Unsupervised Hybrid Feature Extraction Selection for High-Dimensional Non-Gaussian Data Clustering with Variational Inference

Wentao Fan, Nizar Bouguila, and Djemel Ziou

Abstract

Clustering has been a subject of extensive research in data mining, pattern recognition and other areas for several decades. The main goal is to assign samples, which are typically non-Gaussian and expressed as points in high-dimensional feature spaces, to one of a number of clusters. It is well-known that in such high-dimensional settings, the existence of irrelevant features generally compromises modeling capabilities. In this paper, we propose a variational inference framework for unsupervised non-Gaussian feature selection, in the context of finite generalized Dirichlet (GD) mixture-based clustering. Under the proposed principled variational framework, we simultaneously estimate, in a closed-form, all the involved parameters and determine the complexity (i.e. both model an feature selection) of the GD mixture. Extensive simulations using synthetic data along with an analysis of real-world data and human action videos demonstrate that our variational approach achieves better results than comparable techniques.

Index Terms

Mixture models, unsupervised learning, generalized Dirichlet, model selection, feature selection, Bayesian estimation, variational inference, human action videos.

Wentao Fan is with the Department of Electrical and Computer Engineering, Concordia University, QC, Canada H3G 1T7. (email: wenta_fa@encs.concordia.ca)

Nizar Bouguila is with the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC H3G 1T7, Canada (email: bouguila@ciise.concordia.ca).

Djemel Ziou is with the Département d'Informatique, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada (email: djemel.ziou@usherbrooke.ca).

Digital Object Indentifier 10.1109/TKDE.2012.101

I. Introduction

An important traditional step in pattern recognition and data mining is to select the relevant features, with good discriminatory power, for a given application [1–7]. The surge of research interests in the areas of machine learning, data mining, computer vision, statistics and related fields has produced a wide variety of feature selection approaches especially in supervised settings. For an excellent review and in-depth discussions of the research in feature selection the reader is referred to [8–10] and references therein. In recent years there has been considerable interest in formulating the feature selection problem in unsupervised settings using mixture models learned using different optimization algorithms [9–16]. The primary objective is the identification and the reduction of the influence of extraneous (or irrelevant) features which do not contribute information about the true clusters structure. The main assumption in many of these approaches is that the features follow a multivariate normal distribution with diagonal variance-covariance matrix (see, for instance, [9,11,13,17]). This assumption is rarely met, is unrealistic in many cases and is generally violated by real life applications [15, 18]. Indeed, in many applications the per-class distributions are not Gaussian as shown in [10] where a mixture-based approach, relying on GD distribution and benefiting from its interesting mathematical properties and flexibility [19, 20], has been proposed.

The unsupervised feature selection model in [10] has been trained using a minimum message length (MML) [21] objective function with the expectation-maximization (EM) [22] algorithm. Despite the fact that the EM algorithm is the procedure of choice for parameter estimation in the case of incomplete data problems where part of the data is hidden, several studies have shown theoretically and experimentally that the EM algorithm, in deterministic settings (e.g. maximum likelihood estimation), converges either to a local maximum or to a saddle point solution and depends on an appropriate initialization (see, for instance, [22–24]) which may compromise the modeling capabilities. Recently, learning research has been directed towards Bayesian approaches which allow the formal treatment of uncertainty in modeling through the incorporation of prior knowledge about the model's parameters and then the combination of these prior beliefs with the observed data which results in posterior distributions [25]. The calculation and updating of these posteriors is generally untractable and involves high-dimensional integrations. Markov Chain Monte Carlo (MCMC) techniques are the methods of choice in this case and allow to approximate the Bayesian inference, but its computational cost is known to be prohibitive [25, 26]. A deterministic approximation alternative, of posterior

distributions, is now possible thanks to variational methods which at the same time prevent overfitting and allow model selection [27,28]. The main idea of variational Bayes learning is to find an accurate and tractable approximation to the true model's posterior that minimizes the divergence [27–30].

The aim of this paper is to extend our feature selection approach previously proposed in [10] by reformulating it within a variational framework. We are mainly motivated by the good results obtained recently using variational learning techniques in machine learning applications in general [31–34] and for the unsupervised feature selection problem in particular [13, 17]. The rest of the article is organized as follows. Section 2 presents the details of our unsupervised feature selection model and describes it as a probabilistic Bayesian model. In Section 3, we describe our variational approximation procedure for the proposed model learning. Section 4 presents results on synthetic data, real data and a challenging application namely human action videos categorization. Section 5 closes with conclusions, discussions and future directions.

II. Bayesian GD Mixture Model with Feature Selection

We start by briefly reviewing our unsupervised feature selection model previously proposed in [10]. Then, we propose a Bayesian version of this model. Although this paper is self-contained, the reader is urged to refer to [10].

A. Model Specification

Consider a set of N vectors $\mathcal{Y} = \{\vec{Y}_1, \ldots, \vec{Y}_N\}$, where each vector $\vec{Y}_i = (Y_{i1}, \ldots, Y_{iD})$ is represented in a D-dimensional space and assumed to be generated from a finite GD mixture model with M components [20]:

$$p(\vec{Y}_i|\boldsymbol{\pi}, \vec{\boldsymbol{\alpha}}, \vec{\boldsymbol{\beta}}) = \sum_{j=1}^{M} \pi_j \mathrm{GD}(\vec{Y}_i|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$$

where $\mathrm{GD}(\vec{Y}_i|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$ is a GD distribution with parameters $(\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jD})$, $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jD})$, $\vec{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_M)$, $\vec{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$, and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ is the vector of mixing coefficients which are positive and sum to one. Each observed vector $\vec{Y}_i$ is assigned to all the components with posterior probabilities $p(j|\vec{Y}_i) \propto \pi_j \mathrm{GD}(\vec{Y}_i|\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$, which are also known as *responsibilities* [35]. Based on the mathematical properties of the GD thoroughly discussed in [10, 20], it is possible to show that the responsibilities can be written as [10]:

$$p(j|\vec{Y}_i) \propto \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})$$

where $X_{i1} = Y_{i1}$ and $X_{il} = Y_{il}/(1 - \sum_{k=1}^{l-1} Y_{ik})$ for $l > 1$, and $\mathrm{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})$ is a Beta distribution defined as

$$\mathrm{Beta}(X_{il}|\alpha_{jl}, \beta_{jl}) = \frac{\Gamma(\alpha_{jl} + \beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})} X_{il}^{\alpha_{jl}-1}(1 - X_{il})^{\beta_{jl}-1}$$

Therefore, the clustering structure underlying data set $\mathcal{Y}$ can be represented by a new data set $\mathcal{X} = \{\vec{X}_1, \ldots, \vec{X}_N\}$ governed by the mixture model

$$p(\vec{X}_i|\boldsymbol{\pi}, \vec{\boldsymbol{\alpha}}, \vec{\boldsymbol{\beta}}) = \sum_{j=1}^{M} \pi_j \prod_{l=1}^{D} \mathrm{Beta}(X_{il}|\alpha_{jl}, \beta_{jl})$$

This is actually an important property of the GD mixture, since the independence between the features, in the case of the new data set $\mathcal{X}$, becomes a fact and not an assumption as considered in previous unsupervised feature selection Gaussian mixture-based approaches [9, 11, 13, 17]. The approach that we shall present can be viewed then as a hybrid feature extraction selection technique since we identify relevant features in a transformed space and not in the original one [10]. For each vector $\vec{X}_i$, we assign a latent variable $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iM})$, such that $z_{ij} \in \{0, 1\}$, $\sum_{j=1}^{M} z_{ij} = 1$ and $z_{ij} = 1$ if $\vec{X}_i$ belongs to class (or component) j and 0, otherwise. The conditional distribution of the latent variables $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ given $\boldsymbol{\pi}$ is defined as

$$p(\mathcal{Z}|\boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \pi_j^{z_{ij}} \tag{1}$$

Then, the conditional distribution of data set $\mathcal{X}$ given the class labels $\mathcal{Z}$ can be written as

$$p(\mathcal{X}|\mathcal{Z}, \vec{\boldsymbol{\alpha}}, \vec{\boldsymbol{\beta}}) = \prod_{i=1}^{N} \prod_{j=1}^{M} \Big(\prod_{l=1}^{D} \mathrm{Beta}(X_{il}|\alpha_{jl}, \beta_{jl}) \Big)^{z_{ij}}$$

It is noteworthy that the previous model assumes actually that all the features X_{il} are equally important for the clustering task which is not realistic in general since some of the features might be irrelevant and then mask [1] completely the cluster structure and its recovery [36, 38, 39]. The automatic selection of relevant features in the context of unsupervised learning is challenging and is far from trivial because inference has to be made on both the selected features and the clustering structure [7, 9–13, 16, 40]. [9] is an early influential paper advocating the use of finite mixture models for unsupervised feature selection. The main idea is to suppose that a given feature X_{il} is generated from a mixture of two univariate distributions. The first one is assumed to generate relevant features and is different for each cluster and the

[1] Indeed, some authors have referred to irrelevant features as "masking variables" (see, for instance, [36, 37]).

second one is common to all clusters (i.e. independent from class labels) and assumed to generate irrelevant features [2]. In [10] we have extended this work for non-Gaussian features by approximating the feature distribution as following:

$$p(X_{il}|\phi_{il},\alpha_{jl},\beta_{jl},w_{ilk},\lambda_{lk},\tau_{lk}) \simeq \big(\mathrm{Beta}(X_{il}|\alpha_{jl},\beta_{jl})\big)^{\phi_{il}}\big(\prod_{k=1}^{K}\mathrm{Beta}(X_{il}|\lambda_{lk},\tau_{lk})^{w_{ilk}}\big)^{1-\phi_{il}} \tag{2}$$

where ϕ_{il} is a binary latent variable, such that $\phi_{il}=1$ if feature l is relevant (i.e. supposed to follow a Beta distribution, $\mathrm{Beta}(X_{il}|\alpha_{jl},\beta_{jl})$, that depends on the class labels), and $\phi_{il}=0$ if feature l is irrelevant and then supposed to follow a mixture of K Beta distributions independent from the class labels [3]:

$$p(X_{il}|\boldsymbol{\eta}_l,\boldsymbol{\lambda}_l,\boldsymbol{\tau}_l)=\sum_{k=1}^{K}\eta_{lk}\mathrm{Beta}(X_{il}|\lambda_{lk},\tau_{lk}) \tag{3}$$

where $\boldsymbol{\lambda}_l=(\lambda_{l1},\ldots,\lambda_{lK})$, $\boldsymbol{\tau}_l=(\tau_{l1},\ldots,\tau_{lK})$, $\boldsymbol{\eta}_l=(\eta_{l1},\ldots,\eta_{lK})$ such that η_{lk} represents the prior probability that X_{il} comes from the kth component of the Beta mixture representing irrelevant feature, and $\sum_{k=1}^{K}\eta_{lk}=1$. w_{ilk} in Eq. 2 is a binary variable such that $\sum_{k=1}^{K}w_{ilk}=1$ and $w_{ilk}=1$ indicates that X_{il} comes from the kth component of the mixture in Eq. 3. Assuming $\mathcal{W}=\{\vec{\mathbf{w}}_1,\ldots,\vec{\mathbf{w}}_N\}$ with $\vec{\mathbf{w}}_i=(\mathbf{w}_{i1},\ldots,\mathbf{w}_{iD})$ and $\mathbf{w}_{il}=(w_{il1},\ldots,w_{ilK})$, the distribution of the latent variables $\mathcal{W}$ given the mixing probabilities $\vec{\boldsymbol{\eta}}=(\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_D)$ is defined as

$$p(\mathcal{W}|\vec{\boldsymbol{\eta}})=\prod_{i=1}^{N}\prod_{l=1}^{D}\prod_{k=1}^{K}\eta_{lk}^{w_{ilk}} \tag{4}$$

And since each ϕ_{il} is a Bernoulli variable, the distribution of the hidden variables $\vec{\boldsymbol{\phi}}=\{\boldsymbol{\phi_1},\ldots,\boldsymbol{\phi_N}\}$, with elements $\boldsymbol{\phi_i}=\{\phi_{i1},\ldots,\phi_{iD}\}$, is defined as

$$p(\vec{\boldsymbol{\phi}}|\boldsymbol{\epsilon})=\prod_{i=1}^{N}\prod_{l=1}^{D}\epsilon_l^{\phi_{il}}(1-\epsilon_l)^{1-\phi_{il}} \tag{5}$$

where $\boldsymbol{\epsilon}=\{\epsilon_l\}$ represents the features saliencies (i.e. the probabilities that the features are relevant) such that $p(\phi_{il}=1)=\epsilon_l$ and $p(\phi_{il}=0)=1-\epsilon_l$. Having all the model's parameters at hand, the likelihood of the observed data can be written as

$$p(\mathcal{X}|\mathcal{Z},\mathcal{W},\vec{\boldsymbol{\phi}},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\lambda}},\vec{\boldsymbol{\tau}})=\prod_{i=1}^{N}\prod_{j=1}^{M}\left[\prod_{l=1}^{D}\mathrm{Beta}(X_{il}|\alpha_{jl},\beta_{jl})^{\phi_{il}}\left(\prod_{k=1}^{K}\mathrm{Beta}(X_{il}|\lambda_{lk},\tau_{lk})^{w_{ilk}}\right)^{1-\phi_{il}}\right]^{z_{ij}} \tag{6}$$

where $\vec{\boldsymbol{\lambda}}=\{\boldsymbol{\lambda}_1,\ldots,\boldsymbol{\lambda}_D\}$ and $\vec{\boldsymbol{\tau}}=\{\boldsymbol{\tau}_1,\ldots,\boldsymbol{\tau}_D\}$.

