# Discriminative Bayesian Nonparametric Clustering

**Vu Nguyen†, Dinh Phung†, Trung Le†, and Hung Bui‡**
†Deakin University, Australia
‡Adobe Research, USA
†{v.nguyen,dinh.phung,trung.l}@deakin.edu.au
‡bui.h.hung@gmail.com

## Abstract

We propose a general framework for discriminative Bayesian nonparametric clustering to promote the inter-discrimination among the learned clusters in a fully Bayesian nonparametric (BNP) manner. Our method combines existing BNP clustering and discriminative models by enforcing latent cluster indices to be consistent with the predicted labels resulted from probabilistic discriminative model. This formulation results in a well-defined generative process wherein we can use either logistic regression or SVM for discrimination. Using the proposed framework, we develop two novel discriminative BNP variants: the discriminative Dirichlet process mixtures, and the discriminative-state infinite HMMs for sequential data. We develop efficient data-augmentation Gibbs samplers for posterior inference. Extensive experiments in image clustering and dynamic location clustering demonstrate that by encouraging discrimination between induced clusters, our model enhances the quality of clustering in comparison with the traditional generative BNP models.

## 1 Introduction

Clustering is an exploratory data analysis task commonly encountered in machine learning and data mining. The ideal goal of clustering is to group data points into clusters in such a way that data points within a cluster are similar (minimize intra-distance) and at the same time, clusters are as separated as possible (maximize inter-distance). Separated clusters often result in distinct and more interpretable clusters and the number of clusters are smaller. Previous work has argued that discriminative clustering is desirable in various tasks, e.g., subspace selection analysis [De la Torre and Kanade, 2006; Ye *et al.*, 2008], computer vision [Joulin *et al.*, 2010], unsupervised regression [Krause *et al.*, 2010] and factor modeling [Henao *et al.*, 2014].

As the number of clusters is often unknown and growing over time, modern Bayesian nonparametric (BNP) clustering methods have the advantages of automatically identifying the suitable number of clusters - most popular models include the Dirichlet process mixture models (DPM) [Ferguson, 1973; Antoniak, 1974; Sethuraman, 1994; Ghahramani, 2012]. While DPM and models alike have been successful, they are strictly generative and the clustering outcomes favor the similarity of data points within a cluster, but do not explicitly model the discrimination among clusters. We consider in this paper *discriminative clustering* [Kaski *et al.*, 2005; Ye *et al.*, 2008; Krause *et al.*, 2010; Chen *et al.*, 2014] under the formalism of Bayesian nonparametrics where data points in the same cluster not only promote to be *similar*, but also be *discriminated* from other data points from other clusters.

To achieve this goal, we propose a general framework called the *discriminative Bayesian nonparametric clustering*. In our framework, the joint strength of BNP clustering and discriminative modeling is achieved by enforcing the latent cluster indices resulted from BNP clustering process to be consistent with the discriminative decisions or labels predicted by the Bayesian discriminative model. This allows our models to have a clear generative process and the flexibility to choose discriminative loss functions, e.g., logistic loss (as in logistic regression) or hinge-loss (as in SVM). Using the proposed framework, we develop two novel unsupervised discriminative BNP variants: the *discriminative Dirichlet process mixtures* (DDPM) for i.i.d data, and the *discriminative-state infinite HMMs* (DIH) for sequential data. For inference, our posterior nicely turns out to be similar to the Chinese restaurant process in DPM but with an additional (discriminative) likelihood ratio term. To efficiently handle the discriminative loss function, we employ data-augmentation sampling technique proposed in [Polson *et al.*, 2011; Polson *et al.*, 2013]. Together, our inference procedure reduces to an efficient data-augmentation Gibbs-sampler.

The idea of learning a latent structure to solve the problem of a *supervised* discriminative task has been investigated before (e.g., [Zhu *et al.*, 2014; Li *et al.*, 2014; Chen *et al.*, 2014]). These works were built on the regularized posterior Bayesian principle, introducing constraints into (generative) Bayesian models for classification tasks, as well as for supervised sequential labeling problems [Zhang *et al.*, 2014]. In contrast, our work focuses purely on *unsupervised* tasks for both i.i.d and sequential data.

Among others, the closest work to ours is the DP Max-margin GMM [Chen *et al.*, 2014]. This regularized model combines a multi-class setting with a standard posterior of a Bayesian nonparametric clustering models for i.i.d data. In

contrast, our model is derived by directly enforcing consistency between a generative BNP component with a probabilistic discriminative one. Furthermore, our model can readily incorporate any discriminative loss such as the logistic loss or the hinge-loss by turning them into likelihood; it also simplifies the inference procedure. Lastly, beyond the work of [Chen *et al.*, 2014], our discriminative-state infinite HMMs is designed to handle sequential data.

To demonstrate the benefit of our models, we experiment with two different tasks in two domains: image clustering (for clustering task) and dynamic location clustering from user WiFi usage (for sequential state-space clustering task). Our experimental results demonstrate that the additional discriminative aspect of our models offers important benefits: not only that it discovers more discriminative clusters, it also enhances the quality of clustering in comparison with the traditional generative BNP models.

