Stat

The ISI's Journal for the Rapid
Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.103

Uncovering smartphone usage patterns with multi-view mixed membership models

Seppo Virtanen[a]*, Mattias Rost[b], Alistair Morrison[b], Matthew Chalmers[b] and Mark Girolami[a,c]

Received 23 November 2015; Accepted 14 December 2015

We present a novel class of mixed membership models for combining information from multiple data sources inferring inter-view and intra-view statistical associations. An important contemporary application of this work is the meaningful synthesis of data sources corresponding to smartphone application usage, app developers' descriptions and customer feedback. We demonstrate the ability of the model to infer meaningful, interpretable and informative app usage patterns based on the app usage data augmented with rich text data describing the apps. We provide quantitative model evaluations showing the model provides significantly better predictive ability than comparative related existing methods. © 2016 The Authors. *Stat* Published by John Wiley & Sons Ltd.

Keywords: Bayesian methods; categorical data; machine learning; statistical inference; statistical modelling

1 Introduction

Mixed membership modelling is a powerful modelling methodology, which employs exchangeability assumptions (Aldous, 1985) to simplify and speed up posterior computations, for inferring useful structure from rich grouped data sources for summarization and prediction (Airoldi et al., 2014). Observations are grouped, and each group is modelled with a mixture model. Components of the mixtures are shared across groups, whereas groups may have different mixture proportions. Recent research has shown these models may be used to capture meaningful and useful co-occurrence patterns based on, for example, text data (Blei et al., 2003; Griffiths & Steyvers, 2004), image patches (Sivic & Zisserman, 2003; Sivic et al., 2005) and computer (Linstead et al., 2008) or genetic code (Pritchard et al., 2000). For text data, these models, also referred to as topic models with a prominent model called latent Dirichlet allocation, are intended to improve information retrieval and to discover and organize large collections of online content by automatically extracting semantic themes (topics) based on the observed data (Blei et al., 2003). Latent Dirichlet allocation assumes both mixture components (topics), and proportions are drawn from Dirichlet distributions and observations (words tokens over a vocabulary) as well as mixture assignments are drawn from categorical distributions. *A priori*, the topics are deemed to capture recurring patterns (co-occurrences) of observations.

Previous work has noticed single-view topic modelling may infer topics that are not meaningful. Problems stem from large vocabularies and small overlap of semantically/contextually similar words over different groups. Thus, capturing

[a]Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[b]School of Computing Science, University of Glasgow, Glasgow G12 8RZ, UK

[c]The Alan Turing Institute for Data Science, British Library, 96 Euston Road, London NW1 2DB, UK

*Email: s.virtanen@warwick.ac.uk

(wileyonlinelibrary.com) DOI: 10.1002/sta4.103

dependence between two similar words may be impossible because of data sparsity. To overcome this issue, vocabulary information, such as similarity graphs between pairs of words, have been used to improve semantic topical coherence and predictive ability via informative prior distributions for the topics (Newman et al., 2011; Petterson et al., 2010). Newman et al. (2011) replace the Dirichlet distribution with heuristic priors that do not admit fully Bayesian inference undermining inference of smoothing (concentration) parameters. Petterson et al. (2010) retain a Dirichlet prior but use an informative prior for the concentration parameters leading to a complicated inference algorithm, because the cost function to optimize involves a normalizing constant of the Dirichlet distribution.

Multi-view mixed membership models aim to combine information from multiple data sources with co-occurring groups. Learning from multiple data sources may be used to improve predictive ability or interpretability for the primary data source of interest as well as, equally importantly, to uncover statistical associations between multiple data sources. Applications of the model structures defined herein include text documents in multiple languages (Mimno et al., 2009), text documents that co-occur with images (Barnard et al., 2003) and textual review data coupled with categorical ratings (Virtanen & Girolami, 2015). Multi-view topic models, multi-field correlated topic model (Salomatin et al., 2009) and factorized multi-modal topic model (Virtanen et al., 2012) assume multiple co-occurring groups and are able to capture inter-view and intra-view topical correlations. These approaches build on a finite single-view correlated topic model (Blei & Lafferty, 2005) as well as its non-parametric formulation (Paisley et al., 2012), respectively, and are computationally demanding, because they work with unconstrained covariance matrices of Gaussian distributions, further complicating analysis.

In this work, we generalize mixed membership models to capture statistical associations between multiple data sources (views) of an interconnected set of objects via probabilistic hierarchical modelling. We present a structured prior for multivariate continuous probability distributions over a finite set of categories that may be interpreted as an essential building block in multi-view mixed membership modelling. We use the prior to capture dependencies between the categories and especially for combining information from multiple views. The prior assumes a set of continuous-valued latent variables that are linearly combined with category-specific projections in the log-domain regressing from the latent variables to probability distributions. The construction builds on normalized scaled positive Gamma-distributed variables. Our multi-view model formulation adopts the framework of probabilistic hierarchical modelling via sharing the latent variables across groups from multiple views in a flexible manner. We introduce group-specific latent variables for constructing the expectation parameters of categorical distributions for the mixture components and topic proportions for different views for capturing inter-view and intra-view correlations.

