Bayesian Approach of Biomarker Density Estimation Using Pooled Data

---

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master's of Science
Mathematical Sciences

---

by
Chase Joyner
May 2016

Accepted by:
Dr. Christopher McMahan, Committee Chair
Dr. Yingbo Li
Dr. Brook Russell

Abstract

In this paper, we introduce an approach which can be used to estimate individual biomarker densities by using pooled data under a Bayesian framework. Group testing can effectively reduce the cost of testing individuals for certain infectious diseases. In modeling the parameters of interest, we first consider a univariate population in which all individuals are equally likely to be infected with the disease. Then we extend this setup to a regression setting to account for covariate information of each individual, inducing a heterogenous population.

Chapter 1

Introduction

Biomarkers are an indicator of infectious diseases by using biological specimens such as blood, urine, or other DNA. The goal of this paper is to accurately model individuals' biomarkers under a Bayesian paradigm by placing the individuals into groups and performing a single test on the entire group. As a result, the cost of testing all individuals is dramatically reduced because the number of tests performed is lower. When a group tests positive for a disease, all individuals in that group are then retested for that disease to help identify all positives.

Bayesian inference is a method in which Baye's rule is primarily used in order to obtain a posterior distribution that can provide all information on unknown parameters of interest. One benefit of using Bayesian methods rather than Frequentist methods is that instead of obtaining just point estimates and confidence intervals for the parameters, a Bayesian obtains an entire distribution for the parameters of interest. For example, suppose that you flip a fair coin 100 times and record 64 heads and 36 tails. Would you consider the coin to be bias? As you can see, a Frequentist approach requires a larger sample size to obtain a long-run frequency for a better point-estimate. This leads us towards a Bayesian approach, where we can include prior knowledge of the coin to assess it fairness.

In section 2 of this paper, we discuss the methods used to obtain, or empirically estimate, the posterior distributions, which provides information about the unknown parameters used to model the individual biomarkers. These methods include Bayesian inference, sampling techniques to empirically estimate these parameters, and generalized linear models for a regression setting. The posterior distribution, if obtainable, provides all information on the unknown parameters. However, if the posterior distribution is not obtainable, such as it relies on other unknown parameters or

is not recognizable, then Markov chain Monte Carlo (MCMC) can be used to draw samples from the posterior distributions. Two common MCMC algorithms include the Gibbs sampler and the Metropolis-Hastings algorithms. These methods allow for the inclusion of covariate information through a generalized linear model (GLM); e.g. see [1]. The framework under GLMs is built upon the assumption that we have observations distributed according to some exponential family [4]. With these MCMC sampling techniques, we are able to obtain a sample of the parameters as if we drew them directly from the posterior distribution [2].

The remainder of the report is organized as follows. In section 3 we introduce our notation and consider different models based on certain assumptions about the individual biomarkers. In section 4, we provide simulation results to demonstrate that these models can, in realistic situations, accurately estimate the parameters of interest to be used in modeling the biomarkers.

Chapter 2

Methods

2.1 Bayesian Inference

Bayesian techniques combine *a priori* information and observed data through the use of Baye's rule to obtain a posterior distribution. The *a priori* information specifies a prior distribution, denoted $\pi(\boldsymbol{\theta})$, that uses a person's belief of the true value of the parameters. The observed data specifies a likelihood function when given the unknown parameters, denoted $f(\mathbf{y}|\boldsymbol{\theta})$. Notice that these are both functions of the parameters of interest and by applying Baye's rule as follows, we can obtain posterior distribution [3]

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})} \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1}$$

Generally, we simplify the work by finding the distribution that is proportional to the posterior distribution, and then integrating over the parameter space while setting equal to 1 to find the normalizing constant. The posterior distribution is an update of the prior distribution after observing the data. In common situations, the posterior distribution is not obtainable and so we introduce Markov chain Monte Carlo (MCMC) methods used in the modeling section of this paper. These MCMC techniques are used to obtain samples of the unknown parameters as if they were drawn directly from the posterior distribution, and in turn can empirically estimate the distribution.

2.2 Gibbs Sampling

The Gibbs sampling algorithm generates a sequence of samples of the unknown parameters by sampling from the full-conditional distributions, if possible. To do this, we use the posterior distribution in (1) to obtain the full-conditional distributions

$$f(\theta_i|\boldsymbol{\theta}_{(-i)}, \mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\theta_i|\boldsymbol{\theta}_{(-i)}),$$

where $\boldsymbol{\theta}_{(-i)} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p)$. Notice that here we do not need to include the prior distribution $\pi(\boldsymbol{\theta}_{(-i)})$ since it is constant with respect to the density of θ_i and is therefore just a part of the normalizing constant. Given an initial value of each component of $\boldsymbol{\theta}$, denoted $\theta_i^{(0)}$, proceed as follows:

1. Set $t = 1$;
2. Sample $\theta_i^{(t)} \sim f\left(\theta_i \middle| \theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, ..., \theta_p^{(t-1)}, \mathbf{y}\right)$ for $i = 1, ..., p$;
3. Set $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, ..., \theta_p^{(t)})$;
4. Increment t by 1 and return to step 2.

