A Sequential Decision Approach to Ordinal Preferences in Recommender Systems

Truyen Tran[†], **Dinh Q. Phung**[‡] **and Svetha Venkatesh**[‡]
[†]Curtin University, Australia
t.tran2@curtin.edu.au
[‡]Deakin University, Australia
{*dinh.phung,svetha.venkatesh*}@*deakin.edu.au*

Abstract

We propose a novel sequential decision approach to modeling ordinal ratings in collaborative filtering problems. The rating process is assumed to start from the lowest level, evaluates against the latent utility at the corresponding level and moves up until a suitable ordinal level is found. Crucial to this generative process is the underlying utility random variables that govern the generation of ratings and their modelling choices. To this end, we make a novel use of the generalised extreme value distributions, which is found to be particularly suitable for our modeling tasks and at the same time, facilitate our inference and learning procedure. The proposed approach is flexible to incorporate features from both the user and the item. We evaluate the proposed framework on three well-known datasets: MovieLens, Dating Agency and Netflix. In all cases, it is demonstrated that the proposed work is competitive against state-of-the-art collaborative filtering methods.

Introduction

Collaborative filtering is currently the prominent approach to learning users' preferences from rating data. One of the most effective techniques so far is *matrix factorisation* which seeks a low-rank representation of the incomplete rating matrix so that the reconstruction error is minimised (Koren 2010; Takács et al. 2009; Salakhutdinov and Mnih 2008). This effectively treats rating as a Gaussian random variable whilst there is no reason why the rating should obey such law. Rating is a qualitative and ordinal assessment: the order between rating values is of intrinsic significance but not their absolute and relative values. Another technique is to consider rating as an unordered categorical variable from which a multinomial model can be constructed (Koren and Sill 2011; Salakhutdinov, Mnih, and Hinton 2007). However, this is inefficient since the ordinal constraint is lost resulting in models larger than necessary.

Recent work has recognised the importance of ordinal approach for collaborative filtering data (Koren and Sill 2011; Truyen, Phung, and Venkatesh 2009; Paquet, Thomson, and Winther 2011; Yu et al. 2006). This results in a number of advantages: one can model the rating process directly with weaker assumptions, explain the recommendation to users more naturally, and might result in a better predictive model. The most popularly-used *cumulative* approach assumes the existence of a latent *utility* per user-item pair, and an ordinal level is selected given the utility falling into an designated interval (McCullagh 1980). The main drawback of this approach lies in its rather restrictive assumption: while it is suitable for the case where such a utility naturally support the data (e.g., income ranges), it may not rich enough to model more challenging data such as those arise from multi-dimensional rating/assessment processes (Anderson 1984).

Alternatively, this paper advocates a *sequential decision* approach to modelling ordinal ratings (Mare 1980; Albert and Chib 2001). Under the proposed scheme, the user will first evaluate an item against the lowest ordinal level. If the item is deemed to be better than this level, it gets moved up to the next level. The process continues until a suitable level is reached, otherwise the item is assigned to the top level as the default. Different from the cumulative approach, each item in our proposed sequential approach is assumed to 'perceive' against $L-1$ latent utilities and $L-1$ corresponding thresholds. This offers a much more flexible way to model each user and item according to different level-specific criteria.

Crucial to our approach is the modeling choices for the latent utility random variables. In this paper we propose to investigate two classes of distributions: the *mean values* which give rise to normal distributions, and the *extreme values* which lead to *generalized extreme value* (GEV) distributions (Gumbel 1958). Whilst the normal distributions have been well studied in collaborative filtering, to the best of our knowledge, the GEV distributions have not been investigated. GEV distributions arise as a result of modelling the maximal values of some quantities such as the strongest wind in a year at a particular location. As such, the GEV family is suitable for modelling *rare events* such as ratings in the collaborative filtering setting. In particular, it is expected to work well with the sequential approach since we are interested in the utility *exceedance* after each ordinal level. To sum up intuitively, one can think of a generative process follows: for each item and an ordinal level of interest, the user has a number of *random* assessments according to different criteria of the context, each results in a utility value; and depending on the choice of the distribution, either the mean or maximum value will be used as the true utility of that item. This utility then evaluated against, first, the lowest level-specific threshold. If on average the utility exceeds the threshold, we move to the next level and repeat the whole process; otherwise, the current ordinal level is selected as the rating.

We evaluate the proposed methods on three well-known collaborative filtering datasets: the MovieLens (1 million ratings), the Dating agency (17 million ratings) and the Netflix (100 million ratings). In all cases, it is demonstrated that our proposed sequential approach is competitive against state-of-the-art collaborative filtering methods.

