IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 30, 2019, accepted June 19, 2019, date of publication June 24, 2019, date of current version July 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2924651

Nonparametric Hierarchical Bayesian Models for Positive Data Clustering Based on Inverted Dirichlet-Based Distributions

WENTAO FAN[1], **(Member, IEEE), AND NIZAR BOUGUILA**[2], **(Member, IEEE)**
[1]Department of Computer Science and Technology, Huaqiao University, Xiamen 361021, China
[2]Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC H3G 2W1, Canada

Corresponding author: Wentao Fan (fwt@hqu.edu.cn)

This work was supported in part by the Natural Science Foundation of Fujian Province under Grant 2018J01094, and in part by the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University under Grant ZQN-PY510.

ABSTRACT In this paper, we propose nonparametric hierarchical Bayesian models based on two inverted Dirichlet-based distributions and Pitman–Yor process for positive data features clustering. The choice of the inverted Dirichlet and the generalized inverted Dirichlet distributions is motivated by their flexibility and modeling capabilities when dealing with this kind of data, while deploying the Pitman–Yor process prior is justified by its power-law behavior, which makes it a natural choice in real-life application compared with Dirichlet processes for instance. The inference for the resulting models takes into account the challenging problem of feature weighting/selection and is conducted under a Bayesian setting by means of the recently proposed stochastic variational Bayes technique. The efficacy and merits of the proposed approaches are examined using the synthetic data and a challenging real-life application that concerns video background subtraction.

INDEX TERMS Clustering, mixture models, inverted Dirichlet, nonparametric Bayesian model, stochastic variational inference.

I. INTRODUCTION

In recent years, the accelerated growth of digital collections, with increased computing power and electronic storage capacity, has established the need for the development of strong machine learning and data mining techniques. Among these techniques, finite mixture models are being increasingly used in image processing and computer vision applications. Indeed, they can be fitted easily to extracted visual features via both frequentist and Bayesian approaches. Thus, they have continued to receive special attention over the years as technically sound formal approach for visual descriptors clustering [1], [2]. A challenging problem when deploying mixture-based approaches is model selection which consists of determining automatically the model's complexity. It is a crucial task since considering an inappropriate number of mixture component leads to poor generalization capabilities (i.e. under- or over-fitting problems) [3]. Normally, this issue is addressed using the *maximum likelihood* (ML) method in which a model selection criterion is included, such as Akaike information criterion (AIC) [4], Bayes information criterion (BIC) [4], minimum message length (MML) [5], or exact integrated completed likelihood (ICL) [6]. Nevertheless, these approaches are time-consuming due to the fact that they have to evaluate the given selection criterion for multiple numbers of mixture components in order to discover the optimal one.

Recently, nonparametric Bayesian approaches, especially Dirichlet process mixture models have become very popular to tackle the model selection problem by assuming an infinite number of components which avoids the need of deploying tedious selection criteria [2], [7]. In one of our earlier works, we have construed a Dirichlet process mixture of generalized inverted Dirichlet Distributions with feature selection for spatio-temporal video modeling and segmentation [8]. A technically sound alternative to Dirichlet process is Pitman-Yor process which can be viewed as a generalization to the Dirichlet process prior for nonparametric Bayesian

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Chen.

modeling [9]. The Pitman-Yor process allows to produce a large number of populated clusters while reducing the number of sparsely populated ones, which is crucial for several real-life applications [7], [10], [11].

A challenging problem when considering mixture models within nonparametric Bayesian frameworks in general and with a Pitman-Yor process prior in particular is the choice of the per-components data distributions. Gaussian distribution has enjoyed great popularity and has been widely used in several image processing applications. Despite its success and usefulness in many application domains, it suffers from some limitations when dealing with positive data vectors which is generally the case of extracted visual features [12]. In an effort to obtain improved modeling performance and capability when dealing with this kind of features, we propose the consideration of inverted Dirichlet-based distributions [12], [13] within a Pitman-Yor process-based framework. Indeed, we focus on tackling the problem of model-based clustering of grouped positive data using a nonparametric hierarchical Bayesian model namely the hierarchical Pitman-Yor (HPY) process mixture model [7], [10]. In our approach, the data are assumed to be subdivided into a set of groups, where each observation within a group follows a mixture model with an unknown number of components, and where mixture components, represented by inverted Dirichlet-based distributions, are shared among these groups. Within the same group, each observation is drawn independently from a mixture model, and the number of observations within each group may be different. The dependencies among groups are caused by the assumption that the mixture models in different groups may share mixture components. Under the settings of hierarchical modeling [14], parameters are shared among groups, and the randomness of the parameters induces dependencies among different groups. Another crucial problem when dealing with vectors of visual descriptors, that we take into account within our nonparametric Bayesian framework, is feature selection. Indeed, we perform simultaneously clustering and feature selection by adapting a feature selection scheme [15] to our approach. The inference for the resulting models is conducted under a Bayesian setting by means of a stochastic variational Bayes technique namely stochastic variational inference [16]. It exploits natural gradients and is able to handle streaming or large-scale data sets.

The contributions of this work can summarized as follows: we propose two nonparametric hierarchical Bayesian models based on Pitman-Yor process mixture model with both inverted Dirichlet (ID) and generalized inverted Dirichlet (GID) mixture models. We integrate an unsupervised features selection scheme into the proposed nonparametric hierarchical GID mixture models in order to improve clustering performance. We develop efficient learning algorithms to estimate both model parameters and feature saliencies through the framework of stochastic variational inference. The proposed nonparametric hierarchical Bayesian models and the learning algorithms are validated using synthetic data and a real-life application namely video background subtraction.

It is noteworthy that this work can be considered as an extension to our previous work [8]. Compared with [8] in which a hierarchical Dirichlet process (HDP) mixture of GID distributions was proposed for video modeling and segmentation, this work proposes two nonparametric Bayesian models based on HPY process mixture model with both ID and GID distributions. Moreover, different from [8] in which the conventional variational inference was used to learn model parameters, more efficient model learning algorithms based on stochastic variational inference are developed in this work to learn the proposed nonparametric hierarchical Bayesian models. Another related work to the current work is [11] where a HPY process mixture model of generalized Dirichlet (GD) distributions with online variational inference was proposed for clustering proportional data (i.e. normalized vectors) and was applied for scene recognition and video segmentation. In contrast to [11], the current work focusing on clustering positive data based on HPY process mixture models of both ID and GID distributions with stochastic variational inference.

The rest of the paper is organized as follows: Section 2 describes the hierarchical Pitman-Yor mixture model. In Section 3, we develop hierarchical Pitman-Yor mixture models based on inverted Dirichlet-based distributions. The learning of these models is tackled in Section 4. Section 5 reports the experimental results. In Section 6, we end the paper by presenting some concluding remarks.

II. THE HIERARCHICAL PITMAN-YOR PROCESS MIXTURE

The HPY process, which is based on the Pitman-Yor process (also known as the two-parameter Poisson-Dirichlet process) [9], defines a global random probability measure G_0 and an indexed set of random probability measures $\{G_j\}$, one for each group:

$$\begin{aligned} &G_0 \sim \mathrm{PY}(a, b, H), \\ &G_j \sim \mathrm{PY}(a', b', G_0), \quad \text{for each } j \in \{1, \ldots, M\} \end{aligned} \tag{1}$$

where G_0 is the common base (i.e., global-level) measure shared across the different Pitman-Yor processes G_j, and is itself distributed according to a Pitman-Yor process prior.

We can build the HPY process model by applying the stick-breaking construction [17], [18] for the base measure G_0 and the group-level measure G_j, respectively. The stick-breaking construction for G_0 is defined as

$$G_0 = \sum_{k=1}^{\infty} \varpi_k \delta_{\Lambda_k} \tag{2}$$

where

$$\varpi_k = \varpi'_k \prod_{s=1}^{k-1} (1 - \varpi'_s), \quad \varpi'_k \sim \mathrm{Beta}(1-a, ak+b), \;\; \Lambda_k \sim H \tag{3}$$

where $\{\Lambda_k\}$ is a set of independent random variables drawn from H, δ_{Λ_k} is an atom centered at Λ_k, and ϖ_k are the stick-breaking proportions with the constraint that $\sum_{k=1}^{\infty} \varpi_k = 1$.

Notice that since G_0 is the base measure of G_j, the atoms Λ_k are shared among all G_j with different proportions.

The stick-breaking construction for group-level Pitman-Yor process G_j is defined as

$$G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\Omega_{jt}} \tag{4}$$

where we have

$$\pi_{jt} = \pi'_{jt} \prod_{s=1}^{t-1} (1 - \pi'_{js}), \qquad \Omega_{jt} \sim G_0, \tag{5}$$

$$\pi'_{jt} \sim \mathrm{Beta}(1 - a', a't + b') \tag{6}$$

where π_{jt} are the stick-breaking proportions, and each group-level atom Ω_{jt} maps to a global-level atom Λ_k based on the distribution defined by G_0. Then, we introduce a binary indicator variable W_{jtk} such that $W_{jtk} = 1$ if Ω_{jt} maps to the global-level atom Λ_k; otherwise, $W_{jtk} = 0$. As a result, we have $\Omega_{jt} = \Lambda_k^{W_{jtk}}$. The probability distribution of the indicator variable $\vec{W} = (W_{jt1}, W_{jt2}, \ldots)$ is given by

$$p(\vec{W}) = \prod_{j=1}^{M} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\varpi'_k \prod_{s=1}^{k-1} (1 - \varpi'_s) \right]^{W_{jtk}} \tag{7}$$

The HPY process can be used in the grouped mixture model setting by considering a HPY process as the prior distribution over the parameters for grouped data. Let $F(\theta_{ji})$ denotes the distribution of the data point X_{ji} given the parameter θ_{ji}, where the index ji indicates the i-th observation within j-th group. Let G_j represents a HPY process prior distribution for the parameters $\vec{\theta}_j = (\theta_{j1}, \theta_{j2}, \ldots)$ associated with group j. Then, the HPY process mixture model can be defined as

$$\theta_{ji} | G_j \sim G_j, \qquad X_{ji} | \theta_{ji} \sim F(\theta_{ji}) \tag{8}$$

For the HPY process mixture model, we introduce another binary indicator variable Z_{jit}, such that $Z_{jit} = 1$ if θ_{ji} is associated with the t-th component and maps to the group-level atom Ω_{jt}; otherwise, $Z_{jit} = 0$. Then, we have $\theta_{ji} = \Omega_{jt}^{Z_{jit}}$. The probability distribution of the indicator variable $\vec{Z} = (Z_{ji1}, Z_{ji2}, \ldots)$ is given by

$$p(\vec{Z}) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} \prod_{t=1}^{\infty} \left[\pi'_{jt} \prod_{s=1}^{t-1} (1 - \pi'_{js}) \right]^{Z_{jit}} \tag{9}$$

where N_j denotes the number of observations within group j.

III. THE HPY PROCESS MIXTURE MODEL WITH INVERTED DIRICHLET-BASED DISTRIBUTIONS

In this section, we propose two related nonparametric hierarchical Bayesian models based on the HPY process mixture model with two inverted Dirichlet-based distributions, namely the inverted Dirichlet and the generalized inverted Dirichlet distributions. We also incorporate an unsupervised feature selection scheme into the developed HPY process mixture model, and thus form a unified framework for both grouped data modeling and feature selection.

