Axial Data Modeling with Collapsed Nonparametric Watson Mixture Models and Its Application to Depth Image Analysis

Lin Yang[1], Yuhang Liu[1], and Wentao Fan[1,2,3](✉)

[1] Department of Computer Science and Technology,
Huaqiao University, Xiamen, Fujian, China
{19014083028,1625161012}@stu.hqu.edu.cn, fwt@hqu.edu.cn

[2] Xiamen Key Laboratory of Computer Vision and Pattern Recognition,
Huaqiao University, Xiamen, China

[3] Key Laboratory of Computer Vision and Machine Learning (Huaqiao University),
Fujian Province University, Xiamen, China

Abstract. Recently, axial data (i.e. the observations are axes of direction) have been involved with various fields ranging from blind speech separation to gene expression data clustering. In this paper, axial data modeling is performed by proposing a nonparametric infinite Watson mixture model which is constructed in a collapsed space (denoted by Co-InWMM) where the mixing coefficients are integrated out. Then, an effective collapsed variational Bayes (CVB) inference method is theoretically developed to learn the Co-InWMM with closed-from solutions. The proposed Co-InWMM with CVB inference for modeling axial data is validated through both synthetical data sets and a challenging application regarding depth image analysis.

Keywords: Axial data modeling · Mixture model · Watson distributions · Collapsed variational Bayes · Depth image analysis

1 Introduction

In recent years, directional data (i.e. the "direction" of the data is more important than their magnitude) analysis has drawn significant attention in various fields [12,14]. Typical directional data are the data that are normalized to have unit norm, which lie on the surface of the unit sphere. Since directional data

The first author is a student. The completion of this work was supported by the National Natural Science Foundation of China (61876068), the Natural Science Foundation of Fujian Province (2018J01094), the Promotion Program for Young and Middle-aged Teacher in Science and Technology Research of Huaqiao University (ZQNPY510) and Provincial Key Laboratory of Computer Vision and Machine Learning of Educational Department of Fujian Province (201902).

Y. Peng et al. (Eds.): PRCV 2020, LNCS 12306, pp. 17–28, 2020.
https://doi.org/10.1007/978-3-030-60639-8_2

are better represented on a manifold, the nonlinear nature of manifolds implies that common distributions such as the multivariate Gaussian distribution can not be used to model and analyze directional data. Alternatively, distributions that are defined on the unit hypersphere are more appropriate and effective to model directional data.

One of the most basic directional distributions is the von Mises-Fisher (vMF) distribution, which is defined on the unit hyperspahere ($\mathbb{S}^{D-1}$) and has similar characteristics to those of the multivariate Gaussian distribution defined in the Euclidean space $\mathbb{R}^D$. Although vMF distributions were widely involved with directional data modeling, it is not a universal solution to all types directional data. For instance, resent reach works have demonstrated that *axial data* where the observations are axes of direction (i.e. the unit vectors $\pm\boldsymbol{X}$ are indistinguishable) are better modeled with Watson distributions rather than with vMF [1]. As a special type of directional data, axial data have found their applications in various applications, such as blind speech separation [21], speech clustering in distributed microphone arrays [17], differentiation between normal and schizophrenic brains [13], gene expression data clustering analysis [7], etc.

Different methods have been proposed to learn Watson distributions or its natural extension the Watson mixture model (WMM). The major difficulty of learning Watson-based models lies on the fact that no analytically solution to the inference of the concentration parameters of Watson distributions can be found. Thus, approximation methods were proposed to solve this problem. A simple approximation method for large concentrations has been proposed in [13] to learn Watson distributions with the maximum likelihood (ML) estimates. This learning method, however, can not deal axial data with higher dimensions. In [1], an approximation to ML estimates has been proposed within an expectation maximization (EM) framework to learn WMMs. However, this method is prone to the problem of over-fitting. A better alternative method to ML estimates is the variational Bayes (VB) [4,9], a method that approximates posterior distributions through optimization. In [18], a VB inference method was proposed to learn WMMs and demonstrated better performance than the ML estimates. Although closed-form solutions can be obtained by this method, the evaluation of the model complexity (i.e. the number of mixture components model that best fit the data) requires extra effort. Specifically, the VB inference method in [18] treats the mixing coefficients of the WMM as random variables which are assigned with a Dirichlet prior. Then, model selection was performed by removing the components with small responsibilities. A more elegant solution to the model selection problem in modeling WMMs was proposed in [7], where a nonparametric framework known as the *Dirichlet process mixture model* [3,10] was adopted to define the WMM with an infinite number of components. By applying VB inference method to learn the infinite WMM (In-WMM), the number of mixture component can be freely initialized and will be adjusted automatically as the data set increases [7].