Once convergence of the chain $\{\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(s)}\}$ is obtained, the procedure should be repeated until a sequence of desired length is generated, say m. Then, $\{\boldsymbol{\theta}^{(s)}, ..., \boldsymbol{\theta}^{(s+m)}\}$ represents a sample from the posterior distribution. With this sequence of samples, we can use the weak law of large numbers to induce properties such as

$$\frac{1}{s}\sum_{i=1}^{s} \boldsymbol{\theta}^{(s)} \longrightarrow E[\boldsymbol{\theta}|\mathbf{y}]$$

as $s \to \infty$. That is, as the number of iterations increases without bound, the mean of the sample obtained from the Gibbs sampling algorithm converges to the true mean [3]. Next, we introduce the Metropolis-Hastings algorithm, which can be used in the situation where the full-conditionals are analytically obtainable, but not recognizable.

2.3 Metropolis–Hastings

The Metropolis-Hastings algorithm is another MCMC technique that can be used when the posterior distribution is not recognizable and therefore cannot be sampled from. Suppose that we have an initial value of our parameter, $\theta_i^{(0)}$. If we propose a new value $\theta_i^\star$ from a proposal distribution, say J_{θ_i}, then an intuitive idea is to include this value in our sample if the posterior density of this proposed value is larger than the posterior density of the current parameter value. Otherwise, we should accept the proposed value $\theta_i^\star$ with some probability. An instinctive way to achieve this is to calculate the ratio of these densities and the use of a correction factor. The correction factor is the ratio of the proposal distribution used to propose $\theta_i^\star$, where the numerator is the proposal distribution evaluated at the current parameter value and the denominator is the proposal distribution evaluated at the proposed value. That is, the acceptance ratio is [1]

$$r = \frac{f\left(\theta_i^\star \middle| \theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, ..., \theta_p^{(t-1)}, \mathbf{y}\right)}{f\left(\theta_i^{(t)} \middle| \theta_1^{(t)}, ..., \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, ..., \theta_p^{(t-1)}, \mathbf{y}\right)} \frac{J\left(\theta_i^{(t)} \middle| \theta_i^\star\right)}{J\left(\theta_i^\star \middle| \theta_i^{(t)}\right)}. \tag{2}$$

Then, we set our acceptance probability to be $\alpha = \min\{r, 1\}$ and then set

$$\theta_i^{(t+1)} = \begin{cases} \theta_i^\star & : \text{with probability } \alpha \\ \theta_i^{(t)} & : \text{with probability } 1-\alpha. \end{cases}$$

A summary of the Metropolis-Hastings algorithm is as follows:

1. Given an initial value $\boldsymbol{\theta}^{(0)}$, set $t = 1$;
2. For each $i = 1, ..., p$ where $f(\theta_i|\boldsymbol{\theta}_{(-i)}, \mathbf{y})$ is not recognizable,

   2a. Propose $\theta_i^\star$ from J_{θ_i} and compute r;

   2b. Set $\theta_i^{(t)} = \begin{cases} \theta_i^\star & : \text{with probability } \alpha \\ \theta_i^{(t-1)} & : \text{with probability } 1-\alpha. \end{cases}$
3. Set $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, ..., \theta_p^{(t)})$;
4. Increment t by 1 and return to step 2.

A desirable property of the proposal distributions is that the proposed values get accepted between 20% and 50% of the time in order to have low correlation in the sequence but to still allow the chain

to move around the parameter space to converge as efficiently as possible [3]. However, in some situations this can be difficult to achieve, such as in the case of higher dimensions. With this said, in section 2.5 we discuss a smarter way to obtain a proposal distribution that will yield much higher acceptance rates, around 95%.

2.4 Generalized Linear Models (GLMs)

The framework developed under a generalized linear model allows the inclusion of covariate information into a model. A generalized linear model is a generalization of regular linear regression to response types other than normally distributed ones. There are three major components to a GLM. The first, and most obvious component, is the random variable. This will specify the conditional distribution of the response variable, say Y_i, given the covariates in the model. It is assumed that this distribution is a member of the exponential family, i.e. it has the form

$$f(y_i|\theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\},$$

where the covariates are included in θ_i. The second component is a linear predictor, which is most commonly a linear function of the regressors, denoted

$$\eta_i = \mathbf{X}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ir},$$

where $\mathbf{X}_i$ is a vector of covariates for the ith observation. The third requirement for a GLM is a smooth and invertible link function, $g(\cdot)$, which relates the mean of the response variable to the linear predictor. That is to say that if $\mu_i = E[Y_i]$, then

$$g(\mu_i) = \eta_i = \mathbf{X}_i'\boldsymbol{\beta}.$$

A link function that can be considered in every situation is the canonical link $\theta_i = \eta_i$. However, there are many link functions that could be used, which depends on the situation or beliefs of how the true mean structure is related to the predictors.