A Sequential Decision Model for Collaborative Ordinal Regression

Let L denote the number of ordinal levels. Assume that for a particular item i, a user u will perceive a set of utilities $\{x_{uil} \in \mathbb{R}\}_{l=1}^{L-1}$ corresponding to the first $L-1$ ordinal levels[1]. Since the nature of such a perception process is unknown, we assume that each utility is a random variable with additive noise:

$$x_{uil} = \mu_{ui} + \epsilon_{uil} \tag{1}$$

where μ is a function of (u, i) representing the structure of the data[2] and ϵ_{uil} is the random noise. Whilst we cannot directly measure the perceived utilities, the observed data provides evidences to estimate their distributions. To do so, when judging the value of the item i against ordinal levels, we further assume that each user also perceives $L-1$ corresponding thresholds, i.e., $\boldsymbol{\tau}_{ui} = \{\tau_{ui1}, \tau_{ui2}, .., \tau_{ui(L-1)}\}$. The user will first consider if the perceived utility x_{ui1} exceeds the his or her specific ordinal threshold τ_{ui1} defined at the the lowest level of the scale. If it does, then the user will evaluate the next utility x_{ui2} against the threshold τ_{ui2}. The process stops at level l if the utility falls below the threshold τ_{uil}. If none of the first $L-1$ has been found, the the top level will be chosen. This stage-wise process is also known as the *sequential ordering* or *continuation ratio* models in the literature (Mare 1980; Albert and Chib 2001).

Denote by r_{ui} the rating given to item i by user u. The sequential ordering process can be formally described as follows. First, at the lowest level:

$$\begin{aligned} P(r_{ui} = 1) &= P_1\left(x_{ui1} \leq \tau_{ui1}\right) \\ &= F_1\left(\frac{\tau_{ui1} - \mu_{ui}}{s_{ui}}\right) = F_1\left(\tau_{ui1}^{*}\right) \end{aligned}$$

where $F_1\left(\frac{\tau_{uil} - \mu_{ui}}{s_{ui}}\right) = P_1(x_{uil} \leq \tau_{uil})$ is the cumulative distribution function (CDF), $s_{ui} > 0$ is the *scale* parameter, and $\tau_{ui1}^{*} = \frac{\tau_{ui1} - \mu_{ui}}{s_{ui}}$.

At the subsequent levels $l = 2, 3, ..., L-1$, we must ensure that we have failed to pick the previous levels (i.e., $x_{uim} \geq \tau_{uim}$ for $m < l$) and the current level is feasible (i.e., $x_{uil} \leq \tau_{uil}$): $P(r_{ui} = l) =$

$$\begin{aligned} &= P\left(x_{uil} \leq \tau_{uil}, \{x_{uim} \geq \tau_{uim}\}_{m=1}^{l-1}\right) \\ &= \left\{\int_{-\infty}^{\tau_{uil}} P(x_{uil}) dx_{uil}\right\} \prod_{m=1}^{l-1} \left\{\int_{\tau_{uim}}^{\infty} P(x_{uim}) dx_{uim}\right\} \\ &= F_l\left(\tau_{uil}^{*}\right) \prod_{m=1}^{l-1} \left\{1 - F_m\left(\tau_{uim}^{*}\right)\right\} \end{aligned} \tag{2}$$

where we have made use of the fact that $\int_{\tau_{uim}}^{\infty} P(x_{uim}) dx_{uim} = 1 - \int_{-\infty}^{\tau_{uim}} P(x_{uim}) dx_{uim}$.

Finally, at the last level when we have failed at all previous levels (i.e., $x_{uim} \geq \tau_{uim}$ for $m < L$):

$$\begin{aligned} P(r_{ui} = L) &= P\left(\{x_{uim} \geq \tau_{uim}\}_{m=1}^{L-1}\right) \\ &= \prod_{m=1}^{L-1} \left\{1 - F_m\left(\tau_{uim}^{*}\right)\right\} \end{aligned} \tag{3}$$

One can verify that indeed[3] $\sum_{l=1}^{L} P(r_{uil} = l) = 1$. Alternatively, we can summarise the sequential decision process by

$$\begin{aligned} P(r_{ui} = l \mid r_{ui} \geq l) &= \frac{P(r_{ui} = l, r_{ui} \geq l)}{P(r_{ui} \geq l)} \\ &= \frac{P(r_{ui} = l)}{P(r_{ui} \geq l)} = F_l\left(\tau_{uil}^{*}\right) \end{aligned}$$

since $r_{ui} = l$ would imply that $r_{ui} \geq l$ and $P(r_{ui} \geq l) = \prod_{m=1}^{l-1} P(x_{uim} \geq \tau_{uim}) = \prod_{m=1}^{l-1} \left\{1 - F_m\left(\tau_{uim}^{*}\right)\right\}$. Thus, the odds $\frac{P(r_{ui}=l)}{P(r_{ui} \geq l)}$ reflects the name *continuation ratio*.

To sum up, each user assumes his or her own perception about what the utilities should be, e.g., by evaluating against user-specific thresholds. The evaluation is sequential in nature, selecting a suitable level only if the utilities have failed all lower levels. Lastly, we remark that the subscript l in the CDF $F_l(\tau_{uil}^{*})$ for $l = 1, 2, ..., L-1$ is used deliberately since one has the flexibility to choose different distribution family at different ordinal level l, as presented in sequel.

Utility Distribution Families

We now proceed to specify the nature of the noise by defining the form of $P(x_{uil} \leq \tau_{uil}) = F\left(\frac{\tau_{uil} - \mu_{ui}}{s_{ui}}\right)$. Before doing so, it's worth asking where the perceived utility x_{uil} may come from. Let us assume further that when assessing an item i against level l the user u makes a sequence of utility estimates $z_{uil}^{(1)}, z_{uil}^{(2)}, ..., z_{uil}^{(n)}$ and finally choose a value x_{uil}. We investigate two forms of decisions which might result in significantly different modeling behaviors: the *mean values* and the *extreme values*.