[2]Several other quantitative formalisms for relevance in the case of feature selection have been proposed in the past (see, for instance, [41]).

[3]An important distinction between the model in [9] and the one proposed in [10] is that the nonsalient feature is modeled as a mixture of distributions rather than a usual single distribution. Our specific choice is justified by the fact that a Beta mixture allows the accurate approximation of any univariate distribution (e.g. uniform, Gaussian, Gamma, etc.) [42].

B. Bayesian Framework

The EM algorithm can be applied for inferencing the model presented in the previous section [10]. However, it requires the integration of an entropy measure [19] or an information criterion such as MML [20] as done in [10] for the determination of the optimal number of components. In this paper, we adopt a Bayesian variational inference approach which allows simultaneously the estimation of all the involved parameters and model selection (i.e. both feature selection and determination of the optimal number of clusters). Indeed, variational learning has been used recently as an approximation of Bayesian learning and as an alternative to both fully Bayesian MCMC techniques and fully deterministic likelihood-based approaches which can be justified by its computational tractability and good generalization performance [27–30].

In order to perform variational Bayes, we need to introduce conjugate priors over parameters $\vec{\boldsymbol{\alpha}}, \vec{\boldsymbol{\beta}}, \vec{\boldsymbol{\lambda}}$ and $\vec{\boldsymbol{\tau}}$. It is noteworthy that $\boldsymbol{\epsilon}$, $\boldsymbol{\pi}$ and $\vec{\boldsymbol{\eta}}$ will be considered as parameters and not as random variables within our framework, thus priors shall not be imposed on them as we will explain further in next section. The conjugate priors, that can be developed using the fact that the Beta distribution belongs to the exponential family (see, for instance, [43, 44]), are analytically intractable and cannot be used within a variational framework as shown in [45]. Thus, we use Gamma priors as suggested in [45], for the Beta distribution, by assuming that the different model's parameters are independent:

$$p(\vec{\boldsymbol{\alpha}}|\mathbf{u},\mathbf{v}) = \prod_{j=1}^{M}\prod_{l=1}^{D}\frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})}\alpha_{jl}^{u_{jl}-1}e^{-v_{jl}\alpha_{jl}}\ , \qquad p(\vec{\boldsymbol{\beta}}|\mathbf{p},\mathbf{q}) = \prod_{j=1}^{M}\prod_{l=1}^{D}\frac{q_{jl}^{p_{jl}}}{\Gamma(p_{jl})}\beta_{jl}^{p_{jl}-1}e^{-q_{jl}\beta_{jl}} \tag{7}$$

$$p(\vec{\boldsymbol{\lambda}}|\mathbf{g},\mathbf{h}) = \prod_{l=1}^{D}\prod_{k=1}^{K}\frac{h_{lk}^{g_{lk}}}{\Gamma(g_{lk})}\lambda_{lk}^{g_{lk}-1}e^{-h_{lk}\lambda_{lk}}\ , \qquad p(\vec{\boldsymbol{\tau}}|\mathbf{s},\mathbf{t}) = \prod_{l=1}^{D}\prod_{k=1}^{K}\frac{t_{lk}^{s_{lk}}}{\Gamma(s_{lk})}\tau_{lk}^{s_{lk}-1}e^{-t_{lk}\tau_{lk}} \tag{8}$$

where all the hyperparameters $\mathbf{u}=\{u_{jl}\}, \mathbf{v}=\{v_{jl}\}, \mathbf{p}=\{p_{jl}\}, \mathbf{q}=\{q_{jl}\}, \mathbf{g}=\{g_{lk}\}, \mathbf{h}=\{h_{lk}\}, \mathbf{s}=\{s_{lk}\}$ and $\mathbf{t}=\{t_{lk}\}$ of the above conjugate priors are positive. Then, by using Eqs.1, 4, 5, 6, 7, and 8, the joint distribution of all the random variables, conditioned on parameters, is given by

$$\begin{aligned}
& p(\mathcal{X},\mathcal{Z},\mathcal{W},\vec{\boldsymbol{\phi}},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\lambda}},\vec{\boldsymbol{\tau}}|\boldsymbol{\pi},\vec{\boldsymbol{\eta}},\boldsymbol{\epsilon}) = p(\mathcal{X}|\mathcal{Z},\mathcal{W},\vec{\boldsymbol{\phi}},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\lambda}},\vec{\boldsymbol{\tau}})p(\mathcal{Z}|\boldsymbol{\pi})p(\mathcal{W}|\vec{\boldsymbol{\eta}})p(\vec{\boldsymbol{\phi}}|\boldsymbol{\epsilon})p(\vec{\boldsymbol{\alpha}})p(\vec{\boldsymbol{\beta}})p(\vec{\boldsymbol{\lambda}})p(\vec{\boldsymbol{\tau}}) \\
= & \prod_{i=1}^{N}\prod_{j=1}^{M}\Big\{\pi_j\prod_{l=1}^{D}\Big[\frac{\Gamma(\alpha_{jl}+\beta_{jl})}{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}X_{il}^{\alpha_{jl}-1}(1-X_{il})^{\beta_{jl}-1}\Big]^{\phi_{il}}\Big[\prod_{k=1}^{K}\Big(\frac{\Gamma(\lambda_{lk}+\tau_{lk})}{\Gamma(\lambda_{lk})\Gamma(\tau_{lk})}X_{il}^{\lambda_{lk}-1}(1-X_{il})^{\tau_{lk}-1}\Big)^{w_{ilk}}\Big]^{1-\phi_{il}}\Big\}^{z_{ij}} \\
\times & \Big[\prod_{i=1}^{N}\prod_{l=1}^{D}\epsilon_l^{\phi_{il}}(1-\epsilon_l)^{1-\phi_{il}}\Big]\Big[\prod_{i=1}^{N}\prod_{l=1}^{D}\prod_{k=1}^{K}\eta_{kl}^{w_{ilk}}\Big]\Big[\prod_{j=1}^{M}\prod_{l=1}^{D}\frac{v_{jl}^{u_{jl}}}{\Gamma(u_{jl})}\alpha_{jl}^{u_{jl}-1}e^{-v_{jl}\alpha_{jl}}\frac{q_{jl}^{p_{jl}}}{\Gamma(p_{jl})}\beta_{jl}^{p_{jl}-1}e^{-q_{jl}\beta_{jl}}\Big] \\
\times & \prod_{l=1}^{D}\prod_{k=1}^{K}\frac{h_{lk}^{g_{lk}}}{\Gamma(g_{lk})}\lambda_{lk}^{g_{lk}-1}e^{-h_{lk}\lambda_{lk}}\frac{t_{lk}^{s_{lk}}}{\Gamma(s_{lk})}\tau_{lk}^{s_{lk}-1}e^{-t_{lk}\tau_{lk}}
\end{aligned} \tag{9}$$

A directed graphical representation of this model is illustrated in Fig. 1.

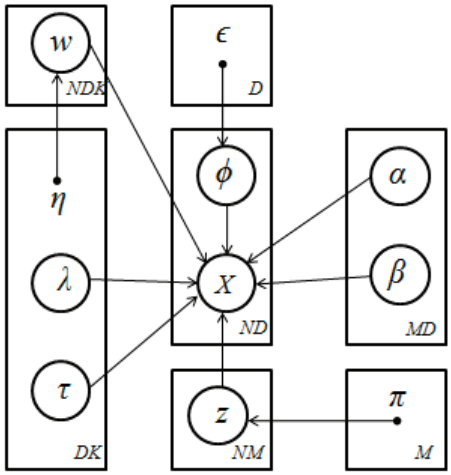


Fig. 1. Graphical model representation of our unsupervised feature selection model. Symbols in circles denote random variables; otherwise, they denote model parameters. Plates indicate repetition (with the number of repetitions in the lower right), and arcs describe conditional dependencies between variables.

III. Variational Learning of the Model

In this section, we describe a variational Bayes learning approach for our model by following the inference methodology proposed in [46]. The proposed variational framework is used to prevent overfitting and allows simultaneously the parameters estimation, the automatic determination of the number of clusters, and the saliencies of the features.

A. Variational Learning

Let us define $\mathbf{\Theta} = \{\mathcal{Z}, \vec{\phi}, \mathcal{W}, \vec{\alpha}, \vec{\beta}, \vec{\lambda}, \vec{\tau}\}$ as the set of random variables and denote $\mathbf{\Lambda} = \{\boldsymbol{\pi}, \vec{\eta}, \boldsymbol{\epsilon}\}$ as the set of parameters. Our goal is to estimate the parameters $\mathbf{\Lambda}$ by maximizing the marginal likelihood $p(\mathcal{X}|\mathbf{\Lambda})$:

$$p(\mathcal{X}|\mathbf{\Lambda}) = \sum_{\mathcal{Z},\vec{\phi},\mathcal{W}} \int_{\vec{\alpha}} \int_{\vec{\beta}} \int_{\vec{\lambda}} \int_{\vec{\tau}} p(\mathcal{X}, \mathbf{\Theta}|\mathbf{\Lambda})\, d\vec{\boldsymbol{\alpha}}\, d\vec{\boldsymbol{\beta}}\, d\vec{\boldsymbol{\lambda}}\, d\vec{\boldsymbol{\tau}}$$

Since this marginalization is intractable, variational approach is then adopted to find a tractable lower bound on $p(\mathcal{X}|\mathbf{\Lambda})$. By applying Jensen's inequality, the lower bound $\mathcal{L}$ of the logarithm of the marginal likelihood $\ln p(\mathcal{X}|\mathbf{\Lambda})$ can be found as [27–30]

$$\ln p(\mathcal{X}|\mathbf{\Lambda}) = \ln \int_{\Theta} p(\mathcal{X}, \mathbf{\Theta}|\mathbf{\Lambda}) d\mathbf{\Theta} = \ln \int_{\Theta} Q(\mathbf{\Theta}) \frac{p(\mathcal{X}, \mathbf{\Theta}|\mathbf{\Lambda})}{Q(\mathbf{\Theta})} d\mathbf{\Theta} \geq \int_{\Theta} Q(\mathbf{\Theta}) \ln \frac{p(\mathcal{X}, \mathbf{\Theta}|\mathbf{\Lambda})}{Q(\mathbf{\Theta})} d\mathbf{\Theta} = \mathcal{L}(Q) \qquad (10)$$

where $Q(\mathbf{\Theta})$ is an approximation to the true posterior distribution $p(\mathbf{\Theta}|\mathcal{X}, \boldsymbol{\pi})$. The lower bound $\mathcal{L}(Q)$ is maximized when $Q(\mathbf{\Theta}) = p(\mathbf{\Theta}|\mathcal{X}, \mathbf{\Lambda})$. However, in practice the true posterior distribution is computationally intractable and can not be directly used for variational inference. Thus, a restricted family of distributions $Q(\mathbf{\Theta})$ needs to be considered. Here, we restrict the form of $Q(\mathbf{\Theta})$ by adopting a factorization assumption that will allow the marginalization to be carried out efficiently as we shall see later. This approximation approach, which has been developed from statistical mechanics [47], is known as *mean field theory* and has been used efficiently by several researchers in the past [48, 49]. With the factorization assumption, the posterior distribution $Q(\mathbf{\Theta})$ can be factorized into disjoint tractable distributions such that $Q(\mathbf{\Theta}) = \prod_a Q_a(\mathbf{\Theta}_a)$. Note that this is the only assumption about the distribution, and no restriction is placed on the functional forms of the individual factors $Q_a(\mathbf{\Theta}_a)$. In order to maximize the lower bound $\mathcal{L}(Q)$, we need to make a variational optimization of $\mathcal{L}(Q)$ with respect to each of the distributions $Q_a(\mathbf{\Theta}_a)$ in turn. For a specific factor $Q_s(\mathbf{\Theta}_s)$ in a standard variational inference approach, the general expression for its optimal solution can be found by [30, 46]

$$Q_s(\mathbf{\Theta}_s) = \frac{\exp\big\langle \ln p(\mathcal{X}, \mathbf{\Theta})\big\rangle_{a\neq s}}{\int \exp\big\langle \ln p(\mathcal{X}, \mathbf{\Theta})\big\rangle_{a\neq s} d\mathbf{\Theta}} \tag{11}$$

where $\langle\cdot\rangle_{a\neq s}$ denotes an expectation with respect to all the distributions $Q_a(\mathbf{\Theta}_a)$ except for $a = s$. Since the expression for the optimal solution $Q_s(\mathbf{\Theta}_s)$ depends on calculating the expectations with respect to the other factors $Q_a(\mathbf{\Theta}_a)$ for $a \neq s$, we need to cycle through all the factors for finding the maximum of the lower bound. In general, in order to perform the variational inference, all the factors $Q_a(\mathbf{\Theta}_a)$ need to be suitably initialized first, then each factor is updated in turn with a revised value obtained by Eq. 11 using the current values for all of the other factors. Convergence is guaranteed since bound is convex with respect to each of the factors $Q_a(\mathbf{\Theta}_a)$ [50]. We apply the variational approach to our model by assuming that $Q(\mathbf{\Theta})$ can be factorized as following $Q(\mathbf{\Theta}) = Q(\mathcal{Z})Q(\vec{\boldsymbol{\phi}})Q(\mathcal{W})Q(\vec{\boldsymbol{\alpha}})Q(\vec{\boldsymbol{\beta}})Q(\vec{\boldsymbol{\lambda}})Q(\vec{\boldsymbol{\tau}})$. By applying Eq. 11 to each factor, we obtain the optimal solutions for the factors of the variational posterior :