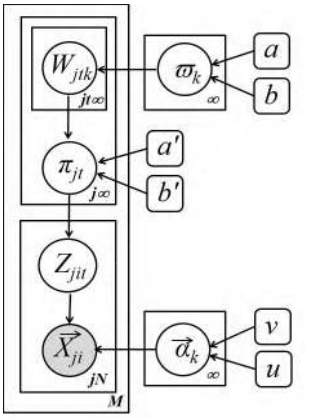


FIGURE 1. Graphical model of the HPY process mixture model with ID distributions. Each node in the graph is associated with a random variable, where shading denotes an observed variable. Plates denote the replication, in which the number of replications is shown in the bottom right corner.

A. THE HPY PROCESS MIXTURE MODEL WITH ID DISTRIBUTIONS

Recently, the inverted Dirichlet (ID) mixture model has shown promising performance in modeling positive vectors and has been applied in various applications, such as object detection, visual scenes analysis and classification [13], [19]. The ID distribution has considerable flexibility which allows both multiple symmetric and asymmetric modes. If a D-dimensional random vector $\vec{X} = (X_1, \ldots, X_D)$ follows a ID distribution with parameter $\vec{\alpha} = (\alpha_1, \ldots, \alpha_{D+1})$, then its probability density function (pdf) is given by [20]:

$$\mathrm{ID}(\vec{X}|\vec{\alpha}) = \frac{\Gamma(\sum_{l=1}^{D+1} \alpha_l)}{\prod_{l=1}^{D+1} \Gamma(\alpha_l)} \prod_{l=1}^{D} X_l^{\alpha_l - 1} \left(1 + \sum_{l=1}^{D} X_l\right)^{-\sum_{l=1}^{D+1} \alpha_l} \tag{10}$$

where $\Gamma(\cdot)$ is the gamma function, $X_l > 0$ for $l = 1, \ldots, D$, and $\alpha_l > 0$ for $l = 1, \ldots, D + 1$.

Given a data set $\mathcal{X}$ that is partitioned into M groups, if each D-dimensional data vector $\vec{X}_{ji} = (X_{ji1}, \ldots, X_{jiD})$ is distributed according to a HPY process mixture model with ID distributions, then its likelihood function is defined by

$$p(\mathcal{X}) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\mathrm{ID}(\vec{X}_{ji}|\vec{\alpha}_k)\right]^{Z_{jit} W_{jtk}} \tag{11}$$

where j and i denote the indices for the group and the data vector, respectively.

In this nonparametric hierarchical Bayesian model, since the parameter $\vec{\alpha}$ is positive, a Gamma prior $\mathcal{G}(\cdot)$ is adopted as

$$p(\vec{\alpha}) = \prod_{k=1}^{\infty} \prod_{l=1}^{D+1} \mathcal{G}(\alpha_{kl} | u_{kl}, v_{kl}) \tag{12}$$

where both u_{kl} and v_{kl} are positive hyperparameters.

A graphical model representation of the HPY process mixture model with ID distributions is shown in Fig. 1.

B. THE HPY PROCESS MIXTURE MODEL WITH GID DISTRIBUTIONS WITH FEATURE SELECTION

Although the ID distribution is a powerful tool for modeling positive vectors, it has a very restrictive covariance structure (i.e. can only be positive) that limits its applicability in real-life applications. Since in real practical cases, the correlation of data may also be negative and then the ID distribution becomes an inappropriate choice. Thus, in this subsection, we propose another nonparametric hierarchical Bayesian model which is based on HPY process mixture model and the generalized inverted Dirichlet distribution (GID) distributions. The GID distribution has a more general covariance structure (can be either positive or negative) than the ID, which therefore may provide more flexibility and better modeling capability [12], [21].

If a D-dimensional random vector $\vec{Y} = (Y_1, \ldots, Y_D)$ follows a GID distribution with parameters $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)$ and $\vec{\beta} = (\beta_1, \ldots, \beta_D)$, then its pdf is given by [21]

$$\mathrm{GID}(\vec{Y}|\vec{\alpha}, \vec{\beta}) = \prod_{l=1}^{D} \frac{\Gamma(\alpha_l+\beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)} \frac{Y_l^{\alpha_l-1}}{(1+\sum_{l=1}^{D} Y_l)^{\alpha_l+\beta_l-\beta_{l+1}}} \tag{13}$$

where $\beta_{D+1} = 0$.

Then, we can build the HPY process mixture model with GID distributions for modeling grouped data. Given an observed data set $\mathcal{Y}$ that contains N D-dimensional random vectors and is divided into M groups, where each vector in one group $\vec{Y}_{ji} = (Y_{ji1}, \ldots, Y_{jiD})$ is sampled from a HPY process mixture model with GID distributions. Then, the likelihood function of this model with latent variables can be defined as

$$p(\mathcal{Y}) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\mathrm{GID}(\vec{Y}_{ji}|\vec{\alpha}_k, \vec{\beta}_k) \right]^{Z_{jit} W_{jtk}} \tag{14}$$

As we may notice, the above HPY process mixture model assumes that all features have the same weight and therefore are equally significant for the clustering task. However, this assumption is not realistic in practice when high-dimensional data is encountered, since some of the features might be irrelevant and then degrade the clustering performance. This fact motivated us to incorporate a feature selection scheme into the proposed nonparametric hierarchical mixture model to take into account this critical issue.

In order to perform feature selection, we adopt an interesting and useful property of the GID distribution as discussed in [21], such that the estimation of a D-dimensional GID distribution may be transformed to D estimations of one-dimensional inverted Beta distributions (also known as Beta prime distributions) with independent features as

$$\mathrm{GID}(\vec{Y}|\vec{\alpha}, \vec{\beta}) = \prod_{l=1}^{D} \mathrm{IB}(X_l|\alpha_l, \beta_l) \tag{15}$$

where the data vector $\vec{Y}$ is geometrically transformed into another D-dimensional data point $\vec{X}$ as: $X_1 = Y_1$ and $X_l = Y_l/(1+\sum_{s=1}^{l-1} Y_s)$ for $l > 1$. $\mathrm{IB}(X_l|\alpha_l, \beta_l)$ is an inverted Beta distribution with parameters $\{\alpha_l, \beta_l\}$ and its pdf is given by

$$\mathrm{IB}(X_l|\alpha_l, \beta_l) = \frac{\Gamma(\alpha_l+\beta_l)}{\Gamma(\alpha_l)\Gamma(\beta_l)} X_l^{\alpha_l-1}(1+X_l)^{-(\alpha_l+\beta_l)} \tag{16}$$

This transformation is well-suited for feature selection. Since the independence between features now becomes a fact rather than an assumption as considered in Gaussian mixture models with unsupervised feature selection [15]. Then, the HPY process mixture model with GID distributions as defined in Eq. (13) is equivalent to the following model with inverted Beta distributions

$$p(\mathcal{X}) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\prod_{l=1}^{D} \mathrm{IB}(X_{jil}|\alpha_{kl}, \beta_{kl}) \right]^{Z_{jit} W_{jtk}} \tag{17}$$

In our work, we exploit an unsupervised feature selection scheme as introduced in [15], so that an irrelevant feature is defined as the one having a common distribution independent from class labels. Thus, the probability distribution of each feature X_{jil} in our model is defined as

$$p(X_{jil}) = \mathrm{IB}(X_{jil}|\alpha_{kl}, \beta_{kl})^{\phi_{jil}} \mathrm{IB}(X_{jil}|\alpha'_l, \beta'_l)^{1-\phi_{jil}} \tag{18}$$

where ϕ_{jil} is a binary variable that indicates the feature relevancy. When ϕ_{jil} equals 0, it indicates that the feature l of the i-th data point with the j-th group is irrelevant, and follows an inverted Beta distribution with parameters α'_l and β'_l) that are common to all clusters. When ϕ_{jil} equals 1, it indicates that the feature X_{jil} is relevant, and is distributed according to $\mathrm{IB}(X_{jil}|\alpha_{kl}, \beta_{kl})$. The marginal distribution of $\vec{\phi}$ is defined as

$$p(\vec{\phi}|\vec{\epsilon}) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} \prod_{l=1}^{D} \epsilon_l^{\phi_{jil}} (1-\epsilon_l)^{1-\phi_{jil}} \tag{19}$$

where $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_D)$ represent feature weights (i.e., the probabilities that the features are relevant).

The prior distribution of $\vec{\epsilon}$ is a Dirichlet distribution parameterized by $\vec{\zeta} = (\zeta_1, \zeta_2)$ with $(\zeta_1, \zeta_2) > 0$ in the form of

$$p(\vec{\epsilon}) = \prod_{l=1}^{D} \mathrm{Dir}(\epsilon_l|\vec{\zeta}) = \prod_{l=1}^{D} \frac{\Gamma(\zeta_1+\zeta_2)}{\Gamma(\zeta_1)\Gamma(\zeta_2)} \epsilon_l^{\zeta_1-1}(1-\epsilon_l)^{\zeta_2-1} \tag{20}$$

Then, the likelihood function of our HPY process mixture model with unsupervised feature selection can be written as

$$p(\mathcal{X}|\vec{Z}, \vec{W}, \vec{\theta}, \vec{\phi}) = \prod_{j=1}^{M} \prod_{i=1}^{N_j} \prod_{t=1}^{\infty} \prod_{k=1}^{\infty} \left[\prod_{l=1}^{D} \mathrm{IB}(X_{jil}|\alpha_{kl}, \beta_{kl})^{\phi_{jil}} \times \mathrm{IB}(X_{jil}|\alpha'_l|\beta'_l)^{(1-\phi_{jil})} \right]^{Z_{jit} W_{jtk}} \tag{21}$$

where $\vec{\theta} = \{\vec{\alpha}, \vec{\beta}, \vec{\alpha}', \vec{\beta}'\}$. Gamma priors $\mathcal{G}(\cdot)$ are used for parameters $\vec{\alpha}$, $\vec{\beta}$, $\vec{\alpha}'$ and $\vec{\beta}'$:

$$p(\vec{\alpha}) = \mathcal{G}(\vec{\alpha}|\vec{u}, \vec{v}), \quad p(\vec{\beta}) = \mathcal{G}(\vec{\beta}|\vec{g}, \vec{h}) \tag{22}$$

$$p(\vec{\alpha}') = \mathcal{G}(\vec{\alpha}'|\vec{u}', \vec{v}'), \quad p(\vec{\beta}') = \mathcal{G}(\vec{\beta}'|\vec{g}', \vec{h}') \tag{23}$$

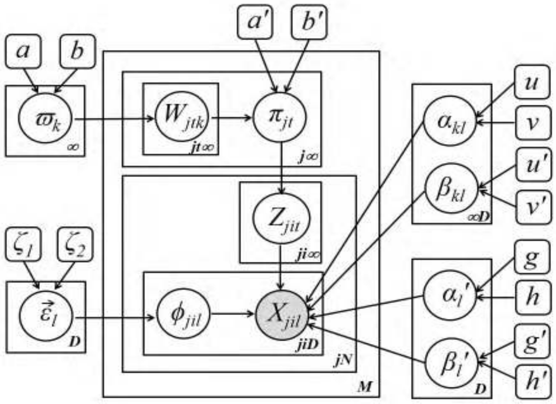


FIGURE 2. Graphical model of the HPY process mixture model with GID distributions with feature selection. Each node in the graph is associated with a random variable, where shading denotes an observed variable. Plates denote the replication, in which the number of replications is shown in the bottom right corner.

