Bayesian Nonparametric Multilevel Clustering with Group-Level Contexts

Vu Nguyen TVNGUYE@DEAKIN.EDU.AU
Center for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Australia

Dinh Phung DINH.PHUNG@DEAKIN.EDU.AU
Center for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Australia

XuanLong Nguyen XUANLONG@UMICH.EDU
Department of Statistics, University of Michigan, Ann Arbor, USA

Svetha Venkatesh SVETHA.VENKATESH@DEAKIN.EDU.AU
Center for Pattern Recognition and Data Analytics (PRaDA), Deakin University, Australia

Hung Hai Bui BUI.H.HUNG@GMAIL.COM
Laboratory for Natural Language Understanding, Nuance Communications, Sunnyvale, USA

Abstract

We present a Bayesian nonparametric framework for multilevel clustering which utilizes group-level context information to simultaneously discover low-dimensional structures of the group contents and partitions groups into clusters. Using the Dirichlet process as the building block, our model constructs a product base-measure with a nested structure to accommodate content and context observations at multiple levels. The proposed model possesses properties that link the nested Dirichlet processes (nDP) and the Dirichlet process mixture models (DPM) in an interesting way: integrating out all contents results in the DPM over contexts, whereas integrating out group-specific contexts results in the nDP mixture over content variables. We provide a Polya-urn view of the model and an efficient collapsed Gibbs inference procedure. Extensive experiments on real-world datasets demonstrate the advantage of utilizing context information via our model in both text and image domains.

1. Introduction

In many situations, content data naturally present themselves in groups, e.g., students are grouped into classes, classes grouped into schools, words grouped into documents, etc. Furthermore, each content group can be associated with additional context information (teachers of the class, authors of the document, time and location stamps). Dealing with grouped data, a setting known as *multilevel analysis* (Hox, 2010; Diez-Roux, 2000), has diverse application domains ranging from document modeling (Blei et al., 2003) to public health (Leyland & Goldstein, 2001).

This paper considers specifically the multilevel clustering problem in multilevel analysis: to jointly cluster both the content data and their groups when there is group-level context information. By *context,* we mean a secondary data source attached to the group of primary *content* data. An example is the problem of clustering documents, where each document is a group of words associated with group-level context information such as time-stamps, list of authors, etc. Another example is image clustering where visual image features (e.g. SIFT) are the content and image tags are the context.

To cluster groups together, it is often necessary to perform dimensionality reduction of the content data by forming content topics, effectively performing clustering of the content as well. For example, in document clustering, using bag-of-words directly as features is often problematic due to the large vocabulary size and the sparsity of the in-document word occurrences. Thus, a typical approach is to first apply dimensionality reduction techniques such as LDA (Blei et al., 2003) or HDP (Teh et al., 2006) to find word topics (i.e., distributions on words), then perform document clustering using the word topics and the document-level context information as features. In such a

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

cascaded approach, the dimensionality reduction step (e.g., topic modeling) is not able to utilize the context information. This limitation suggests that a better alternative is to perform context-aware document clustering and topic modeling jointly. With a joint model, one can expect to obtain improved document clusters as well as context-guided content topics that are more predictive of the data.

Recent work has attempted to jointly capture word topics and document clusters. Parametric approaches (Xie & Xing, 2013) are extensions of the LDA (Blei et al., 2003) and require specifying the number of topics and clusters in advance. Bayesian nonparametric approaches including the nested Dirichlet process (nDP) (Rodriguez et al., 2008) and the multi-level clustering hierarchical Dirichlet Process (MLC-HDP) (Wulsin et al., 2012) can automatically adjust the number of clusters. We note that none of these methods can utilize context data.

This paper propose the *Multilevel Clustering with Context* (MC^2), a Bayesian nonparametric model to jointly cluster both content and groups while fully utilizing group-level context. Using the Dirichlet process as the building block, our model constructs a product base-measure with a nested structure to accommodate both content and context observations. The MC^2 model possesses properties that link the nested Dirichlet process (nDP) and the Dirichlet process mixture model (DPM) in an interesting way: integrating out all contents results in the DPM over contexts, whereas integrating out group-level context results in the nDP mixture over content variables. For inference, we provide an efficient collapsed Gibbs sampling procedure for the model.

The advantages of our model are: (1) the model automatically discovers the (unspecified) number of groups clusters and the number of topics while fully utilizing the context information; (2) content topic modeling is informed by group-level context information, leading to more predictive content topics; (3) the model is robust to partially missing context information. In our experiments, we demonstrate that our proposed model achieves better document clustering performances and more predictive word topics in real-world datasets in both text and image domains.

2. Related Background

There have been extensive works on clustering documents in the literature. Due to limited scope of the paper, we only describe works closely related to probabilistic topic models. We note that standard topic models such as LDA (Blei et al., 2003) or its nonparametric Bayesian counter part, HDP (Teh et al., 2006) exploits the group structure for word clustering. However these models do not cluster documents.

An approach to document clustering is to employ a two-stage process. First, topic models (e.g. LDA or HDP) are applied to extract the topics and their mixture proportion for each document. Then, this is used as feature input to another clustering algorithm. Some examples of this approach include the use of LDA+Kmeans for image clustering (Elango & Jayaraman, 2005) and HDP+Affinity Propagation for clustering human activities (Nguyen et al., 2013).

