Robust Bayesian Max-Margin Clustering

Changyou Chen[†] **Jun Zhu**[‡] **Xinhua Zhang**[♯]

[†]Dept. of Electrical and Computer Engineering, Duke University, Durham, NC, USA

[‡]State Key Lab of Intelligent Technology & Systems; Tsinghua National TNList Lab;

[‡]Dept. of Computer Science & Tech., Tsinghua University, Beijing 100084, China

[♯]Australian National University (ANU) and National ICT Australia (NICTA), Canberra, Australia

cchangyou@gmail.com; dcszj@tsinghua.edu.cn; xinhua.zhang@anu.edu.au

Abstract

We present max-margin Bayesian clustering (BMC), a general and robust framework that incorporates the max-margin criterion into Bayesian clustering models, as well as two concrete models of BMC to demonstrate its flexibility and effectiveness in dealing with different clustering tasks. The *Dirichlet process max-margin Gaussian mixture* is a nonparametric Bayesian clustering model that relaxes the underlying Gaussian assumption of Dirichlet process Gaussian mixtures by incorporating max-margin posterior constraints, and is able to infer the number of clusters from data. We further extend the ideas to present *max-margin clustering topic model*, which can learn the latent topic representation of each document while at the same time cluster documents in the max-margin fashion. Extensive experiments are performed on a number of real datasets, and the results indicate superior clustering performance of our methods compared to related baselines.

1 Introduction

Existing clustering methods fall roughly into two categories. Deterministic clustering directly optimises some loss functions, while Bayesian clustering models the data generating process and infers the clustering structure via Bayes rule. Typical deterministic methods include the well known kmeans [1], nCut [2], support vector clustering [3], Bregman divergence clustering [4, 5], and the methods built on the very effective max-margin principle [6–9]. Although these methods can flexibly incorporate constraints for better performance, it is challenging for them to finely capture hidden regularities in the data, e.g., automated inference of the number of clusters and the hierarchies underlying the clusters. In contrast, Bayesian clustering provides favourable convenience in modelling latent structures, and their posterior distributions can be inferred in a principled fashion. For example, by defining a Dirichlet process (DP) prior on the mixing probability of Gaussian mixtures, *Dirichlet process Gaussian mixture models* [10] (DPGMM) can infer the number of clusters in the dataset. Other priors on latent structures include the hierarchical cluster structure [11–13], co-clustering structure [14], *etc*. However, Bayesian clustering is typically difficult to accommodate external constraints such as max-margin. This is because under the standard Bayesian inference designing some informative priors (if any) that satisfy these constraints is highly challenging.

To address this issue, we propose Bayesian max-margin clustering (BMC), which allows max-margin constraints to be flexibly incorporated into a Bayesian clustering model. Distinct from the traditional max-margin clustering, BMC is fully Bayesian and enables probabilistic inference of the number of clusters or the latent feature representations of data. Technically, BMC leverages the regularized Bayesian inference (RegBayes) principle [15], which has shown promise on *supervised* learning tasks, such as classification [16, 17], link prediction [18], and matrix factorisation [19], where max-margin constraints are introduced to improve the discriminative power of a Bayesian model. However, little exploration has been devoted to the *unsupervised* setting, due in part to the absence of true labels that makes it technically challenging to enforce max-margin constraints. BMC constitutes a first extension of RegBayes to the unsupervised clustering task. Note that distinct from the clustering models using maximum entropy principle [20, 21] or posterior regularisation [22], BMC is more general due to the intrinsic generality of RegBayes [15].

We demonstrate the flexibility and effectiveness of BMC by two concrete instantiations. The first is *Dirichlet process max-margin Gaussian mixture* (DPMMGM), a nonparametric Bayesian clustering model that relaxes the Gaussian assumption underlying DPGMM by incorporating max-margin constraints, and is able to infer the number of clusters in the raw input space. To further discover latent feature representations, we propose the *max-margin clustering topic model* (MMCTM). As a topic model, it performs max-margin clustering of documents, while at the same time learns the latent topic representation for each document. For both DPMMGM and MMCTM, we develop efficient MCMC algorithms by exploiting data augmentation techniques. This avoids imposing restrictive assumptions such as in variational Bayes, thereby facilitating the inference of the true posterior. Extensive experiments demonstrate superior clustering performance of BMC over various competitors.

2 Regularized Bayesian Inference

We first briefly overview the principle of regularised Bayesian inference (RegBayes) [15].

The motivation of RegBayes is to enrich the posterior of a probabilistic model by incorporating additional constraints, under an information-theoretical optimisation formulation. Formally, suppose a probabilistic model has latent variables Θ, endowed with a prior $p(\Theta)$ (examples of Θ will be clear soon later). We also have observations $\mathbf{X} := \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$, with $\mathbf{x}_i \in \mathbb{R}^p$. Let $p(\mathbf{X}|\Theta)$ be the likelihood. Then, posterior inference via the Bayes' theorem is equivalent to solving the following optimisation problem [15]:

$$\inf_{q(\Theta)\in\mathcal{P}} \mathrm{KL}(q(\Theta) \,||\, p(\Theta)) - \mathbb{E}_{\Theta\sim q(\Theta)}\left[\log p(\mathbf{X}|\Theta)\right] \tag{1}$$

where $\mathcal{P}$ is the space of probability distribution[1], $q(\Theta)$ is the required posterior (here and afterwards we will drop the dependency on $\mathbf{X}$ for notation simplicity). In other words, the Bayesian posterior $p(\Theta|\mathbf{X})$ is identical to the optimal solution to (1). The power of RegBayes stems in part from the flexibility of engineering $\mathcal{P}$, which typically encodes constraints imposed on $q(\Theta)$, *e.g.*, via expectations of some feature functions of Θ (and possibly the data $\mathbf{X}$). Furthermore, the constraints can be parameterised by some auxiliary variable $\boldsymbol{\xi}$. For example, $\boldsymbol{\xi}$ may quantify the extent to which the constraints are violated, then it is penalised in the objective through a function U. To summarise, RegBayes can be generally formulated as

$$\begin{aligned} &\inf_{\boldsymbol{\xi}, q(\Theta)} \mathrm{KL}(q(\Theta) \,||\, p(\Theta)) - \mathbb{E}_{\Theta\sim q(\Theta)}\left[\log p(\mathbf{X}|\Theta)\right] + U(\boldsymbol{\xi}) \\ &\text{s.t.} \quad q(\Theta) \in \mathcal{P}(\boldsymbol{\xi}). \end{aligned} \tag{2}$$

To distinguish from the standard Bayesian posterior, the optimal $q(\Theta)$ is called *post-data posterior*. Under mild regularity conditions, RegBayes admits a generic representation theorem to characterise the solution $q(\Theta)$ [15]. It is also shown to be more general than the conventional Bayesian methods, including those methods that introduce constraints on a prior. Such generality is essential for us to develop a Bayesian framework of max-margin clustering. Note that like many sophisticated Bayesian models, posterior inference remains as a key challenge of developing novel RegBayes models. Therefore, one of our key technical contributions is on developing efficient and accurate algorithms for BMC, as detailed below.

3 Robust Bayesian Max-margin Clustering

For clustering, one key assumption of our model is that $\mathbf{X}$ forms a latent cluster structure. In particular, let each cluster be associated with a latent projector $\boldsymbol{\eta}_k \in \mathbb{R}^p$, which is included in Θ and has prior distribution subsumed in $p(\Theta)$. Given any distribution q on Θ, we then define the compatibility score of $\mathbf{x}_i$ with respect to cluster k by using the marginal distribution on $\boldsymbol{\eta}_k$ (as $\boldsymbol{\eta}_k \in \Theta$):

$$F_k(\mathbf{x}_i) = \mathbb{E}_{q(\boldsymbol{\eta}_k)}\left[\boldsymbol{\eta}_k^T \mathbf{x}_i\right] = \mathbb{E}_{q(\Theta)}\left[\boldsymbol{\eta}_k^T \mathbf{x}_i\right] . \tag{3}$$

[1] In theory, we also require that q is absolutely continuous with respect to p to make the KL-divergence well defined. The present paper treats this constraint as an implicit assumption for clarity.

For each example $\mathbf{x}_i$, we introduce a random variable y_i valued in $\mathbb{Z}^+$, which denotes its cluster assignment and is also included in Θ. Inspired by conventional multiclass SVM [7, 23], we utilize $\mathcal{P}(\boldsymbol{\xi})$ in RegBayes (2) to encode the max-margin constraints based on $F_k(\mathbf{x}_i)$, with the slack variable $\boldsymbol{\xi}$ penalised via their sum in $U(\boldsymbol{\xi})$. This amounts to our Bayesian max-margin clustering (BMC):

$$\begin{aligned} &\inf_{\xi_i \geq 0, q(\Theta)} \mathcal{L}(q(\Theta)) + 2c \sum_i \xi_i \\ &\text{s.t.} \quad F_{y_i}(\mathbf{x}_i) - F_k(\mathbf{x}_i) \geq \ell\, \mathbb{I}(y_i \neq k) - \xi_i, \quad \forall i, k \end{aligned} \tag{4}$$

where $\mathcal{L}(q(\Theta)) = \mathrm{KL}(q(\Theta)||p(\Theta)) - \mathbb{E}_{\Theta \sim q(\Theta)}[\log p(\mathbf{X}|\Theta)]$ measures the KL divergence between q and the original Bayesian posterior $p(\Theta|\mathbf{X})$ (up to a constant); $\mathbb{I}(\cdot) = 1$ if $\cdot$ holds true, and 0 otherwise; $\ell > 0$ is a constant scalar of margin. Note we found that the commonly adopted balance constraints in max-margin clustering models [6] either are unnecessary or do not help in our framework. We will address this issue in specific models.

