Scalable Nonparametric Bayesian Multilevel Clustering

Viet Huynh [†] **Dinh Phung** [†] **Svetha Venkatesh** [†]
[†] Center for Pattern Recognition and Data Analytics (PRaDA)
Deakin University, Australia

XuanLong Nguyen
Department of Statistics
University of Michigan, Ann Arbor, USA

Matt Hoffman
Adobe Research
Adobe Systems, Inc.

Hung Hai Bui
Adobe Research
Adobe Systems, Inc.

Abstract

Multilevel clustering problems where the content and contextual information are jointly clustered are ubiquitous in modern datasets. Existing works on this problem are limited to small datasets due to the use of the Gibbs sampler. We address the problem of scaling up multilevel clustering under a Bayesian nonparametric setting, extending the MC2 model proposed in (Nguyen *et al.*, 2014). We ground our approach in structured mean-field and stochastic variational inference (SVI) and develop a tree-structured SVI algorithm that exploits the interplay between content and context modeling. Our new algorithm avoids the need to repeatedly go through the corpus as in Gibbs sampler. More crucially, our method is immediately amendable to parallelization, facilitating a scalable distributed implementation on the Apache Spark platform. We conduct extensive experiments in a variety of domains including text, images, and real-world user application activities. Direct comparison with the Gibbs-sampler demonstrates that our method is an order-of-magnitude faster without loss of model quality. Our Spark-based implementation gains another order-of-magnitude speedup and can scale to large real-world datasets containing millions of documents and groups.

1 INTRODUCTION

A prominent feature in numerous modern datasets tackled in machine learning is how the data are naturally layered into groups in a hierarchical representation: text corpus as collection of documents, which are subdivided into words, user's activities are organized by users, whose sessions divided into actions, electronic medical records (EMR) organized as sets of ICD[1] codes diagnosed for the patient. Probabilistic modeling techniques for grouped data have become a standard tool in machine learning, including topic modeling (Blei *et al.*, 2003; Teh *et al.*, 2006) and multilevel data analysis (Hox, 2010; Diez-Roux, 2000). Another important feature in such datasets is the availability of rich sources of additional information known as contexts and group-specific meta-data (Phung *et al.*, 2012; Nguyen *et al.*, 2014). These include information about authorships, timestamps, various tags associated with texts and images, user's demographics, etc. For consistency, we shall refer to the content groups (e.g., text documents, images, user's activity session) broadly as *documents*, and its associated context as document-specific *context*.

The rich and interwoven nature of raw document contents and their contextual information provides an excellent opportunity for joint modeling and, in particular, clustering the content-units (e.g., forming topics from words) and the content-groups (e.g., forming cluster of documents) — a problem known as *multilevel clustering with context* (Nguyen *et al.*, 2014). There have been several attempts of multilevel clustering in the probabilistic topic modeling literature. A simple approach is to subdivide this task into two phases: first learn a topic model and then perform document clustering using the topic-induced representation of the documents (Lu *et al.*, 2011; Nguyen *et al.*, 2013; Phung *et al.*, 2014). An elegant approach is to unify these two steps into a single framework (Nguyen *et al.*, 2014; Xie & Xing, 2013; Rodriguez *et al.*, 2008; Wulsin *et al.*, 2012). Among these work, the Bayesian nonparametric approach to multilevel clustering with group-level contexts (MC2) (Nguyen *et al.*, 2014) offers a powerful method capable of jointly modeling both content and context in a flexible and nonparametric manner, generalizing on several previous modeling techniques. The key idea of the MC2 model is a special Dirichlet Process (DP) whose base-measure is a product between a context-generating measure and a content-generating DP. This construct enables both clustering of documents associating with their

[1] Stands for International Classification of Disease.

contexts and clustering of words into topics. Nguyen *et al.* (2014) have shown that their MC2 integrates the nested DP (Rodriguez *et al.*, 2008) and DP mixture (DPM) (Antoniak, 1974) into one single unified model wherein marginalizing out the documents' contents results in a DP mixture, and marginalizing out document-specific contexts results in a nested DP mixture.

The need for jointly accounting for both context and content data in a flexible Bayesian nonparametric fashion also underlies a formidable computational challenge for model fitting. In fact, the MC2 model of Nguyen *et al.* (2014) was originally equipped with a Gibbs sampler for inference; hence the usefulness of the model could only be demonstrated on relatively small datasets. This seriously hinders the usefulness and applicability of MC2 in tackling big real world datasets which can contain millions of documents or more, along with it the millions of useful pieces of contextual information.

Our goal in this work is to address the multilevel clustering with contexts problem at scale, by developing effective posterior inference algorithms for the MC2 using techniques from stochastic variational inference. A challenging aspect about inference for MC2 is the computational treatment in the clustering of discrete distributions of contents jointly with the context variables. Unlike either the Dirichlet process or HDP mixtures, the context-content linkage present in the MC2 model makes the model more expressive, while necessitating the inference of the joint context and content atoms. These are mathematically rich objects — while the context atoms take on usual contextual values, the content atoms represent probability distributions over words. To maintain an accurate approximation of the joint context and content atoms, we employ a *tree-structured mean-field* decomposition that explicitly links the model context and content atoms.

The result is a scalable stochastic variational inference (SVI) algorithm for MC2 (SVI-MC2) that, unlike Gibbs sampling, avoids the need to go through the corpus multiple times. Moreover, the SVI computation within each mini-batch can be easily parallelizable. To fully exploit this advantage of the SVI formulation, we further implement our proposed SVI for the MC2 algorithm on the Apache Spark platform. We demonstrate that even a sequential implementation of SVI-MC2 is several times faster than Nguyen *et al.* (2014)'s Gibbs sampler while yielding the same model perplexity. A parallel implementation can gain another order of magnitude improvement in speed; our Spark implementation can simultaneously find topics and clustering millions of documents and their context. Our contributions then can be summarized as: (a) a new theoretical development of stochastic variational inference for an important family of models to address the problem of multilevel clustering with contexts. We note this class of models (MC2) include nested DP, DPM, and HDP as the special cases; (b) a scalable implementation of the proposed SVI-MC2 on Apache Spark; and (c) the demonstration that our new algorithm can scale up to very large corpora.