A more elegant approach is to simultaneously cluster documents and discover topics. The first Bayesian nonparametric model proposed for this task is the nested Dirichlet Process (nDP) (Rodriguez et al., 2008) where documents in a cluster share the same distribution over topic atoms. Although the original nDP does not force the topic atoms to be shared across document clusters, this can be achieved by simply introducing a DP prior for the nDP base measure. The same observation was also made by (Wulsin et al., 2012) who introduced the MLC-HDP, a 3-level extension to the nDP. This model thus can cluster words, documents and document-corpora with shared topic atoms throughout the group hierarchy. Xie et al (Xie & Xing, 2013) recently introduced the Multi-Grain Clustering Topic Model which allows mixing between global topics and document-cluster topics. However, this is a parametric model which requires fixing the number of topics in advance. More crucially, all of these existing models do not attempt to utilize group-level context information.

Modelling with Dirichlet Process

We provide a brief account of the Dirichlet process and its variants. The literature on DP is vast and we refer to (Hjort et al., 2010) for a comprehensive account. Here we focus on DPM, HDP and nDP which are related to our work.

Dirichlet process (Ferguson, 1973) is a basic building block in Bayesian nonparametrics. Let $(\Theta, \mathcal{B}, H)$ be a probability measure space, and γ is a positive number, a Dirichlet process $\mathrm{DP}(\gamma, H)$ is a distribution over discrete random probability measure G on $(\Theta, \mathcal{B})$. Sethuraman (Sethuraman, 1994) provides an alternative constructive definition which makes the discreteness property of a draw from a Dirichlet process explicit via the stick-breaking representation: $G = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$ where $\phi_k \overset{\text{iid}}{\sim} H, k = 1, \ldots, \infty$ and $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty}$ are the weights constructed through a 'stick-breaking' process $\beta_k = v_k \prod_{s<k}(1 - v_s)$ with $v_k \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \gamma)$. It can be shown that $\sum_{k=1}^{\infty} \beta_k = 1$ with probability one, and as a convention (Pitman, 2002), we hereafter write $\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$.

Due to its discrete nature, Dirichlet process has been widely used in Bayesian mixture models as the prior distribution on the mixing measures, each is associated with an atom ϕ_k in the stick-breaking representation of G above. A like-

lihood kernel $F(\cdot)$ is used to generate data $x_i \mid \phi_k \overset{\text{iid}}{\sim} F(\cdot \mid \phi_k)$, resulting in a model known as the *Dirichlet process mixture model* (DPM), pioneered by the work of (Antoniak, 1974) and subsequently developed by many others. In section 3 we provide a precise definition for DPM.

While DPM models exchangeable data within a *single* group, the Dirichlet process can also be constructed hierarchically to provide prior distributions over *multiple* exchangeable groups. Under this setting, each group is modelled as a DPM and these models are 'linked' together to reflect the dependency among them – a formalism which is generally known as dependent Dirichlet processes (MacEachern, 1999). One particular attractive approach is the *hierarchical Dirichlet processes* (Teh et al., 2006) which posits the dependency among the group-level DPM by another Dirichlet process, i.e., $G_j \mid \alpha, G_0 \sim \text{DP}(\alpha, G_0)$ and $G_0 \mid \gamma, H \sim \text{DP}(\gamma, H)$ where G_j is the prior for the j-th group, linked together via a discrete measure G_0 whose distribution is another DP.

Yet another way of using DP to model multiple groups is to construct random measure in a nested structure in which the DP base measure is itself another DP. This formalism is the *nested Dirichlet Process* (Rodriguez et al., 2008), specifically $G_j \overset{\text{iid}}{\sim} U$ where $U \sim \text{DP}(\alpha \times \text{DP}(\gamma H))$. Modeling G_j (s) hierarchically as in HDP and nestedly as in nDP yields different effects. HDP focuses on exploiting statistical strength across groups via sharing atoms ϕ_k (s), but it does not partition groups into clusters. This statement is made precisely by noting that $P(G_j = G_{j'}) = 0$ in HDP. Whereas, nDP emphasizes on inducing clusters on both observations and distributions, hence it partitions groups into clusters. To be precise, the prior probability of two groups being clustered together is $P(G_j = G_{j'}) = \frac{1}{a+1}$. Finally we note that this original definition of nDP in (Rodriguez et al., 2008) does not force the atoms to be shared across clusters of groups, but this can be achieved by simply introducing a DP prior for the nDP base measure, a modification that we use in this paper. This is made clearly in our definition for nDP mixture in section 3.

3. Multilevel Clustering with Contexts

3.1. Model description and stick-breaking

Consider data presented in a two-level group structure as follows. Denote by J the number of groups; each group j contains N_j exchangeable data points, represented by $\mathbf{w}_j = \{w_{j1}, w_{j2}, \ldots, w_{jN_j}\}$. For each group j, the group-specific context data is denoted by x_j. Assuming that the groups are exchangeable, the overall data is $\{(x_j, \mathbf{w}_j)\}_{j=1}^{J}$. The collection $\{\mathbf{w}_1, \ldots, \mathbf{w}_J\}$ represents observations of the group contents, and $\{x_1, \ldots, x_J\}$ represents observations of the group-level contexts.

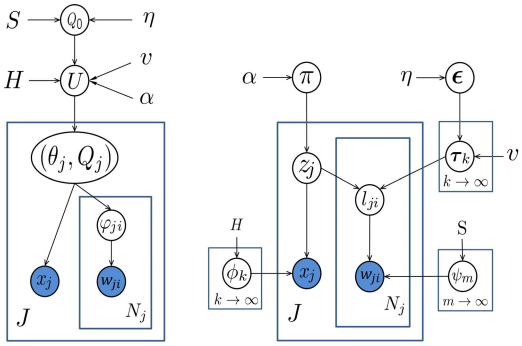

Figure 1. Graphical model representation for the proposed model. Right figure illustrates a stick breaking representation.

