DIRICHLET PROCESS MIXTURE MODELS FOR CLUSTERING I-VECTOR DATA

Shreyas Seshadri *Ulpu Remes* *Okko Räsänen*

Aalto University, Department of Signal Processing and Acoustics, Finland

ABSTRACT

Non-parametric Bayesian methods have recently gained popularity in several research areas dealing with unsupervised learning. These models are capable of simultaneously learning the cluster models as well as their number based on properties of a dataset. The most commonly applied models are using Dirichlet process priors and Gaussian models, called as Dirichlet process Gaussian mixture models (DPGMMs). Recently, von Mises-Fisher mixture models (VMMs) have also been gaining popularity in modelling high-dimensional unit-normalized features such as text documents and gene expression data. VMMs are potentially more efficient in modeling certain speech representations such as i-vector data when compared to the GMM-based models, as they work with unit-normalized features based on cosine distance. The current work investigates the applicability of Dirichlet process VMMs (DPVMMs) for i-vector-based speaker clustering and verification, showing that they indeed show superior performance in comparison to DPGMMs in the tasks. In addition, we introduce an implementation of the DPVMMs with variational inference that is publicly available for use.

***Index Terms*—** Non-parametric methods, speaker clustering, unsupervised learning, variational inference, von Mises-Fisher mixtures

1. INTRODUCTION

Dirichlet process mixture models (DPMM) [1] are nonparametric Bayesian approaches [2] that can determine model size based on data without explicit model comparison. Such approaches are applicable when the number of clusters represented in the observed data is not available as a priori information. For instance, zero-resource speech processing systems [3, 4] and speaker diarisation systems [5] use clustering methods to find structure in unlabelled data. DPMM approaches are applicable to large datasets when the posterior distribution over model parameters and cluster assignments is calculated with methods based on variational inference [6]. The approaches are also flexible as observation model can be chosen based on the data to be clustered. The most common approaches are Gaussian DPMMs, called DPGMMs. However, when observations can be modelled as directional data, a von Mises–Fisher (VMF) [7] distribution would be a more natural choice as it compares observations based on cosine distance that is invariant to magnitude.

VMF mixture models (VMM) have been used to cluster high-dimensional observations such as text documents [8, 9, 10] and gene expression data [8, 11]. VMMs have also been used to find speaker clusters based on utterances represented as speaker-adapted GMM mean supervectors [12]. The previous works have studied both maximum likelihood paremeter estimation [8] and approaches that model VMM parameters and cluster assignments as unobserved random variables [9, 10, 11, 13]. Furthermore, nonparametric extensions have been proposed to model directional data with an infinite mixture model as follows: Bangert et al. [14] proposed a sampling method to estimate the posterior distribution over DPVMM parameters on low-dimensional observation data, Straub et al. [15] used small-variance analysis to derive a deterministic method that can be interpreted as a nonparametric extension to spherical k-means, and Batmanghelich et al. [16] proposed stochastic variational inference (SVI) [17] to estimate a hierarchical DPVMM.

The current work focusses on clustering of utterances that are represented as i-vectors [18]. The representation corresponds to a GMM mean supervector mapped into a lower-dimensional factor domain and captures differences between speakers in the vector direction. Hence applications like speaker diarisation use cosine distance to compare i-vectors [19, 20]. The i-vector representation is also common in speaker verification, where an i-vector calculated based on test utterance is compared to a target i-vector to determine whether it corresponds to the same speaker [21]. While the comparison can be based on cosine distance between test and target i-vectors, most speaker verification systems utilise additional normalisation to model variation between sessions. Examples include probabilistic linear discriminant analysis (PLDA) [22] that partitions variation between i-vectors into within-speaker and between-speaker variation. Here speaker-labelled data is needed to estimate the within-speaker and between-speaker compensation parameters, or when labelled data is not available, i-vectors can be clustered in an unsupervised manner [23].

The experiments conducted in this work evaluate variational DPVMM and DPGMM approaches on 600-dimensional i-vector data that could aid PLDA parameter estimation. More specifically, we investigate 1) whether the non-parametric methods can compete with traditional methods (here: k-means with cosine distance; [19]) that divide observation into clusters based on a priori information on the number of classes and 2) how DPVMM compares to the more common DPGMM when evaluated on i-vector data. We believe one reason GMM-based approaches have been favoured in previous works are the computational difficulties related to DPVMM estimation, since otherwise VMM-based approaches are better suited to model directional data that arises in some speech applications. Therefore, as a third contribution, we provide a MATLAB implementation for variational DPGMM/DPVMM estimation.