2 RELATED WORK

Models for clustering documents

Two of the most well-known probabilistic models for learning from grouped data are Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) and its nonparametric counterpart, Hierarchical Dirichlet process (HDP) (Teh *et al.*, 2006). These models allow us to exploit the group structure for word clustering but not to cluster the groups of data. To clustering documents, some authors employed a two-step process. In the first step, each document is represented by the feature of its topic proportion using topic models, e.g. LDA or HDP. Now each document is considered as an input data point for some clustering algorithm. Elango & Jayaraman (2005) used LDA combined with K-means to cluster images while Nguyen *et al.* (2013) exploited features by HDP and used Affinity Propagation for clustering human activities.

Incorporating topic modeling and clustering in one unique model is a more elegant approach. Nested DP (nDP) (Rodriguez *et al.*, 2008) is the first attempt to handle this challenge in the context of Bayesian nonparametric. The model by Rodriguez *et al.* (2008) has tried to group documents into clusters each of which shares the same distribution over the topics. However, in the original nDP, the documents do not share topics. An extension to nDP, the MLC-HDP model with 3-level clustering, has been done by Wulsin *et al.* (2012). This model can cluster words, documents and document-corpora with shared topic atoms throughout the group hierarchy with this model. Later, Multi-Grain Clustering Topic Model which allows mixing between global topics and document-cluster topics has been introduced by Xie & Xing (2013). The most recent work, the Bayesian nonparametric multilevel clustering with group-level contexts (MC2) (Nguyen *et al.*, 2014), offers a theoretically elegant joint model for both content and context. To our best of knowledge, this model is the current state-of-the-art for this problem.

However, authors in (Nguyen *et al.*, 2014) only provide a Gibbs sampling method for inference. This seriously hinder the usefulness and applicability of MC2 in tackling modern datasets which can contain millions of documents.

Stochastic Variational Inference

Between two main inference approaches for graphical model including MCMC and deterministic variational methods, variational inference is usually preferred due to its predictable convergence. In variational inference scheme, the problem of computing intractable posterior distribution is transformed into an optimization problem by

introducing tractable variational distribution. One of the most popular approximation is mean-field which assumes that the variational distribution is fully factorized. The objective function called Evidence Lower BOund (ELBO) is defined as KL divergence between approximated distribution and the posterior distribution plus a constant. To solve this optimization problem, coordinate descent can be used. However, this optimization method is not suitable for modern datasets with millions of documents since all document are visited in each iteration. To circumvent this challenge, the earliest, but very recent, attempt can be traced back to (Hoffman *et al.*, 2010) where SVI framework for Bayesian nonparametric inference was proposed by combining mean-field approximation and stochastic optimization. SVI for the hierarchical Dirichlet process (HDP) was also presented in (Wang *et al.*, 2011).

Instead of using coordinate descent, stochastic variational inference (SVI) (Hoffman *et al.*, 2013) using stochastic gradient descent to optimize the ELBO. In order to keep optimization process converge faster, SVI uses the coordinate descent for the local update which is related to each data point and update global variables involving multiple data points with stochastic gradient. Moreover, as suggested by Amari (1998), learning with natural gradient may lead to faster convergence. In the SVI framework with exponential family distributions, the natural gradient updates are not only more likely to improve optimization speed but also produces simpler update equations.

We ground our methodology on (Hoffman *et al.*, 2013) and develop the SVI updates for MC2. However, we note at the outset that, unlike HDP, our model is not completely factorized, hence our solution does not simply follow a naive mean field, but rather a variant of structured mean field approximation of Bayesian nonparametric models.

3 MULTILEVEL CLUSTERING WITH CONTEXTS (MC2)

We first describe the MC2 model of (Nguyen *et al.*, 2014). The generative process for MC2 model (see Fig. 1a) is as follows

$$U \sim \text{DP}\left(\gamma\left(H \times \text{DP}\left(\upsilon Q_0\right)\right)\right) \text{ where } Q_0 \sim \text{DP}\left(\eta S\right),$$
$$(\theta_j, Q_j) \sim U \text{ for each group } j$$
$$x_j \sim F\left(\cdot \mid \theta_j\right), \quad \varphi_{ji} \sim Q_j, \quad w_{ji} \sim Y\left(\cdot \mid \varphi_{ji}\right).$$

In the above, U is a DP realization, hence a discrete measure with probability 1, and therefore enforces the clustering of documents. The sample pair $(\theta_j, Q_j) \sim U$ represents the context parameter and content-generating measures of the j-th document. Distinct measures Q_j are effectively drawn from $\text{DP}(\upsilon Q_0)$ where $Q_0 \sim \text{DP}\left(\eta S\right)$, so the samples φ_{ji} share atoms just like in a hierarchical DP (HDP). $F(.|\theta_j)$ and $Y(.|\varphi_{ji})$ are the likelihoods for context and content with parameters θ_j and φ_{ji}. Their base-measures H and S are assumed to be conjugate with the respective likelihoods.