Although both VB inference methods ([18] and [7]) are effective to learn WMMs, to ensure closed-form solutions, the VB inference has to adopt the mean-field assumption [2] where parameters are assumed to be independent. This assumption, however, is not realistic in the WMM or In-WMM in which the mixing coefficients and the latent indicator variables are obviously closely related. This issue can be addressed by applying VB inference in a collapsed space where parameters are marginalized out, which leads to the so-called collapsed VB (CVB) inference framework [20]. As described in [11,20], the mean-field assumption is more satisfied with CVB without the concern of dependencies between parameters. Thus, in this work we focus on developing an effective CVB inference method to learn the In-WMM in a collapsed space where the mixing coefficients are integrated out.

We summarize the contributions of this work as follows. Firstly, a collapsed infinite WMM (Co-InWMM) is proposed for modeling axial data by marginalizing out the mixing coefficients. Secondly, an effective CVB inference method is theoretically developed to learn Co-InWMM with closed-from solutions. Lastly, the proposed Co-InWMM with CVB inference is validated through both synthetical data sets and a challenging application about depth image analysis.

2 The Collapsed Infinite WMM

2.1 Infinite Watson Mixture Models

Given a data set $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ which contains N axial random vectors (i.e. $\boldsymbol{x}$ and $-\boldsymbol{x}$ are equivalent), each D-dimensional data vector can be represented as a unit vector (i.e. $\|\boldsymbol{x}\|_2 = 1$) defined on a $(D-1)$-dimensional unit hypersphere $\mathbb{S}^{D-1}$. If each vector $\boldsymbol{x}$ is a drawn from a mixture of an infinite number of Watson distributions, then the probability density function of this infinite Watson mixture model (InWMM) is given by

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\gamma}) = \sum_{k=1}^{\infty} \pi_k \mathcal{W}(\boldsymbol{x}|\boldsymbol{\mu}_k, \gamma_k) \tag{1}$$

where $\boldsymbol{\pi} = \{\pi_k\}_{k=1}^{\infty}$ represent the mixing coefficients that should be nonnegative and sum to 1; $\boldsymbol{\mu} \in \mathbb{S}^{D-1}$ denotes the *mean direction* with $\|\boldsymbol{\mu}\|_2 = 1$, and $\gamma \in \mathbb{R}$ represents the *concentration*. $\mathcal{W}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \gamma_k)$ indicates the Watson distribution associated with the kth component of the mixture model and is defined by

$$\mathcal{W}(\boldsymbol{x}|\boldsymbol{\mu}_k, \gamma_k) = \frac{\mathit{\Gamma}(D/2)}{2\pi^{D/2} M(\frac{1}{2}, \frac{D}{2}, \gamma_k)} \exp[\gamma_k (\boldsymbol{\mu}_k^T \boldsymbol{x})^2] \tag{2}$$

where $M(a, b, \cdot)$ represents the Kummer function (also known as the confluent hypergeometric function) which is given by

$$M(a, b, \gamma) = \sum_{n=0}^{\infty} \frac{\mathit{\Gamma}(a+n)\mathit{\Gamma}(b)}{\mathit{\Gamma}(a)\mathit{\Gamma}(b+n)} \frac{\gamma^n}{n!} \tag{3}$$

where $\mathit{\Gamma}(\cdot)$ denotes the Gamma function.

Next, each vector $\boldsymbol{x}_i$ is assigned with a latent indicator variable z_i which is used to indicate the component from which $\boldsymbol{x}_i$ is drawn. For the data set $\mathcal{X}$, the distribution of indicator variables $\boldsymbol{z} = \{z_i\}_{i=1}^N$ can be represented by

$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \prod_{i=1}^{N}\prod_{k=1}^{\infty} \pi_k^{\mathbf{1}[z_i=k]} \tag{4}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function which equals 1 when $z_i = k$, otherwise it equals 0.