$$Q(\mathcal{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{M} r_{ij}^{z_{ij}}\ , \qquad Q(\vec{\boldsymbol{\phi}}) = \prod_{i=1}^{N}\prod_{l=1}^{D} f_{il}^{\phi_{il}}(1-f_{il})^{(1-\phi_{il})}\ , \qquad Q(\mathcal{W}) = \prod_{i=1}^{N}\prod_{l=1}^{D}\prod_{k=1}^{K} m_{ilk}^{w_{ilk}} \tag{12}$$

$$Q(\vec{\boldsymbol{\alpha}}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\alpha_{jl}|u_{jl}^{*}, v_{jl}^{*})\ , \qquad Q(\vec{\boldsymbol{\beta}}) = \prod_{j=1}^{M}\prod_{l=1}^{D} \mathcal{G}(\beta_{jl}|p_{jl}^{*}, q_{jl}^{*}) \tag{13}$$

$$Q(\vec{\boldsymbol{\lambda}}) = \prod_{l=1}^{D}\prod_{k=1}^{K} \mathcal{G}(\lambda_{lk}|g_{lk}^{*}, h_{lk}^{*})\ , \qquad Q(\vec{\boldsymbol{\tau}}) = \prod_{l=1}^{D}\prod_{k=1}^{K} \mathcal{G}(\tau_{lk}|s_{lk}^{*}, t_{lk}^{*}) \tag{14}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

where $\mathcal{G}$ represents the Gamma distribution and where we define

$$r_{ij} = \frac{\rho_{ij}}{\sum_{d=1}^{M} \rho_{id}}$$

$$\rho_{ij} = \exp\Big\{ \ln \pi_j + \sum_{l=1}^{D} \big\langle \phi_{il} \big\rangle [\widetilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl} - 1) \ln X_{il} + (\bar{\beta}_{jl} - 1) \ln(1 - X_{il})] \Big\}$$

$$\begin{aligned}
\widetilde{\mathcal{R}} &= \ln \frac{\Gamma(\bar{\alpha} + \bar{\beta})}{\Gamma(\bar{\alpha})\Gamma(\bar{\beta})} + \bar{\alpha}[\psi(\bar{\alpha} + \bar{\beta}) - \psi(\bar{\alpha})](\big\langle \ln \alpha \big\rangle - \ln \bar{\alpha}) + \bar{\beta}[\psi(\bar{\alpha} + \bar{\beta}) - \psi(\bar{\beta})](\big\langle \ln \beta \big\rangle - \ln \bar{\beta}) \\
&+ 0.5\bar{\alpha}^2[\psi'(\bar{\alpha} + \bar{\beta}) - \psi'(\bar{\alpha})]\big\langle (\ln \alpha - \ln \bar{\alpha})^2 \big\rangle + 0.5\bar{\beta}^2[\psi'(\bar{\alpha} + \bar{\beta}) - \psi'(\bar{\beta})]\big\langle (\ln \beta - \ln \bar{\beta})^2 \big\rangle \\
&+ \bar{\alpha}\bar{\beta}\psi'(\bar{\alpha} + \bar{\beta})(\big\langle \ln \alpha \big\rangle - \ln \bar{\alpha})(\big\langle \ln \beta \big\rangle - \ln \bar{\beta})
\end{aligned}$$

$$f_{il} = \frac{\delta_{il}^{(\phi_{il})}}{\delta_{il}^{(\phi_{il})} + \delta_{il}^{(1-\phi_{il})}}$$

$$\delta_{il}^{(\phi_{il})} = \exp\Big\{ \sum_{j=1}^{M} \big\langle z_{ij} \big\rangle \big[\widetilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl} - 1) \ln X_{il} + (\bar{\beta}_{jl} - 1) \ln(1 - X_{il}) \big] + \ln \epsilon_l \Big\}$$

$$\delta_{il}^{(1-\phi_{il})} = \exp\Big\{ \sum_{k=1}^{K} \big\langle w_{ilk} \big\rangle [\widetilde{\mathcal{F}}_{lk} + (\bar{\lambda}_{lk} - 1) \ln X_{il} + (\bar{\tau}_{lk} - 1) \ln(1 - X_{il})] + \ln(1 - \epsilon_l) \Big\}$$

$$\begin{aligned}
\widetilde{\mathcal{F}} &= \ln \frac{\Gamma(\bar{\lambda} + \bar{\tau})}{\Gamma(\bar{\lambda})\Gamma(\bar{\tau})} + \bar{\lambda}[\psi(\bar{\lambda} + \bar{\tau}) - \psi(\bar{\lambda})](\big\langle \ln \lambda \big\rangle - \ln \bar{\lambda}) + \bar{\tau}[\psi(\bar{\lambda} + \bar{\tau}) - \psi(\bar{\tau})](\big\langle \ln \tau \big\rangle - \ln \bar{\tau}) \\
&+ 0.5\bar{\lambda}^2[\psi'(\bar{\lambda} + \bar{\tau}) - \psi'(\bar{\lambda})]\big\langle (\ln \lambda - \ln \bar{\lambda})^2 \big\rangle + 0.5\bar{\tau}^2[\psi'(\bar{\lambda} + \bar{\tau}) - \psi'(\bar{\tau})]\big\langle (\ln \tau - \ln \bar{\tau})^2 \big\rangle \\
&+ \bar{\lambda}\bar{\tau}\psi'(\bar{\lambda} + \bar{\tau})(\big\langle \ln \lambda \big\rangle - \ln \bar{\lambda})(\big\langle \ln \tau \big\rangle - \ln \bar{\tau})
\end{aligned}$$

$$m_{ilk} = \frac{\varphi_{ilk}}{\sum_{d=1}^{K} \varphi_{ild}}$$

$$\varphi_{ilk} = \exp\Big\{ \big\langle 1 - \phi_{il} \big\rangle \big[\widetilde{\mathcal{F}}_{kl} + (\bar{\lambda}_{lk} - 1) \ln X_{il} + (\bar{\tau}_{lk} - 1) \ln(1 - X_{il}) \big] + \ln \eta_{lk} \Big\}$$

$$u_{jl}^{*} = u_{jl} + \sum_{i=1}^{N} \big\langle z_{ij} \big\rangle \big\langle \phi_{il} \big\rangle \bar{\alpha}_{jl} \big[\psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \psi(\bar{\alpha}_{jl}) + \bar{\beta}_{jl}\psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl})(\langle \ln \beta_{jl} \rangle - \ln \bar{\beta}_{jl}) \big]$$

$$v_{jl}^{*} = v_{jl} - \sum_{i=1}^{N} \big\langle z_{ij} \big\rangle \big\langle \phi_{il} \big\rangle \ln X_{il}$$

$$p_{jl}^{*} = p_{jl} + \sum_{i=1}^{N} \big\langle z_{ij} \big\rangle \big\langle \phi_{il} \big\rangle \bar{\beta}_{jl} \big[\psi(\bar{\alpha}_{jl} + \bar{\beta}_{jl}) - \psi(\bar{\beta}_{jl}) + \bar{\alpha}_{jl}\psi'(\bar{\alpha}_{jl} + \bar{\beta}_{jl})(\big\langle \ln \alpha_{jl} \big\rangle - \ln \bar{\alpha}_{jl}) \big]$$

$$q_{jl}^{*} = q_{jl} - \sum_{i=1}^{N} \big\langle z_{ij} \big\rangle \big\langle \phi_{il} \big\rangle \ln(1 - X_{il})$$

$$g^*_{lk} = g_{lk} + \sum_{i=1}^{N} \big\langle 1-\phi_{il} \big\rangle \big\langle w_{ilk} \big\rangle \bar{\lambda}_{lk} \big[\psi(\bar{\lambda}_{lk}+\bar{\tau}_{lk}) - \psi(\bar{\lambda}_{lk}) + \bar{\tau}_{lk}\psi'(\bar{\lambda}_{lk}+\bar{\tau}_{lk}) (\big\langle \ln \tau_{lk} \big\rangle - \ln \bar{\tau}_{lk}) \big]$$

$$h^*_{lk} = h_{lk} - \sum_{i=1}^{N} \big\langle 1-\phi_{il} \big\rangle \big\langle w_{ilk} \big\rangle \ln X_{il}$$

$$s^*_{lk} = s_{lk} + \sum_{i=1}^{N} \big\langle 1-\phi_{il} \big\rangle \big\langle w_{ilk} \big\rangle \bar{\tau}_{lk} \big[\psi(\bar{\lambda}_{lk}+\bar{\tau}_{lk}) - \psi(\bar{\tau}_{lk}) + \bar{\lambda}_{lk}\psi'(\bar{\lambda}_{lk}+\bar{\tau}_{lk}) (\big\langle \ln \lambda_{lk} \big\rangle - \ln \bar{\lambda}_{lk}) \big]$$

$$t^*_{lk} = t_{lk} - \sum_{i=1}^{N} \big\langle 1-\phi_{il} \big\rangle \big\langle w_{ilk} \big\rangle \ln(1-X_{il})$$

$$\bar{\alpha}_{jl} = \big\langle \alpha_{jl} \big\rangle = \frac{u^*_{jl}}{v^*_{jl}}\,, \quad \bar{\beta}_{jl} = \big\langle \beta_{jl} \big\rangle = \frac{p^*_{jl}}{q^*_{jl}}\,, \quad \bar{\lambda}_{lk} = \big\langle \lambda_{lk} \big\rangle = \frac{g^*_{lk}}{h^*_{lk}}\,, \quad \bar{\tau}_{lk} = \big\langle \tau_{lk} \big\rangle = \frac{s^*_{lk}}{t^*_{lk}} \tag{15}$$

where $\big\langle \cdot \big\rangle$ represents an expected value, the $\psi(\cdot)$ is the digamma function and defined as: $\psi(a) = d\ln\Gamma(a)/da$. Notice that, $\widetilde{\mathcal{R}}$ and $\widetilde{\mathcal{F}}$ are the lower bound approximations of $\mathcal{R} = \big\langle \ln \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \big\rangle$ and $\mathcal{F} = \big\langle \ln \frac{\Gamma(\lambda+\tau)}{\Gamma(\lambda)\Gamma(\tau)} \big\rangle$, respectively. Since these expectations are intractable, we use the second-order Taylor series expansion to find their lower bounds as proposed in [45]. The expected values in the above formulas are given by

$$\big\langle\, z_{ij} \big\rangle = r_{ij}, \quad \big\langle\, w_{ilk} \big\rangle = m_{ilk}, \quad \big\langle \phi_{il} \big\rangle = f_{il}, \quad \big\langle 1-\phi_{il} \big\rangle = 1-f_{il}$$

$$\big\langle \ln \alpha \big\rangle = \psi(u^*) - \ln v^*, \quad \big\langle \ln \beta \big\rangle = \psi(p^*) - \ln q^*, \quad \big\langle \ln \lambda \big\rangle = \psi(g^*) - \ln h^*, \quad \big\langle \ln \tau \big\rangle = \psi(s^*) - \ln t^*$$

$$\big\langle (\ln \alpha - \ln \bar{\alpha})^2 \big\rangle = [\psi(u^*) - \ln u^*]^2 + \psi'(u^*), \quad \big\langle (\ln \beta - \ln \bar{\beta})^2 \big\rangle = [\psi(p^*) - \ln p^*]^2 + \psi'(p^*)$$

$$\big\langle (\ln \lambda - \ln \bar{\lambda})^2 \big\rangle = [\psi(g^*) - \ln g^*]^2 + \psi'(g^*), \quad \big\langle (\ln \tau - \ln \bar{\tau})^2 \big\rangle = [\psi(s^*) - \ln s^*]^2 + \psi'(s^*)$$