A graphical model representation of the HPY process mixture model with GID distributions with feature selection is shown in Fig. 2.

IV. MODEL LEARNING

A. STOCHASTIC VARIATIONAL INFERENCE

Variational inference is a well-defined deterministic approach to approximate posterior distributions through optimization [3], [22], [23]. It finds a variational distribution $q(\Theta)$ to approximate the posterior $p(\Theta|\mathcal{X})$ of a model with parameter Θ by minimizing the Kullback-Leibler (KL) divergence between $q(\Theta)$ and $p(\Theta|\mathcal{X})$. This calculation is equivalent to the maximization of the evidence lower bound (ELBO), which is a lower bound on the logarithm of the model evidence $\log p(\mathcal{X})$. The ELBO is equal to the negative KL divergence up to an additive constant and is defined by

$$\text{ELBO}(q) = \langle \ln p(\mathcal{X}, \Theta)\rangle - \langle \ln q(\Theta)\rangle \tag{24}$$

where $\langle\cdot\rangle$ denotes the calculation of expectation. The variational solutions are obtained by maximizing the ELBO through coordinate ascent, where each variational parameter is iteratively optimized while holding other parameters fixed.

Although variational inference is continuously gaining popularity in Bayesian inference, it is inefficient for large data sets. This is due to the reason that it requires iterating through the entire data set at each iteration. The computational cost becomes more expensive as the size of data set grows. In order to cope with large-scale data sets, we develop an efficient learning algorithm to learn the proposed HPY process mixture model with inverted Dirichlet-based distributions based on stochastic variational inference [16].

The main idea of stochastic variational inference is to optimize the ELBO with noisy estimates of its natural gradient through stochastic optimization (also known as Robbins-Monro algorithm) [24]. In stochastic variational inference, the ELBO as described in Eq. (24) can be decomposed into a global term and a sum of local term as

$$\begin{aligned}\text{ELBO}(q) = \big[\langle \ln p(\Upsilon)\rangle - \langle \ln q(\Upsilon)\rangle\big]\\ + \sum_{i=1}^{N}\big[\langle \ln p(\vec{X}_i, \mathcal{Z}_i|\Upsilon)\rangle - \langle \ln \mathcal{Z}_i)\rangle\big]\end{aligned} \tag{25}$$

where Υ is the set of global variables (i.e., variables that are coupled to the entire set of observations), $\mathcal{Z}_i$ denotes the set of local variables (i.e., variables correspond to each observation $\vec{X}_i$). Stochastic variational approach optimized the maximized ELBO by subsampling the data to form noisy estimates of the natural gradient. Now assume that a single observation indexed by n is sampled uniformly at random with $n \sim \text{Unif}(1, \ldots, N)$ (i.e. $\vec{X}_n$ is sampled uniformly from the data set $\{\vec{X}_1, \ldots, \vec{X}_N\}$). Then, we can form a data set by S replicates of observation ($\vec{X}_n$ and local variables $\mathcal{Z}_n$), and the corresponding ELBO is calculated by

$$\begin{aligned}\text{ELBO}^n(q) = \langle \ln p(\Upsilon)\rangle - \langle \ln q(\Upsilon)\rangle\\ + S[\langle \ln p(\vec{X}_n, \vec{Z}_n|\Upsilon)\rangle - \langle \ln q(\vec{Z}_n)\rangle\big]\end{aligned} \tag{26}$$

As explained in [16], the expected value of $\text{ELBO}^n(q)$ is equal to $\text{ELBO}(q)$, so the natural gradient of $\text{ELBO}^n(q)$ with respect to each global variational parameter is a noisy but unbiased estimate of the natural gradient of $\text{ELBO}(q)$ [16]. Therefore, stochastic variational inference does not need to analyze whole data set but only requires computation about one single local context at each iteration, which is much more computationally efficient than conventional variational inference. Since stochastic variational inference is based on subsampling the training data and performs online parameter updates by using each time a single data point or a small "mini-batch", it is analogous to other online learning algorithms. Moreover, we apply the truncation technique as in [25] to truncate the variational distributions of G_0 and G_j at levels K and T, respectively

$$\varpi_K' = 1, \quad \sum_{k=1}^{K} \varpi_k = 1, \;\; \varpi_k = 0 \;\text{ when }\; k > K \tag{27}$$

$$\pi_{jT}' = 1, \quad \sum_{t=1}^{T} \pi_{jt} = 1, \;\; \pi_{jt} = 0 \;\text{ when }\; t > T \tag{28}$$

where the truncation levels K and T will be inferred automatically during the learning process.

B. LEARNING OF HPY PROCESS MIXTURE WITH ID DISTRIBUTIONS

In this part, we propose an algorithm to learn the HPY process mixture model with ID distributions through stochastic variational inference. In order to obtain closed-form solutions, the mean-field assumption [3] is adopted to factorize the variational posterior distribution $q(\Theta)$ into disjoint factors as

$$q(\Theta) = q(\vec{Z})q(\vec{W})q(\vec{\pi}')q(\vec{\varpi}')q(\vec{\alpha}) \tag{29}$$

where $\Theta = \{\mathcal{Z}, \vec{W}, \vec{\pi}', \vec{\varpi}', \vec{\alpha}\}$ is the set of latent and unknown random variables in our model. Suppose an observed data set $\mathcal{X} = \vec{X}_{1:N}$ is partitioned into M groups. At iteration r, a point $\vec{X}_{jn}$ in group j is randomly sampled with $n \sim \text{Unif}(1, \ldots, N_j)$, where N_j indicates the number of data points in group j. Then, the local variational distribution $q(\vec{Z}_{jn})$ is optimized while holding global variational distributions at their values at the $(r-1)$th iteration. As a result, the variational solution to $q(\vec{Z}_n)$ can be calculated by (See A for details)

$$q(\vec{Z}_{jn}) = \prod_{t=1}^{T} \rho_{jnt}^{Z_{jnt}} \tag{30}$$

The next step is to optimize the global variational distributions $q(\vec{W})$, $q(\vec{\pi}')$, $q(\vec{\varpi}')$ and $q(\vec{\alpha})$ for the current iteration (Details are provided in A). The parametric form of $q(\vec{W})$ at the rth iteration is given by

$$q^{(r)}(\vec{W}) = \prod_{j=1}^{M}\prod_{t=1}^{T}\prod_{k=1}^{K} (\vartheta_{jtk}^{(r)})^{W_{jtk}^{(r)}} \tag{31}$$

where the hyperparameter $\vartheta_{jtk}^{(r)}$ can be calculated by leveraging the noisy natural gradient of the ELBO (with step-size λ_r) as

$$\vartheta_{jtk}^{(r)} = \vartheta_{jtk}^{(r-1)} + \lambda_r \partial \vartheta_{jtk}^{(r)}, \tag{32}$$

where $\partial\vartheta_{jtk}^{(r)}$ denotes the noisy natural gradient of the ELBO with respect to ϑ_{jtk} at the rth iteration.

The global variational distributions $q(\vec{\pi}')$, $q(\vec{\varpi}')$ and $q(\vec{\alpha})$ for the rth iteration can be obtained by

$$q^{(r)}(\vec{\pi}') = \prod_{j=1}^{M}\prod_{t=1}^{T} \text{Beta}(\pi_{jt}'^{(r)} | c_{jt}'^{(r)}, d_{jt}'^{(r)}) \tag{33}$$

$$q^{(r)}(\vec{\varpi}') = \prod_{k=1}^{K} \text{Beta}(\varpi_{k}'^{(r)} | c_{k}^{(r)}, d_{k}^{(r)}) \tag{34}$$

$$q^{(r)}(\vec{\alpha}) = \prod_{k=1}^{K}\prod_{l=1}^{D+1} \mathcal{G}(\alpha_{kl}^{(r)} | u_{kl}^{*(r)}, v_{kl}^{*(r)}) \tag{35}$$

where the hyperparameters of the above global variational distributions are given by

$$c_{jt}'^{(r)} = c_{jt}'^{(r-1)} + \lambda_r \partial c_{jt}'^{(r)}, \quad d_{jt}'^{(r)} = d_{jt}'^{(r-1)} + \lambda_r \partial d_{jt}'^{(r)} \tag{36}$$

$$c_{k}^{(r)} = c_{k}^{(r-1)} + \lambda_r \partial c_{k}^{(r)}, \quad d_{k}^{(r)} = d_{k}^{(r-1)} + \lambda_r \partial d_{k}^{(r)} \tag{37}$$

$$u_{kl}^{*(r)} = u_{kl}^{*(r-1)} + \lambda_r \partial u_{kl}^{*(r)}, \quad v_{kl}^{*(r)} = v_{kl}^{*(r-1)} + \lambda_r \partial v_{kl}^{*(r)} \tag{38}$$

In above equations, the step size λ_r at iteration r is defined by

$$\lambda_r = (\eta + r)^{-\varsigma} \tag{39}$$

where $\varsigma \in (0.5, 1]$ denotes the *forgetting rate* and is used to control the forgetting speed in the earlier stage of learning; $\eta \geq 0$ represents the *delay factor* to down-weight early iterations. The stochastic variational inference is guaranteed to converge to a local optimum of the ELBO if the step size satisfies the following conditions [16]:

$$\sum_r \lambda_r = \infty, \quad \sum_r \lambda_r^2 < \infty. \tag{40}$$

The stochastic variational inference of the HPY process mixture with ID distributions is summarized in Algorithm 1.

