Model Description

I am trying to model the transition probability of survey response. Each of the $N = 1700$ survey respondents are asked $K = 12$ questions in two waves, leading to data $Y_{n,k,t}$ with $n = 1, \ldots, N, k = 1, \ldots, K, t = 1, 2$. All answers $Y_{n,k,t}$ are in $1, 2, ..., 5$ scale. Some other demographic information is also collected so we can divide all responses into $L = 488$ cells according to the interaction of those demographic variables. Let $l(n)$ to be the index of cell that respondents n belongs to, and let $y1(n)$ and $y2(n)$ denotes the response of that person on wave 1 and 2.

The transition probability from response i to j on question k in cell i is denoted by $p_{i,j,k,l}$.

The likelihood is given by

$$\prod_{n,k} Pr(Y_{n,k,2}|Y_{n,k,1}) = \prod_n \prod_k p_{[i=y1(n),j=y2(n),k=k,l=l(n)]}$$

The transition probability should satisfy the 4-simplex condition for any given i, k, l:

$$\sum_{j=1}^{5} p_{i,j,k,l} = 1, \quad p_{i,j,k,l} \geq 0$$

We can imagine it should be a nearly symmetric distribution spiked at i. In practice, we model the log probability ratio $q \in R$:

$$q_{i,j,k,l} = \log \left(\frac{p_{i,j,k,l}}{p_{i,i,k,l}} \right)$$

so that the reference probability of making the same response in two waves is always $q_{i,i,k,l} = 0$.

Conversely,

$$p_{i,j,k,l} = \frac{\exp(q_{i,j,k,l})}{\sum_{j=1}^{5} \exp(q_{i,j,k,l})}$$

We model the probability of the second-wave response with decay proportional to the root distance $\sqrt{|i-j|}$. The larger distance, the less possible to make the jump. The main effect $\beta_{i,k,l} \in R$ captures such decay rate. A larger β indicates higher chance to stick to the initial answer.

To allow a potential asymmetric distribution, we further introduce $\gamma_k \in R$ to be the question level shift. It capture how likely to make a positive jump. Intuitively, when γ_k is negative, it makes the distribution of the transition sharper on the right tail, making it less likely to make a positive jump, so there will be a global negative shift, and vise versa.

$$q_{i,j,k} = -\big(\beta_{i,k,l} - \gamma_k I(j > i)\big) \sqrt{|i-j|}$$
$$\gamma_k \sim \mathrm{N}(0, \sigma_\gamma^2)$$

We also allow the main effect β to vary across initial value i, question k, and demographic variable l. To begin with , we decompose the transition probability as the summation of characteristic effects λ_l, question level effect μ_k and attitude-level effect η_i. All these parameters are unconstrained and centered at 0 (for identification). $\mu_0 \in R$ is the extracted constant term.

$$\beta_{i,k,l} \sim \mathrm{N}(\lambda_l + \mu_k + \eta_i + \mu_0, \sigma_\beta^2), \quad \mu_k \sim \mathrm{N}(0, \sigma_\mu^2) \quad \eta_i \sim \mathrm{N}(0, \sigma_\eta^2)$$

The demographic effect λ_l itself is modeled as the summation of the effects of each demographic variables, e.g, age, gender, marriage .... Let $age[l]$, $gender[l]$,..., denotes the corresponding variable values in cell l. All these variables are categorical. For example, age is in 1-6 scale, so $\lambda^{age}[t] (t = 1, \dots, 6)$ represents the age effect in cells with age$= t$.

$$\lambda_l = \lambda^{age}\big[age[l]\big] + \lambda^{gender}\big[gender[l]\big] + \dots,$$

$$\lambda^{age}[t] \sim \mathrm{N}(0, \sigma_{age}^2), \quad \lambda^{gender}[t] \sim \mathrm{N}(0, \sigma_{gender}^2), \dots$$