## 2 Data Augmentation

We first briefly review the data augmentation method for Bayesian discriminative models [Polson *et al.*, 2011; Polson *et al.*, 2013] which has received great interest recently [Nguyen *et al.*, 2016a; Nguyen *et al.*, 2016b]. Let $x_i \in \mathbb{R}^D$ be a feature vector and $y_i \in \{-1, 1\}$ be a corresponding observed label.

### 2.1 Bayesian Support Vector Machine

Minimizing the regularized Hinge-loss objective function for a standard SVM $\mathscr{L}(\eta; C) = \sum_{i=1}^{N} 2\max\{1 - y_i \eta^T x_i, 0\} + C||\eta||_2^2$ (where $\eta$ is the vector of coefficient parameters and $C > 0$ is the regularization hyper-parameter) is equivalent to a MAP estimation for the following pseudo-posterior distribution parameterized by $\eta$ due to the monotonic property of the exponential $\hat{\eta}_{\text{MAP}} = \underset{\eta}{\text{argmax}}\, p(\eta \mid x, y, C)$ where

$$p(\eta \mid x, y, C) \propto \exp\{-C||\eta||_2^2\} \prod_{i=1}^{N} \exp\{-2\max(1 - y_i \eta^T x_i, 0)\}. \quad (1)$$

Thus, $p(y_i \mid x_i, \eta) \propto \exp\{-2\max(1 - y_i \eta^T x_i, 0)\}$ can be viewed as the likelihood and $p(\eta \mid C) \propto \exp\{-C||\eta||_2^2\}$ as a prior distribution over $\eta$; this is a Gaussian form, i.e., $p(\eta \mid C) = \mathcal{N}(\mu_0, \Sigma_0)$ with mean $\mu_0 = 0$ and $\Sigma_0 = w\mathbf{I}$ where $w = \frac{1}{2C}$.

In a Bayesian setting, we would like to sample from Eq. (1). However, the likelihood term renders it difficult to achieve this goal. The key idea from [Polson *et al.*, 2011] is to augment each data point $x_i$ with an auxiliary variables $\lambda_i > 0$ so that the individual likelihood term can be written as

$$p(y_i \mid x_i, \eta) \propto \exp\left\{-2\max\left(1 - y_i \eta^\top x_i, 0\right)\right\} = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{-\frac{\left[\lambda_i + \left(1 - y_i \eta^\top x_i\right)\right]^2}{2\lambda_i}\right\} d\lambda_i.$$

The term inside the integral can then be viewed as the joint distribution $p(y_i, \lambda_i \mid x_i, \eta)$ (over which marginalizing $\lambda_i$ will recover the likelihood for $y_i$). Thus, the posterior $p(\eta \mid x, y, C)$ can be viewed as the marginal from a joint posterior with the auxiliary variables

$$p(\eta, \lambda \mid ...) \propto \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda_i}} \exp\left\{-\frac{\left[\lambda_i + \left(1 - y_i \eta^\top x_i\right)\right]^2}{2\lambda_i}\right\} \times \exp\left(-\frac{1}{2}[\eta - \mu_0]^T \Sigma_0^{-1} [\eta - \mu_0]\right).$$

Gibbs sampling can now be used on this joint posterior by sequentially sampling $\eta$ given $\lambda_i$ and vice versa.

## 3 Discriminative BNP Clustering

We present the proposed discriminative clustering in this section. We first describe our framework for exchangeable data, and then for sequential data, which extend the Dirichlet process mixture (DPM) [Antoniak, 1974] and the infinite hidden Markov model (iHMM) [Beal *et al.*, 2002] with discriminative power.

### 3.1 Discriminative DPM

Fig. 1 presents the graphical model representation for our proposed *discriminative Dirichlet process mixture* (DDPM). We start with the set of $N$ data points $\{x_1, \ldots, x_N\}$ assumed to be exchangeable where $x_i \in \mathbb{R}^D$. Our goal is to cluster them into $K$ clusters through the indicators $z_i$. The number of active clusters $K$ is unknown and will be determined given the data setting. Given the concentration parameter $\alpha > 0$, we generate the infinite mixing proportion $\pi \sim \text{Stick}(\alpha)$ and the cluster label for each data point $z_i \overset{\text{iid}}{\sim} \text{Mult}(\pi), \forall i = 1, .., N$. Then, we randomly draw a collection of topic atoms as $\phi_k \overset{\text{iid}}{\sim} H(\omega)$ and the corresponding data observation $x_i \sim F(\phi_{z_i})$. This part of the model is identical to a DPM [Antoniak, 1974].

To introduce discrimination, for each cluster, we randomly generate the hyperplane $\eta_k \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \Sigma_0)$ whose purpose is to separate cluster $k$ from the other clusters. For each data point, we further introduce the random vector $y_i = [y_{i1}, ..., y_{iK}, y_{iK_{\text{new}}}]$. Each variable $y_{ik} \in \{1, -1\}$ has two parents $x_i$ and $\eta_k$ (cf. Fig. 1) where $y_{ik} = 1$ indicates that $x_i$ lies in the positive half of $\eta_k$ so that $p(y_{ik} \mid x_i, \eta_k)$

$$= \begin{cases} \propto \exp\{-2\max[1 - y_{ik} \eta_k^T x_i, 0]\} & \text{for SVM} \\ \frac{1}{1 + \exp(-y_{ik} x_i^T \eta_k)} & \text{for LR.} \end{cases} \quad (2)$$

In a traditional discriminative setting, during training $y_{i*}$ will be observed as a one hot encoding[1] of the label; and during testing the predicted value is computed as $\hat{k} = \underset{k}{\text{argmax}}\, p(y_{ik} = 1 \mid x_i, \eta_k)$.