We now describe the generative process of MC^2 that generates a two-level clustering of this data. We use a group-level DP mixture to generate an infinite cluster model for groups. Each group cluster k is associated with an atom having the form of a pair (ϕ_k, Q_k^*) where ϕ_k is a parameter that generates the group-level contexts within the cluster and Q_k^* is a measure that generates the group contents within the same cluster.

To generate atomic pairs of context parameter and measure-valued content parameter, we introduce a product base-measure of the form $H \times \text{DP}(vQ_0)$ for the group-level DP mixture. Drawing from a DP mixture with this base measure, each realization is a pair (θ_j, Q_j); θ_j is then used to generate the context x_j and Q_j is used to repeatedly produce the set of content observations w_{ji} within the group j. Specifically,

$$\begin{aligned} U \sim \text{DP}(\alpha(H \times \text{DP}(vQ_0))) \text{ where } Q_0 \sim \text{DP}(\eta S) \\ (\theta_j, Q_j) \overset{\text{iid}}{\sim} U \text{ for each group } j \\ x_j \sim F(.|\theta_j), \quad \varphi_{ji} \overset{\text{iid}}{\sim} Q_j, \quad w_{ji} \sim Y(.|\varphi_{ji}) \end{aligned} \tag{1}$$

In the above, H and S are respectively base measures for context and content parameters θ_j and φ_{ji}. The context and content observations are then generated via the likelihood kernels $F(\cdot \mid \theta_j)$ and $Y(\cdot \mid \varphi_{ji})$. To simplify inference, H and S are assumed to be conjugate to F and Y respectively. The generative process is illustrated in Figure 1.

STICK-BREAKING REPRESENTATION

We now derive the stick-breaking construction for MC^2 where all the random discrete measures are specified by a distribution over integers and a countable set of atoms. The random measure U in Eq. (1) has the stick-breaking form:

$$U = \sum_{k=1}^{\infty} \pi_k \delta_{(\phi_k, Q_k^*)} \tag{2}$$

where $\pi \sim \text{GEM}(\alpha)$ and $(\phi_k, Q_k^*) \overset{\text{iid}}{\sim} H \times \text{DP}(vQ_0)$. Equivalently, this means ϕ_k is drawn i.i.d. from H and Q_k^* drawn i.i.d. from $\text{DP}(vQ_0)$. Since $Q_0 \sim \text{DP}(\eta S)$, Q_0 and Q_k^* have the standard HDP (Teh et al., 2006) stick-breaking forms: $Q_0 = \sum_{m=1}^{\infty} \epsilon_m \delta_{\psi_m}$ where $\boldsymbol{\epsilon} \sim \text{GEM}(\eta)$, $\psi_m \overset{\text{iid}}{\sim} S$; $Q_k^* = \sum_{m=1}^{\infty} \tau_{k,m} \delta_{\psi_m}$ where $\boldsymbol{\tau}_k = (\tau_{k1}, \tau_{k2}, \ldots) \sim \text{DP}(v, \boldsymbol{\epsilon})$.

For each group j we sample the parameter pair $(\theta_j, Q_j) \overset{\text{iid}}{\sim} U$; equivalently, this means drawing $z_j \overset{\text{iid}}{\sim} \pi$ and letting $\theta_j = \phi_{z_j}$ and $Q_j = Q_{z_j}^*$. For the i-th content data within the group j, the content parameter φ_{ji} is drawn $\overset{\text{iid}}{\sim} Q_j = Q_{z_j}^*$; equivalently, this means drawing $l_{ji} \overset{\text{iid}}{\sim} \tau_{z_j}$ and letting $\varphi_{ji} = \psi_{l_{ji}}$. Figure 1 presents the graphical model of this stick-breaking representation.

3.2. Inference and Polya Urn View

We use collapsed Gibbs sampling, integrating out ϕ_k(s), ψ_m(s), π and τ_k (s). Latent variables $\boldsymbol{z}$, $\boldsymbol{l}$, $\boldsymbol{\epsilon}$ and the hyperparameters α, v, η will be resampled. We only describe the key inference steps in sampling $\boldsymbol{z}, \boldsymbol{l}$ and $\boldsymbol{\epsilon}$ here and refer to the supplementary material (Nguyen et al., 2014) for the rest of the details (including how to sample the hyperparameters).

Sampling $\boldsymbol{z}$. The required conditional distribution is $p(z_j = k \mid \boldsymbol{z}_{-j}, \boldsymbol{l}, \boldsymbol{x}, \alpha, H) \propto$

$$p(z_j = k|\boldsymbol{z}_{-j}, \alpha)\, p(x_j|z_j = k, \boldsymbol{z}_{-j}, \boldsymbol{x}_{-j}, H) \times p(l_{j*}|z_j = k, \boldsymbol{l}_{-j*}, \boldsymbol{z}_{-j}, \boldsymbol{\epsilon}, v)$$

The first term can be recognized as a form of the Chinese restaurant process (CRP). The second term is the predictive likelihood for the context observations under the component ϕ_k after integrating out ϕ_k. This can be evaluated analytically due to conjugacy of F and H. The last term is the predictive likelihood for the group content-index $l_{j*} = \{l_{ji} | i = 1 \ldots N_j\}$. Since $l_{ji} \mid z_j = k \overset{\text{iid}}{\sim} \text{Mult}(\boldsymbol{\tau}_k)$ where $\boldsymbol{\tau}_k \sim \text{Dir}(v\epsilon_1, \ldots, v\epsilon_M, \epsilon_{\text{new}})$, the last term can also be evaluated analytically by integrating out $\boldsymbol{\tau}_k$ using the Multinomial-Dirichlet conjugacy property.