Our multi-view model is more flexible than the related multi-view topic models (Salomatin et al., 2009; Virtanen et al., 2012) that are constrained by the assumption that all the groups co-occur. These models are not flexible enough for the application considered in this work, as detailed in the following. Our model assumes a set of shared latent variables between the views, instead of assuming an underlying covariance matrix of a Gaussian distribution, leading to computationally more scalable posterior inference, admitting analysis of the inferred latent variables following factor model conventions (for example, Murphy, 2012) and for constructing model extensions in a computationally tractable and intuitive manner.

We study if it is plausible to incorporate the group-wise correlation structure captured by the latent variables to infer more meaningful and useful mixture components. Our work is related to the aforementioned works by Newman et al. (2011) and Petterson et al. (2010) for constructing more meaningful topics (mixture components). However, our model differs from these single-view constructions in several ways; our approach is not limited to pair-wise similarity information for word tokens; we infer group-specific latent variables in an unsupervised approach instead of conditional (regressive) formulations and use a decomposable model construction leading to simplified posterior inference. Further, unsupervised multi-view formulation of our model enables analysis of statistical dependencies between the views and inter-view prediction.

Variational Bayesian (VB) methods have been recently developed for mixture and mixed membership modelling obtaining closed form approximations to analytically intractable model posteriors in reasonable computation time (Airoldi et al., 2014; Bishop, 2006; Murphy, 2012; Wainwright & Jordan, 2008). In this work, we adopt this inference framework and derive a novel, efficient and scalable VB algorithm for the model. VB methods introduce a factorized distribution for the unknown variables of the model and infer the parameters (as well as functional form, in some cases) of the factorized distribution by minimizing a Kullback–Leibler divergence between the approximation and model posterior, alternatively, maximizing a lower bound for the model evidence. In order to retain analytical tractability of the factorization, one may need to assume a fully factored factorization with respect to the variables. However, the more factored, the less accurate the posterior approximation will be, introducing additional bias and inefficient updates (especially) for strongly correlated variables. We assume less factorization leading to less biased estimates following the inference scheme proposed by Mimno et al. (2012). Despite the factorization becomes analytically intractable to compute, we can draw samples from it and accordingly use Monte Carlo sample estimates for the variables. To further decrease bias, we analytically marginalize out some of the variables of the model. In addition to introducing bias, VB methods, in particular for mixture models, may be prone to local optima. In more detail, for a small enough expected value of the mixture proportion, the value is not likely to be affected by consecutive updates. Instead, the value may become clamped to values close to zero because of properties of the digamma function used to calculate the analytic expectations with respect to the factorization. The digamma function is extremely non-linear for small values. Thus, small perturbations will have profound effects for computations. Further, in hierarchical models, small parameter values affect inferences for the corresponding hyperparameters encouraging excessive sparsity (for example, shrinkage towards zero) and tendency to remain stuck in a local solution. We reduce the risk of local maxima (clamping) by approximating analytical log-expectations of random variables with respect to the factorized distribution. This amounts to approximating the digamma function with the logarithm function that has more stable behaviour for small values. In summary, we assume less factorization assumptions than commonly adopted in the fully factored approach and approximate the complicated digamma function appearing in analytically tractable expectations for obtaining approximate expectations that have robust behaviour for small values.

In the application of this work, we demonstrate that data collected from online data sources can be used to provide important insights, pave the way to improve understanding and prediction of human behaviour with technology and online digital content. Our analysis focuses on, but is not limited to, user-interactions with mobile software applications (referred to as apps), for which online app stores provide abundant descriptive and qualitative information. On one hand, we augment statistical analysis of app usage by combining rich textual information of apps from app stores using descriptions by software developers and feedback by software consumers. On the other hand, we infer statistical associations between actual app usage, descriptions, which reflect desired purpose and functionality, as well as customer feedback, which summarize user experiences and feature requests. There is a need to exploit multiple sources of information in these contexts, and this is the first work to formally quantify app usage with contextual text data of apps.

We show the proposed model provides useful and meaningful summaries of app usage inferring usage patterns (mixture components for app usage) that may be interpreted in terms of associated text topics. We provide quantitative model evaluations using predictive ability and show that our model performs significantly better than alternative related existing models that, by construction, are unable to employ all the data sources available. In addition, we show the VB algorithm that employs the proposed digamma approximations provides notable improvements in predictive ability compared with the standard approach.

The paper structure is as follows. Section 2 presents the structured prior distribution, and Section 3 introduces the statistical description of the model and the VB posterior inference algorithm. Section 4 presents the experiments and results. Finally, Section 5 concludes the paper.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.103

The ISI's Journal for the Rapid
Dissemination of Statistics Research

2 Correlated Dirichlet distribution

We present a construction for multivariate continuous probability distributions $\mathbf{x}$ over $C \geq 3$ categories, $\mathbf{x} \in \Delta^C$, that is able to capture correlations between the categories. The distribution serves as a prior distribution for the expectation parameter of a categorical or multinomial distribution. Although the Dirichlet distribution, Dirichlet$(\mathbf{x}|\boldsymbol{\alpha})$, is a canonical alternative for $\mathbf{x}$ specified by a concentration parameter vector $\boldsymbol{\alpha}$, it is unable to capture non-trivial correlations between the categories. Our construction builds on normalized, scaled and positive variables; we assume Gamma-distributed random variables

$$g_k \sim \text{Gamma}(\alpha_0, \exp(-\mathbf{w}_k^T \mathbf{b} - \mu_k)) \tag{1}$$

