A Gibbs Sampler for Polytomous Rasch Model

Iman

Friday, October 10, 2008

1 Polytomous Rasch Model for Multi-Scale Speech Quality Judgment

We develop a Gibbs sampler for the Polytomous Rasch model of subject responses in a pairwise comparison test with M scales. In this test the subject is asked to compare the **quality** of speech signal x_2 against the quality of speech signal x_1 using M possible response alternatives (scales) represented by integers $\{1, ..., M\}$. For example in a pairwise comparison with two scales ($M = 2$), the responses correspond to 'worse' and 'better', and if we use five scales the responses could be 'worse', 'slightly worse', 'equal', 'slightly better', and 'better'.

First, let us begin by describing a hypothetical mechanism for subject response developed in [1] to describe the physical meaning of the Polytomous Rasch model parameters. This development is used later to develop an efficient Gibbs sampler. To make his decision, we assume that the subject compares M noisy latent (internal) representations of the difference in speech quality against $(M-1)$ thresholds. In other words, the subject decision is characterized by $(M-1)$ latent values. The subject chooses category n as its response, if for the first $(n-1)$ thresholds the corresponding latent values exceed the threshold and for the rest of $(M-n)$ thresholds, the corresponding latent values are smaller.

Having the qualitative response mechanism in mind, we proceed to give a mathematical account of the Polytomous Rasch model with M scales. The $(M-1)$ thresholds are denoted by $\{\tau_1, \cdots, \tau_{M-1}\}$. In our data set, we assume to have N pairwise comparison tests each comprised of test signals (x_i^1, x_i^2) and subject response $d_i \in \{1, ..., M\}$ for $i = 1, \cdots, N$. In the i-th test, we denote the latent representation corresponding to j-th threshold, τ_j, by $z_{i,j}$. We assume $z_{i,j}$ to follow a linear utility model with Gumbel noise,

$$\begin{aligned} z_{i,j} &= \Delta_i + \varepsilon_{i,j} \\ \Delta_i &= \boldsymbol{\omega}^{\mathrm{T}} \boldsymbol{v}_i \\ \boldsymbol{v}_i &= \boldsymbol{\phi}_i^1 - \boldsymbol{\phi}_i^2 \\ \varepsilon_{i,j} &\sim \frac{\exp(-\varepsilon_{i,j})}{(1+\exp(-\varepsilon_{i,j}))^2} \end{aligned} \tag{1}$$

where $\boldsymbol{\omega}$ is the feature weights and $\boldsymbol{v}_i = \boldsymbol{\phi}_i^1 - \boldsymbol{\phi}_i^2$ is the difference of feature vectors $\boldsymbol{\phi}_i^1$ (corresponding to signal x_i^1) and $\boldsymbol{\phi}_i^2$ (corresponding to signal x_i^2). The random noise follows the logistic distribution, with cumulative distribution function

$$p(\varepsilon_{i,j} < x) = \frac{1}{1 + \exp(-x)}. \tag{2}$$

The strange point about the response mechanism is that $\varepsilon_{i,j}$ are 'independent' for different scales [1].

According the response mechanism described earlier, the subject response d_i is chosen as follows

$$\begin{aligned} d_i = n \quad &\Leftrightarrow \quad \begin{cases} y_{i,j} = 1 & \text{for all } j < n \\ y_{i,j} = 0 & \text{for all } j \geq n \end{cases}, \\ y_{i,j} \quad &= \quad \mathrm{I}(z_{i,j} > \tau_j), \end{aligned} \tag{3}$$

where I(.) is the logical indicator function which takes on the value one when the inner statement is true.