2.5 Bayesian Iterative Re–weighted Least Squares

This algorithm mimics the iterative re-weighted least squares used by Frequentists in order to obtain a nice proposal distribution to be used in a Metropolis-Hastings iteration. In regression modeling, the parameters of interest include the regression coefficients vector $\boldsymbol{\beta}$, which can be difficult to find an appropriate proposal distribution. We begin by placing a normal prior distribution on $\boldsymbol{\beta}$, say $N(\mathbf{a}, \mathbf{R})$. Then under the GLM framework, the posterior distribution for $\boldsymbol{\beta}$ takes the form [1]

$$f(\boldsymbol{\beta}|\mathbf{y}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{a})'\mathbf{R}^{-1}(\boldsymbol{\beta}-\mathbf{a}) + \sum_i \frac{y_i\theta_i - b(\theta_i)}{\phi}\right\}, \tag{2}$$

where $\boldsymbol{\beta}$ is included in θ_i. The idea is to approximate this posterior distribution with a normal distribution to be used as the proposal distribution, and hence the high acceptance rate mentioned in section 2.3. By carrying out a second order Taylor expansion of the likelihood term

$$\ell(\boldsymbol{\beta}) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{\phi}$$

around some value of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(t-1)}$, and then combining terms in (2), we obtain a normal distribution with mean vector

$$\mathbf{m}^{(t)} = \left(\mathbf{R}^{-1} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X}\right)^{-1} \times \left(\mathbf{R}^{-1}\mathbf{a} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\widetilde{\mathbf{y}}(\boldsymbol{\beta}^{(t-1)})\right) \tag{2.1}$$

and covariance matrix

$$\mathbf{C}^{(t)} = \left(\mathbf{R}^{-1} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X}\right)^{-1}, \tag{2.2}$$

where $\mathbf{W}(\boldsymbol{\beta}^{(t-1)})$ is a diagonal weight matrix with entries

$$W_{ii}(\boldsymbol{\beta}^{(t-1)}) = \frac{1}{b''(\theta_i)g'(\mu_i)^2}$$

and $\widetilde{\mathbf{y}}(\boldsymbol{\beta}^{(t-1)})$ is a vector of transformed observations with entries

$$\widetilde{y}_i(\boldsymbol{\beta}^{(t-1)}) = \eta_i + (y_i - \mu_i)g'(\mu_i).$$

The function g is the link function. Then, $N(\mathbf{m}^{(t)}, \mathbf{C}^{(t)})$ approximates the true posterior distribution in (2). This is the distribution used as the proposal distribution, i.e. $J_{\boldsymbol{\beta}^{(t-1)}} = N(\mathbf{m}^{(t)}, \mathbf{C}^{(t)})$. This method is summarized as follows:

1. Given an initial value $\boldsymbol{\beta}^{(0)}$, set $t = 1$;

2. Propose $\boldsymbol{\beta}^{\star}$ from $J_{\boldsymbol{\beta}^{(t-1)}}$ and compute r;

3. Set $\boldsymbol{\beta}^{(t)} = \begin{cases} \boldsymbol{\beta}^{\star} & : \text{with probability } \alpha \\ \boldsymbol{\beta}^{(t-1)} & : \text{with probability } 1-\alpha. \end{cases}$

4. Increase t by 1 and return to step 2.

The construction of the proposal parameters $\mathbf{m}^{(t)}$ and $\mathbf{C}^{(t)}$ approximates the posterior mode and posterior covariance matrix for $\boldsymbol{\beta}$. As a result, the acceptance rate in this method is extremely high, usually 90% and higher, while keeping the correlation in the sequence low.

Chapter 3

Models

3.1 Univariate Models

Let $\mathcal{C}_{ij}$ denote the continuous individual biomarker concentration level for the ith specimen in the jth pool of size c_j, where $i = 1, ..., c_j$ and $j = 1, ..., J$. Also, let $\mathcal{C}_j$ denote the continuous biomarker concentration level observed for the jth pool. In order to relate $\mathcal{C}_j$ to $\mathcal{C}_{ij}$, we assume that $\mathcal{C}_j = c_j^{-1} \sum_{i=1}^{c_j} \mathcal{C}_{ij}$. This assumption is ubiquitously made in the statistical literature (see Faraggi et al., 2003; Liu and Schisterman, 2003; Liu et al., 2004; Mumford et al., 2006; Bondell et al., 2007; Vexler et al., 2008; Malinovsky et al., 2012), and we find it to be reasonable as long as the pooled assessments contain like-volume specimens. Given $\mathcal{C}_{ij} \sim f(\cdot|\boldsymbol{\theta})$, we are left to estimate $\boldsymbol{\theta}$ from the observed data $\mathcal{C}_j$, $j = 1, ..., J$. Since measurements of pooled data are taken, the individual concentration levels are never observed, making the $\mathcal{C}_{ij}$ latent variables. If the distribution of the observed data, $f_{\mathcal{C}_j}$, is obtainable, then the posterior distribution of $\boldsymbol{\theta}$ is

$$f(\boldsymbol{\theta}|\boldsymbol{\mathcal{C}}) \propto \prod_{j=1}^{J} f_{\mathcal{C}_j}(\mathcal{C}_j|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}),$$

where $\pi(\boldsymbol{\theta})$ represents the prior distribution that is introduced because of our uncertainty in the true values of $\boldsymbol{\theta}$.

