Dynamic content based ranking

Seppo Virtanen
University of Cambridge

Mark Girolami
University of Cambridge and The Alan Turing Institute

Abstract

We introduce a novel state space model for a set of sequentially time-stamped partial rankings of items and textual descriptions for the items. Based on the data, the model infers text-based themes that are predictive of the rankings enabling forecasting tasks and performing trend analysis. We propose a scaled Gamma process based prior for capturing the underlying dynamics. Based on two challenging and contemporary real data collections, we show the model infers meaningful and useful textual themes as well as performs better than existing related dynamic models.

1 INTRODUCTION

Social media is becoming increasingly prevalent affecting and altering population behaviour. Users of social media interactively expose to and distribute content online, creating communities or networks of similar users. Within these communities users may amplify and disseminate specific content that may become *viral*, analogous to rapid spread of disease. It is of high relevance to analyse exposure uncovering complex dynamics of user interaction with social media, and detecting viral content.

For example, Facebook and Twitter are replacing traditional news providers, whereas Movielens and Netflix are changing consumption and distribution of movies. Users expose to items, such as, news articles or movies, and provide content via sharing news articles or rating movies.

We propose a general and natural framework to analyse exposure and uncover dynamics via analysing partial rankings of items based on their population popularity or relevance within consecutive time frames. Raw counts of *likes*, *shares* or *ratings* may naturally be used to compute the rankings for top-M items reflecting relative order of items. Viral content may, first, directly attain top rank or evolve more slowly to top in the consecutive rankings, second, dominate the top ranks for some time, and, finally, decline towards lower ranks eventually disappearing from the rankings.

In this work, we present a novel model for dynamic partial rankings of items that may be accompanied with text data providing rich information about the item content. Our main goal is to uncover textual content that explain the dynamic rankings and use them for exploration/summarisation in an easily interpretable manner, and for prediction. Based on two real data collections of movies and news, using user-provided tags/keywords and actual news articles, respectively, we show that the proposed model is useful for ranking prediction, trend-analysis and is able to capture complex dynamics of viral content.

In the following sections, we first introduce the model accompanied with a MCMC algorithm for posterior inference. Section 3 discusses related work. Then we present the experiments and results. The final section concludes the paper.

2 MODEL

2.1 Data

In this work, we consider partial or top-M time-stamped rankings, that is, partial permutations, of items $\boldsymbol{y}^{(t)} = \{y_1^{(t)}, \ldots, y_M^{(t)}\}$, where $t = 1, \ldots, T$ and $m = 1, \ldots, M$ denote time stamps and rank positions ($m = 1$ being the highest position), respectively. Each $y_m^{(t)}$ provides an identifier (for example, an integer) to a set of items $\mathcal{I}$.

2.2 Dynamic Plackett-Luce Model

We present a Plackett-Luce (PL)-based (Luce, 2005; Plackett, 1975) state space model for the rankings: the

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

probability of a ranking $\boldsymbol{y}^{(t)}$ is

$$\mathrm{PL}(\boldsymbol{y}^{(t)}) = \prod_{m=1}^{M} \frac{\hat{\boldsymbol{z}}_{y_m^{(t)}}^T \boldsymbol{w}^{(t)}}{\sum_{m'=m}^{M} \hat{\boldsymbol{z}}_{y_{m'}^{(t)}}^T \boldsymbol{w}^{(t)}} = \tag{1}$$

$$\frac{\hat{\boldsymbol{z}}_{y_1^{(t)}}^T \boldsymbol{w}^{(t)}}{\sum_{m'=1}^{M} \hat{\boldsymbol{z}}_{y_{m'}^{(t)}}^T \boldsymbol{w}^{(t)}} \frac{\hat{\boldsymbol{z}}_{y_2^{(t)}}^T \boldsymbol{w}^{(t)}}{\sum_{m'=2}^{M} \hat{\boldsymbol{z}}_{y_{m'}^{(t)}}^T \boldsymbol{w}^{(t)}} \cdots \frac{\hat{\boldsymbol{z}}_{y_{M-1}^{(t)}}^T \boldsymbol{w}^{(t)}}{\sum_{m'=M-1}^{M} \hat{\boldsymbol{z}}_{y_{m'}^{(t)}}^T \boldsymbol{w}^{(t)}},$$

where $\hat{\boldsymbol{z}}_d \in \Delta^K$, for $y_m^{(t)} = d$, denotes K item-specific compositional features, that are positive and sum to one over K, for the dth item and $\boldsymbol{w}^{(t)}$, for $t = 1, \ldots, T$, denote positively-valued feature-specific dynamic regression coefficients. The rankings depend on item-specific (positive) scores $\lambda_d^{(t)} = \hat{\boldsymbol{z}}_d^T \boldsymbol{w}^{(t)}$, that may be interpreted as a weighted average of feature-specific scores $w_k^{(t)}$, for $k = 1, \ldots, K$. Even though the total number of items is not known, the model is internally consistent, meaning that there is no need to take into account all possible items. Instead, only items that appear in the partial rankings are required (Hunter, 2004; Guiver and Snelson, 2009). The likelihood (1) may be intuitively decomposed, to illustrate a sampling process of items without replacement proportionally to the scores, such that the denominator for $m = i$ excludes items that occurred in $\boldsymbol{y}^{(t)}$ for ranks $m < i$. The higher the score, the better the rank.