We are now ready to present the Gibbs sampler, however, we finish this section by deriving the Rasch model for $p(d_i = n)$. For brevity, we drop the test index subscript (i). To compute the probability $p(d = n)$, we first compute the probability $p(y_j = 1)$ as follows

$$\begin{aligned} p(y_j = 1) = p(z_j > \tau_j) \quad &= \quad p(-\Delta + \tau_j < \varepsilon_j) \\ &= \quad \frac{\exp[\Delta - \tau_j]}{1 + \exp[\Delta - \tau_j]}, \end{aligned} \tag{4}$$

where Δ is introduced in Equation (1). Let us define the set $\mathcal{Y} = \{(y_1, \cdots, y_{M-1})\}$ containing 2^{M-1} vectors. Only M of these vectors correspond to a response in the model. The valid vectors start with $(n-1)$ 1's followed by $M - n$ 0's. If $\boldsymbol{y}_n \in \mathcal{Y}$ corresponds to $d = n$, the probability of $\boldsymbol{y}_n$ is

$$\begin{aligned} p(\boldsymbol{y}_1) \quad &= \quad \prod_{l=1}^{M-1} p(y_l = 0), \\ &= \quad \frac{1}{\prod_{l=1}^{M}(1 + \exp(\Delta - \tau_l))}, \\ p(\boldsymbol{y}_{n\neq 1}) \quad &= \quad \prod_{l=1}^{n-1} p(y_l = 1) \prod_{l=n}^{M-1} p(y_l = 0), \\ &= \quad \frac{\exp\left((n-1)\Delta - \sum_{l=1}^{n-1} \tau_l\right)}{\prod_{l=1}^{M}(1 + \exp(\Delta - \tau_l))}. \end{aligned} \tag{5}$$

The rest of vectors in $\mathcal{Y}$ are simply contradictory cases, where we expect the subject to resolve in some way until it comes up with a valid configuration. Hence we have to

compute the probability conditioned on the fact that only one of $\boldsymbol{y}_1, \cdots, \boldsymbol{y}_M$ is chosen. This conditioning amounts to normalizing $p(\boldsymbol{y}_1), \cdots, p(\boldsymbol{y}_M)$

$$\begin{aligned} p(d=n) &= \frac{p(\boldsymbol{y}_n)}{\sum_{l=1}^{l=M} p(\boldsymbol{y}_l)}, \\ &= \frac{\exp\left((n-1)\Delta - \sum_{l=1}^{n-1}\tau_l\right)}{1+\sum_{l=1}^{M}\exp\left((l-1)\Delta - \sum_{k=1}^{l-1}\tau_k\right)}, \end{aligned} \tag{6}$$

which is the probability formula of Polytomous Rasch model.

2 Gibbs Sampler

In this section we develop a Gibbs Sampler. We begin by presenting a new expression for $z_{i,j}$, defined for the first time in Equation (1), using a scale mixture of normal form [2]. This approach is adopted from [3].

$$\begin{aligned} y_{i,j} &= \begin{cases} 1 & \text{if } z_{i,j} > \tau_j \\ 0 & \text{otherwise,} \end{cases} \\ z_{i,j} &= \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{v}_i + \varepsilon_{i,j}, \\ \varepsilon_{i,j} &\sim N(0, \lambda_{i,j}), \\ \lambda_{i,j} &= (2\psi_{i,j})^2, \\ \psi_{i,j} &\sim \text{KS}, \end{aligned} \tag{7}$$

where $i = 1, \cdots, n$ and $j = 1, \cdots, M$ and $\boldsymbol{v}_i$ is defined in Equation (1). Furthermore, $\psi_{i,j}$ are independent random variables following the Kolmogorov-Smirnov (KS) distribution (e.g. [4]). In this case, $\varepsilon_{i,j}$ has a scale mixture of normal form with a marginal logistic distribution [2], so that the marginal likelihood $L(\boldsymbol{\omega}|y_{i,j})$ for models (7) and (1) are equivalent. [This part should be re-written]

