be before observing the phage count data. We confirmed with our collaborative investigators that the proposed priors were reasonable by explaining these implications to them. Figure 1 shows that posteriors of β_{ij} are mainly influenced by the likelihood. Compared with the prior of β_{ij} , the posterior of a particular pair (i, j) with counts $(1, 0, 6)$ places most mass at values larger than 1. We examined the posteriors for all the pairs and they all seemed to agree with the observed counts.

We performed a sensitivity analysis by exploring alternative prior specifications for our model. In Figure 2, we use a different set of parameter values and obtain a prior of β_{ij} that places most mass between -4 and 4 and is flat between -2 and 2 . The alternative prior is more variable. The corresponding posterior still indicates that β_{ij} is positive with a large probability, exhibiting more variability than the one depicted in Figure 1. We subsequently selected a set of phages based on the FDR procedure using the posterior corresponding to each of the two prior distributions. The two sets of selected phages are very similar because the ranks of phage-organ pairs based on both models are very similar. We proceed to report complete results using the prior in Figure 1.

We iterated the Markov chain 50000 times and obtained a posterior sample value every five iterations after a burn-in of 5000 samples. We performed convergence assessment for the proposed parameter values to ensure that MCMC converged well. Based on the convergence criteria in Cowles and Carlin (1996), we found that the Markov chains mixed very well and converged rapidly.

Figure 3 contains the three clusters (indicated by three colors) of the 257 pairs of tri-peptides and organs. The blue cluster includes tri-peptide and organ pairs with ascending patterns in their display. The light blue and the green clusters contain the tri-peptide and organ pairs with non-ascending patterns, i.e., the values of their counts over the stages either oscillate or

decrease in general. The cluster label for each pair (i, j) is determined by computing the posterior probability $\Pr(\lambda_{ijl} = 1 | \mathbf{N})$ for $l = 1, 2, 3$, and assigning the pair to cluster l with the largest posterior probability. We implemented the Bayesian FDR selection procedure in Section 4.1 to identify tri-peptide and organ pairs with the sharpest ascending patterns in their counts. We found 22, 40, and 62 pairs with an FDR value of 0.05, 0.10, and 0.15, respectively. In addition, Figure 4 presents the posterior means and the 95% HPD intervals for all the β_{ij} . Three HPD intervals do not contain zero, indicating that only three pairs have significantly positive slopes. It appears that the 95% HPD interval procedure is more conservative than the FDR procedure with $\delta = 0.05$.

Table 2 contains the list of 62 peptide and organ pairs selected with an FDR $\delta = 0.15$. From this table, more than 10 candidate mouse proteins mimicked by tissue-specific peptides have been verified (Kolonin et al., 2006).

5 Simulation

We conducted a simulation study to investigate the performance of the proposed mixture model. We simulated data with sparse counts of 27 tri-peptide and organ pairs with the same features as the phage display data. We assumed that there were $i = 1, \dots, 9$ tri-peptides and $j = 1, 2, 3$ organs. For each pair (i, j) , we generated three Poisson counts over three sequential stages. Let $\mathbf{N}_{ij} = (N_{ij0}, N_{ij1}, N_{ij2})$ denote the counts generated using the Poisson model,

$$N_{ijk} | \mu_{ij}, \beta_{ij} \sim Poi(\mu_{ij} e^{k\beta_{ij}}), \quad i = 1, \dots, 9, \quad j = 1, 2, 3, \quad \text{and } k = 0, 1, 2.$$

The values of μ_{ij} and β_{ij} are respectively taken to be 0.5 and 1.0 for pairs $i = 1, \dots, 9$ and $j = 1$; 1.0 and 0.0 for pairs $i = 1, \dots, 9$ and $j = 2$; and 4.0 and -1.0 for pairs $i = 1, \dots, 9$ and $j = 3$. Hence the first nine generated pairs have ascending patterns due to the positive value of β_{ij} , and the middle and the last nine have flat and descending patterns, respectively. We implemented

our Bayesian mixture model for the simulated count data and obtained estimates of the μ_{ij} and β_{ij} . We iterated the process of data generating and model fitting 500 times and recorded the resulting 500 posterior means of the β_{ij} and μ_{ij} . Figure 5 shows that the kernel density estimates of posterior means of the β_{ij} and μ_{ij} . For example, the superimposed nine solid curves in the upper panel represent the nine estimated densities for the 500 posterior means of the slopes β_{ij} from pairs $(1, 1), (2, 1), \dots, (9, 1)$. The true value of β_{ij} for these pairs equals 1. It appears that all the densities are centered around their true values, indicating a satisfactory performance of the proposed models.