In this paper, the variational method used in DPVMM posterior estimation is based on the method proposed in [6], but as the variational lower bound does not have a closed-form solution, it is approximated as proposed in [11]. In contrast, DPGMM posterior is computed with the variational method introduced in [6], but as GMMs are not well-suited to model i-vectors, the observations are first mapped into low-dimensional features with principal component analysis (PCA) as proposed in [24]. The evaluation conducted in this work focusses on comparison between the estimated cluster solutions and true speaker classes, but experiments

were also conducted with a PLDA-based speaker verification system [25] to evaluate cluster solutions in an application context.

2. METHODS

2.1. Dirichlet process mixture models

A Dirichlet process (DP) is [26] uniquely defined by a base distribution H and a concentration parameter α. DPs can be combined with an observation model to construct Dirichlet process mixture models [1] where the DP functions as a prior over the model parameters. The current work studies DPMMs constructed as follows. The cluster or mixture component that generated observation $\boldsymbol{x}_n$ is indicated with an unobserved variable z_n. These are modelled as samples from a multinomial distribution $\boldsymbol{\pi}$ that is constructed based on the stick-breaking process [27]:

$$\pi_k = v_k \prod_{i=1}^{k-1}(1 - v_i), \tag{1}$$

where v_k are stick proportions with a beta distribution $Beta(1, \alpha)$. This means that the assignment and observation probabilities associated with observation n can be expressed as

$$p(z_n = k|\boldsymbol{v}) = \prod_{k=1}^{\infty}(1 - v_k)^{\mathbf{1}[z_n>k]} v_k^{\mathbf{1}[z_n=k]}, \tag{2}$$

$$p(\boldsymbol{x}_n|z_n, \boldsymbol{\phi}) = \prod_{k=1}^{\infty} p(\boldsymbol{x}_n|\boldsymbol{\phi}_k)^{\mathbf{1}[z_n=k]}, \tag{3}$$

where $\boldsymbol{\phi}_k$ are parameter vectors with prior distribution H. The parameter distributions used in this work are discussed in Sections 2.1.1–2.1.2.

2.1.1. DPVMM

The observations $\boldsymbol{x}_n$ considered in this work are i-vectors which can be modelled as directional data and compared based on cosine distance [18, 19, 20]. The most common distribution model used with directional data is the von Mises–Fisher (VMF) distribution [7]. The distribution parameters include mean direction $\boldsymbol{\mu}$ ($||\boldsymbol{\mu}|| = 1$) and concentration parameter $\lambda \geq 0$. The observation probabilities are calculated as

$$p(\boldsymbol{x}|\boldsymbol{\phi}) = \frac{\kappa^{D/2-1}}{(2\pi)^{D/2} I_{D/2-1}(\lambda)} \exp(\lambda \boldsymbol{\mu}^{\mathsf{T}} \boldsymbol{x}) \tag{4}$$

where $I_\nu(u)$ denotes the modified Bessel function of the first kind and order ν [28]. $I_\nu(u)$ does not have a closed-form representation, which makes VMF parameter estimation and moment calculation complicated. Alternative parameter estimates are discussed in [8, 29] and numerical issues in parameter estimation and likelihood calculation due to $I_\nu(u)$ are discussed in [30].

The current work studies VMF as an observation model in DPMM. This means that the component-conditioned observation probabilities $p(\boldsymbol{x}_n|\boldsymbol{\phi}_k)$ in Equation (3) are modelled as VMF distributions with mean direction $\boldsymbol{\mu}_k$ and concentration parameter λ_k. The parameters are modelled as unobserved random variables with prior distribution $p(\boldsymbol{\phi}_k)$ (distribution H). The most common approach is to choose a distribution that is conjugate to the likelihood function $p(\boldsymbol{x}_n|\boldsymbol{\phi}_k)$. However, the distribution that could be used as a prior to the concentration parameter has an unknown normalisation term that makes calculations complicated [13]. Hence we choose the alternative prior proposed in [11]. The parameter prior $p(\boldsymbol{\phi}_k) = p(\boldsymbol{\mu}_k|\lambda_k)p(\lambda_k)$, where $p(\boldsymbol{\mu}_k|\lambda_k)$ is a VMF distribution with mean direction $\boldsymbol{m}_0$ and concentration parameter $\beta_0\lambda_k$, and $p(\lambda_k)$ is a gamma distribution with shape parameter a_0 and inverse scale parameter b_0. The prior parameters are modelled as fixed hyperparameters that encode expectations about the observation model. The values utilised in this work are reported in Section 3.3.