2.2 Prior Distributions

The InWMM is constructed using a Bayesian framework, in which each unknown variable is assigned with a prior distribution. A nonparametric prior namely Dirichlet process [10] is considered for the mixing coefficients $\boldsymbol{\pi}$, and is defined in terms of a stick-breaking representation [3] as

$$\pi_k = \pi'_k \prod_{s=1}^{k-1}(1-\pi'_s), \quad \pi'_k \sim \text{Beta}(1, \varpi_k), \quad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}, \quad \theta_k \sim H \tag{5}$$

where G is a drawn from the Dirichlet process $G \sim DP(\varpi, H)$ with the base distribution H and scaling parameter ϖ, where δ_{θ_k} is an atom at θ_k.

Following [7,18], a Watson-Gamma prior is selected for parameters $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ as

$$p(\boldsymbol{\mu}, \boldsymbol{\gamma}) = \prod_{k=1}^{\infty} \mathcal{W}(\boldsymbol{\mu}_k|\boldsymbol{m}_k, \beta_k\gamma_k)\mathcal{G}(\gamma_k|a_k, b_k) \tag{6}$$

where $\mathcal{G}(\cdot)$ indicates the Gamma distribution.

2.3 Collapsed Infinite Watson Mixture Models

According to several recent works in the literature of mixture modeling [5,6], better performance often would be obtained when model learning was conducted in a collapsed space where some or all of the parameters are marginalized out. In our case, inspired from [5,6,11], we re-formulate a collapsed version of InWMM (i.e. the Co-InWMM) by marginalizing out the mixing coefficients $\boldsymbol{\pi}$. Consequently, the latent variable $\boldsymbol{z}$ does not depend on the mixing coefficients $\boldsymbol{\pi}$ anymore and is distributed as

$$p(\boldsymbol{z}) = \prod_{k=1}^{\infty} \frac{\Gamma(1+n_k)\Gamma(\varpi_k + n_{>k})}{\Gamma(1+\varpi_k + n_{\geq k})} \tag{7}$$

where $n_k = \sum_{i=1}^N \mathbf{1}[z_i = k]$ indicates the number of data instances from the kth component, $n_{>k} = \sum_{i=1}^N \mathbf{1}[z_i > k]$, and $n_{\geq k} = n_k + n_{>k}$.

The conditional distribution of $z_i = k$ given the current state of all except one variable z_i is

$$p(z_i = k|\boldsymbol{z}^{\neg i}) \propto (1 + n_k^{\neg i})(\varpi_k + n_{>k}^{\neg i})(1 + \varpi_k + n_{\geq k}^{\neg i})^{-1} \tag{8}$$

where the superscript $\neg i$ indicates the associated ith term is removed.

The joint distribution of all latent and random variables in the Co-InWMM is given by

$$p(\mathcal{X}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\gamma}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\mu}_{z_i}, \gamma_{z_i})p(z_i) \prod_{k=1}^{\infty} p(\boldsymbol{\mu}_k, \gamma_k) \tag{9}$$

In contrast with the InWMM as described in Eq. (1), the Co-InWMM has two major advantages: 1) the explicit dependency between latent variables $\boldsymbol{z}$ and mixing coefficients $\boldsymbol{\pi}$ is broken, which will be in favor of the mean-filed variational Bayes model learning method as developed in the following section; 2) a smaller number of parameters are obtained by integrating out $\boldsymbol{\pi}$, which leads to a faster inference process with better performance.

3 Model Learning

In this section, based on the VB inference methods that were respectively proposed in [7,18] for learning finite WMM and InWMM, we develop an effective method based on collapsed variational Bayes (CVB) [11,20] to learn the proposed Co-InWMM with closed-form solutions.