6 Discussion

We have developed a class of Bayesian mixture models for analyzing the complex high-dimensional count data from a phage experiment with parallel biopanning. The data are typically of large dimensionality and the counts are correlated. The Bayesian mixture model accurately accounts for the correlation of the tri-peptide counts both within and between stages. It distinguishes different patterns in the display of tri-peptide counts and identifies the outlier tri-peptides that have relatively large counts across stages. Compared to the conventional simple statistical methods used by Kolonin et al.(2006), the new method allows us to study the dynamic trends in the counts across different stages of phage display. The selection of significant phages is based on the value of the slope parameter, instead of the counts. Therefore, we can target the phages with an overall increasing trend in their counts, rather than those simply with a large count at the end of the experiment. The new method also applies a simple Bayesian FDR approach that provides investigators with an estimate of the error rate in the selection procedure.

We use a mixture of three normal distributions to capture the ascending, descending and oscillating patterns in the phage display. The normal mixture priors are flexible in accommodating different shapes of distributions for the slopes β_{ij} . An alternative is to use a mixture of a normal,

a negative gamma, and a gamma distribution,

$$p(\beta_{ij}) = -\pi_1 G(\beta_{ij}|g_1, h_1) + \pi_2 N(\beta_{ij}|s_2, \tau_2^2) + \pi_3 G(\beta_{ij}|g_2, h_2), \quad (5)$$

where g_1 , g_2 , h_1 and h_2 are positive parameters for the gamma distributions in the mixture. To appreciate why, note that the negative gamma and the gamma distributions are disjoint, as opposed to the mixture of normal distributions, so that the slopes β_{ij} belonging to either of the gamma distributions are strictly negative or positive. However, the MCMC convergence with (5) is not satisfactory since the Markov chain can only jump between the gamma and the normal, or between the minus gamma and the normal, but not between the two gamma distributions directly. Thus, the chain must overcome two valleys of low probabilities to move from the minus gamma component to the gamma in the mixture. The chain may get “stuck” in the vicinity of one mode of a gamma component. Normal mixtures are more flexible in allowing a sampler to move freely, and therefore can achieve faster and better convergence.

We use different intercepts μ_{ij} for different pairs of tri-peptides and organs because tri-peptides have different baseline counts due to the sampling variation. A large count at the later stages could be due to either the corresponding tri-peptide binding strongly to the organ, or it having a large baseline count to start with and possibly not binding to the organ. Using different μ_{ij} in the model differentiates these two cases and allows for the identification of the peptides with large slopes β_{ij} . In fact, our simulation results (not shown) report poor estimates of the β_{ij} if we only allow for one common intercept μ for all the (i, j) pairs.

Improper priors are often used as long as the likelihood function is integrable with respect to the parameters (Carlin and Louis, 2000). For mixture models, however, one cannot use improper priors for the parameters in the mixture components since improper priors lead to improper posteriors. In the analysis of phage data, we specified the parameter values in the priors by consulting with the investigators and by using the information contained in the data. We presented simulation results to check the model fitting. From additional simulations, we find

that phage counts generated from $Poi(\mu_{ij}e^{k\beta_{ij}})$ with large k may cause overflow. Additionally, a large k value (say, 15) results in noisy simulated data. Thus, the observed posterior means of β_{ij} vary from simulation to simulation, depending on the data generated. Recall that we only have three data points in the Poisson regression. To capture the true relationship between the counts and covariate k , especially when k spans a wide range, we need more observations. When the values of k are relatively small, the proposed model fits well. We explored different k values such as 0, 0.5, 1.6 in the simulations and obtained similar results.