Clearly by absorbing the slack variables $\boldsymbol{\xi}$, the optimisation problem (4) is equivalent to

$$\inf_{q(\Theta)} \mathcal{L}(q(\Theta)) + 2c \sum_i \max\left(0, \max_{k:k\neq y_i} \mathbb{E}_{\Theta\sim q(\Theta)}[\zeta_{ik}]\right) \tag{5}$$

where $\zeta_{ik} := \ell\, \mathbb{I}(y_i \neq k) - (\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_k)^T \mathbf{x}_i$. Exact solution to (5) is hard to compute. An alternative approach is to approximate the posterior by assuming independence between random variables, *e.g.* variational inference. However, this is usually slow and susceptible to local optimal. In order to obtain an analytic optimal distribution q that facilitates efficient Bayesian inference, we resort to the technique of Gibbs classifier [17] which approximates (in fact, upper bounds due to the convexity of max function) the second term in (5) by an expected hinge loss, *i.e.*, moving the expectation out of the max. This leads to our final formulation of BMC:

$$\inf_{q(\Theta)} \mathcal{L}(q(\Theta)) + 2c \sum_i \mathbb{E}_{\Theta\sim q(\Theta)}\left[\max\left(0, \max_{k:k\neq y_i} \zeta_{ik}\right)\right]. \tag{6}$$

Problem (6) is still much more challenging than existing RegBayes models [17], which are restricted to supervised learning with two classes only. Specifically, BMC allows multiple clusters/classes in an unsupervised setting, and the latent cluster membership y_i needs to be inferred. This complicates the model and brings challenges for posterior inference, as addressed below. In a nutshell, our inference algorithms rely on two key steps by exploring data augmentation techniques. First, in order to tackle the multi-class case, we introduce auxiliary variables $s_i := \arg\max_{k:k\neq y_i} \zeta_{ik}$. Applying standard derivations in calculus of variation [24] and augmenting the model with $\{s_i\}$, we obtain an analytic form of the optimal solution to (6) by augmenting Θ (refer to Appendix A for details):

$$q(\Theta, \{s_i\}) \propto p(\Theta|\mathbf{X}) \prod_i \exp(-2c \max(0, \zeta_{is_i}))\,. \tag{7}$$

Second, since the $\max$ term in (7) obfuscates efficient sampling, we apply the augmentation technique introduced by [17], which showed that $q(\Theta, \{s_i\})$ is identical to the marginal distribution of the *augmented* post-data posterior

$$q(\Theta, \{s_i\}, \{\lambda_i\}) \propto p(\Theta|\mathbf{X}) \prod_i \tilde{\phi}_i(\lambda_i|\Theta), \tag{8}$$

where $\tilde{\phi}_i(\lambda_i|\Theta) := \lambda_i^{-\frac{1}{2}} \exp\left(\frac{-1}{2\lambda_i}(\lambda_i + c\zeta_{is_i})^2\right)$. Here λ_i is an augmented variable for $\mathbf{x}_i$ that has an generalised inverse Gaussian distribution [25] given Θ and $\mathbf{x}_i$.

Note that our two steps of data augmentation are exact and incur no approximation. With the augmented variables $(\{s_i\}, \{\lambda_i\})$, we can develop efficient sampling algorithms for the augmented posterior $q(\Theta, \{s_i\}, \{\lambda_i\})$ without restrictive assumptions, thereby allowing us to approach the true target posterior $q(\Theta)$ by dropping the augmented variables. The details will become clear soon in our subsequent clustering models.

4 Dirichlet Process Max-margin Gaussian Mixture Models

In (4), we have left unspecified the prior $p(\Theta)$ and the likelihood $p(\mathbf{X}|\Theta)$. This section presents an instantiation of Bayesian nonparametric clustering for non-Gaussian data. We will present another instantiation of max-margin document clustering based on topic models in next section.

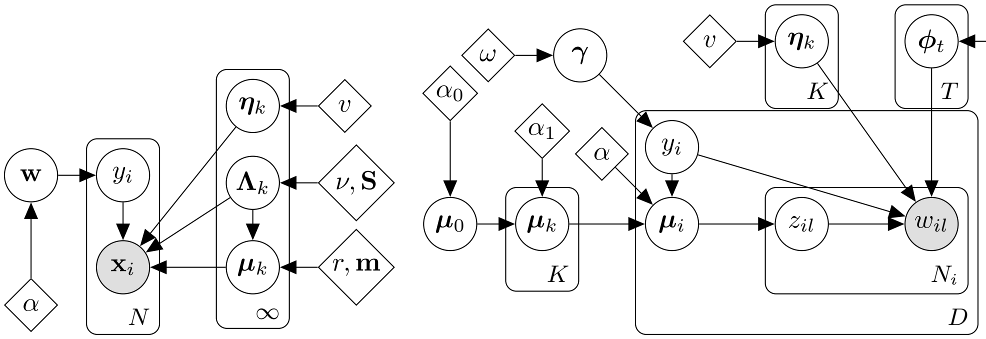


Figure 1: Left: Graphical model of DPMMGM. The part excluding $\boldsymbol{\eta}_k$ and v corresponds to DPGMM. Right: Graphical model of MMCTM. The one excluding $\{\boldsymbol{\eta}_k\}$ and the arrow between y_i and w_{il} corresponds to CTM.

Here a convenient model of $p(\mathbf{X}, \Theta)$ is mixture of Gaussian. Let the mean and variance of the k-th cluster component be $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$. In a nonparametric setting, the number of clusters is allowed to be infinite, and the cluster y_i that each data point belongs to is drawn from a Dirichlet process [10]. To summarize, the latent variables are $\Theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\eta}_k\}_{k=1}^{\infty} \cup \{y_i\}_{i=1}^{n}$. The prior $p(\Theta)$ is specified as: $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$ employ a standard *Normal-inverse Wishart* prior [26]:

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_k; \mathbf{m}, (r\boldsymbol{\Lambda}_k)^{-1}), \text{ and } \boldsymbol{\Lambda}_k \sim \mathcal{IW}(\boldsymbol{\Lambda}_k; \mathbf{S}, \nu). \tag{9}$$

$y_i \in \mathbb{Z}^+$ has a Dirichlet process prior with parameter α. $\boldsymbol{\eta}_k$ follows a normal prior with mean $\mathbf{0}$ and variance $v\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. The likelihood $p(\mathbf{x}_i|\Theta)$ is $\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_{y_i}, (r\boldsymbol{\Lambda}_{y_i})^{-1})$, *i.e.* independent of $\boldsymbol{\eta}_k$. The max-margin constraints take effects in the model via $\tilde{\phi}_i$'s in (8). Note this model of $p(\Theta, \mathbf{X})$, apart from $\boldsymbol{\eta}_k$, is effectively the Dirichlet process Gaussian mixture model [10] (DPGMM). Therefore, we call our post-data posterior $q(\Theta, \{s_i\}, \{\lambda_i\})$ in (8) as *Dirichlet process max-margin Gaussian mixture* model (DPMMGM). The hyperparameters include $\mathbf{m}$, r, $\mathbf{S}$, ν, α, v.

Interpretation as a generalised DP mixture The formula of the augmented post-data posterior in (8) reveals that, compared with DPGMM, each data point is associated with an additional factor $\tilde{\phi}_i(\lambda_i|\Theta)$. Thus we can interpret DPMMGM as a generalised DP mixture with *Normal-inversed Wishart-Normal* as the base distribution, and a generalised *pseudo* likelihood that is proportional to

$$f(\mathbf{x}_i, \lambda_i | y_i, \boldsymbol{\mu}_{y_i}, \boldsymbol{\Lambda}_{y_i}, \{\boldsymbol{\eta}_k\}) := \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_{y_i}, (r\boldsymbol{\Lambda}_{y_i})^{-1}) \tilde{\phi}_i\left(\lambda_i|\Theta\right) \ . \tag{10}$$

To summarise, DPMMGM employs the following generative process with the graphical model shown in Fig. 1 (left):

$$\begin{aligned}
(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\eta}_k) &\sim \mathcal{N}\left(\boldsymbol{\mu}_k; \mathbf{m}, (r\boldsymbol{\Lambda}_k)^{-1}\right) \times \mathcal{IW}\left(\boldsymbol{\Lambda}_k; \mathbf{S}, \nu\right) \times \mathcal{N}\left(\boldsymbol{\eta}_k; 0, v\mathbf{I}\right), \quad k = 1, 2, \cdots \\
\mathbf{w} &\sim \text{Stick-Breaking}(\alpha), \\
y_i|\mathbf{w} &\sim \text{Discrete}(\mathbf{w}), \qquad\qquad\qquad\qquad\qquad\qquad i \in [n] \\
(\mathbf{x}_i, \lambda_i)|y_i, \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\eta}_k\} &\simeq f(\mathbf{x}_i, \lambda_i | y_i, \boldsymbol{\mu}_{y_i}, \boldsymbol{\Lambda}_{y_i}, \{\boldsymbol{\eta}_k\}). \qquad\qquad i \in [n]
\end{aligned}$$

Here $[n] := \{1, \cdots, n\}$ is the set of integers up to n and $\simeq$ means that $(\mathbf{x}_i, \lambda_i)$ is generative from a distribution that is *proportional* to $f(\cdot)$. Since this normalisation constant is shared by all samples $\mathbf{x}_i$, there is no need to deal with it by posterior inference. Another benefit of this interpretation is that it allows us to use existing techniques for non-conjugate DP mixtures to sample the cluster indicators y_i efficiently, and to infer the number of clusters in the data. This approach is different from previous work on RegBayes nonparametric models where truncated approximation is used to deal with the infinite dimensional model space [15, 18]. In contrast, our method does not rely on any approximation. Note that DPMMGM does not need the complicated class balance constraints [6] because the Gaussians in the pseudo likelihood would balance the clusters to some extent.

Posterior inference Posterior inference for DPMMGM can be done by efficient Gibbs sampling. We integrate out the infinite dimension vector $\mathbf{w}$, so the variables needed to be sampled are $\{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\eta}_k\}_k \cup \{y_i, s_i, \lambda_i\}_i$. Conditional distributions are derived in Appendix B. Note that we use an extension of the *Reused Algorithm* [27] to jointly sample (y_i, s_i), which allows it to allocate to empty clusters in Bayesian nonparametric setting. The time complexity is almost the same as DPGMM except for the additional step to sample $\boldsymbol{\eta}_k$, with cost $O(p^3)$. So it would be necessary to put the constraints on a subspace (*e.g.*, by projection) of the original feature space when p is high.

