A justifiable prior for U-INVITE

Jeff Zemla

June 15, 2017

1 The zero-inflated beta-binomial

The prior for an edge, $\mathbb{P}(G_{ij}^{-s})$, is constructed around the observed counts β_o and α_o which denote the number of participants (excluding the participant s being fitted) who either do or do not have that edge (G_{ij}^n) respectively. Specifically,

$$\beta_o = \sum_{n \neq s}^{N} \mathbb{1}_{G_{ij}^n = 1} \tag{1}$$

$$\alpha_o = \sum_{n \neq s}^{N} \mathbb{1}_{G_{ij}^n = 0} \tag{2}$$

where N is the total number of participants, s is the participant whose graph is being estimated, and G_{ij}^n denotes an edge (or lack of edge) between i and j for participant n. Note that α_o is not necessarily $N - \beta_o$ (and vice versa) because a participant's graph might not contain both items i and j; thus the edge is undefined for that participant, rather than zero.

It is assumed that these counts are generated from a zero-inflated beta-binomial process. Zeros are generated with probability p_1, while zeros and ones are estimated from a beta-binomial (BB) process with probability $1-p_1$. The reason for this is that with a sparsity of data for each participant, any estimated graph will be too sparse, i.e., contain too many zeros compared to the true graph. Since each uncensored walk traverses only a small fraction of the total number of edges, there is no data to estimate edges that are not traversed, leading to excess zeros. This is the "zero-inflated" portion of the model, whereas the BB process models the "signal" in the observed counts, i.e., a true estimate of the edge prior from the U-INVITE process.

The observed proportion of zeros is

$$\mu_2 = \frac{\alpha_o}{\alpha_o + \beta_o} \tag{3}$$

and

$$\mu_2 = p_1 + (1 - p_1)\mu_1 \tag{4}$$

so

$$\mu_1 = \frac{\mu_2 - p_1}{1 - p_1} \tag{5}$$

Substituting Equation 3 into Equation 5 we have

$$\mu_1 = \frac{\alpha_o - p_1(\alpha_o + \beta_o)}{(\alpha_o + \beta_o)(1 - p_1)} \tag{6}$$

or equivalently,

$$\mu_1 = \frac{\alpha_o - p_1(\alpha_o + \beta_o)}{\alpha_o + \beta_o - p_1(\alpha_o + \beta_o)} \tag{7}$$

In other words, μ_1 represents the proportion of zeros in the counts after subtracting zeros that were not modeled by the BB process. The prior for an edge is then

$$P(G_{ij}^{-s}) = 1 - \mu_1 = 1 - \left[\frac{\alpha_o - p_1 n}{(1 - p_1)n}\right] \tag{8}$$

where $n = \alpha_o + \beta_o$.

There is an additional constraint such that $p_1 n \leq \alpha_o$; with greater values of p_1 it would not be possible to observe α_o zeros. To enforce this constraint, we scale the probability of the zero-generating process to range from zero to u_2:

$$p_1 = p_2 \frac{\alpha_o}{\alpha_o + \beta_o} \tag{9}$$

where p_2 denotes the proportion of observed zeros that were due to the zero-generating process. Now p_2 is a free parameter that can range from zero to one. By substituting Equation 9 into Equation 8 we can now represent the prior on an edge as:

$$P(G_{ij}^{-s}) = 1 - \left[\frac{\alpha_o(1-p_2)}{n - p_2\alpha_o}\right] \tag{10}$$

With some substitution, we can simplify this equation:

$$P(G_{ij}^{-s}) = \frac{\beta}{\beta + \alpha_o(1-p_2)} \tag{11}$$

2 Unresolved issues

The prior in Equation 11 seems to work fine. $p_2 = .275$ seems to emulate the unjustified "halfa" prior used earlier. There are a few issues that would be nice to resolve if possible. One issue is that the results are fairly sensitive to the choice for p_2, and it would be nice if there were a principled process for choosing p_2 based on the data. Another related issue is that p_2 really shouldn't be a fixed number at all. By assuming a fixed value for p_2 we are saying that, no matter what a graph looks like, a fixed proportion of the non-edges are spurious (i.e., resulting from the zero-generating process). But we are modeling this process, so the graph becomes less sparse over time. Therefore, p_2 should reduce to zero as the density of the estimated graph approaches the density of the true graph. In addition, the process cannot model an excess of ones in the counts, i.e., when the estimated graph is too dense. This might happen if a naïve random walk is used to estimate the initial graph and each participant has a moderate amount of data.

I did try to model this such that p_2 denotes the proportion of excess zeros in the data given the current prior graph density d and the desired or prior density of the true graph, D:

$$p_2 = \frac{D-d}{1-d} \tag{12}$$

p_2 dynamically changes as the graph density changes. This eliminates a free parameter and simultaneously allows for modeling of graphs that are too dense. Unfortunately, it doesn't work. Or rather, it's noticeably worse than fixing the parameter as is done above.