However, *our model is unsupervised*, therefore $y$ is not observed. To connect and ensure the consistency of the generative and the discriminative components in our model, we additionally introduce the consistency variable $c_i$. Specifically, $c_i = 1$ if $y_i$ and $z_i$ are consistent in the sense that $y_{iz_i} = 1$

[1] One hot vector has the usual form [0,...,1,...0] while this one has the form [-1,...1,...-1].

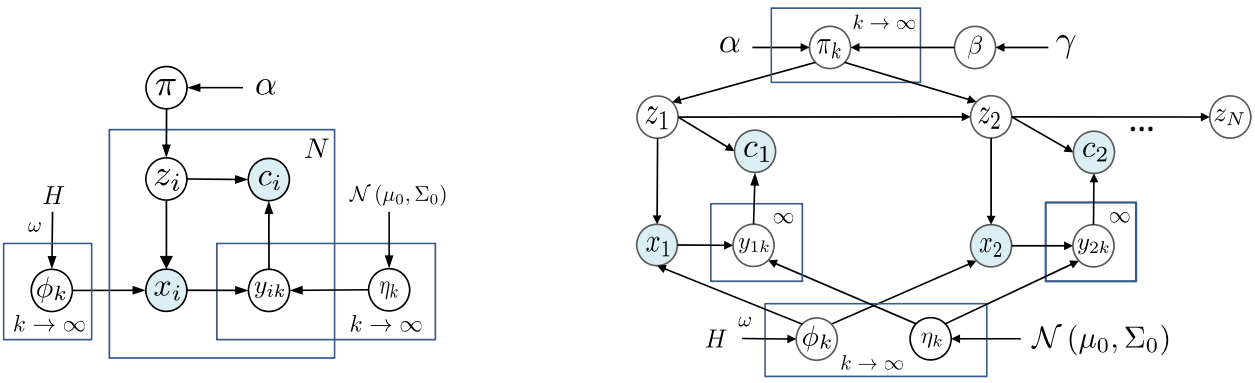


Figure 1: Stick-breaking representation for the discriminative Dirichlet process mixture (DDPM) (Left) and the discriminative-state infinite hidden Markov model (DIH) (Right). The DAG arrows from $z_i$ and the descendants of $x_i$ meet in $c_i$. As a result, conditioning on $c_i$ introduces an extra dependency (in addition to the one already present in the generative model) between $x_i$ and $z_i$. This dependency is contained in the discriminative part of the model.

and $y_{il} = -1, \forall l \neq z_i$ and $c_i = 0$ otherwise. To force the consistency between $y_i$ and $z_i$, $c_i$ is clamped to 1. As a result, the variable $y_i$ will be equated to the one-hot encoding of $z_i$. Hence, there is a dual view of $y_i$: (a) as the hidden mixture component that the data point $x_i$ is assigned to, and (b) as the target for separation by the discriminative component (either a Bayesian SVM or a Bayesian LR).

Given $c_i = 1$ and $z_i = k$, we can rewrite the discriminative likelihood for simplicity as $p\left(y_{i*} = \left[-1, ..., 1_{(k)}, ..., -1\right] \mid x_i, z_i = k, \eta, c_i = 1\right) =$

$$p(y_{ik} = 1 \mid x_i, \eta_k) \prod_{l \neq k} p(y_{il} = -1 \mid x_i, \eta_l)$$
$$= \frac{p(y_{ik} = 1 \mid x_i, \eta_k)}{p(y_{ik} = -1 \mid x_i, \eta_k)} \times Z. \quad (3)$$

where $Z = \prod_{l=1}^{K+1} p(y_{il} = -1 \mid x_i, \eta_l)$. For a new cluster, we sample $\eta_k$ from its prior.

The introduction of the consistency variable $c_i$ is the most crucial part of our model as it provides the glue between the generative and the discriminative components. The truth-value conveyed by $c_i$ (true or false) simply reflects whether the generative component is consistent with the discriminative component. Viewing this way, it is natural to clamp $c_i$ to true since the generative and the discriminative components are then forced to be consistent. There is also a rejection-sampling semantic for our model where the $c_i$ is obtained randomly, but only those samples where $c_i = 1$ are retained.

### Posterior inference

We use collapsed and augmented Gibbs sampler for posterior inference where we integrate out $\pi$ and $\phi_k$ under conjugacy property. Note that all probabilities are conditioned on $c_i = 1$, and hence $y_{i*}$ is a one hot encoding of $z_i$. Thus, we do not need to sample $y_{i*}$ when $z_i$ is known.

The variables which need to be sampled include $z_i, \lambda_i, \forall i = 1, ..., N$ and the discriminative hyperplane $\eta_k, \forall k = 1, ..., K$. Particularly, we provide the posterior inference of $\lambda_i$ and $\eta_k$ for two discriminative cases of SVM and LR, respectively.