5 Max-margin Clustering Topic Model

Although many applications exhibit clustering structures in the raw observed data which can be effectively captured by DPMMGM, it is common that such regularities are more salient in terms of some high-level but latent features. For example, topic distributions are often more useful than word frequency in the task of document clustering. Therefore, we develop a *max-margin clustering topic model* (MMCTM) in the framework of BMC, which allows topic discovery to co-occur with document clustering in a Bayesian and max-margin fashion. To this end, the latent Dirichlet allocation (LDA) [28] needs to be extended by introducing a *cluster label* into the model, and define each cluster as a mixture of topic distributions. This cluster-based topic model [29] (CTM) can then be used in concert with BMC to enforce large margin between clusters in the posterior $q(\Theta)$.

Let V be the size of the word vocabulary, T be the number of topics, and K be the number of clusters, $\mathbf{1}_N$ be a N-dimensional one vector. Then the generative process of CTM for the documents goes as:
1. For each topic t, generate its word distribution $\boldsymbol{\phi}_t$: $\boldsymbol{\phi}_t|\beta \sim \text{Dir}(\beta\mathbf{1}_V)$.

2. Draw a base topic distribution $\boldsymbol{\mu}_0$: $\boldsymbol{\mu}_0|\alpha_0 \sim \text{Dir}(\alpha_0\mathbf{1}_T)$. Then for each cluster k, generate its topic distribution mixture $\boldsymbol{\mu}_k$: $\boldsymbol{\mu}_k|\alpha_1, \boldsymbol{\mu}_0 \sim \text{Dir}(\alpha_1\boldsymbol{\mu}_0)$.

3. Draw a base cluster distribution $\boldsymbol{\gamma}$: $\boldsymbol{\gamma}|\omega \sim \text{Dir}(\omega\mathbf{1}_K)$. Then for each document $i \in [D]$:

- Generate a *cluster label* y_i and a *topic distribution* $\boldsymbol{\mu}_i$: $y_i|\boldsymbol{\gamma} \sim \text{Discrete}(\boldsymbol{\gamma})$, $\boldsymbol{\mu}_i|\alpha, \boldsymbol{\mu}_{y_i} \sim \text{Dir}(\alpha\boldsymbol{\mu}_{y_i})$.
- Generate the observed words w_{il}: $z_{il} \sim \text{Discrete}(\boldsymbol{\mu}_i)$, $\quad w_{il} \sim \text{Discrete}(\boldsymbol{\phi}_{z_{il}}), \forall\, l \in [N_i]$.

Fig. 1 (right) shows the structure. We then augment CTM with max-margin constraints, and get the same posterior as in Eq. (7), with the variables Θ corresponding to $\{\boldsymbol{\phi}_t\}_{t=1}^T \cup \{\boldsymbol{\eta}_k, \boldsymbol{\mu}_k\}_{k=1}^K \cup \{\boldsymbol{\mu}_\mathbf{0}, \boldsymbol{\gamma}\} \cup \{\boldsymbol{\mu}_i, y_i\}_{i=1}^D \cup \{z_{il}\}_{i=1,l=1}^{D,N_i}$.

Compared with the raw word space which is normally extremely high-dimensional and sparse, it is more reasonable to characterise the clustering structure in the latent feature space–the empirical latent topic distributions as in the MedLDA [16]. Specifically, we summarise the topic distribution of document i by $\mathbf{x}_i \in \mathbb{R}^T$, whose t-th element is $\frac{1}{N_i}\sum_{l=1}^{N_i} \mathbb{I}(z_{il} = t)$. Then the compatibility score for document i with respect to cluster k is defined similar to (3) as $F_k(\mathbf{x}_i) = \mathbb{E}_{q(\Theta)}\left[\boldsymbol{\eta}_k^T\mathbf{x}_i\right]$. Note, however, the expectation is also taken over $\mathbf{x}_i$ since it is *not* observed.

Posterior inference To achieve fast mixing, we integrate out $\{\boldsymbol{\phi}_t\}_{t=1}^T \cup \{\boldsymbol{\mu}_0, \boldsymbol{\gamma}\} \cup \{\boldsymbol{\mu}_k\}_{k=1}^K \cup \{\boldsymbol{\mu}_i\}_{i=1}^D$ in the posterior, thus $\Theta = \{y_i\}_{i=1}^D \cup \{\boldsymbol{\eta}_k\}_{k=1}^K \cup \{z_{il}\}_{i=1,l=1}^{D,N_i}$. The integration is straightforward by the Dirichlet-Multinomial conjugacy. The detailed form of the posterior and the conditional distributions are derived in Appendix C. By extending CTM with max-margin, we note that many of the the sampling formulas are extension of those in CTM [29], with additional sampling for $\boldsymbol{\eta}_k$, thus the sampling can be done fairly efficiently.

Dealing with vacuous solutions Different from DPMMGM, the max-margin constraints in MMCTM do not interact with the observed words w_{il}, but with the latent topic representations $\mathbf{x}_i$ (or z_{il}) that are also inferred from the model. This easily makes the latent representation z_i's collapse into a single cluster, a vacuous solution plaguing many other unsupervised learning methods as well. One remedy is to incorporate the cluster balance constraints into the model [7]. However, this does not help in our Bayesian setting because apart from significant increase in computational cost, MCMC often fails to converge in practice[2]. Another solution is to morph the problem into a weakly semi-supervised setting, where we assign to each cluster a few documents according to their true label (we will refer to these documents as *landmarks*), and sample the rest as in the above unsupervised setting. These "labeled examples" can be considered as introducing constraints that are alternative to the balance constraints. Usually only a very small number of labeled documents are needed, thus barely increasing the cost in training and labelling. We will focus on this setting in experiment.

6 Experiments

6.1 Dirichlet Process Max-margin Gaussian Mixture

[2] We observed the cluster sizes kept bouncing with sampling iterations. This is probably due to the highly nonlinear mapping from observed word space to the feature space (topic distribution), making the problem multi-modal, *i.e.*, there are multiple optimal topic assignments in the post-data posterior (8). Also the balance constraints might weaken the max-margin constraints too much.

We first show the distinction between our DPMMGM and DPGMM by running both models on the non-Gaussian *half-rings* data set [30]. There are a number of hyperparameters to be determined, *e.g.*, $(\alpha, r, S, \nu, v, c, \ell)$; see Section 4. It turns out the cluster structure is insensitive to (α, r, S, ν), and so we use a standard sampling method to update α [31], while r, ν, S are sampled by employing Gamma, truncated Poisson, inverse Wishart priors respectively, as is done in [32]. We set $v = 0.01, c = 0.1, \ell = 5$ in this experiment. Note that the clustering structure is sensitive to the values of c and ℓ, which will be studied below. Empirically we find that DPMMGM converges much faster than DPGMM, both converging well within 200 iterations (see Appendix D.4 for examples). In Fig. 2, the clustering structures demonstrate clearly that DPMMGM relaxes the Gaussian assumption of the data distribution, and correctly finds the number of clusters based on the margin boundary, whereas DPGMM produces a too fragmented partition of the data for the clustering task.

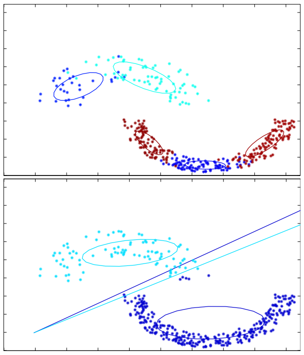
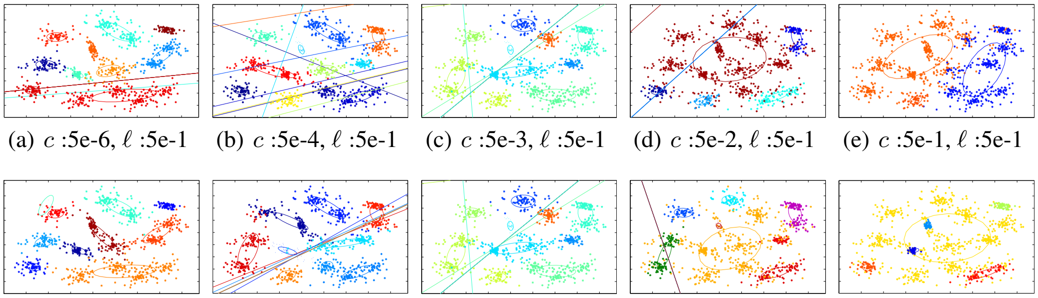


Figure 2: An illustration of DPGMM (up) and DPMMGM (bottom).

Parameter sensitivity We next study the sensitivity of hyperparameters c and ℓ, with other hyperparameters sampled during inference as above. Intuitively the impact of these parameters is as follows. c controls the weight that the max-margin constraint places on the posterior. If there were no other constraint, the max-margin constraint would drive the data points to collapse into a single cluster. As a result, we expect that a larger value of the weight c will result in fewer clusters. Similarly, increasing the value of ℓ will lead to a higher loss for any violation of the constraints, thus driving the data points to collapse as well. To test these implications, we run DPMMGM on a 2-dimensional synthetic dataset with 15 clusters [33]. We vary c and ℓ to study how the cluster structures change with respect to these parameter settings. As can be observed from Fig. 3, the results indeed follow our intuition, providing a mean to control the cluster structure in applications.



(a) c :5e-6, ℓ :5e-1 (b) c :5e-4, ℓ :5e-1 (c) c :5e-3, ℓ :5e-1 (d) c :5e-2, ℓ :5e-1 (e) c :5e-1, ℓ :5e-1

(f) c :5e-3, ℓ :5e-4 (g) c :5e-3, ℓ :5e-2 (h) c :5e-3, ℓ :5e-1 (i) c :5e-3, ℓ :2 (j) c :5e-3, ℓ :5

Figure 3: Clustering structures with varied ℓ and c: (first row) fixed ℓ and increasing c; (second row) fixed c and increasing ℓ. Lines are η's. Clearly the number cluster decreases with growing c and ℓ.

Real Datasets. As other clustering models, we test DPMMGM on ten real datasets (small to moderate sizes) from the UCI repository [34]. Scaling up to large dataset is an interesting future. The first three columns of Table 1 list some of the statistics of these datasets (we used random subsets of the three large datasets – Letter, MNIST, and Segmentation).