The stick-breaking representation for the MC2 model is given Fig. 1b. When integrated out the random stick length, the model has an intuitive Polya-Urn view known as the Chinese Restaurant Franchise Bus (CRF-Bus) (Nguyen *et al.*, 2014). Each word in a document is viewed as a customer in a bus. The buses deliver customers randomly to a set of restaurants following a Chinese Restaurant Process (CRP). After getting off the buses, the customers in the restaurants behave as in the HDP - Chinese Restaurant Franchise (CRF). The MC2 model thus inherits the metaphor of tables at restaurants and global dishes from the CRF. The detailed stick-breaking representations are

- Stick length for *content* generation $\epsilon = \{\epsilon_m\}_{m=1}^{\infty}$ and *content* shared atoms $\{\psi_m\}_{m=1}^{\infty}$

$$\epsilon \sim \text{GEM}\left(1, \gamma\right), \qquad \psi_m \sim S, \qquad Q_0 = \sum_{m=1}^{\infty} \epsilon_m \delta_{\psi_m}.$$

- Stick length for *context* generation $\beta = \{\beta_k\}_{k=1}^{\infty}$ and *context* shared atoms $\{\phi_k\}_{k=1}^{\infty}$

$$\beta \sim \text{GEM}\left(1, \eta\right), \qquad \phi_k \sim H, \qquad G = \sum_{t=1}^{\infty} \beta_k \delta_{\phi_k}.$$

- Choosing document group (restaurant) for document $j = 1, \ldots, J$ and generating *context observation*

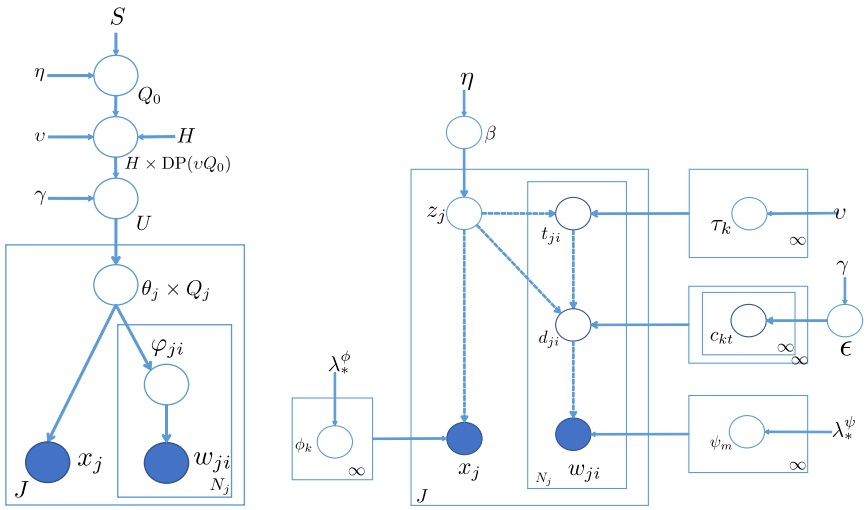
$$z_j \sim \text{Cat}\left(\beta_{1:\infty}\right), \qquad x_j \sim F\left(\cdot \mid \phi_{z_i}\right).$$

- Stick length for each document group $k = 1, \ldots, \infty$, $\{\tau_{kt}\}_{t=1}^{\infty}$, choosing tables t, dishes c and generating *content word,* $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$

$$\tau_k \sim \text{GEM}\left(1, \upsilon\right), \qquad t_{ji} \sim \text{Cat}\left(\tau_{z_j}\right),$$
$$c_{kt} \sim \text{Cat}\left(\epsilon\right), \qquad w_{ji} \sim Y\left(\cdot \mid \psi_{c_{z_j t_{ji}}}\right).$$

We consider general exponential family forms for the likelihoods[2] $Y\left(w \mid \psi\right) = \exp\left(\left\langle T\left(w\right), \psi\right\rangle - A\left(\psi\right)\right)$ and $F\left(x \mid \phi\right) = \exp\left(\left\langle T\left(x\right), \psi\right\rangle - A\left(\phi\right)\right)$. The prior $S\left(\psi \mid \cdot\right)$ and $H\left(\phi \mid \cdot\right)$ have the conjugate forms $p\left(\psi \mid \lambda_*^{\psi}\right) \propto \exp\left(\left\langle \lambda_*^{\psi}, [\psi; -A\left(\psi\right)]\right\rangle\right)$ and $p\left(\phi \mid \lambda_*^{\phi}\right) \propto \exp\left(\left\langle \lambda_*^{\phi}, [\phi; -A\left(\phi\right)]\right\rangle\right)$. The notation $[v; c]$ represents the stacking of two column vectors.

[2]Note that $T\left(w\right)$ and $T\left(x\right)$ may have different forms.



(a) Generative view. (b) Stick-breaking view.

Figure 1: Graphical presentation for Multilevel clustering with contexts models.

4 SVI FOR MC2

4.1 TRUNCATED STICK-BREAKING REPRESENTATIONS

The approximation of DP by truncated stick-breaking representation has been introduced by (Ishwaran & James, 2001) and later used by (Blei & Jordan, 2006) for variational inference in DP mixtures model. In this work, we also use the truncated stick-breaking approximation for all three stick-breaking length variables of the model which are β, ϵ, and τ. As pointed by Ishwaran & James (2001), the truncated stick-breaking is equivalent to the generalized Dirichlet distribution (Connor & Mosiman, 1969; Wong, 1998) which is a distribution on $K-1$ simplex with $2(K-1)$-parameter $\lambda = \left(\lambda_{11}, \ldots \lambda_{(K-1)1},, \lambda_{12}, \lambda_{(K-1)2}\right)$. Each pair of parameters $(\lambda_{k1}, \lambda_{k2})$ corresponds to the parameters for a Beta distribution in stick-breaking process. Generalized Dirichlet (GD) distribution is a member of the exponential family and is conjugate to Multinomial distributions (for more details, see the Appendix). For this reason, the mean-field update of a GD-distributed stick length also has a GD form. We used this conjugacy to compute the variational updates for stick-breaking variables.

