Learning Conditional Latent Structures from Multiple Data Sources

Viet Huynh[1(✉)], Dinh Phung[1], Long Nguyen[2], Svetha Venkatesh[1], and Hung H. Bui[3]

[1] Pattern Recognition and Data Analytics Centre, Deakin University, Geelong, Australia
hvhuynh@deakin.edu.au
[2] Department of Statistics, University of Michigan, Ann Arbor, USA
[3] Adobe Research, San Francisco Bay Area, San Francisco, USA

Abstract. Data usually present in heterogeneous sources. When dealing with multiple data sources, existing models often treat them independently and thus can not explicitly model the correlation structures among data sources. To address this problem, we propose a full Bayesian nonparametric approach to model correlation structures among multiple and heterogeneous datasets. The proposed framework, first, induces mixture distribution over primary data source using hierarchical Dirichlet processes (HDP). Once conditioned on each atom (group) discovered in previous step, *context* data sources are mutually independent and each is generated from hierarchical Dirichlet processes. In each specific application, which covariates constitute content or context(s) is determined by the nature of data. We also derive the efficient inference and exploit the conditional independence structure to propose (conditional) parallel Gibbs sampling scheme. We demonstrate our model to address the problem of latent activities discovery in pervasive computing using mobile data. We show the advantage of utilizing multiple data sources in terms of exploratory analysis as well as quantitative clustering performance.

1 Introduction

We are entering the age of big data. The challenges are that these data not only present in massive amount but also co-exist in heterogeneous forms including texts, hypertexts, images, graphics, videos, speeches and so forth. For example, in dealing with social network analysis, data present in network connection accompanying with users' profiles, their comments, activities. In medical data understanding, the patients' information usually co-exists with medical information such as diagnosis codes, demographics, laboratory tests. This deluge of data requires advanced algorithms for analyzing and making sense out of data. Machine learning provides a set of methods that can automatically discover low-dimensional structures in data which can be used for reasoning, making decision and predicting. Bayesian methods are increasingly popular in machine learning due to their resilience to over-fitting. Parametric models assume a finite number of parameters and this number needs to be fixed in advance, hence hinders its practicality. Bayesian nonparametrics, on the other hand, relax the assumption

T. Cao et al. (Eds.): PAKDD 2015, Part I, LNAI 9077, pp. 343–354, 2015.
DOI: 10.1007/978-3-319-18038-0_27

of parameter space to be infinite-dimensional, thus the model complexity, e.g., the number of mixture components, can grow with the data[1].

Two fundamental building blocks in Bayesian nonparametric models are the (hierarchical) Dirichlet processes [14] and Beta processes [15]. The former is usually used in clustering models, whereas the later is used in matrix factorization problems. Many extensions of them are developed to accommodate richer types of data [12,16]. However, when dealing with multiple covariates, these models often treat them independently, hence fail to explicitly model the correlation among data sources. The presence of rich and naturally correlated covariates calls for the need to model their correlation with nonparametric models.

In this paper, we aim to develop a full Bayesian nonparametric approach to the problem of multi-level and contextually related data sources and modelling their correlation. We use a stochastic process, being DP, to conditionally "index" other stochastic processes. The model can be viewed as a generalization of the hierarchical Dirichlet process (HDP) [14] and the nested Dirichlet process (nDP) [12]. In fact, it provides an interesting interpretation whereas, under a suitable parameterization, integrating out the topic components results in a nested DP, whereas integrating out the context components results in a hierarchical DP. For simplicity, correlated data channels are referred as two categories: *content* and *context(s)*. In each application, which the covariates constitute *content* or *context(s)* is determined by the nature of data. For instance, in pervasive computing application, we choose the *bluetooth co-location* of user as content while contexts are *time and location.*

Our main contributions in this paper include: (1) a Bayesian nonparametric approach to model multiple naturally correlated data channels in different areas of real-world applications such as pervasive computing, medical data mining, etc.; (2) a derivation of efficient parallel inference with Gibbs sampling for multiple contexts; (3) a novel application on understanding latent activities contextually dependent on time and place from mobile data in pervasive applications.

2 Background

A notable strand in both recent machine learning and statistics literature focuses on Bayesian nonparametric models of which Dirichlet process is the crux. Dirichlet process and its existence was established by Ferguson in a seminal paper in 1973 [4]. A Dirichlet process $\mathrm{DP}(\alpha, H)$ is a distribution of a random probability measure G over the measurable space $(\Theta, \mathcal{B})$ where H is a *base* probability measure and $\alpha > 0$ is the *concentration* parameter. It is defined such that, for any finite measurable partition $(A_k : k = 1, \ldots, K)$ of Θ, the resultant random vector $(G(A_1), \ldots, G(A_k))$ is distributed according to a Dirichlet distribution with parameters $(H(A_1), \ldots, H(A_k))$. In 1994, Sethuraman [13] provided an alternative constructive definition which makes the discreteness property of a Dirichlet process explicitly via a stick breaking construction. This is useful while dealing with infinite parametric space and defined as