Algorithm 1 Stochastic Variational Inference of the HPY Process Mixture with ID Distributions

```
Choose the initial truncation level $K$ and $T$.
Initialize the parameters of the prior distributions: $a_k$, $b_k$, $a'_{jt}$, $b'_{jt}$, $u_{kl}$ and $v_{kl}$.
Set the step-size $\lambda_r$ as $\lambda_r = (\eta + r)^{-\varsigma}$.
while TRUE do
    Sample a data point $\vec{X}_{jn}$ uniformly from the $j$-th group of the data set: $n \sim \text{Unif}(1, \ldots, N_j)$.
    Update the local variational distribution $q(\vec{Z}_{jn})$ using Eq.(30).
    Update the current estimate of the global variational distributions using Eqs.(31), (33) ∼(35).
end while
```

C. LEARNING OF HPY PROCESS MIXTURE MODEL WITH GID DISTRIBUTIONS WITH FEATURE SELECTION

This subsection is devoted to learn the HPY process mixture model with GID distributions with stochastic variational inference. First, the variational posterior distribution $q(\Theta)$ is factorized into the products of independent factors by the mean-field assumption as

$$q(\Theta) = q(\vec{Z})q(\vec{W})q(\vec{\phi})q(\vec{\pi}')q(\vec{\varpi}')q(\vec{\alpha})q(\vec{\beta})q(\vec{\alpha}')q(\vec{\beta}')q(\vec{\epsilon}) \tag{41}$$

where $\Theta = \{\vec{Z}, \vec{W}, \vec{\phi}, \vec{\pi}', \vec{\varpi}', \vec{\alpha}, \vec{\beta}, \vec{\alpha}', \vec{\beta}', \vec{\epsilon}\}$ is the set of latent and random variables in the model. Given a data set $\mathcal{X} = \vec{X}_{1:N}$ that is divided into M groups. At the rth iteration, we uniformly sample a data instance $\vec{X}_{jn}$ with $n \sim \text{Unif}(1, \ldots, N_j)$ in group j. Then, we can update the variational solutions to the local variational distributions $q(\vec{Z}_{jn})$ and $q(\vec{\phi}_{jn})$ while the global variational distributions remain fixed to their values at the $(r-1)$th iteration as (See B for details)

$$q(\vec{Z}_{jn}) = \prod_{t=1}^{T} \rho_{jnt}^{Z_{jnt}}, \tag{42}$$

$$q(\vec{\phi}_{jn}) = \prod_{l=1}^{D} \varphi_{jnl}^{\phi_{jnl}} (1 - \varphi_{jnl})^{1-\phi_{jnl}} \tag{43}$$

In the following step, the global variational distributions $q(\vec{W})$, $q(\vec{\epsilon})$, $q(\vec{\pi}')$, $q(\vec{\varpi}')$, $q(\vec{\alpha})$, $q(\vec{\beta})$, $q(\vec{\alpha}')$ and $q(\vec{\beta}')$ are updated for the current r-th iteration as

$$q^{(r)}(\vec{W}) = \prod_{j=1}^{M}\prod_{t=1}^{T}\prod_{k=1}^{K} (\vartheta_{jtk}^{(r)})^{W_{jtk}^{(r)}} \tag{44}$$

$$q^{(r)}(\vec{\epsilon}) = \prod_{l=1}^{D} \mathrm{Dir}(\epsilon_l^{(r)}|\vec{\zeta}^{*(r)}) \tag{45}$$

$$q^{(r)}(\vec{\pi}') = \prod_{j=1}^{M}\prod_{t=1}^{T} \mathrm{Beta}(\pi_{jt}'^{(r)}|c_{jt}'^{(r)}, d_{jt}'^{(r)}) \tag{46}$$

$$q^{(r)}(\vec{\varpi}') = \prod_{k=1}^{K} \mathrm{Beta}(\varpi_k'^{(r)}|c_k^{(r)}, d_k^{(r)}) \tag{47}$$

$$q^{(r)}(\vec{\alpha}) = \prod_{k=1}^{K}\prod_{l=1}^{D} \mathcal{G}(\alpha_{kl}^{(r)}|\tilde{u}_{kl}^{(r)}, \tilde{v}_{kl}^{(r)}) \tag{48}$$

$$q^{(r)}(\vec{\alpha}') = \prod_{l=1}^{D} \mathcal{G}(\alpha_l'^{(r)}|\tilde{u}_l'^{(r)}, \tilde{v}_l'^{(r)}) \tag{49}$$

$$q^{(r)}(\vec{\beta}) = \prod_{k=1}^{K}\prod_{l=1}^{D} \mathcal{G}(\beta_{kl}^{(r)}|\tilde{g}_{kl}^{(r)}, \tilde{h}_{kl}^{(r)}) \tag{50}$$

$$q^{(r)}(\vec{\beta}') = \prod_{l=1}^{D} \mathcal{G}(\beta_l'^{(r)}|\tilde{g}_l'^{(r)}, \tilde{h}_l'^{(r)}) \tag{51}$$

where the hyperparameters of the above global variational distributions at the rth iteration can be calculated by

$$\vartheta_{jtk}^{(r)} = \vartheta_{jtk}^{(r-1)} + \lambda_r \partial \vartheta_{jtk}^{(r)}, \quad \vec{\zeta}^{*(r)} = \vec{\zeta}^{*(r-1)} + \lambda_r \partial \vec{\zeta}^{*(r)} \tag{52}$$

$$c_{jt}'^{(r)} = c_{jt}'^{(r-1)} + \lambda_r \partial c_{jt}'^{(r)}, \quad d_{jt}'^{(r)} = d_{jt}'^{(r-1)} + \lambda_r \partial d_{jt}'^{(r)} \tag{53}$$

$$c_k^{(r)} = c_k^{(r-1)} + \lambda_r \partial c_k^{(r)}, \quad d_k^{(r)} = d_k^{(r-1)} + \lambda_r \partial d_k^{(r)} \tag{54}$$

$$\tilde{u}_{kl}^{(r)} = \tilde{u}_{kl}^{(r-1)} + \lambda_r \partial \tilde{u}_{kl}^{(r)}, \quad \tilde{v}_{kl}^{(r)} = \tilde{v}_{kl}^{(r-1)} + \lambda_r \partial v_{kl}^{(r)} \tag{55}$$

$$\tilde{u}_l'^{(r)} = \tilde{u}_l'^{(r-1)} + \lambda_r \partial \tilde{u}_l'^{(r)}, \quad \tilde{v}_l'^{(r)} = \tilde{v}_l'^{(r-1)} + \lambda_r \partial v_l'^{(r)} \tag{56}$$

$$\tilde{g}_{kl}^{(r)} = \tilde{g}_{kl}^{(r-1)} + \lambda_r \partial \tilde{g}_{kl}^{(r)}, \quad \tilde{h}_{kl}^{(r)} = \tilde{h}_{kl}^{(r-1)} + \lambda_r \partial h_{kl}^{(r)} \tag{57}$$

$$\tilde{g}_l'^{(r)} = \tilde{g}_l'^{(r-1)} + \lambda_r \partial \tilde{g}_l'^{(r)}, \quad \tilde{h}_l'^{(r)} = \tilde{h}_l'^{(r-1)} + \lambda_r \partial h_l'^{(r)} \tag{58}$$

where we adopt the same step size λ_r as defined in Eq.(39) with constraints that are described in Eq.(40). The stochastic variational inference of the HPY process mixture model with GID distributions is summarized in Algorithm 2.

Algorithm 2 Stochastic Variational Inference of the HPY Process Mixture with GID Distributions

```
1: Choose the initial truncation levels $K$ and $T$.
2: Initialize the parameters of the prior distributions: $a_k, b_k$, $a'_{jt}, b'_{jt}, u_{kl}, v_{kl}, g_{kl}, h_{kl}, u'_l, v'_l, g'_l$ and $h'_l$.
3: Set the step-size $\lambda_r$ as $\lambda_r = (\eta + r)^{-\varsigma}$.
4: while TRUE do
5:     Sample a data point $\vec{X}_{jn}$ uniformly from the $j$th group of the data set: $n \sim \mathrm{Unif}(1, \ldots, N_j)$.
6:     Update the local variational distributions $q(\vec{Z}_{jn})$ and $q(\vec{\phi}_{jn})$ using Eqs.(42) and (43).
7:     Update the current estimate of the global variational distributions using Eqs.(44)∼(51).
8: end while
```

V. EXPERIMENTAL RESULTS

In this section, we first validate the proposed two HPY process mixture models with stochastic variational inference learning through synthetic data sets. Then, we apply our models to a challenging application namely video background subtraction. The goal of the synthetic data is to investigate the accuracy of the stochastic variational inference for learning these two models, in terms of parameters estimation. In the real application of video background subtraction, the proposed approach is compared with other well-defined mixture modeling based background subtraction approaches to show its advantages. Since the HPY process mixture model can be considered as a hierarchical infinite mixture model, the developed HPY process mixture model with ID distributions can be referred to as the hierarchical infinite ID mixture model (HIn-IDMM), whereas the HPY process mixture of GID distributions with feature selection can be referred to as the hierarchical infinite GID mixture model with feature selection (HIn-GIDMM). In our experiments, for both HIn-IDMM and HIn-GIDMM, we initialize the global truncation level K and the group truncation level T as 120 and 60, respectively. The parameters ς and η of the learning rate are set to 0.80 and 64, respectively. The hyperparameters of the stick-breaking weights are initialized as: $(a'_{jt}, b'_{jt}, a_k, b_k) = (0.1, 0.5, 0.1, 0.5)$. For HIn-IDMM, we initialize its hyperparameters u_{kl} and v_{kl} as 0.1 and 0.05, respectively. For HIn-GIDMM, its hyperparameters are initialized as $(u_{kl}, v_{kl}, g_{kl}, h_{kl}, u'_l, v'_l, g'_l, h'_l) = (0.1, 0.05, 0.1, 0.05, 0.1, 0.05, 0.1, 0.05)$. The hyperparameters ξ_1 and ξ_2 of the feature saliency are both initialized to 0.5.

TABLE 1. True parameters for generating the synthetic data set. N denotes the total number of elements, N_k denotes the number of elements in cluster k and π_k indicates the mixing proportion for cluster k.

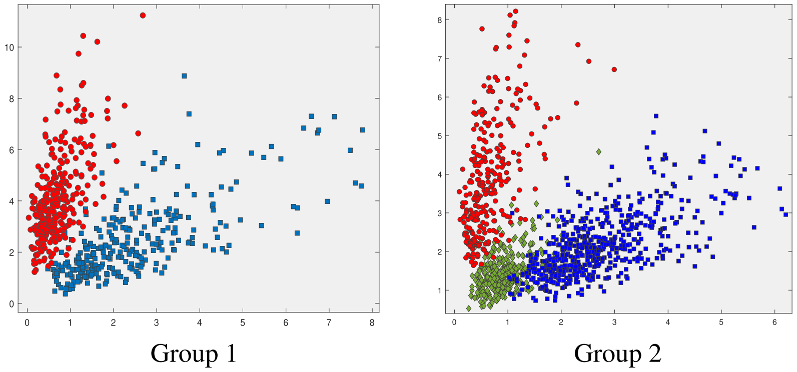
	k	N_k	α_{k1}	α_{k2}	α_{k3}	π_k
Group 1	1	300	10	10	5	0.50
($N = 600$)	2	300	5	30	8	0.50
Group 2	1	300	12	20	15	0.25
($N = 1200$)	2	300	5	30	8	0.25
	3	600	25	20	10	0.50

A. SYNTHETIC DATA SETS

1) SYNTHETIC ID MIXTURES

To validate the proposed variational inference method for learning HIn-IDMM, we generate a two-dimensional synthetic data set that can be divided into two groups. The first group contains two clusters (C_{11} and C_{12}) of data points that are randomly sampled from two ID distributions with different parameters. The second group has three clusters (C_{21}, C_{22} and C_{23}) of data points that are generated based on three different ID distributions. In order to link these two groups statistically, the data points in C_{12} and C_{22} are sampled according to the same ID density (i.e., the ID distribution with same parameters). The detailed setting of parameters for generating this synthetic data set can be viewed in Table 1.

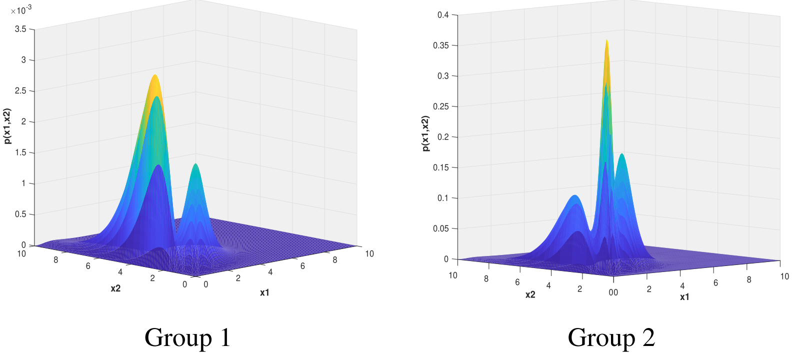
The two groups of synthetic data can be viewed in Fig. 3. The results of parameter estimation for each group of the



Group 1

Group 2

FIGURE 3. The scatter plot of data points of the synthetic data sets. Group 1: Two clusters; Group 2: Three clusters.