There are several possible improvements to the phage experiments that would make the proposed model work better. One is to repeat the parallel biopanning and the enrichment procedure for more than three stages. With additional stages, we would have more power to detect the trends in the phage display. Nevertheless, the parallel biopanning method is new and promising biologically, and it takes much effort to implement each stage of phage display. With the data that we were provided, our mixture model seems to be adequate. Furthermore, if in the future additional data were made available with more stages, the proposed model could still be easily applied. Finally, our model can be extended to accommodate functional data. This requires a possibly nonparametric functional form in modeling the Poisson means. Research in this direction may become useful when counts are observed continuously over time.

Acknowledgment

This research was supported in part by the University of Texas SPORE in Prostate Cancer grant CA90270. We thank the Associate Editor and two anonymous referees for their valuable comments of the paper.

References

- Agresti, A. (2002), *Categorical Data Analysis (2nd ed.)*, John Wiley & Sons, New York, NY.
- Albert, J. (1997), “Bayesian testing and estimation of association in a two-way contingency table”, *Journal of the American Statistical Association* **92**, 685–693.
- Arap, W., Kolonin, M. G., Trepel, M., Lahdenranta, J., Cardo-Vila, M., Giordano, R. J., Mintz, P. J., Ardelt, P. U., Yao, V. J., Vidal, C. I., L., C., A., F., Valtanen, H., Weavind, L. M., Hicks, M. E., Pollock, R. E., Botz, G. H., Bucana, C. D., Koivunen, E., Cahill, D., Troncoso, P., Baggerly, K. A., Pentz, R. D., Do, K.-A., Logothetis, C. J. and Pasqualini, R. (2002), “Steps toward mapping the human vasculature by phage display”, *Nature Medicine* **8**, 121–127.
- Carlin, B. and Louis, T. (2000), *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*, Chapman & Hall/CRC.
- Clackson, T. and Lowman, H. (2004), *Phage display – A practical approach*, Oxford University Press.
- Cowles, M. and Carlin, B. (1996), “Markov chain Monte Carlo convergence diagnostics: A comparative review”, *Journal of the American Statistical Association* **91**, 883–904.
- Geman, S. and Geman, D. (1984), “Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.
- Gilks, W., Richardson, S. and Spiegelhalter, D. (1999), *Introducing Markov chain Monte Carlo*, CRC Press.
- Hastings, W. (1970), “Monte Carlo sampling methods using Markov chains and their applications”, *Biometrika* **57**, 97–109.

- Kolonin, M., Sun, J., Do, K.-A., Vidal, C., Ji, Y., Baggerly, K., Pasqualini, R. and Arap, W. (2006), “Synchronized selection of homing peptides for multiple tissues with *in vivo* phage display”, *FASEB Journal* **20**, 979–981.
- Leonard, T. and Hsu, J. (1999), *Bayesian methods: an analysis for statisticians and interdisciplinary researchers*, Cambridge University Press, Cambridge, UK.
- McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, John Wiley & Sons, New York.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953), “Equations of state calculations by fast computing machines”, *Journal of Chemical Physics* **21**, 1087–1091.
- Müller, P., Parmigiani, G., Robert, C. and Rousseau, J. (2004), “Optimal sample size for multiple testing: The case of gene expression micorarrays”, *Journal of the American Statistical Association* **99**, 990–1001.
- Newton, M., Noueir, A., Sarkar, D. and Ahlquist, P. (2004), “Detecting differential gene expression with a semiparametric hierarchical mixture method”, *Biostatistics* **5**, 155–176.
- Nowakowski, G., Dooner, M., Valinski, H., Mihaliak, A., Quesenberry, P. and Becker, P. (2004), “A specific heptapeptide from a phage display peptide library homes to bone marrow and binds to primitive hematopoietic stem cells”, *Stem Cells* **22**, 1030–1038.
- Pasqualini, R. and Ruoslahti, E. (1996), “Organ targeting *in vivo* using phage display peptide libraries”, *Nature* **380**, 364–366.
- Raftery, A. (1997), “A note on Bayes factors for log-linear contingency table models with vague prior information”, *Journal of the Royal Statistical Society B* **48**, 249–250.
- Spiegelhalter, D. and Smith, A. (1982), “Bayes factors for linear and log-linear models with vague prior information”, *Journal of the Royal Statistical Society B* **44**, 377–387.

