1 The Model

Let $i = 1, ..., n$ denote subjects, $j = 1, ..., p$ cell subsets and $l = 1, ..., L_{ij}$ the observations we have for each subject/subset combination. Assume the following model:

$$\nu \sim \mathrm{N}_p(0, \Sigma), \qquad z \in \{0, 1\} \sim \mathrm{Ising}(\Theta),$$

$$\mathrm{logit}(\mu_{ijl}) = X_{ijl}\beta_j + z_{ij}T_{ijl}\tau_j + \nu_j$$

$$p_{ijl} \sim \mathrm{Beta}(M_j\mu_{ijl}, M_j(1 - \mu_{ijl})), \qquad y_{ijl} \sim \mathrm{Binom}(N_{il}, \mu_{ijl}).$$

In the model X_{ijl} corresponds to fixed effects that are not dependent on the treatment e.g. age, gender etc.. T_{ijl} is a vector of covariates corresponding to the treatment effect, in the simplest case this would just be an indicator for stimulation but in more interesting cases T_{ijl} may be a binary vector for which stimulation was introduced to a sample or indicators for time varying effects.

Conditional on ν_i and z_i, y_{ijl} follows a beta-binomial distribution and so, it has expectation and variance:

$$E(y_{ijl}/N_{il}|\nu_i, z_i) = \mu_{ijl},$$

$$Var(y_{ijl}/N_{il}|\nu_i, z_i) = \frac{\mu_{ijl}(1 - \mu_{ijl})}{N_{il}} + \frac{\mu_{ijl}(1 - \mu_{ijl})}{M_j + 1}.$$

And so, the beta-binomial model captures the fact that regardless of how many cells we get to see (N_{il}), we cannot obtain perfect information regarding the data generating process based on a singe blood sample.

2 Computation

2.1 An EM Formulation

Estimating the mixed mixture model is difficult because the likelihood involves both a high-dimensional integral and a summation over 2^p possible response assignments:

$$\mathcal{L}(\beta, \tau, \Sigma, \Theta) = \prod_{i=1}^{n} \left\{ \sum_{z \in \{0,1\}^p} \int_{\mathcal{R}^p} P(z)\varphi(\nu; \Sigma) \prod_{j=1}^{p} \prod_{l=1}^{L_{ij}} f(y_{ijl}|z, \nu) d\nu \right\}.$$

A common method for maximizing such complex likelihood is the EM algorithm. In the EM algorithm we maximize the complete-information log-likelihood where we replace the unknown quantities with their expectation conditional on the observed data as dictated by the current parameter estimates.

$$Q\left(\{\beta, \tau, \Sigma, \Theta\}^t \mid |\{\beta, \tau, \Sigma, \Theta\}^{t-1}\right) = \sum_{i=1}^{n} E(\log f(y, z, \nu)|y)$$

$$= \sum_{i=1}^{n} \sum_{z \in \{0,1\}^p} P(z|y) E\left[(\log f(y, \nu, z) \mid |z, y\right]$$

$$= \sum_{i=1}^{n} \sum_{z \in \{0,1\}^p} P(z|y) \int_{\mathcal{R}^p} f(\nu|y, z) \log f(y, \nu, z) d\nu.$$

with

$$\log f(y_i, z_i, \nu_i) = \log P(z_i) + \log \varphi(\nu_i; \Sigma) + \sum_{j=1}^{p} \sum_{l=1}^{L_{ij}} \log f(y_{ijl}|z, \nu).$$

The EM complete-data log-likelihood is still intractable because it involves the same high-dimensional integrals and summations, but they are suggestive of the possibility of approximating the full-information log-likelihood using Monte-Carlo integration. Let $(\nu_i^*, z_i^*)_1, ..., (\nu_i^*, z_i^*)_M$ be joint samples from the posterior distribution of ν and z given y. Then, the complete information log-likelihood can be approximated with:

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{n} \log f(y, \nu_{im}^*, z_{im}^*).$$

Replacing the expectation step with a posterior sampling step yields a Stochastic-EM (SEM) algorithm. We discuss the implementation of the SEM algorithm in our setting next.

3 Posterior Sampling for the Stochastic-EM

We sample from the posterior joint distribution of ν_i and z_i in two stages. First, we sample from the marginal posterior distribution of z_i and then sample $\nu_i|z_i$. We start by describing a Gibbs sampler for sampling z_i. For an arbitrary index $j \in \{1, ..., p\}$, the posterior probability of response in the jth subset can be written as:

$$\begin{aligned} P(z_j = 1|z_{-j}, y) &\propto P(z_j = 1|z_{-j}) f(y, \nu|z) \\ &= P(z_j = 1|z_{-j}) f(y_j, \nu_j|z_j) f(y_{-j}, \nu_{-j}|z_{-j}, \nu_j) \end{aligned}$$