and define

$$x_k = \frac{g_k}{\sum_{k'=1}^{C} g_{k'}}.$$

Here, we introduce $R \geq 1$ dimensional continuous-valued latent variables

$$\mathbf{w}_k \sim \text{Normal}(\mathbf{0}, \frac{1}{R}\mathbf{I})$$

and

$$\mathbf{b} \sim \text{Normal}(\mathbf{0}, \mathbf{I})$$

as well as mean variables μ_k drawn from a non-informative prior distribution, for $k = 1, \ldots, C$. Expectation is given by

$$\mathbb{E}[x_k] \propto \mathbb{E}[g_k] = m_k \exp(\mathbf{w}_k^T \mathbf{b}), \tag{2}$$

where $m_k = \alpha_0 \exp(\mu_k)$.[†] Interestingly, setting the rate parameter of the Gamma distribution in (1) to value one, the construction reduces to an alternative formulation for a Dirichlet distribution using normalized Gamma variables that are statistically independent.

The prior, referred to as correlated Dirichlet (CD) distribution, is useful for capturing statistical associations between multiple realizations of $\mathbf{x}_i$, for $i = 1, \ldots, I$ over the common categories, sharing the category-specific latent variables $\mathbf{w}_k$, for $k = 1, \ldots, C$, and assigning a separate object-specific latent variable $\mathbf{b}_i$ for each realization. When $R < \min(I, C)$, the prior induces correlations between the multiple categories as well as realizations, bearing resemblance to a general class of factor models.

3 Model description and data

We observe launches of mobile software applications (apps), referred to as $a_{s,d}$, over a heterogenous population of $s = 1, \ldots, S$ software consumers. The D_s launches for the sth user are collected in a group $\boldsymbol{a}_s = \{a_{s,1}, a_{s,2}, \ldots, a_{s,D_s}\}$, where the launches $a_{s,d}$, for $d = 1, \ldots, D_s$, take discrete values over a population-level vocabulary $\mathcal{A}$ of M apps, $|\mathcal{A}| = M$.

[†] For $x \sim \text{Gamma}(a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$, where $\Gamma(\cdot)$ denotes the Gamma function and shape and rate parameters are denoted by a and b, respectively, expected value is given by $\mathbb{E}[x] = a/b$.

For the mth app in the vocabulary, $\mathcal{A}_m$, we collect textual data consisting of software consumers' feedback as well as descriptions provided by software developers or app stores. The text data for the mth app in the vth view contain a group of word tokens $w_n^{(v)}$, where $v = 1, \ldots, V$ and $n = 1, \ldots, N_m^{(v)}$, over a $V^{(v)}$-dimensional word vocabulary $\mathcal{W}^{(v)}$, $|\mathcal{W}^{(v)}| = V^{(v)}$. Throughout the paper, we use the upper index v to denote the (co-occurring) groups of different views corresponding to descriptions and reviews.

In the following, we present a statistical joint model for the app launches and text data. To alleviate model description, we first explain the generative process for the app launches and, second, describe the model for the co-occurring word groups.

3.1 Model for app launches

We assume *a priori* a set of T distributions specified over $\mathcal{A}$, also referred to as app patterns (mixture components for app launches), $\boldsymbol{\eta}_t$, where $t = 1, \ldots, T$, have high probability for frequently co-occurring apps intendedly capturing common themes according to app functionality and user behaviour. The app patterns correspond to mixture components of a mixed membership model that is used to generate the observations.

The generative process for the app launches $\boldsymbol{a}_s$ proceeds repeatedly by drawing first a mixture assignment

$$c_{s,d} \sim \text{Categorical}(\boldsymbol{\theta}_s),$$

where

$$\boldsymbol{\theta}_s \sim \text{Dirichlet}(\boldsymbol{\alpha})$$

denotes a user-specific mixture proportion, and then, given the assignment,

$$a_{s,d} \sim \text{Categorical}(\boldsymbol{\eta}_{c_{s,d}}).$$

We construct the app patterns $\eta_{m,t}$, for $m = 1, \ldots, M$ and $t = 1, \ldots, T$, using the CD distribution presented in Section 2 introducing Gamma-distributed random variables $\beta_{m,t}$,

$$\begin{aligned} \eta_{m,t} &= \frac{\beta_{m,t}}{\sum_{m'=1}^{M} \beta_{m',t}}, \\ \beta_{m,t} &\sim \text{Gamma}\big(\alpha_0, \exp(-\mathbf{v}_t^T \mathbf{u}_m)\big). \end{aligned} \tag{3}$$

Here, $\mathbf{u}_m$ denotes latent variables specific for the mth app, $\mathbf{v}_t$ denotes latent variables for the tth pattern and α_0 denotes a shape parameter of the Gamma distribution. We set the mean variables included in (1) to zero ($\mu_t = 0$) to avoid over-parameterization.

The construction for the app patterns (3) is able to capture correlations between the apps using a low-rank decomposition. We may interpret $\mathbf{v}$ as latent projections from $\mathbf{u}$ to $\boldsymbol{\eta}$. Any two apps $\mathbf{u}_i$ and $\mathbf{u}_j$, whose distance in the latent space is (relatively) small, are dependent. Depending on the actual values for the latent variables as well as projections, such dependence may imply app co-occurrence.