Wasserman, L. (2000), “Asymptotic inference for mixture models using data dependent priors”,
Journal of the Royal Statistical Society B **62**, 159–180.

Table 1: Selected tri-peptide and organ pairs with ascending display patterns for FDR value $\delta = 0.05$.

Organ	Tri-peptide	Counts (estimated Poisson mean)		
		Stage I	Stage II	Stage III
Brain	GGL	1 (0.81)	0 (2.14)	6 (5.66)
Bowel	DRW	0 (0.42)	0 (1.26)	4 (3.77)
	AGV	0 (0.38)	0 (1.12)	4 (3.67)
	FGG	0 (0.39)	0 (1.21)	4 (3.73)
	GGR	1 (0.85)	0 (2.19)	6 (5.74)
	GLL	0 (0.62)	1 (1.38)	3 (3.06)
Kidney	LRV	0 (0.63)	1 (1.62)	4 (4.20)
	LGS	1 (1.46)	2 (2.71)	5 (5.02)
Muscle	GGT	0 (0.38)	0 (1.34)	5 (4.68)
	FSG	0 (0.62)	1 (1.80)	5 (5.25)
	AGS	0 (0.61)	1 (1.79)	5 (5.26)
	IGS	0 (0.60)	1 (1.77)	5 (5.22)
	AIG	0 (0.41)	0 (1.23)	4 (3.70)
	IAY	0 (0.42)	0 (1.26)	4 (3.77)
	DFS	0 (0.42)	0 (1.26)	4 (3.77)
	RRS	0 (0.58)	1 (1.56)	4 (4.16)
	FRS	0 (0.64)	1 (1.42)	3 (3.10)
	SGV	0 (0.61)	1 (1.38)	3 (3.11)
Pancreas	SSV	1 (0.82)	0 (2.17)	6 (5.74)
	SGV	0 (0.62)	1 (1.37)	3 (3.14)
	GWR	0 (0.62)	1 (1.39)	3 (3.06)
Uterus	AAG	0 (0.63)	1 (1.70)	4 (4.21)

Table 2: Selected tri-peptide and organ pairs for FDR value $\delta = 0.15$ under the criterion $C_{ij} > 3$.

Organ	Tri-peptide
Bowel	DRW AGV FGG GGR GLL GRV AGS GSS RGS GVG LVS
Brain	GGL RGS LLS GGV GSL
Kidney	LRV LGS DSG SLS SRV GGG GSN GLP GSL
Muscle	GGT FSG AGS IGS AIG IAY RRS DFS FRS SGV GDT AGG GSR GAV GCC GRS
Pancreas	SSV GSW GWR LTR RSS LVS GSS GAL LVR FVG LGS AGS
Uterus	AAG GAS GGL RGR GAG GTV ASS GSS GLL

