SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2016, Vol. 59 012201:1–012201:14
doi: 10.1007/s11432-015-5429-0

Unsupervised learning of Dirichlet process mixture models with missing data

Xunan ZHANG[1], Shiji SONG[1*], Lei ZHU[2], Keyou YOU[1] & Cheng WU[1]

[1]*Department of Automation, Tsinghua University, Beijing* 100084*, China;*
[2]*China Ocean Mineral Resources R&D Association, Beijing* 100860*, China*

Received June 3, 2015; accepted July 2, 2015; published online December 2, 2015

Abstract This study presents a novel approach to unsupervised learning for clustering with missing data. We first extend a finite mixture model to the infinite case by considering Dirichlet process mixtures, which can automatically determine the number of mixture components or clusters. Furthermore, we view the missing features as latent variables and compute the posterior distributions using the variational Bayesian expectation maximization algorithm, which optimizes the evidence lower bound on the complete-data log marginal likelihood. We demonstrate the performance on several artificial data sets with missing values. The experimental results indicate that the proposed method outperforms some classic imputation methods. We finally present an application to seabed hydrothermal sulfide color images analysis problem.

Keywords Dirichlet processes, missing data, clustering, variational Bayesian, image analysis

Citation Zhang X N, Song S J, Zhu L, et al. Unsupervised learning of Dirichlet process mixture models with missing data. Sci China Inf Sci, 2016, 59(1): 012201, doi: 10.1007/s11432-015-5429-0

1 Introduction

The aim of unsupervised learning is to find hidden structure in unlabeled data [1,2]. Finite mixture models (FMMs) are a flexible and powerful probabilistic modeling tool for clustering data in various domains, such as image analysis, computer vision, and signal processing [3]. An important issue in FMMs is to determine the appropriate number of components. In general, methods for selecting the optimal number can be classified into deterministic and Bayesian types [4]. However, an excess of components creates an over-fitting problem, while a mixture with very few components might not be flexible enough to approximate the true underlying model [5]. Dirichlet process mixture models (DPMMs) provide a powerful nonparametric Bayesian model. They sidestep setting the correct number of mixture components, and allow the number to increase as new data arrive [6]. Recent development of approximation schemes, such as Markov chain Monte Carlo (MCMC) and variational inference (VI), has enabled the widespread use of DPMMs for clustering, model selection and density estimation [7–9].

However, in real-world scenarios, there are many cases in which the data being collected are incomplete, because some values in special dimensions are unavailable. For example, data values are not recorded

* Corresponding author (email: shijis@mail.tsinghua.edu.cn)

or observed at different stages in medical studies. In surveys and social network recommendation systems, some participants refuse to respond to particular questions [10]. In DNA analysis, gene-expression microarrays might be incomplete because of insufficient resolution, image corruption, or simply dust or scratches on the slide [11]. In sensing applications, a subset of sensors might be absent or fail to operate at certain regions [12]. MCMC sampling or variational methods cannot work directly on DPMMs, if the data collected are incomplete with values missing.

Previous studies on solving machine learning algorithms with incomplete data can be categorized into three groups. The first is listwise deletion, which simply discards samples with missing values. It is the simplest approach but only useful when the amount of missing data is small [13]. The second method is imputation, which substitutes missing values with statistically plausible values. Some classic imputation methods include mean imputation (MI), K nearest neighbor (KNN), expectation maximization (EM) and multiple imputation [14]. However, in clustering problem, these methods take little consideration of the relation and uncertainty between samples. A poor imputation strategy can render clustering algorithms ineffective. The last method addresses missing values during the model-learning procedures without a previous estimation. For example, Chechik and Heitz represent an improved support vector machine (SVM) by re-scaling the margin according to the observed features for each instance [15], and Sanja and Danijel propose an approach that combines reconstructive and discriminative subspace methods for robust classification and regression by subsampling [16]. Both these examples illustrate that handling missing data within the algorithm are more effective than simple imputation methods. The performance of different methods has a bearing on different missing mechanisms [14]. Types of missing data can be classified into three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Most researchers assume the data missing mechanism satisfies MAR or MCAR, which is a more realistic and practical model.

In this study, first FMMs are extended to the infinite model using a stick-breaking construction. Next, we use the deterministic VI algorithm to solve DPMMs with missing data. This can yield a robust and stable estimate with fast convergence. We partition each data point into its observed and missing parts, and view the missing parts as latent variables. The posterior parameters are estimated using the variational Bayesian EM (VBEM) algorithm, which optimizes the evidence lower bound (ELBO) iteratively through a fully-factorized variational distribution. Therefore, the new proposed unsupervised learning method can automatically determine the number of mixture components or clusters. Furthermore, the VI framework can effectively solve the missing data problem without advance imputation. It can also avoid over-fitting by compromising between generalization ability and model complexity [17].

The remainder of this paper is organized as follows. In Section 2, we introduce DPMMs and the VI algorithm. In Section 3, we present the VBEM algorithm for solving DPMMs with missing data. Section 4 presents the results of experimental comparisons. Futhermore, we apply the proposed approach to classify seabed hydrothermal sulfide color images. Conclusions are drawn in Section 5.

2 Learning DPMMs

Let $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ be independent D-dimensional observations arising from a mixture of distributions $F(\theta_k)$, where θ_k is the model parameter independently drawn from some distribution G. In the FMMs, there are a total of K clusters, and $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ is the mixing proportion. Figure 1(a) gives the directed graphical representation of FMMs. However, before using this model, we have to set the cluster number K in advance.

2.1 DPMMs

Recent development of the Dirichlet process (DP) as a nonparametric prior distribution on the components of a mixture model enables automatic identification of the cluster numbers. DPMMs can be derived as the limit of a sequence of FMMs, where the number of mixture components is taken to infinity [18]. In DPMMs, the mixture proportion can be represented by a stick-breaking construction [19]. The graphical

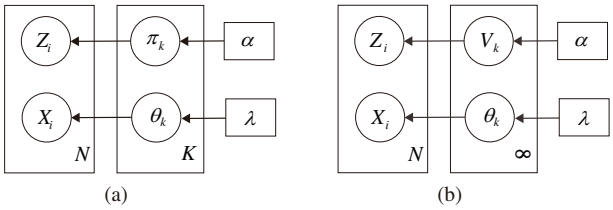

Figure 1 Directed graphical representations of (a) finite mixture models and (b) Dirichlet process mixture models.

representation for construction of this model is shown in Figure 1(b). The conditional distributions of the DPMMs are listed as follows:

$$\begin{gathered} v_k|\alpha \sim \text{Beta}(1,\alpha), \\ z_i|\pi(\boldsymbol{v}) \sim \text{Mult}\left(\pi(\boldsymbol{v})\right), \\ \theta_k|\lambda \sim G_0(\lambda), \\ \boldsymbol{x}_i|z_i, \{\theta_k\}_{k=1}^{\infty} \sim F(\theta_{z_i}). \end{gathered}$$

We use z_i as an indicator variable to specify the cluster associated with $\boldsymbol{x}_i$ and the mixture weight as Dirichlet prior distribution π, which is given by successively breaking a unit length stick into an infinite number of pieces. The size of each successive piece v_i is given by an independent draw from a Beta$(1,\alpha)$ distribution. The mixing proportion π_k then satisfies the following expression:

$$\pi_k(\boldsymbol{v}) = v_k \prod_{i=1}^{k-1}(1-v_i) \in [0,1], \quad \sum_{k=1}^{\infty}\pi_k(\boldsymbol{v}) = 1. \tag{1}$$

2.2 VI for DPMMs

VI provides an alternative, and deterministic method for approximating the intractable posteriors in DPMMs [20]. It provides a lower bound on the log marginal likelihood using a fully factorized variational distribution q. Given a model with observed variable $\boldsymbol{x}$, latent variable $\boldsymbol{z}$, model parameters $\boldsymbol{v}$, θ and hyperparameters $\phi = \{\alpha, \lambda\}$, the optimal q maximizes the ELBO $\mathcal{L}$ as follows

$$\log p(\boldsymbol{x}|\phi) \geqslant \mathcal{L}(q) = \text{E}_q\left[\log p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{z}, \theta|\phi) - \log q(\boldsymbol{v}, \boldsymbol{z}, \theta)\right]. \tag{2}$$