[1] This characteristic is usually called "let the data speak for itself".

$$G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \text{ where } \phi_k \overset{\text{iid}}{\sim} H, k = 1, \ldots, \infty \text{ and } \boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}, \quad (1)$$

$$\beta_k = v_k \prod_{s<k} (1 - v_s) \text{ with } v_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha), \quad k = 1, \ldots, \infty.$$

It can be shown that $\sum_{k=1}^{\infty} \beta_k = 1$ with probability one, and as a convention in [11], we hereafter write $\boldsymbol{\beta} \sim \text{GEM}(\alpha)$. Due to its discreteness, Dirichlet processes is used as a prior for mixing proportion in Bayesian mixture models. Dirichlet processes mixture models (DPM) [1,7] which are nonparametric counterpart of well-known Gaussian mixture models (GMM)[2] with the relaxation of the number of components to be infinite were first introduced by Antoniak [1] and elaborated efficiently computational aspect by Neal [7].

However, in practice, data usually appear into collections which can be modelled together. From statistical perspective, it is interesting to extend the DP to accommodate these collections with dependent models. MacEachern [6] introduced framework that induces dependencies over these collections by using a stochastic process to couple them together. Following this framework, Nested Dirichlet process [12] induces dependency by using base measure as another Dirichlet process shared by collections which are modeled by Dirichlet process mixtures. Another widely used model driven by idea of MacEachern is hierarchical Dirichlet process [14] in which dependency is induced by sharing stick breaking representation of a Dirichlet process. All of these models are supposed to model single variable in data. In topic modeling, for instance, HDP is used as a nonparametric counterpart of Latent Dirichlet Allocation (LDA) to model word distributions over latent topics. In this application, the model ignores other co-existing variables such as time, authors.

When dealing with multiple covariates, one can treat the covariates as independent factors. With such independent assumption, he can not leverage the correlated nature of data. There are several works dealing with these situations. Recently, the work by Nguyen et. al. [8] tried to model secondary data channel (called *context*) attached with primary channel (*content*). In this model, secondary data channel is collected in group-level, e.g time or author for each document (consisting of words) or tags in each image. In the case of other data sets, observations are not at group-level but data point-level. For instance, in pervasive computing, each bluetooth co-location of each user includes several observations such as co-location, time stamp, location, etc. There is a motivation for modelling in these kind of applications. Dubey et. al. [2] tried to model topics over time where time are treated as context. The models can only handle one context while modelling but can not leverage the multiple correlated data channels. Another work by Wulsin et. al. [16] proposed the multi-level clustering hierarchical Dirichlet process (MLC-HDP) for clustering human seizures. In this model, authors assumed that data channels are clustered into multi-level which may not suitable for aforementioned data sets. In

[2] Indeed, DPM models are more general than (infinite) GMM since we can not only use Gaussian distribution but different kinds of distribution, e.g. Multinomial, Bernoulli, etc., to model each component.

consequence, there is the need for nonparametric models to handle naturally correlated data channels with certain dependent assumptions. In this paper, we propose a model that can model jointly the topic and the context distribution. Our method assumes a conditional dependence between two sets of stochastic process (content-context) which are coupled in a fashion similar to nested DP. The content models the primary observation with HDP and the dependent co-observations are modeled as nested DP with group index provided by the stochastic process from the content side. The set of DPs from the context side is further linked hierarchically in the similar fashion to HDP. Since our inference derivations rely on hierarchical Dirichlet processes, we briefly review hierarchical Dirichlet processes and some useful properties for inference. The justification for these properties can be found in [1,14, Proposition 3].

Let consider the case when we have a corpus with J documents. With the assumption that each document is related to several topics, we can model each document as a mixture of latent topics using Dirichlet process mixture. Though different documents may be generated from different topics, they usually share some of topics each others. Hierarchical Dirichlet process (HDP) models this topic sharing phenomenon. In HDP, the topics among documents are coupled using another Dirichlet process mixture G_0. For each document, a Dirichlet process $G_j,\ j = 1, \ldots, J$, is used to model its topic distribution. Formally, generative representation is as below:

$$\begin{aligned} G_0 \mid \gamma, H \sim DP(\gamma, H) \qquad & G_j \mid \alpha, G_0 \sim DP(\alpha, G_0) \\ \theta_{ji} \mid G_j \sim G_j \qquad & x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}). \end{aligned} \tag{2}$$

Similar to DPs, stick breaking representation of HDP is described as follows

$$\beta = \beta_{1:\infty} \sim GEM(\gamma) \qquad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \qquad \pi_j = \pi_{j1:j\infty} \sim DP(\alpha, \beta)$$

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \qquad z_{ij} \sim \pi_j \qquad \phi_k \sim H(\lambda) \quad x_{ji} \sim F(\phi_{z_{ji}}). \tag{3}$$