2.1.2. DPGMM

When Gaussians are used as the observation model, the mixture model distribution parameters are the mean $\boldsymbol{\mu}_k$ and precision $\boldsymbol{\Lambda}_k$ of the Gaussian, as $\boldsymbol{\phi}_k = \{\boldsymbol{\mu}_k, \Lambda_k\}$. The prior, H, is chosen as the conjugate distribution: the normal Wishart distribution (NW($\mu_0, \kappa_0, \psi_0, \nu_0$)) [31].

2.2. Variational inference

The complete model derived in the previous section includes latent variables v_k and $\boldsymbol{\phi}_k$ that are associated with mixture components k and z_n that are associated with observations n. Since the posterior distribution over latent variables in DPMM cannot be determined in closed form, inference relies on approximate methods. The current work focusses on variational methods that approximate the posterior distribution with a tractable distribution called the variational distribution. The variational distribution is chosen so that an evidence lower bound (ELBO) can be evaluated under the variational model, and the variational distribution parameters are determined as parameters that maximise the bound [32]. Hence variational methods convert the intractable inference problem into a conventional optimisation problem.

The DPMM approaches evaluated in this work use the variational distribution proposed in [6]. The latent variables v_k, $\boldsymbol{\phi}_k$, and z_n are assumed independent and the stick-breaking representation is truncated at truncation limit T so that the complete variational distribution

$$q(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{\phi}) = \prod_{n=1}^{N} q(z_n) \prod_{k=1}^{T-1} q(v_k) \prod_{k=1}^{T} q(\boldsymbol{\phi}_k), \tag{5}$$

where $q(z_n)$ are multinomial distributions, $q(v_k)$ are beta distributions, and $q(\boldsymbol{\phi}_k)$ have the same parametric form as prior distributions $p(\boldsymbol{\phi}_k)$. When observations are modelled as DPGMM, the distribution parameters can be determined with a coordinate ascent method as proposed in [6]. In practice, coordinate ascent under the variational model in Equation (5) iterates between updates to maximise ELBO with respect to $q(z_n)$ when $q(\boldsymbol{v})$ and $q(\boldsymbol{\phi})$ are fixed and updates to maximise ELBO with respect to the component-conditioned variational distribution parameters when $q(z_n)$ are fixed.

When the observations are modelled as a DPVMM, direct optimisation is not possible because ELBO includes $E\{\ln I_{D/2-1}(\lambda_k)\}$ which does not have a closed-form expression under the variational distribution $q(\lambda_k)$. A common solution, used for example in [9], is to include the concentration parameters λ_k as hyperparameters so that the troublesome function $f(x) = \ln I_\nu(x)$ is constant with respect to the variational model parameters and $E\{f(x)\} = f(x)$. The current work utilises the alternative solution proposed in [11]. Here the function $f(x)$ is substituted with upper and lower bounds so that coordinate ascent can be used to derive variational updates that optimise a lower bound to ELBO [11] (see [33] for details). Alternative solutions also include sampling the concentration parameter [10, 16].

3. EXPERIMENTAL SETUP

3.1. Evaluation

The speaker clustering experiments conducted in this work use DPVMM, DPGMM and k-means with cosine distance to partition i-vector data into clusters. The cluster solutions are evaluated (1) based external measures that compare estimated partition to true speaker classes and (2) based on application performance. The external measures used in this work are adjusted Rand index (ARI) [34] and an accuracy (ACC) measure calculated as geometric mean between average cluster purity (ACP) and average speaker purity (ASP) [5, Section 7.1]. In addition, cluster solutions are evaluated as substitute to labelled parameter estimation data in PLDA-based speaker verification [22]. The evaluation was conducted with the standard PLDA and evaluation measures implemented in [25]. PLDA model trained based on labelled data is included in the evaluation as reference. The evaluation measures include equal error rate (EER) and minimum decision cost function (DCF). EER is calculated at an operation point t where false acceptance and false rejection errors occur at equal rate, whereas DCF emphasises false acceptance errors, $\text{DCF}(t) = \text{FRR}(t) + 100 \times \text{FAR}(t)$.

3.2. Data

The experiments were conducted on the NIST SRE 2014 development partition that contains 600-dimensional i-vector features extracted from 4958 speakers [21]. Experiments conducted with DPVMM and k-means used observations that were normalised to unit length whereas experiments conducted with DPGMM used observations that were compressed into $D = \{50, 10\}$ dimensions with PCA. The clustering methods were evaluated on test datasets that included $M = \{10, 100, 500, 650\}$ speakers with most observations. The datasets included $N = \{474, 2931, 10495, 12786\}$ observations. The speaker verification experiments were conducted on the complete dataset partitioned as follows. The system parameters were determined based on the dataset that included 650 speakers with most data. The enrolled speaker set included 1031 speakers. The speakers included in this set had recorded 10–15 utterances. 5 utterances were used to model the speaker, as in [21, 23] and the rest were used test data. 3277 impostors with 1–10 utterances were also represented in the test dataset.