3.1 Mean-Field Collapsed Variational Inference

VB inference is an effective method for approximating posterior dentistries in Bayesian models. In our case, VB is adopted to approximate the true posterior $p(\Theta|\mathcal{X})$ with an approximated posterior $q(\Theta)$ (also referred to as *variational posterior*), where $\Theta = \{\boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\gamma}\}$ denotes the set of all latent and random variables of the Co-InWMM. VB inference solves the problem of approximation though optimization, by minimizing the Kullback-Leibler (KL) divergence between $q(\Theta)$ and $p(\Theta|\mathcal{X})$, which is equivalent to maximizing the lower bound of $\ln p(\mathcal{X})$ that is defined by

$$\mathcal{L}(q) = \int q(\Theta) \ln[p(\mathcal{X}, \Theta)/q(\Theta)] d\Theta \tag{10}$$

To perform VB inference for learning Co-InWMM which contains an infinite number of mixture components, a common technique is to truncate the stick-breaking representation of Co-InWMM at a finite value K as

$$\pi'_K = 1, \quad \sum_{k=1}^{K} \pi_k = 1, \quad \pi_k = 0 \text{ when } k > K \tag{11}$$

where K can be freely initialized and would be inferred automatically through VB inference.

To obtain closed-from solutions, *mean-field* assumption [2] is often adopted in VB inference to factorize the variational posterior as the product of independent factors, where each factor represents variational posterior of the corresponding variable. In [7], the variational posterior of InWMM with truncation was factorized as

$$q(\Theta) = q(\boldsymbol{\pi})q(\boldsymbol{z})q(\boldsymbol{\mu}, \boldsymbol{\gamma}) \tag{12}$$

This factorization assumption, however, clearly violates the fact that latent variables $\boldsymbol{z}$ and mixing coefficients $\boldsymbol{\pi}$ are closely related with strong dependency as demonstrated in Eq. (4). The mean-field assumption is more satisfied in Co-InWMM where $\boldsymbol{\pi}$ are marginalized out as

$$q(\Theta) = \prod_{i=1}^{N}\left[q(z_i)\right]\prod_{k=1}^{K}\left[q(\boldsymbol{\mu}_k, \gamma_k)\right] \tag{13}$$

Then, we can obtain the following update equations by maximizing the lower bound $\mathcal{L}(q)$ with respect to each variational posterior

$$q(\boldsymbol{z}) = \prod_{i=1}^{N}\prod_{k=1}^{K} r_{ik}^{\mathbf{1}[z_i=k]} \tag{14}$$

$$q(\boldsymbol{\mu}, \boldsymbol{\gamma}) = \prod_{k=1}^{K} \mathcal{W}(\boldsymbol{\mu}_k|\boldsymbol{m}_k^*, \beta_k^*\gamma_k)\mathcal{G}(\gamma_k|a_k^*, b_k^*) \tag{15}$$

where the hyperparameters in the above variational posteriors are calculated by

$$r_{ik} = \frac{\widetilde{r}_{ik}}{\sum_{s=1}^{K}\widetilde{r}_{is}}, \tag{16}$$

$$\begin{aligned}
\widetilde{r}_{ik} =& \ln\Gamma(\frac{D}{2}) - \frac{D}{2}\ln 2\pi + \frac{D}{2}\langle\ln\gamma_k\rangle - \ln[\bar{\gamma}_k^{\frac{D}{2}} M(\frac{1}{2}, \frac{D}{2}, \bar{\gamma}_k)] \\
& - \frac{\partial}{\partial\bar{\gamma}_k}\left[\ln\bar{\gamma}_k^{\frac{D}{2}} M(\frac{1}{2}, \frac{D}{2}, \bar{\gamma}_k)\right](\langle\gamma_k\rangle - \bar{\gamma}_k) \\
& + \bar{\gamma}_k\vartheta(\beta_k^*\bar{\gamma}_k) + \big\{\bar{\gamma}_k[\vartheta(\beta_k^*\bar{\gamma}_k) + \beta_k^*\bar{\gamma}_k\vartheta'(\beta_k^*\bar{\gamma}_k)] \\
& \times (\langle\ln\gamma_k\rangle + \ln\beta_k^* - \ln\beta_k^*\bar{\gamma}_k)\big\}(\boldsymbol{m}_k^{*T}\boldsymbol{X}_i)^2 \\
& + \langle\ln(1 + n_k^{\neg i})\rangle - \langle\ln(1 + \varpi_k + n_{\geq k}^{\neg i})\rangle \\
& + \sum_{j<k}\left[\langle\ln(\varpi_j + n_{>j}^{\neg i})\rangle - \langle\ln(1 + \varpi_j + n_{\geq j}^{\neg i})\rangle\right]
\end{aligned} \tag{17}$$