[1] The last level is not modelled explicitly because if no suitable ordinal level can be chosen for the first $L-1$ levels then the last level L will be chosen by definition.

[2] One may enrich the model by making μ depend on the ordinal levels as well but we choose otherwise to keep the model compact in this paper. Statistically speaking, since the parameters are shared among ordinal levels, this usually results in a more robust estimation.

[3] Since $P(r_{ui} \leq L-1) = \sum_{m=1}^{L-1} P(r_{ui} = m)$ we need to show that $P(r_{ui} = L) = P(r_{ui} > L-1)$. We can prove by induction as follows. Assume that $P(r_{ui} > l-1) = \prod_{m=1}^{l-1} \{1 - F_m(\tau_{uim}^{*})\}$. Using $P(r_{ui} > l) = P(r_{ui} > l - 1) - P(r_{ui} = l)$ and Eq. (2) we would yield $P(r_{ui} > l) = \prod_{m=1}^{l} \{1 - F_m(\tau_{uim}^{*})\}$. When $l = L-1$, this is essentially $P(r_{ui} = L)$ in Eq. (3).

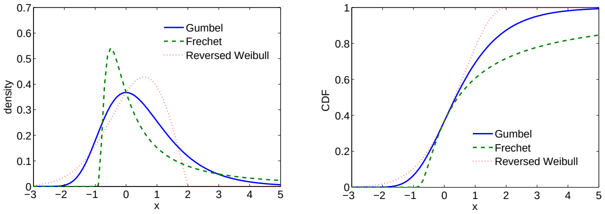

Figure 1: The GEV family. *Left*: the density function. *Right*: the cumulative distribution function.

Mean Values. Mean-value distributions arise when one assume the perceived utility has the form of an average over the utility estimates:

$$x_{uil} = \frac{1}{n}\sum_{m=1}^{n} z_{uil}^{(m)}$$

The asymptotic distribution of x_{uil} is the well known Gaussian distribution (CLT theorem)[4], i.e., $P(x_{uil} \leq \tau_{uil}) = \Phi\left(\frac{\tau_{uil}-\mu_{ui}}{s_{ui}}\right)$ where s_{ui} is the standard deviation. A closely related distribution is the logistic, where

$$P(x_{ui} \leq \tau_{uil}) = \frac{1}{1+\exp\left(-\frac{\tau_{uil}-\mu_{ui}}{s_{ui}}\right)}$$

The shape of the density functions of these two distributions is symmetric, but the logistic has higher variance given the same scale parameter.

Extreme Values. As in the sequential decision model we are interested in the probability at which the perceived utility surpasses a level-specific threshold, e.g., $P(x_{uil} > \tau_{uil}) = 1 - P(x_{uil} \leq \tau_{uil})$, it may be natural to suggest that the utility is a maximum value:

$$x_{uil} = \max_m \{z_{uil}^{(m)}\}_{m=1}^{n} \tag{4}$$

The *extreme value theory* states that the asymptotic probability $P(x_{uil} \leq \tau_{uil})$ must belong to a family known as *generalized extreme value* (GEV) distributions (Gumbel 1958):

$$P(x_{uil} \leq \tau_{uil}) = \exp\left(-g\left(\frac{\tau_{uil}-\mu_{ui}}{s_{ui}}\right)\right) \quad \text{where}$$

$$g(y) = \begin{cases} (1+\xi y)^{-1/\xi} & \xi \neq 0 \\ e^{-y} & \xi = 0 \end{cases}$$

under the support domain of $\xi y + 1 \geq 0$ with the *shape* ξ, *location* μ_{ui} and *scale* s_{ui}. When $\xi = 0$, the model is known as the Gumbel distribution:

$$P(x_{uil} \leq \tau_{uil}) = \exp\left(-e^{-(\tau_{uil}-\mu_{ui})/s_{ui}}\right) \tag{5}$$

For other cases, the model is the Fréchet distribution for $\xi > 0$, and the reversed Weibull distribution for $\xi < 0$. See Fig. 1 for the distributions of these three cases. In practice, the case of $\xi = 0$ (Gumbel) appears to be the most common.

The main difference from the mean value distributions is that the shape of the density function of the GEV family is asymmetric. For example, the Gumbel and the Fréchet distributions are heavily-tailed, that is, they allocate more probability mass for those utility values far from the median.

Model Specification and Estimation

We can expect that the location parameter μ_{ui} captures the data structure (e.g., the mean or median utility), and the scale parameter s_{ui} specifies the variation in the quality perception and the rating choices[5]. More specifically, s_{ui} reflects the fact that some users may choose to rate only a few high quality items but some others are willing to input a wide range of them. Similarly, some items may receive general agreement on quality among users, but some others may cause controversy.