B. Variational Lower Bound

In order to monitor the convergence and check the correctness of the proposed variational learning approach, we can evaluate its variational lower bound. After obtaining the functional forms for the variational factors, the lower bound (Eq. 10) of the model can be obtained as following

$$\begin{aligned}
\mathcal{L}(Q) &= \int Q(\boldsymbol{\Theta}) \ln \frac{p(\boldsymbol{\Theta}, \mathcal{X}|\boldsymbol{\Lambda})}{Q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} = \big\langle \ln p(\mathcal{X}, \boldsymbol{\Theta}|\boldsymbol{\Lambda}) \big\rangle - \big\langle \ln Q(\boldsymbol{\Theta}) \big\rangle \\
&= \big\langle \ln p(\mathcal{X}|\mathcal{Z}, \mathcal{W}, \vec{\boldsymbol{\phi}}, \vec{\boldsymbol{\alpha}}, \vec{\boldsymbol{\beta}}, \vec{\boldsymbol{\lambda}}, \vec{\boldsymbol{\tau}}) \big\rangle + \big\langle \ln p(\mathcal{Z}|\boldsymbol{\pi}) \big\rangle + \big\langle \ln p(\mathcal{W}|\vec{\boldsymbol{\eta}}) \big\rangle + \big\langle \ln p(\vec{\boldsymbol{\phi}}|\boldsymbol{\epsilon}) \big\rangle + \big\langle \ln p(\vec{\boldsymbol{\alpha}}) \big\rangle \\
&+ \big\langle \ln p(\vec{\boldsymbol{\beta}}) \big\rangle + \big\langle \ln p(\vec{\boldsymbol{\lambda}}) \big\rangle + \big\langle \ln p(\vec{\boldsymbol{\tau}}) \big\rangle - \big\langle \ln Q(\mathcal{Z}) \big\rangle - \big\langle \ln Q(\vec{\boldsymbol{\phi}}) \big\rangle - \big\langle \ln Q(\mathcal{W}) \big\rangle - \big\langle \ln Q(\vec{\boldsymbol{\alpha}}) \big\rangle \\
&- \big\langle \ln Q(\vec{\boldsymbol{\beta}}) \big\rangle - \big\langle \ln Q(\vec{\boldsymbol{\lambda}}) \big\rangle - \big\langle \ln Q(\vec{\boldsymbol{\tau}}) \big\rangle
\end{aligned} \tag{16}$$

Here, each expectation is evaluated with respect to all of the random variables in its argument. It is straightforward to obtain these expectations according to the results from previous section:

$$\begin{aligned}\big\langle \ln p(\mathcal{X}|\mathcal{Z},\mathcal{W},\vec{\boldsymbol{\phi}},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}},\vec{\boldsymbol{\lambda}},\vec{\boldsymbol{\tau}})\big\rangle \approx \sum_{i=1}^{N}\sum_{j=1}^{M} r_{ij}\Bigg\{ &\sum_{l=1}^{D} f_{il}\big[\widetilde{\mathcal{R}}_{jl} + (\bar{\alpha}_{jl}-1)\ln X_{il} + (\bar{\beta}_{jl}-1)\ln(1-X_{il})\big] \\ &+ \sum_{l=1}^{D}(1-f_{il})\sum_{k=1}^{K} m_{ilk}\big[\widetilde{\mathcal{F}}_{lk} + (\bar{\lambda}_{lk}-1)\ln X_{il} + (\bar{\tau}_{lk}-1)\ln(1-X_{il})\big]\Bigg\} \quad (17)\end{aligned}$$

$$\big\langle \ln p(\mathcal{Z}|\boldsymbol{\pi})\big\rangle = \sum_{i=1}^{N}\sum_{j=1}^{M} r_{ij}\ln\pi_j\ , \qquad \big\langle \ln p(\mathcal{W}|\vec{\boldsymbol{\eta}})\big\rangle = \sum_{i=1}^{N}\sum_{l=1}^{D}\sum_{k=1}^{K} m_{ilk}\ln\eta_{lk} \quad (18)$$

$$\big\langle \ln p(\vec{\boldsymbol{\phi}}|\boldsymbol{\epsilon})\big\rangle = \sum_{i=1}^{N}\sum_{l=1}^{D} f_{il}\ln\epsilon_l + (1-f_{il})\ln(1-\epsilon_l) \quad (19)$$

$$\big\langle \ln p(\vec{\boldsymbol{\alpha}})\big\rangle = \sum_{j=1}^{M}\sum_{l=1}^{D}\big[u_{jl}\ln v_{jl} - \ln\Gamma(u_{jl}) + (u_{jl}-1)\big\langle\ln\alpha_{jl}\big\rangle - v_{jl}\bar{\alpha}_{jl}\big] \quad (20)$$

$$\big\langle \ln p(\vec{\boldsymbol{\beta}})\big\rangle = \sum_{j=1}^{M}\sum_{l=1}^{D}\big[p_{jl}\ln q_{jl} - \ln\Gamma(p_{jl}) + (p_{jl}-1)\big\langle\ln\beta_{jl}\big\rangle - q_{jl}\bar{\beta}_{jl}\big] \quad (21)$$

$$\big\langle \ln p(\vec{\boldsymbol{\lambda}})\big\rangle = \sum_{l=1}^{D}\sum_{k=1}^{K}\big[g_{lk}\ln h_{lk} - \ln\Gamma(g_{lk}) + (g_{lk}-1)\big\langle\ln\lambda_{lk}\big\rangle - h_{lk}\bar{\lambda}_{lk}\big] \quad (22)$$

$$\big\langle \ln p(\vec{\boldsymbol{\tau}})\big\rangle = \sum_{l=1}^{D}\sum_{k=1}^{K}\big[s_{lk}\ln t_{lk} - \ln\Gamma(s_{lk}) + (s_{lk}-1)\big\langle\ln\tau_{lk}\big\rangle - t_{lk}\bar{\tau}_{lk}\big] \quad (23)$$

$$\big\langle \ln Q(\mathcal{Z})\big\rangle = \sum_{i=1}^{N}\sum_{j=1}^{M} r_{ij}\ln r_{ij}\ , \qquad \big\langle \ln Q(\mathcal{W})\big\rangle = \sum_{i=1}^{N}\sum_{l=1}^{D}\sum_{k=1}^{K} m_{ilk}\ln m_{ilk} \quad (24)$$

$$\big\langle \ln Q(\vec{\boldsymbol{\phi}})\big\rangle = \sum_{i=1}^{N}\sum_{l=1}^{D} f_{il}\ln f_{il} + (1-f_{il})\ln(1-f_{il}) \quad (25)$$

$$\big\langle \ln Q(\vec{\boldsymbol{\alpha}})\big\rangle = \sum_{j=1}^{M}\sum_{l=1}^{D}\big[u^*_{jl}\ln v^*_{jl} - \ln\Gamma(u^*_{jl}) + (u^*_{jl}-1)\big\langle\ln\alpha_{jl}\big\rangle - v^*_{jl}\bar{\alpha}_{jl}\big] \quad (26)$$

$$\big\langle \ln Q(\vec{\boldsymbol{\beta}})\big\rangle = \sum_{j=1}^{M}\sum_{l=1}^{D}\big[p^*_{jl}\ln q^*_{jl} - \ln\Gamma(p^*_{jl}) + (p^*_{jl}-1)\big\langle\ln\beta_{jl}\big\rangle - q^*_{jl}\bar{\beta}_{jl}\big] \quad (27)$$

$$\big\langle \ln Q(\vec{\boldsymbol{\lambda}})\big\rangle = \sum_{l=1}^{D}\sum_{k=1}^{K}\big[g^*_{lk}\ln h^*_{lk} - \ln\Gamma(g^*_{lk}) + (g^*_{lk}-1)\big\langle\ln\lambda_{lk}\big\rangle - h^*_{lk}\bar{\lambda}_{lk}\big] \quad (28)$$

$$\big\langle \ln Q(\vec{\boldsymbol{\tau}})\big\rangle = \sum_{l=1}^{D}\sum_{k=1}^{K}\big[s^*_{lk}\ln t^*_{lk} - \ln\Gamma(s^*_{lk}) + (s^*_{lk}-1)\big\langle\ln\tau_{lk}\big\rangle - t^*_{lk}\bar{\tau}_{lk}\big] \quad (29)$$

C. Optimizing the Mixing Coefficients and Complete Algorithm

Now, that we have obtained a variational lower bound $\mathcal{L}(Q)$ which approximates the true marginal log likelihood $\ln p(\mathcal{X}|\mathbf{\Lambda})$, the model parameters $\mathbf{\Lambda}$ can be estimated by maximizing $\mathcal{L}(Q)$ with respect to $\boldsymbol{\pi}, \vec{\boldsymbol{\eta}}$ and $\boldsymbol{\epsilon}$. Setting the derivative of the lower bound with respect to π_j, η_{lk} and ϵ_l to zero, we get

$$\pi_j = \frac{1}{N}\sum_{i=1}^{N} r_{ij} \ , \qquad \eta_{lk} = \frac{1}{N}\sum_{i=1}^{N} m_{ilk} \ , \qquad \epsilon_l = \frac{1}{N}\sum_{i=1}^{N} f_{il} \tag{30}$$

Since the solutions for the variational posterior Q and the value of the lower bound depend on the values of $\boldsymbol{\pi}, \vec{\boldsymbol{\eta}}$ and $\boldsymbol{\epsilon}$, the optimization of the model can be solved in a way analogous to the EM algorithm. In the variational equivalent of the E-step, we optimize the variational solutions for each variational factor (Eq. 12 to Eq. 14). Then, in the subsequent variational equivalent of the M-step, we maximize the lower bound $\mathcal{L}(Q)$ with respect to the current values of $\boldsymbol{\pi}, \vec{\boldsymbol{\eta}}$ and $\boldsymbol{\epsilon}$. These two steps are repeated until convergence. The complete algorithm can be summarized as follows [4]:

1) Initialization
   - Choose initial number of components for M and K, and the initial values for hyper-parameters u_{jl}, v_{jl}, p_{jl}, q_{jl}, g_{lk}, h_{lk}, s_{lk} and t_{lk}.
   - Initialize the value of r_{ij} and m_{ilk} by K-means algorithm.
2) The variational E-step: Update the variational solutions through Eq. 12 to Eq. 14.
3) The variational M-step: maximize lower bound $\mathcal{L}(Q)$ with respect to the current values of $\boldsymbol{\pi}, \vec{\boldsymbol{\eta}}$ and $\boldsymbol{\epsilon}$ using Eq. 30.
4) Repeat steps 2 and 3 until convergence.
5) Detect the correct M and K by eliminating the components with small mixing coefficients (less than 10^{-5}).

It is noteworthy that the proposed algorithm allows implicitly and simultaneously model selection with parameter estimation and feature selection. This is different from classic approaches which perform model selection using model selection rules, derived generally under asymptotical assumption and information theoretic reasoning, such as MML, MDL and AIC [51]. A major drawback of these traditional approaches is that they require the entire learning process to be repeated for different models (i.e. different values of M and K in our case).

[4]The complete source codes are available from the authors.

IV. Experimental Results

In this section, we shall illustrate our results with a collection of simulation studies involving both artificial and real-world data, and a real challenging application namely human action videos categorization. The goal of the synthetic data is to investigate the accuracy of the variational approach. The applications involving real data have two main goals. The first goal is to compare our approach which we refer to as varFsGD to the MML-based unsupervised feature selection approach (MMLFsGD) previously proposed in [10]. The second goal is to compare varFsGD with the GD mixture learned in a variational way without feature selection (we refer to this approach as varGD). Please note that it is rather difficult to make a fruitful comparison among the many unsupervised feature selection techniques that have been proposed in the literature and this is not actually the aim of the paper. We have compared, however, our results with the variational Gaussian mixture-based unsupervised feature selection approach (we shall refer to as varFsGau) proposed in [13]. In all our experiments, we initialize the number of components M and K with large values (15 and 10, respectively) with equal mixing coefficients, and the feature saliency values are initialized at 0.5. In order to provide broad non-informative prior distributions, the initial value of u, p, g and s for the conjugate priors are set to 1, and v, q, h, t are set to 0.01. Then, the initial values of $\bar{\alpha}$, $\bar{\beta}$, $\bar{\lambda}$ and $\bar{\tau}$ can be calculated using Eq. 15.

A. Artificial Data

First, we have tested the performance of our variational algorithm in terms of estimation and selection, on six eleven-dimensional ground-truth synthetic data sets (three-dimensional relevant features and eight-dimensional irrelevant features). Following the scheme used in [10], the relevant features are generated in the transformed space from mixtures of Beta distributions with well-separated components and irrelevant ones from mixtures of overlapped components. The synthetic data sets are constructed based on different values of M (the number of components in the mixture of relevant features) and K (the number of components in the mixture of irrelevant features) with the corresponding parameters of each component. For instance, data set 1 is built with 900 instances. Its mixture of the relevant features has three components (i.e. $j = \{1, 2, 3\}$). Each component j has n_j instances with its own parameters $\{\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j\}$ and its corresponding mixing coefficient π_j. Eight irrelevant features which are sampled from a $Beta(X_{il}|1.5, 0.8)$ density are then appended to this data, yielding a set of 900 eleven-dimensional patterns. The mixture of the irrelevant features

TABLE I

PARAMETERS OF THE DIFFERENT GENERATED DATA SETS. N DENOTES THE TOTAL NUMBER OF ELEMENTS, n_j DENOTES THE NUMBER OF ELEMENTS IN CLUSTER j FOR THE RELEVANT FEATURES. α_{j1}, β_{j1}, α_{j2}, β_{j2}, α_{j3}, β_{j3} AND π_j ARE THE REAL PARAMETERS OF THE MIXTURE MODELS OF RELEVANT FEATURES. $\hat{\alpha}_{j1}$, $\hat{\beta}_{j1}$, $\hat{\alpha}_{j2}$, $\hat{\beta}_{j2}$, $\hat{\alpha}_{j3}$, $\hat{\beta}_{j3}$ AND $\hat{\pi}_j$ ARE THE ESTIMATED PARAMETERS FROM VARIATIONAL INFERENCE.