A heuristic approach for model selection. Model selection is generally hard for unsupervised clustering. Most existing algorithms simply fix the hyperparameters without examining their impacts on model performance [10, 35]. In DPMMGM, the hyperparameters c and ℓ are critical to clustering quality since they control the number of clusters. Without training data in our setting they can not be set using cross validation. Moreover, they are not feasible to be estimated use Bayesian sampling as well because they are not parameters from a proper Bayesian model. we thus introduce a time-efficient heuristic approach to selecting appropriate values. Suppose the dataset is known to have K clusters. Our heuristic goes as follows. First initialise c and ℓ to 0.1. Then at each iteration, we compare the inferred number of clusters with K. If it is larger than K (otherwise we do the converse), we choose c or ℓ randomly, and increase its value by $\frac{u}{n}$, where u is a uniform random variable in $[0, 1]$ and n is the number of iterations so far. According to the *parameter sensitivity* studied above, increasing c or ℓ tends to decrease the number of clusters, and the model eventually

Dataset	Data property			NMI				
	n	p	K	kmeans	nCut	DPGMM	DPMMGM	DPMMGM*
Glass	214	10	7	0.37±0.04	0.22±0.00	0.37±0.05	**0.46**±0.01	0.45±0.01
Half_circle	300	2	2	0.43±0.00	1.00±0.00	0.49±0.02	0.67±0.02	0.51±0.07
Iris	150	4	3	0.72±0.08	0.61±0.00	0.73±0.00	**0.73**±0.00	0.73±0.00
Letter	1000	16	10	0.33±0.01	0.04±0.00	0.19±0.09	**0.38**±0.04	0.23±0.04
MNIST	1000	784	10	0.50±0.01	0.38±0.00	0.55±0.03	**0.56**±0.01	0.55±0.02
Satimage	4435	36	6	0.57±0.06	0.55±0.00	0.21±0.05	0.51±0.01	0.30±0.00
Segment'n	1000	19	7	0.52±0.03	0.34±0.00	0.23±0.09	**0.61**±0.05	0.52±0.10
Vehicle	846	18	4	0.10±0.00	0.14±0.00	0.02±0.02	**0.14**±0.00	0.05±0.00
Vowel	990	10	11	0.42±0.01	0.44±0.00	0.28±0.03	0.39±0.02	0.41±0.02
Wine	178	13	3	0.84±0.01	0.46±0.00	0.56±0.02	**0.90**±0.02	0.59±0.01

Table 1: Comparison for different methods on NMI scores. K: true number of clusters.

stabilises due to the stochastic decrement by $\frac{u}{n}$. We denote the model learned from this heuristic as DPMMGM. In the case where the true number of clusters is unknown, we can still apply this strategy, except that the number of clusters K needs to be first inferred from DPGMM. This method is denoted as DPMMGM*.

Comparison. We measure the quality of clustering results by using the standard normalised mutual information (NMI) criterion [36]. We compare our DPMMGM with the well established KMeans, nCut and DPGMM clustering methods[3]. All experiments are repeated for five times with random initialisation. The results are shown in Table 1. Clearly DPMMGM significantly outperforms other models, achieving the best NMI scores. DPMMGM*, which is not informed of the true number of clusters, still obtains reasonably high NMI scores, and outperforms the DPGMM model.

6.2 Max-margin Clustering Topic Model

Datasets. We test the MMCTM model on two document datasets: *20NEWS* and *Reuters-R8* . For the *20NEWS* dataset, we combine the training and test datasets used in [16], which ends up with 20 categories/clusters with roughly balanced cluster sizes. It contains 18,772 documents in total with a vocabulary size of 61,188. The *Reuters-R8* dataset is a subset of the Reuters-21578 dataset[4], with of 8 categories and 7,674 documents in total. The size of different categories is biased, with the lowest number of documents in a category being 51 while the highest being 2,292.

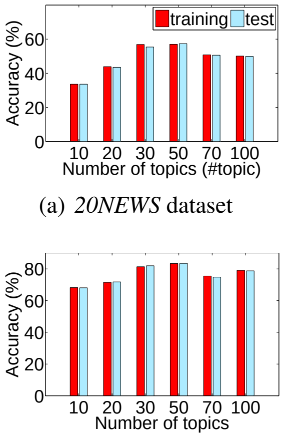
Comparison We choose $L \in \{5, 10, 15, 20, 25\}$ documents randomly from each category as the *landmarks*, use 80% documents for training and the rest for testing. We set the number of topics (i.e., T) to 50, and set the Dirichlet prior in Section 5 to $\omega = 0.1$, $\beta = 0.01$, $\alpha = \alpha_0 = \alpha_1$ $= 10$, as clustering quality is not sensitive to them. For the other hyperparameters related to the max-margin constraints, *e.g.*, v in the Gaussian prior for $\boldsymbol{\eta}$, the balance parameter c, and the cost parameter ℓ, instead of doing cross validation which is computationally expensive and not helpful for our scenario with few labeled data, we simply set $v = 0.1, c = 9, \ell = 0.1$. This is found to be a good setting and denoted as MMCTM. To test the robustness of this setting, we vary c over $\{0.1, 0.2, 0.5, 0.7, 1, 3, 5, 7, 9, 15, 30, 50\}$ and keep $v = \ell = 0.1$ (ℓ and c play similar roles and so varying one is enough). We choose the best performance out of these parameter settings, denoted as MMCTM*, which can be roughly deemed as the setting for the optimal performance. We compared MMCTM with state-of-the-art SVM and semi-supervised SVM (S3VM) models. They are efficiently implemented in [37], and the related parameters are chosen by 5-fold cross validation. As in [16], raw word frequencies are used as input features. We also compare MMCTM with a Bayesian baseline–cluster based topic model (CTM) [29], the building block of MMCTM without the max-margin constraints. Note we did not compare with the standard MedLDA [16] because it is supervised. We measure the performance by cluster accuracy, which is the proportion of correctly clustered documents. To accelerate MMCTM, we simply initialise it with CTM, and find it converges surprisingly fast in term of accuracy, *e.g.*, usually within 30 iterations (refer to Appendix

[3] We additionally show some comparison with some existing max-margin clustering models in Appendix D.2 on two-cluster data because their code only deals with the case of two clusters. Our method performs best.

[4] Downloaded from csmining.org/index.php/r52-and-r8-of-reuters-21578.html.

L	CTM	SVM	S3VM	MMCTM	MMCTM*
			20NEWS		
5	17.22± 4.2	37.13± 2.9	39.36± 3.2	**56.70**± **1.9**	**57.86**± **0.9**
10	24.50± 4.5	46.99± 2.4	47.91± 2.8	**54.92**± **1.6**	**56.56**± **1.3**
15	22.76± 4.2	52.80± 1.2	52.49± 1.4	55.06± 2.7	**57.80**± **2.2**
20	26.07± 7.2	56.10± 1.5	54.44± 2.1	56.62± 2.2	**59.70**± **1.4**
25	27.20± 1.5	59.15± 1.4	57.45± 1.7	55.70± 2.4	**61.92**± **3.0**
			Reuters-R8		
5	41.27± 16.7	78.12± 1.1	78.51± 2.3	79.18± 4.1	80.86± 2.9
10	42.63± 7.4	80.69± 1.2	79.15± 1.2	80.04± 5.3	**83.48**± **1.0**
15	39.67± 9.9	83.25± 1.7	81.87± 0.8	85.48± 2.1	86.86± 2.5
20	58.24± 8.3	85.66± 1.0	73.95± 2.0	82.92± 1.7	83.82± 1.6
25	51.93± 5.9	84.95± 0.1	82.39± 1.8	86.56± 2.5	**88.12**± **0.5**

Table 2: Clustering acc. (in %). **Bold** means significantly different.



(a) *20NEWS* dataset

(b) *Reuters-R8* dataset

Figure 4: Accuracy vs. #topic

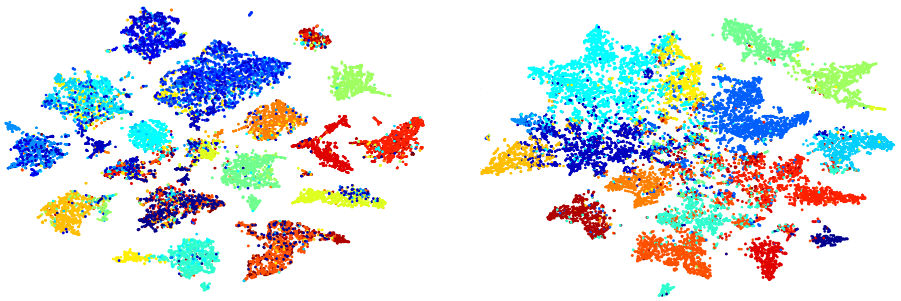


Figure 5: 2-D tSNE embedding on *20NEWS* for MMCTM (left) and CTM (right). Best viewed in color. See Appendix D.3 for the results on *Reuters-R8* datasets.

D.5). The accuracies are shown in Table 2, and we can see that MMCTM outperforms other models (also see Appendix D.4), except for SVM when $L = 20$ on the Reuters-R8 dataset. In addition, MMCTM performs almost as well as using the optimal parameter setting (MMCTM*).

Sensitivity to the number of topics (i.e., T). Note the above experiments simply set $T = 50$. To validate the affect of T, we varied T from 10 to 100, and the corresponding accuracies are plotted In Fig. 4 for the two datasets. In both cases, $T = 50$ seems to be a good parameter value.

Cluster embedding. We finally plot the clustering results by embedding them into the 2-dimensional plane using tSNE [38]. In Fig. 5, it can be observed that compared to CTM, MMCTM generates well separated clusters with much larger margin between clusters.

7 Conclusions

We propose a robust Bayesian max-margin clustering framework to bridge the gap between max-margin learning and Bayesian clustering, allowing many Bayesian clustering algorithms to be directly equipped with the max-margin criterion. Posterior inference is done via two data augmentation techniques. Two models from the framework are proposed for Bayesian nonparametric max-margin clustering and topic model based document clustering. Experimental results show our models significantly outperform existing methods with competitive clustering accuracy.

Acknowledgments

This work was supported by an Australia China Science and Research Fund grant (ACSRF-06283) from the Department of Industry, Innovation, Climate Change, Science, Research and Tertiary Education of the Australian Government, the National Key Project for Basic Research of China (No. 2013CB329403), and NSF of China (Nos. 61322308, 61332007). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

References

[1] J. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Math., Stat., and Prob.*, page 281, 1967.