3.2 Model for co-occurring app text groups

For the text collections, we assume a set of $K^{(v)}$ topics (mixture components for word tokens) $\boldsymbol{\nu}_k^{(v)}$ capture significant word co-occurrences over the word dictionaries. The generative process for the word tokens $w_{m,n}^{(v)}$, for $v = 1, \ldots, V$, $m = 1, \ldots, M$ and $n = 1, \ldots, N_m^{(v)}$, is

(wileyonlinelibrary.com) DOI: 10.1002/sta4.103

$$w_{m,n}^{(v)} \sim \text{ Categorical}\left(\boldsymbol{\nu}_{z_{m,n}^{(v)}}^{(v)}\right),$$

where

$$z_{m,n}^{(v)} \sim \text{ Categorical}\left(\boldsymbol{\omega}_m^{(v)}\right)$$

and

$$\boldsymbol{\nu}_k^{(v)} \sim \text{ Dirichlet}\left(\gamma^{(v)}\mathbf{1}\right).$$

We construct the mixture (topic) proportions $\omega_{m,k}^{(v)}$, for $m = 1, \ldots, M$ and $k = 1, \ldots, K^{(v)}$, introducing additional Gamma-distributed random variables $y_{m,k}^{(v)}$,

$$\begin{aligned} \omega_{m,k}^{(v)} &= \frac{y_{m,k}^{(v)}}{\sum_{k'=1}^{K^{(v)}} y_{m,k'}^{(v)}}, \\ y_{m,k}^{(v)} &\sim \text{Gamma}\left(\beta_0^{(v)}, \exp\left(-\mathbf{u}_m^T \boldsymbol{\xi}_k^{(v)} - \mu_k^{(v)}\right)\right), \end{aligned} \tag{4}$$

where $\mathbf{u}_m$ denotes app-specific latent variables, $\boldsymbol{\xi}_k^{(v)}$ denotes topic-specific latent variables, $\boldsymbol{\mu}^{(v)}$ denotes a vector of mean variables and $\beta_0^{(v)}$ is a shape parameter of the Gamma distribution.

The model employs CD distributions (4) and captures correlations between and within topics of views via the common latent variables $\mathbf{u}_m$, for $m = 1, \ldots, M$. Contextually, similar apps (based on the app descriptions and customer reviews) have high probability to share similar values for the latent variables.

3.3 Interpretation of the joint model

The generative models for the app launches and text groups share the app-specific latent variables $\mathbf{u}_m$, for $m = 1, \ldots, M$, appearing in (3) and (4). The app-specific latent variables act as an information channel between the app patterns and topic proportions capturing statistical associations (correlations) not only between but also within app launches and co-occurring text data of apps in an unconstrained and unsupervised manner. Two (sets of) apps that are close by in the latent space have high probability not only for usage co-occurrence but also for sharing the same functionality and context.

The mean variables $\mu_k^{(v)}$, for $v = 1, \ldots, V$ and $k = 1, \ldots, K^{(v)}$, function the same role as the app pattern concentration parameters $\boldsymbol{\alpha}$ capturing topic or pattern probabilities, correspondingly. For (relatively) small values, the corresponding patterns (or topics) have little probability to occur.

A joint distribution of the observed data $\mathcal{D} = \left\{\mathbf{a}_s, \mathbf{w}_m^{(v)}\right\}_{s,m,v}$ and unobserved variables Θ is

$$\begin{aligned} p(\mathcal{D}, \Theta) = \prod_{s=1}^{S} \prod_{d=1}^{D_s} \prod_{v=1}^{V} \prod_{m=1}^{M} \prod_{n=1}^{N_m^{(v)}} \prod_{t=1}^{T} \prod_{k=1}^{K^{(v)}} & p(a_{s,d}|c_{s,d}, \boldsymbol{\beta}) p(\beta_{m,t}|\mathbf{v}_t, \mathbf{u}_m, \alpha_0) p(c_{s,d}|\boldsymbol{\theta}_s) p(\boldsymbol{\theta}_s, \boldsymbol{\alpha}) \\ & p\left(w_{m,n}^{(v)}|z_{m,n}^{(v)}, \boldsymbol{\nu}^{(v)}\right) p\left(z_{m,n}^{(v)}|\mathbf{y}_m^{(v)}\right) p\left(y_{m,k}^{(v)}|\boldsymbol{\xi}_k^{(v)}, \mathbf{u}_m, \mu_k^{(v)}, \beta_0^{(v)}\right) p\left(\boldsymbol{\nu}_k^{(v)}, \gamma^{(v)}\right) p\left(\mathbf{u}_m, \mathbf{v}_t, \boldsymbol{\xi}_k^{(v)}, \alpha_0, \beta_0^{(v)}\right). \end{aligned}$$

To complete the model description, we set a vaguely informative $\text{Gamma}(1,1)$ prior for the scale parameters α_0 as well as $\beta_0^{(v)}$ and non-informative priors for the Dirichlet concentration parameters $\boldsymbol{\alpha}$ and $\gamma^{(v)}$, for $v \in (1, V)$. Figure 1 illustrates a graphical plate diagram of the model.