We begin by assuming that $\mathcal{C}_{ij} \sim N(\mu, \sigma^2)$. Under the assumption that the pooled biomarker concentration levels are the arithmetic mean of the individual biomarkers, it follows that

$\mathcal{C}_j \sim N(\mu, c_j^{-1}\sigma^2)$. We introduce the following prior distributions

$$\mu|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{n_0}\right) \quad \text{and} \quad \sigma^2 \sim IG\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right).$$

Under these prior formulations, the conditional posterior distributions are

$$\mu|\boldsymbol{\mathcal{C}}, \sigma^2 \sim N\left(\frac{\sum_{j=1}^{J} c_j\mathcal{C}_j + \mu_0 n_0}{N + n_0}, \frac{\sigma^2}{N + n_0}\right)$$
$$\sigma^2|\boldsymbol{\mathcal{C}} \sim IG\left(\frac{\alpha_0 + J}{2}, \frac{\beta_0 + \sum_{j=1}^{J} c_j\mathcal{C}_j^2 + n_0\mu_0^2}{2} - \frac{(\sum_{j=1}^{J} c_j\mathcal{C}_j + \mu_0 n_0)^2}{2(N + n_0)}\right),$$

where $\boldsymbol{\mathcal{C}} = (\mathcal{C}_1, ..., \mathcal{C}_j)$ and $N = \sum_{j=1}^{J} c_j$. Here, we implement a Gibbs sampler to estimate $\boldsymbol{\theta}$. When the concentration levels are non-negative and right-skewed as is the case in common practice, a different model should be considered.

Now assume that $\mathcal{C}_{ij} \sim \text{Gamma}(\alpha, \beta)$. As a consequence, we have the reasonable likelihood $\mathcal{C}_j \sim \text{Gamma}(c_j\alpha, c_j\beta)$. We let the model parameters $\boldsymbol{\theta} = (\alpha, \beta)$ have independent prior distributions

$$\alpha \sim \text{Exp}(\lambda) \quad \text{and} \quad \beta \sim \text{Gamma}(a, b).$$

The full conditional posterior distribution of β has a closed form while the conditional posterior distribution of α does not. Specifically, we have

$$f(\alpha|\beta, \boldsymbol{\mathcal{C}}) \propto \frac{\left[\beta^N e^{-\lambda} \prod_{j=1}^{J} \left(c_j\mathcal{C}_j\right)^{c_j}\right]^{\alpha}}{\prod_{j=1}^{J} \Gamma(c_j\alpha)}$$
$$\beta|\alpha, \boldsymbol{\mathcal{C}} \sim \text{Gamma}\left(N\alpha + a, \sum_{j=1}^{J} c_j\mathcal{C}_j + b\right),$$

in which Gibbs sampling and Metropolis-Hastings algorithm should be used.

Notice that the preceding setups have nice properties, i.e. that the posterior distributions are easily obtainable. However, this is not always the case. Assume that the $\mathcal{C}_{ij}$ independently arise from a common probability density function $f_{\mathcal{C}}(\cdot|\boldsymbol{\theta})$; but $f_{\mathcal{C}_j}(\cdot|\boldsymbol{\theta})$, which denotes the probability density function of the observed pooled measurements, is not obtainable. To circumvent this issue, we proceed by introducing the individual biomarker concentration levels as latent variables. More specifically, introducing $c_j - 1$ latent variables $\widetilde{\mathcal{C}}_{ij}$ for $i = 1, ..., (c_j - 1)$ for each group $j = 1, ..., J$

remedies the unusable previous setup, which we can now express the conditional posterior distribution as

$$f(\boldsymbol{\theta}, \widetilde{\boldsymbol{\mathcal{C}}}|\boldsymbol{\mathcal{C}}) \propto \prod_{j=1}^{J} \left[f_{\mathcal{C}}\left(c_j \mathcal{C}_j - \sum_{i=1}^{c_j-1} \widetilde{\mathcal{C}}_{ij} \middle| \boldsymbol{\theta} \right) \cdot \prod_{i=1}^{c_j-1} f_{\mathcal{C}}(\widetilde{\mathcal{C}}_{ij}|\boldsymbol{\theta}) \right] \times \pi(\boldsymbol{\theta}),$$

where $\widetilde{\boldsymbol{\mathcal{C}}} = (\widetilde{\boldsymbol{\mathcal{C}}}_1, ..., \widetilde{\boldsymbol{\mathcal{C}}}_J)$ and $\widetilde{\boldsymbol{\mathcal{C}}}_j = (\widetilde{\mathcal{C}}_{1j}, ..., \widetilde{\mathcal{C}}_{(c_j-1)j})$. We use a hybrid of Gibbs sampling and Metropolis-Hastings algorithms to sample from the joint posterior distributions one variable at a time. In each iteration of the MCMC, for each parameter (or latent variable) that has a conditional posterior distribution in a common distribution family, we draw a random sample from that distribution conditional on all other parameters fixed to their current values; while for each parameter (or latent variable) that only has its conditional posterior density available up to some normalizing constant, we use univariate Metropolis-hastings algorithm.