[2] J. Shi and J. Malik. Normalized cuts and image segmentation. *TPAMI*, 22(8):705–767, 2000.

[3] A. Ben-Hur, D. Horn, H. Siegelmann, and V. Vapnik. Support vector clustering. *JMLR*, 2:125–137, 2001.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *JMLR*, 6: 1705–1749, 2005.

[5] H. Cheng, X. Zhang, and D. Schuurmans. Convex relaxations of Bregman divergence clustering. In *UAI*, 2013.

[6] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Max-margin clustering. In *NIPS*, 2005.

[7] B. Zhao, F. Wang, and C. Zhang. Efficient multiclass maximum margin clustering. In *ICML*, 2008.

[8] Y. F. Li, I. W. Tsang, J. T. Kwok, and Z. H. Zhou. Tighter and convex maximum margin clustering. In *AISTATS*, 2009.

[9] G. T. Zhou, T. Lan, A. Vahdat, and G. Mori. Latent maximum margin clustering. In *NIPS*, 2013.

[10] C. E. Rasmussen. The infinite Gaussian mixture model. In *NIPS*, 2000.

[11] K. A. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *ICML*, 2005.

[12] Y. W. Teh, H. Daumé III, and D. Roy. Bayesian agglomerative clustering with coalescents. In *NIPS*, 2008.

[13] Y. Hu, J. Boyd-Graber, H. Daumé, and Z. I. Ying. Binary to bushy: Bayesian hierarchical clustering with the Beta coalescent. In *NIPS*, 2013.

[14] P. Wang, K. B. Laskey, C. Domeniconi, and M. I. Jordan. Nonparametric Bayesian co-clustering ensembles. In *SDM*, 2011.

[15] J. Zhu, N. Chen, and E. P. Xing. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *JMLR*, 2014.

[16] J. Zhu, A. Ahmed, and E. P. Xing. MedLDA: Maximum margin supervised topic models. *JMLR*, 13(8): 2237–2278, 2012.

[17] J. Zhu, N. Chen, H. Perkins, and B. Zhang. Gibbs max-margin topic models with fast sampling algorithms. In *ICML*, 2013.

[18] J. Zhu. Max-margin nonparametric latent feature models for link prediction. In *ICML*, 2012.

[19] M. Xu, J. Zhu, and B. Zhang. Fast max-margin matrix factorization with data augmentation. In *ICML*, 2013.

[20] L. Wang, X. Li, Z. Tu, and J. Jia. Discriminative cllustering via generative feature mapping. In *AAAI*, 2012.

[21] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *NIPS*, 2010.

[22] K.‘Ganchev, J. Graa, J. Gillenwater, and B. Taskar. Posterior regularization for structured latent variable models. *JMLR*, 11:2001–2049, 2010.

[23] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001.

[24] C. Fox. *An introduction to the calculus of variations*. Courier Dover Publications, 1987.

[25] B. Jorgensen. *Statistical properties of the generalized inverse Gaussian distribution*. Lecture Notes in Statistics, 1982.

[26] K. P. Murphy. Conjugate Bayesian analysis of the Gaussian distribution. Technical report, UCB, 2007.

[27] S. Favaro and Y. W. Teh. MCMC for normalized random measure mixture models. *Stat. Sci.*, 2013.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[29] H. M. Wallach. *Structured topic models for language*. PhD thesis, University of Cambridge, 2008.

[30] A. Jain and M. Law. Data clustering: A user's dilemma. *Lecture Notes in Comp. Sci.*, 3776:1–10, 2005.

[31] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101(476):1566–1581, 2006.

[32] M. Davy and J. Y. Tourneret. Generative supervised classification using Dirichlet process priors. *TPAMI*, 32(10):1781–1794, 2010.

[33] P. Franti and O. Virmajoki. Iterative shrinking method for clustering problems. *PR*, 39(5):761–765, 2006.

[34] K. Bache and M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

[35] A. Shah and Z. Ghahramani. Determinantal clustering process – a nonparametric bayesian approach to kernel based semi-supervised clustering. In *UAI*, 2013.

[36] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *JMLR*, (11):2837–2854, 2010.

[37] V. Sindhwani, P. Niyogi, and M. Belkin. SVMlin: Fast linear SVM solvers for supervised and semi-supervised learning. In *NIPS Workshop on Machine Learning Open Source Software*, 2006. http://vikas.sindhwani.org/svmlin.html.

[38] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-SNE. *JMLR*, 9(11): 2579–2605, 2008.

A Derivation of Eq.(7)

First we introduce a basic lemma from calculus of variation [24] below:

Lemma 1 (The Euler-Lagrange equation) *Let $F(\alpha, \beta, \gamma)$ be a function with continuously first and second partial derivatives with respect to (α, β, γ), $\mathcal{S} = \{x \in \mathcal{C}^1[a, b] | x(a) = y_a, x(b) = y_b\}$ where $\mathcal{C}^1[a, b]$ denotes the space of one time continuously differentiable functions in $[a, b]$. Let $\mathcal{I} : \mathcal{S} \to \mathbb{R}$ be a function of the form:*

$$\mathcal{I}(x) = \int_a^b F(x(t), x'(t), t)\mathrm{d}t \, .$$

If $\mathcal{I}$ has an extremum at $x_0 \in \mathcal{S}$, then x_0 satisfies the Euler-Lagrange equation:

$$\frac{\partial F}{\partial \alpha}(x_0(t), x_0'(t), t) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial F}{\partial \beta}(x_0(t), x_0'(t), t)\right) = 0,$$

for $t \in [a, b]$.

Derivation of Eq.(7) Substituting the $\mathcal{L}$ into (6), it is easy to see that (6) is equivalent to

$$\inf_{q(\Theta)} \mathrm{KL}(q(\Theta)||p(\Theta)) - \mathbb{E}_{\Theta\sim q(\Theta)}\left[\log p(\mathbf{X}|\Theta)\right] + 2c\sum_i \mathbb{E}_{\Theta\sim q(\Theta)}\left[\max\left(0, \max_{k:k\neq y_i} \zeta_{ik}\right)\right]$$

$$\Longrightarrow \inf_{q(\Theta)} \int_\Theta \log\frac{q(\Theta)}{p(\Theta)}q(\Theta)\mathrm{d}\Theta - \int_\Theta \log p(\mathbf{X}|\Theta)q(\Theta)\mathrm{d}\Theta + 2c\sum_i \int_\Theta \left[\max\left(0, \max_{k:k\neq y_i} \zeta_{ik}\right)\right] q(\Theta)\mathrm{d}\Theta$$

Let F in Lemma 1 be:

$$F(q(\Theta), q'(\Theta), \Theta) = \log\frac{q(\Theta)}{p(\Theta)}q(\Theta) - \log p(\mathbf{X}|\Theta)q(\Theta) + 2c\sum_i\left[\max\left(0, \max_{k:k\neq y_i} \zeta_{ik}\right)\right] q(\Theta) \, .$$

Applying Lemma 1 and simplifying we get

$$q(\Theta) \propto p(\Theta|\mathbf{X})\prod_i \exp(-2c\max(0, \max_{k:k\neq y_i} \zeta_{ik})) \, .$$

Now with similar idea as the slice sampler [39], we augment the state Θ with some auxiliary variables $(s_i \in [K])_{i=1}^n$ such that the conditional probability of s_i is:

$$p(s_i = s|\text{others}) = \delta_{\arg\max_{k:k\neq y_i} \zeta_{ik}}(s) \, ,$$

where $\delta_x()$ is a spike at location x. Combining the above probability, the augmented posterior is then given by

$$q(\Theta, \{s_i\}) \propto p(\Theta|\mathbf{X})\prod_i \exp(-2c\max(0, \zeta_{is_i})) \, ,$$

which is exactly Eq.(7). Note this augmentation allows us to sample Θ and $\{s_i\}$ via Gibbs sampler:

- Conditioned on $\{s_i\}$, sample Θ. This results in simple conditional distributions similar to the 2-class case.
- Conditioned on Θ, update $\{s_i\}$ (possibly jointly sample (y_i, s_i) as is done in this paper), with s_i simply takes the value: $s_i = \arg\max_{k:k\neq y_i} \zeta_{ik}$.

B Posterior formulas for the DPMMGM model

The posterior of the DPMMGM model is obtained by multiplying the pseudo likelihood and the prior in the generative process in Section 4. We use MCMC for posterior inference since it has been shown quite efficient [17]. After integrating out the mixing weights $\mathbf{w}$, the variables that need to be sampled are $C := \Theta \cup \{s_i, \lambda_i\}_i = \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\eta}_k\}_k \cup \{y_i, s_i, \lambda_i\}_i$. Substituting the DPGMM posterior into (8) we can derive the following updates for the variables.

Sampling $\boldsymbol{\eta}_k$ This can be done by expanding related $\tilde{\phi}_i(\lambda_i|\Theta)$ terms in (8). Clearly the terms associated with $\boldsymbol{\eta}_k$ are those i's such that $y_i = k$ or $s_i = k$. First expanding the squared term in $\tilde{\phi}_i(\lambda_i|\Theta)$, we have

$$(\lambda_i + c\zeta_{is_i})^2 = \left(\lambda_i + c\ell - c\left(\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i}\right)^T \mathbf{x}_i\right)^2$$
$$= (\lambda_i + c\ell)^2 - 2c\left(\lambda_i + c\ell\right)\left(\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i}\right)^T \mathbf{x}_i + c^2\boldsymbol{\eta}_{y_i}^T\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\eta}_{y_i} + c^2\boldsymbol{\eta}_{s_i}^T\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\eta}_{s_i} - 2c^2\boldsymbol{\eta}_{y_i}^T\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\eta}_{s_i}$$

Now sum over all i's such that $y_i = k$ or $s_i = k$ and multiply by $-\frac{1}{2\lambda_i}$ inside $\tilde{\phi}_i(\lambda_i|\Theta)$, we have the terms inside the exponent being:

$$\Longrightarrow \left(\sum_{i:y_i=k}\left(\frac{c}{\lambda_i}\left(\lambda_i + c\ell\right)\mathbf{x}_i^T + \frac{c^2}{\lambda_i}\boldsymbol{\eta}_{s_i}^T\mathbf{x}_i\mathbf{x}_i^T\right) - \left(\sum_{i:s_i=k}\left(\frac{c}{\lambda_i}\left(\lambda_i + c\ell\right)\mathbf{x}_i^T - \frac{c^2}{\lambda_i}\boldsymbol{\eta}_{y_i}^T\mathbf{x}_i\mathbf{x}_i^T\right)\right)\right)\boldsymbol{\eta}_k$$
$$- \frac{c^2}{2}\boldsymbol{\eta}_k^T\left(\sum_{i:y_i=k||s_i=k}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\lambda_i}\right)\boldsymbol{\eta}_k\ .$$