Sampling $\boldsymbol{l}$. Let w_{-ji} be the same set as $\mathbf{w}$ excluding w_{ji}, let $w_{-ji}(m) = \{w_{j'i'} | (j'i') \neq (ji) \wedge l_{j'i'} = m\}$ and $\boldsymbol{l}_{-ij}(k) = \{l_{j'i'} | (j'i') \neq (ji) \wedge z_{j'} = k\}$. Then $p(l_{ji} = m \mid l_{-ji}, z_j = k, z_{-j}, v, \mathbf{w}, \boldsymbol{\epsilon}, S) \propto$

$$p(w_{ji}|\boldsymbol{l}, w_{-ji}, S)\, p(l_{ji} = m|\boldsymbol{l}_{-ji}, z_j = k, z_{-j}, \boldsymbol{\epsilon}, v) = p(w_{ji} \mid w_{-ji}(m), S)\, p(l_{ji} = m \mid \boldsymbol{l}_{-ji}(k), \boldsymbol{\epsilon}, v)$$

The first term is the predictive likelihood under mixture component ψ_m after integrating out ψ_m, which can be evaluated analytically due to the conjugacy of Y and S. The second term is in the form of a CRP similar to the one that arises during inference for HDP (Teh et al., 2006).

Sampling $\boldsymbol{\epsilon}$. Sampling $\boldsymbol{\epsilon}$ requires information from both $\boldsymbol{z}$ and $\boldsymbol{l}$.

$$p(\epsilon \mid \boldsymbol{l}, \boldsymbol{z}, v, \eta) \propto p(\boldsymbol{l} \mid \boldsymbol{\epsilon}, v, \boldsymbol{z}, \eta) \times p(\boldsymbol{\epsilon} \mid \eta) \quad (3)$$

Using a similar strategy in HDP, we introduce auxiliary variables (o_{km}), then alternatively sample together with $\boldsymbol{\epsilon}$:

$$p(o_{km} = h \mid \cdot) \propto \text{Stirl}(h, n_{km})\, (v\epsilon_m)^h,\ h = 0, 1, \ldots, n_{km}$$

$$p(\boldsymbol{\epsilon} \mid \cdot) \propto \epsilon_{\text{new}}^{\eta - 1} \prod_{m=1}^{M} \epsilon_m^{\sum_k o_{km} - 1}$$

where $\text{Stirl}(h, n_{km})$ is the Stirling number of the first kind, n_{km} is the count of seeing the pair $(z_j = k, l_{ji} = m)$: $\forall i, j$, and finally M is the current number of active content topics. It clear that o_{km} can be sampled from a Multinomial distribution and ϵ from an $(M+1)$-dim Dirichlet distribution.

POLYA URN VIEW

Our model exhibits a Polya-urn view using the analogy of a fleet of buses, driving customers to restaurants. Each bus represents a group and customers on the bus are data points within the group. For each bus j, z_j acts as the index to the restaurant for its destination. Thus, buses form clusters at their destination restaurants according to a CRP: a new bus drives to an existing restaurant with the probability proportional to the number of other buses that have arrived at that restaurant, and with probability proportional to α, it goes to a completely new restaurant.

Once all the buses have delivered customers to the restaurants, *all customers at the restaurants start to behave in the same manner as in a Chinese restaurant franchise (CRF) process*: customers are assigned tables according to a restaurant-specific CRP; tables are assigned with dishes ψ_m (representing the content topic atoms) according to a global franchise CRP. In addition to the usual CRF, at restaurant k, a single dessert ϕ_k (which represents the context-generating atom, drawing $\overset{\text{iid}}{\sim}$ from H) will be served to all the customers at that restaurant. Thus, every customer on the same bus j will be served the same dessert ϕ_{z_j}. We observe three sub-CRPs, corresponding to the three DP(s) in our model: the CRP at the dish level is due to the $\text{DP}(\eta S)$, the CRP forming tables inside each restaurant is due to the $\text{DP}(vQ_0)$, and the CRP aggregating buses to restaurants is due to the $\text{DP}(\alpha(H \times \text{DP}(vQ_0)))$.

3.3. Marginalization property

We study marginalization property for our model when either the content topics φ_{ji} (s) or context topics θ_j (s) are

marginalized out. Our main result is established in Theorem 3 where we show an interesting link to nested DP and DPM via our model.

Let H be a measure over some measurable spaces (Θ, Σ). Let $\mathbb{P}$ be the set of all measures over (Θ, Σ), suitably endowed with some σ-algebra. Let $G \sim \text{DP}(\alpha H)$ and $\theta_i \overset{\text{iid}}{\sim} G$. The collection (θ_i) then follows the DP mixture distribution which is defined formally below.

(DPM) A DPM is a probability measure over $\Theta^n \ni (\theta_1, \ldots, \theta_n)$ with the usual product sigma algebra Σ^n such that for every collection of measurable sets $\{(S_1, \ldots, S_n) : S_i \in \Sigma, i = 1, \ldots, n\}$:

$$\text{DPM}(\theta_1 \in S_1, \ldots, \theta_n \in S_n | \alpha, H) = \int \prod_{i=1}^{n} G(S_i) \, \text{DP}(dG \mid \alpha H)$$

