Ordinal Mixed Membership Models

Seppo Virtanen S.VIRTANEN@WARWICK.AC.UK
Mark Girolami M.GIROLAMI@WARWICK.AC.UK
Department of Statistics, University of Warwick, CV4 7AL Coventry UK

Abstract

We present a novel class of mixed membership models for joint distributions of groups of observations that co-occur with ordinal response variables for each group for learning statistical associations between the ordinal response variables and the observation groups. The class of proposed models addresses a requirement for predictive and diagnostic methods in a wide range of practical contemporary applications. In this work, by way of illustration, we apply the models to a collection of consumer-generated reviews of mobile software applications, where each review contains unstructured text data accompanied with an ordinal rating, and demonstrate that the models infer useful and meaningful recurring patterns of consumer feedback. We also compare the developed models to relevant existing works, which rely on improper statistical assumptions for ordinal variables, showing significant improvements both in predictive ability and knowledge extraction.

1. Introduction

There exist large repositories of user-generated assessment, preference or review data consisting of free-form text data associated with ordinal variables for quality or preference. Examples include product reviews, user feedback, recommendation systems, expert assessments, clinical records, survey questionnaires, economic or health status reports, to name a few. The ubiquitous need to statistically model the underlying processes and analyse such data collections presents significant methodological research challenges necessitating the development of proper statistical models and inference approaches.

In this work, our interest focuses on, but is not limited to, analysing reviews of mobile software applications provided by consumers. Such analysis is useful for both software developers and consumers, inferring and understanding themes or properties of mobile applications that consumers comment about. These themes may involve consumers' preferences and experiences on properties they (dis)appreciate or direct feature requests or problems directed to the software developers.

Our work belongs in the field of mixed membership modelling, which is a powerful and important statistical modelling methodology. Observations are grouped and each group is modelled with a mixture model; mixture components are common to all groups, whereas mixture proportions are group-specific. The components are deemed to capture recurring patterns of observations and each group to exhibit a subset of components. The class of models has been shown to be able to extract interpretable meaningful themes, also referred to as topics, based on, for example, text data (Blei et al., 2003). These models, however, are not able to capture statistical associations between the groups and co-occurring quantitative information, that is, response variables, related to each group.

Previous work on joint models utilising both the textual data and response variables (Blei & McAuliffe, 2007; Dai & Storkey, 2015; Lacoste-Julien et al., 2009; Nguyen et al., 2013; Ramage et al., 2009; Wang et al., 2009) has demonstrated the utility of joint modelling by inferring topics that are predictive of the response leading to increased interpretability. However, these models lack proper statistical formulations suitable for ordinal response variables and it is not at all straightforward to correct this shortcoming. In this work, we remove this hindrance by presenting a novel class of joint mixed membership models.

The proposed class of models builds on our new statistical generative response model for ordinal variables. In more detail, we introduce a certain stick-breaking formulation to parameterise underlying data-generating probabilities over the ordinal variables. The response model contains group-specific latent scores as well as mean variables that transform the scores into ordinal variables using the developed construction. We compare the response model with exist-

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

ing alternatives for ordinal variables (Albert & Chib, 1993; Chu & Ghahramani, 2005) and show that our formulation provides favourable statistical properties.

We present two different novel model formulations that couple the developed response model with mixed membership models. Specifically, the formulations hierarchically couple the latent scores of the response model with the mixing components of a mixed membership model either via the mixture proportions or observation assignments capturing associations between the components and responses. The first construction infers a correlation structure between (as well as, within) the mixture proportions and latent scores based on the observed data, not enforcing *a priori* any correlation structure or specifying which of the components are associated with the responses. We derive a scalable variational Bayesian inference algorithm to approximate the model posterior distribution. The model is motivated by *unsupervised* correlated topic models by Blei & Lafferty (2006) and Paisley et al. (2012). The second construction assumes the latent scores of the response model are given by a weighted linear combination of the mean assignments over each group, such that the component-specific combination weights *a posteriori* provide a means to inspect components that have predictive value. We present a Markov Chain Monte Carlo (MCMC) sampling scheme for posterior inference. The model is related to supervised LDA (SLDA; Blei & McAuliffe, 2007); our model can be seen as an extension of SLDA to ordinal responses.

We demonstrate the developed models on a collection of reviews of mobile software applications. We compare the models to the relevant previous work and show that the proper ordinal response model is valuable for learning statistical associations between the responses and text data providing significant improvements in terms of both predictive ability and knowledge extraction by inferring interpretable and useful themes of consumer feedback.

The paper is structured as follows. Section 2 presents the methodological contributions of this work: Section 2.1 presents our proposed generative model for ordinal variables, whereas the next two Sections 2.2 and 2.3 present model formulations and inference approaches for joint mixed membership modelling of groups of observations and group-specific ordinal response variables. Related work is reviewed in Section 3. Section 4 describes the experiments and contains the results. Section 5 concludes the paper.

2. Joint Mixed Membership Models

The mth group of observations $\mathbf{w}^{(m)}$ is paired with an ordinal response variable $y^{(m)}$. The response variables, also referred to as ratings, take values in $R \in \mathbb{Z}_+ > 2$ ordered categories ranging between poor (1) and excellent (R). We note that for a simple case, when $R = 2$, $y^{(m)}$ is binary and may be modelled by a Bernoulli distribution. The $\mathbf{w}^{(m)}$ contains an unordered sequence of $D^{(m)}$ words $w_d^{(m)}$ over a V-dimensional vocabulary, $\mathbf{w}^{(m)} = \{w_1^{(m)}, w_2^{(m)}, \ldots, w_{D^{(m)}}^{(m)}\}$.