This corresponds to the first order and second order terms of a Multivariate normal distribution. Re-arranging terms it is easy to see that $\boldsymbol{\eta}_k$ is a multivariate normal distribution as:

$q(\boldsymbol{\eta}_k|C\backslash\boldsymbol{\eta}_k) \sim \mathcal{N}(\boldsymbol{\eta}_k; \mathbf{B}^{-1}\mathbf{A}, \mathbf{B}^{-1})$, where

$$\mathbf{A} = c^2\left(\left(\sum_{i:y_i=k}(\ell\kappa_i + \frac{1}{c})\mathbf{x}_i\right) - \left(\sum_{i:s_i=k}(\ell\kappa_i + \frac{1}{c})\mathbf{x}_i\right) + \sum_{i:y_i=k}\kappa_i\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\eta}_{s_i} + \sum_{i:s_i=k}\kappa_i\mathbf{x}_i\mathbf{x}_i^T\boldsymbol{\eta}_{y_i}\right)$$
$$\mathbf{B} = \frac{1}{v}\mathbf{I} + c^2\left(\sum_{i:y_i=k \text{ or } s_i=k}\kappa_i\mathbf{x}_i\mathbf{x}_i^T\right), \tag{11}$$

with $\kappa_i := \lambda_i^{-1}$ for notation simplicity, and K is the current number of active clusters.

Sample λ_i It is easy to see that λ_i has posterior

$$q(\lambda_i|C\backslash\lambda_i) \sim \mathcal{GIG}\Big(\lambda_i; -\frac{1}{2}, 1, c^2\Big(\ell - \left(\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i}\right)^T\mathbf{x}_i\Big)^2\Big),$$

where $\mathcal{GIG}(x; p, a, b) \propto x^{p-1}\exp\left(-\frac{1}{2}\left(\frac{b}{x} + ax\right)\right)$ is a generalised inverse Gaussian distribution with parameters (p, a, b). So λ_i can be sampled efficiently via the technique proposed recently by [40].

Sample $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ With a *Gauss-inverse Wishart* prior for $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ in (9), the posterior is still a *Gauss-inverse Wishart* [26]. Let $\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i:y_i=k}\mathbf{x}_i$ where $n_k = \sum_i \mathbb{I}(y_i = k)$ is the current number of observations in cluster k. Then the update formulae of parameters are

$$r_k = r + n_k,\quad \nu_k = \nu + n_k,\quad \mathbf{m}_k = \frac{r\mathbf{m} + n_k\bar{\mathbf{x}}_k}{r + n_k},$$
$$\mathbf{S}_k = \mathbf{S} + \frac{r \cdot n_k}{r + n_k}(\mathbf{m} - \bar{\mathbf{x}}_k)(\mathbf{m} - \bar{\mathbf{x}}_k)^T + \sum_{i:y_i=k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T.$$

Sample (y_i, s_i) jointly with $\mathbf{w}$ integrated out Our model is nonparametric Bayesian, allowing new components to be born while sampling y_i, as is done in DPGMM. We sample y_i in a collapsed fashion by integrating out the weights $\mathbf{w}$, which has been shown more efficient [41]. Because of the non-conjugacy between the pseudo likelihood and the prior, we adopt an auxiliary state MCMC algorithm called *Reused Algorithm*, developed by [27] recently. The idea is similar to *Algorithm 8* of [41] by introducing some augmented empty components into the MCMC states, but the efficiency is much improved. Since s_i depends on y_i, we will sample (y_i, s_i) jointly. The algorithm is summarised in Algorithm 1.

Algorithm 1: Reused Algorithm for sampling y_i

```
Initialise $\{y_1, \cdots, y_n\}$. Assign $H$ (set to 3 in the experiments) auxiliary empty clusters $(\boldsymbol{\mu}_t, \boldsymbol{\Lambda}_t, \boldsymbol{\eta}_t)_{t=1}^{T}$.
Remove $y_i$ from the statistics (denote the current cluster parameters as $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\eta}_k)$):
if the current cluster contains only $\mathbf{x}_i$ then
    Randomly sample a cluster from the $H$ auxiliary empty clusters;
    Assign the selected auxiliary empty clusters with parameters $(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{\eta}_k)$;
end if
Sample $(y_i, s_i)$ jointly from its posterior distribution:
    $$q(y_i = k, s_i = \tau(k) | C \backslash (y_i, s_i)) \propto \beta_k f(\mathbf{x}_i | \cdots),$$
    where $\tau(k) = \arg\max_{k':k' \neq k} \zeta_{ik'}$, $f(\cdot)$ is defined in (10), $\beta_k = n_k$ if $n_k > 0$, and $\frac{\alpha}{H}$ otherwise.
if $y_i$ is assigned to an auxiliary empty cluster then
    Move the cluster out of the auxiliary empty cluster.
    Draw the cluster parameters for the corresponding auxiliary empty cluster from the prior distribution.
end if
```

C Posterior formulas for the MMCTM model

We first derive the posterior for the CTM (the one without maxi-margin constraint). This corresponds to Figure 1 (right) with $\{\boldsymbol{\eta}_k\}$ and the arrow between y_i and $w_{i\ell}$ removed. We can write down the posterior of CTM by first integrating out the $\boldsymbol{\gamma}$ and $\{\boldsymbol{\phi}_t\}$ latent variables using the Dirichlet-Multinomial conjugacy. The hierarchical Dirichlet distribution part does not embody a close form posterior, but we can represent the posterior in a sequential way, *e.g.*, we assume the words are generated sequentially, then we multiply the probabilities of generating each word to form the posterior. Before giving the posterior formula, we first define the following statistics:

- m_{tw}: the number of times that a word w is sampled from topic t.
- C_k: the number of documents in cluster k.
- n_{it}: the number of words in document i with topic t.
- g_{kt}: the number of documents in cluster k with topic t.
- h_t: the number of clusters included in topic t.

We further use 'dot' to represent marginal sum, *e.g.*, $n_{i\cdot} := \sum_{t=1}^{T} n_{it}$. Superscript "$\neg l$" means the statistics excluding word l, *e.g.*, $\mathbf{Z}^{\neg l} := \{z_{il'}\}_{i,l' \neq l}$. Based on these statistics, the posterior of CTM can be written as

$$\begin{aligned} & p(\mathbf{W}, \{m_{tw}\}, \{C_k\}, \{n_{it}\}, \{g_{yt}\}, \{h_t\} | \alpha_0, \alpha_1, \alpha, \omega, \nu, \beta) \\ = & \frac{\prod_{t=1}^{T} \Gamma\left(m_{t\cdot} + \beta\right)}{\Gamma\left(m_{\cdot\cdot} + \beta V\right)} \frac{\prod_{k=1}^{K} \Gamma\left(C_k + \omega\right)}{\Gamma\left(C_{\cdot} + \omega K\right)} \prod_{i=1}^{D} \prod_{l=1}^{N_d} p\left(z_{il} | y_i, \text{ others}\right), \end{aligned} \tag{12}$$

where

$$p\left(z_{il} | y_i, \text{ others}\right) = \frac{n_{iz_{il}}^{<} + \alpha \frac{g_{y_i z_{il}}^{<} + \alpha_1 \frac{h_{z_{il}}^{<} + \alpha_0 / T}{h_{\cdot}^{<} + \alpha_0}}{g_{y_i \cdot}^{<} + \alpha_1}}{n_{i\cdot}^{<} + \alpha}.$$

Here superscript "$<$" means the statistics up to the current word in the current document, and $p\left(z_{il} | y_i, \text{ others}\right)$ represents the probability of generating the word w_{il}.

According to the Bayesian max-margin clustering framework in Section 3, the posterior of the proposed MMCTM is simply multiplying the above CTM posterior (12) with the terms $\tilde{\phi}_i(\lambda_i | \Theta)$ defined in (8) induced by the max-margin constraints, and with the features $\mathbf{x}_i$ defined as latent: $\mathbf{x}_i \triangleq \bar{\mathbf{z}}_i \in \mathbb{R}^T$, where the t-th element is simply $\frac{1}{N_i} \sum_{l=1}^{N_i} \mathbb{I}(z_{il} = t)$.

C.1 Posterior inference

The variables needed to be sampled are $C = (\{\boldsymbol{\eta}_k\}, \{\lambda_i\}, \{z_{il}\}, \{y_i, s_i\})$, the conditional distributions can be derived from the posterior $q(\{\boldsymbol{\eta}_k\}, \{\lambda_i\}, \{z_{il}\}, \{y_i, s_i\})$. Based on the posterior analysis above, the sampling proceeds simply goes as:

Sampling $\boldsymbol{\eta}_k$: This is similar to the DPMMGM case since the original posterior $p(\mathbf{W}, \{m_{tw}\}, \{C_k\}, \{n_{it}\}, \{g_{kt}\}, \{h_t\} | \alpha_0, \alpha_1, \alpha, \omega, \nu, \beta)$ does not impact the $\boldsymbol{\eta}_k$ directly. So the form of the conditional distribution for $\boldsymbol{\eta}_k$ is the same as the DPMMGM (but note here $\mathbf{x}_i$ is latent):

$$
\begin{aligned}
& q(\boldsymbol{\eta}_k | C \backslash \boldsymbol{\eta}_k) \sim \mathcal{N}(\boldsymbol{\eta}_k; \mathbf{B}^{-1}\mathbf{A}, \mathbf{B}^{-1}), \text{ where} \\
& \mathbf{A} = c^2 \left(\left(\sum_{i:y_i=k} \frac{c\ell + \lambda_i}{c\lambda_i} \mathbf{x}_i \right) - \left(\sum_{i:s_i=k} \frac{c\ell + \lambda_i}{c\lambda_i} \mathbf{x}_i \right) \right. \\
& \qquad \left. + \sum_{i:y_i=k} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\lambda_i} \boldsymbol{\eta}_{s_i} + \sum_{i:s_i=k} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\lambda_i} \boldsymbol{\eta}_{y_i} \right) \\
& \mathbf{B} = \frac{1}{v}\mathbf{I} + c^2 \left(\sum_{i:y_i=k || s_i=k} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\lambda_i} \right),
\end{aligned}
$$

and K is the current number of active clusters.