Given the HDP model as described in Equation (3) and $\theta_{j1}, \ldots \theta_{jN_j}$ be i.d.d samples from G_j for all $j = 1, \ldots, J$. All of these samples of each group G_j are grouped into M^j factors $\psi_{j1}, \ldots, \psi_{jM^j}$. These factors from all groups can be grouped into K sharing atoms $\phi_1, \ldots, \phi_K$. Then the posterior distributions stick breaking of G_0(denoted as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K, \beta_{new})$ is

$$(\beta_1, \ldots, \beta_K, \beta_{new}) \sim \mathrm{Dir}\,(m_1, \ldots, m_K, \gamma)\,, \tag{4}$$

where $m_k = \sum_{j=1}^{J} \sum_{i=1}^{M^j} \mathbf{1}\,(\psi_{ji} = \phi_k)$.

Another useful property for posterior of number of cluster K of a Dirichlet process is that if $G \sim DP(\alpha, H)$ and $\theta_1, \ldots, \theta_N$ be N i.i.d samples from G . These θ's values can be grouped into K clusters where $1 \leq K \leq N$. The conditional probability of K given α and N is

$$p(K = k \mid \alpha, N) = \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} s(N, k), \tag{5}$$

where $s(N, k)$ is the unsigned Stirling number of the first kind.

3 Framework

3.1 Context Sensitive Dirichlet Processes

Model Description: Suppose we have J documents in our corpus, and each has N_j words of which observed values are x_{ji}'s. From topic modeling perspective, there are a (specified or unspecified) number of topics among documents in corpus where each document may relate to several topics. We have an assumption that each of these topics is correlated with a number of realizations of context(s)[3] (e.g. time). To link the context with topic models we view context as distributions over some index spaces, governed by the topics discovered from the primary data source (content), and model both content and contexts jointly. We impose a conditional structure in which contents provide the topics, upon which contexts are conditionally distributed. Loosely speaking, we use a stochastic process to model content, being DP, and to conditionally "index" other stochastic processes which models contexts.

In details, we model the content side with a HDP, where x_{ji}'s are given in J groups. Each of group is modeled by a random probability distribution G_j, which shares a global random G_0 probability distribution. G_0 is draw from a DP with a base distribution H and concentration parameter γ. The distribution G_0 plays as a base distribution in a DP with concentration parameter α to construct G_j's for groups. The specification for this HDP is similar to Equation (2) in which the θ_{ji}'s are grouped into global atoms $\phi_k (k = 1, 2, \ldots)$.

For each observation x_{ji}, there is an associated context observation s_{ji} which is assumed to depend on the topic atom θ_{ji} of x_{ji}. Furthermore, the context observations of a given topic $S_k = \{s_{ji} \mid \theta_{ji} = \phi_k\}$ are assumed to be distributed a mixture Q_k. Given the number of topics K, there are the same number of context groups. Now to link these context groups, we again use the hierarchical structure that have the similar manner with HDP [14] where Q_k's share the global random probability distribution Q_0. Formally, generative specification for conditional independent context is as follows

$$\begin{aligned} Q_0 &\sim \mathrm{DP}\,(\eta, R) & Q_k &\sim \mathrm{DP}\,(\nu, Q_0) \\ \varphi_{ji} &\sim Q_k, \text{ s.t } \theta_{ji} = \phi_k & s_{ji} &\sim Y\,(\cdot \mid \varphi_{ji})\,. \end{aligned} \tag{6}$$

The stick breaking construction for content side is similar to the HDP, however, for the context size we have to take into account of the partition as induced by the content atoms. The stick breaking construction for context is

$$\begin{aligned} \epsilon &\sim \mathrm{GEM}\,(\eta) & \tau_k &\sim \mathrm{DP}\,(\nu, \epsilon) & \psi &\sim R \\ Q_0 &= \sum_{m=1}^{\infty} \epsilon_m \delta_{\psi_m} & Q_k &= \sum_{m=1}^{\infty} \tau_{km} \delta_{\psi_m} & l_{ji} &\sim \tau_{z_{ji}} & s_{ji} &\sim Y\left(\psi_{l_{ji}}\right) \end{aligned} \tag{7}$$

The graphical model for generative representation is depicted in Figure (1a).

Inference: we illustrate the auxiliary conditional approach using stick breaking scheme for inference. We briefly describe inference result of model. We also

[3] For simplicity, we will consider one context and generalize to multiple contexts.

assume conjugacy between F and H for content distributions as well as Y and R for context distributions since the conjugacy allows us to integrate out the atoms ϕ_k and τ_m. The sampling state space now consists of $\{\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{l}, \boldsymbol{\epsilon}\}$. Furthermore, we endow Gamma distributions as priors for hyperparameters $\{\gamma, \alpha, \eta, \nu\}$ and sample through each Gibbs iteration. During sampling iterations, we maintain the following counting variables: n_{jk} - the number of content observations in document j belong to content topic k, the marginal counts are denoted as $n_{j.} = \sum_k n_{jk}$, and $n_{.k} = \sum_j n_{jk}$; w_{km} - the number of context observations given the topic k belong to context m. The marginal counts are denoted similarly to n_{jk}. Sampling equations for *content side* are described below.