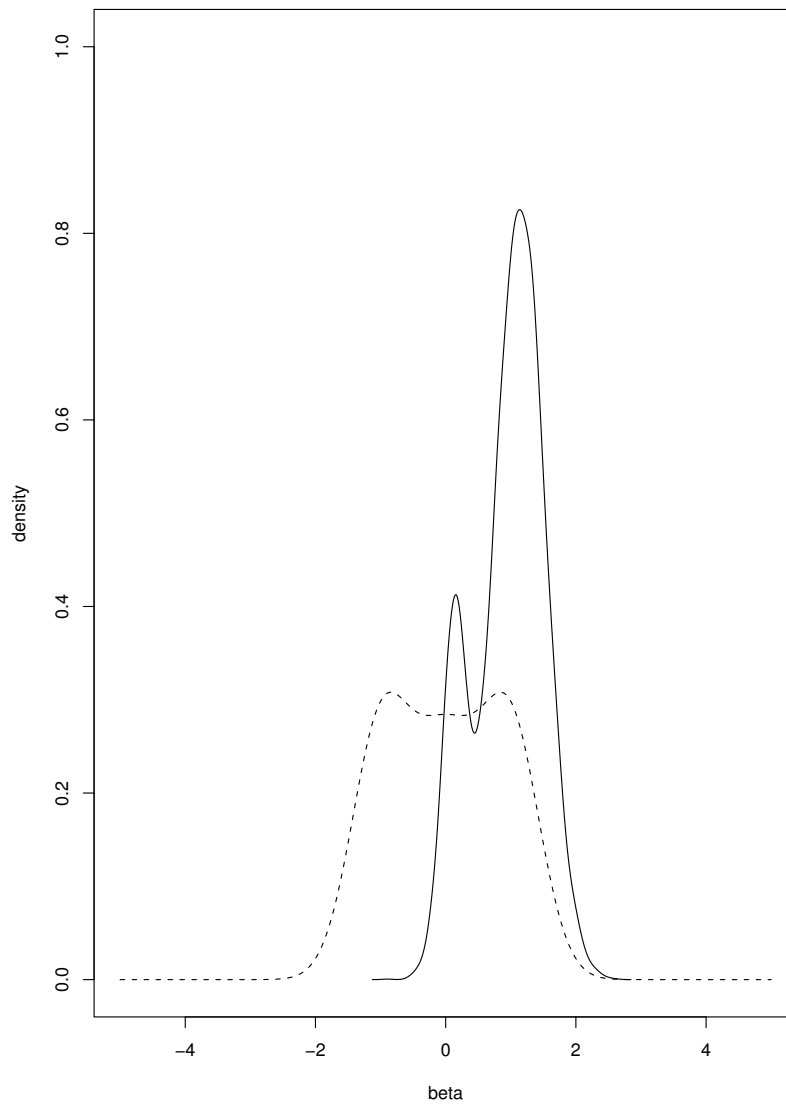


Figure 1: A mixture prior for β_{ij} (dashed line) and the kernel density estimate of the posterior distribution of β_{ij} (solid line) for a randomly chosen peptide and organ pair (i, j) with counts $(1, 0, 6)$.