**Sampling $\eta_k$.** Given data points in cluster $k$ $\{x_i \mid z_i = k\}$ and data points in other clusters $\{x_j \mid z_j \neq k\}$, the conditional distribution of the discriminative hyperplane $\eta_k$ (with the prior distribution $\eta_k \sim \mathcal{N}(\mu_0, \Sigma_0)$), is

$$\eta_k \mid z, x \propto \mathcal{N}(\eta_k \mid \mu_0, \Sigma_0) \prod_{i \mid z_i = k} p(y_{ik} = 1 \mid x_i, \eta_k) \times \prod_{j \mid z_j \neq k} p\left(y_{jk} = -1 \mid x_j, \eta_k\right). \quad (4)$$

Using data augmentation technique (cf. Sec. 2), the conditional distribution of $\eta_k$ has the form $\mathcal{N}(\eta_k \mid \mu_N, \Sigma_N)$. For discriminative setting with SVM, we have $\Sigma_N^{-1} = \mu_0 \Sigma_0^{-1} + \sum_{i=1}^N \frac{x_i x_i^T}{\lambda_i}$ and $\mu_N = \Sigma_N \left(\sum_{i=1}^N \frac{\lambda_i + 1}{\lambda_i} x_i y_{ik}\right)$. For LR case, $\Sigma_N^{-1} = \mu_0 \Sigma_0^{-1} + \sum_{i=1}^N \lambda_i x_i x_i^T$ and $\mu_N = \Sigma_N \left(\sum_{i=1}^N \frac{1}{2} x_i y_{ik} + \Sigma_0^{-1} \mu_0\right)$. To sample $\eta_k \sim \mathcal{N}(\mu_N, \Sigma_N)$, we compute $\Sigma_N^{-1}$ from the data, but not $\Sigma_N$. To avoid computational inefficiency of matrix inversion, we compute $\mu_N$ by solving the system of linear equations with the complexity of $\mathcal{O}\left(D^{2.376}\right)$ [Ambainis and Filmus, 2014] where $D$ is the feature dimension.

**Sampling $\lambda_i$.** We sample the auxiliary variable as follows: $\lambda_i^{-1} \sim IG\left(\left|1 - x_i^T \eta_{z_i}\right|^{-1}, 1\right)$ for SVM and $\lambda_i \sim PG\left(1, x_i^T \eta_{z_i}\right)$ for LR.

**Sampling $z_i$.** Using the graphical model for DDPM in Fig. 1, the conditional distribution for sampling $z_i$ is $p(z_i = k \mid c_i = 1, x, z_{-i}, \eta, y, \alpha, H) \propto p(z_i = k \mid z_{-i}, \alpha, x_i, x_{-i}, H)\, p(y_{i*} \mid c_i = 1, z_i = k, \eta_k, x_i)$

$$= \underbrace{p(z_i = k \mid z_{-i}, \alpha)}_{\text{CRP}} \times \underbrace{p(x_i \mid z_i = k, z_{-i}, x_{-i}, H)}_{\text{Similarity LK}} \times \underbrace{p(y_{i*} \mid c_i = 1, z_i = k, x_i, \eta_k)}_{\text{Discriminative LK}}. \quad (5)$$

The first term corresponds to a standard Chinese restaurant process (CRP). The second term is the similarity likelihood, reflecting the similarity of $x_i$ w.r.t. other data points

$\{x_j \mid z_j = k, \forall j \neq i\}$ in cluster $k$. Here, we utilize the conjugacy property to integrate out the parameter $\phi_k$. The final term is the discriminative likelihood of the observation $x_i$ (defined in Eq. (3)) which is high if $x_i$ lies on the positive half of $\eta_k$ and low vice versa.

### 3.2 Discriminative-state Infinite HMM

The iHMM [Beal *et al.*, 2002], also known as the HDP-HMM [Teh *et al.*, 2006], is developed to identify the suitable number of states (clusters) for sequential data. Based on the HDP-HMM and the discriminative clustering framework proposed earlier, we develop the *discriminative-state infinite hidden Markov model* (DIH) to perform discriminative clustering on sequential data. The use of Markov property ensures that the temporal dynamics nature of the data is taken into consideration. We again introduce the variables $y_{ik}$, $c_i$ for each observation (cf. Fig. 1 Right) and the discriminative hyperplane $\eta_k$ for capturing the discrimination between clusters.

**Sampling $z_i$.** We follow Eq. (5) to derive the posterior inference over $z_i$ as follows $p(z_i = k \mid .) \propto$

$$p(y_{i*} \mid c_i = 1, z_i = k, x_i, \eta_k) \times p(z_i = k \mid z_{-i}, \alpha, \beta) \times p(x_i \mid z_i = k, z_{-i}, x_{-i}, H).$$