Sampling $\boldsymbol{z}$: the sampling of z_{ji} have to take into account of influence from the context apart from cluster assignment probability and likelihood.

$$p(z_{ji} = k \mid \boldsymbol{z}_{-ji}, \boldsymbol{l}, \boldsymbol{x}, \mathbf{s}) \propto p(z_{ji} = k \mid \boldsymbol{z}_{-ji}). \\ p(x_{ji} = k \mid z_{ji} = k, \boldsymbol{z}_{-ji}, \boldsymbol{x}_{-ji}) p(l_{ji} \mid z_{ji} = k, \boldsymbol{l}_{-ji}). \quad (8)$$

The first term of above equation in the RHS is the predictive likelihood of prior at the content side similar to HDP in [14] while the second term indicates the predictive likelihood of the observation for content topic k (except x_{ji}), denoted as $f_k^{-x_{ji}}(x_{ji})$. The last term is the context predictive likelihood given the content topic k. As a result, conditional sampling for z_{ji} is

$$p(z_{ji} = k \mid \boldsymbol{z}_{-ji}, \boldsymbol{l}, \boldsymbol{x}, \mathbf{s}) = \begin{cases} \left(n_{.k}^{-ji} + \alpha\beta_k\right) \frac{w_{km} + \nu\epsilon_m}{w_{k.} + \nu} f_k^{-x_{ji}}(x_{ji}) & \text{if } k \text{ previously used} \\ \alpha\beta_{new}\epsilon_m f_{new}^{-x_{ji}}(x_{ji}) & \text{if } k = k_{new} \end{cases}$$

Sampling $\boldsymbol{\beta}$: we use the posterior stick breaking of HDP in Equation (4).

In order to sample $\boldsymbol{m}$, we use the result from Equation (5), i.e. $m_{jk} \propto (\alpha\beta_k)^m s(n_{jk}, m)$ for $m = 1 \ldots n_{jk}$ where $s(n_{jk}, m)$ is the unsigned Stirling number of the first kind and compute $m_k = \sum_{j1}^{J} m_{jk}$.

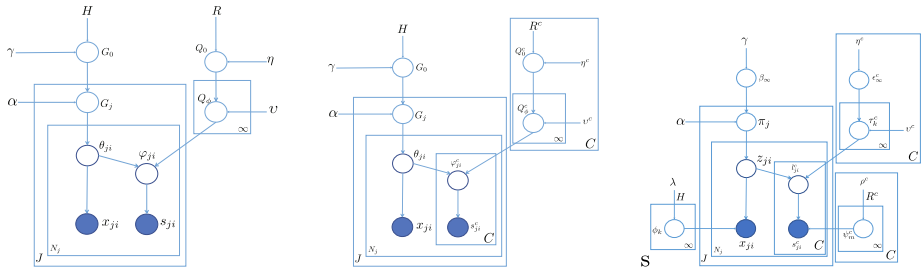
Next, we present sampling derivations for *context variables.*

Sampling $\boldsymbol{l}$: given the cluster assignment of content observations ($\boldsymbol{z}$), context observations are grouped into K groups of context. Let $\boldsymbol{s}_k$ be the set of context observations indexed by the same content cluster k. i.e. $\boldsymbol{s}_k \triangleq \{s_{ji} : z_{ji} = k, \forall j, i\}$, while $\boldsymbol{s}_k^{-ji}$ is the same set as $\boldsymbol{s}_k$ but excluding s_{ji}. The posterior probability of l_{ji} is computed as follows

$$p(l_{ji} = m \mid \boldsymbol{l}_{-ji}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{\nu}, \boldsymbol{\epsilon}) \propto p(l_{ji} = m \mid \boldsymbol{l}_{-ji}, z_{ji} = k, \boldsymbol{\epsilon}). \\ p(s_{ji} \mid l_{ji} = m, \boldsymbol{l}_{-ji}, z_{ji} = k, \boldsymbol{s}_{-ji}). \quad (9)$$

The first term is the conditional Chinese restaurant process given content cluster k while the second term, denoted as $y_{k,m}^{-s_{ji}}(s_{ji})$, is recognized to be a form of predictive likelihood in a standard Bayesian setting of which likelihood function is Y, conjugate prior S and a set of observation $\boldsymbol{s}_k^{-ji}(m) \triangleq \left\{s_{j'i'} : l_{j'i'} = m,\ z_{j'i'} = k, j' \neq j, i' \neq i\right\}$. The sampling equation for l_{ji} is

$$p(l_{ji} = m \mid \boldsymbol{l}_{-ji}, \boldsymbol{z}, \boldsymbol{s}, \boldsymbol{\nu}, \boldsymbol{\epsilon}) = \begin{cases} (w_{km} + \nu\epsilon_m) y_{k,m}^{-s_{ji}}(s_{ji}) & \text{if } m \text{previously used} \\ \epsilon_{\text{new}} y_{k,m_{\text{new}}}^{-s_{ji}}(s_{ji}) & \text{if } m = m_{\text{new}} \end{cases}$$