Parameterisation

Assume that we have N users and M items. Let $\mathcal{I}(u)$ be the set of items rated by the user u, and $\mathcal{U}(i)$ the set of users who rate the item i. The location structure is decomposed as follows[6]

$$\mu_{ui} = \alpha_u + \beta_i + \boldsymbol{p}_u'\boldsymbol{q}_i + \\ + \sum_{j\in\mathcal{I}(u),j\neq i} w_{ij} f(r_{uj}) + \sum_{v\in\mathcal{U}(i),v\neq u} \omega_{uv} g(r_{vi}) \tag{6}$$

where $\alpha_u, \beta_i \in \mathbb{R}$ are user and item *biases*, $\boldsymbol{p}_u, \boldsymbol{q}_i \in \mathbb{R}^K$ are *latent features*, w_{ij} is the item-item correlation weight, ω_{uv} is the user-user correlation weight and $f(r_{uj})$ and $g(r_{vi})$ are some feature functions. The biases reflect the notions that some users are biased in their ratings (e.g., they tend to rate higher than average, or rate only what they like), and some items are inherently high (or low) quality. The inner product $\boldsymbol{p}_u'\boldsymbol{q}_i$ measures the *compatibility* between user u and item i in the latent feature space. It can also be seen as a *low-rank approximation* to the location matrix.

The last two components captures the known fact that items co-rated by the same user and users who co-rate the same item tend to correlate[7] (Resnick et al. 1994; Sarwar et al. 2001). Put differently, the co-rated item set $\{j \mid j \in \mathcal{I}(u), j \neq i\}$ and those co-rating user set $\{v \mid v \in \mathcal{U}(i), v \neq u\}$ provide *contextual features* for the pair (u, i). To keep the number of contextual features manageable, we keep only $J \ll \min\{N, M\}$ most popular items and users[8]. In our implementation, we choose the feature functions as $f(r) = g(r) = \frac{r}{L} - 0.5$.

The thresholds are defined as

$$\tau_{uil} = \gamma_{ul} + \lambda_{il} \tag{7}$$

for $l = 1, 2, ..., L-1$. Since the scale parameters are positive, we parameterise as

$$s_{ui} = e^{\nu_u + \eta_i} \tag{8}$$

[4] This is not to say that the rating is normally distributed, only its underlying generating variable is.

[5] For example, standard statistics for the Gumbel distribution: the mean is approximately $\mu_{ui} + 0.5772 s_{ui}$, the median is $\mu_{ui} - s_{ui} \log\log 2$ and the variance is $s_{ui}^2 \pi^2/6$.

[6] The inclusion of neighbour ratings essentially defines a conditional distribution $P\left(r_{ui}|\mathcal{I}(u), \mathcal{U}(i)\right)$. This is somewhat similar to the pseudo-likelihood model and the dependency networks (Heckerman et al. 2001).

[7] To the best of knowledge, these specific features are our contributions.

[8] Typically we would want J to be in the order of hundreds or thousands as opposed to the number of users or items, which can be in the order of hundreds of thousands, or even millions

For robustness, we can specify the minimum scale by ensuring that $s_{ui} \geq s_0$ – this effectively prevents the distribution from collapsing into a single point mass, which is often a sign of overfitting.

We also assume Gaussian priors over parameters, i.e., $\alpha_u \sim \mathcal{N}(0, \sigma_\alpha)$, $\beta_i \sim \mathcal{N}(0, \sigma_\beta)$, $\boldsymbol{p}_u, \boldsymbol{q}_i \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}^K)$, $w_{ij} \sim \mathcal{N}(0, \sigma_w)$, $\omega_{uv} \sim \mathcal{N}(0, \sigma_\omega)$, $\boldsymbol{\gamma}_u \sim \mathcal{N}(\mathbf{0}, \sigma_\gamma \mathbf{I}^{L-1})$, $\boldsymbol{\lambda}_i \sim \mathcal{N}(\mathbf{0}, \sigma_\lambda \mathbf{I}^{L-1})$, $\nu_u \sim \mathcal{N}(0, \sigma_\nu)$ and $\eta_i \sim \mathcal{N}(0, \sigma_\eta)$, where $\mathbf{I}^{L-1}$ and $\mathbf{I}^K$ are the identity matrices of size $(L-1)$ and K, respectively. Denote by $\boldsymbol{\theta}$ the vector of all parameters[9], thus $\boldsymbol{\theta} \in \mathbb{R}^{(N+M)(K+L+J+1)}$.

Learning

Learning using maximum a posterior (MAP) maximises the regularised log-likelihood with respect to parameters $\boldsymbol{\theta}$

$$\mathcal{L} = \sum_u \sum_{i \in R(u)} \log P(r_{ui} \mid \boldsymbol{\theta}) + \log P(\boldsymbol{\theta})$$

The gradient of the log-likelihood reads

$$\partial \mathcal{L} = \sum_u \sum_{i \in R(u)} \partial \log P(r_{ui} \mid \boldsymbol{\theta}) + \partial \log P(\boldsymbol{\theta})$$

where

$$\partial \log P(r_{ui} = l \mid \boldsymbol{\theta}) = \frac{\partial F_l\left(\tau^*_{uil}\right)}{F_l\left(\tau^*_{uil}\right)} - \sum_{m=1}^{l-1} \frac{\partial F_m\left(\tau^*_{uim}\right)}{1 - F_m\left(\tau^*_{uim}\right)}$$