	n_j	j	α_{j1}	β_{j1}	α_{j2}	β_{j2}	α_{j3}	β_{j3}	π_j	$\hat{\alpha}_{j1}$	$\hat{\beta}_{j1}$	$\hat{\alpha}_{j2}$	$\hat{\beta}_{j2}$	$\hat{\alpha}_{j3}$	$\hat{\beta}_{j3}$	$\hat{\pi}_j$
Data set 1	300	1	30	15	20	40	33	18	0.33	27.94	14.32	18.65	41.27	32.13	17.52	0.32
(N = 900)	300	2	25	33	30	50	14	62	0.33	23.71	31.15	28.16	48.88	13.57	59.93	0.34
	300	3	40	30	35	26	27	12	0.34	39.54	29.36	36.22	24.51	25.33	11.89	0.34
Data set 2	200	1	30	15	20	20	33	18	0.23	28.68	14.14	19.01	19.55	31.76	17.54	0.24
(N = 900)	300	2	25	33	30	50	14	62	0.34	25.03	32.72	28.11	48.39	14.58	64.39	0.34
	400	3	40	30	19	21	15	10	0.43	35.57	26.34	18.73	20.58	15.77	9.81	0.42
Data set 3	800	1	45	55	62	47	54	39	0.53	46.01	57.86	60.15	45.29	51.04	41.68	0.54
(N = 1500)	700	2	59	60	50	65	35	45	0.47	58.10	58.16	48.43	61.89	34.51	47.84	0.46
Data set 4	200	1	15	16	20	15	17	36	0.16	15.31	17.09	19.23	15.21	16.33	38.19	0.16
(N = 1200)	200	2	18	35	10	25	20	13	0.16	18.95	37.17	10.15	23.94	22.18	12.57	0.15
	400	3	40	28	33	46	18	40	0.33	39.30	27.65	31.17	47.56	19.22	43.83	0.33
	400	4	30	44	25	40	35	22	0.35	30.24	45.79	23.61	38.39	33.37	24.15	0.36
Data set 5	300	1	16	33	10	28	25	17	0.25	16.70	35.20	9.75	26.12	27.31	16.55	0.24
(N = 1200)	300	2	19	17	33	14	15	18	0.25	18.05	15.71	35.06	16.21	14.48	18.52	0.25
	300	3	30	15	22	15	14	30	0.25	31.50	15.41	20.11	16.29	14.25	29.84	0.25
	300	4	26	32	11	19	34	21	0.25	24.67	31.50	11.69	18.38	35.51	20.24	0.26
Data set 6	200	1	16	16	20	19	36	20	0.13	14.56	15.80	18.82	19.36	35.53	20.41	0.13
(N = 1500)	200	2	18	35	33	46	20	13	0.13	19.34	33.53	30.44	43.15	18.18	12.65	0.13
	300	3	40	28	36	10	21	22	0.20	38.97	29.15	38.71	10.12	20.08	23.36	0.19
	300	4	30	44	18	30	32	29	0.20	29.83	44.09	16.92	31.16	33.42	30.08	0.20
	500	5	25	20	43	15	12	19	0.34	23.19	19.28	44.89	15.49	11.73	18.66	0.35

for data set 1 contains only one component (i.e. $K = 1$). Notice that, in this experiment, we choose that irrelevant features are generated, per-component, from the same density (i.e. $(\lambda_{1k}, \tau_{1k}) = (\lambda_{2k}, \tau_{2k}) = \ldots = (\lambda_{8k}, \tau_{8k})$).

Table I illustrates the different synthetic data sets with their real and estimated parameters of the relevant mixture models. According to Table I, we can observe that our approach is able to estimate accurately the parameters of the relevant mixture models. The real and estimated parameters of the mixture models representing irrelevant features are shown

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.

in Tables II and III. Within these two tables, K denotes the number of components of the irrelevant mixture models,

TABLE II

ESTIMATED PARAMETERS OF THE MIXTURE MODELS REPRESENTING THE IRRELEVANT FEATURES FOR DATA SETS 1, 2 AND 3.

	Data set 1 K =1		Data set 2 K =2				Data set 3 K =3					
	$\lambda_{l1}=1.5$	$\tau_{l1}=0.8$	$\lambda_{l1}=0.3$	$\tau_{l1}=0.1$	$\lambda_{l2}=0.6$	$\tau_{l2}=0.7$	$\lambda_{l1}=2$	$\tau_{l1}=1$	$\lambda_{l2}=0.5$	$\tau_{l2}=0.5$	$\lambda_{l3}=0.3$	$\tau_{l3}=0.1$
l	$\hat{\lambda}_{l1}$	$\hat{\tau}_{l1}$	$\hat{\lambda}_{l1}$	$\hat{\tau}_{l1}$	$\hat{\lambda}_{l2}$	$\hat{\tau}_{l2}$	$\hat{\lambda}_{l1}$	$\hat{\tau}_{l1}$	$\hat{\lambda}_{l2}$	$\hat{\tau}_{l2}$	$\hat{\lambda}_{l3}$	$\hat{\tau}_{l3}$
1	1.46	0.79	0.33	0.11	0.56	0.74	2.23	1.16	0.46	0.48	0.32	0.11
2	1.44	0.77	0.31	0.12	0.64	0.73	2.11	1.04	0.51	0.52	0.27	0.10
3	1.49	0.79	0.32	0.08	0.59	0.67	1.91	0.95	0.52	0.53	0.31	0.12
4	1.51	0.81	0.29	0.13	0.57	0.66	1.98	0.97	0.54	0.51	0.35	0.11
5	1.55	0.82	0.32	0.12	0.52	0.67	2.01	1.06	0.49	0.49	0.33	0.11
6	1.48	0.83	0.31	0.09	0.56	0.71	2.07	0.93	0.45	0.51	0.28	0.11
7	1.52	0.71	0.27	0.11	0.61	0.71	2.11	1.92	0.51	0.57	0.32	0.10
8	1.54	0.78	0.28	0.11	0.63	0.73	1.84	1.11	0.49	0.54	0.29	0.11

TABLE III

ESTIMATED PARAMETERS OF THE MIXTURE MODELS REPRESENTING THE IRRELEVANT FEATURES FOR DATA SETS 4, 5 AND 6.

Data set 4 K =2				Data set 5 K =3						Data set 6 K =3					
$\lambda_1=2$	$\tau_1=2$	$\lambda_2=3$	$\tau_2=5$	$\lambda_1=1$	$\tau_1=3$	$\lambda_2=2$	$\tau_2=4$	$\lambda_3=5$	$\tau_3=1$	$\lambda_1=3$	$\tau_1=3$	$\lambda_2=2$	$\tau_2=5$	$\lambda_3=5$	$\tau_3=2$
$\hat{\lambda}_1$	$\hat{\tau}_1$	$\hat{\lambda}_2$	$\hat{\tau}_2$	$\hat{\lambda}_1$	$\hat{\tau}_1$	$\hat{\lambda}_2$	$\hat{\tau}_2$	$\hat{\lambda}_3$	$\hat{\tau}_3$	$\hat{\lambda}_1$	$\hat{\tau}_1$	$\hat{\lambda}_2$	$\hat{\tau}_2$	$\hat{\lambda}_3$	$\hat{\tau}_3$
2.17	1.98	2.88	5.11	1.02	2.85	2.12	4.01	5.27	1.07	2.87	3.02	1.81	4.94	5.15	2.13
2.06	2.19	2.93	4.98	0.97	2.94	1.95	3.95	5.16	1.05	2.95	2.89	1.85	4.91	5.11	2.07
1.86	1.87	2.98	4.87	1.02	3.14	1.96	3.87	4.87	1.03	2.94	2.91	2.13	5.02	4.92	1.98
2.13	2.02	2.82	4.79	0.98	2.93	1.89	3.98	5.04	0.98	2.81	2.87	2.02	5.17	5.08	1.89
1.92	1.90	2.91	4.82	1.09	3.07	1.91	4.16	5.17	1.02	3.13	2.98	1.94	4.95	4.97	1.93
1.89	1.91	3.01	4.98	0.91	2.96	2.15	4.08	4.96	0.97	2.86	2.95	1.82	4.91	5.04	2.97
2.12	2.13	2.78	4.75	1.07	2.93	1.95	3.89	5.05	1.01	3.08	3.04	1.97	5.12	4.87	2.14
2.07	2.11	3.06	5.16	1.16	3.08	2.06	4.07	5.07	1.01	2.95	3.11	2.01	4.89	5.05	2.02

$\{\lambda_{lk}, \tau_{lk}\}$ are the real parameters and $\{\hat{\lambda}_{lk}, \hat{\tau}_{lk}\}$ are the estimated parameters by variational inference. Table IV shows the true and estimated mixing coefficients of the irrelevant mixtures. According to these three tables, we can see clearly that the parameters of the irrelevant mixture models are estimated accurately by our algorithm.

TABLE IV

REAL AND ESTIMATED MIXING COEFFICIENTS OF THE IRRELEVANT MIXTURES FOR DIFFERENT DATA SETS. η_k IS THE REAL MIXING COEFFICIENT OF COMPONENT k. $\hat{\eta}_k$ IS THE ESTIMATED VALUE.

	Data set 1	Data set 2		Data set 3			Data set 4		Data set 5			Data set 6		
k	1	1	2	1	2	3	1	2	1	2	3	1	2	3
η_k	1.00	0.50	0.50	0.33	0.33	0.34	0.50	0.50	0.33	0.33	0.34	0.27	0.33	0.40
$\hat{\eta}_k$	1.00	0.49	0.51	0.32	0.34	0.34	0.50	0.50	0.33	0.33	0.34	0.27	0.34	0.39

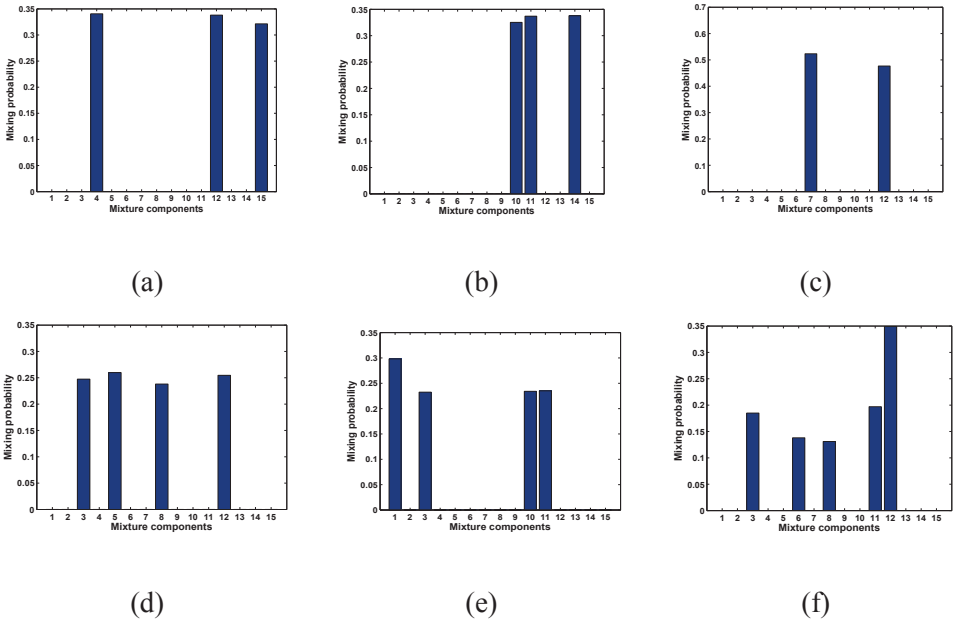


Fig. 2. Mixing probabilities of components, π_j, found for the different generated data sets after convergence. (a) Data set 1, (b) Data set 2, (c) Data set 3, (d) Data set 4, (e) Data set 5, (f) Data set 6.

Figure 2 shows the estimated mixing coefficients of the mixture components, in each data set, after convergence. By removing the components with very small mixing coefficients in each data set, we obtain the correct number of components for the mixtures representing relevant features. The estimated mixing coefficients for the irrelevant mixtures of each synthetic data are shown in Fig. 3. Based on this figure, it is obvious that the appropriate number of components for the irrelevant features can be obtained by eliminating the coefficients that are close to 0 for each generated data.

The feature saliencies of all the 11 features for each generated data set are shown in Fig. 4. It is clear that features 1, 2 and 3 have been assigned a high degree of relevance which is consistent with the ground-truth. Therefore, we can

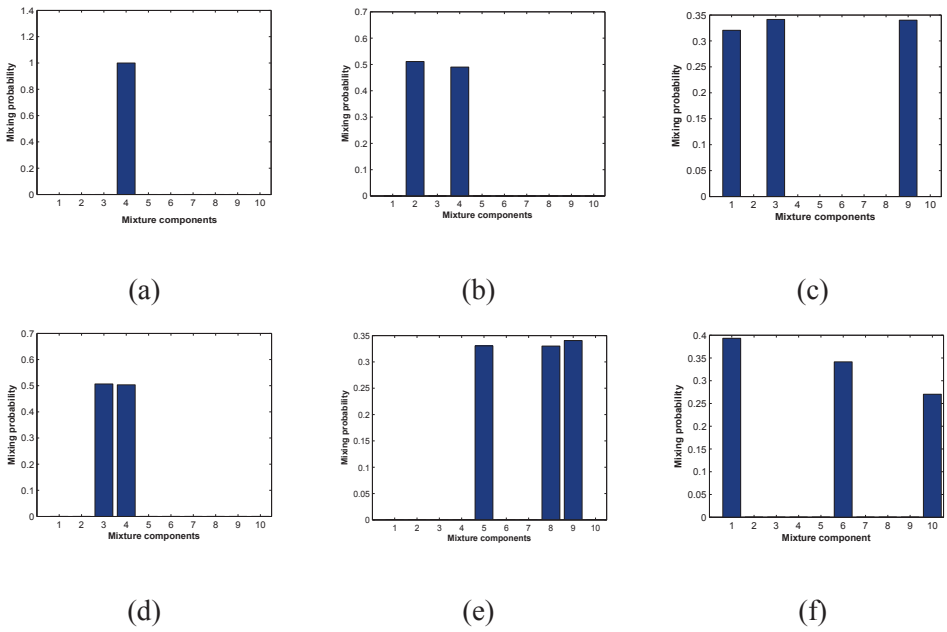


Fig. 3. Mixing probabilities of components, η_k, found for the irrelevant features of the different generated data sets after convergence. (a) Data set 1, (b) Data set 2, (c) Data set 3, (d) Data set 4, (e) Data set 5, (f) Data set 6.