3.2 Modeling Covariates in a Regression Setting

In common practice, it can be beneficial to include covariate information into models. Up to this point, we have only modeled the biomarker densities with the a single observation, the pooled assessment for each group. Now, we introduce a covariate information matrix $\mathbf{X}$ whose ith row is a vector of k covariates for individual i, $i = 1, ..., N$, where $N = \sum_{j=1}^{J} c_j$.

We begin by assuming that the individual concentration levels have the distribution $\mathcal{C}_{ij} \sim N\left(\mathbf{X}'_{ij}\boldsymbol{\beta}, \sigma^2\right)$, where $\mathbf{X}_{ij}$ is a vector of k covariates for the ith individual in the jth group. Then, we have that

$$\mathcal{C}_j \sim N\left(\overline{\mathbf{X}}'_j \boldsymbol{\beta}, \frac{\sigma^2}{c_j}\right),$$

where $\overline{\mathbf{X}}_j = \frac{1}{c_j} \sum_{i=1}^{c_j} \mathbf{X}_{ij}$ for $j = 1, ..., J$. It follows that the joint likelihood function is

$$\boldsymbol{\mathcal{C}}|\boldsymbol{\beta}, \sigma^2 \sim MVN\left(\overline{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 \mathbf{D}\right),$$

where $\boldsymbol{\mathcal{C}} = (\mathcal{C}_1, ..., \mathcal{C}_J)$, $\overline{\mathbf{X}} = \left(\mathbf{1}, \overline{\mathbf{X}}_1, ..., \overline{\mathbf{X}}_J\right)$, and $\mathbf{D}$ is a diagonal matrix with entries $c_1^{-1}, ..., c_J^{-1}$. Analogous to the univariate setting, we specify the prior distributions

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \sim MVN\left(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{\Sigma}\right)$$
$$\pi\left(\sigma^2\right) \sim IG\left(\frac{\alpha_0}{2}, \frac{\beta_0}{2}\right).$$

Therefore, we find that the full-conditional distribution of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}|\boldsymbol{\mathcal{C}},\sigma^2 \sim MVN\left(\left(\overline{\mathbf{X}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\mathcal{C}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\beta}_0'\boldsymbol{\Sigma}^{-1}\right)',\sigma^2\left(\overline{\mathbf{X}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\right).$$

Also, by integrating out $\boldsymbol{\beta}$ from the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2, we obtain the posterior for σ^2 to be

$$\sigma^2|\boldsymbol{\mathcal{C}} \sim IG\left(\frac{J+\alpha_0}{2},\frac{1}{2}\left[\boldsymbol{\mathcal{C}}'\mathbf{D}^{-1}\boldsymbol{\mathcal{C}}+\boldsymbol{\beta}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0+\beta_0-\mathbf{A}_0\mathbf{A}^{-1}\mathbf{A}_0'\right]\right),$$

where $\mathbf{A}=\overline{\mathbf{X}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\Sigma}^{-1}$ and $\mathbf{A}_0=\boldsymbol{\mathcal{C}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\beta}_0'\boldsymbol{\Sigma}^{-1}$. In conclusion, we have

$$\begin{aligned}\boldsymbol{\beta}|\boldsymbol{\mathcal{C}},\sigma^2 &\sim MVN\left(\left(\overline{\mathbf{X}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\mathcal{C}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\beta}_0'\boldsymbol{\Sigma}^{-1}\right)',\sigma^2\left(\overline{\mathbf{X}}'\mathbf{D}^{-1}\overline{\mathbf{X}}+\boldsymbol{\Sigma}^{-1}\right)^{-1}\right)\\ \sigma^2|\boldsymbol{\mathcal{C}} &\sim IG\left(\frac{J+\alpha_0}{2},\frac{1}{2}\left[\boldsymbol{\mathcal{C}}'\mathbf{D}^{-1}\boldsymbol{\mathcal{C}}+\boldsymbol{\beta}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0+\beta_0-\mathbf{A}_0\mathbf{A}^{-1}\mathbf{A}_0'\right]\right).\end{aligned}$$

Here we implement a Gibbs sampling algorithm to quickly estimate the parameters $\boldsymbol{\beta}$ and σ^2.

Analogously to the univariate case, under the reasonable assumption that the pooled observations are non-negative and right-skewed, we now assume that $\mathcal{C}_{ij} \sim \text{Gamma}(\alpha, \mu_{ij}/\alpha)$, where μ_{ij}/α is the scale parameter, $i = 1, ..., c_j$ and $j = 1, ..., J$. Notice that this distribution can be written as

$$\begin{aligned}f_{\mathcal{C}_{ij}} &= \frac{1}{\Gamma(\alpha)}\left(\frac{\mu_{ij}}{\alpha}\right)^{-\alpha}\mathcal{C}_{ij}^{\alpha-1}\exp\left\{-\frac{\alpha\mathcal{C}_{ij}}{\mu_{ij}}\right\}\\ &= \exp\left\{\frac{-\frac{1}{\mu_{ij}}\mathcal{C}_{ij}-\log\mu_{ij}}{1/\alpha}+\alpha\log\alpha-\log\Gamma(\alpha)+(\alpha-1)\log\mathcal{C}_{ij}\right\}.\end{aligned}$$