$$a_k^* = a_k + \frac{D}{2}(1 + \sum_{i=1}^{N}\langle z_{i=k}\rangle) + \beta_k^*\bar{\gamma}_k\frac{\partial}{\partial\beta_k^*\bar{\gamma}_k}\ln M\big(\frac{1}{2}, \frac{D}{2}, \beta_k^*\bar{\gamma}_k\big) \tag{18}$$

$$\begin{aligned}
b_k^* =& b_k + \sum_{i=1}^{N}\langle z_{i=k}\rangle\frac{\partial}{\partial\bar{\gamma}_k}\left[\ln\bar{\gamma}_k^{\frac{D}{2}} M(\frac{1}{2}, \frac{D}{2}, \bar{\gamma}_k)\right] \\
& + \beta_k\frac{\partial}{\partial\beta_k\bar{\gamma}_k}\left[\ln(\beta_k\bar{\gamma}_k)^{\frac{D}{2}} M(\frac{1}{2}, \frac{D}{2}, \beta_k\bar{\gamma}_k)\right]
\end{aligned} \tag{19}$$

$$A = \beta_k \boldsymbol{m}_k \boldsymbol{m}_k^T + \sum_{i=1}^{N} \langle z_{i=k} \rangle \boldsymbol{x}_i \boldsymbol{x}_i^T \tag{20}$$

where $\vartheta(x) = \frac{\partial}{\partial x} \ln M\left(\frac{1}{2}, \frac{D}{2}, x\right)$, β_k^* is the largest eigenvalue of A, $\boldsymbol{m}_k^*$ represents the corresponding eigenvector to β_k^*. The expected values in above equations are given by

$$\langle z_{i=k} \rangle = r_{ik}, \qquad \bar{\gamma}_k = a_k^*/b_k^*, \qquad \langle \ln \gamma_k \rangle = \psi(a_k^*) - \ln b_k^* \tag{21}$$

$$\langle \ln(1 + n_k^{\neg i}) \rangle \approx \ln(1 + \langle n_k^{\neg i} \rangle), \tag{22}$$

$$\langle \ln(\varpi_k + n_{>k}^{\neg i}) \rangle \approx \ln(\varpi_k + \langle n_{>k}^{\neg i} \rangle) \tag{23}$$

$$\langle \ln(1 + \varpi_k + n_{\geq k}^{\neg i}) \rangle \approx \ln(1 + \varpi_k + \langle n_{\geq k}^{\neg i} \rangle) \tag{24}$$

$$\langle n_k^{\neg i} \rangle = \sum_{i' \neq i} r_{i'k}, \qquad \langle n_{>k}^{\neg i} \rangle = \sum_{i' \neq i} \sum_{s=k+1}^{K} r_{i's}, \qquad \langle n_{\geq k}^{\neg i} \rangle = \langle n_k^{\neg i} \rangle + \langle n_{>k}^{\neg i} \rangle \tag{25}$$

where the expected values of $\ln(1 + n_k^{\neg i})$, $\ln(\varpi_k + n_{>k}^{\neg i})$, and $\ln(1 + \varpi_k + n_{\geq k}^{\neg i})$ were acquired according to Gaussian approximations [20] with 0th-order Taylor approximation [15]. Our CVB inference method for learning the Co-InWMM is analogous to the maximum likelihood expectation maximization (EM) algorithm, which is summarized in Algorithm 1.

Algorithm 1. CVB Inference of the Co-InWMM.

```
Initialize the truncation level $K$.
Initialize the hyper-parameters $a_k$, $b_k$, $\varpi_k$, and $\beta_k$.
Apply $K$-Means algorithm to initialize $r_{ik}$.
repeat
    The variational E-step:
    Estimate the expected values in (21)∼(25), use the current distributions over
    the model parameters.
    The variational M-step:
    Update the variational posteriors with (14) and (15) based on the estimated
    expected values.
until The convergence criterion is satisfied.
```

4 Experimental Results

The proposed Co-InWMM with CVB inference is evaluated through two experiments involved with both simulated data and a application about depth image analysis. In our experiments, the truncation level K is initialized to 10, ϖ_k and β_k are set to 1, a_k and b_k are initialized to 1 and 0.01, respectively. These initial values were found through cross validation.