(a) Generative view for proposed model with single context

(b) Generative view for proposed model with C contexts

(c) The stick breaking view for the proposed model C contexts

Fig. 1. Graphical representation for the proposed model. (a) & (b) Generative view for single and multiple contexts which conditional independent given content topic. (c) Stick breaking view with C contexts, for single context, one can set $C = 1$.

Sampling $\boldsymbol{\epsilon}$: different from HDP, sampling $\boldsymbol{\epsilon}$ requires more works as it is dependent on both $\boldsymbol{z}$ and $\boldsymbol{l}$. Let isolate context variables l_{ji}^k's generated by the same topic $z_{ji} = k$ into one group $\boldsymbol{l}^k \triangleq \{l_{ji} : z_{ji} = k, \forall j, i\}$, context observations are also isolated in the similar way $\boldsymbol{s}^k \triangleq \{s_{ji} : z_{ji} = k, \forall j, i\}$. Now the context side is modeled with the structure similar to HDP in which the observations related Q_k are $\boldsymbol{s}^k$. We can sample $\boldsymbol{\epsilon}$ as follows $(\epsilon_1, \ldots, \epsilon_M, \epsilon_{new}) \sim \text{Dir}(h_{.1}, \ldots, h_{.M}, \eta)$ where $h_{.m}$, $m = 1 \ldots M$ are auxiliary variables which represent number of active context factors associated with atom m. Similar to sampling $\boldsymbol{m}$, the value of each $h_{.m}$ will be computed using samples $h_{km} \propto (\nu\epsilon_m)^h s(w_{km}, h)$ for $h = 1 \ldots w_{km}$ and summed up as $h_{.m} = \sum_{k=1}^{K} h_{km}$.

Moreover, there are four hyper-parameters in our model: α, γ, ν, η. Sampling α and γ is identical to HDP and therefore we refer to [14] for details. Sampling other hyperparameters is also doable, one can refer to [10] for details.

3.2 Context Sensitive Dirichlet Processes with Multiple Contexts

Model Description. When multiple contexts exist for a topic, the model can easily be extended to accommodate this. The generative and stick breaking specifications for content side remain the same as in Equation (2) and (3). The specification for multiple contexts will be duplicated from one context in Equation (6). Figure (1) depicts the graphical model for context sensitive Dirichlet process with multiple contexts. The generative model is

$$Q_0^c \sim \text{DP}(\eta^c, R^c) \qquad Q_k^c \sim \text{DP}(\nu^c, Q_0^c) \qquad \varphi_{ji}^c \sim Q_k^c, \text{ where } \theta_{ji} = \phi_k$$
$$x_{ji} \sim F(\cdot \mid \theta_{ji}) \qquad s_{ji}^c \sim Y^c(\cdot \mid \varphi_{ji}^c) \text{ for all } c = 1, \ldots, C.$$

The stick breaking construction for the context side is duplicated the specifications of context side in Equation (7) for C contexts which is provided below for all $c = 1, \ldots, C$:

Algorithm 1. Multiple Context CSDP Gibbs Sampler

procedure mCSDPGibbsSampler($\mathcal{D}$) ▷ $\mathcal{D}$: input including x_{ij} and s^c_{ij}
  repeat ▷ J: the number of groups
    for $j \leftarrow 1, J; i \leftarrow 1, N_j$ **do** ▷ N_j: the number of data in j-th group
      Sample z_{ji} using Equation (10) ▷ Sampling content side
      for $c \leftarrow 1, C$ **do** ▷ Sampling context side (can be parallised)
        Sample l^c_{ji} using Equation (9)
      end for
    end for
    Sample β and ϵ using Equation (4) and hyperparameters
  until Convergence
  return $z, l^{1:C}, \beta, \epsilon$ ▷ return learned parameters of model
end procedure

$$\epsilon^c \sim \text{GEM}\,(\eta) \qquad \tau^c_k \sim \text{DP}\,(\nu, \epsilon) \qquad \psi^c \sim R^c$$

$$Q^c_0 = \sum_{m=1}^{\infty} \epsilon^c_m \delta_{\psi^c_m} \qquad Q_k = \sum_{m=1}^{\infty} \tau^c_{km} \delta_{\psi^c_m} \qquad l^c_{ji} \sim \tau^c_{z_{ji}} \qquad s^c_{ji} \sim Y^c\left(\psi^c_{l^c_{ji}}\right).$$