The first term is the discriminative likelihood, defined in Eq. (3) where $y_{ik} = 1$ only at $z_i = k$ and $y_{il} = -1$ otherwise. The second term is the similarity likelihood of the observation $x_i$ in cluster $k$ where the atom component $\phi_{z_i}$ is integrated out. The third term is the CRP for Markov transition $p(z_i = k \mid z_{-i}, \alpha, \beta)$

$$\propto \begin{cases} \left(n_{z_{i-1},k} + \alpha\beta_k\right) \frac{n_{k,z_{i+1}} + \alpha\beta_{z_{i+1}}}{n_{k*} + \alpha} & k \leq K, k \neq z_{i-1} \\ \left(n_{z_{i-1},k} + \alpha\beta_k\right) \frac{n_{k,z_{i+1}} + 1 + \alpha\beta_{z_{i+1}}}{n_{k*} + 1 + \alpha} & z_{i-1} = k = z_{i+1} \\ \left(n_{z_{i-1},k} + \alpha\beta_k\right) \frac{n_{k,z_{i+1}} + \alpha\beta_{z_{i+1}}}{n_{k*} + 1 + \alpha} & z_{i-1} = k \neq z_{i+1} \\ \alpha\beta_{\text{new}}\beta_{z_{i+1}} & k = K + 1. \end{cases}$$

where $n_{i,j}$ is the number of transitions from state $i$ to state $j$ and $n_{i,*}$ is for all transitions from state $i$.

**Sampling $\lambda_i$ and $\eta_k$.** Similar to the approach presented in Section 3.1 for DDPM.

## 4 Experiments

We aim to illustrate that our models can reasonably estimate the unknown number of discriminative clusters and improve the clustering performance.

**Experimental setup.** All experiments are running in the same Windows machine Core i7, 16GB of RAM. The implementation of the proposed model and the baselines are in Matlab. Throughout the experiments, the prior distribution for $\eta_k$ is set as $p(\eta_k \mid \mu_0, \Sigma_0) \sim \mathcal{N}(0, \mathbf{I})$. We use symmetric Dirichlet with parameter 0.01. We set the hyperparameters as $\alpha, \gamma \sim \text{Gamma}(1,1)$, then we resample them in each iteration (for robustness) following the approaches presented in [Escobar and West, 1995; Teh *et al.*, 2006].

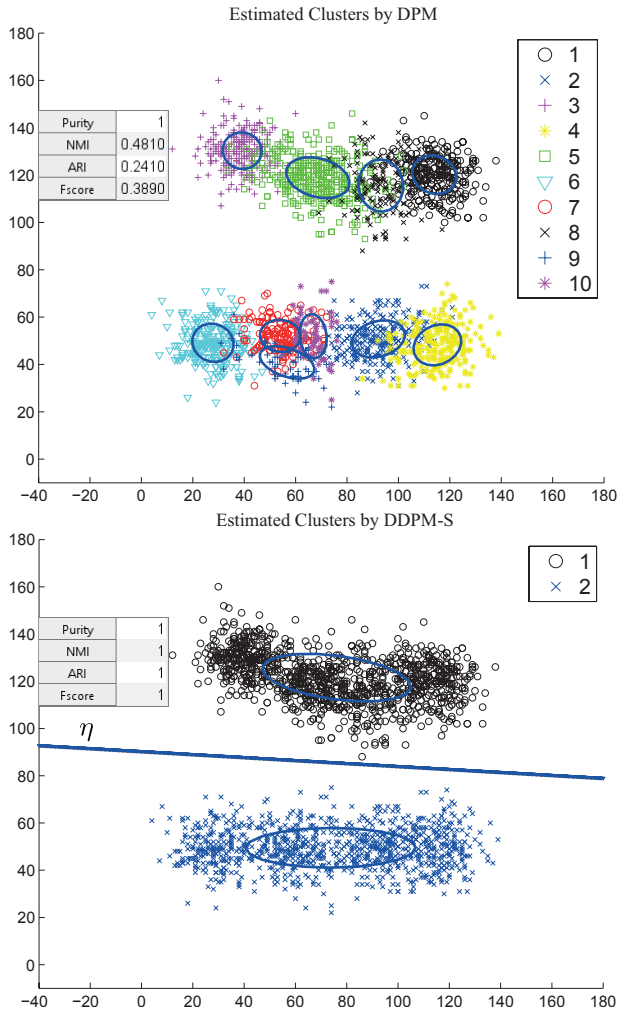


Figure 2: Clustering on the linearly chained Gaussians. We note that if we exhaustively tune the Gaussian covariance hyperparameter controlling the horizontal shape, DPM will estimate correctly. In contrast, by utilizing the discriminative property, DDPM can learn the pattern successfully without tuning the covariance.

To have a good initialization for each discriminative models, in the first 10 iterations of the collapsed Gibbs sampler, we run our proposed models without using discriminative property. We call our discriminative DPM with SVM (DDPM-S) and with LR (DDPM-L). Similarly, DIH-S and DIH-L refer to the two variants of discriminative state iHMM. For evaluation, the clustering scores are measured using 10 independent runs, for each run we record the results from the last 5 Gibbs samples to reduce the sensitivity of MCMC.

### 4.1 DDPM for Data Clustering

We conduct experiments for DDPM on synthetic linear-shape Gaussian 2D data and image clustering task where the data include a collection of i.i.d. observations.