2.1. Ordinal Response Variables

We assume $y^{(m)}$ is drawn from a categorical distribution over R categories. The probability that $y^{(m)}$ takes an integer value $r \in \{1, \ldots, R\}$ is denoted by $p(y^{(m)} = r)$. Since the categories are ordered, we propose a stick-breaking parameterisation for the probabilities; a unit length stick is split into R smaller sticks that sum to one. We refer to these smaller sticks as stick weights $v_r^{(m)}$ for the mth group and rth category. We parameterise the $v_r^{(m)}$ using a function $\sigma(\cdot)$ mapping its argument to a value between zero and one and introducing continuous-valued latent variables or scores $t^{(m)}$ for each group as well as mean parameters μ_r for each category. The generative model for the $y^{(m)}$ is

$$p(y^{(m)} = r) = v_r^{(m)} \prod_{r'=1}^{r-1} (1 - v_{r'}^{(m)}), \tag{1}$$
$$v_r^{(m)} = \sigma(t^{(m)} - \mu_r).$$

Each $v_r^{(m)}$ represents a binary decision boundary, specified by the mean variables, for the $t^{(m)}$. The mean variables are ordered, that is, $\mu_1 < \mu_2 < \cdots < \mu_R$, representing boundaries between the ordered categories. For computational simplicity, we use $\sigma(x) = (1 + \exp(-x))^{-1}$ corresponding to a logit (or sigmoid) function, for which $1 - \sigma(x) = \sigma(-x)$. Alternative choices include probit, log log or Cauchy functions, to name a few. The stick-breaking formulation guarantees that the probabilities $p(y^{(m)} = r)$, for $r = 1, \ldots, R$, are positive and sum to one for any value of the $t^{(m)}$. More importantly, the formulation leads to a simple posterior inference algorithm; the ordering of the mean variables is implicitly inferred based on the observed data without enforcing explicit constraints. For identifiability, we set, without loss generality, $v_R^{(m)} = 1$. Figure 1 demonstrates the construction of probabilities based on the $t^{(m)}$ for simulated mean variables $\boldsymbol{\mu}$.

Based on a collection of observed responses $y^{(m)}$, where $m = 1, \ldots, M$, the model log likelihood is

$$\mathcal{L} = \sum_m \ln(v_{y^{(m)}}^{(m)}) + \sum_{r'=1}^{y^{(m)}-1} \ln(1 - v_{r'}^{(m)}). \tag{2}$$

Point estimates for the latent scores as well as mean variables may be inferred by maximising the log likelihood using unconstrained gradient-based optimisation techniques.

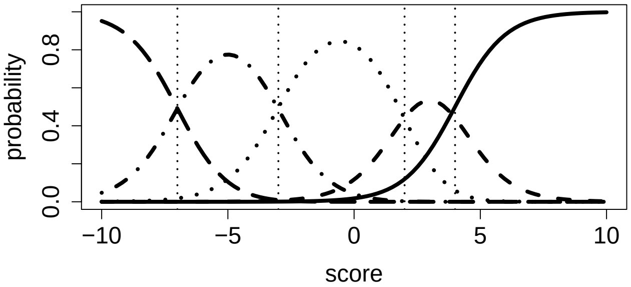

Figure 1. Visual demonstration of category probabilities. Here, x-axis denotes a range of values for the latent variable or score $t^{(m)}$, whereas the vertical lines denote the category cut-off points, referred to as mean variables $\boldsymbol{\mu}$.

In the following sections, we present two approaches for parameterising the latent scores constructing statistical associations between the responses and groups. Main statistical interest focuses on the parameterisation, whereas the mean variables are relevant mainly for computing predictions. For this reason, in the following, we assign a uniform prior for the mean variables.

2.2. Joint Correlated Topic Model

In this section, we present a novel joint model (referred to as, JTM) for the $y^{(m)}$ and $\mathbf{w}^{(m)}$, where $m = 1, \ldots, M$. At the core of the model are group-specific latent variables $\mathbf{u}^{(m)}$ that are common for $y^{(m)}$ and $\mathbf{w}^{(m)}$ capturing statistical associations between them.

For the responses we introduce a linear mapping or projection $\boldsymbol{\xi}$ and construct the data-generating latent score (Equation 1) as $t^{(m)} = \boldsymbol{\xi}^T \mathbf{u}^{(m)}$, computing a cross product between the $\mathbf{u}^{(m)}$ and the mapping $\boldsymbol{\xi}$.

The generative process for the $\mathbf{w}^{(m)}$ (groups of observations), for $m = 1, \ldots, M$, is given by

$$\begin{aligned} w_d^{(m)} &\sim \text{Categorical}(\boldsymbol{\eta}_{c_d^{(m)}}), \qquad (3) \\ c_d^{(m)} &\sim \text{Categorical}(\boldsymbol{\theta}^{(m)}), \end{aligned}$$

where $\boldsymbol{\eta}_k$, for $k = 1, \ldots, K$, denotes mixture components (topics), $c_d^{(m)}$, for $d = 1, \ldots, D_m$, denotes observation assignments and $\boldsymbol{\theta}^{(m)}$ mixture (topic) proportions over the K topics.

