Group's heterogeneity in rating tasks: a Bayesian semi-parametric approach

Giuseppe Mignemi[1], Ioanna Manolopoulou[2], Antonio Calcagnì[1]
[1]Università di Padova (email: giuseppe.mignemi@phd.unipd.it, antonio.calcagni@unipd.it)
[2]University College London (email:i.manolopoulou@ucl.ac.u)

Abstract In several observational contexts where different raters evaluate a set of items, it is common to assume that all raters draw their scores from the same underlying distribution. A common distributional assumption in this setting is that hierarchical effects as independent and identically distributed from a normal with the mean parameter fixed to zero and unknown variance. The present work aims to overcome this strong assumption in the inter-rater agreement estimation by assigning a Dirichlet Process (DP) mixture as the hierarchical effects' prior distribution. A new semi-parametric index λ is proposed to quantify raters polarization in presence of group heterogeneity. The model is applied to a real context.

Key words: rating process, inter-rater agreement, Dirichlet mixture process, Bayesian nonparametrics

1 A semi-parametric model proposal

Several methods and statistical models that aim to account for inter-rater variability have appeared in the literature [3]. Despite the popularity of work on this issue, less attention has been paid to possible latent dissimilarities among raters within inter-rater agreement studies[4]. From a psychometric point of view, it can be appealing to assess the extent to which different raters could have different latent opinions for specific rating processes.

To this aim, Hierarchical Generalized Linear Models (HGLM) are a natural choice, since they can account for the individual-variability specifying the effect of m covariates. The HGLM assumption regarding the distribution of the hierarchical effects is crucial in characterising different possible clusters or latent patterns of heterogeneity among raters. To this aim a Dirichlet Process Prior is specified over the hierarchical effects and the model is specified as follow.

The rating $y_{ij} \in \{0,1\}$ of the item $j \in \{1,..,J\}$ carried out by rater $i \in \{1,..,I\}$, considering a set $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ of covariates for the different effects, respectively, is modelled as follows:

$$\begin{aligned}
P(y_{ij}=1) &= F(\mathbf{x}_{ij}\beta+\mathbf{z}_i\mathbf{u}_i+\varepsilon_{ij}),\\
\mathbf{u}_i|\mu_c,\mathbf{Q} &\sim N_q(\mu_c,\mathbf{Q}),\\
\mu_c|G &\sim G,\\
G &\sim DP(\alpha,G_0).
\end{aligned}$$

Here $F(\cdot)$ is a cumulative distribution function (e.g., Normal or Logistic), $N_q(\cdot)$ stands for a q-variate normal distribution, β is a $p\times 1$ vector of non hierarchical effects and $\mathbf{u}_i$ is a q vector of hierarchical effects. Here, $DP(\alpha,G_0)$ is a Dirichlet Process Mixture with $\alpha>0$ *precision parameter* and *base measure* G_0. It is assumed that $\mathbf{u}_i$ and ε_{ij} are independent.
The hierarchical effects distribution considering a stick breaking construction of the DPM might be then specified the as follow:

$$\mathbf{u}_i|\mu_c,\mathbf{Q},\alpha \overset{iid}{\sim} \sum_{c=1}^{R}\pi_c N_q(\mu_c,\mathbf{Q}), \quad i=i,\ldots,I$$

$$\mu_c \overset{iid}{\sim} G_0,$$

$$\pi_c = \nu_c\prod_{l<c}(1-\nu_l),$$

$$\nu_c \overset{iid}{\sim} Be(1,\alpha),\;\; c=1\ldots,R.$$

Where $R\in\mathbf{N}$ and large enough [1].

1.1 The λ index

The marginal posterior distribution of the hierarchical effects in the model outlined above captures information about the dissimilarity or disagreement among raters (on the assumption that the model captures the data adequately). To this end the full estimated distribution of $\mathbf{u}$ resulting from the model might be useful. At each iteration t, the density of $\mathbf{u}$ is given by the corresponding mixture model given the parameters at iteration t. Following the formulation of [1] , the set of modes and antimodes (i.e., the least frequent values between two consecutive modes) is identified; the latent disagreement λ is then defined as the log ratio between the mean density of the modes and the that of the antimodes:

$$\lambda = \ln\left(\frac{\frac{1}{M}\sum_{m=1}^{M} f_{\mathbf{u}}(\gamma_m)}{\frac{1}{A}\sum_{a=1}^{A} f_{\mathbf{u}}(\zeta_a)}\right)$$