To handle the infinite set of components available under the DP prior tractably, Blei proposed a truncated stick-breaking representations [7] by fixing a value K and letting $q(v_K = 1) = 1$, which implies $q(z_n = k) = 0$ for $k > K$. Therefore, inference for the variational parameters can focus on a finite set of K components. When a fully-factorized distribution with individual factors is considered, the variational distributions $q(\boldsymbol{v}, \theta, \boldsymbol{z})$ can be written as follows:

$$q(\boldsymbol{v}, \theta, \boldsymbol{z}) = \prod_{k=1}^{K-1} q_{\gamma_k}(v_k) \prod_{k=1}^{K} q_{\tau_k}(\theta_k) \prod_{n=1}^{N} q_{\omega_n}(z_n). \tag{3}$$

The free variational parameters are $\boldsymbol{\nu} = \{\gamma_1, ..., \gamma_{K-1}, \tau_1, ..., \tau_K, \omega_1, ..., \omega_N\}$. For DPMMs of exponential family distributions, Blei gives an explicit coordinate ascent algorithm [7], which can optimize the bound in Eq. (2) with respect to the variational parameters. The algorithms can be described in terms of two updates, a variational Bayesian expectation (VBE) step for local parameters (assignments of data to components z_n) and a variational Bayesian maximization (VBM) step for global parameters (stick-breaking proportions v_k and data-generating parameters θ_k) [21]. However, when the data used for clustering are not complete, the VI cannot work on DPMMs directly. In our study, we propose an effective method for solving the missing data problem.

3 Learning DPMMs with missing data

In standard DPMMs, it is assumed that all components of the feature vectors are available without missing data. However, in real-world scenarios, there are many cases in which the data being collected

are incomplete, where some values in special dimensions are unavailable. In general, we can delete incomplete samples or impute the missing values before using the model for clustering. In contrast to supervised learning, the listwise deletion method cannot be used directly on the clustering problem, because samples with missing values do not participate in the clustering process, and no label will be assigned to them. Imputation methods take little consideration of the relation and uncertainty between samples. For example, MI method might classify all the missing samples to the same cluster.

To resolve these issues, we propose a novel learning procedure that can handle missing data within the algorithm. In this study, we restrict the observable data drawn form Gaussian distributions with a possibly infinite number of components. We partition each feature vector $\boldsymbol{x}_i$ into its observed and missing parts. Efficient inference is implemented using the VBEM to solve missing data clustering problem, where we assume the data are missing under the MAR or MCAR assumption.

3.1 Notation and category of missing data

Referring to the standard representation for missing data given by Little and Rubin [14], we introduce a response indicator variable $\boldsymbol{r}_i$ for each sample $\boldsymbol{x}_i$. If x_{ij} is observed $r_{ij} = 1$; otherwise $r_{ij} = 0$ means x_{ij} is missing. Let $\boldsymbol{x}_i$ be partitioned into two components $\boldsymbol{x}_i = [\boldsymbol{x}_i^{o_i}; \boldsymbol{x}_i^{m_i}]$, where $\boldsymbol{x}_i^{o_i}(D^{o_i} \times 1)$ and $\boldsymbol{x}_i^{m_i}(D^{m_i} \times 1)$ are the observed and missing components of $\boldsymbol{x}_i$, respectively. We use $z_i = k$ to indicate $\boldsymbol{x}_i$ is generated by the kth component in a Gaussian mixture model, we can write the equation as follows:

$$p(\boldsymbol{x}_i^{o_i}, \boldsymbol{x}_i^{m_i}) = (2\pi)^{-D/2}|\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{ -\frac{1}{2} \begin{bmatrix} \boldsymbol{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i} \\ \boldsymbol{x}_i^{m_i} - \boldsymbol{\mu}_k^{m_i} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \boldsymbol{\Sigma}_k^{o_i o_i} & \boldsymbol{\Sigma}_k^{o_i m_i} \\ \boldsymbol{\Sigma}_k^{m_i o_i} & \boldsymbol{\Sigma}_k^{m_i m_i} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i} \\ \boldsymbol{x}_i^{m_i} - \boldsymbol{\mu}_k^{m_i} \end{bmatrix} \right\}.$$

Notation of the form $\boldsymbol{\Sigma}_k^{o_i m_i}$ denotes a sub-matrix of $\boldsymbol{\Sigma}$ obtained by selecting the rows corresponding to the observed dimensions and the columns corresponding to the missing dimensions of $\boldsymbol{x}_i$, respectively. For convenience of calculations, we introduce two types of binary indicator matrices, $[O_i]_{D_i^o \times D}$ and $[M_i]_{D_i^m \times D}$, satisfying $\boldsymbol{x}_i^{o_i} = O_i \boldsymbol{x}_i$ and $\boldsymbol{x}_i^{m_i} = M_i \boldsymbol{x}_i$. The matrices O_i and M_i are $D_i^{o_i} \times D$ and $D_i^{m_i} \times D$ matrices extracted from a D-dimensional identity matrix I_D. We provide two propositions for convenience of calculation during the VBEM procedure.

Lemma 1. Assume that $\boldsymbol{x}_i$ is partitioned into two components $\boldsymbol{x}_i = [\boldsymbol{x}_i^{o_i}; \boldsymbol{x}_i^{m_i}]$, given indicator matrices O_i and M_i, we have

$$\boldsymbol{x}_i = \begin{cases} \boldsymbol{x}_i^{o_i}, & D_i^{o_i} = D, \\ O_i^{\mathrm{T}} \boldsymbol{x}_i^{o_i} + M_i^{\mathrm{T}} \boldsymbol{x}_i^{m_i}, & 1 \leqslant D_i^{o_i} < D, \end{cases} \quad \text{and} \quad O_i^{\mathrm{T}} O_i + M_i^{\mathrm{T}} M_i = I_D.$$

Proof. The proof is straightforward and hence is omitted.

Lemma 2. Assume that $\boldsymbol{x}_i \sim \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where the observed and missing components are $\boldsymbol{x}_i^{o_i}$ and $\boldsymbol{x}_i^{m_i}$. The marginal distribution of observed variable $\boldsymbol{x}_i^{o_i} \sim \sum_{k=1}^K w_k \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{\mu}_k^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i})$, where

$$\mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{\mu}_k^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i}) = (2\pi)^{-D_i^{o_i}/2} |\boldsymbol{\Sigma}_k^{o_i o_i}|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{\Delta}_k^{o_i} \right),$$

$$\boldsymbol{\mu}_k^{o_i} = O_i \boldsymbol{\mu}_k,\ \boldsymbol{\Sigma}_k^{o_i o_i} = O_i \boldsymbol{\Sigma}_k O_i^{\mathrm{T}},\ \boldsymbol{\Delta}_k^{o_i} = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{W}_k^{o_i o_i} (\boldsymbol{x}_i - \boldsymbol{\mu}_k),\ \boldsymbol{W}_k^{o_i o_i} = O_i^{\mathrm{T}} (O_i \boldsymbol{\Sigma}_k O_i^{\mathrm{T}})^{-1} O_i.$$