We denote the probability distribution over the model parameters by $p(\boldsymbol{\omega}, \boldsymbol{\lambda}, \boldsymbol{z}, \boldsymbol{\tau}|\boldsymbol{V}, \boldsymbol{T})$, where

$$\begin{aligned} \boldsymbol{\lambda}_{N(M-1)\times 1} &= (\lambda_{1,1}, \lambda_{2,1}, \cdots, \lambda_{N,1}, \lambda_{1,2}, \cdots, \lambda_{N,M-1})^{\mathrm{T}}, \\ \boldsymbol{z}_{N(M-1)\times 1} &= (z_{1,1}, z_{2,1}, \cdots, z_{N,1}, z_{1,2}, \cdots, z_{N,M-1})^{\mathrm{T}}, \\ \boldsymbol{\tau}_{(M-1)\times 1} &= (\tau_1, \cdots, \tau_{M-1})^{\mathrm{T}}, \\ \boldsymbol{V}_{N\times 3} &= (\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N)^{\mathrm{T}}, \\ \boldsymbol{Y}_{N\times(M-1)} &= (\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N)^{\mathrm{T}}, \end{aligned}$$

where $\boldsymbol{y}_i = (y_{i,1}, \cdots, y_{i,M-1})^{\mathrm{Y}}$ are vectors of $y_{i,j}$ as defined as in (3).

We start by the probability of $\boldsymbol{z}$ conditioned on the rest of variables,

$$z_{i,j}|\boldsymbol{\omega}, \boldsymbol{v}_i, y_{i,j}, \lambda_{i,j}, \tau_{i,j} \propto \begin{cases} N(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{v}_i, \lambda_{i,j})I(z_{i,j} > \tau_{i,j}) & y_{i,j} = 1 \\ N(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{v}_i, \lambda_{i,j})I(z_{i,j} < \tau_{i,j}) & y_{i,j} = 0 \end{cases}. \tag{8}$$

Next, we proceed to the probability of $\boldsymbol{\omega}$ conditioned on the rest of variables. Given a normal prior $p(\boldsymbol{\omega}) = N(\mathbf{M}, \mathbf{S})$, the conditional posterior is normal

$$\begin{aligned} \boldsymbol{\omega}|\boldsymbol{z}, \boldsymbol{\lambda}, \boldsymbol{V} &\sim N(\mathbf{B}, \mathbf{V}), \\ \mathbf{B} &= \mathbf{V}(\mathbf{S}^{-1}\mathbf{M} + \mathbf{x}^{\mathrm{T}}\mathbf{W}\boldsymbol{z}), \\ \mathbf{V} &= (\mathbf{S}^{-1} + \mathbf{x}^{\mathrm{T}}\mathbf{W}\mathbf{x})^{-1}, \\ \mathbf{W} &= diag(\lambda_{1,1}^{-1}, \lambda_{2,1}^{-1}, \cdots, \lambda_{N,1}^{-1}, \lambda_{1,2}^{-1}, \cdots, \lambda_{N,M-1}^{-1}), \end{aligned} \tag{9}$$

where $\mathbf{x}_{N(M-1)\times 3} = (\boldsymbol{V}^{\mathrm{T}}, \cdots, \boldsymbol{V}^{\mathrm{T}})^{\mathrm{T}}$ consists of $M-1$ replications of $\boldsymbol{V}$.Afterwards, we compute the probability of $\boldsymbol{\tau}$ conditioned on the rest of variables. Given an improper flat prior we have

$$\begin{aligned} \tau_j|\boldsymbol{z} &\sim \begin{cases} \frac{1}{b_j - a_j} & a_j < \tau_i < b_j \\ 0 & \text{otherwise.} \end{cases}, \\ a_j &= \max(z_{i,j}|y_{i,j} = 0), \\ b_j &= \min(z_{i,j}|y_{i,j} = 1). \end{aligned} \tag{10}$$

Finally, we have to sample the distribution $p(\lambda_{i,j}|z_{i,j}, \boldsymbol{\omega})$ using rejection sampling as described in [3].