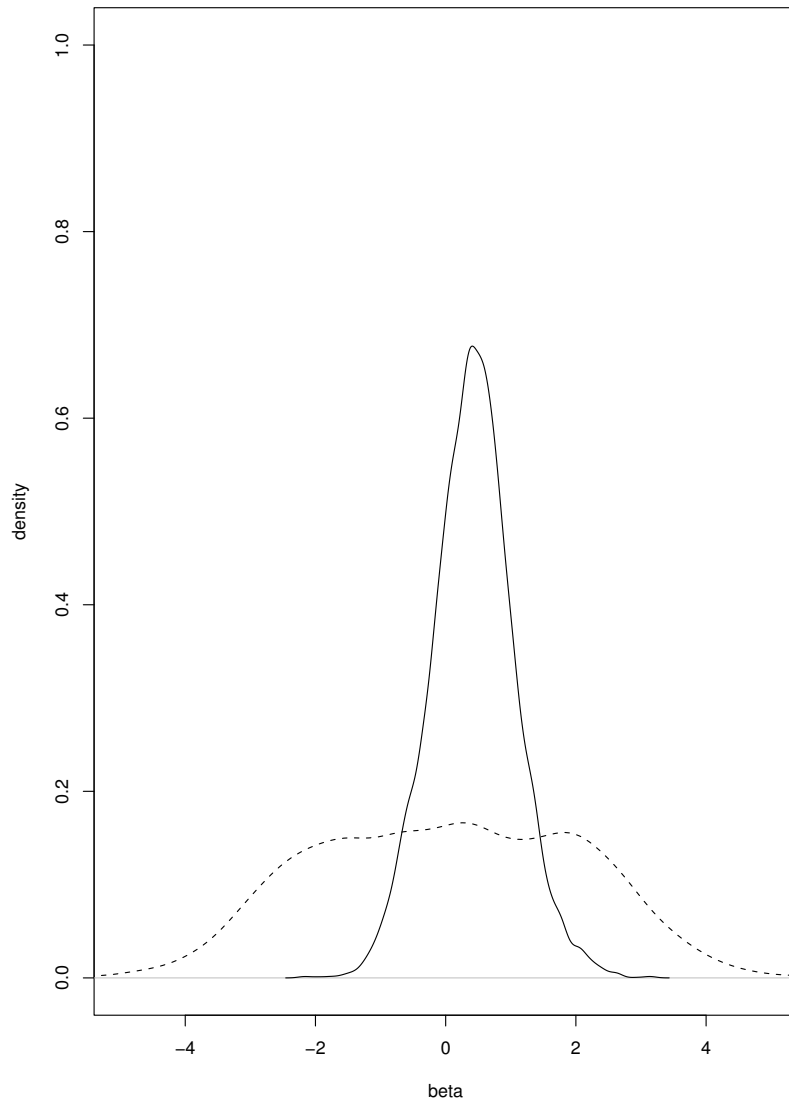


Figure 2: An alternative mixture prior for β_{ij} (dashed line) and the kernel density estimate of the posterior distribution of β_{ij} (solid line) for the peptide and organ pair (i, j) with counts $(1, 0, 6)$.