**Synthetic experiment**

We offer a simple illustration to understand the advantage of discriminative clustering (against the standard clustering

| | Score | Kmeans | GMM | AP | DPmeans | DPM | MMGM | DDPM-L | DDPM-S |
|---|---|---|---|---|---|---|---|---|---|
| NUS | K | 10-25 | 10-25 | 18 | 17 | 19 | 15 | 15 | 17 |
| | NMI | .18(.01) | .19(.01) | .166 | .161 | .191(.006) | .184(.007) | ***.197(.012)*** | **.199(.005)** |
| | F1 | .16(.01) | .16(.01) | .145 | .166 | .172(.003) | ***.177(.01)*** | .175(.011) | **.180(.007)** |
| Scenes | K | 10-25 | 10-25 | 18 | 17 | 12 | 10 | 11 | 12 |
| | NMI | .16(.02) | .17(.02) | .154 | .155 | .208(.006) | .186(.008) | ***.218(.01)*** | **.217(.02)** |
| | F1 | .13(.01) | .12(.01) | .124 | .148 | .172(.006) | .171(.009) | **.178(.01)** | ***.176(.02)*** |
| Mnist | K | 10-25 | 10-25 | 18 | 17 | 20 | 18 | 16 | 14 |
| | NMI | 0.56(.02) | .52(.03) | .471 | .518 | ***.603(.02)*** | .583(.02) | **.617(.02)** | .597(.04) |
| | F1 | .40(.03) | .38(.02) | .327 | .379 | .395(.01) | .420(.02) | ***.481(.02)*** | **.457(.03)** |

Table 1: Image clustering comparison. The best and second best scores are in **bold** and ***italic***.

methods) in Fig. 2. DDPM-S and DDPM-L recover correctly $K = 2$ clusters while DPM cannot recover the true clusters. Both DDPM and DPM use Gaussian likelihood to encourage the similarity of data points within a cluster. In addition, the discriminative hyperplane $\eta_k$ in DDPM helps to separate data points in the $k$-th cluster to be away from other clusters. This property makes the algorithm converge to a fewer number of clusters while gaining better clustering performance. We note that if we exhaustively tune the covariance of the Gaussian distribution, DPM might estimate these clusters correctly; whereas our model achieves this with minimum effort.

### Image clustering

We evaluate DDPM on image clustering using NUS Wide (K=13, N=3411, D=500), Fifteen Scenes (K=15, N=2245, D=128) and MNIST (K=10, N=1000, D=784) datasets where K is the number of true cluster, N is the number of data points and D is the feature dimension. We use SIFT histogram as a feature which is assumed to follow Multinomial distribution [Nguyen *et al.*, 2014; Nguyen *et al.*, 2015].

We compare our models with the existing methods:

*DPM* [Antoniak, 1974]: we use the same configuration with our proposed model.

*MMGM* [Chen *et al.*, 2014]: Since Gaussian-Wishart for high dimensional count data of SIFT vectors is not effective, we use Multinomial-Dirichlet likelihood.

*DPmeans* [Kulis and Jordan, 2012]: We follow [Kulis and Jordan, 2012] method to select $\lambda$ using farthest-first heuristic where the expected number of cluster is set to 17 to estimate $\lambda$. This yields $\lambda = 5690$.

*K-means*, *GMM* and *AP* [Frey and Dueck, 2007]: We use Matlab built-in function with Euclidean distance. As K-means and GMM require the number of clusters provided in advance, we vary the number of clusters from 10 to 25, then report the mean and standard deviation.

Table 1 shows the clustering results. DP-means [Kulis and Jordan, 2012] achieves much lower scores than DPM and DDPM in clustering quality. DP-means uses small-variance asymptotic analysis and uses the mean of the data to represent a center. This trade-off assumption makes the DP-means deterministic as traditional k-means, hence run faster, but it degrades the clustering performance. GMM and Kmeans require the number of clusters provided in advance. To have a fair comparison with other BNP approaches which can identify the suitable number of clusters, we vary the number of clusters used in GMM and Kmeans, then report the mean and standard deviation. Our discriminative models consistently achieve the best clustering quality as they outperform the other baselines in all NMI and F1-score. Among our two discriminative variants, DDPM-S with the hinge loss performs slightly better than DDPM-L with the logistic loss.

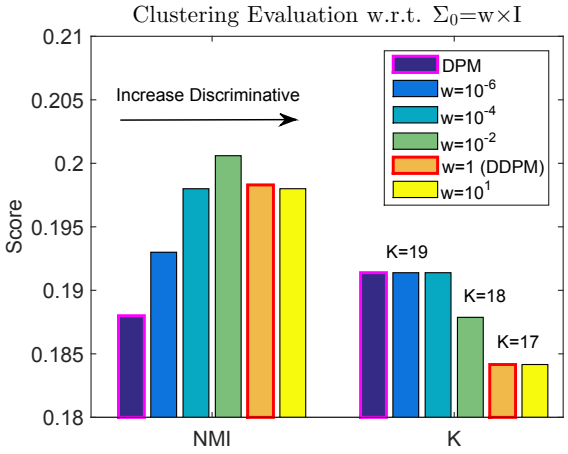


Figure 3: NUS Wide: model behavior w.r.t. $\Sigma_0 = w \times \mathbf{I}$. As $w$ increases from $[10^{-6}, 10^{1}]$, the norm of $\eta$ is likely to increase making the model more discriminative and reduce the number of cluster $K$ from 19 to 17.