We connect the $\boldsymbol{\theta}^{(m)}$ to the latent variables $\mathbf{u}^{(m)}$ by introducing topic-specific mappings $\mathbf{v}_k$ and gamma-distributed variables $z_k^{(m)}$ (parameterised suitably) such that *a priori*

$$\mathbb{E}[\theta_k^{(m)}] \propto \tilde{\beta}_k \exp(\mathbf{v}_k^T \mathbf{u}^{(m)}), \qquad (4)$$

where $\tilde{\beta}_k$, for $k = 1, \ldots, K$, are positive concentration parameters. The latent mappings capture statistical associations between any two topics indexed by k and k'. If the $\mathbf{v}_k$ and $\mathbf{v}'_k$ are similar, the topics $\boldsymbol{\eta}_k$ and $\boldsymbol{\eta}_{k'}$, respectively, tend to co-occur, assuming that $\tilde{\beta}_k$ and $\tilde{\beta}_{k'}$ are sufficiently large. We use (normalised) gamma-distributed variables to construct the topic proportions thus parameterising a mapping from the continuous latent variables to the discrete topic proportions. For simplified posterior inference we define

$$\tilde{\beta}_k = \beta \exp(m_k). \qquad (5)$$

The process is

$$\theta_k^{(m)} \propto z_k^{(m)} \sim \text{Gamma}\big(\beta, \exp(-\mathbf{v}_k^T \mathbf{u}^{(m)} - m_k)\big),$$

where the β denotes the shape parameter and the $\exp(-\mathbf{v}_k^T \mathbf{u}^{(m)} - m_k)$ denotes the rate parameter of the gamma distribution, respectively. We see that

$$\mathbb{E}[\theta_k^{(m)}] \propto \beta \exp(\mathbf{v}_k^T \mathbf{u}^{(m)} + m_k),$$

as desired (4), using equation (5)[1]. Figure 2 illustrates a graphical plate diagram of the model.

We complete the model description specifying distributions for the model hyper-parameters, the root nodes in Figure 2. We assign

$$\begin{aligned} \beta &\sim \text{Gamma}(\alpha_0, \beta_0), \\ \mathbf{u}^{(m)} &\sim \text{Normal}(\mathbf{0}, \mathbf{I}) \\ \boldsymbol{\xi}, \mathbf{v}_k &\sim \text{Normal}(\mathbf{0}, l^{-1}\mathbf{I}), \end{aligned}$$

where l denotes a precision (inverse variance) parameter of a (zero-mean) Gaussian distribution,

$$\boldsymbol{\eta}_k \sim \text{Dirichlet}(\gamma \mathbf{1}),$$

where γ is a concentration parameter of a Dirichlet distribution, and a non-informative prior for the m_k.

2.2.1. Interpretation

After specifying the model, we highlight the role of the latent variables and the corresponding mappings for the responses and topics, $\boldsymbol{\xi}$ and $\mathbf{v}_k$, where $k = 1, \ldots, K$, respectively. We may compute a measure for similarity between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ defining a function

$$l(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\sqrt{(\mathbf{x}_i^T \mathbf{x}_i)(\mathbf{x}_j^T \mathbf{x}_j)}}$$

that outputs a value between 1 and -1 indicating similarity or dissimilarity between the vectors. We may compute $l(\boldsymbol{\xi}, \mathbf{v}_k)$, where $k = 1, \ldots, K$, and use the (dis)similarity

[1] We note that for a gamma-distributed random variable $x \sim \text{Gamma}(a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$, where $\Gamma(\cdot)$ denotes the gamma function, $\mathbb{E}[x] = a/b$.

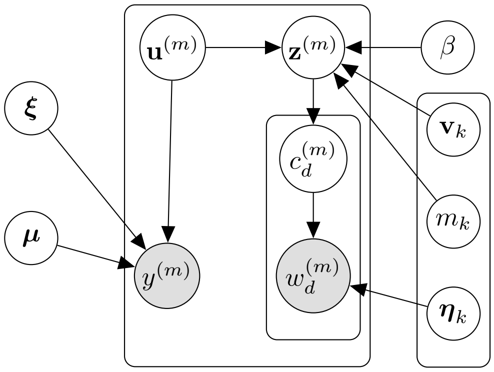

Figure 2. Graphical plate diagram of the joint correlated topic model. Unshaded nodes correspond to unobserved variables, whereas shaded nodes correspond to observed variables. Hyperparameters for the root nodes, whose values need to be fixed prior to posterior inference, are omitted from the visualisation. Plates indicate replication over topics, groups and words. The hidden variables may be divided into local group-specific variables and global variables common to all groups. That is, the unnormalised topic proportions $\mathbf{z}^{(m)}$, topic indicators $c_j^{(m)}$ and latent variables $\mathbf{u}^{(m)}$ are defined for each group, whereas the set of topics $\boldsymbol{\eta}_k$, mappings from latent variables to data domains, $\boldsymbol{\xi}$ and $\mathbf{v}_k$, are common to all groups.

scores to infer whether the topics that are positively or negatively associated with *excellent* or *poor* ratings.

Next, we present a theoretical justification for the similarity measure. Marginalisation of the latent variables $\mathbf{u}$ is analytically tractable leading to a joint Gaussian distribution for the $t^{(m)}$ and auxiliary variables $h_k^{(m)}$ (replacing the $\mathbf{v}_k^T\mathbf{u}^{(m)}$). The covariance matrix of the Gaussian distribution is

$$\boldsymbol{\Sigma} = \mathbf{W}\mathbf{W}^T + \mathbf{I},$$

where $\mathbf{W}^T = \begin{pmatrix}\boldsymbol{\xi} & \mathbf{v}_1 & \dots & \mathbf{v}_K\end{pmatrix}$.We see that the similarity values defined above correspond to correlations between the response and topical mappings, respectively. We also note that the distribution is able to capture correlations between any two topics. Hence, we refer to this model as joint correlated topic model.