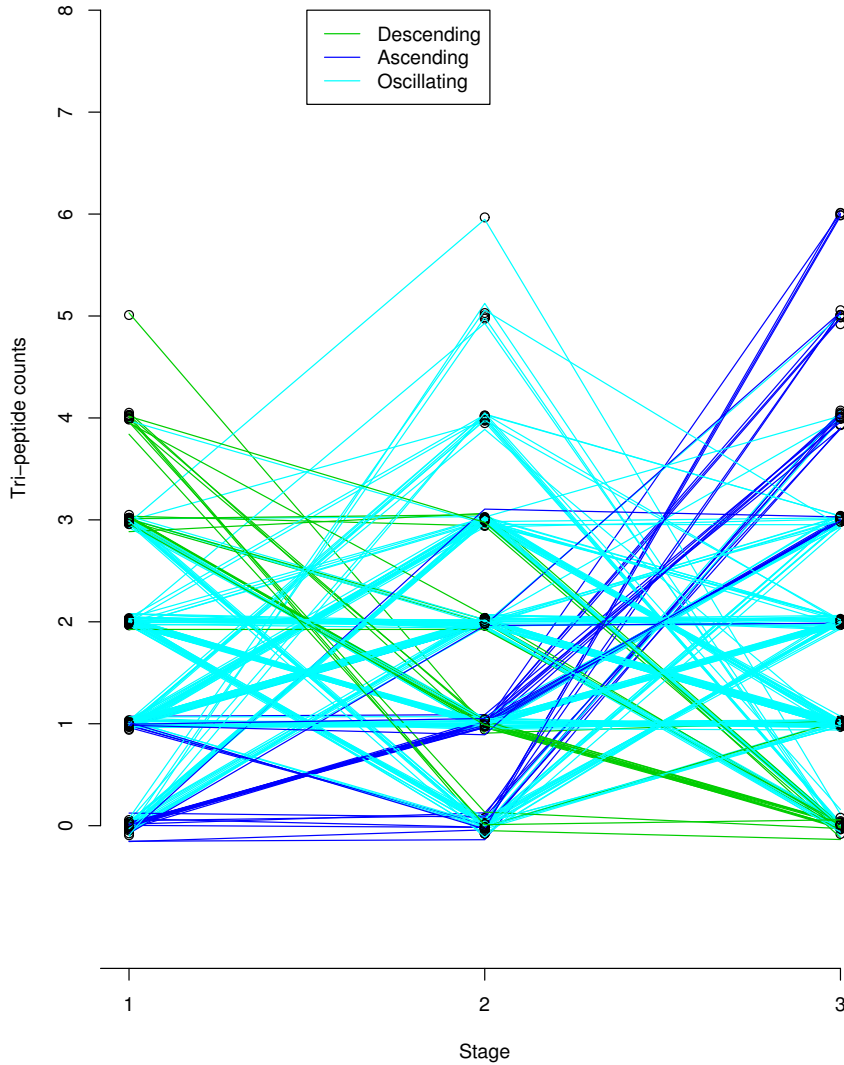


Figure 3: A visual display of 257 vectors of \mathbf{N}_{ij} . The vertical axis represents the actual count values and the horizontal axis represents the three stages. Each circle corresponds to one N_{ijk} , and the three circles belonging to the same peptide and organ pair (i, j) are connected with two lines. The three clusters for the 257 \mathbf{N}_{ij} captured by the Bayesian model are shown in different colors.

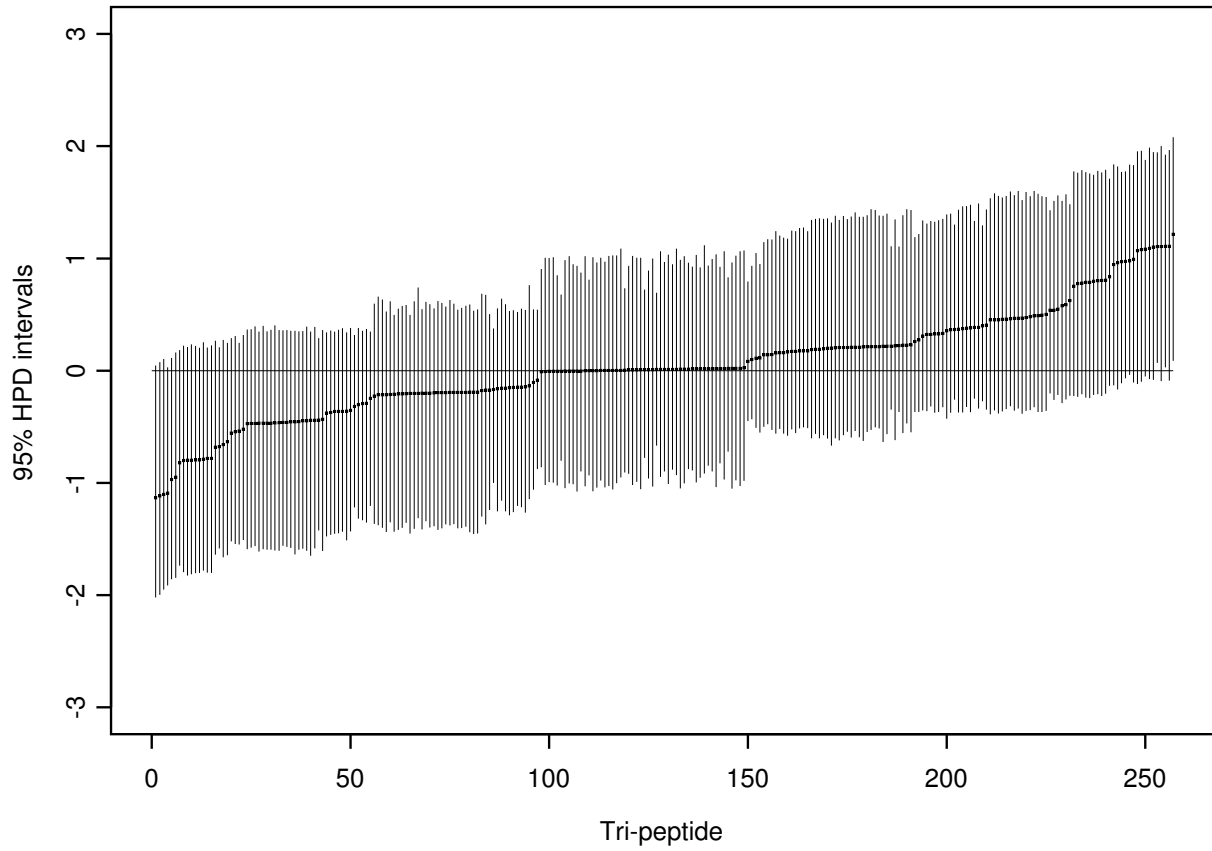


Figure 4: 95% HPD intervals for β_{ij} with sorted posterior means of β_{ij} represented by the dots in the middle.