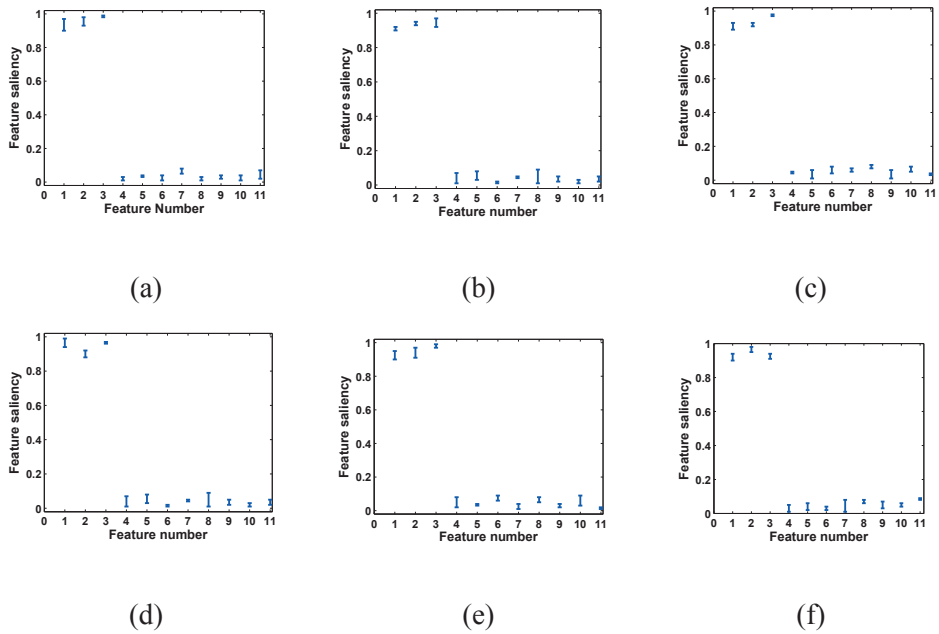


Fig. 4. Feature saliency for synthetic data sets with one standard deviation over ten runs. (a) Data set 1, (b) Data set 2, (c) Data set 3, (d) Data set 4, (e) Data set 5, (f) Data set 6.

conclude that, for synthetic data sets, the proposed algorithm successfully detects the true number of components and correctly assigns the importance of features.

Next, two more synthetic data sets are generated to validate our algorithm when $N \leq D$. Data set A consists of 50 sixty-dimensional (10 relevant and 50 irrelevant features) instances where each relevant feature is generated from

a mixture of two Beta distributions having equal weights: $Beta(X_{il}|15, 20)$ and $Beta(X_{il}|25, 13)$, and the 50 irrelevant features are generated from $Beta(X_{il}|0.2, 0.8)$. Data set B contains 60 sixty-dimensional instances where 10 features are relevant and synthesized from a mixture of three Beta components with equal mixing coefficients: $Beta(X_{il}|15, 20)$, $Beta(X_{il}|25, 13)$ and $Beta(X_{il}|35, 18)$. Each irrelevant feature in data set B is generated from a mixture of two Beta distributions having equal weights: $Beta(X_{il}|0.5, 2)$ and $Beta(X_{il}|1.3, 0.8)$. The corresponding estimated mixing coefficients and feature saliencies of each data set are illustrated in Fig 5 which clearly shows that, although $N \leq D$ in each test set, our variational approach still can learn the corresponding models correctly.

Furthermore, we have evaluated the performance of the proposed varFsGD on a known $10,000$-dimensional synthetic

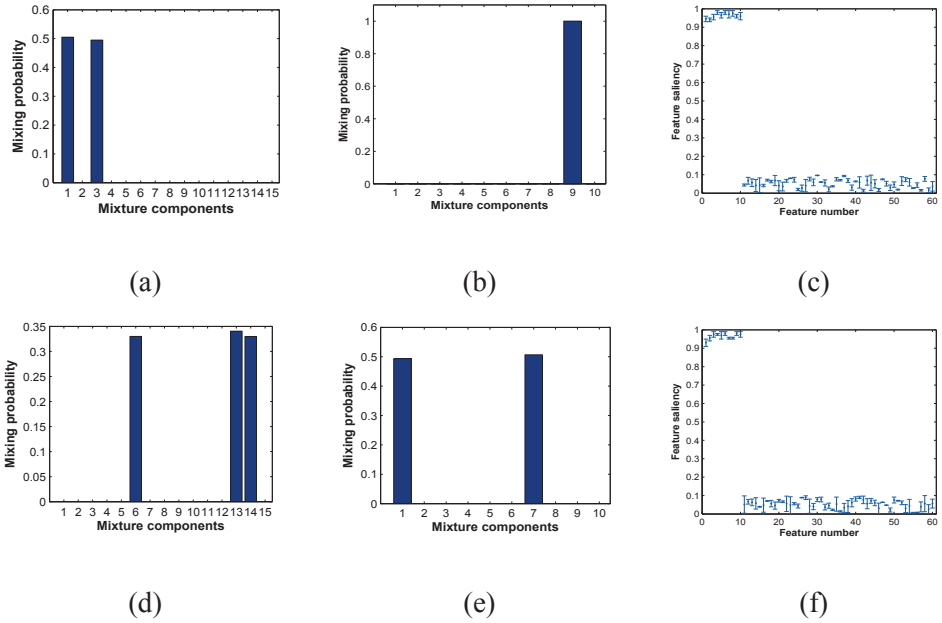


Fig. 5. Experiments on data sets A and B by using the varFsGD: (a), (b) and (c) represent the estimated relevant mixing coefficients, irrelevant mixing coefficients and feature saliencies for data set A, respectively; (d), (e) and (f) show the estimated relevant mixing coefficients, irrelevant mixing coefficients and feature saliencies for data set B, respectively.

example which has been previously proposed in [52] and adopted, for instance, in our previous work [10]. This data set contains 100 instances where the relevant features are sampled from a mixture of two 150-dimensional Gaussian distributions: $\mathcal{N}(1.5, 0.2)$ and $\mathcal{N}(0, 1)$, while the irrelevant features are sampled from $\mathcal{N}(0, 1)$ with a dimensionality of 9,850. After running the proposed algorithm on this synthetic data set, we have observed that our varFsGD can accurately detect the 150 relevant features by assigning to them high degrees of relevancy ($\epsilon_l > 0.9$). The remaining irrelevant features have been correctly recognized in our test, since low values of saliency ($\epsilon_l < 0.1$) have been

assigned to them.

B. Real Data

In this section, we evaluate the proposed algorithm on four real-world data sets with different properties, as shown in Table V. The spambase data set (SP) contains a collection of spam and non-spam e-mails. The aim is to determine

TABLE V

THE THREE REAL DATA SETS. N, D AND M DENOTE THE NUMBERS OF INSTANCES, FEATURES AND CLASSES, RESPECTIVELY.

Data set	N	D	M
Spambase	4601	57	2
Statlog	6435	36	6
Image Segmentation	2320	19	7
Handwritten Digits	5620	64	10

if an e-mail is spam or legitimate. It contains 4601 57-dimensional vectors divided into two classes. The statlog data set (ST) consists of the multi-spectral values of pixels in 3×3 neighborhoods in a satellite image, and the classification associated with the central pixel in each neighborhood. It contains 6,435 36-dimensional vectors from six classes: read soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble and very damp grey soil. The image segmentation data set (IS) contains 2320 instances each of which is defined by 18 features. Each instance describes a 3×3 region drawn from seven types of outdoor images: brickface, sky, foliage, cement, window, path and grass. The handwritten digits data set (HD) contains 5,620 vectors with 64 features from ten classes: '0' to '9'. All the data sets are taken from the UCI machine learning repository [5]. Since the features of all data sets are within some specific range, normalization is performed as a preprocessing step to transform all the data points into the range of [0,1]. We evaluate the performance of the proposed algorithm by running it 20 times with 30 initial components. We also initialize the number of components for irrelevant features to 15. For comparison, we also apply MMLFsGD, varGD and varFsGau on the same data sets. The results are summarized in Table VI. According to this table, we can observe that the improvement is immediately obvious with our method since it decreases the error rate and selects

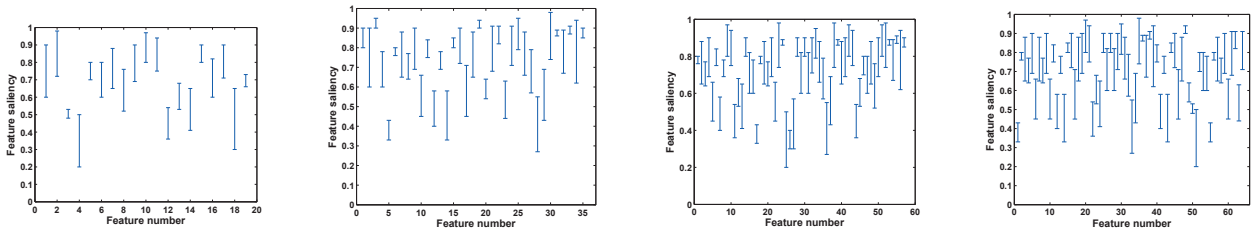
[5] http://www.ics.uci.edu/∼mlearn/MLRepository.html

TABLE VI

THE AVERAGE ERROR AND AVERAGE NUMBER OF COMPONENTS ($\hat{M}$) COMPUTED USING VARFSGD, MMLFSGD, VARGD AND VARFSGAU OVER 20 RANDOM RUNS.

	varFsGD		**MMLFsGD**		**varGD**		**varFsGau**	
Data set	error (%)	$\hat{M}$	error (%)	$\hat{M}$	error (%)	$\hat{M}$	error (%)	$\hat{M}$
SP	6.54 ± 1.53	2.08 ± 0.18	7.15 ± 1.38	2.12 ± 0.75	9.27 ± 2.01	2.06 ± 0.86	9.04 ± 1.47	2.26 ± 0.61
ST	9.97 ± 0.86	6.88 ± 1.02	11.38 ± 1.28	7.02 ± 1.57	16.20 ± 0.92	6.56 ± 1.19	15.82 ± 1.53	7.38 ± 1.36
IS	15.29 ± 1.95	6.53 ± 1.54	15.71 ± 2.13	7.49 ± 1.77	21.13 ± 1.02	7.25 ± 1.82	22.54 ± 1.87	7.62 ± 0.97
HD	11.86 ± 1.49	10.78 ± 0.84	13.53 ± 0.58	11.05 ± 1.11	18.64 ± 0.73	10.02 ± 1.64	17.97 ± 0.68	11.14 ± 0.22

accurately the correct number of components for all the data sets. The features saliencies for the different tested data sets are given in Fig. 6. According to this figure, it is obvious that the different features do not contribute equally to the clustering since different relevancy degrees are affected to them. Take the IS data set for instance, there are three features (features number 4, 12 and 18) that have saliencies less than 0.5, and therefore have less impact on clustering structure.



(a) (b) (c) (d)

Fig. 6. Feature saliencies for the different real world data sets over 20 runs. (a) IS data set, (b) ST data set, (c) SP data set, (d) HD data set.

C. Human Action Videos Categorization

With the rapid development of digital technologies, the increase in the availability of multimedia data such as images and videos is tremendous. With thousands of videos on hand, grouping them according to their contents is highly important for a variety of visual tasks such as event analysis [53], video indexing, browsing and retrieval, and digital

libraries organization [54]. How to provide efficient videos categorization approaches has attracted many research efforts and has been addressed by several researchers in the past (see, for instance, [55–60]). Videos categorization remains, however, an extremely challenging task due to several typical scenarios such as unconstrained motions, cluttered scenes, moving backgrounds, object occlusions, non-stationary camera, geometric changes and deformation of objects and variations of illumination conditions and viewpoints. In this section, we present an unsupervised learning method, based on our variational algorithm, for categorizing human action videos. The performance of the proposed method is evaluated on a challenging video data set namely the KTH [61] human action data set.