Following Guiver and Snelson (2009), we adopt Gamma-based prior for the scores. We present a dynamic Gamma process for the regression coefficients (that is, feature-specific scores),

$$w_k^{(t)} \sim \mathrm{Gamma}\left(\tau, \frac{\tau}{w_k^{(t-1)}}\right), \tag{2}$$
$$w_k^{(1)} \sim \mathrm{Gamma}(\alpha_0, \beta_0),$$

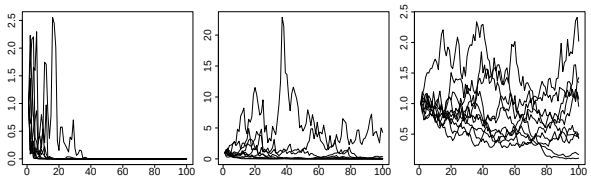
for $k = 1, \ldots, K$ and $t = 2, \ldots, T$, and assume

$$\tau \sim \mathrm{Gamma}\left(\alpha_0^{(\tau)}, \beta_0^{(\tau)}\right).$$

For $t \geq 2$, the process may be intuitively interpreted via the scaling property of Gamma-distributed random variables; for

$$\delta_k^{(t)} \sim \mathrm{Gamma}(\tau, \tau) \text{ and } w_k^{(t-1)} > 0,$$
$$w_k^{(t)} = \delta_k^{(t)} w_k^{(t-1)} \sim \mathrm{Gamma}\left(\tau, \frac{\tau}{w_k^{(t-1)}}\right).$$

Here, $\delta_k^{(t)}$ represents random multiplicative constants of the dynamical process. The prior mean of $w_k^{(t)}$ is given by $\mathbb{E}\left[w_k^{(t)}\right] = w_k^{(t-1)}$, irrespective of τ, whereas τ affects the prior variation, written as, $\mathrm{Var}\left[w_k^{(t)}\right] = \left(w_k^{(t)}\right)^2/\tau$, as well as specifies the shape of the distribution through controlling skewness $\frac{2}{\sqrt{\tau}}$ and kurtosis $\frac{6}{\tau}$. Figure 1 illustrates simulated coefficients. Depending on the value for τ the prior is flexible to account for smoothness, burstiness and self-excitation. For large values of τ, the coefficients are smooth and vary slowly in time. For smaller τ, the coefficients may experience burstiness and encourage self-excitation via increased variance.



(a) $\tau = 1$ (b) $\tau = 10$ (c) $\tau = 100$

Figure 1: Simulated regression coefficients for different values for τ for an initial value of $w^{(1)} = 1$. For $\tau \leq 1$, the curves fluctuate strongly and eventually fall close to zero and for $\tau \geq 100$ the curves become smooth with values increasingly closer to 1 (not shown).

Further, for $1 < t < T$, the prior conditional distribution of $w_k^{(t)}$ given $w_k^{(t')}$, for $t' \neq t$, proportional to

$$p(w_k^{(t)} | \{w_k^{(t')} : t' \neq t\}) \propto p(w_k^{(t+1)} | w_k^{(t)}) p(w_k^{(t)} | w_k^{(t-1)}), \tag{3}$$

corresponds to a generalised inverse Gaussian (GIG) distribution,

$$w_k^{(t)} \sim \mathrm{GIG}\left(0, \frac{2\tau}{w_k^{(t-1)}}, 2\tau w_k^{(t+1)}\right) \propto$$
$$\left(w_k^{(t)}\right)^{-1} \exp\left[-\tau\left(\frac{w_k^{(t)}}{w_k^{(t-1)}} + \frac{w_k^{(t+1)}}{w_k^{(t)}}\right)\right], \tag{4}$$

that illustrates how the prior takes naturally into account coefficient ratios between forward and backward time steps and also imposes sparsity due to the (logarithmic) penalty term, $-\log(w_k^{(t)})$.

2.3 Constructing Features

Assuming $K = |\mathcal{I}|$ and a diagonal feature matrix such that $\hat{z}_{d,d} = 1$ and zero for all other off-diagonal elements, we obtain item-specific dynamic scores, $\lambda_d^{(t)} = w_d^{(t)}$. However, for this simple choice the number of parameters increases rapidly for increasing $|\mathcal{I}|$. Further, the model is unable to generalise to new items and fails to properly leverage statistical strength between the coefficients. To overcome these limitations we assume text-based data exist for the items.

For $y_m^{(t)} = d$, let $\boldsymbol{x}_d = \{x_{d,1}, \ldots, x_{d,N_d}\}$ denote a set of bag-of-word text data, containing N_d word tokens $x_{d,n} \in \mathcal{V}$ over a word vocabulary $\mathcal{V}$. In the following, we refer to $\boldsymbol{x}_d$ as a (textual item-specific) document. The text data is available for each unique item, $d = 1, \ldots, |\mathcal{I}|$.

A straightforward approach would be to regress directly based on the empirical word-distributions, that is, $K = |\mathcal{V}|$ and

$$\hat{z}_{d,k} = \frac{1}{N_d} \sum_{n=1}^{N_d} \#[x_{d,n} = k],$$

for $k = 1, \ldots, |\mathcal{V}|$, inferring word-specific scores. However, text data is often high dimensional and sparse, complicating inference. Additionally, we know that word usage exhibits complex dependencies and, hence, treating words independently would not be optimal.

