Baskets containing similar items

Suppose the p_0 and p_1 is the return probability of a basket that does not/does contains similar items, respectively. Let's define the following random variables:

- A: $A = 1$ indicates a basket contains similar items, and $A = 0$ indicates a basket does not conain similar items;
- R: $R = 1$ indicates a return, and $R = 0$ indicates no return
- Θ: a hidden random variable that includes all the other factors that can affect the return rates. In the algorithm, a target item shares the same Θ with its neigbours.

Therefore,

$$P\{R = 1|A = 0\} = p_0;$$

$$p\{R = 1|A = 1\} = p_1 = \kappa p_0.$$

Assumption: the return probability of a neighbor basket contains similar items and that of a neighbor basket does not contrain similar items still preserve the same ratio κ.

Mathematically, the two probabilities are defined as:

$$P\{R = 1|\Theta = \theta, A = 1\} = \frac{P\{\Theta = \theta|R = 1, A = 1\}P\{R = 1|A = 1\}}{P\{\Theta = \theta|A = 1\}} = q_1(\theta)$$

and

$$P\{R = 1|\Theta = \theta, A = 0\} = \frac{P\{\Theta = \theta|R = 1, A = 0\}P\{R = 1|A = 0\}}{P\{\Theta = \theta|A = 0\}} = q_0(\theta)$$

We need to assume that Θ is only independent or weakly dependent on A and R, then

$$\frac{q_1}{q_0} = \frac{P\{R = 1|\Theta = \theta_0, A = 1\}}{P\{R = 1|\Theta = \theta_0, A = 0\}} \approx \frac{P\{R = 1|A = 1\}}{P\{R = 1|A = 0\}} = \frac{p_1}{p_0} = \kappa$$

Estimate of q_0 and q_1:

Suppose that $R_i = r_i$ is observed under condition $\Theta = \theta_0$, and we need to estimate q_0 and q_1. The likelyhood function is the joint distribution of Bernoulli trails:

$$L = \prod_{i \in \{k|A_k=0\}} P\{R = r_i|\Theta = \theta_0, A_i = 0\} \prod_{j \in \{k|A_k=1\}} P\{R = r_j|\Theta = \theta_0, A_j = 1\}$$

or

$$L = (1 - q_0)^{\sum I_{r_i=0,a_i=0}} q_0^{\sum I_{r_i=1,a_i=0}} (1 - q_1)^{\sum I_{r_i=0,a_i=1}} q_1^{\sum I_{r_i=1,a_i=1}}$$

The log-likelyhood is then

$$l = \sum I_{r_i=0,a_i=0} \ln(1-q_0) + \sum I_{r_i=1,a_i=0} \ln q_0 + \sum I_{r_i=0,a_i=1} \ln(1-q_1) + \sum I_{r_i=1,a_i=1} \ln q_1$$

Maximum likelyhood is obtained by setting the derivate with respect to q_0 equal to 0. Derviative with respect to q_0 is given as

$$\frac{\partial l}{\partial q_0} = -\frac{\sum I_{r_i=0,a_i=0}}{1-q_0} + \frac{\sum I_{r_i=1,a_i=0}}{q_0} - \frac{\sum I_{r_i=0,a_i=1}}{1-q_1}\frac{\partial q_1}{\partial q_0} + \frac{\sum I_{r_i=1,a_i=1}}{q_1}\frac{\partial q_1}{\partial q_0}$$

The solution is then

$$q_0 = \frac{(1-a)\kappa + 1 - c - \sqrt{(1-a)^2\kappa^2 + (1-c)^2 + 2\,(ac + (a+c) - 1)\kappa}}{2\,\kappa}$$

where

$$a = \frac{\sum I_{r_i=0,a_i=0}}{\sum I}$$

$$c = \frac{\sum I_{r_i=0,a_i=1}}{\sum I}$$