2.2.2. Regularisation

During posterior inference the model infers statistical associations between the groups and responses. The inferred topics summarise recurring word co-occurrences over the corpus into interpretable themes some of which may have significant associations with the ratings. However, for finite sample sizes the correlation structure may be weak. Accordingly we introduce a user-defined parameter $\lambda > 0$, that balances for the limited sample sizes. Even though, we expect, when the sample size M increases for fixed vocabulary size V, the role of λ diminishes, since there are more data to estimate the underlying correlation structure. The joint likelihood of the model is

$$p(\mathcal{D},\Theta) = \prod_{m=1}^{M}\prod_{d=1}^{D^{(m)}}\prod_{k=1}^{K} p(w_d^{(m)})p(c_d^{(m)})p(z_k^{(m)})$$
$$\left(\exp(\mathcal{L})p(\boldsymbol{\xi})\right)^{\lambda} p(\mathbf{u}^{(m)})p(\mathbf{v}_k)p(\beta),$$

where $\mathcal{D} = \{\mathbf{w}^{(m)}, y^{(m)}\}_{m=1}^{M}$, Θ denotes unknown quantities of the model and $\mathcal{L}$ is given in Equation (2). For $\lambda < 1$ the model focuses more on explaining the text.

2.2.3. Variational Bayesian Inference

We present a variational Bayesian (VB) (Wainwright & Jordan, 2008) posterior inference algorithm for the model that scales well for large data collections and can readily be extended to stochastic online learning (Hoffman et al., 2013). We approximately marginalise over the topic assignments and proportions using non-trivial factorised distributions, whereas we use point distributions (estimates) for several variables to simplify computations, in essence, adopting an empirical Bayes approach for these variables. The corresponding inference algorithm is able to prune out irrelevant topics from the model based on the observed data. Full variational inference would be possible using techniques presented by Böhning (1992); Jaakkola & Jordan (1997) and Wang & Blei (2013), for example, lower bounding analytically intractable log sigmoid function appearing in the log likelihood function (2). Alternatively, MCMC sampling strategies may provide appealing approaches for posterior inference. However, it is far from trivial to design suitable proposal distributions for the latent variables.

We introduce a factorised posterior approximation

$$q(\Theta) = \prod_{m=1}^{M}\prod_{d=1}^{D^{(m)}}\prod_{k=1}^{K} q(c_d^{(m)})q(z_k^{(m)}),$$

omitting the point distributions for clarity, and minimise the KL-divergence between the factorisation $q(\Theta)$ and the posterior $p(\Theta|\mathcal{D})$. Alternatively, we maximise a lower bound of the model evidence with respect to the parameters of the $q(\Theta)$,

$$\ln p(\mathcal{D}) \geq \mathcal{L}^{VB} = \mathbb{E}[\ln p(\mathcal{D},\Theta)] - \mathbb{E}[p(\Theta)\ln p(\Theta)],$$

where expectations are taken with respect to the $q(\Theta)$.

We choose the following distributions for the topic assignments and unnormalised topic proportions

$$q(c_d^{(m)}) = \text{Categorical}(c_d^{(m)}|\boldsymbol{\phi}_d^{(m)}),$$
$$q(z_k^{(m)}) = \text{Gamma}(z_k^{(m)}|a_k^{(m)}, b_k^{(m)}),$$

whose parameters are

$$\phi_{w,k}^{(m)} \propto \eta_{k,w} \exp(\mathbb{E}[\ln z_k^{(m)}]),$$
$$a_k^{(m)} = \beta + \sum_{j=1}^{D^{(m)}} \phi_{j,k}^{(m)},$$
$$b_k^{(m)} = \exp(-\mathbf{v}_k^T \mathbf{u}^{(m)} - m_k) + \frac{D^{(m)}}{\sum_{k=1}^{K} \mathbb{E}[z_k^{(m)}]}.$$

In the derivations, we applied Jensen's inequality lower bounding analytically intractable $\mathbb{E}[\ln \sum_{k=1}^{K} z_k^{(m)}]$ needed for normalisation of $z_k^{(m)}$, for $k = 1, \ldots, K$, by introducing additional auxiliary parameters for each group. The expectations appearing above with respect to the variational factorisation are

$$\mathbb{E}[\ln z_k^{(m)}] = \psi(a_k^{(m)}) - \ln b_k^{(m)},$$
$$\mathbb{E}[z_k^{(m)}] = \frac{a_k^{(m)}}{b_k^{(m)}},$$

where $\psi(\cdot)$ denotes the digamma function.