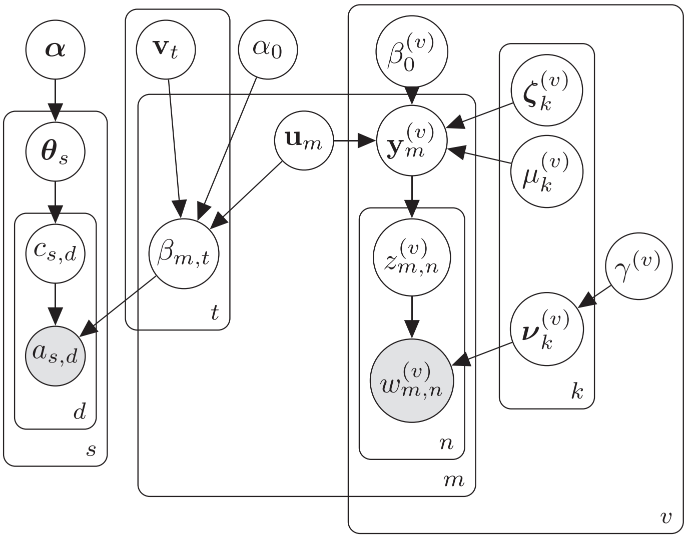

Figure 1. A graphical plate diagram of the model. Unshaded nodes correspond to unobserved variables Θ, whereas shaded nodes correspond to observed variables $\mathcal{D} = \left\{\boldsymbol{a}_s, \mathbf{w}_m^{(v)}\right\}_{s,m,v}$. Hyperparameters for the root nodes, whose values need to be fixed before posterior inference, are omitted from the diagram for clarity of representation. From left to right, plates indicate replication over users, app launches, patterns, apps, word tokens, topics and views. The nodes on the left-hand side of the $\boldsymbol{u}_m$ correspond to the model for the app launches (Section 3.1), whereas the nodes on the right-hand side of the $\boldsymbol{u}_m$ represent the multi-view mixed membership model of co-occurring groups (Section 3.2).

3.4 Structured relaxed variational Bayesian inference

We introduce a factorized distribution for the unknown variables of the model (unshaded variables appearing in Figure 1) $q(\Theta)$ and infer the parameters of the $q(\Theta)$ by minimizing a Kullback–Leibler divergence between the $q(\Theta)$ and model posterior $p(\Theta|\mathcal{D})$, alternatively, maximizing a lower bound for the model evidence

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

where expectations are taken with respect to the $q(\Theta)$. Optimization proceeds coordinate-wise for updating the parameters of each factored distribution given the (current) values for the remaining distributions.

We assume the factorized distribution

$$q(\Theta) = \prod_{s,m,v,t,k} q(\beta_{m,k}) q\left(y_{m,k}^{(v)}\right) q(\mathbf{c}_s) q\left(\mathbf{z}_m^{(v)}\right), \tag{5}$$

analytically marginalizing out the app pattern proportions $\boldsymbol{\theta}_s$, for $s = 1, \ldots, S$, and topics $\boldsymbol{\nu}_k^{(v)}$, for $v = 1, \ldots, V$ and $k = 1, \ldots, K^{(v)}$. To simplify computations, we use point distributions (omitted from the factorization for notational clarity) for $\boldsymbol{\alpha}, \gamma^{(v)}, \alpha_0^{(v)}, \beta_0^{(v)}, \mathbf{u}_m, \mathbf{v}_t$ and $\boldsymbol{\zeta}_k^{(v)}$. The factorized distributions for $\beta_{m,t}$ and $y_{m,k}^{(v)}$ are

$$q(\beta_{m,t}|a_{m,t}, b_{m,t}) = \text{Gamma}(a_{m,t}, b_{m,t}),$$
$$q\left(y_{m,k}^{(v)}|\tilde{a}_{m,k}^{(v)}, \tilde{b}_{m,k}^{(v)}\right) = \text{Gamma}\left(\tilde{a}_{m,k}^{(v)}, \tilde{b}_{m,k}^{(v)}\right).$$

The factorized joint distributions for the mixture assignments $\mathbf{c}_s$ and $\mathbf{z}_m^{(v)}$ cannot be evaluated in closed form, but we can still draw samples from these factorizations and use Monte Carlo sample estimates to compute required statistics. The parameters of the $q(\Theta)$, Eq. (5), are

$$p(c_{s,d}=t) \propto \frac{\mathbb{E}[\,\beta_{a_{s,d},t}]}{\sum_{m'=1}^{M}\mathbb{E}[\,\beta_{m',t}]}\left(N_{a_{s,d},t}^{-(c_{s,d})}+\alpha_t\right), a_{m,t} = \quad \alpha_0+\widehat{N}_{m,t}, \tag{6}$$

$$b_{m,t} = \frac{\sum_{m'}\widehat{N}_{m',t}}{\sum_{m'}\mathbb{E}[\,\beta_{m',t}]} + \exp\left(-\mathbf{v}_t^T\mathbf{u}_m\right), \tag{7}$$

$$\begin{aligned} p\left(z_{m,n}^{(v)}=k\right) \propto & \frac{N_{k,w_{m,n}^{(v)}}^{-\left(z_{m,n}^{(v)},v\right)}+\gamma^{(v)}}{N_k^{-\left(z_{m,n}^{(v)},v\right)}+V^{(v)}\gamma^{(v)}}\mathbb{E}\left[y_{m,k}^{(v)}\right], \\ \tilde{a}_{m,k}^{(v)} =& \beta_0^{(v)}+\widehat{N}_{m,k}^{(v)}, \\ \tilde{b}_{m,k}^{(v)} =& \frac{\sum_{k'}\widehat{N}_{m,k'}^{(v)}}{\sum_{k'}\mathbb{E}\left[y_{m,k'}^{(v)}\right]} + \exp\left(-\mathbf{u}_m^T\boldsymbol{\zeta}_k^{(v)}-\mu_k^{(v)}\right). \end{aligned} \tag{8}$$