4.2 MEAN-FIELD VARIATIONAL APPROXIMATION

The objective of inference problem with the model is to estimate the posterior distribution $p(\Theta \mid x, w)$ where Θ is the collection of parameter variable of the model, $\Theta \triangleq \{\beta, \epsilon, \tau, c, z, t, \psi, \phi\}$. In variational Bayes inference, this intractable posterior will be approximated with a tractable distribution called variational distribution, $q(\Theta)$. In order to ensure that $q(\Theta)$ is tractable, one usually uses mean-field assumption which assumes that all variational variables in Θ are independent. However, because of the nature of the MC2 model, two group of variables z_i (restaurant) and $t_{j1}, \ldots, t_{jn_j}$(tables) are highly correlated. We will maintain the joint distribution of these variables in as a collection of tree-structure graphical model. Thus, the variational distribution q is factorized as

$$q(\Theta) = q(\beta)\, q(\epsilon)\, q(\tau)\, q(c)\, q(z,t)\, q(\psi)\, q(\phi)\,.$$

All the factorized q's have exponential family form and for convenience we shall use either the natural or mean parameterization when appropriate. We use the following convention in naming the variational parameters: λ denotes a natural parameter, μ denotes a mean parameter, superscript denotes the collection of random variables of being parameterized and subscript denotes the index of variables. For instance, under this convention, λ^{ϕ}_{k} is the natural parameter for $q(\phi_k)$. The actual parameterization of q's are

- For the group of stick-breaking variables $q(\beta) = \mathrm{GD}\left(\beta \mid \lambda^{\beta}\right)$, $q(\epsilon) = \mathrm{GD}\left(\epsilon \mid \lambda^{\epsilon}\right)$, and $q(\tau) = \prod_{k=1}^{K} \mathrm{GD}\left(\tau_k \mid \lambda^{\tau}_{k}\right)$ where λ^{β}, λ^{ϵ}, and λ^{τ}_{k} are $2K-2$, $2M-2$, and $2T-2$ dimension vector, respectively. K, M and T are the truncated levels for restaurants, dishes and tables in the CRF-Bus process respectively.
- For the group of content and context atoms $q(\psi) = \prod_{m=1}^{M} q\left(\psi_m \mid \lambda^{\psi}_{m}\right)$ and $q(\phi) = \prod_{k=1}^{K} q\left(\phi_k \mid \lambda^{\phi}_{k}\right)$.
- For the group of indicator variables $q(c) = \prod_{k=1}^{K}\prod_{t=1}^{T} \mathrm{Mult}\left(c_{kt} \mid \mu^{c}_{kt}\right)$ and $q(z,t) = \prod_{j}\left[\mathrm{Mult}\left(z_j \mid \mu^{z}_{j}\right)\prod_{i=1}^{n_j}\mathrm{Mult}\left(t_{ji} \mid \mu^{t}_{jiz_j}\right)\right]$ where μ^{z}_{j}, μ^{c}_{kt}, and μ^{t}_{jik} are K, M, T-dimension vectors,

correspondingly. Note that two groups of variables z and t are not fully factorized but form a forest of trees, with each tree rooted at z_j.

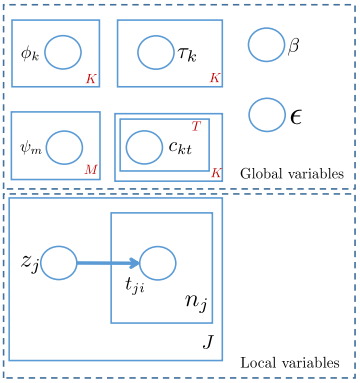


Figure 2: Variational factorization and global vs. local variables for SVI.

4.3 MEAN-FIELD UPDATES

All the individual q's in our model are in the exponential family and are locally conjugate. Thus, standard naive mean-field updates (Bishop *et al.*, 2006; Blei & Jordan, 2006), can be derived for all the variational parameters in a straight-forward manner. We provide more details on the update for the variational parameters of z and t since these are coupled and structured mean-field is needed (Wainwright & Jordan, 2008). At a high-level, for each tree rooted at z_j, exact inference needs to be done to convert from natural to mean parameters. The actual updates equation for these parameters are

$$\mu^t_{jikl} \propto \tilde{\mu}^t_{jikl}, \tag{1}$$

$$\mu^z_{jk} \propto \exp(\mathbb{E}\left[\ln \beta_k p(x_j \mid \phi_k)\right] + \sum_i \ln(\sum_{l=1}^{T} \tilde{\mu}^t_{jikl})),$$

where $\tilde{\mu}^t_{jikl}$ is the unnormalized value of μ^t_{jikl} and is $\exp\left(\sum_{m=1}^{M} \mu^c_{klm} \mathbb{E}\left[\ln p\left(w_{ji} \mid \psi_m\right)\right] + \mathbb{E}\left[\ln \tau_{kl}\right]\right)$.

The update for the rest of the parameters uses naive mean-field.

Two groups of variables, stick-breaking and atoms, contain similar variables. One variable in each group will be presented, the others have a similar forms and can be found in the appendix. The following equations includes updates for the content side of the stick-breaking and atom variables.