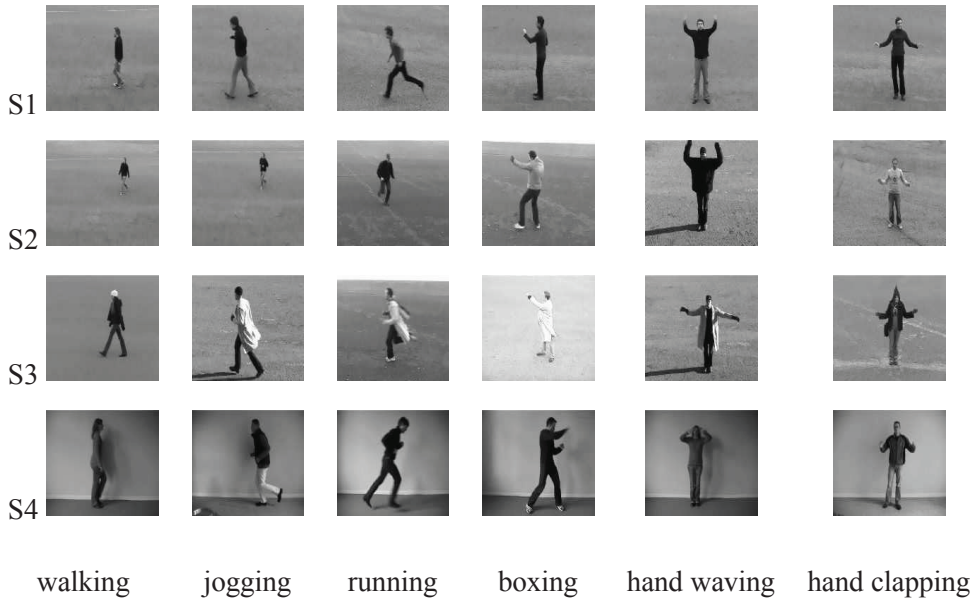


Fig. 7. Examples of frames, representing different human actions in different scenarios, from video sequences in the KTH data set.

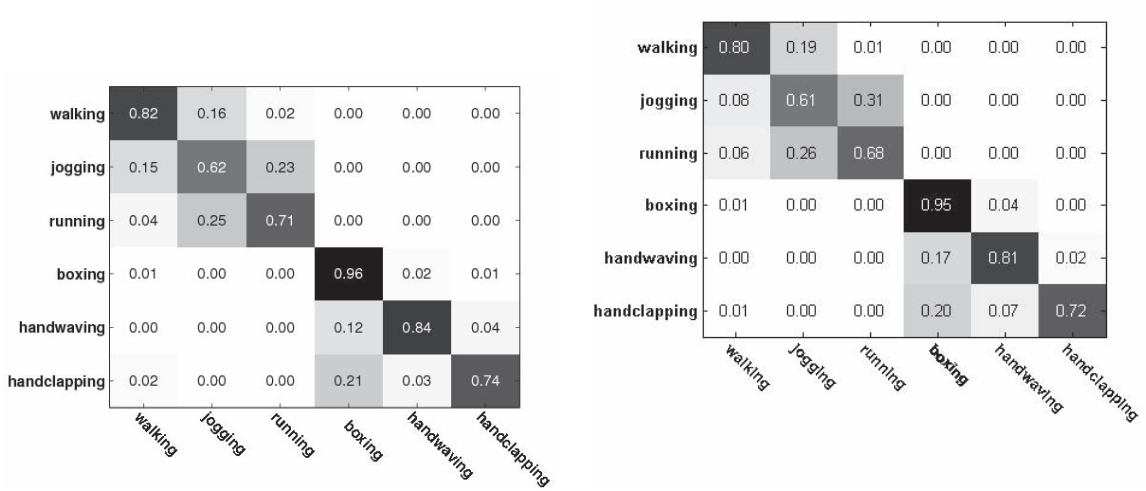
1) Experimental Methodology: Several studies have been conducted to provide models and visual features in order to consistently (i.e. regardless changes in viewpoint angles, position, distance, size, orientation, or deformation) categorize objects and visual scenes. These studies have shown that a good model is required, and it must be able to select relevant visual features to improve categorization performance [62, 63]. Recently several works have been based on the notion of visual vocabulary constructed via a quantization process, according to a coding rule such as K-Means, of local features (spatio-temporal features in the case of videos) extracted from a set of detected interest points (space-time interest points in the case of videos). This approach allows the representation of images and videos as histograms of visual words and have convincingly proven its effectiveness in several applications (see, for instance, [64]). Here we consider this approach and our methodology for unsupervised videos categorization can be summarized as follows.

First, local spatio-temporal features from each video sequence are extracted from their detected space-time interest points. Among many of the existing space-time interest points detectors and local spatio-temporal features [55, 65–68], we employ the space-time interest point detector proposed in [57] [6], which is actually a space-time extension of the well-known Harris operator, and histograms of optic flow (HoF) as proposed in [65]. Next, a visual vocabulary is constructed by quantizing these spatio-temporal features into visual words using K-means algorithm and each video is then represented as a frequency histogram over the visual words. Then, we apply the pLSA model [69] to the obtained histograms as done in [70] in the case of still images. As a result each video is represented now by a D-dimensional proportional vector where D is the number of latent aspects. Finally, we employ our varFsGD model as a classifier to categorize videos by assigning the video sequence to the group which has the highest posterior probability according to Bayes' decision rule.

2) KTH Human Action Data Set: The KTH human action data set is one of the largest available video sequences data sets of human actions [61]. It contains six types of human action classes including: walking, jogging, running, boxing, hand waving and hand clapping. Each action class is performed several times by 25 subjects in four different scenarios: outdoors (S1), outdoors with scale variation (S2), outdoors with different clothes (S3) and indoors (S4). This data set contains 2391 video sequences and all sequences were taken over homogenous backgrounds with a static camera with 25ftps frame rate. All video samples were downsampled to the spatial resolution of 160×120 pixels and have a length of four seconds in average. Examples of frames from video sequences of each category are shown in Fig. 7. In this experiment, we considered a training set composed of actions related to 16 subjects to construct the visual vocabulary, by setting the number of clusters in the K-Means algorithm (i.e. number of visual words) to 1000, as explained in the previous section. The pLSA model was applied by considering 40 aspects and each video in the database was then represented by a 40-dimensional vector of proportions. Last, the resulting vectors were clustered by our varFsGD model. The entire procedure was repeated 20 times for evaluating the performance of our approach. The optimal number of components was estimated as around 6 while the number of irrelevant Beta components was identified as $K = 2$. The confusion matrix for the KTH data set is shown in Fig. 8.a. We note that, most of the

[6] We have also tested another popular feature detector namely the Cuboid detector proposed in [55]. However, we have not noticed a significant improvement according to our experiments.

confusion takes place between “walking” and “jogging”, “jogging” and “running”, as well as between “hand clapping” and “boxing”. This is due to the fact that similar actions contain similar types of local space-time events. In order



(a) (b)

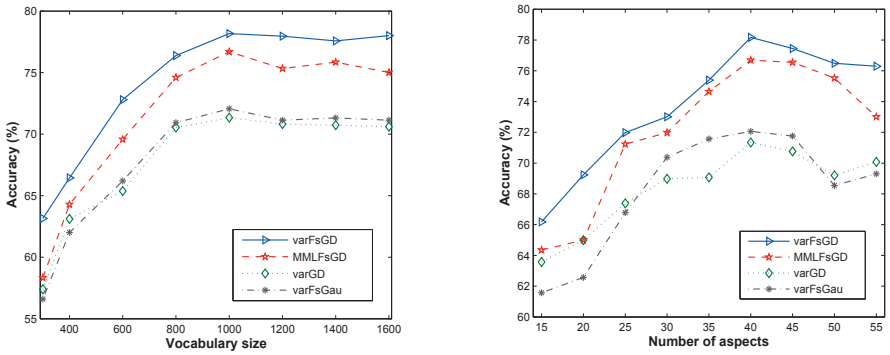
Fig. 8. Confusion matrix for the KTH data set. (a) K is determined automatically via the learning algorithm, (b) K is set to one.

to demonstrate the importance of using mixtures for modeling irrelevant features in this experiment, we have also implemented our algorithm by fixing K equal to 1. The resultant confusion matrix is shown in Fig. 8.b which shows clearly that the average classification accuracy has decreased. Table VII shows the average classification accuracy and the average number of relevant components obtained by varFsGD, MMLFsGD, varGD and varFsGau. It clearly shows that our algorithm outperforms the other approaches for clustering KTH human action videos. For instance, the fact that the varFsGD performs better than the varFsGau is actually expected since videos are represented by vectors of proportions for which the GD mixture is one of the best modeling choices unlike the Gaussian mixture which implicitly assumes that the features vectors are Gaussian which is far from the case . We have also tested the effect of different sizes of visual vocabulary on classification accuracy for varFsGD, MMLFsGD, varGD and varFsGau, as illustrated in Fig. 9(a). As we can see, the classification rate peaks around 1000. The choice of the number of aspects also influences the accuracy of classification. As shown in Fig. 9(b), the optimal accuracy can be obtained when the number of aspects is set to 40. These aspects may contribute with different degrees in the categorization of the action videos. The corresponding feature saliencies of the 40-dimensional aspects together with their standard deviations (error bars) are shown in Fig. 10. As illustrated in this figure, the features have different relevance degrees and then

TABLE VII

THE AVERAGE CLASSIFICATION ACCURACY AND THE NUMBER OF COMPONENTS ($\hat{M}$) COMPUTED ON THE KTH DATA SET USING VARFSGD, MMLFSGD, VARGD AND VARFSGAU OVER 20 RANDOM RUNS.

Algorithm	$\hat{M}$	Accuracy (%)
varFsGD	5.96	78.17
MMLFsGD	5.87	76.69
varGD	5.53	71.34
varFsGau	5.67	72.06



(a) (b)

Fig. 9. (a) Classification accuracy vs. vocabulary size for the KTH data set; (b) Classification accuracy vs. the number of aspects for the KTH data set.

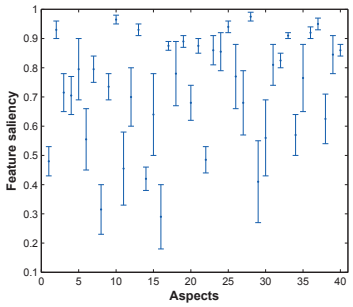


Fig. 10. Feature saliencies of the different aspect features over 20 runs for the KTH data set.

contribute differently to clustering. For instance, there are seven features (features number 1, 8, 11, 14, 16, 22, 29) that have saliencies lower than 0.5, and then provide less contribution in clustering. This is because these aspects are associated to all categories and have less discrimination power. By contrast, eight features (features number 2, 10, 13,

25, 28, 33, 36 and 37) have high relevance degrees with feature saliencies greater than 0.9 which can be explained by the fact that these features are mainly associated with specific action categories. Moreover, as we can see in Fig. 10, the standard deviations of the features with high relevance degrees are generally smaller than those for the irrelevant features, which means that our algorithm for selecting relevant features is efficient and consistent.

V. Conclusion

Most of the feature selection algorithms based on mixture models assume that the data in each component follow Gaussian distribution, which is seldom the case in real-life applications. Unlike these approaches, we have proposed in this paper a principled variational framework for unsupervised feature selection in the case of non-Gaussian data which naturally appear in many application from different domains and disciplines. Variational frameworks offer a deterministic alternative for Bayesian approximate inference by maximizing a lower bound on the marginal likelihood which main advantage is computational efficiency and guaranteed convergence that can be easily assessed as compared to MCMC-based approaches which make posterior approximation in a stochastic sense. We have shown that the variational approach can be used to obtain a closed form parameters posteriors for our model. The proposed approach has been applied to both synthetic and real data, and to a challenging application which concerns human action videos categorization, with encouraging results. It is noteworthy that the proposed selection model is also applicable to many other challenging problems involving non-Gaussian proportional data such as text mining and compression, and protein sequences modeling in biology. There are several interesting possible future works such as handling the important problem of online learning in the case of dynamic data sets where the relevancy of the features may vary as new data arrive or disappear. A possible solution to this problem could be the extension of the proposed model to the infinite case by integrating it with the non-parametric Bayesian framework recently proposed in [71].

Acknowledgment

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC). The authors would like to thank the anonymous referees and the associate editor for their helpful comments. The complete source code of this work is available upon request.

REFERENCES

[1] J. H. Friedman and J. W. Tukey, "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Transactions on Computers*, vol. 23, no. 9, pp. 881–890, 1974.

[2] P. E. Green, F. J. Carmone and J. Kim, "A Preliminary Study of Optimal Variable Weighting in K-Means Clustering," *Journal of Classification*, vol. 7, no. 2, pp. 271–285, 1990.

[3] F. Z. Brill, D. E. Brown and W. N. Martin, "Fast Genetic Selection of Features for Neural Network Classifiers," *IEEE Transactions on Neural Networks*, vol. 3, no. 2, pp. 324–328, 1992.

[4] F. E. Shaudys and T. K. Leen, "Feature Selection for Improved Classification," in *Proc. of IEEE International Joint Conference on Neural Networks (IJCNN)*, pp. 697–702, 1992.

[5] P. V. Verveer and R. P. W. Duin, "An Evaluation of Intrinsic Dimensionality Estimators," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 1, pp. 81–86, 1995.

[6] C. E. Brodley and P. E. Utgoff, "Multivariate Decision Trees," *Machine Learning*, vol. 19, no. 1, pp. 45–77, 1995.