Inference: using the same routing and assumptions on conjugacy of H and F, R^c and Y^c, we derive the sampling equations for variables as follows

Sampling $\boldsymbol{z}$: in multiple context setting, the sampling equation of z_{ji} involves the influence from multiple context rather than one:

$$p(z_{ji} = k \mid \boldsymbol{z}_{-ji}, \boldsymbol{l}, \boldsymbol{x}, \mathbf{s}) \propto p(z_{ji} = k \mid \boldsymbol{z}_{-ji}). \tag{10}$$

$$p(x_{ji} = k \mid z_{ji} = k, \boldsymbol{z}_{-ji}, \boldsymbol{x}_{-ji}) \prod_{c=1}^{C} p(l^c_{ji} = m^c \mid z_{ji} = k, \boldsymbol{l}^c_{-ji}).$$

It is straightforward to apply the result for one context case. The final sampling equation for z_{ji} is

$$p(z_{ji} = k \mid \boldsymbol{z}_{-ji}, \boldsymbol{l}, \boldsymbol{x}, \mathbf{s}) = \begin{cases} \left(n^{-ji}_{.k} + \alpha\beta_k\right) f^{-x_{ji}}_k\left(x_{ji}\right) \prod_{c=1}^{C} \frac{w^c_{km^c} + \nu^c \epsilon^c_{m^c}}{w^c_{k.} + \nu^c} & \text{if } k \text{ used} \\ \alpha\beta_{new} f^{-x_{ji}}_{k_{new}}\left(x_{ji}\right) \prod_{c=1}^{C} \epsilon^c_{m^c} & \text{if } k = k_{new}. \end{cases}$$

Sampling derivation of $\boldsymbol{\beta}$ is unchanged compared with one context.

Sampling equations of $\boldsymbol{l}^{1\ldots C}, \boldsymbol{\epsilon}^{1\ldots C}$ are similar to one context case where each set of context variables $\{\boldsymbol{l}^c, \boldsymbol{\epsilon}^c\}$ is dependent given sampled values of $\boldsymbol{z}$. We can perform sampling for each context in parallel thus the computation complexity in this case should remain the same as in the single context case given enough number of core processors to execute in parallel. We summarize sampling procedure for the model in Algorithm 1.

4 Experiments

In this section we demonstrate the application of our model to discover latent activities from social signals which is a challenging task in pervasive computing. We implemented model using C# and ran on Intel i7-3.4GHz machine with

installed Windows 7. We then used Reality Mining, a well-known data set collected at MIT Media Lab [3] to discover latent group activities. The model not only improves grouping performance but also reveals when and where these activities happened. In the following sections, we briefly describe data set, data preparation, parameter settings for the model and exploratory results as well as clustering performance using our proposed model.

4.1 Reality Mining Data Set

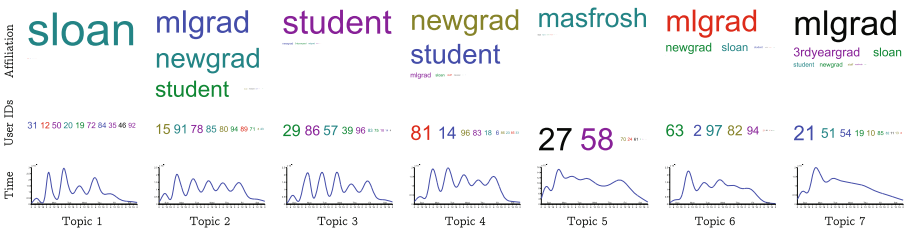
Reality Mining [3] is a well-known mobile data set collected by MIT Media Lab on 100 users over 9 months (approximately 450.000 hours). The collected information includes proximity using Bluetooth devices, cell tower IDs, call logs, application usage, and phone status. To illustrate the capability of proposed model, we extract proximity data recorded by Bluetooth devices and users' location via cell tower IDs. In order to compare with the results from [9], we preprocessed to filter users whose affiliations are missing or who do not share affiliation with others and then sampled proximity data for every 10 minutes. In the end we had 69 users. For each user, at every 10 minutes, we obtained a data point of 69-dimension which represents co-location information with other users. Each data point is an indicator binary vector of which i-th element set to 1 if the i-th user is co-located and 0 otherwise (self-presence set to 1). In addition, we also obtain the time stamp and cell ID data vectors. As a consequence, we have 69-user data groups. Each data point in group includes three observations: co-location vector, time stamp, cell tower ID.