For the stick-breaking variational distribution $q(\epsilon)$

$$\lambda^\epsilon_{m1} = 1 + \sum_{k,t} \mu^c_{ktm} \quad \lambda^\epsilon_{m2} = \gamma + \sum_{k,t} \sum_{l=m+1}^{M} \mu^c_{ktl}. \tag{2}$$

For the content-atom variational distribution $q(\psi)$

$$\lambda^\psi_m = \lambda^\psi_* + \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(\sum_{k=1}^{K} \mu^z_{jk} \sum_{l=1}^{T} \mu^c_{ktm} \mu^t_{jikl} \right) \left[T\left(w_{ji}\right); 1\right].$$

4.4 STOCHASTIC VARIATIONAL INFERENCE

We follow the SVI framework (Hoffman *et al.*, 2013) and divide the set of variables Θ in the posterior into the set of *local* variables $\{z, t\}$ with the rest as *global* variables (see Fig. 2). The variational Evidence Lower BOund (ELBO) function is

where $\Theta^g \triangleq \Theta \backslash \{z, t\}$ is the global parameters of the model.

We will reuse the coordinate descent updates for local variational parameters μ^z_j and μ^t_{jik} given in section 4.2. To derive the stochastic gradient descent update for the global parameters, instead of taking the gradient of $\mathcal{L}$ which would result in messages being passed from all the documents, we take the gradient of $\mathcal{L}_j$ which is sufficient to yield a stochastic gradient of $\mathcal{L}$. The gradients are multiplied with the inverse Fisher information matrix to obtain the natural gradients (denoted as $\frac{\partial^{(\text{ng})}}{\partial}$). The gradient with respect to the content atom and stick breaking variational parameters λ^ψ_m and $\lambda^\epsilon_{m1,2}$ are

$$\frac{\partial^{(\text{ng})} \mathcal{L}_j}{\partial \lambda^\psi_m} = \frac{-\lambda^\psi_m + \lambda^\psi_*}{J} + \sum_{i=1}^{n_j} (\sum_{k,l} \mu^z_{jk} \mu^c_{klm} \mu^t_{jikl}) \left[T\left\langle w_{ji} \right\rangle; 1\right]. \tag{3}$$

$$\frac{\partial^{(\text{ng})} \mathcal{L}_j}{\partial \lambda^\epsilon_{m1}} = \frac{-\lambda^\epsilon_{m1} + 1}{J} + \sum_{k,t} \mu^c_{ktm}, \tag{4}$$

$$\frac{\partial^{(\text{ng})} \mathcal{L}_j}{\partial \lambda^\epsilon_{m2}} = \frac{-\lambda^\epsilon_{m2} + \gamma}{J} + \sum_{k,t} \sum_{r=m+1}^{M} \mu^c_{ktr}$$

Computing the gradient w.r.t. $q(c_{kt})$ is easier using the minimal natural parameterization of the multinomial. Let λ^c_{kt} be the minimal natural parameter corresponding to the mean parameter μ^c_{kt}, the gradient w.r.t λ^c_{kt} is

$$\frac{\partial^{(\text{ng})} \mathcal{L}_j}{\partial \lambda^c_{ktm}} = \frac{-\lambda^c_{ktm} + \mathbb{E}\left[\ln \frac{\epsilon_m}{\epsilon_M}\right]}{J} + (a_{ktm} - a_{ktM}) \tag{5}$$

where $a_{ktm} = \mu^z_{jk} \sum_{i=1}^{n_j} \mu^t_{jikl} \mathbb{E}\left[\ln p\left(w_{ji} \mid \psi_m\right)\right]$ for $m = 1 \ldots M$. Conversion from natural to mean parameters for the multinomials are standard

$$\mu^c_{ktm} = \frac{\exp\left(\lambda^c_{ktm}\right)}{1 + \sum_{m=1}^{M-1} \exp\left(\lambda^c_{ktm}\right)}, m = 1, \ldots, M-1$$

and $\mu^c_{ktM} = 1 - \sum_{m=1}^{M-1} \mu_{ktm}$.

With above derivations, we can summarize the procedure of stochastic variational inference for MC2 model in Algorithm 1.

In the above, the stochastic gradient is obtained for each document. In practice, mini-batch of documents are used

Algorithm 1 Stochastic variational inference for MC2

Require: forgetting rate ι and delay ϱ
Initialize $\lambda_m^{\psi(0)}, \lambda_k^{\phi(0)}$ and set $t = 1$;
repeat
Choose uniformly document j from data
Compute μ_{jik}^t and μ_j^z with Eq. (1)
Set $\varpi_t = (t + \varrho)^{-\iota}$
Update stick-breaking variable hyperparameters $\lambda^\beta, \lambda^\epsilon, \lambda_k^\tau$ using corresponding gradient similar to Eq. (4) as follows $\lambda^{(t+1)} = \lambda^{(t)} + J\varpi_t \frac{\partial^{(\text{ng})} \mathcal{L}_j}{\partial \lambda}$
Update content and context atom hyperparameters $\lambda^\psi, \lambda^\phi$ using corresponding gradient similar to Eq. (3) as follows $\lambda^{(t+1)} = \lambda^{(t)} + J\varpi_t \frac{\partial^{(\text{ng})} \mathcal{L}_j}{\partial \lambda}$
Update "dish-table" indicator variable hyperparameters μ_{ktm}^c using gradient in Eq. (5).
until convergence

in each update to reduce the variance (Hoffman *et al.*, 2010, 2013). In this case, the gradients with a single document in are replaced by the average gradients of all the documents in a mini-batch.

5 EXPERIMENTS

We evaluate our inference algorithm on real datasets with two different scale settings: small datasets with thousands of documents which can also be run using Gibbs sampler; large-scale data with millions of documents which can not be practically run with sampling methods. For small-scale settings, we illustrate competitive perplexity of our inference methods compare to Gibbs sampler but with much less computation time. We also report the running time and performance of our model for large-scale data sets.

5.1 DATASETS

As aforementioned, we use two groups of different scales of datasets. For the *small scale setting*, in order to compare with Gibbs sampler, we use the same datasets in (Nguyen *et al.*, 2014): a text dataset, NIPS, and image dataset, NUS-WIDE.