Here, the different matrices N collect counts as in collapsed Gibbs sampling (Griffiths & Steyvers, 2004) and $\widehat{N}$ denotes Monte Carlo expectation estimates (after burn in). We have used first-order Taylor expansion to approximate, firstly, the log-expectation avoiding the aforementioned digamma problem

$$\mathbb{E}[\ln\beta_{m,t}] = \psi(a_{m,t}) - \ln b_{m,t} \geq \ln\mathbb{E}[\,\beta_{m,t}] = \ln(a_{m,t}) - \ln(b_{m,t}),$$

where $\psi(\cdot)$ denotes the digamma function, and, secondly, the analytically intractable log normalizing constant for computational tractability

$$\mathbb{E}[\ln\sum_{m'=1}^{M}\beta_{m',t}] \geq \ln\sum_{m'=1}^{M}\mathbb{E}[\,\beta_{m',t}],$$

introducing variational auxiliary variables. The digamma approximation affects expectations in the assignment updates, Eqs. (6) and (8). Because of the approximation, small expected values are not pushed strongly towards zero by the digamma function.

To update the latent variables (as well as the scale variables α_0 and $\beta_0^{(v)}$ of the Gamma distributions), we maximize the lower bound, retaining the relevant terms,

$$\widetilde{\mathcal{L}} = \sum_{m,v,t,k}\left(\mathbb{E}\left[\ln p(\beta_{m,t}|\mathbf{v}_t,\mathbf{u}_m,\alpha_0)\right] + \mathbb{E}\left[\ln p\left(y_{m,k}^{(v)}|\boldsymbol{\zeta}_k^{(v)},\mathbf{u}_m,\mu_k,\beta_0^{(v)}\right)\right] + \ln p\left(\mathbf{u}_m,\mathbf{v}_t,\boldsymbol{\zeta}_k^{(v)}\right)\right)$$

using a multi-purpose unconstrained gradient-based optimization technique called limited-memory Broyden–Fletcher–Goldfarb–Shanno (Byrd et al., 1995).

The digamma approximation affects also updates of the Gamma scale variables. The problem with analytical updates is that small values for the β (or y) encourage α_0 (β_0) to decrease strongly. This behaviour may result in excessive model sparsity and a tendency to get stuck in a local optimum. The digamma approximation addresses this problem.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.103

We update the pattern proportion and topic Dirichlet concentration parameters maximizing the partially marginalized evidence using Minka's fixed-point iteration (Minka, 2000).

4 Experiments and results

4.1 Data collection

We have collected app launches over a heterogenous and diverse population of $S = 2000$ software consumers. The number of unique apps, for which we have collected textual descriptions and reviews from app stores, is $M = 24,274$. We construct the word vocabularies retaining words that occur in at least one percentage of the apps. For the descriptions, the vocabulary size is $|\mathcal{W}^{(1)}| = 1500$ and for the reviews $|\mathcal{W}^{(2)}| = 3800$. Depending on availability, we note that text data groups for each app may contain both descriptions and reviews or either one. Accordingly, we modify the model to cope with missing (empty) groups.

4.2 Experimental setting

Given the computational resources available, we set an upper bound for the number of both app patterns and text topics to $T = K^{(v)} = 50$ and the dimensionality of the latent variables $R = 20$. We initialize unknown quantities of the models randomly and run the posterior inference algorithm for 200 iterations. We adopt standard tools for checking algorithm convergence and find the number of iterations to suffice. We provide an implementation of the model in the R language at https://github.com/svirtanen/mvmmm.

4.3 Visualization of the latent structure

Distances between the latent variables $\mathbf{u}_m$, for $m = 1, \dots, M$, may be used to represent similarities between any pairs of apps useful for information retrieval and visualization. We use t-distributed stochastic neighbour embedding (Van Der Maaten, 2014) to infer app locations in a lower-dimensional space aiming to preserve the true neighbourhoods described by the similarities by minimizing information retrieval measure recall. Figure 2 shows a kernel density

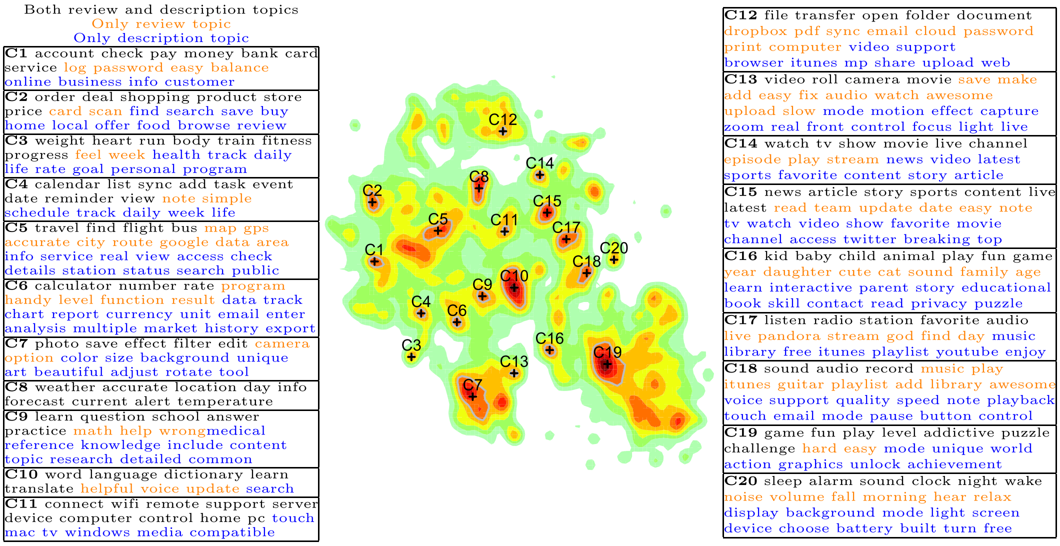

Figure 2. Visualization of the inferred app structure. See Section 4.3 for more details.