Given $\boldsymbol{x}_i^{o_i}$, the conditional distribution for $\boldsymbol{x}_i^{m_i} | \boldsymbol{x}_i^{o_i} \sim \sum_{k=1}^K \tilde{w}_k \mathcal{N}(\boldsymbol{x}_i^{m_i} | \boldsymbol{\mu}_k^{m_i|o_i}, \boldsymbol{\Sigma}_k^{m_i|o_i})$, where

$$\mathcal{N}(\boldsymbol{x}_i^{m_i} | \boldsymbol{\mu}_k^{m_i|o_i}, \boldsymbol{\Sigma}_k^{m_i|o_i}) = (2\pi)^{-D_i^{m_i}/2} |\boldsymbol{\Sigma}_k^{m_i|o_i}|^{-1/2} \exp\left(-\frac{1}{2} \boldsymbol{\Delta}_k^{m_i|o_i} \right),$$

$$\tilde{w}_k = w_k \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{\mu}_k^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i}) \Big/ \sum_{l=1}^K w_l \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{\mu}_l^{o_i}, \boldsymbol{\Sigma}_l^{o_i o_i}),$$

$$\boldsymbol{\mu}_k^{m_i|o_i} = M_i(\boldsymbol{\mu}_k + \boldsymbol{\Sigma}_k \boldsymbol{W}_k^{o_i o_i} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)),\ \boldsymbol{E}_{ik} = M_i(I_D - \boldsymbol{\Sigma}_k \boldsymbol{W}_k^{o_i o_i}),\ \boldsymbol{\Sigma}_k^{m_i|o_i} = \boldsymbol{\Sigma}_k M_i^{\mathrm{T}},$$

$$\boldsymbol{\Delta}_k^{m_i|o_i} = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{W}_k^{m_i|o_i} (\boldsymbol{x}_i - \boldsymbol{\mu}_k),\ \boldsymbol{W}_k^{m_i|o_i} = \boldsymbol{E}_{ik}^{\mathrm{T}} (\boldsymbol{E}_{ik} \boldsymbol{\Sigma}_k M_i^{\mathrm{T}})^{-1} \boldsymbol{E}_{ik}.$$

Proof. The derivation is similar to those of the conditional and marginal Gaussian distributions. The details are provided in [22, 23].

We use ξ to denote parameters characterizing the distribution of the missing indicator matrix $\mathcal{R}$. There are three different missing mechanisms. If missingness does not depend on the values of the data X, i.e., if $f(\mathcal{R}|X, \xi) = f(\mathcal{R}|\xi)$, then the data are MCAR. MAR is a less restrictive assumption than MCAR to the effect that missingness depends only on the observed part of X, satisfying $f(\mathcal{R}|X, \xi) = f(\mathcal{R}|X^o, \xi)$. If the distribution of $\mathcal{R}$ is non-random and depends on the missing values in the matrix X, the missing-data mechanism is NMAR, in which case one must explicitly specify a model for the missingness variable $\mathcal{R}$. However, this is difficult to achieve in most cases, and we would rather assume MCAR or MAR. It has been demonstrated that although these two assumptions might be invalid, they do not lead to substantial bias in the inference result [24]. The joint distribution of observed features and the missingness variable can be obtained by integrating out the missing features $\boldsymbol{x}^m$,

$$p(\boldsymbol{x}^o, \boldsymbol{r}|\theta, \xi) = \int p(\mathbf{x}|\theta) p(\boldsymbol{r}|\boldsymbol{x}, \xi) \mathrm{d}\boldsymbol{x}^m. \tag{4}$$

Under the MCAR or MAR assumption $p(\boldsymbol{r}|\boldsymbol{x}, \xi) = p(\boldsymbol{r}|\boldsymbol{x}^o, \xi)$, and the joint distribution reduces to

$$p(\boldsymbol{x}^o, \boldsymbol{r}|\theta, \xi) = p(\boldsymbol{r}|\boldsymbol{x}^o, \xi) \int p(\boldsymbol{x}|\theta) d\boldsymbol{x}^m = p(\boldsymbol{r}|\boldsymbol{x}^o, \xi) p(\boldsymbol{x}^o|\theta)). \tag{5}$$

The posterior distribution of parameters is as follows:

$$p(\theta, \xi|\boldsymbol{x}^o, \boldsymbol{r}) \propto p(\boldsymbol{r}|\boldsymbol{x}^o, \xi) p(\xi) p(\boldsymbol{x}^o|\theta) p(\theta). \tag{6}$$

3.2 VBEM for DPMMs with missing data

When the dataset is incomplete with missing values for some features, the missingness variable $X^m = \{\boldsymbol{x}_i^{m_i}\}_{i=1}^N$ is added as a new parameter in the variational distribution [12]. We use the same truncated stick-breaking representations by fixing a value K and letting $q(v_K = 1) = 1$, which implies $q(z_n = k) = 0$ for $k > K$. Therefore, based on the factorized approximation, the variational family can factorize to be independent variables as follows:

$$q(\boldsymbol{v}, \theta, \boldsymbol{z}, X^m) = \prod_{k=1}^{K-1} q_{\gamma_k}(v_k) \prod_{k=1}^{K} q_{\tau_k}(\theta_k) \prod_{i=1}^{N} q_{\phi_i}(z_i, \boldsymbol{x}_i^{m_i}), \tag{7}$$

where $\theta_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$. We let variable $\Psi = \{v, \theta\}$ and $\Phi = \{Z, X^m\}$, and use Jensens inequality on the log probability of the observed data based on the factorized approximation,

$$\begin{aligned} \log\ p(\boldsymbol{x}_i^{o_i}) &= \log \int p(\boldsymbol{x}_i^{o_i}, \Phi, \Psi) \mathrm{d}\Phi \mathrm{d}\Psi = \log \int q(\phi, \theta) \frac{p(x_i^{o_i}, \Phi, \Psi)}{q(\Phi, \Psi)} \mathrm{d}\Phi \mathrm{d}\Psi \\ &\geqslant \int q(\Phi, \Theta) \log \frac{p(x_i^{o_i}, \Phi, \Psi)}{q(\Phi, \Psi)} \mathrm{d}\Phi \mathrm{d}\Psi \approx \int q(\Phi) q(\Psi) \log \frac{p(x_i^{o_i}, \Phi, \Psi)}{q(\Phi, \Psi)} \mathrm{d}\Phi \mathrm{d}\Psi. \end{aligned} \tag{8}$$

The optimal q maximizes the evidence lower bound objective $\mathcal{L}$ on the observed data,

$$\log p(X^o|\phi) \geqslant \mathcal{L}(q) = \mathrm{E}_q \left[\log p(\Phi, \Psi, X^o) - \log q(\Phi, \Psi)\right]. \tag{9}$$

Next, we provide the VBEM algorithm to maximize the ELBO with respect to the variational distributions $q(\Phi)$ and $q(\Psi)$ with missing data. The resulting VBE and VBM steps are:

$$\text{VBE}: \quad q(\Phi) \propto \exp\left\{\int \log p(x_i^{o_i}, \Phi|\Psi) q(\Psi) \mathrm{d}\Psi\right\}, \tag{10}$$

$$\text{VBM}: \quad q(\Psi) \propto p(\Psi) \exp\left\{\int \log p(x_i^{o_i}, \Phi|\Psi) q(\Phi) \mathrm{d}\Phi\right\}. \tag{11}$$

As for v_k, it is sampled from $\mathrm{Beta}(\gamma_{k1}, \gamma_{k2})$. We choose an independent Gaussian-Wishart prior distribution to govern the mean and precision of each Gaussian component, given by the following equation:

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k | \boldsymbol{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k | \boldsymbol{S}_k, \kappa_0).$$

Proposition 1. In each iteration of the VBE step, we update the quantity for $q(\boldsymbol{x}_i^{m_i}, z_i = k)$ as follows:

$$q(\boldsymbol{x}_i^{m_i}, z_i = k) = \tilde{\delta}_k^i \mathcal{N}(\boldsymbol{x}_i^{m_i} | \boldsymbol{m}_k^{m_i|o_i}, \boldsymbol{S}_k^{m_i|o_i}), \tag{12}$$

where we defined

$$\tilde{\delta}_k^i = \frac{A_k \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{m}_k^{o_i}, \kappa_k^{-1} (\boldsymbol{S}_k^{-1})^{o_i o_i}) \mathcal{N}(\boldsymbol{x}_i^{m_i} | \boldsymbol{m}_k^{m_i|o_i}, \boldsymbol{S}_k^{m_i|o_i})}{\sum_{l=1}^{K} A_l \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{m}_l^{o_i}, \kappa_l^{-1} (\boldsymbol{S}_l^{-1})^{o_i o_i})}, \tag{13}$$

$$A_k = \exp\left\{ \mathrm{E}_q[\log w_k] + \frac{D}{2}\log 2 - \frac{1}{2}\log \kappa_k + \frac{1}{2}\sum_{d=1}^{D} \psi\left(\frac{\kappa_k + 1 - d}{2}\right) - \mathbf{tr}(\beta_k \mathbf{I_D}) \right\}, \tag{14}$$