- NIPS[3] consists of 1740 document with the vocabulary size 13,649. To evaluate predictive performance, we randomly split into 90% training and 10% held-out for computing perplexity.
- NUS-WIDE (Chua *et al.*, 2009) contains a subset of 13 animal classes which totally include 3411 images. Held-out data includes 1357 images and the rest is used for training the model. For the image features, we use bag-of-word SIFT vector with dimension 500. For the context observations, we use the tags for each image which are 1000-dimension spare vectors.

[3]http://www.cs.nyu.edu/~roweis/data.html

For the *large-scale setting,* we use three different datasets including *Wikipedia*, *Pubmed*, and *Application Usage Activity (AUA).*

- *Wikipedia* includes about 1.1 million documents downloaded from wikipedia.com. We pre-process data using a vocabulary list taken from the top 10,000 words in Project Gutenberg and remove all words less than three characters (Hoffman *et al.*, 2013). For the context features we use the (first) writer of the articles and the (top level) categories inferred from tagged categories in each article as described in (De Vries *et al.*, 2010).
- *PubMed* comprises 1.4 million abstracts acquired from pubmed.gov. These documents are filtered with the published year from 2000 onward. Similar to *Wikipedia,* we also extracted the vocabulary from the whole dataset and only kept words with more than 2 characters. A top list of 10.000 words is used as vocabulary list for computing bag-of-word. We further extract the Medical Subject Headings (MeSH) and consider as the context.
- *Application Usage Activity (AUA)*: This dataset contains the usage behavior from more than one million users of a popular software application. Each user is treated as a document in which a word refers to a specific functionality of the application and word frequency refers to the number of times the user interact with the corresponding functionality. The total number of functionalities (vocabulary size) is roughly 10,000. In addition to the current application, each user also uses a host of other related software products which can be used as the context of the user. Applying MC2 to this data effectively cluster the set of users into different segments. To measure the clustering quality, we simply use a ground-truth of two clusters of paid and free users. Note that this information is not present in the context or the word content.

5.2 EXPERIMENT SETUPS

Since our observed data are discrete, we assume that they are generated from either Categorical or Multinomial distributions endowed with Dirichlet priors. The learning rate for stochastic learning at iteration t is $\varpi_t = (t + \varrho)^{-\iota}$where $\varrho \geq 0$ is the delay parameter, and $\iota \in (.5, 1]$ is the forgetting rate which controls how quickly previous statistics is forgotten. In the experiment for computing perplexity, we fixed $\varrho = 1$ and $\iota = 0.8$. The hyperparameters for Dirichlet distributions are set to 0.01 and 0.1 for content and context, respectively.

Small-scale setting

The experiments for NIPS and NUSWIDE datasets are carried out on an Intel Xeon 2.6GHz machine with 16 cores, 16GB RAM using a C# implementation running on Windows 7. SVI method can be parallelized when computing local updates. We run the experiment using both datasets in serial and parallel modes. The parallel implementation is accomplished using the Task Parallel Library (TPL) in .NET framework.

Large-scale setting

In order to handle big datasets, we implement our algorithms on Apache Spark platform[4]. The experiments for *Wikipedia*, *Pubmed*, *AUA* are run in two main settings with no context observations, and with full context observations for each corresponding context. Since HDP implementation is not available on Spark, we use the LDA implementation provided by Spark machine learning library (MLLIB) to compare perplexity with our algorithm. We set the number of topics for LDA equal to the number of topic truncated in the MC2 model.

5.3 EVALUATION METRICS

Perplexity. We use perplexity as the evaluation metric to compare the modelling performance between inference algorithms (Gibbs vs. SVI) or between model (our model vs. LDA). The perplexity is defined as perplexity $(w^{\text{test}}) = \exp\left\{-\frac{\sum_j \ln p(w_{j.}|\mathcal{D})}{\sum_j n_j}\right\}$ where w^{test} is the content words in the test set and $\mathcal{D}$ is the training data. Since we wish to compare our SVI algorithm with Gibbs sampler, we implemented importance sampling (Wallach *et al.*, 2009) to computed $\ln p\left(w_{j.} \mid \mathcal{D}\right)$ in both cases. In Spark MLLIB, there is no implementation for computing perplexity with importance sampling, we instead used the code given by Wallach *et al.* (2009).

Clustering performance. Since our model can carry out clustering for documents, we wish to compare clustering performance. However, documents usually do not have a "strong" ground truth and most of them are with multiple-cluster. For instance, with PubMed data, we use MeSH for each article as ground truth cluster but each article usually associates with several MeSH terms. Some popular clustering performance metrics including purity, Random Index(RI), Normalized Mutual Information (NMI), Fscore (Manning *et al.*, 2008, Chap16) are designed for single cluster ground truth. Whenever there is single cluster ground truth, for example, in the *AUA* dataset, we use the above four metrics. In other cases, we use the extended Normalized Mutual Information (eNMI) which is defined as follows. Let suppose that we have N objects each of which is belong to one of K clusters. A clustering algorithm will assign this object to one of T clusters. With N objects, we denote W as an $N \times K$ ground truth matrix where each row of this matrix represent a (transposed) one-hot vector encoding of the cluster it belongs. Similarly, we have $N \times T$ matrix as a result matrix. The joint probability when an object has the ground truth cluster k and is assigned to cluster t is $p(w, c) = W^{\mathsf{T}}C$. The mutual information between discovered clusters and ground truth cluster is $\text{MI}(W, C) = \sum_{k,t} p(w = k, c = t) \ln \frac{p(w=k,c=t)}{p(c=t)p(w=k)}$. The normalized mutual information is $\text{NMI}(W, C) = \frac{2\text{MI}(W,C)}{H(C)H(W)}$ where $H(\cdot)$ is the entropy of histogram of clusters. In the case of multiple clustering, we have the matrix W and C where each row is not one-hot vector but a vector with the sum as 1. We use some equations above for computing extended NMI (eNMI).