To leverage and capture word-usage dependencies, we assume the documents are generated from a mixed membership model, also referred to as, a topic model or Latent Dirichlet Allocation (LDA) (Blei et al., 2003). The model explains words in each document, under the bag-of-words assumption, as a latent mixture of topics (or themes) that are distributions over a word vocabulary. The topics are shared across documents and each document employs a mixture over a set of topics according to document-specific topic proportions. In more detail, the model assumes a set of K topics $\boldsymbol{\eta}_k$, for $k = 1, \ldots, K$, distributions over the elements in $\mathcal{V}$, and each document is assigned a distribution over the topics $\boldsymbol{\theta}_d$, for $d = 1, \ldots, |\mathcal{I}|$. The words appearing in $\boldsymbol{x}_d$ are explained by drawing assignment variables

$$z_{d,n} \sim \text{Categorical}(\boldsymbol{\theta}_d) \tag{5}$$

and assuming

$$x_{d,n} \sim \text{Categorical}(\boldsymbol{\eta}_{z_{d,n}}),$$

for $n = 1, \ldots, N_d$. The topics and topic proportions are generated from Dirichlet distributions,

$$\boldsymbol{\eta}_k \sim \text{Dir}(\gamma \mathbf{1}), \ \boldsymbol{\theta}_d \sim \text{Dir}(\alpha \mathbf{1}),$$

for $k = 1, \ldots, K$ and $d = 1, \ldots, |\mathcal{I}|$, correspondingly. We assume the features are given by empirical topic proportions,

$$\hat{z}_{d,k} = \frac{1}{N_d} \sum_{n=1}^{N_d} \#[z_{d,n} = k].$$

The rankings may change slowly or contain similar items at any time-stamp, affecting the (effective) size of $|\mathcal{I}|$. The small sample size problem complicates modelling of either rankings or items separately and, hence, it would be useful to share statistical strength via joint modelling. Thus, we present a joint model where the topic assignments depend both on the text data and rankings via (5) and (1), respectively.

2.4 MCMC for Posterior Inference

Following (Griffiths and Steyvers, 2004), we adopt (partially) collapsed Gibbs sampling for inference for the topic assignments analytically marginalising out topics and topic proportions. The probability that the word $x_{d,n} = i$ is assigned to the kth topic is given by

$$p(z_{d,n} = k) \propto \left(N_{d,k}^{-(dn)} + \alpha\right) \frac{G_{k,i}^{-(dn)} + \gamma}{\sum_{j=1}^{|\mathcal{V}|} \left(G_{k,j}^{-(dn)} + \gamma\right)} \times \cdots$$
$$\prod_{\{t: d \in \boldsymbol{y}^{(t)}\}} \text{PL}\left(\boldsymbol{y}^{(t)} | z_{d,n} = k, \boldsymbol{z}^{-(dn)}, \boldsymbol{w}^{(t)}\right),$$

where $\mathbf{N}$ collects document-topic counts, $\mathbf{G}$ collects topic-word counts. The set $\{t : d \in \boldsymbol{y}^{(t)}\}$ contains all the time stamps for which the rankings contain the dth document at any position, coupling the rankings across relevant time-stamps and documents appearing in the rankings. Here, the count matrices as well as topic assignments $\boldsymbol{z}$ exclude the contribution of the current variable $z_{d,n}$, denoted via $(\cdot)^{-(dn)}$.

For the coefficients $w_k^{(t)}$, for $t = 1, \ldots, T$ and $k = 1, \ldots, K$ we employ robust single-site slice sampling (Neal, 2003) in log domain to account for the positivity constraints. Update of $w_k^{(t)}$ for $2 \leq t \leq T-1$ given the remaining variables involves computing the prior contribution as in (3), and in particular (4), and the corresponding likelihood (1) for $\boldsymbol{y}^{(t)}$. For $t = 1$ and $t = T$ the prior computation simplifies: for $t = 1$, we compute

$$\text{Gamma}\left(w_k^{(2)} | \tau, \frac{\tau}{w_k^{(1)}}\right) \text{Gamma}(w_k^{(1)} | \alpha_0, \beta_0)$$

and, for $t = T$,

$$\text{Gamma}\left(w_k^{(T)} | \tau, \frac{\tau}{w_k^{(T-1)}}\right).$$

The algorithm contains two steps. First, sample a horizontal slice, whose y-coordinate is a generated value under the full likelihood (including the prior contribution and Jacobian of the log transformation) of the current state $\log\left(w_k^{(t)}\right)$. The end points of the slice are given by $\log\left(w_k^{(t)}\right) - us$ and $\log\left(w_k^{(t)}\right) + (1-u)s$, where $u \sim \text{Uniform}(0, 1)$ and s is a stick length parameter. Second, repeatedly sample a point along this slice until the full likelihood of the point is above the slice.

For updating τ using for example slice sampling we would need to be able to compute

$$p(\boldsymbol{w}|\tau) = \frac{1}{Z(\tau)} \prod_{t=2}^{T} \prod_{k=1}^{K} \text{Gamma}\left(w_k^{(t)}|\tau, \frac{\tau}{w_k^{(t-1)}}\right)$$

in addition to the prior term $p(\tau)$. Here, the normalisation constant $Z(\tau)$ resulting from integration over $\boldsymbol{w}$ depends on τ. Unfortunately, because we are unable to compute the constant analytically, we propose to use the exchange algorithm (Murray et al., 2006), that does not require computing $Z(\tau)$, for sampling τ. The algorithm proceeds by generating a proposal τ^* and auxiliary variables $\boldsymbol{w}^*$, of the same dimensions as $\boldsymbol{w}$, given τ^*. We note that we are naturally able to generate exact samples from the prior $p(\boldsymbol{w}|\tau)$, as required, following the dynamic generative process. We accept the proposal with probability

$$a = \frac{p(\boldsymbol{w}|\tau^*)}{p(\boldsymbol{w}|\tau)} \frac{p(\boldsymbol{w}^*|\tau)}{p(\boldsymbol{w}^*|\tau^*)} \frac{p(\tau^*)}{p(\tau)},$$

assuming a symmetric proposal distribution.