$$\mathrm{E}_q[\log w_k(\boldsymbol{v})] = \psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2}) + \sum_{j=1}^{k-1} \left(\psi(\gamma_{j,2}) - \psi(\gamma_{j,1} + \gamma_{j,2})\right), \tag{15}$$

$$\boldsymbol{m}_k^{m_i|o_i} = \boldsymbol{m}_k^{m_i} + (\boldsymbol{S}_k^{-1})^{m_i o_i} ((\boldsymbol{S}_k^{-1})^{o_i o_i})^{-1} (\boldsymbol{x}_i^{o_i} - \boldsymbol{m}_k^{o_i}), \tag{16}$$

$$\boldsymbol{S}_k^{m_i|o_i} = -\kappa_k^{-1} \left((\boldsymbol{S}_k^{-1})^{m_i o_i} ((\boldsymbol{S}_k^{-1})^{o_i o_i})^{-1} ((\boldsymbol{S}_k^{-1})^{m_i o_i})^{\mathrm{T}} \right) + \kappa_k^{-1} (\boldsymbol{S}_k^{-1})^{m_i m_i}. \tag{17}$$

Proof. With the truncation $q(z_i > K) = 0$, the inference can be made tractable for infinite components. The posterior $p(z_i | \boldsymbol{x}_i, \theta, \boldsymbol{v})$ over assignments for each item i is approximated by a discrete multinomial distribution over K components. First, we provide an independent variational $q_{\gamma_k}(v_k)$ to each fraction v_k with the updating equation as follows:

$$q_{\gamma_k}(v_k) = \mathrm{Beta}(\gamma_{k,1}, \gamma_{k,2}),\ \gamma_{k,1} = 1 + N_k,\ \gamma_{k,2} = \alpha + \sum_{l=k+1}^{K} N_l. \tag{18}$$

Given $\gamma_{k,1}, \gamma_{k,2}$ for all components, the expected log mixture weights are

$$\mathrm{E}_q[\log v_k] = \psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2}),\ \mathrm{E}_q[\log(1 - v_k)] = \psi(\gamma_{k,2}) - \psi(\gamma_{k,1} + \gamma_{k,2}), \tag{19}$$

$$\mathrm{E}_q[\log w_k(\boldsymbol{v})] = \mathrm{E}_q[\log v_k] + \sum_{l=1}^{k-1} \mathrm{E}_q[\log(1 - v_l)], \tag{20}$$

where $\psi(\cdot)$ is the digamma function defined as $\psi(z) = \frac{\mathrm{d}}{\mathrm{d}z}\log\Gamma(z)$. The VBE step is then written as follows:

$$\begin{aligned} q(\boldsymbol{x}_i^{m_i}, z_i = k) &\propto \exp\left\{ \mathrm{E}_q\left[\log p(\boldsymbol{x}_i^{o_i}, \boldsymbol{x}_i^{m_i}, z_i = k | \Psi)\right] \right\} \\ &= \exp\left\{ \mathrm{E}_q[\log v_k] + \sum_{j=1}^{k-1} \mathrm{E}_q[\log(1 - v_j)] + \frac{1}{2}\mathrm{E}_q[\log|\boldsymbol{\Lambda}_k|] - \frac{D}{2}\log 2\pi \right. \\ &\quad \left. -\frac{1}{2}\mathbf{tr}\left(\mathrm{E}_q[\boldsymbol{\Lambda}_k \mathrm{E}_q[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)]]\right) \right\}, \end{aligned} \tag{21}$$

where

$$\mathrm{E}_q[\log \boldsymbol{\Lambda}_k] = \sum_{d=1}^{D} \psi\left(\frac{\kappa_k + 1 - d}{2}\right) + D\log 2 + \log|\boldsymbol{S}_k|, \tag{22}$$

$$\mathrm{E}_q[(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Lambda}_k (\boldsymbol{x}_i - \boldsymbol{\mu}_k)] = \mathbf{tr}(\beta_k \boldsymbol{I}_D) + (\boldsymbol{x}_i - \boldsymbol{m}_k)^{\mathrm{T}} \kappa_k \boldsymbol{S}_k (\boldsymbol{x}_i - \boldsymbol{m}_k). \tag{23}$$

We define A_k as equation (14), and upon normalizing, the updated quantity is

$$\begin{aligned} q(\boldsymbol{x}_i^{m_i}, z_i = k) &= \frac{A_k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{m}_k, \kappa_k^{-1} \boldsymbol{S}_k^{-1})}{\sum_{l=1}^{K} A_l \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{m}_l^{o_i}, \kappa_l^{-1} (\boldsymbol{S}_l^{-1})^{o_i o_i})} \\ &= \frac{A_k \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{m}_k^{o_i}, \kappa_k^{-1} (\boldsymbol{S}_k^{-1})^{o_i o_i}) \mathcal{N}(\boldsymbol{x}_i^{m_i} | \boldsymbol{m}_k^{m_i|o_i}, \boldsymbol{S}_k^{m_i|o_i})}{\sum_{l=1}^{K} A_l \mathcal{N}(\boldsymbol{x}_i^{o_i} | \boldsymbol{m}_l^{o_i}, \kappa_l^{-1} (\boldsymbol{S}_l^{-1})^{o_i, o_i})} \\ &= \tilde{\delta}_k^i \mathcal{N}(\boldsymbol{x}_i^{m_i} | \boldsymbol{m}_k^{m_i|o_i}, \boldsymbol{S}_k^{m_i|o_i}). \end{aligned} \tag{24}$$

Referring to Lemma 2, we can give the expressions for $\boldsymbol{m}_k^{m_i|o_i}$ and $\boldsymbol{S}_k^{m_i|o_i}$ as shown in Eqs. (16) and (17). We can enhance computational efficiency by introducing binary indicator matrices O_i and M_i for each sample $\boldsymbol{x}_i$. The computations for $(\boldsymbol{S}_k^{-1})^{o_io_i}$, $(\boldsymbol{S}_k^{-1})^{m_io_i}$ and $(\boldsymbol{S}_k^{-1})^{m_im_i}$ are given as follows:

$$(\boldsymbol{S}_k^{-1})^{o_io_i} = O_i\boldsymbol{S}_k^{-1}O_i^{\mathrm{T}},\ (\boldsymbol{S}_k^{-1})^{m_io_i} = M_i\boldsymbol{S}_k^{-1}O_i^{\mathrm{T}} \text{ and } (\boldsymbol{S}_k^{-1})^{m_im_i} = M_i\boldsymbol{S}_k^{-1}M_i^{\mathrm{T}}. \tag{25}$$

For each incomplete sample, we can let $\boldsymbol{m}_k^{m_i|o_i}$ replace the missing values. Therefore, the estimated parameters $\tilde{\boldsymbol{x}}_i^k$, $\bar{\boldsymbol{x}}^k$ and $\tilde{\boldsymbol{S}}_k^i$ are:

$$\tilde{\boldsymbol{x}}_i^k = \boldsymbol{m}_k + \boldsymbol{S}_k^{-1}(\boldsymbol{W}_k^{-1})^{o_io_i}(\boldsymbol{x}_i - \boldsymbol{m}_k), \tag{26}$$

$$\bar{\boldsymbol{x}}^k = \frac{(\beta_k^0)^{-1}\boldsymbol{m}_0 + \sum_{i=1}^N \tilde{\delta}_k^i\tilde{\boldsymbol{x}}_i^k}{(\beta_k^0)^{-1} + \sum_{i=1}^N \tilde{\delta}_k^i}, \tag{27}$$

$$\tilde{\boldsymbol{S}}_k^i = \kappa_k^{-1}\left((\boldsymbol{x}_i^k - \boldsymbol{m}_k)(\boldsymbol{x}_i^k - \boldsymbol{m}_k)^{\mathrm{T}} + (\boldsymbol{I}_D - \boldsymbol{S}_k^{-1}(\boldsymbol{W}_k^{-1})^{o_io_i})\boldsymbol{S}_k^{-1}\right), \tag{28}$$

$$(\boldsymbol{W}_k^{-1})^{o_io_i} = O_i^{\mathrm{T}}(O_i\boldsymbol{S}_k^{-1}O_i^{\mathrm{T}})^{-1}O_i. \tag{29}$$

With auxiliary matric O_i previously defined, it is not necessary to consider the missing values of $\boldsymbol{x}_i^k$ in these equations.