[4]http://spark.apache.org/

	Running time (s)	
	Sequential	Parallel
NIPS	11213	1431
NUSWIDE	8373	682

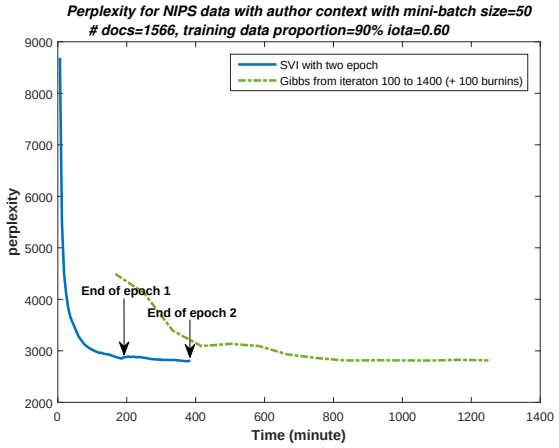
Table 1: Running time of two implementation version

5.4 EXPERIMENTAL RESULT

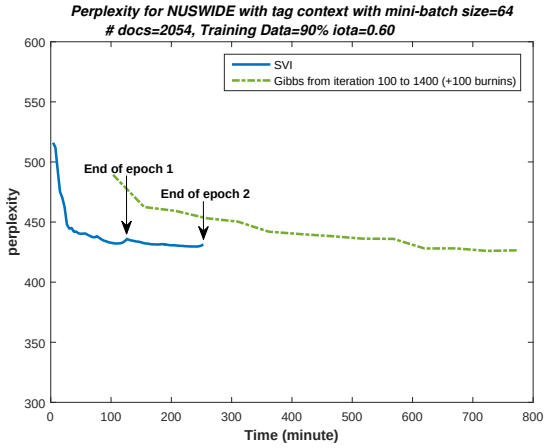
Results on small -scale setting

First, we demonstrate the performance of our proposed methods (SVI) compared with Gibbs sampler of (Nguyen *et al.*, 2014). For Gibbs samplers, we ran 1500 iterations and SVI with 50 documents in each mini-batch and compute perplexity. The Fig. 3 showed the predictive performance of them over running time. In both datasets, SVI can approach the performance of Gibbs sampler within one epoch[5]; after the first epoch, the perplexity only improved a little. To obtain the competitive performance with Gibbs sampler, our algorithm needs only one-fourth of the amount of running time. Furthermore, SVI algorithm is parallelizable. As shown in Table 1, running time with parallelized version on a single machine with 16 cores is further reduced significantly, 8 and 12 times for NIPS and NUS-WIDE, respectively. Note that our parallel SVI-MC2 only parallelize the local updates, hence, the per-core speedup also depends on the fraction of parallelizable local updates and the global update. In the case of NIPS data set, the dimension of the (global) content and context topic are 13,649 and 2037, respectively, while those of NUS-WIDE are 500 and 1000 which could explain why parallelization is more effective for NUS-WIDE.

[5]Each epoch is an iteration in which algorithm visited all data points.



(a) NIPS - context: author



(b) NUS-WIDE - context: tag

Figure 3: Perplexity with respect to running time on two datasets: NIPS and NUS-WISE. The blue line denotes the change of perplexity over running time with two epochs of data for SVI learning algorithm while the green line depicts perplexity running with Gibbs samplers. The results for Gibbs sampler is shown for every 100 iteration from 100-th to 1400-th iteration (excluding 100 burn-in iteration).

	Context availability		LDA
	100%	0%	
Wikipedia - writer	**2,167**	2,280	2,635
Pubmed - MeSH	**2,294**	2.448	3,178
AUA - other products	**142.3**	149.7	209.3

Table 2: Log perplexity of Wikipedia and PubMed data

	Context availability	
	1%	0%
eNMI	**0.084**	0.065

Table 3: Extended Normalized mutual information (NMI) for Pubmed data

Results on large-scale setting

In this setting, we validate the robustness of our algorithm with large-scale datasets. We ran our inference algorithm with *Wikipedia, PubMed*, and *AUA* datasets together with the LDA baseline on an 8-node Spark cluster. We used writer, MeSH, and other products used as contexts for *Wikipedia*, *PubMed*, and *AUA,* respectively. For each dataset, we ran data with full observations of context and without context. Table 2 depict the perplexity for these datasets with and without context compared with LDA. The predictive performance of SVI-MC2 improved remarkably compared to LDA.

For *PubMed* dataset, we used MeSH as the ground truth for clustering evaluation. As each document contains several MeSH terms, we use extended NMI (eNMI) for computing clustering performance. For each mini-batch, we compute eNMI of this mini-batch with its ground truth. The table 3 depict the average eNMI for all mini-batches in an epoch. With a very little availability of the ground truth as context, our algorithm can considerably improve clustering performance.

For *AUA* dataset, three different levels of context availability are used including no context, 1%, and full context. Since the ground truth clusters do not overlap, we can use the conventional metrics for clustering evaluation such as NMI, RI, purity, and Fscore. We also compute the average of the above indices for all mini-batches in an epoch. The clustering results are are shown in table 4. All clustering metrics showed the advantage of context observation (very small percentage is needed) to improve the clustering performance.