Sampling λ_i: This is also similar to the DPMMGM case. By taking related terms out of $\tilde{\phi}_i(\lambda_i | \Theta)$ induced by the max-margin constraint, we get the posterior as

$$
q(\lambda_i | C \backslash \lambda_i) \sim \mathcal{GIG}\Big(\lambda_i; -\frac{1}{2}, 1, c^2 \Big(\ell - (\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i})^T \mathbf{x}_i \Big)^2 \Big).
$$

Sampling z_{il}: When sampling z_{il}, we first remove the word l from document i out of the statistics, and then add it in topic $t = (1, 2, \cdots, T)$. This would change the statistics, and so as the latent feature $\mathbf{x}_i$ of document i in the max-margin term $\tilde{\phi}_i(\lambda_i | \Theta)$. By careful simplification, we get the resulting posterior as:

$$
\begin{aligned}
& q(z_{il} = t | w_{il} = w, \text{others}) \\
\propto & p\left(z_{il} | y_i, \mathbf{Z}^{\neg l}, \text{others} \right) \frac{(m_{tw}^{\neg l} + \beta)}{m_{t\cdot}^{\neg l} + \beta V} \exp\left(\frac{A_t}{N_i} + \frac{B_{tt} + 2\sum_{t'=1}^{T} B_{tt'} n_{it'}^{\neg l}}{N_i^2} \right),
\end{aligned}
\tag{13}
$$

where

$$
\begin{aligned}
\mathbf{A} &= \left(c + \frac{c^2 \ell}{\lambda_i} \right) (\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i}), \\
\mathbf{B} &= -\frac{c^2}{2\lambda_i} (\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i}) (\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i})^T, \\
p\left(z_{il} | y_i, \mathbf{Z}^{\neg l}, \text{others} \right) &= \frac{n_{iz_{il}}^{\neg l} + \alpha \frac{g_{y_i z_{il}}^{\neg l} + \alpha_1 \frac{h_{z_{il}}^{\neg l} + \alpha_0 / T}{h_{\cdot}^{\neg l} + \alpha_0}}{g_{y_i \cdot}^{\neg l} + \alpha_1}}{n_{i\cdot}^{\neg l} + \alpha}.
\end{aligned}
$$

Here the first and second terms in (13) come from (12) by variable elimination, the third term comes from the $\tilde{\phi}_i(\lambda_i | \Theta)$ terms induced by the max-margin constraints.

Sampling (y_i, s_i): Because y_i relates to all the words in document i, to get its conditional posterior, we first remove all the words in document i out of the statistics, then add them in one by one again following their topic assignments $\{z_{il}\}$. This would cause a likelihood change, *i.e.*, $\prod_{l=1}^{N_i} p\left(z_{il} | y_i, \text{ others} \right)$ in (12). Another term what would change in (12) is the term $\frac{\prod_{k=1}^{K} \Gamma(C_k + \omega)}{\Gamma(C_\cdot + \omega K)}$.

Finally, combining the terms in $\tilde{\phi}_i(\lambda_i|\Theta)$ by the max-margin constraints, we get the joint conditional posterior of (y_i, s_i) as:

$$q(y_i = k, s_i = \tau(k)|\text{others}) \propto (C_k + \omega) f(y_i|\Theta) \ ,$$

where $\tau(k) = \arg\max_{k':k' \neq k} \zeta_{ik'}$, $f(y_i|\Theta)$ is the pseudo likelihood defined as

$$f(y_i|\Theta) := \prod_{l=1}^{N_i} p(z_{il}|y_i, Z_{<(i,l)}, \text{others}) \times \\ \exp\left\{-\frac{1}{2\lambda_i}\left(\lambda_i + c\left(\ell - (\boldsymbol{\eta}_{y_i} - \boldsymbol{\eta}_{s_i})^T \mathbf{x}_i\right)\right)^2\right\} \ ,$$

where the subscript $< (i, l)$ means the statistics up to word l in document i because here the words are added in sequentially.

D Additional results

D.1 Influence of NMI scores by hyperparameters in Bayesian nonparametric cluster models

We demonstrate how the hyperparameters affect the NMI scores in two Bayesian nonparametric clustering models in Figure 6 and Figure 7.

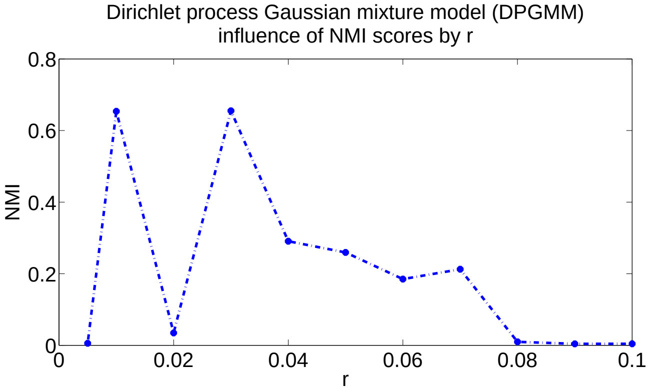


Figure 6: Influence of NMI score by hyparameter r (see Figure. 1 (left) in the main text for the definition) in DPMMGM [10] on the *Wheat Kernel* dataset used in [35] (the whole dataset). Other hyperparameters (refer to Figure. 1 (left) in our paper) are set as $\mathbf{S} = I_7, \mathbf{m} = \mathbf{0}, \nu = 37$, which are empirically performed well. We can see that the NMI performance varies significantly with respect to r. It was not optimal even if we sampled it by placing another hierarchical prior on it.

D.2 CPMMC vs. DPMMGM

In this section we compare the max-margin clustering model [7] with our DPMMGM in Table 3.

D.3 Cluster embedding

For better view of the document clustering embedding results with the MMCTM, Figure 8 re-plots the figures in a larger size, from which we can see clearly that MMCTM clusters documents by inducing larger margins between clusters.

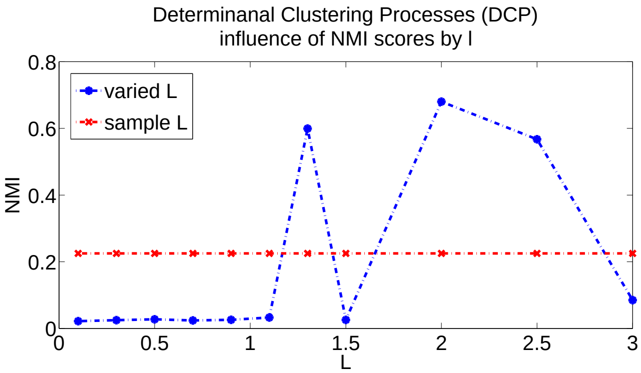


Figure 7: Influence of NMI score by hyparameter L (see Section 2.1.4 in [35]) in DCP [35] on the same dataset using the code from the authors. We varied the L hyperparameter between 0.1 to 3, and sampled all the other other hyperparameters. Again we can see that the NMI performance varies significantly with respect to L. It was not optimal even if we sampled it (the red curve).

Datasets	CPMMC	DPMMGM
Digits	0.5350	**0.5386**
Glass	0.0693	**0.1173**
Half_cicle	0.4291	**0.4726**
Iris	0.0186	**1**
Letter	0.2008	**0.6987**
MNIST	**0.9229**	0.8506
Satimage	0.8558	**1**
Segment'n	N/A	**1**
Vehical	0.0607	**0.0626**
Vowel	0.0044	**0.0129**
Wine	0.8428	**0.9128**

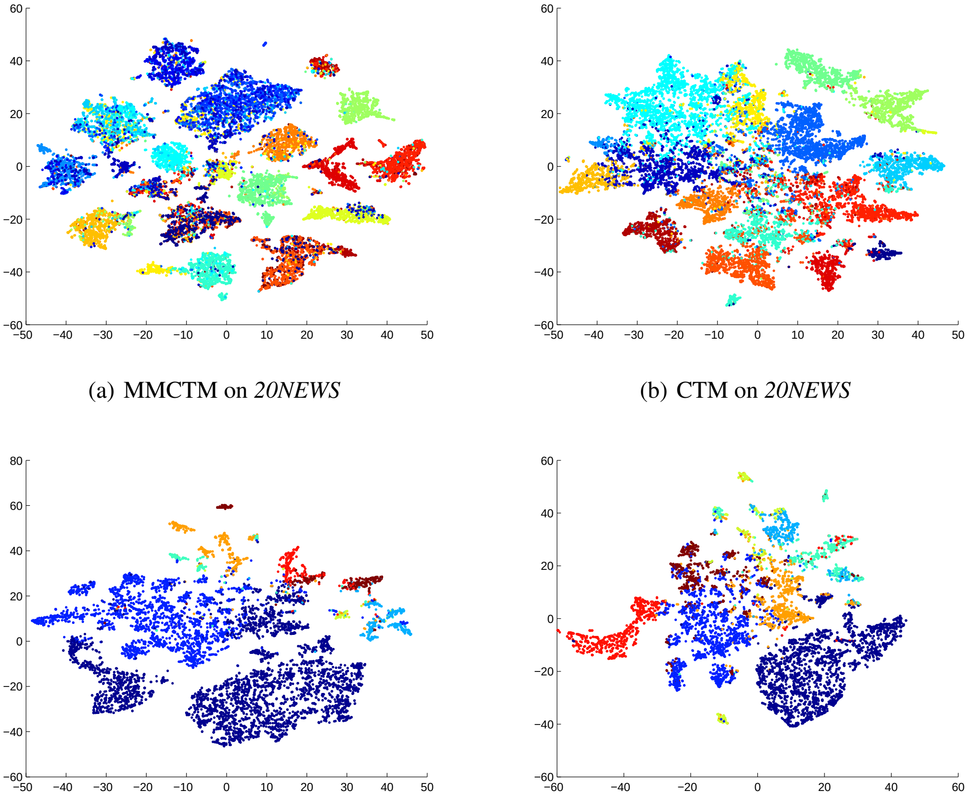
Table 3: Comparison of CPMMC [7] and DPMMGM in term of NMI scores. For each dataset, we randomly chose 2 classes out of the datasets. The first dataset is the one used in the demonstration of their code by the authors, the rest are the datasets used in our paper. For each dataset, related parameters of CPMMC were manually tuned so that it obtained the best NMI scores. "N/A" in the table means infinite loop for the dataset thus no result produced. Obviously our DPMMGM outperforms CPMMC.