The lower bound of the model evidence, a cost function to maximise, with respect to the $\mathbf{u}^{(m)}$ is

$$\mathcal{L}_{\mathbf{u}}^{VB} = \lambda\mathcal{L} + \sum_{m,k} \mathbb{E}[\ln p(z_k^{(m)}|\mathbf{u}^{(m)}, \mathbf{v}_k, \mathbf{m}, \beta)] + \ln p(\mathbf{u}^{(m)}),$$

whereas for $\mathbf{v}$, $\mathbf{m}$ and β the cost function is

$$\mathcal{L}_{\mathbf{v},\mathbf{m},\beta}^{VB} = \sum_{m,k} \mathbb{E}[\ln p(z_k^{(m)}|\mathbf{u}^{(m)}, \mathbf{v}_k, \mathbf{m}, \beta)] + \ln p(\mathbf{v}_k, \mathbf{m}, \beta).$$

To infer the mapping $\boldsymbol{\xi}$ we maximise $\mathcal{L}_{\boldsymbol{\xi}}^{VB} = \mathcal{L} + \ln p(\boldsymbol{\xi})$. Unconstrained gradient-based optimisation techniques may be used to infer point estimates for these unobserved quantities (optimising β in log-domain). Finally, the topics are updated as

$$\eta_{k,w} \propto \sum_{d,m} \phi_{d,k}^{(m)} + \gamma - 1.$$

2.3. Ordinal Supervised Topic Model

In this section, we propose a novel topic model for the ordinal responses and groups of observations. The model assumes a generative process for the words similar to that in Equation 3 introducing topic assignments $c_d^{(m)}$ for words $w_d^{(m)}$, where $d = 1, \ldots, D^{(m)}$, and topic proportions $\boldsymbol{\theta}^{(m)}$ for the mth group. Here, the generative model for the ratings depends on the $c_d^{(m)}$, where $d = 1, \ldots, D^{(m)}$. In more detail, we define

$$\widetilde{c}_k^{(m)} = \frac{1}{D^{(m)}} \sum_{j=1}^{D^{(m)}} \mathbb{I}[c_j^{(m)} = k],$$

where $\mathbb{I}[\cdot]$ denotes the indicator function equaling 1 if the argument is true and zero otherwise, representing an empirical topic distribution for the mth group. We use the quantity to construct a linear mapping to the ratings. The model (see Figure 3 for an illustration of a graphical plate diagram) is

$$t^{(m)} = \boldsymbol{\xi}^T \widetilde{\mathbf{c}}^{(m)}$$
$$w_d^{(m)} \sim \text{Categorical}(\boldsymbol{\eta}_{c_d^{(m)}}),$$
$$c_d^{(m)} \sim \text{Categorical}(\boldsymbol{\theta}^{(m)}),$$
$$\boldsymbol{\theta}^{(m)} \sim \text{Dirichlet}(\boldsymbol{\alpha}),$$
$$\boldsymbol{\eta}_k \sim \text{Dirichlet}(\gamma\mathbf{1}),$$
$$\xi_k \sim \text{Normal}(0, \zeta).$$

Based on the observed data $\mathcal{D}$ the model infers a set of topics that explain not only word co-occurrences but also the responses.

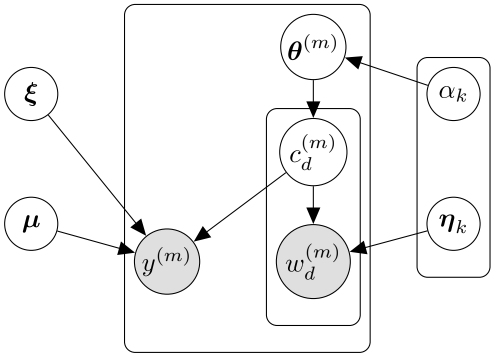

Figure 3. Graphical plate diagram for the ordinal supervised topic model. The topic proportions $\boldsymbol{\theta}^{(m)}$ are group-specific and generated from an asymmetric Dirichlet distribution. The ordinal generative model for the ratings depends on topic assignments $c_d^{(m)}$, that specify the topical content (textual themes via topics $\boldsymbol{\eta}_k$) of the mth group.

2.3.1. MCMC Sampling Scheme

We present a MCMC sampling scheme for the model. We consecutively sample the topic assignments given current value of $\boldsymbol{\xi}$ using collapsed Gibbs sampling, building on the work by Griffiths & Steyvers (2004), analytically marginalising out topics as well as topic proportions. Then, given the newly sampled assignments we update the value for the $\boldsymbol{\xi}$ as well as the concentration parameters $\boldsymbol{\alpha}$. The topic assignment probabilities are given by

$$p(c_d^{(m)} = k) \propto \frac{N_{w,k}^{-c_d^{(m)}} + \gamma}{N_k^{-c_d^{(m)}} + V\gamma} (N_{k,d}^{-c_d^{(m)}} + \alpha_k) \times$$
$$p(y^{(m)}|\{c_j^{(m)}\}_{j=1,j\neq d}^{D^{(m)}}, c_d^{(m)} = k),$$

where $N_{w,k}$ denotes the counts word w (here, $w_d^{(m)} = w$) is assigned to the kth topic, $N_k = \sum_{w=1}^{V} N_{w,k}$ and $N_{k,d}$ denotes counts tokens in document d are assigned to the kth topic. Upper index $-c_d^{(m)}$ means excluding the current count. The parameters of the response distribution are inferred by maximising $\mathcal{L}_{\boldsymbol{\xi}} = \mathcal{L} + \ln p(\boldsymbol{\xi})$. The concentration parameters are updated recursively

$$\alpha_k = \frac{\alpha_k \sum_{m=1}^{M} \psi(N_{k,m} + \alpha_k) - M\psi(\alpha_k)}{\sum_{m=1}^{M} \ln\left(\sum_j N_{j,m} + \alpha_j - \frac{1}{2}\right) - M\psi(\sum_j \alpha_j)},$$

building on Minka's fixed point iteration (Minka, 2000). In the denominator, we approximate $\psi(x) \approx \ln(x - 1/2)$, that is accurate when $x > 1$. This is the case, since all $\mathbf{w}^{(m)}$, for $m = 1, \ldots, M$, contain at least one word token. The asymmetric Dirichlet prior enables pruning irrelevant topics based on the observed data (Wallach et al., 2009).