Since τ^*_{uil} is a function of model parameters, we can use the following relation

$$\partial_{\tau^*_{uil}} F\left(\tau^*_{uil}\right) = P(\tau^*_{uil})$$

to compute $\partial_{\tau^*_{uil}} \log P(r_{ui} = l \mid \boldsymbol{\theta})$ and the chain rule to estimate the other derivatives:

$$\partial_{\boldsymbol{p}_u} \log P(r_{ui} \mid \boldsymbol{\theta}) = -\frac{\boldsymbol{q}_i}{s_{ui}} \partial_{\tau^*_{uil}} \log P(r_{ui} \mid \boldsymbol{\theta})$$

$$\partial_{\boldsymbol{q}_i} \log P(r_{ui} \mid \boldsymbol{\theta}) = -\frac{\boldsymbol{q}_u}{s_{ui}} \partial_{\tau^*_{uil}} \log P(r_{ui} \mid \boldsymbol{\theta})$$

$$\partial_{\nu_u} \log P(r_{ui} \mid \boldsymbol{\theta}) = -\tau^*_{uil} \partial_{\tau^*_{uil}} \log P(r_{ui} \mid \boldsymbol{\theta})$$

The regularised log-likelihood $\mathcal{L}$ is unfortunately non-convex in both $\boldsymbol{p}_u$ and $\boldsymbol{q}_i$. Thus, we suggests an alternating approach by looping through:

1. Fix $\boldsymbol{\beta}, \{\boldsymbol{q}_i, \boldsymbol{\lambda}_i, \eta_i\}_{i=1}^M, \boldsymbol{w}$, maximise $\mathcal{L}$ with respect to $\boldsymbol{\alpha}, \{\boldsymbol{p}_u, \boldsymbol{\gamma}_u, \nu_u\}_{u=1}^N, \boldsymbol{\omega}$, and
2. Fix $\boldsymbol{\alpha}, \{\boldsymbol{p}_u, \boldsymbol{\gamma}_u, \nu_u\}_{u=1}^N, \boldsymbol{\omega}$, maximise $\mathcal{L}$ with respect to $\boldsymbol{\beta}, \{\boldsymbol{q}_i, \boldsymbol{\lambda}_i, \eta_i\}_{i=1}^M, \boldsymbol{w}$.

In our implementation, a simple gradient ascent is used at each step. To speed up, the parameter update is made after every block of items at step 1 or every block of users at step 2. Typical block size is $100 - 1000$.

Prediction

Prediction of ordinal level of an unseen item j is then

$$\hat{r}_{uj} = \arg\max_{r_{uj}} P(r_{uj} \mid \boldsymbol{\theta}) \tag{9}$$

and $P(\hat{r}_{uj} \mid \boldsymbol{\theta})$ can be used as a *confidence measure* of the prediction. Alternatively, we can compute the expected rating as follows

$$\hat{r}_{uj} = \sum_{r_{uj}} P(r_{uj} \mid \boldsymbol{\theta}) r_{uj} \tag{10}$$

Related Work and Discussion

The additive model in Eq. (1) falls into the general category of random utility models (RUMs) (e.g., see (McFadden 1980)). Under RUMs, utilities play the role of underlying property of observed variables (such as ratings). This class is general and encompassing all the models considered in this paper. There are two related aspects of ordinal modelling under the RUM framework: the *ordinal assumption* and the *distribution family* of the underlying utilities.

The standard matrix factorisation approach can be explained in the RUM framework by fixing the utility to rating values, i.e., $x_{ui} = r_{ui} = \boldsymbol{p}'_u \boldsymbol{q}_i + \epsilon_{ui}$, and letting the error term be Gaussian, i.e., $\epsilon_{ui} \sim \mathcal{N}(0, 1)$. This is convenient computationally but makes it hard to interpret the qualitative nature of human preferences. For example, we cannot simply assign numbers to expressions such as {*very good, somewhat good, neither good or bad, somewhat bad, very bad*}.

Another treatment of ordinal ratings is by using standard categorical variables in a multinomial framework. In particular, for each rating level l, we assumes an independent utility $x_{uil} = \mu_{uil} + \epsilon_{uil}$, and the selection of a level is according to the maximum utility: $r_{ui} = l$ if $x_{uil} \geq \max_{m \neq l}\{x_{uim}\}$. It can be shown that for the choice of Gumbel distributions (a.k.a. GEV, $\xi = 0$), we actually arrive at the standard soft-max form (McFadden 1973): $P(r_{ui} = l \mid \boldsymbol{\mu}) = \exp(\mu_{uil}) / \sum_m \exp(\mu_{uim})$. The main drawback of this approach is that it usually requires $L - 1$ times as many parameters as standard ordinal methods, leading to slower learning and inference and less robust model estimation.