D.4 T-test on the clustering results of SVM and MMCTM

To further illustrate the differences between our proposed MMCTM model, we do t-test with the state of the art non-probilistic model – SVM. The results are shown in Table 4, from which we can see that MMCTM* with optimal parameter settings is mostly significantly better than SVM, while MMCTM with fixed hyperparameters is at least as good as SVM, and in some cases it is significantly better.

D.5 Convergence speed

This section shows the convergence speed of the MMCTM in term of clustering accuracy on the two datasets. Figure 9 plots the curves on the *20NEWS* dataset, and Figure 10 plots the curves on the *Reuters-R8* dataset. The results are means and standard derivations over 5 repeated runs with random



(a) MMCTM on *20NEWS*

(b) CTM on *20NEWS*

(c) MMCTM on *Reuters-R8*

(d) CTM on *Reuters-R8*

Figure 8: Clusters embedded in 2-dimensional space with tSNE. Different colours correspond to different clusters. Best view in colour.

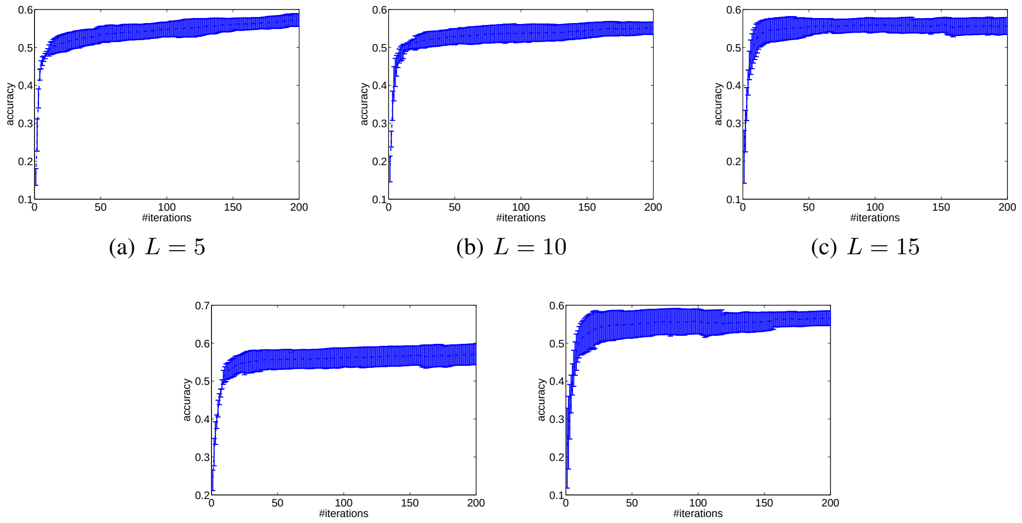
initialisations. We can see that in general the accuracies converge after 30 iterations, demonstrating the efficiency of our sampler.

D.6 Topic illustrations

In this section we show 5 clusters and their associated top 5 topics in each cluster on the *Reuters-R8* dataset in Table 5. We can see from the table that in general each cluster corresponds to specific theme, with several junk topics consists of stop words. Note we did not remove the stop words, and believe it performs even better after removing them.

L	SVM	MMCTM	MMCTM*
		20NEWS	
5	37.13± 2.9	**56.70**± **1.9** ($h=1, p=0.0006$)	**57.86**± **0.9** ($h=1, p=0.0001$)
10	46.99± 2.4	**54.92**± **1.6** ($h=1, p=0.0102$)	**56.56**± **1.3** ($h=1, p=0.0018$)
15	52.80± 1.2	55.06± 2.7 ($h=0, p=0.1962$)	**57.80**± **2.2** ($h=1, p=0.0152$)
20	56.10± 1.5	56.62± 2.2 ($h=0, p=0.7057$)	**59.70**± **1.4** ($h=1, p=0.0347$)
25	59.15± 1.4	55.70± 2.4 ($h=0, p=0.0668$)	**61.92**± **3.0** ($h=1, p=0.0456$)
		Reuters-R8	
5	78.12± 1.1	79.18± 4.1 ($h=0, p=0.6597$)	80.86± 2.9 ($h=0, p=0.1943$)
10	80.69± 1.2	80.04± 5.3 ($h=0, p=0.8265$)	**83.48**± **1.0** ($h=1, p=0.0341$)
15	83.25± 1.7	85.48± 2.1 ($h=0, p=0.0683$)	86.86± 2.5 ($h=0, p=0.0823$)
20	85.66± 1.0	82.92± 1.7 ($h=0, p=0.0742$)	83.82± 1.6 ($h=0, p=0.1711$)
25	84.95± 0.1	86.56± 2.5 ($h=0, p=0.2213$)	**88.12**± **0.5** ($h=1, p=0.0005$)

Table 4: T-test results for the best non-probalistic algorithm SVM and the proposed MMCTM. $h = 1$ means significantly different from SVM, while $h = 0$ means not significantly different, p is the corresponding p-value in the t-test. **Bold** means significantly better.



(a) $L = 5$

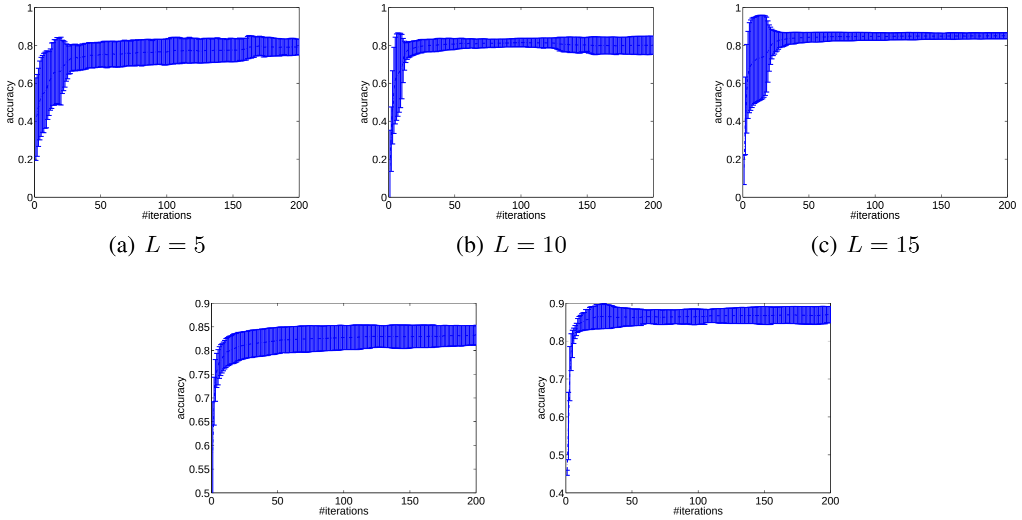
(b) $L = 10$

(c) $L = 15$

(d) $L = 20$

(e) $L = 25$

Figure 9: Accuracy vs. the number of iterations for MMCTM on the *20NEWS* dataset. L is the number of landmarks.



(a) $L = 5$

(b) $L = 10$

(c) $L = 15$

(d) $L = 20$

(e) $L = 25$

Figure 10: Accuracy vs. the number of iterations for MMCTM on the *Reuters-R8* dataset. L is the number of landmarks.

	Cluster 1: tax									
T_1	mln	cts	Net	loss	shr	Reuter	dlrs	qtr	year	revs
T_2	dlrs	share	for	cts	and	per	quater	company	earnings	results
T_3	year	and	dlrs	for	company	will	quater	mln	pct	earnings
T_4	mln	billion	profit	Net	dlrs	loss	tax	shr	pretax	extraordinary
T_5	cts	Net	shr	loss	mln	revs	profit	qtr	Reuter	inc
	Cluster 2: stock									
T_1	for	and	offer	dlrs	share	company	that	will	not	inc
T_2	and	shares	pct	company	stock	group	dlrs	Reuter	common	stake
T_3	and	corp	sale	unit	mln	inc	Reuter	will	company	dlrs
T_4	that	analysts	not	market	analyst	stock	American	spokesman	comment	last
T_5	general	and	gencorp	inc	for	partners	dlrs	taft	mln	acquire
	Cluster 3: trade									
T_1	that	and	for	not	will	told	this	had	meeting	world
T_2	trade	and	reagan	house	bill	countries	gatt	congress	talks	states
T_3	Japan	Japanese	trade	officials	pact	united	Washington	tariffs	semiconductor	last
T_4	trade	billion	exports	dlrs	and	surplus	imports	deficit	Taiwan	year
T_5	and	foreign	for	companies	government	industry	year	sources	this	years
	Cluster 4: agriculture									
T_1	grain	Soviet	agriculture	farm	tonnes	usda	certificates	mln	corp	and
T_2	that	and	for	not	will	told	this	had	meeting	world
T_3	year	and	dlrs	for	company	will	quarter	mln	pct	earnings
T_4	and	foreign	for	companies	government	industry	year	sources	this	years
T_5	trade	and	reagan	house	bill	countries	gatt	congress	talks	states
	Cluster 5: bank									
T_1	dollar	exchange	currency	and	baker	west	Paris	rates	accord	German
T_2	that	and	for	not	will	told	this	had	meeting	world
T_3	bank	stg	mln	market	money	bills	England	pct	today	Reuter
T_4	and	markets	budget	government	dollar	that	economy	policy	exchange	prices
T_5	and	foreign	for	companies	government	industry	year	sources	this	years

Table 5: Top 5 topics in each cluster.

Auxiliary References

[39] R. M. Neal. Slice sampling. *Ann. Statist.*, 31(3):705–767, 2003.

[40] L. Devroye. Random variate generation for the generalized inverse Gaussian distribution. *Stat. Comput.*, 2012.

[41] R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.*, 2000.