Therefore, this distribution is a member of the exponential family. We also wish to include covariate information about each individual, and thus relate the mean of this distribution to the covariates via a link function. Since the mean μ_{ij} must be positive valued, we use a log link as follows

$$\log\mu_{ij} = \mathbf{X}_{ij}'\boldsymbol{\beta}.$$

The likelihood function can then be written in the form

$$f_{\mathcal{C}_{ij}} = \exp\left\{\frac{-e^{-\mathbf{X}_{ij}'\boldsymbol{\beta}}\mathcal{C}_{ij}-\mathbf{X}_{ij}'\boldsymbol{\beta}}{1/\alpha}+\alpha\log\alpha-\log\Gamma(\alpha)+(\alpha-1)\log\mathcal{C}_{ij}\right\}.$$

Note that the distribution of the pooled assessments is unobtainable here and so we introduce the latent variables as before and use the joint posterior found below

$$f\left(\boldsymbol{\theta}, \widetilde{\boldsymbol{\mathcal{C}}} | \boldsymbol{\mathcal{C}}\right) \propto \prod_{j=1}^{J} \left[f_{\mathcal{C}}\left(c_j \mathcal{C}_j - \sum_{i=1}^{c_j - 1} \widetilde{\mathcal{C}}_{ij} \middle| \boldsymbol{\theta} \right) \cdot \prod_{i=1}^{c_j - 1} f_{\mathcal{C}}\left(\widetilde{\mathcal{C}}_{ij} | \boldsymbol{\theta} \right) \right] \times \pi(\boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$. Let α and $\boldsymbol{\beta}$ be independently distributed as before. Since we use a log link, $\boldsymbol{\beta}$ can be on the entire real line, so we are able to use a multivariate normal prior. Also, as before, we use Exp(λ) as the prior for α. That is, specify

$$\boldsymbol{\beta} \sim MVN\left(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}\right)$$
$$\alpha \sim \text{Exp}(\lambda).$$

Denoting $a_j = c_j \mathcal{C}_j - \sum_{i=1}^{c_j - 1} \widetilde{\mathcal{C}}_{ij}$ and $\mathbf{X}_j$ as the vector of covariates for the last individual in group j, the joint posterior distribution is

$$\begin{aligned} f\left(\alpha, \boldsymbol{\beta}, \widetilde{\boldsymbol{\mathcal{C}}} | \boldsymbol{\mathcal{C}}\right) \propto & \prod_{j=1}^{J} \exp\left\{ \frac{e^{-\mathbf{X}_j' \boldsymbol{\beta}} a_j + \mathbf{X}_j' \boldsymbol{\beta}}{-1/\alpha} + \alpha \log \alpha - \log \Gamma\left(\alpha\right) + (\alpha - 1) \log a_j \right\} \cdot \\ & \prod_{i=1}^{c_j - 1} \exp\left\{ \frac{e^{-\mathbf{X}_{ij}' \boldsymbol{\beta}} \mathcal{C}_{ij} + \mathbf{X}_{ij}' \boldsymbol{\beta}}{-1/\alpha} + \alpha \log \alpha - \log \Gamma\left(\alpha\right) + (\alpha - 1) \log \mathcal{C}_{ij} \right\} \cdot \\ & \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\} \cdot \exp\left\{ -\frac{\alpha}{\lambda} \right\}. \end{aligned}$$

Then, we have the posterior distribution for α is

$$f\left(\alpha \middle| \boldsymbol{\beta}, \widetilde{\boldsymbol{\mathcal{C}}}, \boldsymbol{\mathcal{C}}\right) \propto \exp\left\{ -\alpha \gamma + N(\alpha \log \alpha - \log \Gamma(\alpha)) \right\},$$

where

$$\gamma = \sum_{j=1}^{J} e^{-\mathbf{X}_j' \boldsymbol{\beta}} a_j + \sum_{j=1}^{J} \mathbf{X}_j' \boldsymbol{\beta} - \sum_{j=1}^{J} \log a_j + \sum_{j=1}^{J} \sum_{i=1}^{c_j - 1} e^{-\mathbf{X}_{ij}' \boldsymbol{\beta}} \mathcal{C}_{ij} + \sum_{j=1}^{J} \sum_{i=1}^{c_j - 1} \mathbf{X}_{ij}' \boldsymbol{\beta} - \sum_{j=1}^{J} \sum_{i=1}^{c_j - 1} \log \mathcal{C}_{ij} + \frac{1}{\lambda}.$$

The posterior for $\boldsymbol{\beta}$ is

$$f\left(\boldsymbol{\beta}|\alpha,\widetilde{\boldsymbol{\mathcal{C}}},\boldsymbol{\mathcal{C}}\right) \propto \exp\Bigg\{ -\alpha\left(\sum_{j=1}^{J} e^{-\mathbf{X}_j'\boldsymbol{\beta}}a_j + \sum_{j=1}^{J}\mathbf{X}_j'\boldsymbol{\beta} + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1} e^{-\mathbf{X}_{ij}'\boldsymbol{\beta}}\mathcal{C}_{ij} + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1}\mathbf{X}_{ij}'\boldsymbol{\beta}\right) - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\Bigg\}.$$