It is not possible to run the Gibbs sampler for these large datasets; even the serial version of SVI took too much time, hence we only reported running time for Spark SVI-MC2. With the mini-batch size of 500, the best running times are achieved using an 8-node cluster: Wikipedia: 17 hours; Pubmed: 18.5 hours; AUA: 18 hours. However, using a single-node (with 16-core) could also suffice with running time roughly 1.5 times slower than on a full 8-node cluster. We note that the size of the mini-batch (500) in this case strongly affects the effectiveness of the distributed-cluster setting. For example, with a mini-batch size of 1000, the speed-up factor on an 8-node cluster (compared to single-node) increases from 1.5 to 1.8.

Context	Avail.	NMI	Purity	RI	Fscore
Other products used	0%	0.027	0.14	0.284	0.12
	1%	**0.035**	0.174	0.286	0.128
	100%	0.033	**0.179**	**0.287**	**0.131**

Table 4: Clustering performance for *AUA* data

6 CONCLUSION

We have presented a scalable method for Bayesian nonparametric multilevel clustering with contextual side information. We proposed a tree-structured SVI approximation for an efficient approximation of the model's posterior. The approach can be directly parallelizable, and we provide parallelized implementations that work both on a single machine and on a distributed Apache Spark cluster. The experimental results demonstrate that our method is several orders of magnitude faster than existing the Gibb-sampler while yield the same model quality. Most importantly, our work enables the applicability of multilevel clustering to modern real-world datasets which can contain millions of documents.

References

Amari, Shun-Ichi. 1998. Natural gradient works efficiently in learning. *Neural computation*, **10**(2), 251–276.

Antoniak, C.E. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, **2**(6), 1152–1174.

Bishop, Christopher M, *et al.* 2006. *Pattern recognition and machine learning*. Vol. 1. springer New York.

Blei, D.M., & Jordan, M.I. 2006. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, **1**(1), 121–143.

Blei, D.M., Ng, A.Y., & Jordan, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Chua, Tat-Seng, Tang, Jinhui, Hong, Richang, Li, Haojie, Luo, Zhiping, & Zheng, Yantao. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. *Page 48 of: Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM.

Connor, R. J., & Mosiman, J. E. 1969. Concepts of independence for proportions with a generalization of the Dirichlet distribution. *Journal of the American Statistical Association*, **64**, 194–206.

De Vries, Christopher M, Nayak, Richi, Kutty, Sangeetha, Geva, Shlomo, & Tagarelli, Andrea. 2010. Overview of the INEX 2010 XML mining track: Clustering and classification of XML documents. *Pages 363–376 of: Comparative evaluation of focused retrieval*. Springer Berlin Heidelberg.

Diez-Roux, Ana V. 2000. Multilevel analysis in public health research. *Annual review of public health*, **21**(1), 171–192.

Elango, Pradheep K, & Jayaraman, Karthik. 2005. Clustering Images Using the Latent Dirichlet Allocation Model. *University of Wisconsin*.

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

Hoffman, M.D., Blei, D.M., & Bach, F. 2010. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, **23**, 856–864.

Hox, Joop. 2010. *Multilevel analysis: Techniques and applications*. Routledge.

Ishwaran, H., & James, L.F. 2001. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**(453), 161–173.

Lu, Yue, Mei, Qiaozhu, & Zhai, ChengXiang. 2011. Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. *Information Retrieval*, **14**(2), 178–203.

Manning, C.D., Raghavan, P., & Schütze, H. 2008. *Introduction to Information Retrieval*. Vol. 1. Cambridge University Press Cambridge.

Nguyen, T. C., Phung, D., Gupta, S., & Venkatesh, S. 2013. Extraction of Latent Patterns and Contexts from Social Honest Signals Using Hierarchical Dirichlet Processes. *Pages 47–55 of: 2013 IEEE International Conference on Pervasive Computing and Communications (PerCom)*.

Nguyen, V., Phung, D., Venkatesh, S. Nguyen, X.L., & Bui, H. 2014. Bayesian Nonparametric Multilevel Clustering with Group-Level Contexts. *Pages 288–296 of: Proc. of International Conference on Machine Learning (ICML)*.

Phung, D., Nguyen, X., Bui, H., Nguyen, T.V., & Venkatesh, S. 2012. *Conditionally Dependent Dirichlet Processes for Modelling Naturally Correlated Data Sources*. Tech. rept. Pattern Recognition and Data Analytics, Deakin University.

Phung, D., Nguyen, T. C., Gupta, S., & Venkatesh, S. 2014. Learning Latent Activities from Social Signals with Hierarchical Dirichlet Process. *Pages 149–174 of:* et al.

Gita Sukthankar (ed), *Handbook on Plan, Activity, and Intent Recognition*. Elsevier.

Rodriguez, A., Dunson, D.B., & Gelfand, A.E. 2008. The nested Dirichlet process. *Journal of the American Statistical Association*, **103**(483), 1131–1154.

Teh, Y.W., Jordan, M.I., Beal, M.J., & Blei, D.M. 2006. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**(476), 1566–1581.

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

Wallach, H.M., Murray, I., Salakhutdinov, R., & Mimno, D. 2009. Evaluation methods for topic models. *Pages 1105–1112 of: Procs. of Int. Conference on Machine Learning (ICML)*. ACM.

Wang, C., Paisley, J., & Blei, D.M. 2011. Online variational inference for the hierarchical Dirichlet process. *In: Artificial Intelligence and Statistics*.

Wong, T.-T. 1998. Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation*, **97**, 165–181.

Wulsin, D., Jensen, S., & Litt, B. 2012. A Hierarchical Dirichlet Process Model with Multiple Levels of Clustering for Human EEG Seizure Modeling. *In: Proc. of International Conference on Machine Learning (ICML)*.

Xie, Pengtao, & Xing, Eric P. 2013. Integrating Document Clustering and Topic Modeling. *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*.