4.1 Synthetic Data

The principal purpose of conducting experiments on synthetic axial data is to validate the "correctness" of the proposed CVB inference algorithm in learning the proposed Co-InWMM. This is fulfilled by verifying the discrepancy between computed values of the parameters and their true values. A synthetic data set was generated to conduct the experiments. This data set contains 900 3-dimensional data instances which are drawn from 3 Watson distributions (as demonstrated in Fig. 1).

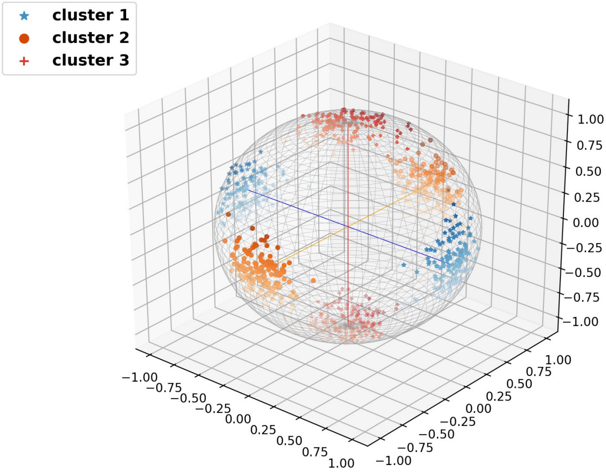

Fig. 1. The synthetic data set.

The true parameters that were used to generate the data set and the estimated parameters by CVB inference method are shown in Table 1. According to this table, the proposed learning algorithm is able to effectively learn the Co-InWMM with estimated values of parameters that are vary close to the true ones.

4.2 Depth Image Analysis

In this experiment, we apply the proposed Co-InWMM to a challenging application namely depth image analysis. We use the NYU-V2 depth data set [16] to conduct our experiments. This data set includes 1449 rgb-d images collected from three different cities in the United States, consisting of 464 indoor different scenes across 26 scene classes in commercial buildings and residences. Following [8], we compute surface normals of depth images and then apply Co-InWMM for clustering the normals. It is worth noting that the axially symmetric property of WMM can naturally overcome the ambiguity signals caused by the normal vector which calculated by plane fitting method.

Table 1. Parameters estimation of the synthetic data set.

N_k	k	μ_{k1}	μ_{k2}	μ_{k3}	κ_k	$\hat{\mu}_{k1}$	$\hat{\mu}_{k2}$	$\hat{\mu}_{k3}$	$\hat{\kappa}_k$
300	1	0	0	1	15	0.01	0.02	0.99	15.46
300	2	0	1	0	22	0.01	0.99	−0.01	22.77
300	3	1	0	0	17	0.99	0.00	−0.01	17.19

Figure 2 shows the number of estimated clusters for all NYU-V2 depth data set obtained by finite WMM with the Integrated Completed Likelihood (ICL) criteria [8] and the proposed Co-InWMM. As can we can see from the figure, most of the images contain 3–4 clusters. It is note worthy that, the WMM method in [8] has to calculate the ICL criteria with different number of clusters in order to determine the optimal number. In contrast, our model can detect the number of clusters automatically with a single run.

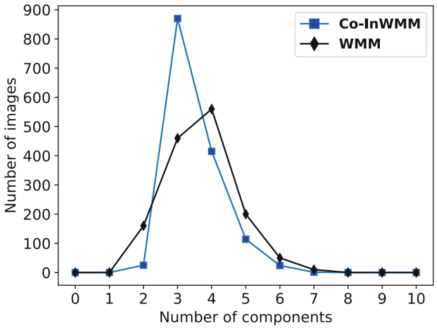

Fig. 2. Estimated number of clusters for NYU-V2 depth data set.

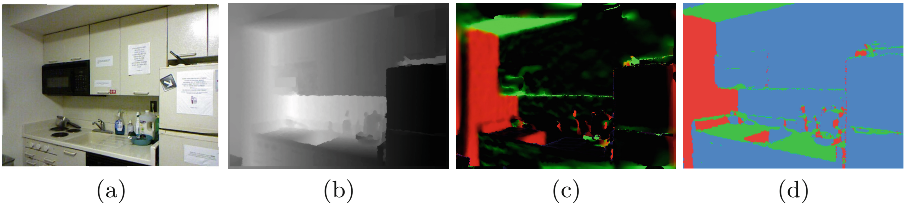
(a) (b) (c) (d)

Fig. 3. Cluster example in the NYU-V2 depth data set. (a) rgb image; (b) depth image; (c) normals; (d) results by Co-InWMM.