Proposition 2. At each iteration of the VBM step, the posterior parameters are updated as follows, based on the parameters estimated in the VBE step.

$$\hat{\gamma}_{k,1} = 1 + \sum_{i=1}^N \tilde{\delta}_k^i,\ \hat{\gamma}_{k,2} = \alpha + \sum_{i=1}^N\sum_{j=k+1}^K \tilde{\delta}_j^i, \tag{30}$$

$$\hat{\boldsymbol{m}}_k = \frac{(\beta_k^0)^{-1}\boldsymbol{m}_0 + \sum_{i=1}^N \tilde{\delta}_k^i\tilde{\boldsymbol{x}}_i^k}{(\beta_k^0)^{-1} + \sum_{i=1}^N \tilde{\delta}_k^i}, \tag{31}$$

$$\hat{\beta}_k^{-1} = (\beta_k^0)^{-1} + \sum_{i=1}^N \tilde{\delta}_k^i, \tag{32}$$

$$\hat{\boldsymbol{S}}_k^{-1} = (\boldsymbol{S}_k^0)^{-1} + \sum_{i=1}^N \tilde{\delta}_k^i\tilde{\boldsymbol{S}}_k^i + (\beta_k^0)^{-1}\boldsymbol{m}_0(\boldsymbol{m}_0)^{\mathrm{T}} + \sum_{i=1}^N \tilde{\delta}_k^i\tilde{\boldsymbol{x}}_i^k(\tilde{\boldsymbol{x}}_i^k)^{\mathrm{T}} - \left((\beta_k^0)^{-1} + \sum_{i=1}^N \tilde{\delta}_k^i\right)\bar{\boldsymbol{x}}_i^k(\bar{\boldsymbol{x}}_i^k)^{\mathrm{T}}, \tag{33}$$

$$\hat{\kappa}_k = \kappa_k^0 + \sum_{i=1}^N \tilde{\delta}_k^i. \tag{34}$$

Proof. With the estimated parameters $\tilde{\boldsymbol{x}}_i^k$, $\bar{\boldsymbol{x}}^k$, and $\tilde{\boldsymbol{S}}_k^i$ in the VBE step, for each sample $\boldsymbol{x}_i$, the expression for equation Eq. (11) is given as follows:

$$\begin{aligned} q(\Psi) &\propto p(\Psi)\exp\left(\mathrm{E}_\Phi[\log p(x_i^{o_i}, x_i^{m_i}, z_i = k|\Psi)]\right) \\ &= p(\Psi)\prod_{k=1}^K\left[v_k\prod_{j=1}^{k-1}(1-v_j)\right]^{\tilde{\delta}_k^i}|\Lambda_k|^{\tilde{\delta}_k^i/2}\exp\left\{-\frac{1}{2}\mathbf{tr}(\Lambda_k\tilde{\delta}_k^i\tilde{\boldsymbol{S}}_k^i)\right\} \\ &\quad\times\exp\left\{-\frac{1}{2}\tilde{\delta}_k^i(\tilde{\boldsymbol{x}}_i^k - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\tilde{\boldsymbol{x}}_i^k - \boldsymbol{\mu}_k)\right\}. \end{aligned} \tag{35}$$

When considering the entire data set, the variational distribution is as

$$q(\Psi) \propto p(\Psi)\prod_{i=1}^N \exp\left(\mathrm{E}_\Phi[\log p(x_i^{o_i}, x_i^{m_i}, z_i = k|\Psi)]\right). \tag{36}$$

The constants can be neglected when calculating the updated posterior parameters. After calculating the derivative of the variational function for each parameter, we can obtain the updated parameters shown as Eqs. (30)–(34).

Thus, the optimization of the variational posterior distributions involves repeated iterations of two stages, the VBE and VBM steps. In the variational equivalent of the expectation step, we use the current distributions over the model parameters to evaluate the expectation $\mathrm{E}(z_{ik})$, and then in the subsequent variational equivalent of the maximization step, we use the responsibilities estimated in the VBE step to re-compute the posterior parameters. Next, we show the complexity analysis of the VBEM inference method. It has a complexity of $O(l(DK + KND^3 + K^2N))$, where l denotes the number of iterations, N is the number of samples, and D is the dimensionality. For each VBE step, the complexity is $O(l(K^2 + DK + KND^3))$, where K^2 is obtained from the computation of $\mathrm{E}_q[\log w_k]$, DK is from the parameter A_k and KND^3 accounts for the aided parameters $(\boldsymbol{S}_k^{-1})^{o_i o_i}$, $(\boldsymbol{S}_k^{-1})^{m_i o_i}$ and $(\boldsymbol{S}_k^{-1})^{m_i m_i}$. For each VBM step, the complexity is $O(l(K^2N + KND^2))$, where K^2N is for the update parameter $\hat{\gamma}_{k,2}$ and K^2N for the posterior estimation of $\hat{\boldsymbol{S}}_k^{-1}$. Typically, the algorithm is very efficient when the dimensionality D and the sample numbers N are not excessively large. Practical applications of variational methods must address initialization of the variational distribution, because poor choices will cause it to fall into local maxima. We use the method shown in Blei and Jordan [7] by incrementally updating the parameters according to a random permutation of the data points. Furthermore, missing data can be imputed by using expectation conditional maximization (ECM) algorithm [25] before parameters initialization.

4 Experiments

Artificial datasets were used in experiments to illustrate the performance of the proposed method in solving unsurprised learning of DPMMs with missing data. All these data are complete without missing values. We generate the missing values randomly under the MCAR mechanism. The missing rate is defined as $M_{\mathrm{r}} = N_{\mathrm{m}}/N \times 100\%$, where N_{m} is the number of missing samples. For purpose of comparison, we use listwise deletion-nearest neighbour (LDNN) and some imputation methods such as MI, KNN, and model-based EM methods. All these methods fill in the missing values before using DPMMs for clustering. First, we give a brief description of these imputation methods.

(a) LDNN: Feature vectors with missing values are simply discarded. However, this can not work directly on the clustering case. The purpose of clustering is to classify each unlabeled sample into a category. If we delete the samples with missing values, then only the complete data X^o are used to construct the model and no category will be assigned to the samples with missing data. To avoid this situation, we first provide the clustering results on complete data set X^o, and for each incomplete sample, we find its nearest neighbor in X^o using the observed dimensions and assign its label to this incomplete sample.

(b) MI: For each dimension $d \in \{1, ..., D\}$, we use the observed data in d to calculate the mean value M_d. All the missing data in the d-dimension are replaced by M_d.

(c) KNN: For each sample $\boldsymbol{x}_i$, we calculate the Euclidean distance between sample $\boldsymbol{x}_i$ and the remaining samples on the same observed dimensions. We then find the K nearest neighbors whose values exist for feature $d \in D^{m_i}$ and use the mean value to fill in the missing value for x_{id}.

(d) EM: We assume the data follow a multi-Gaussian distribution. EM capitalizes on the interdependence of missing data X_m and parameters Θ. They contain mutually relevant information to each other. The algorithm converges by iterating the expectation and maximization steps.