4.2 Experimental Settings and Results

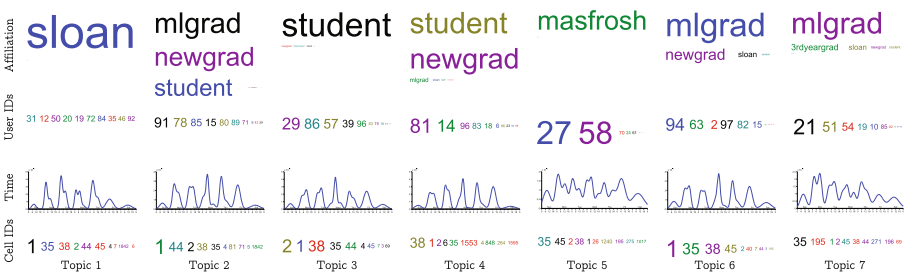
In proposed model, one data source will be chosen as content, the rest will be considered as contexts. We use two different settings in our experiment.

In the first setting, co-location data source is modelled as content which is (69-dimension) *Multinomial* distribution (corresponding to F distribution in model), time and cell tower IDs are modelled as *Gaussian* and *Multinomial* distributions respectively (corresponding to Y^1 and Y^2 distribution in model). We use the conjugate prior H as *Dirichlet* distribution, while R^1 and R^2 are *GaussianGamma* and *Dirichlet* distributions, respectively. We run the data set with 4 different settings for comparison: *HDP* - standard use of HDP on co-location observations (similar to [9]); *CSDP-50% time* - co-location and 50% time stamp data (supposing 50% missing) used for CSDP; *CSDP-time* - similar to *CSDP-50% time,* except that whole time stamp data are used; *CSDP-celltower* - resembling to *CSDP-time* but additional cell tower ID observations are used.

When modelling with *HDP* as in [9], the model merely discovered hidden activities of users. It fails to answer more refined questions such as *when and where these activities happened?* Our proposed model can naturally be used to model the additional data sources to address these questions. In Figure (2a), the topic 1 (*sloan* students) usually happened at specific time on Monday, Tuesday and Thursday while topic 5 (*master frosh* students) mainly gathered on Monday and Friday (less often on the other days). Similarly, when we modelled cell tower IDs data, the results



(a) Top 7 topics explored with *CSDP-time*



(b) Top 7 topics explored with *CSDP-celltower*

Fig. 2. Corresponding top 7 topics discovered by proposed model

revealed a deeper understanding on latent activities. In Figure (2b), we can observe the places (cell phone tower IDs)[4] where the activities took place. For topic 1 - *sloan* student group activities, apart from Sloan School building (*cell no.1* or *40*), they sometimes gathered at the restaurants (*cell no. 44*).

When using more contextual information, it does not only provide more exploratory information but also help the classification to be more discriminated. When using only time as context in Figure (2a), the user *no. 94* is (confusingly) recognized in both topic 2 and 6. But when location data is incorporated into our proposed model, the user *no. 94* is now dominantly classified into topic 6. To quantitatively evaluate proposed model when using more context data, we use the same setting with the work in [9]. First, we ran the data model to discover the latent activities among users. We then used the Affinity Propagation (AP) algorithm [5] to perform clustering among users with similar activities. We evaluated clustering performance using popular metrics: F-measure, cluster purity, rand index (RI) and normalized mutual information (NMI). As it can be clearly seen in Table 1, with more contexts we observed, CSDP achieves better clustering results. *Purity* and *NMI* are significantly improved when more contextual data are observed while other metrics slightly improved when modelling with contextual data.

[4] Since Reality Mining does not provide exact information about these cell towers however we can infer information about some of them by using users' descriptions. For example, cell no.1 and 40 are MIT Lab and Sloan School of Management which are two adjacent buildings. While cell no. 35 is located near Student Center and cell no. 44 is around some restaurants outside MIT campus.

Table 1. Clustering performance improved when more contextual data used in the proposed model

	Purity	NMI	RI	F-measure
HDP	0.7101	0.6467	0.9109	0.7429
CSDP-*50% time*	0.7391	0.6749	**0.9186**	**0.7651**
CSDP-*100% time*	0.7536	0.6798	0.9169	0.7503
CSDP-*celltower*	**0.7826**	**0.6953**	**0.9186**	0.7567

In the second setting, we model time as content and the rest (co-locations, cell towers) as contexts. The conjugate pairs are remained the same in previous setting. In Figure (3), we demonstrate top 4 time topics including Friday, Thursday (upper row), Tuesday, and Monday (lower row) which are Gaussian forms. The groups of users who gathered in that time stamp are depicted under each Gaussian. It is easy to notice that the group with user *27, 58* usually gathered on Friday and Monday whereas other groups met on all four time slots.