We note that due to recursive sampling of the topic assignments computational cost of inference may become considerable for large data sets. The recursive property carries also to a corresponding variational Bayesian treatment, since the topic assignments are dependent on each other.

3. Related Work

Previous works on statistical models for ordinal data (Albert & Chib, 1993; Chu & Ghahramani, 2005) assume

$$y^{(m)} = j \quad \text{if} \quad \mu_{j-1} < z^{(m)} \leq \mu_j,$$
$$z^{(m)} \sim \text{Normal}(t^{(m)}, 1),$$

where $z^{(m)}$, for $m = 1, \ldots, M$, denote Gaussian-distributed auxiliary variables. Marginalisation of the $z^{(m)}$ leads to an ordinal probit model. The corresponding inference algorithm relies on truncated Gaussian distributions and takes into account *explicit* ordering constraints for the mean variables leading to a complicated inference algorithm that is sensitive to initialisation thus potentially leading to local minima.

The original supervised LDA model (SLDA; Blei & McAuliffe, 2007) uses *canonical* exponential family distributions for the response model. Under the canonical formulations the expectation of a response variable is $\mathbb{E}[y^{(m)}] = g(t^{(m)})$, where $g(\cdot)$ denotes a link function specific for each member of the family. Examples of the most common members of this family include Gaussian, Bernoulli and Poisson distributions suitable for continuous-valued, binary or count variables, respectively. However, more importantly, the formulation does not support ordinal variables.

Previous applications of SLDA by Blei & McAuliffe (2007); Dai & Storkey (2015) and Nguyen et al. (2013) for ordinal responses, such as product or movie reviews, have made a strong model mis-specification; they treat ordinal variables as continuous-valued. In this approach, the ordinal variables are represented as distinct values in the real domain with arbitrary user-defined intervals between them, enabling use of a Gaussian response model. The model is $y^{(m)} \sim \text{Normal}(t^{(m)} + \mu, \tau^{-1})$, where μ is a mean variable and τ is a precision (inverse variance) parameter. There are a number of statistical flaws in this approach undermining interpretability. First, we note that the mean parameter of the Gaussian distribution, in general, may lead to results that make no sense in terms of the ordinal categories, especially for non-equidistant between-category intervals. Second, observed ratings still take discrete values but the predictions will not correspond to these values. Third, the Gaussian error assumption is not supported by discrete data.

Wang et al. (2009) present an important and non-trivial extension of SLDA to *unordered*, that is, nominal response variables, motivated by classification tasks. The nominal variables represent logically separate concepts that do not permit ordering.

Ramage et al. (2009) and Lacoste-Julien et al. (2009) present alternative joint topic models, where functions of the nominal response variables (class information) affect topic proportions. The response variables are not explicitly modelled using generative formulations. The approach by Mimno & McCallum (2008) uses a similar model formulation suitable for a wide range of observed response variables (or features, in general) performing linear regression from the responses, which are treated as covariates, to the concentration parameters of Dirichlet distributions of the topic proportions. However, it is not obvious how to use these formulations for ordinal response variables.

4. Experiments and Results

We collect consumer-generated reviews of mobile software applications (apps) from Apple's App Store. The review data for each app contains an ordinal rating taking values in five categories ranging from poor to excellent as well as free-flowing text data. We select the vocabulary using tf-idf scores. After simple pre-processing, the data collection contains $M = 5511$ apps with vocabulary size $V = 3995$ and total number of words $\sum_{m=1}^{M} D^{(m)} = 1.5 \times 10^6$. The relatively small data collection is chosen to keep algorithm running times reasonable especially for the sampling-based inference approaches.

4.1. Experimental Setting

We compare the joint correlated topic model (JTM; Section 2.2) and ordinal supervised topic model (SLDA) (Section 2.3) to SLDA with a Gaussian response model as adopted in previous work by Blei & McAuliffe (2007); Dai & Storkey (2015) and Nguyen et al. (2013) (see Section 3 for more details) as well as to sparse ordinal and Gaussian linear regression models. For Gaussian response models we represent the ratings as unit-spaced integers starting from one. The likelihood-specific parameters for the Gaussian model are mean and precision. We adopt the inference procedure described in Section 2.3 using collapsed Gibbs sampling also for Gaussian SLDA. For the regression models we infer a linear combination of the word counts and assign a sparsity-inducing prior distribution for the regression weights over the vocabulary in order to improve predictive ability. We maximise the corresponding joint log likelihood of the model for a fixed prior precision[2]. For all the models that use a Gaussian response model, the mean variable is inferred by computing an empirical response mean. We initialise the models randomly.