We propose two types of evaluation indices, clustering accuracy I_{a} and mean absolute error (MAE) I_{m}, to evaluate the performance of these methods. Given a data set $X = \{\boldsymbol{x}_i\}_{i=1}^N$, we assume the true cluster number is K_{T}, and the mean value for each cluster is $\boldsymbol{m}_i$. The definitions of I_{a} and I_{m} are as follows:

$$I_{\mathrm{a}} = \frac{\sum_{i=1}^{K_{\mathrm{T}}} N_i}{N} \quad \text{and} \quad I_{\mathrm{m}} = \frac{1}{K_{\mathrm{T}}} \sum_{i=1}^{K_{\mathrm{T}}} \sqrt{||\tilde{\boldsymbol{m}}_i - \boldsymbol{m}_i||}, \tag{37}$$

Table 1 Clustering accuracy (%) and average time (s) for Gaussian dataset with different missing rates

Methods	10%	20%	30%	40%	50%	60%	70%	time
LDNN-VI	96.31	94.27	92.13	90.56	87.49	85.38	81.46	12.52
MI-VI	94.25	84.64	77.48	71.33	65.67	58.78	33.27	1.82
KNN-VI	95.32	92.18	89.72	82.52	78.56	76.67	71.36	14.38
EM-VI	95.24	85.21	82.44	74.35	69.37	61.45	57.68	2.78
VBEM-VI	96.28	94.17	92.62	92.39	88.87	86.72	84.56	200.42

Table 2 The MAE for Gaussian dataset with different missing rates

Methods	10%	20%	30%	40%	50%	60%	70%
LDNN-VI	0.2038	0.2045	0.2178	0.2364	0.2912	0.3578	0.4872
MI-VI	0.1940	0.2672	0.3048	0.3312	0.3693	0.6881	0.9135
KNN-VI	0.2412	0.2746	0.2945	0.3122	0.3687	0.5236	0.6423
EM-VI	0.2056	0.2523	0.2956	0.3317	0.4238	0.5645	0.6928
VB-EM	0.2012	0.2245	0.2393	0.2782	0.3705	0.3948	0.4013

where N_i is the accurate clustering number and $\tilde{\boldsymbol{m}}_i$ is the estimated mean in cluster i.

4.1 Artificial datasets

We first demonstrate the proposed approach on three synthetic datasets. In the first example, we use 900 samples from a three-component bivariate mixture with proportions $w_1 = w_2 = w_3 = 1/3$, mean vectors at $[0, -2]^{\mathrm{T}}, [0, 0]^{\mathrm{T}}, [0, 2]^{\mathrm{T}}$, and equal covariance matrices diag$\{2, 0.2\}$ [26]. We generate missing values randomly on the original data set with different rates from 10% to 70%. For each missing rate, we repeat the experiment for 100 times. The average I_{a} and run time for these different methods are shown in Table 1, and I_{m} is shown in Table 2. For KNN imputation, we set the number of neighbors as five. Figure 2(a), in which a black "$*$" added to the original signs "$\cdot$" represents the missing samples, shows an example with missing rate reaches 30%, and Figure 2(b)–(f) shows the clustering results with different methods based on VI.

From the experimental results, we see that when the missing rate is less than 30%, all the methods perform well. This is because sufficient information is retained for each cluster using the observed data. When the missing rate increases to more than 30%, the MI, KNN, and EM imputation methods do not perform well. The clustering accuracy of MI-VI drops sharply from 77.48% to 33.27%. With a decrease in the complete samples, the difference between missing samples is insignificant, with most samples clustering in the same category. From Table 2, we see that the MAE increased when the proportion of missing samples increased. Furthermore, the predicted cluster components could not reach the true component numbers using the MI or EM method; hence, the absolute mean error could become large. Although the proposed method consumes more run time, it can achieve the highest clustering accuracy.

The second dataset includes 1000 samples from a six-component Gaussian-Wishart distribution. The precision $\boldsymbol{\Sigma}^{-1}$ and mean $\boldsymbol{\mu}$ is sampled from

$$\mathcal{W}(\boldsymbol{\Sigma}^{-1}|\boldsymbol{w}^0, \kappa^0) \ \text{ and } \ \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{m}^0, (\beta^0\boldsymbol{\Sigma})^{-1})$$

for each class, respectively, where $\boldsymbol{w}^0$, κ^0, $\boldsymbol{m}^0$ and β^0 are the true values of the generating parameters set first. The plot of the synthetic data is shown in Figure 3(a), and Figure 3(b) shows the clustering result and the changing curve of the component numbers with the number of iteration using VI methods. We generated missing values on the first or second dimension of the original data set with different rates from 10% to 70%. For KNN imputation, we set the number of neighbors as one, reducing it to the nearest neighbor method. Figure 4(a) shows a 50% missing rate example, in which a black "$*$" adding to the original signs "$\cdot$". Figure 4(b)–(f) shows the clustering results with different methods based on VI method. We repeated the experiment 100 times and recorded the average clustering accuracy and run time for each method in Table 3, which illustrates the effectiveness of the methods proposed here, in particular, when the missing rate is high.

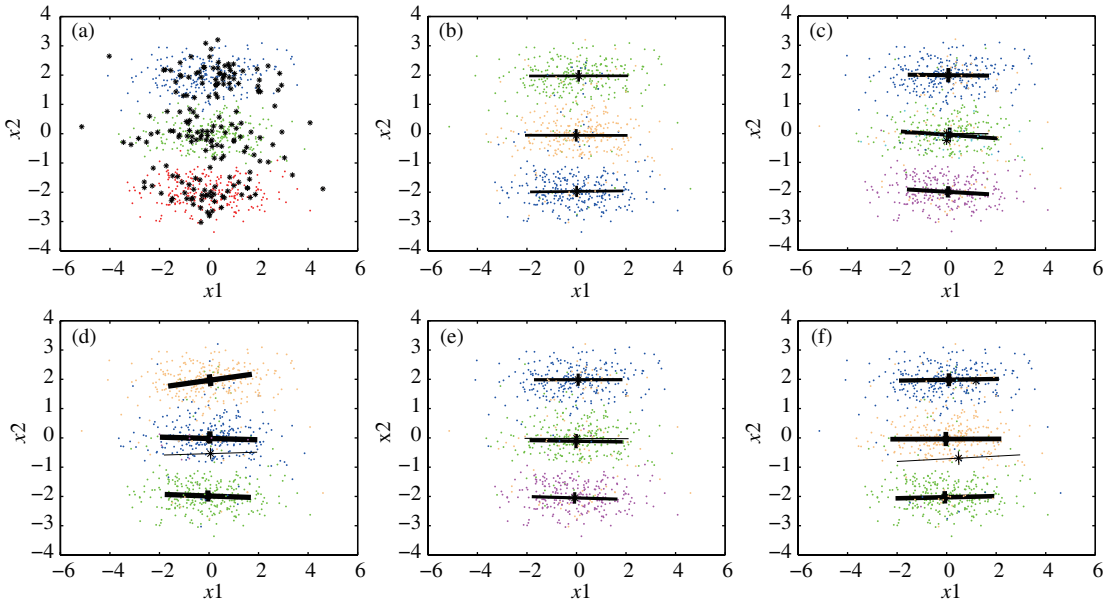

Figure 2 Different methods for clustering synthetic two-dimensional data set with missing rate $M_{\mathrm{r}} = 30\%$. (a) Incomplete data set where black "$*$" represents missing data and "$\cdot$" represents the complete samples; (b) LDNN-VI method; (c) MI-VI method; (d) KNN-VI method; (e) EM-VI method; and (f) VBEM method.

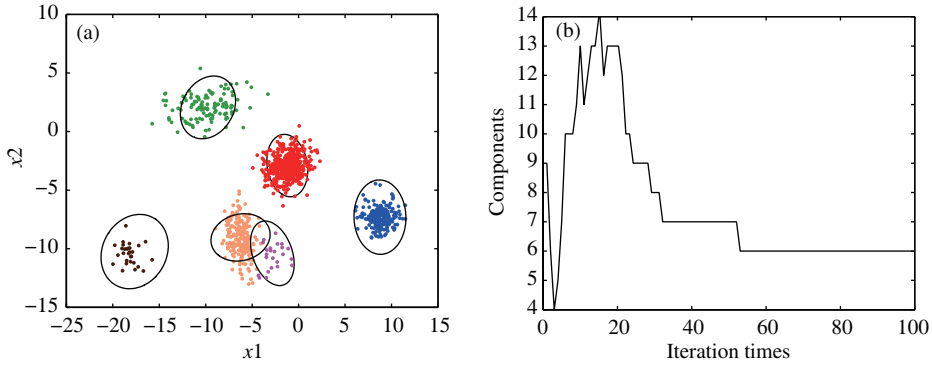

Figure 3 Fitting a six-component Gaussian-Wishart mixture (a) VI clustering result with $K = 6$, and (b) changing curve of the component numbers with the iteration times.