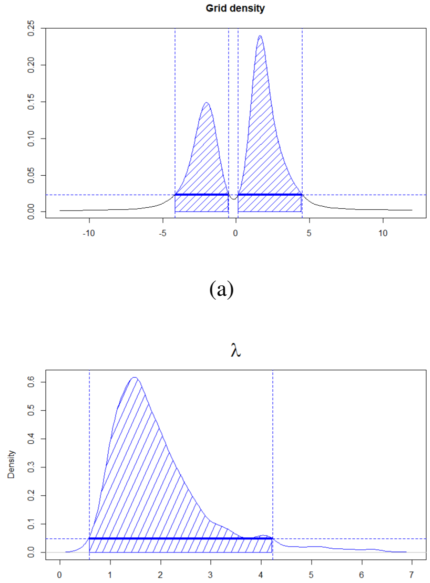
where M is the number of modes γ_m and A the number of antimodes ζ_a of $f(\mathbf{u})$. Larger values of λ indicate strongly multimodal distribution of the hierarchical effects, whereas smaller values are evidence of weak multimodality, thus the estimated hierarchical effects are less concentrated. In this sense this index is informative about the latent group polarization. Which in this context is assumed as a way of disagreement.

2 Posterior sampling and numerical example

As a numerical example a real data set from the social sciences context was analysed. Fifty-two personnel selectors were asked to rate 40 different applicants per rater on a binary scale (0=not selected, 1=selected). In this case, y_{ij} is the binary score given to applicant i by selector j. Selectors' years of experience and applicants' age were two covariate considered in the model. The effect of the latter was specified as hierarchical with the distributional assumption outlined in the previous section.

Since most of the parameters in the model have conjugate prior distributions a blocked Gibbs sampling algorithm was used for the posterior sampling. An underline latent variable approach accounting for the probit link function of the HGLM was adopted. Weakly informative priors were elicited following [2]. As suggested by [1], in order to estimate the density of $\mathbf{u}$ the approach of monitoring $\mathbf{u} \overset{iid}{\sim} \sum_{c=1}^{R} \pi_c N_q(\mu_c, Q)$ at each iteration over a dense grid of u values was adopted. At each iteration t, the density of the parametric mixture was computed at each point of the grid. As result of some prior predictive check, a dense grid of 481 equally-spaced values from -12 to 12 (i.e., with a fixed interval of 0.05) was used to monitoring the mixture density of the hierarchical effects. The maximum number of mixture component R through the stick-breaking construction was fixed to 25. In all the computations 80.000 iteration with 8.000 burn-in were used, the Markov chains were thinned the by a factor of 80, resulting in samples of size 1000.

As shown in table 1 selector's years of experience has a positive effect on th probability of being selected. The marginal posterior distribution of the hierarchical effect of applicant's age showed a bimodal distribution. More precisely the effect of this predictor is positive for a subgroup of the overall sample, whereas it is negative in the other one. The presence of this heterogeneity is shown also by the λ- index which HPD interval is far from zero and includes large values.



(a)

(b)

Fig. 1 95% HPD intervals of the grid density (a) and λ-index.

Table 1 95% HPD intervals

95% HPD intervals	
β	$(1.58, 3.33)$
b_β	$(-0.56, 3.41)$
σ_{B_β}	$(0.16, 4.17)$
μ_0	$(-0.28, 0.75)$
σ_{D_0}	$(2.14, 5.25)$
Q	$(0.09, 0.29)$
σ_ε	$(0.86, 3.08)$
α	$(6.69, 16.06)$
Grid density	$(-4.15, -0.5) \cup (0.15, 4.50)$

References

1. Gelman, A., Carlin, J., Stern, H., Dunson, D., and Vehtari, A.and Rubin, D. *Bayesian Data Analysis*. Chapman and Hall/CRC, 11 2013.
2. Heinzl, F., Kneib, T., and Fahrmeir, L. Additive mixed models with dirichlet process mixture and p-spline priors. *AStA Advances in Statistical Analysis 96* (05 2012).
3. Nelson, K., and Edwards, D. Measures of agreement between many raters for ordinal classifications. *Statistics in medicine 34* (06 2015).
4. Wirtz, M. A. *Interrater Reliability*. Springer International Publishing, Cham, 2020, pp. 2396–2399.