Lastly, we have the posterior distribution for $\widetilde{\boldsymbol{\mathcal{C}}}$ to be

$$f\left(\widetilde{\boldsymbol{\mathcal{C}}}|\alpha,\boldsymbol{\beta},\boldsymbol{\mathcal{C}}\right) \propto \exp\Bigg\{ -\alpha\left(\sum_{j=1}^{J} e^{-\mathbf{X}_j'\boldsymbol{\beta}}a_j + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1} e^{-\mathbf{X}_{ij}'\boldsymbol{\beta}}\mathcal{C}_{ij}\right) + (\alpha-1)\left(\sum_{j=1}^{J}\log a_j + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1}\log\mathcal{C}_{ij}\right)\Bigg\}.$$

In conclusion, the posterior distributions are

$$f\left(\alpha|\boldsymbol{\beta},\widetilde{\boldsymbol{\mathcal{C}}},\boldsymbol{\mathcal{C}}\right) \propto \exp\left\{-\alpha\gamma + N(\alpha\log\alpha - \log\Gamma(\alpha))\right\},$$

$$f\left(\boldsymbol{\beta}|\alpha,\widetilde{\boldsymbol{\mathcal{C}}},\boldsymbol{\mathcal{C}}\right) \propto \exp\Bigg\{ -\alpha\left(\sum_{j=1}^{J} e^{-\mathbf{X}_j'\boldsymbol{\beta}}a_j + \sum_{j=1}^{J}\mathbf{X}_j'\boldsymbol{\beta} + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1} e^{-\mathbf{X}_{ij}'\boldsymbol{\beta}}\mathcal{C}_{ij} + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1}\mathbf{X}_{ij}'\boldsymbol{\beta}\right) - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\Bigg\}$$

$$f\left(\widetilde{\boldsymbol{\mathcal{C}}}|\alpha,\boldsymbol{\beta},\boldsymbol{\mathcal{C}}\right) \propto \exp\Bigg\{ -\alpha\left(\sum_{j=1}^{J} e^{-\mathbf{X}_j'\boldsymbol{\beta}}a_j + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1} e^{-\mathbf{X}_{ij}'\boldsymbol{\beta}}\mathcal{C}_{ij}\right) + (\alpha-1)\left(\sum_{j=1}^{J}\log a_j + \sum_{j=1}^{J}\sum_{i=1}^{c_j-1}\log\mathcal{C}_{ij}\right)\Bigg\}.$$

As for the proposal distribution for $\boldsymbol{\beta}$, we use the methodology discussed in section 2.5. From the likelihood function, we see

$$\theta_{ij} = -\frac{1}{\mu_{ij}} \quad \text{and} \quad b(\theta_{ij}) = \log\mu_{ij} = -\log(-\theta_{ij}).$$

Therefore, we have that

$$b''(\theta_{ij}) = \frac{1}{\theta_{ij}^2} = \mu_{ij}^2 = \exp\{2\mathbf{X}_{ij}'\boldsymbol{\beta}\} \quad \text{and} \quad g'(\mu_{ij})^2 = \exp\{-2\mathbf{X}_{ij}'\boldsymbol{\beta}\},$$

and so the weight matrix $\mathbf{W}(\boldsymbol{\beta}) = I_{N\times N}$. Lastly, we find the transformed observations to be

$$\widetilde{\mathcal{C}}_{ij}(\boldsymbol{\beta}) = \eta_{ij} + (\widetilde{\mathcal{C}}_{ij} - \mu_{ij})g'(\mu_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta} + \big(\widetilde{\mathcal{C}}_{ij} - \exp\{\mathbf{X}'_{ij}\boldsymbol{\beta}\}\big)\frac{1}{\exp\{\mathbf{X}'_{ij}\boldsymbol{\beta}\}}.$$

This gives the proposal distribution for $\boldsymbol{\beta}$ to be a normal distribution with parameters

$$\mathbf{m}^{(t)} = \left(\mathbf{R}^{-1} + \alpha\mathbf{X}'\mathbf{X}\right)^{-1} \times \left(\mathbf{R}^{-1}\mathbf{a} + \alpha\mathbf{X}'\widetilde{\boldsymbol{\mathcal{C}}}(\boldsymbol{\beta}^{(t-1)})\right)$$

and

$$\mathbf{C}^{(t)} = \left(\mathbf{R}^{-1} + \alpha\mathbf{X}'\mathbf{X}\right)^{-1}.$$

From here, we implement a Metropolis-Hastings algorithm to update $\boldsymbol{\theta}$ and $\widetilde{\boldsymbol{\mathcal{C}}}$.