Group 1

Group 2

FIGURE 4. The resulting mixture model learned by HIn-IDMM.

synthetic data set are shown in Table 2, based on the stochastic variational inference learning algorithm as developed in Section IV-B. According to this table, the proposed learning algorithm is able to effectively learn HIn-IDMM with estimated values of parameters that are very close to the true ones. The resulting mixture models with estimated parameters for each group are shown in Fig. 4.

Since we have truncated the number of mixture components to 60 within each group in the initialization step, the proposed HIn-IDMM estimated the correct number of mixture components by removing the redundant components with very small mixing coefficients. In the proposed framework, the number of mixture components in each group was identified by removing the components with the estimated mixing coefficients $\widehat{\pi}_k$ that were close to 0 (less than 10^{-4}). As we can observe from Table 2, our algorithm has detected the correct number of mixture components for each group with accurate mixing proportions.

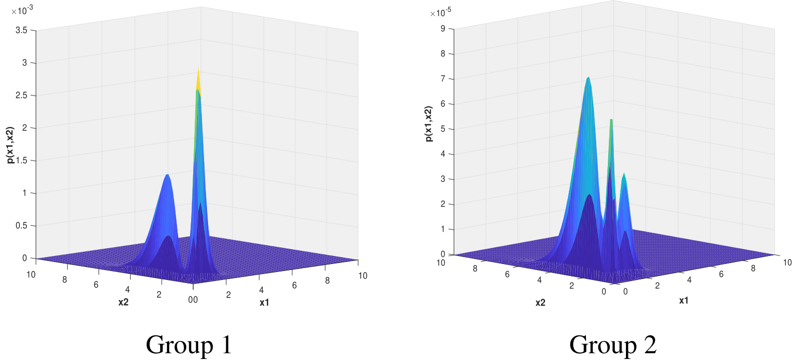
Furthermore, another experiment was performed to test the effectiveness of the proposed stochastic variational algorithm for learning HIn-IDMM with a larger data set that contains 6 million data points in total as described in Table 3. The estimated values of the parameters for generating this data set is given in Table 4. By comparing the results shown in this table with the true parameters as provided in Table 3, it is clear that the proposed stochastic variational algorithm can learn HIn-IDMM accurately for large-scale data set.

TABLE 2. Average estimated parameters of the synthetic data set in 15 runs by the proposed stochastic variational inference method. Values are rounded to two digits after the decimal point. The numbers in parenthesis are the standard deviation of the corresponding quantities.

k	$\widehat{\alpha}_{k1}$	$\widehat{\alpha}_{k2}$	$\widehat{\alpha}_{k3}$	$\widehat{\pi}_k$
1	9.94 (0.06)	9.83 (0.12)	5.12 (0.21)	0.499 (0.003)
2	5.11 (0.04)	31.15 (0.33)	7.90 (0.09)	0.501 (0.001)
1	11.73 (0.23)	20.84 (0.17)	14.90 (0.06)	0.252 (0.002)
2	4.90 (0.02)	29.14 (0.43)	7.90 (0.15)	0.249 (0.001)
3	24.56 (0.28)	20.74 (0.19)	10.57 (0.08)	0.499 (0.003)

2) SYNTHETIC GID MIXTURES

In this part, to evaluate the effectiveness of the proposed learning algorithm in terms of both parameter estimation and feature selection, we sample a 10-dimensional synthetic data set from HIn-GIDMM with two relevant features and eight irrelevant features. Please notice that, the geometric transformation is performed as described in Section III-B, in order to generate data with indecent features. Thus, the two relevant features of our data are generated in the transformed space from a mixture of inverted Beta distributions, whereas the eight irrelevant features are generated according to a common inverted Beta distribution IB(2, 5). This synthetic



Group 1

Group 2

FIGURE 5. The resulting mixture model learned by HIn-GIDMM.

TABLE 3. True parameters for generating the large-scale data set. N denotes the total number of elements, N_k denotes the number of elements in cluster k and π_k indicates the mixing proportion for cluster k.

	k	N_k	α_{k1}	α_{k2}	α_{k3}	π_k
Group 1	1	1 m	15	10	5	0.50
(N=2 m)	2	1 m	8	25	8	0.50
Group 2	1	1 m	10	18	5	0.25
(N=4 m)	2	1 m	8	25	8	0.25
	3	2 m	20	15	10	0.50

TABLE 4. Average estimated parameters of the large-scale data set in 15 runs by the proposed stochastic variational inference method. The numbers in parenthesis are the standard deviation of the corresponding quantities.

k	$\widehat{\alpha}_{k1}$	$\widehat{\alpha}_{k2}$	$\widehat{\alpha}_{k3}$	$\widehat{\pi}_k$
1	14.35 (0.31)	10.23 (0.12)	4.59 (0.17)	0.503 (0.005)
2	8.03 (0.22)	26.56 (0.23)	7.96 (0.09)	0.497 (0.003)
1	9.81 (0.34)	17.69 (0.16)	5.10 (0.02)	0.252 (0.002)
2	7.95 (0.10)	24.38 (0.15)	8.15 (0.21)	0.246 (0.004)
3	19.82 (0.32)	15.73 (0.26)	9.20 (0.11)	0.502 (0.002)

TABLE 5. True parameters for generating the synthetic data set. N denotes the total number of elements, N_k denotes the number of elements in cluster k and π_k indicates the mixing proportion for cluster k.

	k	N_k	α_{k1}	α_{k2}	β_{k1}	β_{k2}	π_k
Group 1	1	500	15	10	15	15	0.50
(N=1000)	2	500	5	25	10	10	0.50
Group 2	1	1000	10	15	20	16	0.25
(N=4000)	2	1000	5	25	10	10	0.25
	3	2000	25	14	17	22	0.50

data set contains 3,000 data points from four different clusters and can be divided into two groups. The first group includes 1,000 data points from two clusters. The second group has 2,000 data points in total from three clusters. The second cluster in both groups are generated using the GID density function with same parameters.

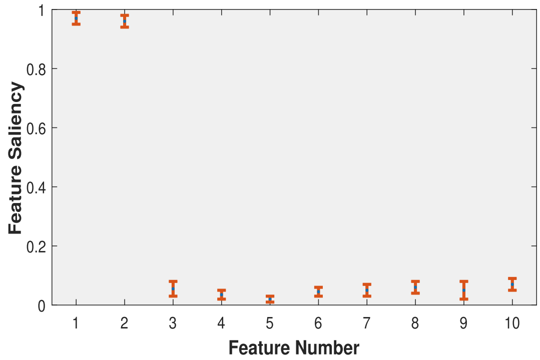


FIGURE 6. Average feature saliency obtained by HIn-GIDMM.

TABLE 6. Average estimated parameters of the synthetic data set in 15 runs by the proposed stochastic variational inference method. The numbers in parenthesis are the standard deviation of the corresponding quantities.

k	$\widehat{\alpha}_{k1}$	$\widehat{\alpha}_{k2}$	$\widehat{\beta}_{k1}$	$\widehat{\beta}_{k2}$	$\widehat{\pi}_k$
1	14.79 (0.27)	9.87 (0.14)	15.22 (0.39)	15.19 (0.25)	0.500 (0.002)
2	5.03 (0.12)	25.90 (0.34)	10.42 (0.07)	9.76 (0.18)	0.500 (0.003)
1	10.13 (0.29)	14.65 (0.31)	21.01 (0.43)	16.53 (0.11)	0.248 (0.002)
2	4.91 (0.07)	26.06 (0.41)	10.15 (0.26)	10.05 (0.14)	0.251 (0.001)
3	26.15 (0.23)	14.23 (0.13)	17.36 (0.32)	21.79 (0.49)	0.501 (0.004)

Table 5 shows the parameters of the distributions representing the relevant features for generating this data set. The estimated parameters for each group of the synthetic data set are illustrated in Table 6, using the proposed stochastic variational inference algorithm. Based on this table, we can observe that the proposed learning algorithm can accurately estimate the parameters of the model representing relevant features, and its mixing coefficients. The resulting mixture models with estimated parameters for each group are shown in Fig. 5. The average results of the estimated features saliencies of all 10 features for the synthetic data set are shown in Fig. 6, based on 15 runs. It obviously shows that features 1 and 2 have been assigned a high degree of relevance (greater than 0.9), which matches the ground-truth.

B. VIDEO BACKGROUND SUBTRACTION

Video background subtraction is the process of discriminating foreground subjects from the background in a sequence of video frames with static cameras. It is a critical problem in video analysis and has been applied as an interest detector

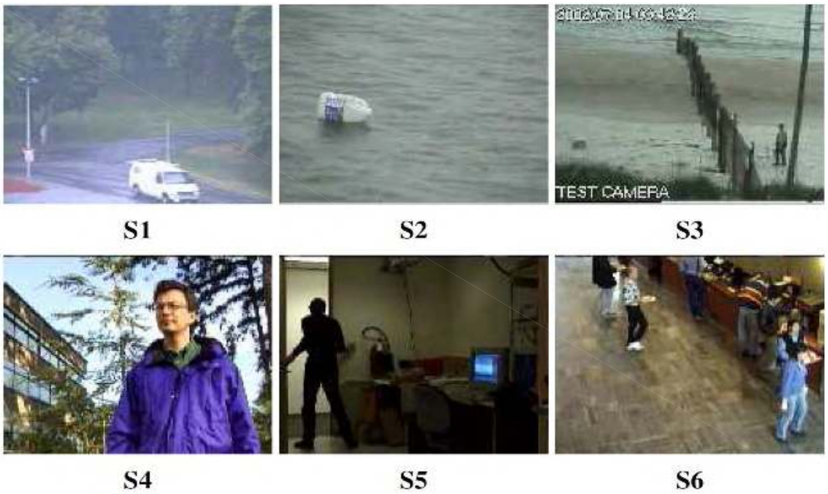


FIGURE 7. Sample frames from each video sequence.

in various higher level problems, such as video surveillance, human motion analysis, object tracking and traffic monitoring [26]–[28]. Among various approaches that have been proposed for video background subtraction in the past, the ones based on mixture models have shown their merits and promising performance [1], [2], [29]–[31]. In mixture modeling approaches, each pixel is represented by a mixture of density functions, and the goal is to differentiate whether the testing pixel belongs to the background or some foreground objects. Thus, background subtraction approaches based on mixture models are considered as pixel-level evaluations, and they are more robust and able to handle multi-modal background distributions in contrast to other approaches. In this experiment, we apply the proposed HIn-IDMM and HIn-GIDMM to video background subtraction using a statistical framework.

