Latent Class Analysis

T diagnostic tests on P populations

Jonathan Marshall

13 November 2015

Let $M = (01001)$ be a binary string representing a test result across multiple tests, where a 1 specifies a positive test and a 0 specifies a negative. The length of this vector is T, the number of tests.

Let P be the number of populations in which the tests were performed. We want to estimate the true prevalence in the population p_i and the test sensitivity s_t and specificity c_t for each test.

The data we have is the number y_{iM} of each observed test pattern M in population i. e.g. $y_{1,(01)}$ would represent the number of units in population 1 that tested negative to the first test and positive to the second.

Then the likelihood of y is multinomial given p, s and c and may be written

$$L(y|p,s,c) = \prod_{i=1}^{P}\prod_{M}\left[p_i\prod_{t=1}^{T}s_t^{M_t}(1-s_t)^{1-M_t} + (1-p_i)\prod_{t=1}^{T}(1-c_t)^{M_t}c_t^{1-M_t}\right]^{y_{iM}}$$

where the product M runs over all 2^T possible test patterns.

To get a Gibbs sampler for this, we introduce latent variables x_{iM} to be the number of true results for each test pattern M in population i (i.e. if the tests had perfect sensitivity and specificity) of true positives that report test pattern M in population i.

The inner term can be expanded out using the binomial formula $(a+b)^n = \sum_{i=1}^{n}\binom{n}{i}a^ib^{n-i}$ to give

$$L(y|p,s,c) = \prod_{i=1}^{P}\prod_{M}\sum_{x_{iM}=0}^{y_{iM}}\binom{y_{iM}}{x_{iM}}\left[p_i\prod_{t=1}^{T}s_t^{M_t}(1-s_t)^{1-M_t}\right]^{x_{iM}}\left[(1-p_i)\prod_{t=1}^{T}(1-c_t)^{M_t}c_t^{1-M_t}\right]^{y_{iM}-x_{iM}}$$

Notice here that x_{iM} represents the unknown true number of positives in population i for test pattern M. If we add these as latent variables, then conditional on those the likelihood reduces to

$$\begin{aligned}
L(y|p,s,c,x) &= \prod_{i=1}^{P}\prod_{M}\binom{y_{iM}}{x_{iM}}\left[p_i\prod_{t=1}^{T}s_t^{M_t}(1-s_t)^{1-M_t}\right]^{x_{iM}}\left[(1-p_i)\prod_{t=1}^{T}(1-c_t)^{M_t}c_t^{1-M_t}\right]^{y_{iM}-x_{iM}}\\
&= \prod_{i=1}^{P}\prod_{M}\binom{y_{iM}}{x_{iM}}p_i^{x_iM}(1-p_i)^{y_{iM}-x_{iM}}\prod_{t=1}^{T}s_t^{M_tx_{iM}}(1-s_t)^{(1-M_t)x_{iM}}\prod_{t=1}^{T}(1-c_t)^{M_t(y_{iM}-x_{iM})}c_t^{(1-M_t)(y_{iM}-x_{iM})}
\end{aligned}$$

Assuming Beta priors $p_i \sim \mathsf{Beta}(\alpha_{p_i},\beta_{p_i})$, $s_t \sim \mathsf{Beta}(\alpha_{s_t},\beta_{s_t})$, $c_t \sim \mathsf{Beta}(\alpha_{c_t},\beta_{c_t})$, the conditional posterior of p, s and c are also Beta:

$$\begin{aligned}
P(p_i|s,c,x,y) &\sim \mathsf{Beta}(\sum_{M}x_{iM}+\alpha_{p_i}, \sum_{M}y_{iM}-x_{iM}+\beta_{p_i})\\
P(s_t|p,c,x,y) &\sim \mathsf{Beta}(\sum_{M}\sum_{i}M_tx_{iM}+\alpha_{s_t}, \sum_{M}\sum_{i}(1-M_t)x_{iM}+\beta_{s_t})\\
P(c_t|s,c,x,y) &\sim \mathsf{Beta}(\sum_{M}\sum_{i}(1-M_t)(y_{iM}-x_{iM})+\alpha_{c_t}, \sum_{M}\sum_{i}M_t(y_{iM}-x_{iM})+\beta_{c_t})
\end{aligned}$$

For the latent variables x_{iM} the conditional posterior is binomial

$$p(x_{iM}|p, s, c, y) \sim \mathsf{Binomial}(y_{iM}, \frac{a_{iM}}{a_{iM} + b_{iM}})$$

where

$$a_{iM} = p_i \prod_{t=1}^{T} s_t^{M_t}(1 - s_t)^{1-M_t}$$

$$b_{iM} = (1 - p_i) \prod_{t=1}^{T} (1 - c_t)^{M_t} c_t^{1-M_t}$$