3.3. Methods

The component-conditioned distribution parameters in DPVMM and DPGMM were modelled as random variables with prior distributions discussed in Section 2.1.1–2.1.2. The prior parameters in DPVMM were set as follows. The mean direction in prior distribution $p(\boldsymbol{\mu}_k|\lambda_k)$ was set to the observed mean $\boldsymbol{\mu}_0 = \sum_n \boldsymbol{x}_n / ||\sum_n \boldsymbol{x}_n||$ while β_0 was set to 0.01 to indicate a low trust on the prior mean direction. The prior distribution for concentration parameter was chosen to favour unconcentrated solutions: the gamma distribution shape parameter was set to 1 and the inverse scale parameter was set $1/50$. DPGMM component means and covariances were modelled with prior distribution $\text{NW}(\mu_0, \kappa_0, \psi_0, \nu_0)$ whose parameters were set similar to those recommended in [31]: μ_0 and ψ_0 were set to the dataset mean and precision, $\kappa_0 = 1$, and $\nu_0 = D + 2$. The concentration parameter α was fixed to 1.

DPVMM and DPGMM posteriors were approximated with a truncated variational model, and experiments were conducted with truncation limit $T = 5000$ unless otherwise mentioned. The model parameters were initialised based on observations assigned to clusters at random or based on k-means solution, and variational updates were continued until the difference between evidence lower bound in consecutive iterations did not exceed 0.001 per observation or when 500 iterations were reached. For evaluation, observations were assigned to clusters and labelled based on variational distribution probabilities $q(z_n)$. Since optimisation can converge to local maxima, experiments were repeated 20 times.

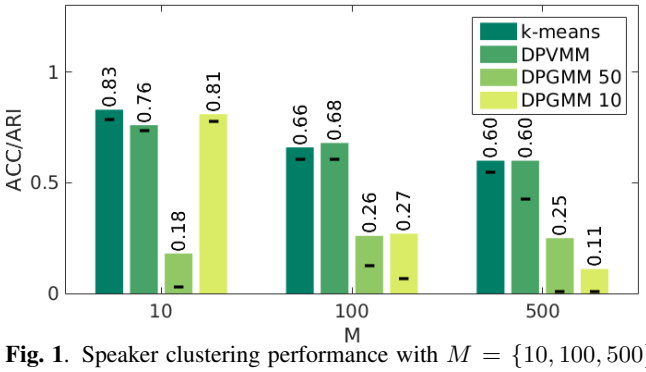

Fig. 1. Speaker clustering performance with $M = \{10, 100, 500\}$ speakers, based on accuracy (bars) and ARI (-).

The experiments were conducted in MATLAB with variational DPVMM and DPGMM implementations[1] that are made available under an open-source license. Related to the DPVMM implementation, we note that while the first-order approximation proposed in [11] (Section 2.2) makes the evidence lower bound and parameter updates calculable, $I_\nu(u)$ is still an issue when calculations are implemented in finite precision. The implementation evaluated in this work uses both MATLAB function `besseli` and an approximation based on the simple bound proposed in [35]. The bound is substituted when `besseli` does not produce a finite value. The speaker verification experiments were conducted with FASTPLDA MATLAB implementation [25]. PLDA model was used with 300-dimensional speaker and channel latent variables and 20 iterations were used in training.

4. RESULTS

The mean accuracy and ARI calculated based on 20 random initialisations for 10, 100 and 500 speakers are reported in Figure 1. DPVMM and k-means were evaluated on full i-vector features, and DPGMM on features compressed to 50 and 10 dimensions using PCA. The k-means solutions were estimated with $K = M$ to have a baseline that uses a priori information regarding the data set. Evaluation based on accuracy indicates that DPVMM solutions are comparable to k-means solutions, whereas DPGMM solutions are less accurate than k-means and DPVMM solutions in most conditions. The exception is with the 10 speaker case where DPGMM with 10-dimensional features resulted in competitive accuracies. The feature dimension better suited to DPGMM depends on the dataset: 1) 10-dimensional features resulted in better DPGMM clusters than 50-dimensional features in the 10-speaker condition, 2) the 50-dimensional features resulted in better DPGMM clusters in the conditions with 500 speakers and 3) their performance very close in the 100 speaker case. ARI has a similar trend but punishes DPMMs for excess clusters.