We now state a result regarding marginalization of draws from a DP mixture with a joint base measure. Consider two measurable spaces (Θ_1, Σ_1) and (Θ_2, Σ_2) and let (Θ, Σ) be their product space where $\Theta = \Theta_1 \times \Theta_2$ and $\Sigma = \Sigma_1 \times \Sigma_2$. Let H^* be a measure over the product space $\Theta = \Theta_1 \times \Theta_2$ and let H_1 be the marginal of H^* over Θ_1 in the sense that for any measurable set $A \in \Sigma_1$, $H_1(A) = H^*(A \times \Theta_2)$. Then drawing $(\theta_i^{(1)}, \theta_i^{(2)})$ from a DP mixture with base measure αH and marginalizing out $(\theta_i^{(2)})$ is the same as drawing $(\theta_i^{(1)})$ from a DP mixture with base measure H_1. Formally

Proposition 1. *Denote by θ_i the pair $\left(\theta_i^{(1)}, \theta_i^{(2)}\right)$, there holds*

$$DPM\left(\theta_1^{(1)} \in S_1, \ldots, \theta_n^{(1)} \in S_n \mid \alpha H_1\right) = DPM\left(\theta_1 \in S_1 \times \Theta_2, \ldots, \theta_n \in S_n \times \Theta_2 \mid \alpha H^*\right)$$

for every collection of measurable sets $\{(S_1, \ldots, S_n) : S_i \in \Sigma_1, i = 1, \ldots, n\}$.

Proof. see supplementary material (Nguyen et al., 2014). □

Next we give a formal definition for the nDP mixture: $\varphi_{ji} \overset{\text{iid}}{\sim} Q_j$, $Q_j \overset{\text{iid}}{\sim} U$, $U \sim \text{DP}(\alpha \text{DP}(vQ_0))$, $Q_0 \sim DP(\eta S)$.

Definition 2. (nested DP Mixture) An nDPM is a probability measure over $\Theta^{\sum_{j=1}^{J} N_j} \ni (\varphi_{11}, \ldots, \varphi_{1N_1}, \ldots, \varphi_{JN_J})$ equipped with the usual product sigma algebra $\Sigma^{N_1} \times \ldots \times \Sigma^{N_J}$ such that for every collection of measurable sets $\{(S_{ji}) : S_{ji} \in \Sigma, j = 1, \ldots, J, i = 1 \ldots, N_j\}$:

$$\text{nDPM}(\varphi_{ji} \in S_{ji}, \forall i, j | \alpha, v, \eta, S) = \int \int \left\{ \prod_{j=1}^{J} \int \prod_{i=1}^{N_j} Q_j(S_{ji}) \, U(dQ_j) \right\} \times \text{DP}(dU \mid \alpha \text{DP}(vQ_0)) \, \text{DP}(dQ_0 \mid \eta, S)$$

We now have the sufficient formalism to state the marginalization result for our model.

Theorem 3. *Given α, H and α, v, η, S, let $\boldsymbol{\theta} = (\theta_j : \forall j)$ and $\boldsymbol{\varphi} = (\varphi_{ji} : \forall j, i)$ be generated as in Eq (1). Then, marginalizing out $\boldsymbol{\varphi}$ results in DPM $(\boldsymbol{\theta} \mid \alpha, H)$, whereas marginalizing out $\boldsymbol{\theta}$ results in nDPM $(\boldsymbol{\varphi} | \alpha, v, \eta, S)$.*

Proof. We sketch the main steps, supplementary material (Nguyen et al., 2014) provides more detail. Let $H^* = H_1 \times H_2$, we note that when either H_1 or H_2 are random, a result similar to Proposition 1 still holds by taking the expectation on both sides of the equality. Now let $H_1 = H$ and $H_2 = \text{DP}(vQ_0)$ where $Q_0 \sim \text{DP}(\eta S)$ yields the proof for the marginalization of $\boldsymbol{\varphi}$; let $H_1 = \text{DP}(vQ_0)$ and $H_2 = H$ yields the proof for the marginalization of $\boldsymbol{\theta}$. □

4. Experiments

We first evaluate the model via simulation studies, then demonstrate its applications on text and image modeling using three real-world datasets. Throughout this section, unless explicitly stated, discrete data is modeled by Multinomial with Dirichlet prior, while continuous data is modeled by Gaussian (unknown mean and unknown variance) with Gaussian-Gamma prior.

4.1. Simulation studies

The main goal is to investigate the posterior consistency of the model, i.e., its ability to recover the true group clusters, context distribution and content topics. To synthesize the data, we use $M = 13$ topics which are the 13 unique letters in the ICML string "INTERNATIONAL CONFERENCE MACHINE LEARNING". Similar to (Griffiths & Steyvers, 2004), each topic ψ_m is a distribution over 35 words (pixels) and visualized as a 7×5 binary image. We generate $K = 4$ clusters of 100 documents each. For each cluster, we choose a set of topics corresponding to letters in the each of 4 words in the ICML string. The topic mixing distribution τ_k is an uniform distribution over the chosen topic letters. Each cluster is also assigned a context-generating univariate Gaussian distribution. These generating parameters are shown in Figure 2 (left). Altogether we have $J = 400$ documents; for each document we sample $N_j = 50$ words and a context variable x_j drawing from the cluster-specific Gaussian.