The hyperparameters of the model include $\{\alpha, \gamma, \alpha_0, \beta_0, \alpha_0^\tau, \beta_0^\tau\}$. For the topic model we use weakly informative priors; $\alpha = 0.1$ and $\gamma = 0.01$. For the coefficients we use $\alpha_0 = \beta_0 = 1$ and for τ we adopt $\alpha_0^{(\tau)} = \beta_0^{(\tau)} = 10^{-3}$.

3 RELATED WORK

The proposed Gamma prior (2) for the coefficients is a novel contribution. Cemgil and Dikmen (2007) present a related hierarchical Gamma chain construction introducing auxiliary variables. Recently, Jerfel et al. (2017) applies the prior by Cemgil and Dikmen (2007) for (conjugate) Poisson matrix factorisation. Acharya et al. (2015) propose a related construction parameterising the coefficients via shape parameters of the Gamma distribution; Schein et al. (2016) and Gong and Huang (2017) adopt this prior choice due to attractive computational properties for (conjugate) Poisson count models.

Paisley et al. (2011, 2012) and Dongwoo and Oh (2014) apply the scaling property of Gamma distribution for mixed membership modelling. Cemgil and Dikmen (2007) also employ the scaling property to construct hierarchical Gamma-Markov chains.

Supervised LDA (sLDA) assumes each document is paired with a response variable from the exponential family of distributions (Blei and McAuliffe, 2008). Zhu et al. (2009) provide a maximum margin based algorithm for the sLDA model. Chong et al. (2009) apply sLDA for responses generated from multinomial distributions. Virtanen and Girolami (2015) go beyond the exponential family assumption but still assume document-specific responses presenting a proper response model for ordinal ratings. Agarwal and Chen (2010) present multilabel extension, whereas Perotte et al. (2011) assume a hierarchy of classes. All of these extensions assume document-specific responses and are not suitable for more structured responses, as considered in this work.

Guiver and Snelson (2009) and Caron and Doucet (2012) present Bayesian inference approaches for PL models, assuming simple iid Gamma priors for the scores in a static setting with the main goal of rank aggregation or uncovering a consensus ranking. Azari et al. (2012) present partially Bayesian inference for a generalised class of PL models. Gormley and Murphy (2008), Caron et al. (2014) and recently Liu et al. (2019) assume a mixture model for the scores, adding modelling flexibility. The mixture model is suitable when the rankings stem from a population of accessors/judges. However, these static approaches are not useful for our dynamic setting.

Caron and Teh (2012) present a Markov-type dependence construction for dynamic scores of the PL model. However, the construction is unable to share information between items and to make predictions for unseen items, undermining predictive inference. Also, often only a small amount of information may be available for each item, depending on the number of rankings where items occur, complicating inference of the dependence structure. We overcome these issues by leveraging associated text data and assuming dynamics for the regression coefficients instead of individual items.

To make valid predictions for unseen items and share statistical strength between items, Cao et al. (2007) and Tkachenko and Lauw (2016) construct parametric mappings from observed item-specific features (that is, covariates) to the scores of the PL model for static (mixture) regression modelling. Our construction includes regression coefficients for every time stamp and is more flexible than a static variant with a single common regressor. Also we assume the features are latent and inferred based on the data instead of being observed and fixed.

Blei and Lafferty (2006) extend LDA (Blei et al., 2003) for a dynamic corpus, where documents are grouped by time stamps and the model assumes the documents within a group are similar by sharing parameters and close-by groups in time follow Markov dependence structure. Wang and McCallum (2006) present a LDA-type dynamic extension, where documents have continuously-valued time-stamps and each topic captures a distribution over timestamps, showing when topics are active. Wei et al. (2007) present a LDA-kind

dynamic model for a sequence of documents, such that the topic proportions between consecutive documents are dependent. The main difference to our model is that these models use the sequential order of documents (time-stamps) or document groupings and word counts to infer topic timelines. We use the ranking information, that reflects document relevance.

4 EXPERIMENTS AND RESULTS

4.1 Data Collections

We apply the model for two real data collections. The news data collection contains daily top-10 online news articles between 2019-01-31 and 2019-09-09 ranked according to social media activity, as measured by so called likes and shares. To detect viral content we use a sliding window approach; for each day we take previous seven days also into account for ranking. The data set statistics are: $T = 219$, $|\mathcal{I}| = 483$ and $|\mathcal{V}| = 2583$. The second data collection contains top-10 movies ranked according to popularity (as measured by raw rating count) during sequential non-overlapping time windows of one week. We use time-stamped ratings of movies by users provided by Movielens[1] between 2012-12-25 and 2018-09-19. For text data we use tags provided by the users. We obtain $T = 300$, $|\mathcal{I}| = 320$ and $|\mathcal{V}| = 2425$.

4.2 Experimental Settings

To assess performance of our prior (referred to as, P1), we carry out an extensive prior comparison against the state-of-the-art alternative related existing priors discussed in Section 3: the priors by Acharya et al. (2015) and Schein et al. (2016), are referred to as P2 and P3, respectively. We note that the prior proposed by Cemgil and Dikmen (2007) and Jerfel et al. (2017) is not suitable for our model because the coefficients may increase or decrease without bound causing numerical instability. For P2, we use

$$w_k^{(t)} \sim \text{Gamma}(w_k^{(t-1)}, c_t),$$

assuming $c_t \sim \text{Gamma}(e_0, f_0)$. For P3, we use

$$w_k^{(t)} \sim \text{Gamma}(\tau_0 w_k^{(t-1)}, \tau_0),$$

assuming an identity transition matrix. For completeness, the prior by Jerfel et al. (2017) corresponds to

$$w_k^{(t)} \sim \text{Gamma}(\iota, u_k^{(t)}),\ u_k^{(t)} \sim \text{Gamma}(\epsilon, w_k^{(t-1)}),$$

where $u_k^{(t)}$ are auxiliary Gamma variables. For inference, we employ, similarly for P1, the inference approach described in Section 2.4, using single-site slice sampling for $w_k^{(t)}$, noting that the proposed inference approaches are not useful here because our PL model is not conjugate to Gamma priors. The algorithms for the different priors differ only for computation of the prior contribution for $w_k^{(t)}$ as in Equation (3).