The cumulative ordinal regression model of (McCullagh 1980; Koren and Sill 2011; Paquet, Thomson, and Winther 2011) assumes one utility variable per user-item pair, and

$$r_{ui} = l \quad \text{if} \quad x_{ui} \in [\tau_{l-1}, \tau_l]$$

or equivalently

$$P(r_{ui} = l \mid \mu_{ui}, \boldsymbol{\tau}) = F(\tau_l - \mu_{ui}) - F(\tau_{l-1} - \mu_{ui})$$

for some CDF $F(\cdot)$ and $-\infty < \tau_1 < ... < \tau_L = \infty$. In (Koren and Sill 2011) the thresholds are user-specific, i.e., τ_{ul} is used instead of the standard τ_l.

It has also been shown that the cumulative approach and the sequential approach, whether there is a single utility variable per user-item pair, are identical under the Gompertz distribution (Läärä and Matthews 1985)

$$P(x_{ui} \leq \tau_{uil}) = 1 - \exp\left(-e^{(\tau_{uil} - \mu_{ui})/s_{ui}}\right)$$

In fact, the Gompertz distribution also belongs to the GEV family but with in a reverse way: we model the the minimum utilities, i.e., $x_{ui} = \min_m \{z_{ui}^{(m)}\}_{m=1}^n$, as opposed to the maximum utilities as in Eq. (4) (Gumbel 1958). The sequential model is however more flexible since we neither have to limit to a single utility variable per user-item pair nor use the same distribution family for all ordinal levels.

[9] These seem to be too many parameters, but without contextual features, they are the same to other matrix factorisation methods.

Experiments

Data and Setting

We evaluate our proposed approach on three rating datasets: the MovieLens[10], the Dating[11] and the Netflix[12]. The MovieLens dataset contains roughly 1 million ratings on the 5-star scale by 6 thousand users on nearly 4 thousand movies. The Dating dataset has nearly 17 million ratings on the 10-point scale by 135 thousand users on nearly 169 thousand profiles. The Netflix dataset consists of 100 millions ratings on the 5-star scale by 480 thousand users on nearly 18 thousand movies. To create more uniform evaluation, we convert the Dating dataset into the 5-point integer scale.

For each dataset, we keep only those users who have no less than 30 ratings. Of these we reserve 5 items for validation (e.g., for hyper-parameter setting & stopping criterion), 10 items for testing, and the rest for training. As the MovieLens and the Netflix datasets also has time-stamps, the training items are rated before the validation and validation before the testing. For the Dating, the selection is random.

Three performance metrics are reported on test data: *the root-mean square error* (RMSE), the *mean absolute error* (MAE) and the *log-likelihood* (LL). The use of RMSE, probably popular due to the Netflix challenge, implicitly assumes that the rating is Gaussian, which may not be very desirable.

Implementation

For comparison, we implement the standard probabilistic matrix factorisation with following rating assumptions: Gaussian (Salakhutdinov and Mnih 2008), multinomial assumption (see also (Koren and Sill 2011)), and cumulative (Koren and Sill 2011; Paquet, Thomson, and Winther 2011). The multinomial and cumulative models share the same location structure as in Eq. (6) but without the context features (matrix factorisation methods are essentially about low-rank decomposition only). All methods are trained by maximising their regularised log-likelihood, also known as MAP estimation. Prediction in discrete methods (multinomial, cumulative and sequential) can be performed using either the optimal choice as in Eq. (9), or expected rating as in Eq. (10).

Our implementation of the cumulative model differs from that (Koren and Sill 2011) in a number of ways: first, we investigate the case where the threshold depend both on user and item:

$$\tau_{uil} = \tau_{ui1} + \sum_{m=2}^{l-1} e^{\gamma_{um} + \lambda_{im}}$$

for $l = 2, 3, ..., L-1$. To prevent arbitrary shifting in utility, we fix $\tau_{ui1} = -L/2$. Second, we introduce the scale parameter $s_{ui} \geq s_0$ as in Eq. (8). And third, we evaluate a variety of utility distribution families in addition to the standard logistic in (Koren and Sill 2011).

Hidden features are initialised randomly from $\mathcal{N}(0, 0.01)$. Unless specified otherwise, all other parameters are initialised from zeros. When the contextual features in Eq. (6) are used, their weights are not updated until the improvement rate after each iteration for validation falls below 10^{-4}, i.e., the earlier iterations only improve the biases and the latent features. Algorithms are stopped if the improvement rate of likelihood of the validation data falls below 10^{-5}.

Method	*RMSE*	*MAE*	*LL*
Matrix Fac. (Gaussian)	0.932	0.737	-1.353
Multinom. (GEV, $\xi = 0$)	0.942	0.742	-1.252
Cumul. (Logistic)	**0.911**	**0.693**	-1.238
Cumul. (Gaussian)	0.928	0.715	-1.248
Cumul. (Gompertz)	0.925	0.714	-1.256
Cumul. (GEV, $\xi = 0$)	0.937	0.750	-1.278
*Sequent. (Logistic)**	*0.913*	*0.706*	***-1.223***
Sequent. (Logistic)	**0.911**	0.703	**-1.217**
Sequent.(Gaussian)	**0.911**	0.704	**-1.216**
Sequent. (Gompertz)	0.924	0.713	**-1.233**
Sequent. (GEV, $\xi = -0.2$)	**0.910**	0.705	**-1.217**
Sequent. (GEV, $\xi = 0.0$)	0.912	0.709	**-1.224**
Sequent. (GEV, $\xi = 0.2$)	0.918	0.718	**-1.237**
Sequent. (GEV, $\xi = 0.5$)	0.927	0.739	-1.258

Table 3: Results on MovieLens (using latent features with $K = 50$). The RMSE column is generated using Eq. 10 for prediction, and the MAE column using Eq. 9. (*) This is when the scale parameter is fixed to $s_{ui} = 1$. See also Table 1.