We model the word w_{ji} with Multinomial and Gaussian for context x_j. After 100 Gibbs iterations, the number of context and content topics ($K = 4, M = 13$) are recovered correctly: the learned context atoms ϕ_k and topic ψ_m are almost identical to the ground truth (Figure 2, right) and the model successfully identifies the 4 clusters of documents with topics corresponding to the 4 words in the

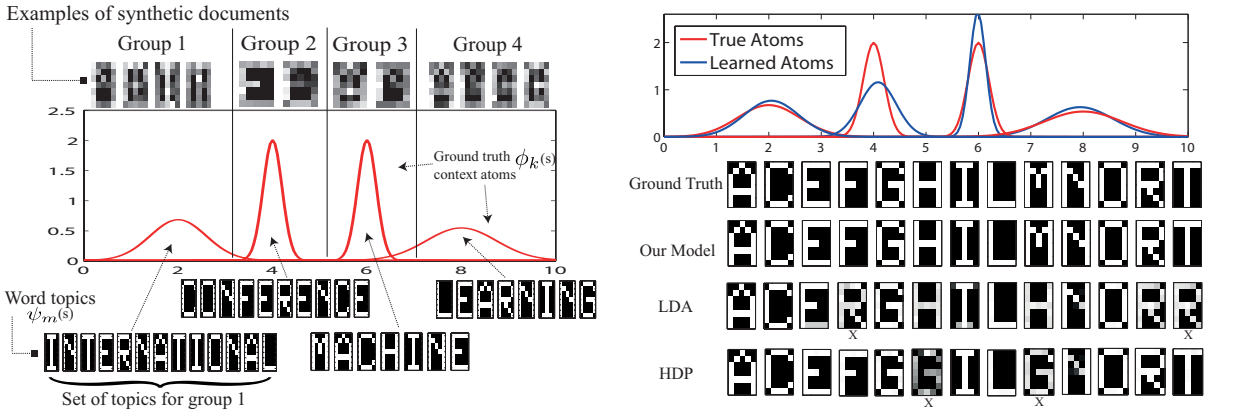

Figure 2. Results from simulation study. Left: illustration of data generation with ground truth for context atoms are 4 univariate Gaussians centered at $2, 4, 6$ and 8 respectively (different variances). Right: Our model recovers the correct 4 group clusters, their context distributions and the set of shared topics. LDA and HDP are unable to recover the true content topics without using contexts.

ICML string.

To demonstrate the importance of context observation, we then run LDA and HDP with only the word observations (ignoring context) where the number of topic of LDA is set to 13. As can be seen from Figure 2 (right), LDA and HDP have problems in recovering the true topics. They cannot distinguish small differences between the overlapping character topics (e.g M vs N, or I vs T). Further analysis of the role of context in MC^2 is provided in supplementary material (Nguyen et al., 2014) due to lacking of space.

4.2. Experiments with Real-World Datasets

We use two standard NIPS and PNAS text datasets, and the NUS-WIDE image dataset.

NIPS contains 1,740 documents with vocabulary size 13,649 (excluding stop words); timestamps (1987-1999), authors (2,037) and title information are available and used as group-level context. *PNAS* contains 79,800 documents, vocab size = 36,782 with publication timestamp (915-2005). For *NUS-WIDE* we use a subset of the 13-class animals [1] comprising of 3,411 images (2,054 images for training and 1357 images for testing) with off-the-shelf features including 500-dim bag-of-word SIFT vector and 1000-dim bag-of-tag annotation vector.

Text Modeling with Document-Level Contexts
We use NIPS and PNAS datasets with 90% for training and 10% for held-out perplexity evaluation. We compare the perplexity with HDP (Teh et al., 2006) where no group-level context can be used, and npTOT (Dubey et al., 2012) where only timestamp information can be used. We note that unlike our model, npTOT requires replication of document timestamp for *every* word in the document, which is somewhat unnatural.

We use perplexity score (Blei et al., 2003) on held-out data as performance metric, defined as[2] $\exp\left\{-\sum_{j=1}^{J}\log p\left(\mathbf{w}_j^{\text{test}} \mid \boldsymbol{x}^{\text{train}}, \mathbf{w}^{\text{train}}\right) / \left(\sum_j N_j^{\text{test}}\right)\right\}$. To ensure fairness and comparable evaluation, *only words* in held-out data is used to compute the perplexity. We use univariate Gaussian for timestamp and Multinomial distributions for words, tags and authors. We ran collapsed Gibbs for 500 iterations after 100 burn-in samples.

Table 1 shows the results where MC^2 achieves significant better performance. This shows that group-level context information during training provide useful guidance for the modelling tasks. Regarding the informative aspect of group-level context, we achieve better perplexity with timestamp information than with titles and authors. This may be explained by the fact that 1361 authors (among 2037) show up only once in the data while title provides little additional information than what already in that abstracts. Interestingly, without the group-level context information, our model still predicts the held-out words better than HDP. This suggests that inducing partitions over documents simultaneously with topic modelling is beneficial.

Beyond the capacity of HDP and npTOT, our model can induce clusters over documents (value of K in Table 1). Figure 3 shows an example of one such document cluster discovered from NIPS data with authors as context.

Our proposed model also allows flexibility in deriving useful understanding into the data and to evaluate on its predictive capacity (e.g., who most likely wrote this article, which authors work in the same research topic and so on). Another possible usage is to obtain *conditional* distribu-

[1] downloaded from http://www.ml-thu.net/~jun/data/

[2] Supplementary material (Nguyen et al., 2014) provides further details on how to derive this score from our model

Method	Perplexity (*on words only*)				Feature used
	PNAS	(K,M)	NIPS	(K,M)	
HDP (Teh et al., 2006)	3027.5	$(-, 86)$	1922.1	$(-, 108)$	words
npTOT (Dubey et al., 2012; Phung et al., 2012)	2491.5	$(-, 145)$	1855.33	$(-, 94)$	words+timestamp
MC^2 without context	1742.6	$(40, 126)$	1583.2	$(19, 61)$	words
MC^2 with titles	–	–	1393.4	$(32, 80)$	words+title
MC^2 with authors	–	–	1246.3	$(8, 55)$	words+authors
MC^2 with timestamp	**895.3**	$(12, 117)$	**984.7**	$(15, 95)$	words+timestamp

Table 1. Perplexity evaluation on PNAS and NIPS datasets. (K,M) is (#cluster,#topic). (Note: missing results are due to title and author information not available in PNAS dataset).