Figure 3 shows the example of depth image analysis. From the results we observe that, different clusters represent different image regions and also represent the segment plane associated with the scene with a specific axis. Other

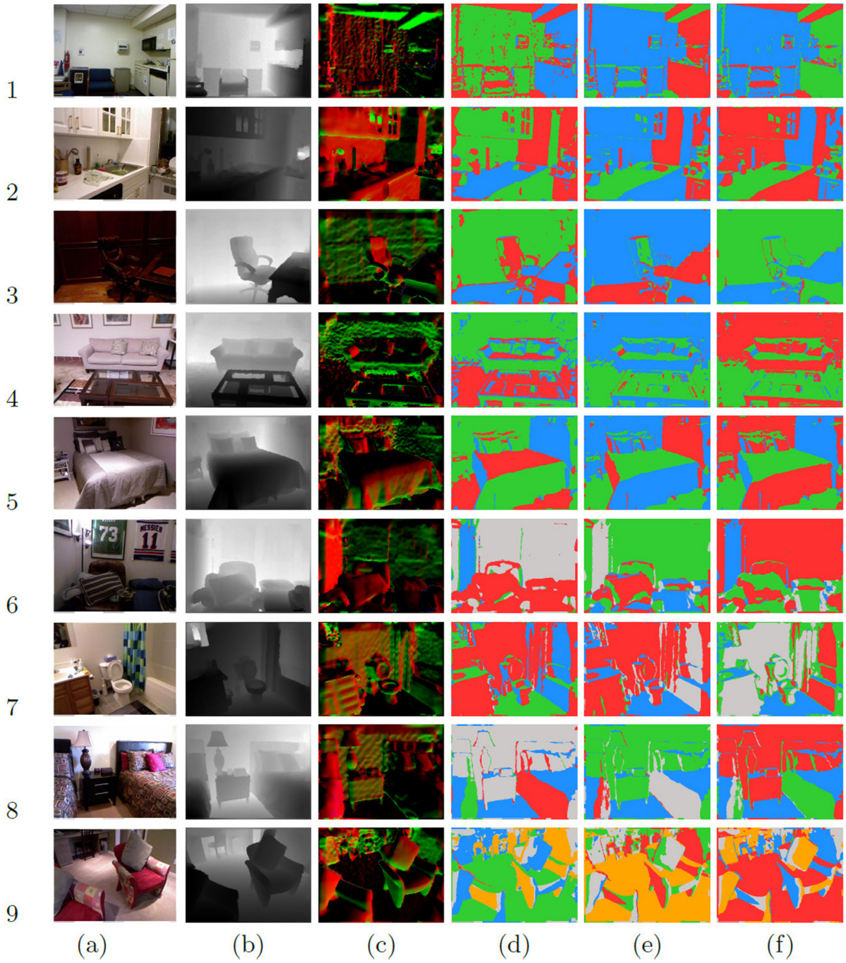

Fig. 4. Cluster results on the NYU-V2 depth data set. (a) rgb images; (b) depth images; (c) normals; (d) results by WMM; (e) results by In-WMM; (f) results by Co-InWMM.

results can be seen in Fig. 4. Through the results, we can see that some classes represent some nonplanar objects (see case-7 and case-9 of Fig. 4), which means that our method can find nonplanar objects. From case-3 and case-5, we can see a lot of noise on the normal vector, but our method can still identify plane and nonplaner objects well. In addition, similar to [8], we also find that the data with lower prior probability will be divided into fewer clusters. In order to solve this problem, a reasonable solution is to highlight each cluster by preprocessing the normal vector to make the clustering more accurate.

In order to show the superiority of our model, we compare it with Kmeans, finite vMFMM [19], finite WMM [8] and the In-WMM proposed in [7] in terms of clustering performance on normals and computational runtime. It should be

Table 2. Results obtained by different methods in terms of MI and computational runtime (in min.)