Model Verification

We verify the proposed sequential methods on the MovieLens and Dating datasets under different settings of mean/max-values utility distributions and the use of latent/contextual features. Table 1 reports the results for the MovieLens. The performance of utility distributions appears to be similar. Location parameters composed of bias features alone are quite effective – however this should not be too surprising because the models also rely other parameters such as thresholds and scales. The latent features and the context features are equally predictive but the latent features are faster to compute. And finally, the combination of features offers further improvement.

On the Dating dataset, only item-based contextual features are used (Table 2). It appears that for all three utility distributions, the contextual features are quite powerful: they already outperform the latent features when $J = 100$, and adding $J = 5000$ contextual features improves the MAE by approximately 9%, given that we have used only 5% of possible features (the number of items is $M \sim 10^5$).

Models Comparison

In this set of experiments, we compare the sequential approach with other approaches, namely the matrix factorisation, the multinomial and the cumulative. For efficiency, we employ only latent features in Eq. (6) and set the dimensionality to $K = 50$. The results on the three datasets are reported in Tables 3, 4 and 5, respectively. Several observations can be made from Tables 3, 4 and 5: First, the standard

[10] http://www.grouplens.org/node/12

[11] http://www.occamslab.com/petricek/data/

[12] http://netflixprize.com/

	Biases only			$K = 50$, $J = 0$			$K = 0$, $J = 200$			$K = 50$, $J = 200$		
Utility CDF	*RMSE*	*MAE*	*LL*	*RMSE*	*MAE*	*LL*	*RMSE*	*MAE*	*LL*	*RMSE*	*MAE*	*LL*
Logistic	0.949	0.757	-1.253	0.911	0.703	-1.217	0.918	0.705	-1.221	0.909	0.699	-1.216
Gaussian	0.942	0.746	-1.244	0.911	0.704	-1.216	0.914	0.707	-1.216	0.906	0.698	-1.213
GEV, $\xi = 0.0$	0.942	0.748	-1.248	0.912	0.709	-1.224	0.913	0.710	-1.222	0.906	0.702	-1.220

Table 1: Performance of the sequential model under different settings on the MovieLens dataset. The RMSE columns are generated using Eq. 10 for prediction, and the MAE column using Eq. 9. Recall that K is the number of latent features and $2J$ is number of context features (both item-based and user-based).

	$K = 50$, $J = 0$			$K = 0$, $J = 100$			$K = 50$, $J = 500$			$K = 50$, $J = 5000$		
Utility CDF	*RMSE*	*MAE*	*LL*	*RMSE*	*MAE*	*LL*	*RMSE*	*MAE*	*LL*	*RMSE*	*MAE*	*LL*
Logistic	0.874	0.555	-0.877	0.862	0.542	-0.863	0.843	0.522	-0.856	0.824	0.505	-0.841
Gaussian	0.864	0.543	-0.865	0.847	0.528	-0.849	0.836	0.517	-0.849	0.813	0.495	-0.832
GEV, $\xi = 0.0$	0.863	0.544	-0.872	0.846	0.528	-0.855	0.837	0.518	-0.856	0.812	0.497	-0.839

Table 2: Performance of the sequential model under different settings on the Dating dataset. Here we use the item-based contextual features only. Adding user-based features does not improve the performance.

	RMSE	*MAE*	*LL*
Matrix Fac. (Gaussian)	0.843	0.589	-1.274
Multinom. (GEV, $\xi = 0$)	0.892	0.532	-0.901
Cumul. (Logistic)	0.895	0.538	-0.907
Cumul. (Gaussian)	0.906	0.548	-0.918
Cumul. (Gompertz)	0.904	0.554	-0.928
Cumul. (GEV, $\xi = 0$)	0.910	0.568	-0.936
Sequent. (Logistic)	0.874	0.555	**-0.877**
Sequent.(Gaussian)	0.864	0.543	**-0.865**
Sequent. (Gompertz)	0.876	0.551	**-0.881**
Sequent. (GEV, $\xi = 0$)	0.863	0.544	**-0.872**
Sequent. (GEV, $\xi = 0$)*	**0.812**	**0.497**	**-0.839**

Table 4: Results on Dating (using latent features with $K = 50$, discrete methods use expected rating). (*) Adding $J = 5000$ item-based contextual features. See also Table 2.

matrix factorisation (MF) does not fit the data well, strongly indicating that rating is neither numerical nor Gaussian, and generally falls behind on the MAE metric. Its RMSE scores are reasonable but this is not surprising because the method optimises the RMSE score directly. Second, the multinomial method fits relatively well. This can be expected since it has more parameters than necessary, but at the cost of slower training time. In term of RMSE, the multinomial is not very competitive, possibly because it is purely discrete. In the case of the Dating dataset, it fares well on the MAE metric.