1) METHODOLOGY

Suppose that we have observed a sequence of M frames $\mathcal{X}^1, \ldots, \mathcal{X}^M$, where each frame $\mathcal{X}$ contains N pixels $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$. In our approach, each frame can be considered as a group, thus each pixel in the frame can be modeled as a mixture of infinite ID or GID distributions, where mixture components are shared among groups (i.e., frames). This setting satisfies the construction of the HPY process mixture model. Another factor that impacts the performance of back subtraction is the choice of features for representing each pixel. Among various types of features, color features are invariant with respect to brightness changes, and therefore illumination changes and shadows. Although the RGB color space is a popular choice in background subtraction [29], [31], [32], other color spaces such as HSV [33] or YCbCr [34] are also exploited. In our case, we adopt the combination of 9 individual color components, taken from the 3 color spaces (RGB, HSV and YCbCr) to build the background model, where each color component falls into the range $[0, 255]$. After computing color features for each pixel in the given frame, the background model is learned using the proposed HPY process mixture models through stochastic variational inference as developed in Section IV. In our mixture model, some of the mixture components are used to model background whereas other components are exploited to model foreground objects. Therefore, the last step in our approach is to determine whether the pixel $\vec{X}_i$ belongs to foreground or background. Here, we adopt the assumption that a mixture component belongs background if it occurs more frequently (high mixing probability π_k) and does not vary significantly (low standard deviation σ_k) [29]. Based on this assumption, all estimated components are ranked according to the ratio $\pi_k/\|\sigma_k\|$ and the first B components are chosen as background components, where B is obtained by

$$B = \arg\min_{s} \sum_{k=1}^{s} \pi_k > H \tag{59}$$

where H is the threshold that represents the minimum portion of the data that is accounted for the background. Thus, we can perform background subtraction for an observed frame by determining if the testing pixel $\vec{X}_i$ belongs to one of the components in set B. In our experiments, different values of H were tested, promising performance was obtained when H was set to $[0.75, 0.8]$ for different tested video sequences. The background model would normally be unimodal if we set H to a very small value. If a higher value of H was chosen, a multimodal distribution caused by repeated background motion may lead to multiple colors in the background model, then a transparent effect may be obtained which allows the background to accept two or more individual colors.

2) DATA SETS

The performance of the proposed background subtraction approach is evaluated through six publicly available video sequences that have been used previously in [35], [36]. These video sequences have different characteristics and are selected to assess the effectiveness of our approach under various scenarios, such as: illumination changes, dynamic backgrounds, etc. The description of each video sequence is listed as following:

- **S1**: In this video sequence, it shows some vehicles driving in the rain;
- **S2**: In this video sequence, a plastic drum is floating on the surface of sea;
- **S3**: A person is walking on a beach in this video sequence;
- **S4**: A person is walking in front of swaying trees in this video sequence;
- **S5**: In this video sequence, people enter and leave a room with light switches on and off;
- **S6**: The video sequence consists of several minutes of an overhead view of a cafeteria.

The size of each frame in these video sequences is 160×120 and sample frames can be viewed in Fig. 7.

3) EXPERIMENTAL RESULTS

Since the performance of our background subtraction approach is evaluated on pixel-level, it is straightforward to consider the detection of foreground objects as a binary classification problem for each pixel. In our experiment, the threshold H is set to $[0.75, 0.8]$ for different videos. Representative results of our background subtraction approaches through HIn-IDMM and HIn-GIDMM can be visualized in Fig. 8, in terms of foreground masks. As illustrated in this figure, both HIn-GIDMM and HIn-IDMM are able to obtain promising results for all sequences, which demonstrate the effectiveness of using hierarchical Pitman-Yor process mixture models to the problem of background subtraction. Even though both HIn-GIDMM and HIn-IDMM can identify foreground objects clearly as demonstrated in Fig. 8, as we may visually notice, HIn-GIDMM has acquired better performance than HIn-IDMM does, in terms of better robustness against noise, particularly for sequences **S1** $\sim$ **S4** due to dynamic backgrounds. This fact shows the ability of the proposed approach to model dynamic backgrounds, and also demonstrate the advantages of using HIn-GIDMM together with a feature selection scheme to model backgrounds compared to HIn-IDMM in which all features are used. Nevertheless, HIn-IDMM is more computational effective than HIn-GIDMM, since no extra computational source is required for HIn-IDMM to spend on the feature selection process. Furthermore, the appealing results of video sequence $S5$ obtained by both HIn-GIDMM and HIn-IDMM demonstrate the robustness of our approach against illumination changes.

We also evaluate the performance of our background subtraction approach quantitatively in terms of F-Measure, which is computed through both *recall* and *precision* as

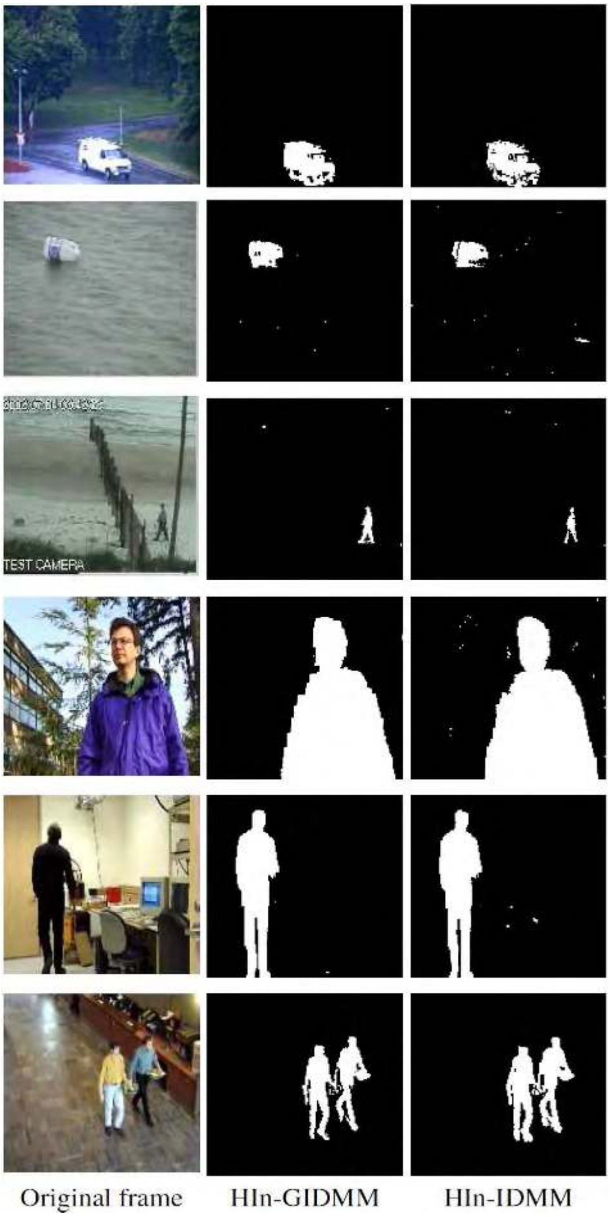


FIGURE 8. Foreground masks obtained by HIn-GIDMM and HIn-IDMM for each video sequence.

$$\text{Recall} = \frac{\text{number of correctly identified foreground pixels}}{\text{number of foreground pixels in ground truth}}$$

$$\text{Precision} = \frac{\text{number of correctly identified foreground pixels}}{\text{number of foreground pixels detected}}$$

$$\text{F-measure} = 2\frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

The F-measure is the harmonic average of the precision and recall, where the F-measure reaches its best value at 1 (i.e., perfect precision and recall) and worst at 0. In our experiment,

TABLE 7. The average results of background subtraction of each approach, in terms of F-measures.

Videos	GMM	In-GMM	In-DMM	HIn-IDMM	HIn-GIDMM
S1	0.653	0.785	0.769	0.817	0.832
S2	0.618	0.760	0.754	0.798	0.803
S3	0.633	0.774	0.762	0.825	0.839
S4	0.598	0.763	0.747	0.781	0.799
S5	0.335	0.613	0.583	0.601	0.617
S6	0.581	0.759	0.737	0.776	0.787

the results of recall and precision are calculated based on the averages over all measured frames for each video sequence.

For comparison, we have also applied other three mixture-based background subtraction approaches: the Gaussian mixture model (GMM) [30], the infinite Gaussian mixture model (In-GMM) [2], and the infinite Dirichlet mixture model (In-DMM) [31]. The hyperparameters associated with each tested background subtraction approach were initialized to the same values as in their original works. The threshold H was set to the same value for all tested approaches for the sake of fair comparison.

The comparison results obtained by the testing approaches for each video sequence are shown Table 7, in terms of F-measures. According to the results shown in this table, both the proposed HIn-IDMM and HIn-GIDMM have provided better performance than other tested approaches for most video sequences with higher F scores. It demonstrates the advantages of using the HPY process mixtures to model backgrounds. Comparably worse performance has obtained by both HIn-IDMM and HIn-GIDMM for the video sequence **S5**, with lower F scores compared to the results for other video sequences. This fact shows that the proposed background subtraction approach has sort of limitations in dealing with videos that contain light switch. Even though In-GMM has obtained better result than HIn-IDMM, and comparable performance to HIn-GIDMM for the video sequence **S5**, HIn-GIDMM is still able to provide the highest F score among all tested approaches.

VI. CONCLUSION

In this paper, we proposed a unified nonparametric Bayesian framework for simultaneous clustering and feature selection in the case of positive vectors of visual descriptors. Our statistical framework is based on inverted Dirichlet-based distributions as parent densities to describe the data and the consideration of hierarchical Pitman-Yor process prior as an alternative to the widely used Dirichlet process. We derived elegant efficient inference algorithms using the recently proposed stochastic variational Bayes learning approach, and we examined the merits of our models using both synthetic histograms and a real application namely video background subtraction. The experimental results suggest that our nonparametric Bayesian framework is promising and offers significant advantages if we take into account comparable mixture-based techniques especially when integrating feature selection to improve generalization capabilities.

APPENDIX A
STOCHASTIC VARIATIONAL INFERENCE OF THE HPY PROCESS MIXTURE MODEL WITH ID DISTRIBUTIONS

A. THE OPTIMIZATION OF LOCAL VARIATIONAL DISTRIBUTION

To update the local variational distribution $q(\vec{Z}_{jn})$ after sampling the data point $\vec{X}_{jn}$, the global variational distributions are remained to their values at the $(r-1)$th iteration. The variational solution to $q(\vec{Z}_{jn})$ can be calculated by

$$q(\vec{Z}_{jn}) = \prod_{t=1}^{T} \rho_{jnt}^{Z_{jnt}} \tag{60}$$

where we have

$$\rho_{jnt} = \frac{\exp(\widetilde{\rho}_{jnt})}{\sum_{s=1}^{T}\exp(\widetilde{\rho}_{jns})} \tag{61}$$

and

$$\begin{aligned}\widetilde{\rho}_{jnt} = \sum_{k=1}^{K}\langle W_{jtk}^{(r-1)}\rangle[\widetilde{\mathcal{R}}_k^{(r-1)} + \sum_{l=1}^{D}(\bar{\alpha}_{kl}^{(r-1)} - 1)\ln X_{jnl} \\ + \sum_{s=1}^{t-1}\langle \ln(1 - \pi_{js}^{\prime(r-1)})\rangle + \langle \ln \pi_{jt}^{\prime(r-1)}\rangle \\ -(\sum_{l=1}^{D+1}\bar{\alpha}_{kl}^{(r-1)})\ln(1 + \sum_{l=1}^{D} X_{jnl})]\end{aligned} \tag{62}$$

where $\widetilde{\mathcal{R}}_k = \left\langle \ln \frac{\Gamma(\sum_{l=1}^{D+1}\bar{\alpha}_{kl})}{\prod_{l=1}^{D+1}\Gamma(\bar{\alpha}_{kl})} \right\rangle$ is intractable and thus no closed-form solution can be found. Therefore, a second-order Taylor expansion is used to approximate its value.