Table 3 Clustering accuracy (%) and average time (s) for Gaussian-Wishart dataset with different missing rates

Methods	10%	20%	30%	40%	50%	60%	70%	time
LDNN-VI	98.61	95.87	93.12	92.37	89.52	87.25	84.91	13.25
MI-VI	87.89	82.34	75.25	62.06	57.67	51.75	44.35	2.98
KNN-VI	98.32	96.17	93.45	91.23	89.68	84.62	80.05	15.31
EM-VI	91.43	86.82	82.54	74.63	54.22	50.61	41.70	2.41
VB-VI	98.25	96.93	95.23	93.85	93.11	92.67	91.32	315.26

The third dataset includes 1000 three-dimensional shrinking spiral datasets [5]. The data are generated according to

$$\begin{bmatrix} x_1^i \\ x_2^i \\ x_3^i \end{bmatrix} = \begin{bmatrix} (13 - 0.5t_i)\mathrm{cos}t_i \\ (10.5t_i - 13)\mathrm{sin}t_i \\ t_i \end{bmatrix} + \begin{bmatrix} n_1^i \\ n_2^i \\ n_3^i \end{bmatrix},$$

where t_i is uniformly distributed in $[0, 4\pi]$, and n_1^i, n_2^i, and n_3^i are independent and identically distributed with $\mathcal{N}(0, 1)$. When the data are complete, Figure 5(a) shows the clustering results using VI method.

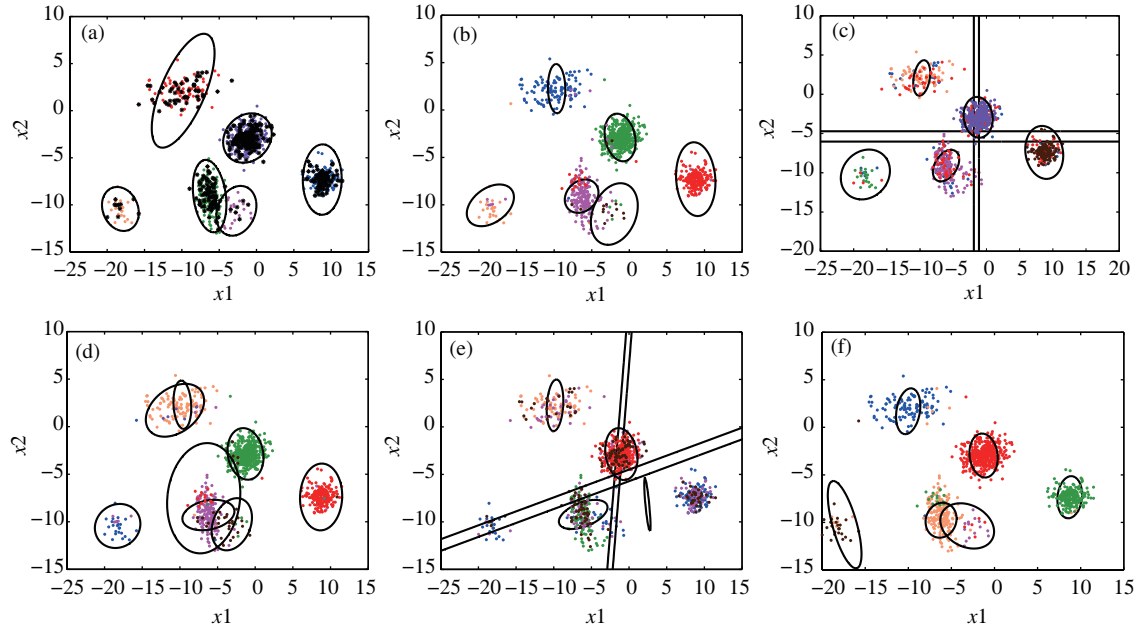

Figure 4 Different methods for (a) synthetic two-dimensional six components data set with missing rate $M_{\mathrm{r}} = 50\%$ in which black "$*$" represents missing data and "." represents the complete samples; (b) LDNN-VI method; (c) MI-VI method; (d) KNN-VI method; (e) EM-VI method; and (f) VBEM method.

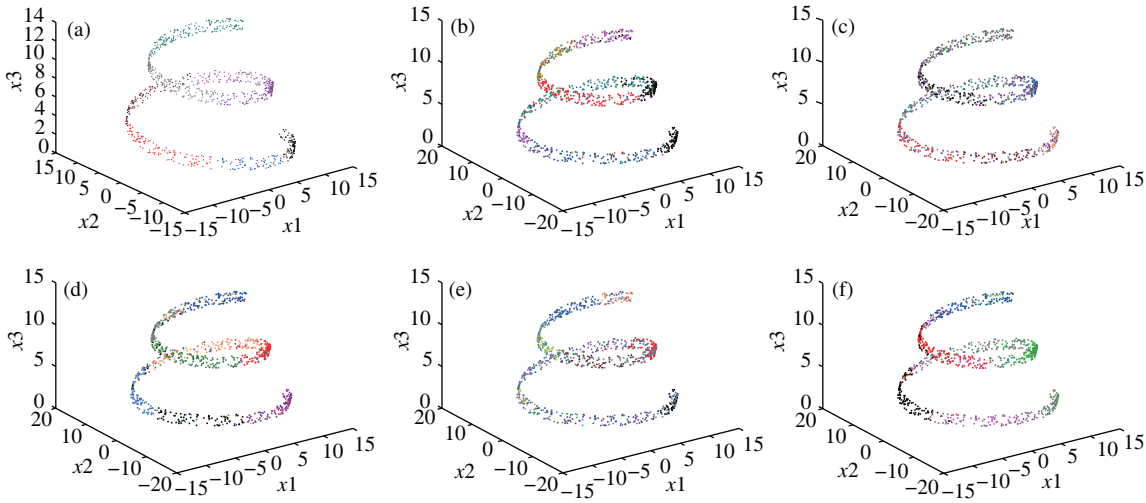

Figure 5 Different methods for fitting a complete three-dimensional shrinking spiral data set and incomplete data set with missing data $M_{\mathrm{r}} = 30\%$. (a) VI method on the complete data set; (b) LDNN-VI method; (c) MI-VI method; (d) KNN-VI method; (e) EM-VI method; and (f) VBEM method.

For missing rates as high as 50%, the clustering results are shown in Figure 5(b)–(f) by different methods. From the experimental result we see that our proposed method for treating shrinking spirals with missing data is more effective.

4.2 Image analysis

Gaussian mixture models are widely used to in computer vision to model natural images to accomplish the tasks of automatic clustering, image retrieval, and classification [27]. The number of mixture components must be set first, but the appropriate number is generally unknown. In this section, our aim is to demonstrate the applicability and robustness of our method for clustering hydrothermal mineral

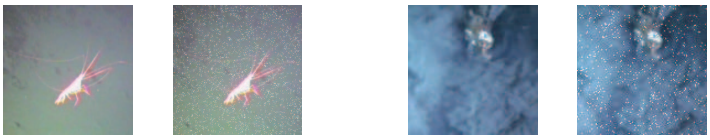

Figure 6 Seafloor hydrothermal sulfide images with random noisy data.

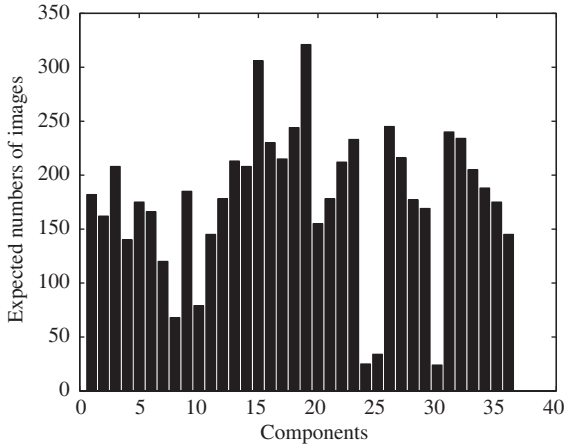

Figure 7 The expected number of images allocated to each component.