For the joint correlated topic model, referred to as, JTM, we bound the maximum number of active topics to $K = 100$, set dimensionality of the latent variables to $L = 30$, $\alpha_0 = 1$, $\beta_0 = 10^{-6}$ and prior precision to $l = L$. The results are shown for $\lambda = 0.001$, although, $\lambda \leq 0.1$ provided also good performance with little statistical variation. We terminated the algorithm (both in training and testing phase), when the relative difference of the (corresponding) lower bound fell below 10^{-4}. The SLDA models were also computed for $K = 100$ and we used $\zeta = 1$. We used 500 sweeps of sampling for inferring the topics and response parameters. For testing we used 500 sweeps of collapsed Gibbs sampling. Although we omit formal time comparisons due to difficulties in comparing VB to MCMC approaches, we find that the sampling approach is roughly one order of magnitude slower. In general, determining convergence for MCMC approaches remains an open research problem, whereas VB provides a local bound for model evidence. For all the topic models we used $\gamma = 0.01$. For JTM, this (effectively) equals a topic Dirichlet concentration parameter value $\gamma + 1$ due to point estimate shifting the value by minus one. For the regression models we sidestep proper cross-validation of the prior precision and show results for the values providing the best performance, potentially leading to over-optimistic results.

4.2. Rating Prediction

We evaluate the models quantitatively in terms of predictive ability. Even though the developed joint mixed membership models are formulated primarily for exploring statistical associations between the ratings and text data, they can readily be used as predictive models. More specifically, we predict the ordinal rating based on the text. We partition available data into multiple training and test sets using 10-fold cross validation. For each model (and fold) we compute the test-set log likelihood (probability) of ratings (the higher, the better) and use these values for comparison. Despite various predictive criteria have been proposed, the selected measure is well motivated by statistical modelling. In the test phase, for JTM, we infer the latent variables $\mathbf{u}$, topic proportions (unnormalised gamma-distributed variables $z_k^{(m)}$) and topic assignments $c_d^{(m)}$ given the values for the remaining parameters inferred in the training phase. For SLDA models the test phase corresponds to estimating the topic assignments using standard LDA model algorithm (using collapsed Gibbs sampling) with fixed topics inferred based on the training data. Finally, we compute the corresponding latent scores $t^{(m)}$ for the models, obtaining the predictions.

Table 1 shows the test-set log likelihoods for the models. The ordinal linear regression model resulted in significantly better predictions than the Gaussian regression model (paired one-sided Wilcoxon; $p < 10^{-3}$) showing that it is important to substitute a statistically poorly motivated Gaussian response distribution with a proper generative model. For both models the sparsity assumption improves predictive ability. For the ordinal regression model, the most relevant words predictive of low (*poor*) ratings include *waste* and *free* and those of high (*excellent*) ratings include *amazing* and *perfect*. The model, however, falls short in providing in-depth interpretations, necessitating the use of topic models.

All the topic models perform substantially better than the regression models. The ordinal SLDA model provides the best predictive performance, JTM is the second best and Gaussian SLDA is the worst. All (pair-wise) comparisons are statistically significant (paired one-sided Wilcoxon; $p < 0.005$). We discovered $K = 100$ is a sufficiently large threshold value for the number of topics; some of the inferred topics are inactive. This, together with good predictive accuracy, establish evidence the developed models have captured the relevant statistical variation in the observed data. For JTM, we also performed a sensitivity analysis of the dimensionality of the latent variables L and found little statistical variation for $30 \leq L \leq 100 = K$. The test log likelihoods range between a minimum of $-669.42(9.68)$ for $L = 80$ and a maximum of $-661.98(11.73)$ for $L = 50$.

Next, we compared the inferred topics of different models quantitatively using a measure, referred to as, *semantic coherence* proposed by Mimno et al. (2011) for quantifying topic trustworthiness. Table 2 shows the average topic

[2] We use $t^{(m)} = \boldsymbol{\xi}^T \mathbf{x}^{(m)}$, where $\mathbf{x}^{(m)}$ denotes word counts over the V-dimensional vocabulary, and $p(\boldsymbol{\xi}|\epsilon) \propto \prod_{d=1}^{V} \exp(-\epsilon \ln(\cosh(\xi_d)))$, where ϵ denotes a precision parameter of the prior distribution.

Table 1. Rating prediction test set log likelihoods for different methods. The table shows values for mean and standard deviation computed over 10 folds obtained by cross-validation.

model	log likelihood
Ordinal SLDA	$-638.53(13.38)$
JTM	$-667.79(15.91)$
Gaussian SLDA	$-681.71(17.69)$
Ordinal regression	$-704.30(13.21)$
Gaussian regression	$-735.40(14.70)$

coherences (the higher, the better). The topics inferred by JTM have significantly larger coherence (two sample one-sided Wilcoxon, $p < 0.0002$).

Table 2. Average semantic coherence values for the inferred topics of different models.

model	coherence
JTM	$-52.64(19.94)$
Oridinal SLDA	$-66.30(26.43)$
Gaussian SLDA	$-67.84(26.54)$

4.3. Inspection of Inferred Topics

Finally, we visualise and interpret the topics inferred by the JTM model. Figures 4 and 5 visualise nine topics associated with high (*excellent*) and low (*poor*) ratings, respectively. As explained in Section 2.2.1, the associations (both sign and strength) are given by computing the similarity scores (that is, correlations).