B. THE OPTIMIZATION OF GLOBAL VARIATIONAL DISTRIBUTIONS

The following step is to optimize the global variational distributions $q^{(r)}(\vec{\pi}')$, $q^{(r)}(\vec{W})$, $q^{(r)}(\vec{\varpi}')$ and $q^{(r)}(\vec{\alpha})$ for the current iteration. First, we need to calculate intermediate global variational solutions based on N_j replicates of the sampled data $\vec{X}_{jn}$ in group j. As a result, the intermediate variational hyperparameters of the global variational variables are

$$\hat{\vartheta}_{jtk} = \frac{\exp(\widetilde{\vartheta}_{jtk})}{\sum_{s=1}^{K}\exp(\widetilde{\vartheta}_{jts})} \tag{63}$$

$$\begin{aligned}\widetilde{\vartheta}_{jtk} = N_j\langle Z_{jnt}\rangle\Bigg[\widetilde{\mathcal{R}}_k - \Bigg(\sum_{l=1}^{D+1}\bar{\alpha}_{kl}\Bigg)\ln\Bigg(1 + \sum_{l=1}^{D} X_{jnl}\Bigg) \\ + \sum_{l=1}^{D}(\bar{\alpha}_{kl} - 1)\ln X_{jnl}\Bigg] + \langle \ln \varpi_k'\rangle + \sum_{s=1}^{k-1}\langle \ln(1 - \varpi_s')\rangle\end{aligned} \tag{64}$$

$$\hat{c}_{jt}' = 1 + N_j\langle Z_{jnt}\rangle - a_{jt}', \quad \hat{c}_k = 1 + \sum_{j=1}^{M}\sum_{t=1}^{T}\langle W_{jtk}\rangle - a_k \tag{65}$$

$$\hat{d}'_{jt} = b'_{jt} + ta'_{jt} + N_j \sum_{s=t+1}^{T} \langle Z_{jns} \rangle \tag{66}$$

$$\hat{d}_k = b_k + ka_k + \sum_{j=1}^{M} \sum_{t=1}^{T} \sum_{s=k+1}^{K} \langle W_{jts} \rangle \tag{67}$$

$$\begin{aligned}\hat{u}^*_{kl} = u_{kl} + N_j \sum_{j=1}^{M} \sum_{t=1}^{T} \langle W_{jtk} \rangle \langle Z_{jnt} \rangle \Big[\psi(\sum_{l=1}^{D+1} \bar{\alpha}_{kl}) - \psi(\bar{\alpha}_{kl}) \\ + \sum_{s \neq l}^{D+1} \bar{\alpha}_{ks} \psi'(\sum_{l=1}^{D+1} \bar{\alpha}_{kl})(\langle \ln \alpha_{ks} \rangle - \ln \bar{\alpha}_{ks}) \Big] \bar{\alpha}_{kl}\end{aligned} \tag{68}$$

$$\hat{v}^*_{kl} = v_{kl} - N_j \sum_{j=1}^{M} \sum_{t=1}^{T} \langle W_{jtk} \rangle \langle Z_{jnt} \rangle [\ln X_{jnl} - \ln(1 + \sum_{l=1}^{D} X_{jnl})] \tag{69}$$

where $\psi(\cdot)$ is the digamma function.

Then, we can obtain the noisy but unbiased natural gradient of the ELBO with respect to each hyperparameter as

$$\partial \vartheta^{(r)}_{jtk} = \hat{\vartheta}^{(r)}_{jtk} - \vartheta^{(r-1)}_{jtk} \tag{70}$$

$$\partial c'^{(r)}_{jt} = \hat{c}'^{(r)}_{jt} - c'^{(r-1)}_{jt}, \quad \partial d'^{(r)}_{jt} = \hat{d}'^{(r)}_{jt} - d'^{(r-1)}_{jt} \tag{71}$$

$$\partial c^{(r)}_{k} = \hat{c}^{(r)}_{k} - c^{(r-1)}_{k}, \quad \partial d^{(r)}_{k} = \hat{d}^{(r)}_{k} - d^{(r-1)}_{k} \tag{72}$$

$$\partial u^{*(r)}_{kl} = \hat{u}^{*(r)}_{kl} - u^{*(r-1)}_{kl}, \quad \partial v^{*(r)}_{kl} = \hat{v}^{*(r)}_{kl} - v^{*(r-1)}_{kl} \tag{73}$$

By substituting these natural gradients into (32), (36)~(38), we then obtain the required hyperparameters for updating the global variational distributions as in (31), (33)~(35).

APPENDIX B
STOCHASTIC VARIATIONAL INFERENCE OF THE HPY PROCESS MIXTURE MODEL WITH GID DISTRIBUTIONS

A. THE OPTIMIZATION OF LOCAL VARIATIONAL DISTRIBUTIONS

The first step to learn the HPY process mixture model with feature selection using stochastic variational inference is to optimize local variational distributions $q(\vec{\phi}_{jn})$ and $q(\vec{Z}_{jn})$ after sampling the data instance $\vec{X}_{jn}$ at the rth iteration, while the global variational dilutions are fixed to their values at the $(r-1)$th iteration. Then, we have

$$q(\vec{\phi}_{jn}) = \prod_{l=1}^{D} \varphi_{jnl}^{\phi_{jnl}} (1 - \varphi_{jnl})^{1-\phi_{jnl}} \tag{74}$$

$$q(\vec{Z}_{jn}) = \prod_{t=1}^{T} \rho_{jnt}^{Z_{jnt}}, \tag{75}$$

where the associated hyperparameters are updated as follows

$$\varphi_{jnl} = \frac{\exp(\widetilde{\varphi}_{jnl})}{\exp(\widetilde{\varphi}_{jnl}) + \exp(\widehat{\varphi}_{jnl})} \tag{76}$$

$$\begin{aligned}\widetilde{\varphi}_{jnl} = \langle \ln \epsilon^{(r-1)}_{l_1} \rangle + \sum_{t=1}^{T} \sum_{k=1}^{K} [\widetilde{\mathcal{R}}^{(r-1)}_{kl} + (\bar{\alpha}^{(r-1)}_{kl} - 1) \ln X_{jnl} \\ - (\bar{\alpha}^{(r-1)}_{kl} + \bar{\beta}^{(r-1)}_{kl}) \ln(1 + X_{jnl})] \langle Z_{jnt} \rangle \langle W^{(r-1)}_{jtk} \rangle\end{aligned} \tag{77}$$

$$\begin{aligned}\widehat{\varphi}_{jnl} = (\bar{\alpha}'^{(r-1)}_{l} - 1) \ln X_{jnl} + \langle \ln \epsilon^{(r-1)}_{l_2} \rangle + \widetilde{\mathcal{R}}'^{(r-1)}_{l} \\ - (\bar{\alpha}'^{(r-1)}_{l} + \bar{\beta}'^{(r-1)}_{l}) \ln(1 + X_{jnl})\end{aligned} \tag{78}$$

$$\rho_{jnt} = \frac{\exp(\widetilde{\rho}_{jnt})}{\sum_{s=1}^{T} \exp(\widetilde{\rho}_{jns})} \tag{79}$$

$$\begin{aligned}\widetilde{\rho}_{jnt} = \langle \ln \pi'^{(r-1)}_{jt} \rangle + \sum_{k=1}^{K} \langle W^{(r-1)}_{jtk} \rangle \sum_{l=1}^{D} \langle \phi_{jnl} \rangle [\widetilde{\mathcal{R}}^{(r-1)}_{kl} \\ + (\bar{\alpha}^{(r-1)}_{kl} - 1) \ln X_{jnl} - (\bar{\alpha}^{(r-1)}_{kl} + \bar{\beta}^{(r-1)}_{kl}) \\ \times \ln(1 + X_{jnl})] + \sum_{s=1}^{t-1} \langle \ln(1 - \pi'^{(r-1)}_{js}) \rangle\end{aligned} \tag{80}$$

where $\widetilde{\mathcal{R}}_{kl} = \Big\langle \ln \frac{\Gamma(\alpha_{kl}+\beta_{kl})}{\Gamma(\alpha_{kl})\Gamma(\beta_{kl})} \Big\rangle$ and $\widetilde{\mathcal{R}}'_{l} = \Big\langle \ln \frac{\Gamma(\alpha'_{l}+\beta'_{l})}{\Gamma(\alpha'_{l})\Gamma(\beta'_{l})} \Big\rangle$ are intractable and we use second-order Taylor expansion to approximate their values.

B. THE OPTIMIZATION OF GLOBAL VARIATIONAL DISTRIBUTIONS

Then, we need to optimize the global variational distributions $q^{(r)}(\vec{\epsilon}), q^{(r)}(\vec{\pi}'), q^{(r)}(\vec{W}), q^{(r)}(\vec{\varpi}')$ and $q^{(r)}(\vec{\alpha}), q^{(r)}(\vec{\alpha}'), q^{(r)}(\vec{\beta})$ and $q^{(r)}(\vec{\beta}')$ for the current rth iteration. The intermediate variational hyperparameters of the global variational variables are obtained based on N_j replicates of the sampled data $\vec{X}_{jn}$ in group j as

$$\hat{\vartheta}_{jtk} = \frac{\exp(\widetilde{\vartheta}_{jtk})}{\sum_{s=1}^{K} \exp(\widetilde{\vartheta}_{jts})} \tag{81}$$

$$\begin{aligned}\widetilde{\sigma}_{jtk} = N_j \langle Z_{jnt} \rangle \sum_{l=1}^{D} \langle \phi_{jnl} \rangle [\widetilde{\mathcal{R}}_{kl} + (\bar{\alpha}_{kl} - 1) \ln X_{jnl} \\ - (\bar{\alpha}_{kl} + \bar{\beta}_{kl}) \ln(1 + X_{jnl})] + \sum_{s=1}^{k-1} \langle \ln(1 - \varpi'_s) \rangle \\ + \langle \ln \varpi'_k \rangle\end{aligned} \tag{82}$$

$$\hat{\zeta}^*_1 = \zeta_1 + N_j \sum_{j=1}^{M} \langle \phi_{jnl} \rangle, \quad \hat{\zeta}^*_2 = \zeta_2 + N_j \sum_{j=1}^{M} \langle 1 - \phi_{jnl} \rangle \tag{83}$$

$$\hat{c}'_{jt} = 1 + N_j \langle Z_{jnt} \rangle - a'_{jt}, \quad \hat{c}_k = 1 + \sum_{j=1}^{M} \sum_{t=1}^{T} \langle W_{jtk} \rangle - a_k \tag{84}$$

$$\hat{d}'_{jt} = b'_{jt} + ta'_{jt} + N_j \sum_{s=t+1}^{T} \langle Z_{jns} \rangle \tag{85}$$

$$\hat{d}_k = b_k + ka_k + \sum_{j=1}^{M} \sum_{t=1}^{T} \sum_{s=k+1}^{K} \langle W_{jts} \rangle \tag{86}$$