We can improve the mixing properties of the sampler by joint updating $\{\boldsymbol{z}, \boldsymbol{\lambda}\}$ given $\{\boldsymbol{\omega}, \boldsymbol{\tau}\}$,

$$p(\boldsymbol{z}, \boldsymbol{\lambda}|\boldsymbol{\omega}, \boldsymbol{\tau}, \boldsymbol{V}, \boldsymbol{Y}) = p(\boldsymbol{z}|\boldsymbol{\omega}, \boldsymbol{\tau}, \boldsymbol{V}, \boldsymbol{Y}) p(\boldsymbol{\lambda}|\boldsymbol{z}, \boldsymbol{\omega}),$$

where $p(\boldsymbol{\lambda}|\boldsymbol{z}, \boldsymbol{\omega})$ is unchanged but $\boldsymbol{z}$ is updated from its marginal distribution having integrated over $\boldsymbol{\lambda}$ [3]. The marginal density for $z_{i,j}$ is

$$z_{i,j}|\boldsymbol{\omega}, \tau_{i,j}, \boldsymbol{\upsilon}_i, y_{i,j} \propto \begin{cases} \text{Logistic}(z_{i,j}|\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\upsilon}_i, 1)\mathrm{I}(z_{i,j} > \tau_{i,j}) & y_{i,j} = 1 \\ \text{Logistic}(z_{i,j}|\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\upsilon}_i, 1)\mathrm{I}(z_{i,j} < \tau_{i,j}) & y_{i,j} = 0 \end{cases}, \tag{11}$$

where $\text{Logistic}(\epsilon|\mu, \gamma)$ is

$$\text{Logistic}(\epsilon|\mu, \gamma) = \frac{\exp(-(\epsilon - \mu)/\gamma)}{\gamma[1 + \exp(-(\epsilon - \mu)/\gamma)]^2}, \tag{12}$$

where $\gamma > 0$, and $-\infty < \epsilon < \infty$. In this case, joint updating $\{\boldsymbol{z}, \boldsymbol{\lambda}\}$ is followed by an update of $\{\boldsymbol{\omega}$ and $\boldsymbol{\tau}\}$, as before.

3 Maximum Likelihood for Polytomous Rasch Model

In this section we compute the first and the second derivatives of the likelihood function

$$S = \sum_{i=1}^{N} \log(p(d_i)), \tag{13}$$

where N is the total number of data points. The first derivatives are

$$\frac{\partial S}{\partial \omega_k} = \sum_{i=1}^{N} 1/p(d_i) \frac{\partial p(d_i)}{\partial \omega_k} \tag{14}$$

$$\frac{\partial S}{\partial \tau_m} = \sum_{i=1}^{N} 1/p(d_i) \frac{\partial p(d_i)}{\partial \tau_m} \tag{15}$$

$$\frac{\partial S}{\partial \alpha_m} = \sum_{i=1}^{N} 1/p(d_i) \frac{\partial p(d_i)}{\partial \alpha_m} \tag{16}$$

where $k = 1, \cdots, p$ for p features, and $m = 1, \cdots, M-1$ for M thresholds.

$$p_n = \frac{\exp(\beta_n)}{1 + \sum_{l=2}^{M} \exp(\beta_l)} \tag{17}$$

$$\beta_n = \sum_{k=1}^{n-1} \alpha_k (\Delta - \tau_k) \tag{18}$$

$$\frac{\partial p_n}{\partial \Delta} = p_n(\gamma_n - \sum_{l=2}^{M} \gamma_k p_k) \tag{19}$$

$$\sigma_n = (\gamma_n - \sum_{l=2}^{M} \gamma_k p_k) \tag{20}$$

$$\frac{\partial^2 p_n}{\partial \Delta^2} = p_n(\sigma_n^2 - \sum_{k=2}^{M} \sigma_k \gamma_k p_k) \tag{21}$$

4 Extensions, Conclusions and Future Work

4.1 Discrimination Factor

The discrimination factor in the original Polytomous Rasch model is an important parameter as it permits assigning different resolutions for different thresholds. As it is noted in [1]:

- For example, in response to an item which involves thresholds between neutral, agree, and strongly agree categories say, it may be easier to discriminate between neutral and agree than between agree and strongly agree. Or in rating a performance into one of the three categories fail, pass and credit pass, it may be easier to discriminate between fail and pass than between pass and credit pass."