To assess the benefit of joint modelling, we compare our model (referred to as, M1) to a two-step approach (M2), where first standard topic model is used to obtain lower-dimensional features of the items that are then used as the observed empirical topic proportions (features) to infer the dynamic ranking model variables. We also compare against using empirical word distributions directly as features for the model (M3). For computational tractability for M3 we constrain the word distributions to top 10^3 most frequent terms, inferring word-specific timelines instead of topic-specific timelines. We adopt the same inference approach for M2/3 as for M1 using slice sampling for the coefficients and collapsed Gibbs sampling, omitting the contribution of the rankings.

Our main motivation is to evaluate the models quantitatively for ranking smoothing and forecasting tasks. For the forecasting task, we leave out the last time stamp and, for the smoothing task, we leave out 20% of the time-stamps for $1 < t < T - 1$, treating the held-out rankings as missing data at random. Although, we require $t = 1$ and $t = T - 1$ to be observed and consider both smoothing and forecasting tasks jointly. We measure model performance by computing the log PL model likelihood, (average) NDCG and Kendall's τ correlation for the held-out rankings using MC averages for the scores $\hat{\lambda}_d^{(t)} = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\boldsymbol{z}}_d^{(s)}\right)^T \boldsymbol{w}^{(t,s)}$ for $S = 50$, after thinning of 10 samples and burnin of 9×10^3 samples. We experimented with $10 \leq K \leq 50$ and show representative results for $K = 30$. We also note that the results generalise over multiple different held-out partitions. For computing NDCG we assign decreasing relevance scores as $M = 10$ to one according to rank.

We sample the regression coefficients for the held-out rankings (i.e., time stamps) based on the posterior and for unseen documents, that only appear in the held-out rankings, we infer the topic assignments based on the posterior of topics. When computing $\hat{\lambda}_d^{(t)}$, we normalise the coefficients $\boldsymbol{w}^{(t,s)}$ to sum to one for each time stamp, without loss of generality, because the evaluation measures are scale-invariant with respect to the coefficients. We note that traditional evaluation measures based on precision and recall are not suitable, because the total set of items is unknown.

[1] https://grouplens.org/datasets/movielens/

4.3 Quantitative Results

Tables 1 and 2 show that the proposed prior (P1) performs much better than the comparison priors P2/3 irrespective of the feature construction choice (M1/2/3) for both data sets. The difference is even more profound when regressing based on the empirical word distributions (M3). For all the measures, higher values indicate better performance. Our joint model (M1-P1) performs always better than the two-step model (M2-P1), showing that solely text-based topic model is unable to infer meaningful topics (features) for predicting rankings highlighting the importance of inferring the topic assignments jointly (M1). We see that both our joint model (M1-P1) and word-based large regression model (M3-P1) perform well; for the news data M1-P1 has better correlation and NDCG, whereas for the movies data, M1-P1 has better likelihood.

Table 1: Results for the news data set.

PL	P1	P2	P3
M1	-611	-625	-615.9
M2	-620	-636.5	-635.1
M3	-562.8	-651.7	-639.2
Kendall	**P1**	**P2**	**P3**
M1	0.6101	0.398	0.3586
M2	0.3616	0.2758	0.299
M3	0.535	0.216	0.3063
NDCG	**P1**	**P2**	**P3**
M1	0.9575	0.9125	0.9174
M2	0.911	0.8983	0.8916
M3	0.9452	0.8897	0.9123

Table 2: Results for the movies data set.

PL	P1	P2	P3
M1	-875.8	-901.6	-893.7
M2	-890.6	-898	-894.8
M3	-885.8	-900.2	-898.3
Kendall	**P1**	**P2**	**P3**
M1	0.2585	0.1919	0.183
M2	0.2022	0.1859	0.1615
M3	0.2763	0.1519	0.1667
NDCG	**P1**	**P2**	**P3**
M1	0.8979	0.8797	0.8827
M2	0.8851	0.877	0.8761
M3	0.9103	0.8677	0.873

4.4 Qualitative Results

Figure 2 shows the score evolution $\hat{\lambda}_d^{(t)}$, for $t = 1, \ldots, T$ for a set of 10 movies that have most top-1 positions (sorted respectively in decreasing order from top to bottom) for our model M1 and different priors P1/2/3. Black dots indicate the first time stamp for each movie when it enters the rankings. The scale for the scores is omitted because only relative values are relevant. Based on the figure, we see our prior enables inferring fluctuating and complex dynamics, whereas the comparison priors smooth excessively explaining poor quantitative performance.