Algorithm	MI	Time
Kmeans	0.293	241.39
vMFMM [19]	0.329	273.19
WMM [8]	0.335	270.00
In-WMM [7]	0.347	164.22
Co-InWMM	**0.355**	**156.97**

noted that the first three algorithms use ICL criteria to calculate the optimal cluster number. We use mutual information (MI) to evaluate the performance of clustering. The specific results are shown in Table 2. Based on the results shown in this table, it is obvious that the Co-InWMM is able to provide better clustering performance in terms of the highest MI value. Moreover, the Co-InWMM is more computational efficient than other tested methods in terms of the shortest computational runtime. This result demonstrates the advantages of constructing the nonparametric infinite WMM in a collapsed space, where mixing coefficients are integrated out and thus leads to a smaller number of parameters that have to be estimated.

5 Conclusion

In this paper, we proposed a collapsed infinite Watson mixture model for modeling axial data where the mixing coefficients are integrated out. We developed an effective collapsed variational Bayes inference method to learn the proposed model with closed-from solutions. The effectiveness of the proposed Co-InWMM with CVB inference for modeling axial data was verified through experiments that were conducted on both synthetical data sets and a challenging application regarding depth image analysis.

References

1. Bijral, A.S., Breitenbach, M., Grudic, G.Z.: Mixture of watson distributions: a generative model for hyperspherical embeddings. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, pp. 35–42 (2007)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)
3. Blei, D.M., Jordan, M.I.: Variational inference for Dirichlet process mixtures. Bayesian Anal. **1**, 121–144 (2005)
4. Blei, D.M., Kucukelbir, A., Mcauliffe, J.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017)
5. Fan, W., Bouguila, N.: Modeling and clustering positive vectors via nonparametric mixture models of Liouville distributions. IEEE Trans. Neural Netw. Learn. Syst. 1–11 (2019). https://doi.org/10.1109/TNNLS.2019.2938830

6. Fan, W., Bouguila, N.: Simultaneous clustering and feature selection via nonparametric Pitman-Yor process mixture models. Int. J. Mach. Learn. Cybernet. **10**(10), 2753–2766 (2019)
7. Fan, W., Bouguila, N., Du, J., Liu, X.: Axially symmetric data clustering through Dirichlet process mixture models of Watson distributions. IEEE Trans. Neural Netw. Learn. Syst. **30**(6), 1683–1694 (2019)
8. Hasnat, M.A., Alata, O., Trémeau, A.: Unsupervised clustering of depth images using watson mixture model. In: 2014 22nd International Conference on Pattern Recognition, pp. 214–219 (2014)
9. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. Mach. Learn. **37**(2), 183–233 (1999)
10. Korwar, R.M., Hollander, M.: Contributions to the theory of Dirichlet processes. Ann. Probab. **1**, 705–711 (1973)
11. Kurihara, K., Welling, M., Teh, Y.W.: Collapsed variational Dirichlet process mixture models. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI), pp. 2796–2801 (2007)
12. Ley, C., Verdebout, T.: Applied Directional Statistics: Modern Methods and Case Studies. Chapman and Hall/CRC, Boca Raton (2018)
13. Mardia, K.V., Dryden, I.L.: The complex Watson distribution and shape analysis. J. Roy. Stat. Soc.: Ser. B (Stat. Methodol.) **61**(4), 913–926 (1999)
14. Mardia, K.V., Jupp, P.E.: Directional Statistics. Wiley, Hoboken (2000)
15. Sato, I., Nakagawa, H.: Rethinking collapsed variational bayes inference for LDA. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012 (2012)
16. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33715-4_54
17. Souden, M., Kinoshita, K., Nakatani, T.: An integration of source location cues for speech clustering in distributed microphone arrays. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 111–115 (2013)
18. Taghia, J., Leijon, A.: Variational inference for Watson mixture model. IEEE Trans. Pattern Anal. Mach. Intell. **38**(9), 1886–1900 (2016)
19. Taghia, J., Ma, Z., Leijon, A.: Bayesian estimation of the von-Mises fisher mixture model with variational inference. IEEE Trans. Pattern Anal. Mach. Intell. **36**(9), 1701–1715 (2014)
20. Teh, Y.W., Newman, D., Welling, M.: A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In: Proceedings of Advances in Neural Information Processing Systems (NIPS), pp. 1353–1360 (2007)
21. Vu, D.H.T., Haeb-Umbach, R.: Blind speech separation employing directional statistics in an expectation maximization framework. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 241–244 (2010)