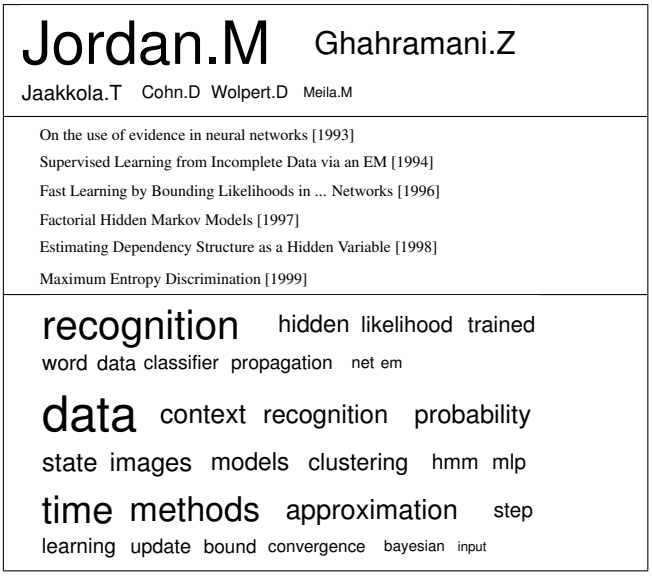

Figure 3. An example of document cluster from NIPS. Top: distribution over authors. Middle: examples of paper titles. Bottom: examples of word topics in this cluster.

tions among context topics ϕ_k (s) and content topics ψ_m (s). For example if the context information is timestamp, the model immediately yields the distribution over time for a topic, showing when the topic rises and falls. Figure 4 illustrates an example of a distribution over time for a content topic discovered from PNAS dataset where timestamp was used as context. This topic appears to capture a congenital disorder known as *Albinism.* This distribution illustrates research attention to this condition over the past 100 years from PNAS data. To seek evidence for this result, we search the term "Albinism" in Google Scholar, using the top 50 searching results and plot the histogram over time in the same figure. Surprisingly, we obtain a very close match between our results and the results from Google Scholar as evidenced in the figure.

Image Clustering with Image-Level Tags

We evaluate the clustering capacity of MC^2 using contexts

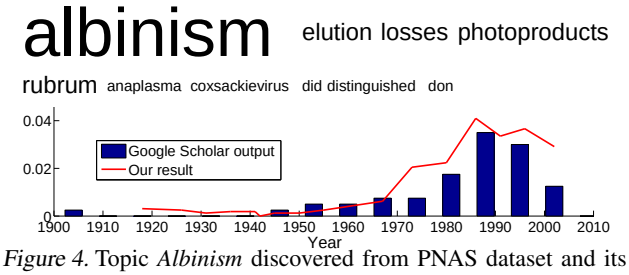

Figure 4. Topic *Albinism* discovered from PNAS dataset and its conditional distribution over time using our model; plotted together with results independently searched from Google Scholar using the top 50 hits.

Method	Perplexity	Feature used
HDP	175.62	SIFT
MC^2 without context	162.74	SIFT
MC^2 with context	**152.32**	Tags+SIFT

Table 2. NUS-WIDE dataset. Perplexity is evaluated on SIFT feature.

on an image clustering task. Our dataset is NUS-WIDE described earlier. We use bag-of-word SIFT features from each image for its content. Since each image in this dataset comes with a set of tags, we exploit them as context information, hence each context observation x_j is a bag-of-tag annotation vector.

First we perform the perplexity evaluation for this dataset using a similar setting as in the previous section. Table 2 presents the results where our model again outperforms HDP even when no context (tags) is used for training.

Next we evaluate the clustering quality of the model using the provided 13 classes as ground truth. We report performance on four well-known clustering evaluation metrics: Purity, Normalized Mutual Information (NMI), Rand-Index (RI), and Fscore (detailed in (Rand, 1971; Cai et al., 2011)). We use the following baselines for comparison:

- Kmeans and Non-negative Matrix Factorization (NMF)(Lee & Seung, 1999). For these methods, we need to specify the number of clusters in advance, hence we vary this number from 10 to 40. We then

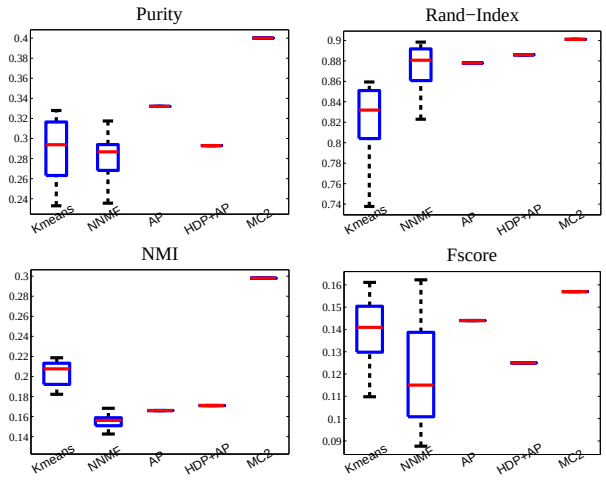

Figure 5. Clustering performance measured in purity, NMI, Rand-Index and F-score using NUS-WIDE dataset.