estimate of the locations revealing explicit clustering structure. The colour scale from green to red corresponds to low to high probability values, respectively, whereas white denotes values zero. We label some of the clusters and associate each labelled cluster with most probable inferred review and description topics showing lists of representative top 20 colour-coded words in tables accompanying the density estimate in Figure 2. The words that appear in both views are black, whereas orange and blue colour is used to distinguish between words appearing in reviews and descriptions, correspondingly. We note that the top words of the associated review topics often correspond to words with positive semantics, such as *love* and *great*; they are omitted from the illustration because of space constraints. Based on Figure 2, we see that the apps form highly informative and useful clusters according to various different app functionalities or themes, such as banking, shopping, fitness, calendar, travelling, gaming, weather, video streaming, photography, news, education and music apps.

4.4 Quantitative evaluation of app patterns

Because performing a qualitative evaluation of the inferred app patterns is difficult, we resort to performing a quantitative model evaluation, following a scheme provided by Wallach et al. (2009). We compare our model with a single-view model for the app launches; no alternative multi-view model formulations exist for the setting proposed, and no similarity-graph information for the app vocabulary is available. We evaluate approximate predictive log likelihoods for held-out folds of cross-validation and note that a better model results in higher predictive values.

In more detail, we partition the users into train and test sets using 10-fold cross-validation. We infer the factorized posterior approximation based on the train users. Following Asuncion et al. (2009), we further partition the app launches for the test users into two equally sized groups and evaluate the models in a group completion task and based on the first group of test launches, predict which apps are likely to be launched, essentially, computing an approximate predictive likelihood of the other group of test data given the first. For the sth test user, we use the first group to infer the user-specific app pattern proportions $\mathbb{E}[\theta_{s,t}]$, for $t = 1, \ldots, T$, and then use the inferred $\mathbb{E}[\boldsymbol{\theta}_s]$ to obtain a user-specific distribution over the apps,

$$p_{s,m} = \sum_t \mathbb{E}[\theta_{s,t}]\,\mathbb{E}[\eta_{m,t}],$$

and finally compute the log likelihood (of a multinomial distribution) of the second group given the $\mathbf{p}_s$. We repeat the process for all the test users and sum the individual test log likelihoods.

The comparison single-view model is obtained by replacing the proposed prior for the app patterns in Eq. (3) with a symmetric Dirichlet prior distribution over the app vocabulary. We set the rate parameter of the Gamma distribution in (3) to one and use the normalized Gamma construction for the app patterns (Section 3.1). For inference, we use the updates provided in Section 3.4; we simply set $\mathbf{v}_t^T \mathbf{u}^{(m)} = 0 \forall (t, m)$ in Eq. (7) and discard variables related to the text data sources. Such a comparison allows us to isolate effects of different inference algorithms and model structures to focus on the effect of the latent variable construction of the app patterns.

Our model has better performance than the single-view model (one-sided paired Wilcoxon; $p < 0.04$). This result indicates that the app reviews and descriptions contain useful information for explaining app usage and that the proposed joint model is able to capture that information and infer more meaningful app patterns.

We also note that the variational inference algorithm with the digamma approximations (Section 3.4) results in better performance compared with the algorithm that uses analytically exact updates (one-sided paired Wilcoxon; $p < 10^{-3}$). This methodological improvement is due to more robust expectation evaluations required not only for the update of the app patterns as well as the text mixture assignments but also for the update of the scale parameters of the

Gamma distributions controlling the amount of smoothing. We find that the inferred smoothing parameters based on the exact updates correspond to smaller values (that is, less smoothing) than the ones obtained based on the digamma approximations.

4.5 Quantitative evaluation of text topics

We also perform a quantitative evaluation for the text topics. We partition apps for which both descriptions and reviews are available into 10 training and test folds (again using cross-validation) and compute log likelihood for held-out reviews, following an evaluation scheme proposed by Virtanen et al. (2012) for inter-view prediction (completion) of groups. We infer the models in a transductive setting; the test text review groups are empty. In the test phase, we first compute review topic proportions for the test apps via the inferred latent variables **u** and corresponding mappings using Eq. (2). Then, we linearly combine the computed proportions with the expected review topics to calculate the log likelihood of the held-out reviews.

The comparison model is obtained similarly to the single-view model for the app launches as described earlier but now using the remaining model (and inference updates) for the descriptions and reviews (Sections 3.2 and 3.4), omitting the app launches. The model is essentially a multi-view mixed membership model of co-occurring text groups that is unable to combine relevant information from the app usage. The comparison allows us to study the predictive effect of the latent variables that connect app usage to text data.

Our model performs better than the comparison model (paired one-sided Wilcoxon; $p < 10^{-3}$). The result establishes strong evidence that the app usage is associated with app reviews (as well as descriptions) and that our model is able to uncover that underlying predictive information inferring more meaningful text topics.