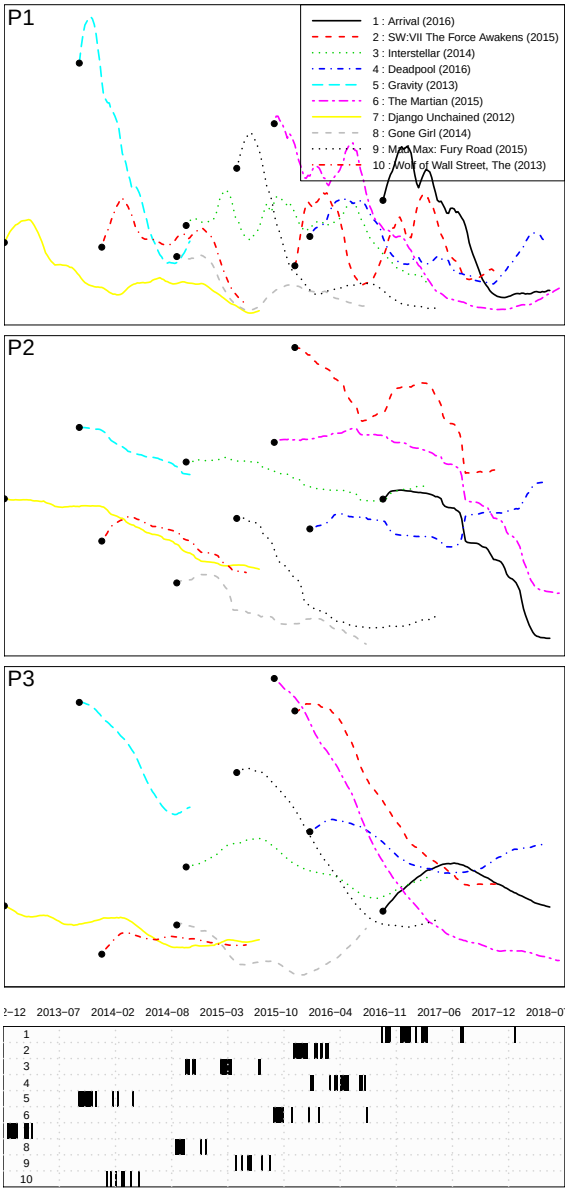


Figure 2: Set of item-specific scores $\hat{\lambda}_d^{(t)}$ across time for a set of 10 top-ranking movies (top) and time stamps when the movies have top-1 position (bottom) for our model M1 and different prior choices P1/2/3.

Figure 4 shows a similar figure for the news data. Our prior is again able to infer complex dynamics: for some articles the scores show a decaying trend illustrating that these articles obtain top position immediately and then drift towards lower rankings. For others, we see increasing or peaking trends. The comparison priors

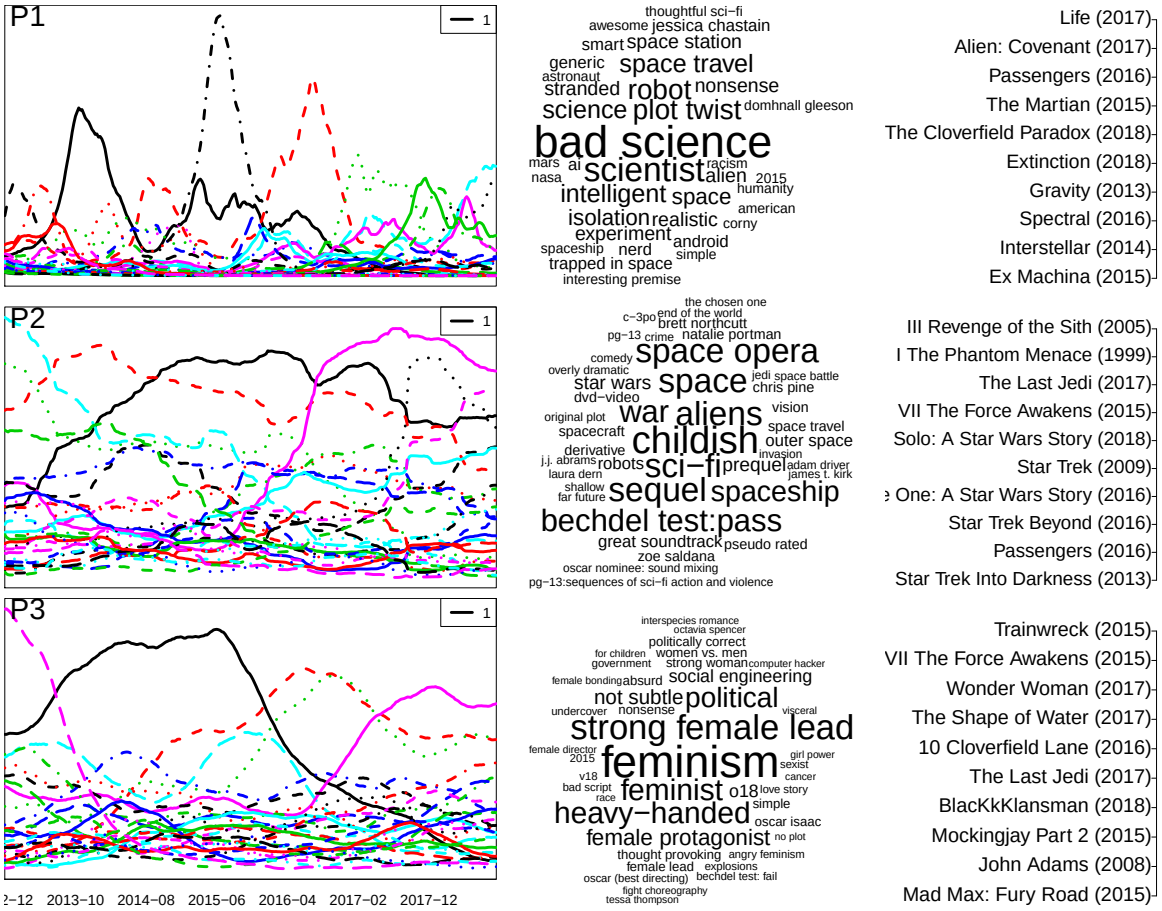


Figure 3: Inspection of topics inferred based on the movies data for our model M1 and different priors P1/2/3.

infer trivial dynamics; the scores are either constant or evolve very slowly.