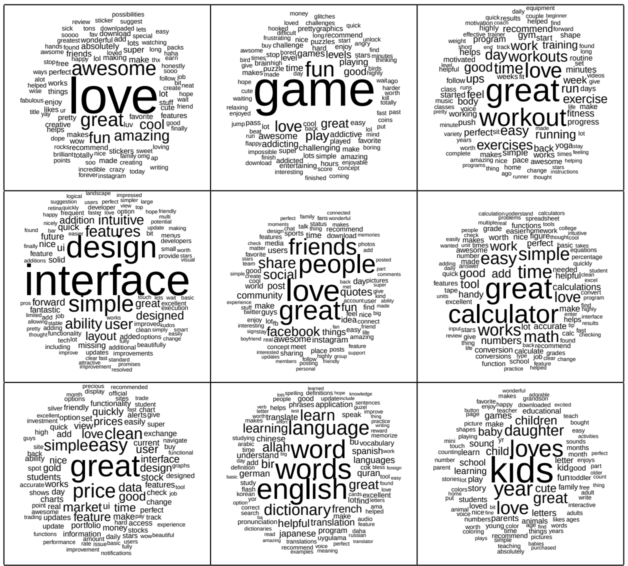

Figure 4. Visual illustration of topics associated with **high ratings**.

One of the topics associated with high ratings (Figure 4) captures word co-occurrence patterns containing adjectives with positive semantics. The remaining topics capture themes customers appreciate, such as games, health monitoring, calculations (for example, for unit conversions), learning languages, social networking and education. One of the topics captures positive customer feedback about app interface and design.

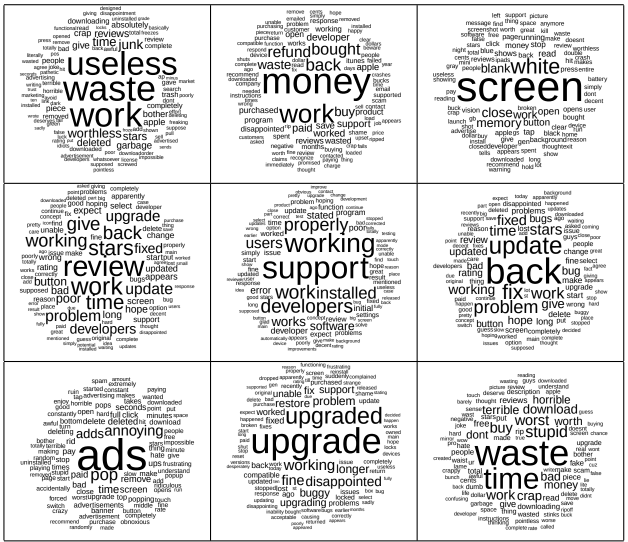

Figure 5. Visual illustration of topics associated with **low ratings**.

The topics associated with low ratings (Figure 5) contain customers' negative experiences or feature requests such as removal of adds, software updates and problems with functionality.

5. Discussion

In this work, we develop a new class of ordinal mixed membership models suitable for capturing statistical associations between groups of observations and co-occuring ordinal response variables for each group. We depart from the existing dominant approach that relies on improper model assumptions for the ordinal response variables. We successfully demonstrate the developed models for analysing reviews of mobile software applications provided by consumers. The proposed class of models as well as inference approaches are applicable for a wide range of present-day applications. In the future, we expect to see improvements in statistical inference including fully Bayesian treatments and nonparametric Bayesian formulations. Stochastic online learning or model formulations for streaming data may be applied to scale the statistical inference to cope with current data repositories containing review data for a few millions of groups.

Acknowledgement

This research is supported by an EPSRC program grant *A Population Approach to Ubicomp System Design* (EP/J007617/1).

References

Albert, James H and Chib, Siddhartha. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

Blei, David and Lafferty, John. Correlated topic models. In *Advances in Neural Information Processing Systems*, 2006.

Blei, David M and McAuliffe, Jon D. Supervised topic models. In *Advances in Neural Information Processing Systems*, 2007.

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

Böhning, Dankmar. Multinomial logistic regression algorithm. *Annals of the Institute of Statistical Mathematics*, 44(1):197–200, 1992.

Chu, Wei and Ghahramani, Zoubin. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6:1019–1041, 2005.

Dai, Andrew and Storkey, Amos J. The supervised hierarchical Dirichlet process. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(2):243–255, 2015.

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

Hoffman, Matthew D, Blei, David M, Wang, Chong, and Paisley, John. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Jaakkola, T and Jordan, Michael I. A variational approach to Bayesian logistic regression models and their extensions. In *Artificial Intelligence and Statistics*, 1997.

Lacoste-Julien, Simon, Sha, Fei, and Jordan, Michael I. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Advances in Neural Information Processing Systems*, 2009.

Mimno, David and McCallum, Andrew. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Uncertainty in Artificial Intelligence*, 2008.

Mimno, David, Wallach, Hanna M, Talley, Edmund, Leenders, Miriam, and McCallum, Andrew. Optimizing semantic coherence in topic models. In *Empirical Methods in Natural Language Processing*, 2011.

Minka, Thomas. Estimating a Dirichlet distribution, 2000.

Nguyen, Viet-An, Boyd-Graber, Jordan L, and Resnik, Philip. Lexical and hierarchical topic regression. In *Advances in Neural Information Processing Systems*, 2013.

Paisley, John, Wang, Chong, Blei, David M, et al. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(2):235–272, 2012.

Ramage, Daniel, Hall, David, Nallapati, Ramesh, and Manning, Christopher D. Labelled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Empirical Methods in Natural Language Processing*, 2009.

Wainwright, Martin J and Jordan, Michael I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Wallach, Hanna M, Minmo, David, and McCallum, Andrew. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, 2009.

Wang, Chong and Blei, David M. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14(1):1005–1031, 2013.

Wang, Chong, Blei, David, and Li, Fei-Fei. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition*, 2009.