Chapter 4

Simulations and Results

The design and results of our simulations are presented here. For each univariate model, we simulated a sample of 1,000 individual concentration levels from the assumed true distribution of individuals. Then, we placed these individuals into groups of size $c = 4$ and averaged the values to create the pooled assessments. Next, we performed an MCMC to obtain 10,000 samples of each parameter to generate parameter estimates, the lower 2.5% and upper 97.5% quantiles, and the standard error of the parameter estimates. We repeated this for 1,000 simulated data sets and averaged the estimates and standard errors, as well as calculated 95% coverage probabilities. The results for the univariate normal model can be found in table 4.1 and the results for the univariate gamma model can be found in table 4.2.

Parameter	True	Estimates	CP95	SE
μ	3.4	3.3994	0.958	0.0289
σ^2	0.8	0.8047	0.947	0.0718

Table 4.1: Results of univariate normal case.

Parameter	True	Estimates	CP95	SE
α	2.1	2.2807	0.929	0.2454
β	0.8	0.8663	0.931	0.0912

Table 4.2: Results of univariate gamma case.

As we can see, both univariate models appeared to estimate the parameters well. In both situations, the coverage probabilities were roughly around 95% and the standard errors are low. Although,

the standard error for α appears to be higher than the other standard errors, this can be caused by the posterior distribution for α not being recognizable, which can induce more variability in the estimates. However, increasing the sample size effectively reduced the standard error and also led to a better estimate. Next, we look at how the models in the regression setting performed.

For the two models in the regression setting, we generated the covariate matrix of dimension $N \times 3$ consisting of a column of all 1s for the intercept, a column of generated normal random variables, and a random binary column of 0s and 1s. For the normal model we considered 400 individuals and for the gamma model we used 1,000 individuals. In both situations, we used a common group size of $c = 4$ as before. We simulated 1,000 data sets and drew 10,000 samples each parameter. Tables 4.3 and 4.4 found below display the results

Parameter	True	Estimates	CP95	SE
σ^2	0.8	0.7945	0.956	0.1098
β_1	3	3.0074	0.953	0.23
β_2	-2	-2.005	0.957	0.0878
β_3	5	4.9867	0.952	0.2039

Table 4.3: Results of normal in regression setting.

Parameter	True	Estimates	CP95	SE
α	5	4.693	0.876	0.4256
β_1	-3	-2.983	0.972	0.0731
β_2	0.75	0.749	0.973	0.0267
β_3	1.1	1.083	0.968	0.0583

Table 4.4: Results of gamma in regression setting.

The results of the simulations performed appear to validate both models. The estimates for σ^2 and $\boldsymbol{\beta}$ are very close to the true values used to generate the individual concentration levels. Also, the coverage probabilities are about 95% and the standard errors are low. As for the gamma model, we were able to estimate $\boldsymbol{\beta}$ very well, but the estimate for α was low. This can be due to the high standard error, which we saw occur in the univariate gamma model. However, increasing the sample size and letting our chain run longer drastically improved the results.

Chapter 5

Discussion

In this paper, we have introduced a way to model individual biomarkers in a Bayesian paradigm using pooled observations of grouped individuals. We have motivated the usefulness of Bayesian inference in this framework by simulations under reasonable assumptions. As a result, group testing allows the cost of testing individuals for infectious diseases to be dramatically reduced by lowering the number of tests performed. While Bayesian inference is an extremely important and useful tool, it comes with drawbacks as does any statistical method. The computation time used in these sampling techniques can often be highly expensive and thus efficient coding is vital. Also, including incorrect prior knowledge into the models can greatly affect inference by drawing samples in the wrong part of the parameter space, potentially leading to false positive or false negative results for individuals. Another drawback worth mentioning was the starting value for $\boldsymbol{\beta}$ was crucial in the Gamma model under the regression setting. However, there are many solutions to this. To this end, our methodology provides a Bayesian perspective on this evolving area of Biomarker density estimation.

5.1 Future Work

The extension to be done in the near future is to use the estimated parameters and the individual likelihood function to find the optimal threshold for diagnosing individuals as either positive or negative for the disease. This can be done by the use of Youden's index, which is a function of the specificity (true negative rate) and sensitivity (true positive rate). This index is

between 0 and 1 and values near 1 indicate that the biomarker is very effective and values near 0 indicate that the biomarker has limited usefulness [5]. After successfully implementing Youden's index, it is desirable to compare certain models that we have discussed in this paper, along with others such as Weibull, Log-Normal, etc. Another extension worth looking at is the use of different group sizes, such as $c = 2$ or $c = 3$, to see if there are any effects. It seems reasonable to believe that there exists an optimal group size, which would lead to better results.

Bibliography

[1] Gamerman D. Sampling from the posterior distribution in generalized linear mixed models. *Statistics and computing*, 1996.

[2] Grimmer J. An introduction to bayesian inference via variational approximations. *Political analysis advance*, 2010.

[3] Hoff P. A first course in bayesian statistical methods. 2010.

[4] Nelder J. and Wedderburn R. Generalized linear models. 1972.

[5] Perkins N. Schisterman E. and Bondell H. Optimal cutpoint and its corresponding youden index to discriminate individuals using pooled blood samples. *Epidemiology*, 2005.