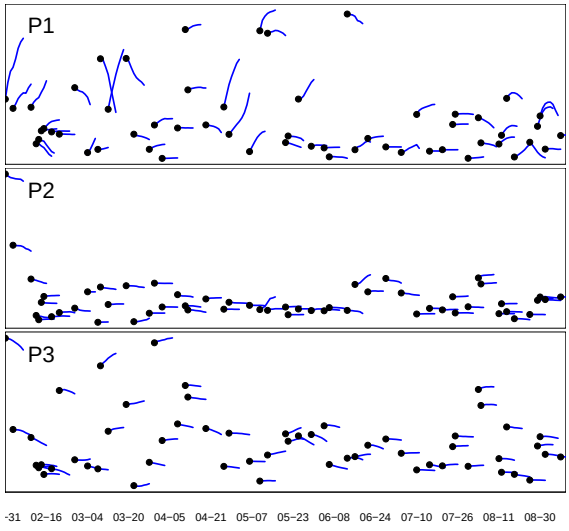


Figure 4: Set of item-specific scores $\hat{\lambda}_d^{(t)}$ across time for a set of top-ranking news articles for our model M1 and different priors P1/2/3.

Figure 3 illustrates topic timelines (regression coefficients) inferred based on the movies data for our model M1 and different priors. For each prior (row) we inspect one prominent topic (as indicated in the legend) showing the word cloud and a set of movie titles where the topic is most active. Based on the figure, we see that our prior is able to infer interpretable topic timelines as well as coherent topics and that the prior is flexible enough to capture complex dynamics. For the comparison priors, the timelines are too smooth and fail to pick emerging trends.

Figure 5 show a similar figure for the news data, inspecting two topics for each prior (row) as indicated by the figure legends. For our prior the timelines are bursty illustrating how news themes quickly reach top position and then decay. Further the topics capture evident themes of climate and Brexit. While the comparison priors are also able to infer meaningful topics as verified by inspecting the corresponding word clouds, they carry little relevance for the timelines and rankings. To summarise, the comparison priors smooth excessively and are less able to capture spikes and self-excitation, that are crucial for our applications. Hence the models based on the comparison priors are unable to capture dependencies between the rankings and text data, explaining poor quantitative performance.

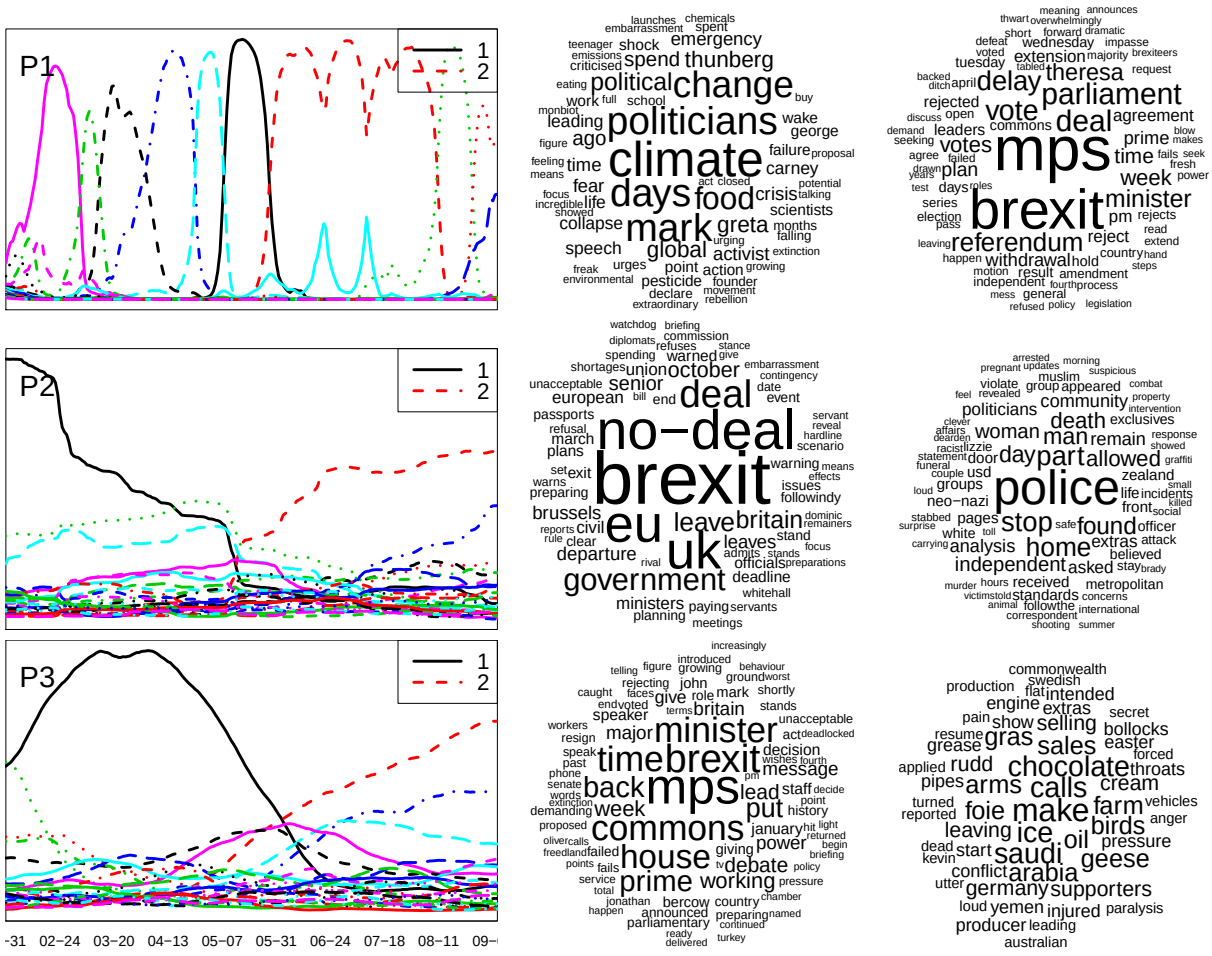


Figure 5: Inspection of topics inferred based on the news data for our model M1 and different priors P1/2/3.