$$\begin{aligned}\hat{\hat{u}}_{kl} = u_{kl} + N_j \sum_{j=1}^{M} \sum_{t=1}^{T} \langle W_{jtk} \rangle \langle Z_{jnt} \rangle \langle \phi_{jnl} \rangle \bar{\alpha}_{kl} \big[\psi(\bar{\alpha}_{kl} + \bar{\beta}_{kl}) \\ - \psi(\bar{\alpha}_{kl}) + \bar{\beta}_{kl} \psi'(\bar{\alpha}_{kl} + \bar{\beta}_{kl})(\langle \ln \beta_{kl} \rangle - \ln \bar{\beta}_{kl}) \big]\end{aligned} \tag{87}$$

$$\hat{\hat{v}}_{kl} = v_{kl} - N_j \sum_{j=1}^{M} \sum_{t=1}^{T} \langle W_{jtk} \rangle \langle Z_{jnt} \rangle \langle \phi_{jnl} \rangle \ln \frac{X_{jnl}}{1 + X_{jnl}} \tag{88}$$

$$\begin{aligned}\hat{\tilde{g}}_{kl} = g_{kl} + N_j \sum_{j=1}^{M}\sum_{t=1}^{T} \langle W_{jtk}\rangle \langle Z_{jnt}\rangle \langle \phi_{jnl}\rangle \bar{\beta}_{kl}\big[\psi(\bar{\alpha}_{kl} + \bar{\beta}_{kl}) \\ - \psi(\bar{\beta}_{kl}) + \bar{\alpha}_{kl}\psi'(\bar{\alpha}_{kl} + \bar{\beta}_{kl})(\langle \ln \alpha_{kl}\rangle - \ln \bar{\alpha}_{kl})\big]\end{aligned} \tag{89}$$

$$\hat{\tilde{h}}_{kl} = h_{kl} - N_j \sum_{j=1}^{M}\sum_{t=1}^{T} \langle W_{jtk}\rangle \langle Z_{jnt}\rangle \langle \phi_{jnl}\rangle \ln \frac{1}{1 + X_{jnl}} \tag{90}$$

$$\begin{aligned}\hat{\tilde{u}}'_l = u'_l + N_j \sum_{j=1}^{M} \langle 1 - \phi_{jnl}\rangle \bar{\alpha}'_l\big[\psi(\bar{\alpha}'_l + \bar{\beta}'_l) - \psi(\bar{\alpha}'_l) \\ + \bar{\beta}'_l\psi'(\bar{\alpha}'_l + \bar{\beta}'_l)(\langle \ln \beta'_l\rangle - \ln \bar{\beta}'_l)\big]\end{aligned} \tag{91}$$

$$\hat{\tilde{v}}'_l = v'_l - N_j \sum_{j=1}^{M} \langle 1 - \phi_{jnl}\rangle \ln \frac{X_{jnl}}{1 + X_{jnl}} \tag{92}$$

$$\begin{aligned}\hat{\tilde{g}}'_l = g'_l + N_j \sum_{j=1}^{M} \langle 1 - \phi_{jnl}\rangle \bar{\beta}'_l\big[\psi(\bar{\alpha}'_l + \bar{\beta}'_l) - \psi(\bar{\beta}'_l) \\ + \bar{\alpha}'_l\psi'(\bar{\alpha}'_l + \bar{\beta}'_l)(\langle \ln \alpha'_l\rangle - \ln \bar{\alpha}'_l)\big]\end{aligned} \tag{93}$$

$$\hat{\tilde{h}}'_l = h'_l - N_j \sum_{j=1}^{M} \langle 1 - \phi_{jnl}\rangle \ln \frac{1}{1 + X_{jnl}} \tag{94}$$

Then, the noisy but unbiased natural gradients of the ELBO with respect to the hyperparameters of the global variational distributions can be calculated by

$$\partial \vartheta_{jtk}^{(r)} = \hat{\vartheta}_{jtk}^{(r)} - \vartheta_{jtk}^{(r-1)}, \quad \partial \vec{\zeta}^{*(r)} = \hat{\vec{\zeta}}^{*(r)} - \vec{\zeta}^{*(r-1)} \tag{95}$$

$$\partial c_{jt}^{\prime(r)} = \hat{c}_{jt}^{\prime(r)} - c_{jt}^{\prime(r-1)}, \quad \partial d_{jt}^{\prime(r)} = \hat{d}_{jt}^{\prime(r)} - d_{jt}^{\prime(r-1)} \tag{96}$$

$$\partial c_{k}^{(r)} = \hat{c}_{k}^{(r)} - c_{k}^{(r-1)}, \quad \partial d_{k}^{(r)} = \hat{d}_{k}^{(r)} - d_{k}^{(r-1)} \tag{97}$$

$$\partial \tilde{u}_{kl}^{(r)} = \hat{\tilde{u}}_{kl}^{(r)} - \tilde{u}_{kl}^{(r-1)}, \quad \partial \tilde{v}_{kl}^{(r)} = \hat{\tilde{v}}_{kl}^{(r)} - \tilde{v}_{kl}^{(r-1)} \tag{98}$$

$$\partial \tilde{u}_{l}^{\prime(r)} = \hat{\tilde{u}}_{l}^{\prime(r)} - \tilde{u}_{l}^{\prime(r-1)}, \quad \partial \tilde{v}_{l}^{\prime(r)} = \hat{\tilde{v}}_{l}^{\prime(r)} - \tilde{v}_{l}^{\prime(r-1)} \tag{99}$$

$$\partial \tilde{g}_{kl}^{(r)} = \hat{\tilde{g}}_{kl}^{(r)} - \tilde{g}_{kl}^{(r-1)}, \quad \partial \tilde{h}_{kl}^{(r)} = \hat{\tilde{h}}_{kl}^{(r)} - \tilde{h}_{kl}^{(r-1)} \tag{100}$$

$$\partial \tilde{g}_{l}^{\prime(r)} = \hat{\tilde{g}}_{l}^{\prime(r)} - \tilde{g}_{l}^{\prime(r-1)}, \quad \partial \tilde{h}_{l}^{\prime(r)} = \hat{\tilde{h}}_{l}^{\prime(r)} - \tilde{h}_{l}^{\prime(r-1)} \tag{101}$$

Thus, the hyperparameters for updating the global variational distributions as in (44)∼(51) are obtained by substituting the natural gradients into (52)∼(58).

REFERENCES

[1] Z. Zivkovic and F. van der Heijden, "Recursive unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 651–656, May 2004.

[2] T. S. F. Haines and T. Xiang, "Background subtraction with Dirichletprocess mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 670–683, Apr. 2014.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[4] G. McLachlan and D. Peel, *Finite Mixture Models*. New York, NY, USA: Wiley, 2000.

[5] N. Bouguila and D. Ziou, "High-dimensional unsupervised selection and estimation of a finite generalized Dirichlet mixture model based on minimum message length," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1716–1731, Oct. 2007.

[6] M. Bertoletti, N. Friel, and R. Rastelli, "Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion," *METRON*, vol. 73, no. 2, pp. 177–199, Aug. 2015.

[7] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," in *Bayesian Nonparametrics: Principles Practice* N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[8] W. Fan, N. Bouguila, and X. Liu, "A hierarchical Dirichlet process mixture of GID Distributions with feature selection for spatio-temporal video modeling and segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2771–2775.

[9] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Ann. Probab.*, vol. 25, no. 2, pp. 855–900, 1997.

[10] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," in *Proc. 21st Int. Conf. Comput. Linguistics*, Jul. 2006, pp. 985–992.

[11] W. Fan, H. Sallay, and N. Bouguila, "Online learning of hierarchical pitman–yor process mixture of generalized Dirichlet distributions with feature selection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 9, pp. 2048–2061, Sep. 2017.

[12] T. Bdiri, N. Bouguila, and D. Ziou, "Variational Bayesian inference for infinite generalized inverted Dirichlet mixtures with feature selection and its application to clustering," *Appl. Intell.*, vol. 44, no. 3, pp. 507–525, Apr. 2016.

[13] T. Bdiri and N. Bouguila, "Positive vectors clustering using inverted Dirichlet finite mixture models," *Expert Syst. Appl.*, vol. 39, no. 2, pp. 1869–1882, Feb. 2012.

[14] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[15] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[16] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, pp. 1303–1347, 2013.

[17] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statist. Sinica*, vol. 4, pp. 639–650, Mar. 1994.

[18] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," *J. Mach. Learn. Res.*, vol. 15, pp. 752–760, Jun. 2011.

[19] T. Bdiri and N. Bouguila, "Bayesian learning of inverted Dirichlet mixtures for SVM kernels generation," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1443–1458, 2013.

[20] G. G. Tiao and I. Cuttman, "The inverted Dirichlet distribution with applications," *J. Amer. Stat. Assoc.*, vol. 60, no. 311, pp. 793–805, 1965.

[21] M. A. Mashrgy, T. Bdiri, and N. Bouguila, "Robust simultaneous positive data clustering and unsupervised feature selection using generalized inverted Dirichlet mixture models," *Knowl.-Based Syst.*, vol. 59, pp. 182–195, Mar. 2014.

[22] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2, pp. 183–233, Nov. 1999.

[23] H. Attias, "A variational Bayes framework for graphical models," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 1999, pp. 209–215.

[24] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[25] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Anal.*, vol. 1, pp. 121–144, 2006.

[26] L. Maddalena and A. Petrosino, "A self-organizing approach to background subtraction for visual surveillance applications," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1168–1177, Jul. 2008.

[27] C. Stiller, "Object-based estimation of dense motion fields," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 234–250, Feb. 1997.

[28] T. Meier and K. N. Ngan, "Automatic segmentation of moving objects for video object plane generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, pp. 525–538, Sep. 1998.

[29] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1999, pp. 246–252.

[30] D. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 827–832, May 2005.

[31] W. Fan and N. Bouguila, "Video background subtraction using Online infinite Dirichlet mixture models," in *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2013, pp. 1–5.

[32] O. Barnich and M. Van Droogenbroeck, “ViBe: A universal background subtraction algorithm for video sequences,” *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.

[33] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, “Detecting moving objects, ghosts, and shadows in video streams,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[34] F. Kristensen, P. Nilsson, and V. Öwall, “Background segmentation beyond RGB,” in *Proc. ACCV*, P. J. Narayanan, S. K. Nayar, and H.-Y. Shum, Eds. 2006, pp. 602–612.

[35] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, “Wallflower: Principles and practice of background maintenance,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 1999, pp. 255–261.

[36] J. Huang, X. Huang, and D. Metaxas, “Learning with dynamic group sparsity,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 64–71.

WENTAO FAN received the M.Sc. and Ph.D. degrees in electrical and computer engineering from Concordia University, Montreal, QC, Canada, in 2009 and 2014, respectively. He is currently an Associate Professor with the Department of Computer Science and Technology, Huaqiao University, Xiamen, China. His research interests include machine learning, computer vision, and pattern recognition.

NIZAR BOUGUILA received the degree in engineering from the University of Tunis, in 2000, and the M.Sc. and Ph.D. degrees from Sherbrooke University, in 2002 and 2006, respectively, all in computer science. He is currently a Professor with the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada. His research interests include image processing, machine learning, 3D graphics, computer vision, and pattern recognition.