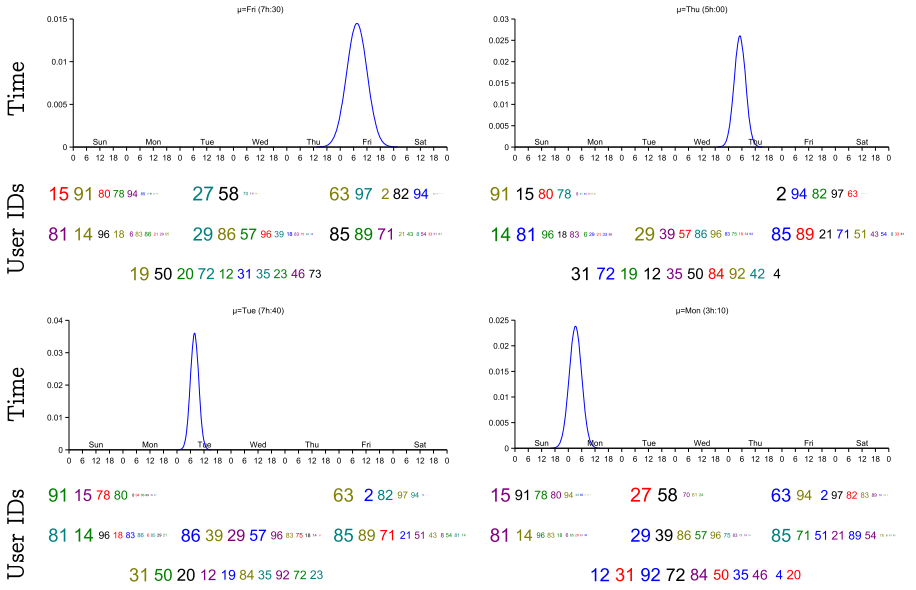


Fig. 3. Top 4 time topics and their corresponding conditional user-IDs groups discovered by proposed model

5 Conclusions

We propose a full Bayesian nonparametric approach to model explicit correlation structures in heterogeneous data sources. Our key contribution is the development of a context sensitive Dirichlet processes, its Gibbs inference and its parallelability. We have further demonstrated the proposed model to discover latent activities from mobile data to answer who (co-location), when (time) and

where (cell-tower ID) – a central problem in context-aware computing applications. With its expressiveness, our model not only discovers latent activities (topics) of users but also reveals time and place information. Qualitatively, it was shown that better clustering performance than without them. Finally, although the building block of our proposed model is the Dirichlet process, based on HDP, it is straightforward to apply other stochastic processes such as nested Dirichlet processes or hierarchical Beta processes to provide alternative representation expressiveness for data modelling tasks.

References

1. Antoniak, C.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics **2**(6), 1152–1174 (1974)
2. Dubey, A., Hefny, A., Williamson, S., Xing, E.P.: A non-parametric mixture model for topic modeling over time (2012). arXiv preprint arXiv:1208.4411
3. Eagle, N., Pentland, A.: Reality mining: Sensing complex social systems. Personal and Ubiquitous Computing **10**(4), 255–268 (2006)
4. Ferguson, T.: A Bayesian analysis of some nonparametric problems. The Annals of Statistics **1**(2), 209–230 (1973)
5. Frey, B., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
6. MacEachern, S.: Dependent nonparametric processes. In: ASA Proceedings of the Section on Bayesian Statistical Science. pp. 50–55 (1999)
7. Neal, R.: Markov chain sampling methods for Dirichlet process mixture models. Journal of computational and graphical statistics **9**(2), 249–265 (2000)
8. Nguyen, V., Phung, D., Venkatesh, S. Nguyen, X., Bui, H.: Bayesian nonparametric multilevel clustering with group-level contexts. In: Proc. of International Conference on Machine Learning (ICML), pp. 288–296. Beijing, China (2014)
9. Phung, D., Nguyen, T.C., Gupta, S., Venkatesh, S.: Learning latent activities from social signals with hierarchical Dirichlet process. In: Sukthankar, G., et al. (ed.) Handbook on Plan, Activity, and Intent Recognition, pp. 149–174. Elsevier (2014)
10. Phung, D., Nguyen, X., Bui, H., Nguyen, T., Venkatesh, S.: Conditionally dependent Dirichlet processes for modelling naturally correlated data sources. Tech. rep., Pattern Recognition and Data Analytics, Deakin University(2012)
11. Pitman, J.: Poisson-Dirichlet and GEM invariant distributions for split-and-merge transformations of an interval partition. Combinatorics, Probability and Computing **11**(05), 501–514 (2002)
12. Rodriguez, A., Dunson, D., Gelfand, A.: The nested Dirichlet process. Journal of the American Statistical Association **103**(483), 1131–1154 (2008)
13. Sethuraman, J.: A constructive definition of Dirichlet priors. Statistica Sinica **4**(2), 639–650 (1994)
14. Teh, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. Journal of the American Statistical Association **101**(476), 1566–1581 (2006)
15. Thibaux, R., Jordan, M.: Hierarchical Beta processes and the Indian buffet process. In: Proc. of Int. Conf. on Artificial Intelligence and Statistics (AISTAT), vol. 11, pp. 564–571 (2007)
16. Wulsin, D., Jensen, S., Litt, B.: A hierarchical dirichlet process model with multiple levels of clustering for human eeg seizure modeling. In: Proc. of International Conference on Machine Learning (ICML) (2012)