Our model also outperforms MMGM in clustering accuracy (cf. Table 1) and computationally more efficient. MMGM [Chen *et al.*, 2014] requires heavier computation than our models due to its max-margin multiclass formulation that uses the Reused Algorithm to jointly estimate two variables: the best and the second best label assignments. The complexity is hence roughly quadratic in the number of current clusters $\mathcal{O}(K^2)$, while sampling $z_i$ in our model only takes $\mathcal{O}(K)$.

### Model analysis

We study the sensitivity of our discriminative model through the parameter $\Sigma_0$ which plays an interesting role in controlling our model discriminative property. Let $\Sigma_0 = w \times \mathbf{I}$ for $w > 0$. As $w \to 0$ (cf. Eq. (4)),$\forall k$, $\eta_k$ approaches zero; hence there is no effect of discrimination, and our model reduces to the standard generative model of DPM. As $w$ increases, the

| Datasets | | | Metric | Non-sequential | | | Sequential | | |
|---|---|---|---|---|---|---|---|---|---|
| User | N | D | | DPM | DDPM-S | DDPM-L | IHMM | DIH-S | DIH-L |
| A | 3932 | 1114 | NMI | .275(.01) | .253(.008) | .246(.026) | **.305(.028)** | .254(.035) | ***.276(.010)*** |
| | | | F1 | .473(.03) | **.526(.009)** | .466(.024) | .441(.018) | .480(.047) | ***.483(.004)*** |
| B | 1793 | 185 | NMI | .245(.006) | .253(.006) | .253(.003) | .235(.023) | ***.254(.002)*** | **.257(.003)** |
| | | | F1 | .535(.007) | ***.553(.006)*** | .551(.001) | .481(.071) | **.559(.007)** | .543(.003) |
| C | 6623 | 1095 | NMI | .199(.01) | .216(.023) | .208(.014) | ***.221(.021)*** | **.222(.006)** | .219(.004) |
| | | | F1 | ***.672(.01)*** | .657(.010) | .641(.031) | .645(.003) | .659(.008) | **.673(.034)** |

Table 2: Clustering evaluation on time-series data using non-sequential and sequential methods.

norm of $\eta$ is also likely to increase, making the discrimination effect stronger.

We vary the value of $w$ and measure performance on the NUS Wide dataset. As can be seen from Fig. 3, the NMI score increases as soon as a small discriminative factor is introduced ($w = 10^{-6}$). In terms of the number of clusters, when we force our model to be more discriminative, our DDPM tends to get less clusters, reducing $K$ from 19 to 17. We also notice the *robustness* of our model in the choice of $\Sigma_0$. For a wide range of $w \in [10^{-6}, 10^{1}]$, our model consistently obtains better performance than the DPM.

### 4.2 DIH for Time-Series Analysis

We evaluate the proposed discriminative-state iHMM using sequential data on location dynamic discovery from the MDC dataset [Laurila *et al.*, 2012]. We pick three users, denoted as A, B and C. Each user has the WiFi access points recorded overtime using a smart phone.

We aim to discover the sequence of routine locations of each user. Each inferred location is represented in our model a state or a cluster of WiFi usage patterns. The clustering performance is then measured using the ground truth provided. We compare discriminative-state iHMM (DIH-S and DIH-L) with the iHMM, as well as models for i.i.d. data (DDPM and DPM). Our main competitor is the iHMM [Beal *et al.*, 2002] which utilizes the Markov property to exploit the sequential dependency and can identify the suitable number of clusters for time series data.

The clustering performance is reported in Table 2. The results show that our discriminative-state iHMMs consistently outperforms the standard iHMM. The DIH-S performs slightly better than DIH-L, and is generally the best method against all the baselines.

We further illustrate the advantage of our discriminative models by visualizing how the clustering performance (F1-score) improves as a function of CPU times in Fig. 4. Since our method might take longer time in each Gibbs iteration comparing to the iHMM, to have a fair comparison, we take the time spent per iteration into the assessment. We observe that our discriminative iHMM take less time to achieve a sufficiently good and stable clustering result. Although not definitive, this behavior suggests that the augmented Gibbs sampler developed for our discriminative iHMM is mixing just as well as the collapsed Gibbs sampler on iHMM, if not faster.

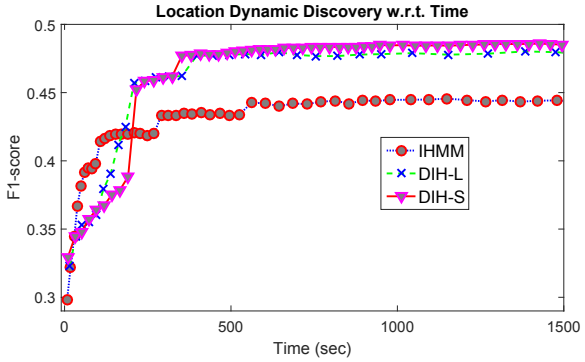


Figure 4: The performance w.r.t. CPU time on User A.