The cumulative methods perform relatively well on the RMSE and MAE metrics but do not exhibit very good likelihood fitting. The sequential methods perform best on likelihood metrics – this is a sign of appropriate data modelling. On other metrics, they are also very competitive against other methods.

Method	*RMSE*	*MAE*	*LL*
Matrix Fac. (Gaussian)	0.913	0.706	-1.344
Multinom. (GEV, $\xi = 0$)	0.934	0.699	-1.211
Cumul. (Logistic)	0.908	**0.658**	-1.211
Cumul. (Gaussian)	0.917	0.677	-1.224
Cumul. (Gompertz)	0.919	0.693	-1.236
Cumul. (GEV, $\xi = 0$)	0.920	0.679	-1.225
Sequent. (Gaussian)	**0.903**	0.665	**-1.184**
Sequent. (Logistic)	**0.905**	0.665	**-1.187**
Sequent. (Gompertz)	0.913	0.676	**-1.197**
Sequent. (GEV, $\xi = 0$)	**0.905**	0.669	**-1.194**

Table 5: Results on Netflix (using only latent features with $K = 50$).

Conclusion and Future Work

We have proposed an approach to address two largely ignored issues in collaborative filtering recommender systems: First, rating is inherently qualitative and ordinal in nature; and second, the process from which ratings are generated should be modelled explicitly. We started from the utility assumption that for each (user,item) pair there exist underlying, latent utility variables that govern such generation. We then investigated two related aspects: the ordinal assumption (e.g., normally distributed, multinomial, cumulative or sequential) and the distribution family of the underlying utilities (e.g., mean values or extreme values).

More explicitly, we advocated seeing ratings as being produced in a sequential decision process: the user starts from the lowest ordinal level by evaluating the perceived item utility against the level-specific threshold. If the utility exceeds the threshold then the next level will be considered, otherwise, the current level is selected. We investigated two major families of utility distributions, namely the mean values and the extreme values under the sequential decision framework. We have demonstrated that the approach is competitive against state-of-the-arts on several large-scale datasets.

There are a number of directions for future work. First, we have shown that the sequential decision assumption leads to better data fitting (on test data), as measured by the likelihood criterion. However, using likelihood as a performance measure is not a common practice in collaborative filtering, and the degree to which it is correlated to the end goals (e.g., usefulness for end users) is not clear. Second, an intrinsic drawback with this approach is that in that it does not offer any mechanism to *reverse* the decision. Modelwise, the probabilistic treatment proposed in this paper lends itself naturally to Bayesian analysis, which may lead to more robust prediction. And finally, an omitted aspect of collaborative filtering data is the high level of missing rate, and it is of high interest to incorporate the missingness mechanism into the model.

References

Albert, J., and Chib, S. 2001. Sequential ordinal modeling with applications to survival data. *Biometrics* 57(3):829–836.

Anderson, J. 1984. Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)* 1–30.

Gumbel, E. 1958. *Statistical of extremes*. Columbia University Press, New York.

Heckerman, D.; Chickering, D.; Meek, C.; Rounthwaite, R.; and Kadie, C. 2001. Dependency networks for inference, collaborative filtering, and data visualization. *The Journal of Machine Learning Research* 1:49–75.

Koren, Y., and Sill, J. 2011. OrdRec: an ordinal model for predicting personalized item rating distributions. In *Proceedings of the fifth ACM conference on Recommender systems*, 117–124. ACM.

Koren, Y. 2010. Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4(1):1.

Läärä, E., and Matthews, J. 1985. The equivalence of two models for ordinal data. *Biometrika* 72(1):206–207.

Mare, R. 1980. Social background and school continuation decisions. *Journal of the American Statistical Association* 295–305.

McCullagh, P. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* 109–142.

McFadden, D. 1973. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics* 105–142.

McFadden, D. 1980. Econometric models for probabilistic choice among products. *Journal of Business* 13–29.

Paquet, U.; Thomson, B.; and Winther, O. 2011. A hierarchical model for ordinal matrix factorization. *Statistics and Computing* 1–13.

Resnick, P.; Iacovou, N.; Suchak, M.; Bergstorm, P.; and Riedl, J. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM Conference on Computer Supported Cooperative Work*, 175–186. Chapel Hill, North Carolina: ACM.

Salakhutdinov, R., and Mnih, A. 2008. Probabilistic matrix factorization. *Advances in neural information processing systems* 20:1257–1264.

Salakhutdinov, R.; Mnih, A.; and Hinton, G. 2007. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 791–798.

Sarwar, B.; Karypis, G.; Konstan, J.; and Reidl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, 285–295. ACM Press New York, NY, USA.

Takács, G.; Pilászy, I.; Németh, B.; and Tikk, D. 2009. Scalable collaborative filtering approaches for large recommender systems. *The Journal of Machine Learning Research* 10:623–656.

Truyen, T.; Phung, D.; and Venkatesh, S. 2009. Ordinal Boltzmann machines for collaborative filtering. In *Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*.

Yu, S.; Yu, K.; Tresp, V.; and Kriegel, H. 2006. Collaborative ordinal regression. In *Proceedings of the 23rd international conference on Machine learning*, 1096. ACM.