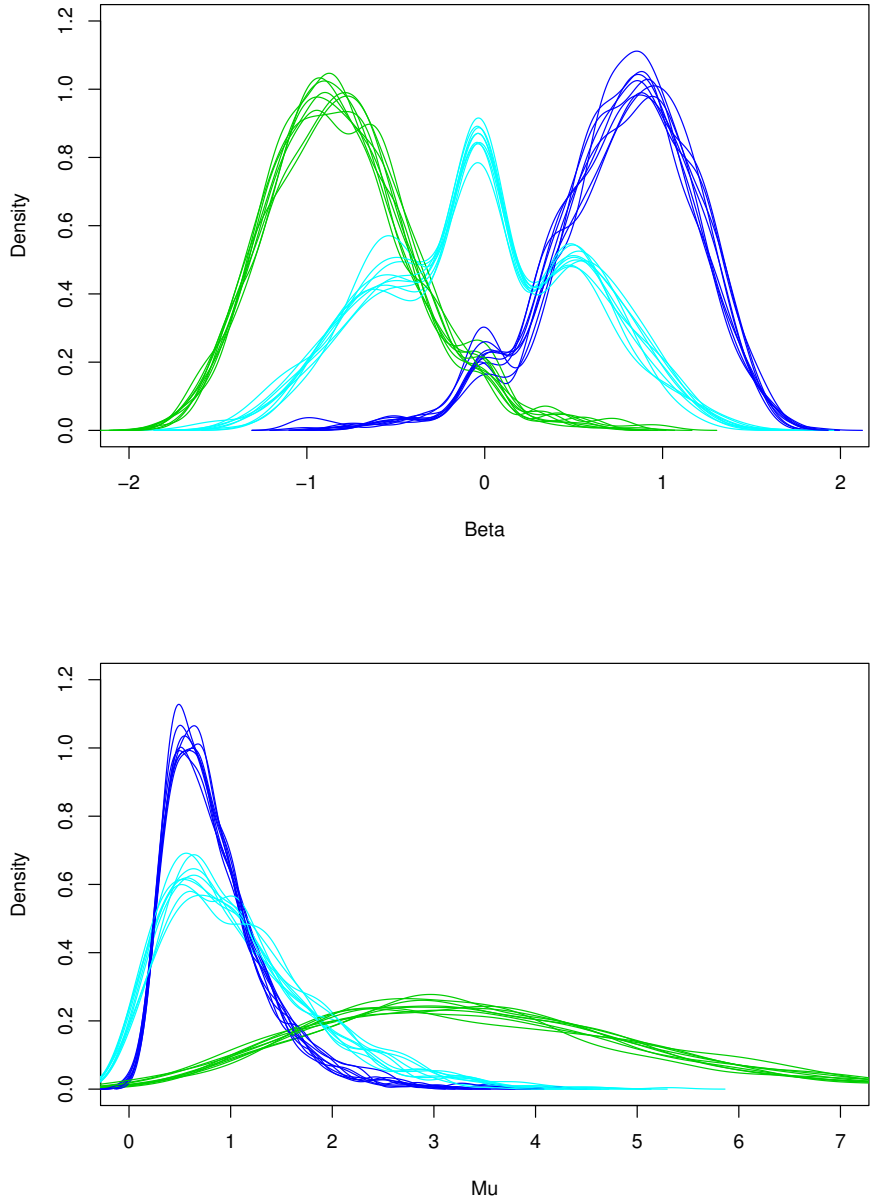


Figure 5: Posterior density estimates of the slopes β_{ij} (upper panel) and the intercepts μ_{ij} (lower panel) from the simulation. The three colored lines represent the first nine pairs with $\beta_{ij} = 1.0$ and $\mu_{ij} = 0.5$, the middle nine pairs with $\beta_{ij} = 0.0$ and $\mu_{ij} = 1.0$, and the last nine pairs with $\beta_{ij} = -1.0$ and $\mu_{ij} = 4.0$.