## 5 Conclusion

Extending current generative nonparametric Bayesian clustering models, we proposed a novel framework for discriminative Bayesian nonparametric clustering. Our framework introduces a discriminative component into the existing generative models by enforcing consistency between the generative clusters and the self-induced discriminative labels. This results in two new discriminative BNP clustering models: discriminative DPM and discriminative-state iHMM suitable for i.i.d. and sequential data respectively. By employing a recent data-augmented Gibbs sampler technique for Bayesian probabilistic discriminative classification models (i.e., SVM and LR), we further developed an efficient inference algorithms for our models. Under the new framework, we automatically identify the (unknown) number of clusters while encouraging the similarity within a cluster as well as the separation between the clusters. Experiments with image clustering and sequential WiFi data clustering tasks demonstrated that our model consistently improved in clustering quality compared to existing methods. To scale up this framework, our future work is readily applicable to use recent advances in scalable inference for BNP such as stochastic variational techniques.

## Acknowledgements

This work was partially supported by the Australian Research Council (ARC) and the CoE in Machine Learning and Big Data.

## References


[Ambainis and Filmus, 2014] Andris Ambainis and Yuval Filmus. On the coppersmith–winograd method. 2014.

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

[Beal *et al.*, 2002] M.J. Beal, Z. Ghahramani, and C.E. Rasmussen. The infinite hidden Markov model. In *Advances in Neural Information Processing Systems (NIPS)*, volume 1, pages 577–584. MIT, 2002.

[Chen *et al.*, 2014] Changyou Chen, Jun Zhu, and Xinhua Zhang. Robust bayesian max-margin clustering. In *Advances in Neural Information Processing Systems*, pages 532–540, 2014.

[De la Torre and Kanade, 2006] Fernando De la Torre and Takeo Kanade. Discriminative cluster analysis. In *Proceedings of the 23rd international conference on Machine learning*, pages 241–248. ACM, 2006.

[Escobar and West, 1995] M.D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

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

[Frey and Dueck, 2007] B.J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, February 2007.

[Ghahramani, 2012] Zoubin Ghahramani. Nonparametric bayesian modelling, 2012. Lecture at Machine Learning Summer School, http://videolectures.net/mlss2012-ghahramani-nonparametric-bayesian.

[Henao *et al.*, 2014] Ricardo Henao, Xin Yuan, and Lawrence Carin. Bayesian nonlinear support vector machines and discriminative factor modeling. In *Advances in Neural Information Processing Systems*, pages 1754–1762, 2014.

[Joulin *et al.*, 2010] Armand Joulin, Francis Bach, and Jean Ponce. Discriminative clustering for image co-segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1943–1950. IEEE, 2010.

[Kaski *et al.*, 2005] Samuel Kaski, Janne Sinkkonen, and Arto Klami. Discriminative clustering. *Neurocomputing*, 69(1):18–41, 2005.

[Krause *et al.*, 2010] Andreas Krause, Pietro Perona, and Ryan G Gomes. Discriminative clustering by regularized information maximization. In *Advances in neural information processing systems*, pages 775–783, 2010.

[Kulis and Jordan, 2012] B. Kulis and M. I. Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, Edinburgh, UK, 2012.

[Laurila *et al.*, 2012] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489, 2012.

[Li *et al.*, 2014] Chengtao Li, Jun Zhu, and Jianfei Chen. Bayesian max-margin multi-task learning with data augmentation. pages 415–423, 2014.

[Nguyen *et al.*, 2014] T.V. Nguyen, D. Phung, S. Venkatesh, X.L. Nguyen, and H. Bui. Bayesian nonparametric multilevel clustering with group-level contexts. In *Proc. of International Conference on Machine Learning (ICML)*, pages 288–296, Beijing, China, 2014.

[Nguyen *et al.*, 2015] Vu Nguyen, Dinh Phung, Trung Le, and Svetha Venkatesh. Large sample asymptotic for nonparametric mixture model with count data. In *Workshop on Advances in Approximate Bayesian Inference at Neural Information Processing Systems (NIPS)*, 2015.

[Nguyen *et al.*, 2016a] Tu Dinh Nguyen, Vu Nguyen, Trung Le, and Dinh Phung. Distributed data augmented support vector machine on spark. In *23st International Conference on Pattern Recognition (ICPR)*, pages 498–503. IEEE, 2016.

[Nguyen *et al.*, 2016b] Vu Nguyen, Tu Dinh Nguyen, Trung Le, Svetha Venkatesh, and Dinh Phung. One-pass logistic regression for label-drift and large-scale classification on distributed systems. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1113–1118. IEEE, 2016.

[Polson *et al.*, 2011] Nicholas G Polson, Steven L Scott, et al. Data augmentation for support vector machines. *Bayesian Analysis*, 6(1):1–23, 2011.

[Polson *et al.*, 2013] Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

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

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

[Ye *et al.*, 2008] Jieping Ye, Zheng Zhao, and Mingrui Wu. Discriminative k-means for clustering. In *Advances in neural information processing systems*, pages 1649–1656, 2008.

[Zhang *et al.*, 2014] Aonan Zhang, Jun Zhu, and Bo Zhang. Max-margin infinite hidden markov models. pages 315–323, 2014.

[Zhu *et al.*, 2014] Jun Zhu, Ning Chen, Hugh Perkins, and Bo Zhang. Gibbs max-margin topic models with data augmentation. *The Journal of Machine Learning Research*, 15(1):1073–1110, 2014.