Note that Unlike the scale in simple logistic regression, we have the sufficient statistics to infer it (this was the basic requirement of Rasch). Discrimination factor is currently absent in the current model. This is the Polytomous Rasch formula with discrimination factors $\alpha_1, \cdots, \alpha_{M-1}$

$$p(d=n) \quad = \quad \frac{\exp\left((\sum_{l=1}^{n-1}\alpha_l)\Delta - \sum_{l=1}^{n-1}\alpha_l\tau_l\right)}{1+\sum_{l=1}^{my}\exp\left((\sum_{k=1}^{l-1}\alpha_k)\Delta - \sum_{k=1}^{l-1}\alpha_k\tau_k\right)}. \tag{22}$$

This should be obtained by changing Equation (1) as follows:

$$\varepsilon_{i,j} \quad \sim \quad \alpha_j \exp(-\alpha_j\varepsilon_{i,j})\exp(-\exp(-\alpha_j\varepsilon_{i,j})), \tag{23}$$

The model based on scale mixture of normal form does not change except for λ

$$\lambda_{i,j} = (2\sqrt{\alpha_j}\psi_{i,j})^2$$

To include the discrimination factor in the Gibbs sampler we have to consider the followings:

- The existing update rules of Gibbs sampler will not change except for updating λ.
- An update rule for $p(\alpha_j|\lambda_{1,j}, \cdots, \lambda_{n,j})$ as well as a prior $p(\alpha_1, \cdots, \alpha_M)$ should be developed, also, it may change the mixing formula.

4.2 Analysis of Henriëtte's Data

The comment that I have for the current analysis of Henriëtte's Data has three aspects:

- I think we should allow for negative thresholds, unlike the current analysis.
- I think the feature gains $\boldsymbol{\omega}$ are positive,
- We should check if the inferred threshold are ordered as expected. This is actually a test of the hypothesis that response classes are ordered.

4.3 Information Theoretic Question

Instead of focusing on relating $\boldsymbol{\omega}$ for different scales, we should possibly focus on discrimination factors and consistency rate across scales for two reasons.

First, because it looks a reasonable hypothesis that across different scales the $\boldsymbol{\omega}$ is fixed (If discrimination factors are included in the model). This hypothesis amounts to

the interpretation that the utility is invariant W.R.T the scale. If discrimination are not included in the model, then it looks reasonable that ratios $\frac{\omega_1}{\omega_2}$ and $\frac{\omega_1}{\omega_3}$ are fixed across scales. These ratios essentially tell us about the relative importance of features against each other.

Second, for fewer number of scales the choices are well distinguished and for large number of scales, they are less distinguished. On the other hand, larger scales potentially permit larger information content per response (e.g. for $my = 2$, each response is equal to at most one bit of information, and for $M = 8$ we gain at most 3 bits of information per response). There seem to exist an information theoretic trad-off here, similar to M-ary communication channels. An information theoretic formulation of this problems is interesting, and I am going to study it in more depth.

4.4 Future Work

- Analysing Henriëtte's data again with no limitations on thresholds and checking the hypothesis that that ratios $\frac{\omega_1}{\omega_2}$ and $\frac{\omega_1}{\omega_3}$ are fixed across scales.
- Analysing Henriëtte's using ML, with a fixed $\boldsymbol{\omega}$ and discrimination factors.
- Checking the Gibbs sampler by simulation on a simple problem.
- Contrasting our sampler against existing similar samplers in the literature survey and correct the introduction accordingly.
- Adding the discrimination factor to our Sampler. We should discuss updating λ in detail with each other, it is appeared just before Appendix A4 of Holmes and Held.
- Formalizing the information theoretic problem rigourously.

References

[1] D. Andrich, "A rating formulation for ordered response categories," *Psychometrika*, vol. 43, no. 4, pp. 561–573, 1978.

[2] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *J. Royal Stat. Soc*, vol. 36, pp. 99–102, 1974.

[3] C. Holmes and L. Held, "Bayesian auxiliary variable models for binary and multinomial regression," *Bayesian Analysis*, vol. 1, no. 1, pp. 145–168, 2006.

[4] L. Devroye, *Non-uniform random variate generation*. Springer-Verlag New York, 1986.