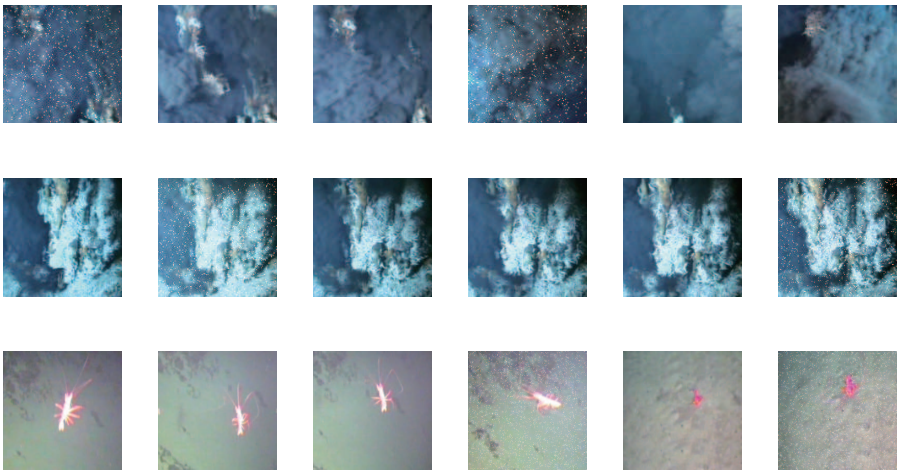

Figure 8 Three sample clusters from a DP mixture analysis of 3200 images from seafloor hydrothermal sulfide video screenshots. These clusters capture patterns in the data, such as black smoker chimney, hydrothermal sulfide deposit and bathyplankton, using the VBEM method.

images in the Trans-Atlantic geotraverse (TAG) area, which is situated in Mid-Atlantic Ridge at latitude 26.08° N. It is one of the largest sea-floor massive sulfide deposits in the sediment-free mid-ocean ridge [28].

We analyzed a collection of 6400 images obtained from video clips and set the sampling interval $ts = 0.5$ s. Because of the complex seafloor environment, combined with low contrast, uneven lighting, and blurred texture details, some images are obscure with noisy data. To highlight the applicability and effectiveness of our method for dealing with missing data, we add noisy (missing) data randomly for half the images on different pixels with missing rate $M_{\mathrm{r}} = 30\%$. Figure 6 shows the case with noisy data added to the original images. Each image was reduced to a real-valued vector using average red, green, and blue values. During the process of parameter estimation, we use diagonal matrix $\sigma^2 \boldsymbol{I}$ instead of covariance matrix $\boldsymbol{\Sigma}$ and the truncation level is set to $K = 50$ for the variational distribution. The algorithm requires nearly 20 min for each iteration.

Figure 7 shows the expected number of images allocated to each component under variational approximation to the posterior. We give an illustration in Figure 8, which shows seafloor images in the

same cluster with homologous approximate posterior probabilities. There are three components, each component including six pictures. These clusters appear to capture the characteristics of black smoker chimney, hydrothermal sulfide deposit and bathyplankton, respectively.

5 Conclusion

In this study, we proposed a novel approach to unsupervised learning for clustering with missing data. The infinite DPMMs are used to automatically determine the number of mixture components or clusters. This sidesteps setting the correct number of mixture components in advance, and allows the number to increase as new data arrive. As the cornerstone of nonparametric Bayesian method, DPMMs can avoid over-fitting. Furthermore, we view the missing features as latent variables and apply the VBEM algorithm to optimize the ELBO. The novelty in our approach is that the VI framework can solve the missing data problem effectively without advance imputation. Instead, we update the missing values at each iteration, considering the relation and uncertainty between samples. Experimental results indicated that the performance of our proposed method is better than those of classic imputation methods. We also applied it to classify seabed hydrothermal sulfide color images to capture different features in a hostile deep sea environment.

Acknowledgements This work was supported by Project of China Ocean Association (Grant No. DY125-25-02), Major Scientific Instrument Development Project of National Natural Science Foundation of China (Grant No. 41427806), National Natural Science Foundation of China (Grant No. 61273233), and Research Foundation for the Doctoral Program of Higher Education (Grant No. 20130002130010).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Li C Z, Xu Z B, Qiao C, et al. Hierarchical clustering driven by cognitive features. Sci China Inf Sci, 2014, 57: 012109
2 Wu C M, Chou S C, Liaw H T. A trend based investment decision approach using clustering and heuristic algorithm. Sci China Inf Sci, 2014, 57: 092117
3 McLachlan G, Peel D. Finite Mixture Models. Hoboken: John Wiley and Sons, 2004
4 Fan W, Bouguila N. Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection. Pattern Recogn, 2013, 46: 2754–2769
5 Figueiredo M A T, Jain A K. Unsupervised learning of finite mixture models. IEEE Trans Pattern Anal Mach Intell, 2002, 24: 381–396
6 Neal R M. Markov chain sampling methods for Dirichlet process mixture models. J Comput Graph Stat, 2000, 9: 249–265
7 Blei D M, Jordan M I. Variational inference for Dirichlet process mixtures. Bayesian Anal, 2006, 1: 121–143
8 Kim S, Tadesse M G, Vannucci M. Variable selection in clustering via Dirichlet process mixture models. Biometrika, 2006, 93: 877–893
9 Orbanz P, Buhmann J M. Nonparametric Bayesian image segmentation. Int J Comput Vision, 2008, 77: 25–45
10 García-Laencina P J, Sancho-Gómez J L, Figueiras-Vidal A R. Pattern classification with missing data: a review. Neural Comput Appl, 2010, 19: 263–282
11 Wang C, Liao X, Carin L, et al. Classification with incomplete data using Dirichlet process priors. J Mach Learn Res, 2010, 11: 3269–3311
12 Williams D, Liao X J, Xue Y, et al. On classification with incomplete data. IEEE Trans Pattern Anal Mach Intell, 2007, 29: 427–436
13 Schafer J L, Graham J W. Missing data: our view of the state of the art. Psychol Method, 2002, 7: 147–177
14 Little R J A, Rubin D B. Statistical Analysis with Missing Data. 2nd ed. Hoboken: John Wiley and Sons, 2002
15 Chechik G, Heitz G, Elidan G, et al. Max-margin classification of data with absent features. J Mach Learn Res, 2008, 9: 1–21
16 Fidler S, Skocaj D, Leonardis A. Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling. IEEE Trans Pattern Anal Mach Intell, 2006, 28: 337–350
17 Chan K, Lee T W, Sejnowski T J. Variational learning of clusters of undercomplete nonsymmetric independent components. J Mach Learn Res, 2003, 3: 99–114
18 Teh Y W, Jordan M I, Beal M J, et al. Hierarchical dirichlet processes. J Amer Stat Assoc, 2006, 101: 1566–1581
19 Sethuraman J. A constructive definition of Dirichlet priors. Stat Sin, 1994, 4: 639–650

20 Ghahramani Z, Beal M J. Propagation algorithms for variational Bayesian learning. In: Leen T K, Dietterich T, Tresp V, eds. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2001. 507–513
21 Hughes M C, Sudderth E. Memoized online variational inference for Dirichlet process mixture models. In: Burges C J C, Bottou L, Welling M, et al, eds. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2013. 1133–1141
22 Bishop C M. Pattern Recognition and Machine Learning. New York: springer, 2006
23 Lin T I, Lee J C, Ho H J. On fast supervised learning for normal mixture models with missing information. Pattern Recogn, 2006, 39: 1177–1187
24 Collins L M, Schafer J L, Kam C M. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychol Method, 2001, 6: 330–351
25 Meng X L, Rubin D B. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika, 1993, 80: 267–278
26 Ueda N, Nakano R. Deterministic annealing EM algorithm. Neural Netw, 1998, 11: 271–282
27 Barnard K, Duygulu P, Forsyth D, et al. Matching words and pictures. J Mach Learn Res, 2003, 3: 1107–1135
28 Herzig P M, Hannington M D. Polymetallic massive sulfides at the modern seafloor a review. Ore Geol Rev, 1995, 10: 95–115