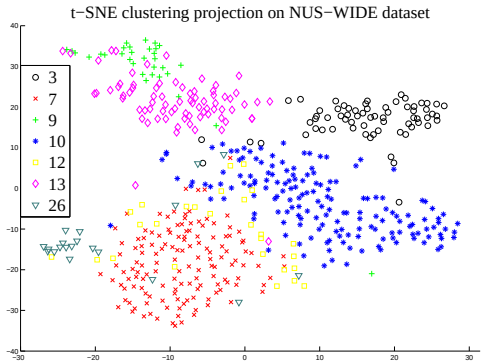

Figure 6. Projecting 7 discovered clusters (among 28) on 2D using t-SNE (Van der Maaten & Hinton, 2008).

report the min, max, mean and standard deviation.

- Affinity Propagation (AP) (Frey & Dueck, 2007): AP requires a similarity score between two documents and we use the Euclidean distance for this purpose.
- Hierarchical Dirichlet Process (HDP) + AP: we first run HDP using content observations, and then apply Affinity Propagation with similarity score derived from the symmetric KL divergence between the mixture proportions from two documents.

Figure 5 shows the result in which our model consistently delivers highest performance across all four metrics. For purity and NMI, our model beats all by a wide margin.

To gain some understanding on the clusters of images induced by our model, we run t-SNE (Van der Maaten & Hinton, 2008), projecting the feature vectors (both content and context) onto a 2D space. For visual clarity, we randomly select 7 out of 28 clusters and display in Figure 6 where it can be seen that they are reasonably well separated.

Effect of partially observed and missing data

Missing and unlabelled data is commonly encountered in practical applications. Here we examine the effect of context observability on document clustering performance. To do so, we again use the NUS-WIDE 13-animal subset as described previously, then vary the amount of observing context observation x_j with missing proportion ranges from 0% to 100%.

Missing (%)	Purity	NMI	RI	F-score
0 %	0.407	0.298	0.901	0.157
25 %	0.338	0.245	0.892	0.149
50 %	0.320	0.236	0.883	0.137
75 %	0.313	0.187	0.860	0.112
100 %	0.306	0.188	0.867	0.119

Table 3. Clustering performance with different missing proportion of context observation x_j.

Table 3 reports the result. We make two observations: a) utilizing context results in a big performance gain as evidenced in the difference between the top and bottom row of the table, and b) as the proportion of missing context starts to increase, the performance degrades gracefully up to 50% missing. This demonstrates the robustness of model against the possibility of missing context data.

5. Conclusion

We have introduced an approach for multilevel clustering when there are group-level context information. Our MC^2 provides a single joint model for utilizing group-level contexts to form group clusters while discovering the shared topics of the group contents at the same time. We provide a collapsed Gibbs sampling procedure and perform extensive experiments on three real-world datasets in both text and image domains. The experimental results using our model demonstrate the importance of utilizing context information in clustering both at the content and at the group level. Since similar types of contexts (time, tags, locations, ages, genres) are commonly encountered in many real-world data sources, we expect that our model will also be further applicable in other domains.

Our model contains a novel ingredient in DP-based Bayesian nonparametric modeling: we propose to use a base measure in the form of a product between a context-generating prior H and a content-generating prior $\mathrm{DP}(vQ_0)$. Doing this results in a new model with one marginal being the DPM and another marginal being the nDP mixture, thus establishing an interesting bridge between the DPM and the nDP. Our product base measure construction can be generalized to yield new models suitable for data presenting in more complicated nested group structures (e.g., more than 2-level deep).

References

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

Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003. ISSN 1533-7928.

Cai, Deng, He, Xiaofei, and Han, Jiawei. Locally consistent concept factorization for document clustering. *Knowledge and Data Engineering, IEEE Transactions on*, 23(6):902–913, 2011.

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

Dubey, Avinava, Hefny, Ahmed, Williamson, Sinead, and Xing, Eric P. A non-parametric mixture model for topic modeling over time. *arXiv preprint arXiv:1208.4411*, 2012.

Elango, Pradheep K and Jayaraman, Karthik. Clustering images using the latent dirichlet allocation model. *University of Wisconsin*, 2005.

Ferguson, T.S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

Frey, B.J. and Dueck, D. Clustering by passing messages between data points. *Science*, 315:972–976, 2007. doi: 10.1126/science.1136800.

Griffiths, Thomas L and Steyvers, Mark. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1): 5228–5235, 2004.

Hjort, N.L., Holmes, C., Müller, P., and Walker, S.G. *Bayesian nonparametrics*. Cambridge University Press, 2010.

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

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Leyland, Alastair H and Goldstein, Harvey. *Multilevel modelling of health statistics*. Wiley, 2001.

MacEachern, S.N. Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pp. 50–55, 1999.

Nguyen, T.C., Phung, D., Gupta, S., and Venkatesh, S. Extraction of latent patterns and contexts from social honest signals using hierarchical dirichlet processes. *IEEE International Conference on Pervasive Computing and Communications*, 2013.

Nguyen, V., Phung, D., Nguyen, X., Venkatesh, S., and Bui, H. Bayesian nonparametric multilevel clustering with group-level contexts. Technical report, 2014. URL http://arxiv.org/abs/1401.1974.

Phung, D., Nguyen, X., Bui, H., Nguyen, T.V., and Venkatesh, S. Conditionally dependent Dirichlet processes for modelling naturally correlated data sources. Technical report, Pattern Recognition and Data Analytics, Deakin University, 2012.

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. ISSN 1469-2163.

Rand, William M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

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

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

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

Van der Maaten, Laurens and Hinton, Geoffrey. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

Wulsin, D., Jensen, S., and Litt, B. A hierarchical dirichlet process model with multiple levels of clustering for human eeg seizure modeling. *Proceedings of the 29th International Conference on Machine learning*, 2012.

Xie, Pengtao and Xing, Eric P. Integrating document clustering and topic modeling. 2013.