5 Discussion

In this work, we set out with the aim of combining information from online app store data repositories for uncovering statistical smartphone software application usage patterns based on trace logs of mobile applications collected over a heterogenous and diverse population of software consumers. We develop novel modelling methodology suitable for learning from multiple data sources (views) leading to significant improvements in predictive ability compared with inflexible modelling frameworks that are unable to exploit all relevant information sources. We demonstrate the inferred app patterns provide important and meaningful insights and are useful for summarization and exploration. Future work includes applying the developed modelling framework to alternative application scenarios, such as statistical analysis of customer purchase history of commercial products, relevant for understanding and predicting online commerce, and improving performance for several applications interconnecting image, video, text and speech data sources, for example.

Acknowledgements

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

References

Airoldi, EM, Blei, D, Erosheva, EA & Fienberg, SE (2014), *Handbook of Mixed Membership Models and Their Applications*, *CRC Press*, Boca Raton, Florida, USA.

Aldous, DJ (1985), *Exchangeability and Related Topics*, *Springer* , Berlin, Germany.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.103

Asuncion, A, Welling, M, Smyth, P & Teh, YW (2009), *On smoothing and inference for topic models,* Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, 27–34.

Barnard, K, Duygulu, P, Forsyth, D, De Freitas, N, Blei, DM & Jordan, MI (2003), 'Matching words and pictures', *The Journal of Machine Learning Research*, **3**, 1107–1135.

Bishop, CM (2006), *Pattern Recognition and Machine Learning*, *Springer* , New York, USA.

Blei, DM, Ng, AY & Jordan, MI (2003), 'Latent Dirichlet allocation', *The Journal of Machine Learning Research*, **3**, 993–1022.

Blei, D & Lafferty, J (2005), *Correlated topic models,* Neural Information Processing Systems, Vancouver, Canada, 147–154.

Byrd, RH, Lu, P, Nocedal, J & Zhu, C (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM Journal on Scientific Computing*, **16**(5), 1190–1208.

Griffiths, TL & Steyvers, M (2004), 'Finding scientific topics', *Proceedings of the National Academy of Sciences*, **101**(suppl 1), 5228–5235.

Linstead, E, Lopes, C & Baldi, P (2008), *An application of latent Dirichlet allocation to analyzing software evolution,* Machine Learning and Applications, San Diego, California, USA, pp. 813–818.

Mimno, D, Wallach, HM, Naradowsky, J, Smith, DA & McCallum, A (2009), *Polylingual topic models,* Empirical Methods in Natural Language Processing, Singapore, 880–889.

Mimno, DM, Hoffman, M & Blei, DM (2012), *Sparse stochastic inference for latent Dirichlet allocation,* International Conference on Machine Learning, Edinburgh, UK, 1599–1606.

Minka, T (2000). *Estimating a Dirichlet distribution*, Technical Report , MIT, Cambridge Massachusetts, USA.

Murphy, KP (2012), *Machine Learning: A Probabilistic Perspective*, *MIT Press* , Cambridge Massachusetts, USA.

Newman, D, Bonilla, EV & Buntine, W (2011), *Improving topic coherence with regularized topic models,* Neural Information Processing Systems, Granada, Spain, 496–504.

Paisley, J, Wang, C & Blei, DM (2012), 'The discrete infinite logistic normal distribution', *Bayesian Analysis*, **7**(2), 235–272.

Petterson, J, Buntine, W, Narayanamurthy, SM, Caetano, TS & Smola, AJ (2010), *Word features for latent Dirichlet allocation,* Neural Information Processing Systems, Vancouver, Canada, 1921–1929.

Pritchard, JK, Stephens, M & Donnelly, P (2000), 'Inference of population structure using multilocus genotype data', *Genetics*, **155**(2), 945–959.

Salomatin, K, Yang, Y & Lad, A (2009), *Multi-field correlated topic modeling,* Society for Industrial and Applied Mathematics (SIAM) International Conference on Data Mining, Sparks Nevada, USA, 628–637.

Sivic, J & Zisserman, A (2003), *Video Google: a text retrieval approach to object matching in videos,* International Conference on Computer Vision, Vol. 2, Nice, France, pp. 1470–1478.

Sivic, J, Russell, BC, Efros, AA, Zisserman, A & Freeman, WT (2005), *Discovering objects and their location in images,* International Conference on Computer Vision, Vol. 1, Beijing, China, pp. 370–377.

Van Der Maaten, L (2014), 'Accelerating t-SNE using tree-based algorithms', *The Journal of Machine Learning Research*, **15**(1), 3221–3245.

Virtanen, S, Jia, Y, Klami, A & Darrell, T (2012), *Factorized multi-modal topic model,* Uncertainty in Artificial Intelligence, Catalina Island, USA, 843–851.

Virtanen, S & Girolami, M (2015), *Ordinal mixed membership models,* International Conference on Machine Learning, Lille, France, 588–596.

Wainwright, MJ & Jordan, MI (2008), 'Graphical models, exponential families, and variational inference', *Foundations and Trends in Machine Learning*, **1**(1-2), 1–305.

Wallach, HM, Murray, I, Salakhutdinov, R & Mimno, D (2009), *Evaluation methods for topic models,* International Conference on Machine Learning, Montreal Quebec Canada, 1105–1112.