[7] Y. Kim, W. N. Street and F. Menczer, "Feature Selection in Unsupervised Learning via Evolutionary Search," in *Proc. of the International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 365–369, 2000.

[8] H. Liu and H. Motoda, *Computational Methods of Feature Selection*. Chapman & Hall, 2008.

[9] M. H. C. Law, M. A. T. Figueiredo and A. K. Jain, "Simultaneous Feature Selection and Clustering Using Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1154–1166, 2004.

[10] S. Boutemedjet, N. Bouguila and D. Ziou, "A Hybrid Feature Extraction Selection Approach for High-Dimensional Non-Gaussian Data Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1429–1443, 2009.

[11] M. W. Graham and D. J. Miller, "Unsupervised Learning of Parsimonious Mixtures on Large Spaces with Integrated Feature and Component Selection," *IEEE Transactions on Signal Processing*, vol. 54, no. 4, pp. 1289–1303, 2006.

[12] A. E. Raftery and N. Dean, "Variable Selection for Model-Based Clustering," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 168–178, 2006.

[13] C. Constantinopoulos, M. K. Titsias and A. Likas, "Bayesian Feature and Model Selection for Gaussian Mixture Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 6, pp. 1013–1018, 2006.

[14] N. Bouguila, "A Model-Based Approach for Discrete Data Clustering and Feature Weighting using MAP and Stochastic Complexity," *IEEE Transactions Knowledge and Data Engineering*, vol. 21, pp. 1649–1664, Dec 2009.

[15] C. Maugis, G. Celeux and M-L. Martin-Magniette, "Variable Selection for Clustering with Gaussian Mixture Models," *Biometrics*, vol. 65, no. 3, pp. 701–709, 2009.

[16] H. Lian, "Sparse Bayesian Hierarchical Modeling of High-Dimensional Clustering Problems," *Journal of Multivariate Analysis*, vol. 101, no. 7, pp. 1728–1737, 2010.

[17] Y. Li, M. Dong and J. Hua, "Simultaneous Localized Feature Selection and Model Detection for Gaussian Mixtures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 953–960, 2009.

[18] C. Fraley and A. E. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[19] N. Bouguila and D. Ziou, "A Hybrid SEM Algorithm for High-Dimensional Unsupervised Learning Using a Finite Generalized Dirichlet Mixture," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2657– 2668, 2006.

[20] N. Bouguila and D. Ziou, "High-Dimensional Unsupervised Selection and Estimation of a Finite Generalized Dirichlet Mixture Model Based on Minimum Message Length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1716–1731, 2007.

[21] C. S. Wallace, *Statistical and Inductive Inference by Minimum Message Length*. Springer-Verlag, 2005.

[22] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*. New York: Wiley-Interscience, 1997.

[23] A. L. Yuille, P. Stolorz and J. Utans, "Statistical Physics, Mixtures of Distributions, and the EM Algorithm," *Neural Computation*, vol. 6, no. 2, pp. 334–340, 1994.

[24] Z. Lu and H. H. S. Ip, "Generalized Competitive Learning of Gaussian Mixture Models," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 39, no. 4, pp. 901–909, 2009.

[25] C.P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer-Verlag, 1999.

[26] O. Cappé and C. P. Robert, "Markov Chain Monte Carlo: 10 Years and Still Running!," *Journal of the American Statistical Association*, vol. 95, no. 452, pp. 1282–1286, 2000.

[27] T. S. Jaakkola and M. I. Jordan, "Bayesian Parameter Estimation via Variational Methods," *Statistics and Computing*, vol. 10, no. 1, pp. 25–37, 2000.

[28] P. Hall, K. Humphreys and D. M. Titterington, "On the Adequacy of Variational Lower Bound Functions for Likelihood-Based Inference in Markovian Models with Missing Values," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64, no. 3, pp. 549–564, 2002.

[29] N. de Freitas, P. Højen-Sørensen, M. I. Jordan and S. Russel, "Variational MCMC," in *Proc. of the Conference in Uncertainty in Artificial Intelligence (UAI)*, pp. 120–127, 2001.

[30] C. M. Bishop, D. Spiegelhalter and J. Winn, "VIBES: A Variational Inference Engine for Bayesian Networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 777–784, 2002.

[31] T. S. Jaakkola and M. I. Jordan, "Variational Probabilistic Inference and the QMR-DT Network," *Journal of Artificial Intelligence Research*, vol. 10, pp. 291–322, 1999.

[32] S. J. Roberts and W. D. Penny, "Variational Bayes for Generalized Autoregressive Models," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002.

[33] N. Nasios and A. G. Bors, "Variational Expectation-Maximization Training for Gaussian Networks," in *Proc. of IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pp. 339–348, 2003.

[34] N. Nasios and A. G. Bors, "Variational Segmentation of Color Images," in *Proc. of IEEE International Conference on Image Processing (ICIP)*, pp. 614–617, 2005.

[35] G.J. McLachlan and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[36] G. W. Milligan, "A Validation Study of a Variable Weighting Algorithm for Cluster Analysis," *Journal of Classification*, vol. 6, no. 1, pp. 53–71, 1989.

[37] M. J. Brusco and J. D. Cradit, "A Variable-Selection Heuristic for K-Means Clustering," *Psychometrika*, vol. 66, no. 2, pp. 249–270, 2001.

[38] E. B. Fowlkes, R. Gnanadesikan and J. R. Kettenring, "Variable Selection in Clustering," *Journal of Classification*, vol. 5, no. 2, pp. 205–228, 1988.

[39] J. Z. Huang, M. K. Ng, H. Rong and Z. Li, "Automated Variable Weighting in K-Means Type Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 657–668, 2005.

[40] W. S. DeSarbo, J. D. Carroll, L. A. Clark and P. E. Green, "Synthesized Clustering: A Method for Amalgamating Alternative Clustering Based With Differential Weighting of Variables," *Psychometrika*, vol. 49, no. 1, pp. 57–78, 1984.

[41] D. A. Bell and H. Wang, "A Formalism for Relevance and Its Application in Feature Subset Selection," *Machine Learning*, vol. 41, no. 2, pp. 175–195, 2000.

[42] N. Bouguila, D. Ziou and E. Monga, "Practical Bayesian Estimation of a Finite Beta Mixture Through Gibbs Sampling and its Applications," *Statistics and Computing*, vol. 16, no. 2, pp. 215–225, 2006.

[43] E. Castillo, A.S. Hadi, C. Solares, "Learning and Updating of Uncertainty in Dirichlet Models," *Machine Learning*, vol. 26, no. 1, pp. 43–63, 1997.

[44] N. Bouguila, D. Ziou and R. I. Hammoud, "On Bayesian Analysis of a Finite Generalized Dirichlet Mixture Via a Metropolis-within-Gibbs Sampling," *Pattern Analysis and Applications*, vol. 12, no. 2, pp. 151–166, 2009.

[45] Z. Ma and A. Leijon, "Bayesian Estimation of Beta Mixture Models with Variational Inference," *IEEE Transactions on PAMI*, vol. 33, no. 11, pp. 2160 – 2173, 2011.

[46] A. Corduneanu and C. M. Bishop, "Variational Bayesian Model Selection for Mixture Distributions," in *Proc. of the 8th International Conference on Artificial Intelligence and Statistics (AISTAT)*, pp. 27–34, 2001.

[47] G. Parisi, *Statistical Field Theory*. Addison-Wesley, 1988.

[48] L. K. Saul and M. I. Jordan, "Exploiting Tractable Substructures in Intractable Networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 486–492, 1995.

[49] T. S. Jaakkola and M. I. Jordan, "Computing Upper and Lower Bounds on Likelihoods in Intractable Networks," in *Proc. of the Conference in Uncertainty in Artificial Intelligence (UAI)*, pp. 340–348, 1996.

[50] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[51] N. Bouguila and D. Ziou, "Unsupervised Selection of a Finite Dirichlet Mixture Model: An MML-Based Approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 8, pp. 993–1009, 2006.

[52] J. H. Friedman and J. J. Meulman, "Clustering Objects on Subsets of Attributes (with discussions)," *Journal of the Royal Statistical Society, Series B*, vol. 66, no. 4, pp. 815–849, 2004.

[53] X. Zhou, X. Zhuang, S. Yan, S-F. Chang, M. Hasegawa-Johnson and T. S. Huang, "SIFT-Based Kernel for Video Event Analysis," in *Proc. of the ACM International Conference on Multimedia (MM)*, pp. 229–238, 2008.

[54] D. Zhong, H. Zhang and S-F. Chang, "Clustering Methods for Video Browsing and Annotation," in *Proc. of the SPIE Conference on Storage and Retrieval for Video and Image Databases*, pp. 239–246, 1997.

[55] P. Dollár, V. Rabaud, G. Cottrell and S. Belongie, "Behavior Recognition via Sparse Spatio-Temporal Feature," in *Proc. of the IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pp. 65–72, 2005.

[56] J. C. Niebles, H. Wang and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," in *Proc. of the British Machine Vision Conference (BMVC)*, pp. 1249–1258, 2006.

[57] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning Realistic Human Actions from Movies," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, 2008.

[58] H. Jhuang, T. Serre, L. Wolf and T. Poggio, "A Biologically Inspired System for Action Recognition," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

[59] Y. Ke, R. Sukthankar and M. Hebert, "Efficient Visual Event Detection Using Volumetric Features," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 166–173, 2005.

[60] P. Turaga, R. Chellappa, V.S. Subrahmanian, and O. Udrea, "Machine Recognition of Human Activities: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[61] C. Schüldt, I. Laptev and B. Caputo, "Recognizing Human Actions: A Local SVM Approach," in *Proc. of the 17th International Conference on Pattern Recognition (ICPR)*, pp. 32–36, 2004.

[62] S. Becker, "Learning to Categorize Objects using Temporal Coherence," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 361–368, 1992.

[63] A. Frome, Y. Singer and J. Malik, "Image Retrieval and Classification using Local Distance Functions," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 417–424, 2006.

[64] G. Csurka, C. R. Dance, L. Fan, J. Willamowski and C. Bray, "Visual Categorization with Bags of Keypoints," in *Workshop on Statistical Learning in Computer Vision, 8th European Conference on Computer Vision (ECCV)*, 2004.

[65] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, no. 2/3, pp. 107–123, 2005.

[66] S. F. Wong and R. Cipolla, "Extracting Spatiotemporal Interest Points using Global Information," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

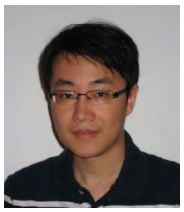
[67] A. Oikonomopoulos, I. Patras and M. Pantic, "Spatiotemporal Salient Points for Visual Recognition of Human Actions," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 36, no. 3, pp. 710–719, 2008.

[68] G. Willems, T. Tuytelaars and L. Van Gool, "An Efficient Dense and Scale-invariant Spatio-temporal Interest Point Detector," in *Proc. of 9th European Conference on Computer Vision (ECCV)*, pp. 650–663, 2008.

[69] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," *Machine Learning*, vol. 42, no. 1/2, pp. 177–196, 2001.

[70] A. Bosch, A. Zisserman and X. Munoz, "Scene Classification Via pLSA," in *Proc. of 9th European Conference on Computer Vision (ECCV)*, pp. 517–530, 2006.

[71] N. Bouguila and D. Ziou, "A Dirichlet Process Mixture of Generalized Dirichlet Distributions For Proportional Data Modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 107–122, 2010.



Wentao Fan(S'11) received the BA degree in computer science and the M.Sc degree in information systems security from Concordia University, Montreal, Quebec, Canada in 2007 and 2009, respectively. He is currently pursuing a PhD degree at Concordia University. His research interests include machine learning, data mining, and pattern recognition.

Nizar Bouguila (S'05-M'06-SM'11) received the engineer degree from the University of Tunis, Tunis, Tunisia, in 2000, and the M.Sc. and Ph.D. degrees in computer science from Sherbrooke University, Sherbrooke, QC, Canada, in 2002 and 2006, respectively. He is currently an Associate Professor with the Concordia Institute for Information Systems Engineering (CIISE) at Concordia University, Montreal, Qc, Canada. His research interests include image processing, machine learning, data mining, 3D graphics, computer vision, and pattern recognition. Prof. Bouguila received the best Ph.D Thesis Award in Engineering and Natural Sciences from Sherbrooke University in 2007. He was awarded the prestigious Prix d'excellence de l'association des doyens des études supèrieures au Québec (best Ph.D Thesis Award in Engineering and Natural Sciences in Québec), and was a runner-up for the prestigious NSERC doctoral prize. He is an IEEE senior member.

Djemel Ziou received the B.Eng. degree in Computer Science from the University of Annaba (Algeria) in 1984, and Ph.D degree in Computer Science from the Institut National Polytechnique de Lorraine (INPL), France in 1991. From 1987 to 1993 he served as lecturer in several universities in France. During the same period, he was a researcher in the Centre de Recherche en Informatique de Nancy (CRIN) and the Institut National de Recherche en Informatique et Automatique (INRIA) in France. Presently, he is full Professor at the department of computer science, Université de Sherbrooke, QC, Canada. He is holder of the NSERC/Bell Canada Research Chair in personal imaging. He has served on numerous conference committees as member or chair. He heads the laboratory MOIVRE and the consortium CoRIMedia which he founded. His research interests include image processing, information retrieval, computer vision and pattern recognition.