5 DISCUSSION

We present a novel joint model of several consecutively time-stamped rankings of items that are accompanied with text data. The model provides an innovative approach for analysis of streaming text collections inferring topical timelines that reflect item relevance based on the rankings. Alternatively, the model presents a novel state space or factor model for the dynamic rankings. A key contribution is a novel dynamic Gamma process prior for the regression coefficients of the model, showcasing a prominent application of the exchange algorithm for posterior inference for more complicated priors. We demonstrate the model is able to infer meaningful topics and timelines based on two real data collections, corresponding to contemporary and challenging applications, and has better predictive performance than related comparison methods, showing the need for this kind of models.

The model is applicable for a wide range of different application areas for analysis of time-stamped lists of top-M rankings of items that may additionally contain rich textual data. In general, depending on the context, the ranks may represent quality, relevance, preference or popularity and the items may correspond to bestselling or most popular products, such as, books, movies or smartphone apps, most relevant or recent news articles or most visited Wikipedia pages, for instance. Here, the rankings often naturally vary in time in a complicated manner. We expect the model to be of great utility and interest, especially, for industrial applications, where the rankings may be based on top-grossing product information. Here, the model may be used to infer interpretable trends via textual content that are associated with high profit to guide developers/producers and investors. We also expect our model to spark interest in the (dynamic) PL ranking models in the machine learning community.

Acknowledgements

The authors were supported by the EPSRC grant EP/P020720/1, Inference COmputation and Numerics for Insights into Cities (ICONIC), `https://iconicmath.org/`.

References

Ayan Acharya, Joydeep Ghosh, and Mingyuan Zhou. Nonparametric Bayesian factor analysis for dynamic count matrices. In *AISTATS*, 2015.

Deepak Agarwal and Bee-Chung Chen. fLDA: Matrix factorization through latent Dirichlet allocation. In *Proceedings of the third ACM international conference on Web search and data mining*, 2010.

Hossein Azari, David Parks, and Lirong Xia. Random utility theory for social choice. In *NIPS*, 2012.

David M Blei and John D Lafferty. Dynamic topic models. In *ICML*, 2006.

David M Blei and Jon D McAuliffe. Supervised topic models. In *NIPS*, 2008.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*, 2007.

Francois Caron and Arnaud Doucet. Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

François Caron and Yee W Teh. Bayesian nonparametric models for ranked data. In *NIPS*, 2012.

François Caron, Yee W Teh, and Thomas B Murphy. Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics*, 8(2): 1145–1181, 2014.

A Taylan Cemgil and Onur Dikmen. Conjugate Gamma Markov random fields for modelling nonstationary sources. In *International Conference on Independent Component Analysis and Signal Separation*, 2007.

Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *CVPR*, 2009.

Kim Dongwoo and Alice Oh. Hierarchical Dirichlet scaling process. In *ICML*, 2014.

Chengyue Gong and Win-Bin Huang. Deep dynamic Poisson factorization model. In *NIPS*, 2017.

Isobel C Gormley and Thomas B Murphy. Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027, 2008.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235, 2004.

John Guiver and Edward Snelson. Bayesian inference for Plackett-Luce ranking models. In *ICML*, 2009.

David R Hunter. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32 (1):384–406, 2004.

Ghassen Jerfel, Mehmet Basbug, and Barbara Engelhardt. Dynamic collaborative filtering with compound Poisson factorization. In *AISTATS*, 2017.

Ao Liu, Zhibing Zhao, Chao Liao, Pinyan Lu, and Lirong Xia. Learning Plackett-Luce mixtures from partial preferences. In *AAAI*, 2019.

R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

Iain Murray, Zoubin Ghahramani, and David JC MacKay. MCMC for doubly-intractable distributions. In *UAI*, 2006.

Radford M Neal. Slice sampling. *Annals of Statistics*, pages 705–741, 2003.

John Paisley, Chong Wang, and David Blei. The discrete infinite logistic normal distribution for mixed-membership modeling. In *AISTATS*, 2011.

John Paisley, Chong Wang, and David M Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(4):997–1034, 2012.

Adler J Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. Hierarchically supervised latent dirichlet allocation. In *NIPS*, 2011.

Robin L Plackett. The analysis of permutations. *Applied Statistics*, 1975.

Aaron Schein, Hanna Wallach, and Mingyuan Zhou. Poisson-Gamma dynamical systems. In *NIPS*, 2016.

Maksim Tkachenko and Hady W Lauw. Plackett-Luce regression mixture model for heterogeneous rankings. In *CIKM*, 2016.

Seppo Virtanen and Mark Girolami. Ordinal mixed membership models. In *ICML*, 2015.

Xuerui Wang and Andrew McCallum. Topics over time: A non-Markov continuous-time model of topical trends. In *SIGKDD*, 2006.

Xing Wei, Jimeng Sun, and Xuerui Wang. Dynamic mixture models for multiple time series. In *IJCAI*, 2007.

Jun Zhu, Amr Ahmed, and Eric P Xing. MedLDA: Maximum margin supervised topic models for regression and classification. In *ICML*, 2009.