Table 1 shows the results for the speaker verification experiments, where the methods were used to cluster a 650-speaker

[1] http://github.com/shreyas253/variational_NP_BMM/

Table 1. Speaker clustering and verification performance.

	ACP	ASP	ACC	ARI	EER	DCF
labelled	1.00	1.00	1.00	1.00	1.67	0.35
k-means	0.61	0.60	0.60	0.55	2.70	0.43
DPVMM	0.52	0.58	0.55	0.40	2.53	0.46
DPGMM	0.13	0.43	0.24	0.01	5.77	0.64

dataset that was used to train PLDA parameters. DPGMM results are reported with PCA $D = 50$ which was better than $D = 10$ for large M. As expected, PLDA trained on clustered data does not achieve as good performance as the labelled supervised case. When unsupervised approaches are used, speaker verification results improve when the cluster solution is improved, so that the best results are achieved with k-means and DPVMM clustered data. However, while k-means clusters are evaluated as better than DPVMM clusters, DPVMM clustered data resulted in lower EER.

The previous experiments used DPVMM and DPGMM with random initialisation and fixed truncation limit. Experiments with DPVMM based on random and k-means initialisation at $T = \{1000, 2000, 5000\}$ are reported in Table 2. Experiments were conducted on the 650-speaker dataset, and the truncation limit T was also used as K in k-means initialisation. The results show that DPVMM with k-means initialisation (Table 2 (c)) resulted in better cluster solutions than DPVMM with random initialisation (Table 2 (b)). However, k-means initialisation also makes DPVMM results depend on T so that the best results are achieved when $T = 1000$, while cluster solutions calculated based on random initialisation are comparable across T. K-means initialisation was also tested on DPGMM, but in this case the DPGMM did not update the solution found by the k-means (Table 2 (a)).

5. DISCUSSION AND CONCLUSIONS

This paper compared variational DPVMM and DPGMM in speaker clustering and verification task using i-vector features. The comparison indicates that despite the approximations required to make variational inference tractable, DPVMM can produce more accurate speaker clusters than DPGMM. Moreover, while the DPMM approaches had no information on how many speakers are represented in the data, DPVMM solutions were able to compete with the k-means -based reference solutions calculated using information on the correct number of speakers. While DPVMM generally performed well, DPGMM solutions were also comparable to k-means solutions in the 10-speaker condition when 10-dimensional features were used. However in the other conditions, DPGMM could not model the data properly, and better results were achieved with DPVMM and k-means. We also observed that DPVMM performance can be improved with k-means initialisation, especially when K is close to M.

DPVMMs are expected to model i-vector data better than DPGMMs since i-vectors are directional (see also [12]), but the experiments show that this is not the case in the 10-speaker condition when DPGMM is applied on 10-dimensional data. This indicates that DPGMMs can model speaker clusters when i-vectors can be mapped to low-dimensional features without too much information loss. The observation is in line with previous work, as Shum et al. [24] also described that GMM-based solutions were better than or comparable to VMM-based solutions in i-vector-based speaker-diarisation task with 3-dimensional features and less than 10 speakers. However, low-dimensional features cannot capture differences between speakers in larger datasets such as datasets needed to train PLDA parameters. GMM-based models are not well-suited to model high-dimensional data, unlike VMMs which handle this well since they work on cosine distance.

Table 2. Speaker clustering performance of (a) k-means, and DPVMM with (b) random and (c) k-means initialisation when the number of k-means clusters equals to the truncation limit ($K = T$.)

	K/T	ACP	ASP	ACC	ARI	EER	DCF
(a)	1000	0.75	0.55	0.64	0.60	2.63	0.40
	2000	0.82	0.38	0.55	0.46	2.76	0.38
	5000	0.91	0.17	0.39	0.20	2.95	0.36
(b)	1000	0.55	0.55	0.55	0.41	2.72	0.46
	2000	0.54	0.58	0.56	0.43	2.57	0.45
	5000	0.52	0.58	0.55	0.40	2.54	0.47
(c)	1000	0.60	0.66	0.63	0.58	2.47	0.41
	2000	0.53	0.64	0.58	0.51	2.42	0.41
	5000	0.50	0.61	0.55	0.43	2.38	0.44

Related to comparison between DPGMM and DPVMM approaches, the DPGMM experiments reported here used full covariance matrices as proposed in [24]. To evaluate whether this is a problem when less data is available per speaker, we also experimented with more constrained covariances. While diagonal or spherical covariances did not improve the performance, DPGMM with task-optimised covariance parameters was able to outperform DPVMM. A fixed concentration parameter could also be used in the DPVMM, in which case there would be no need to use an approximation on the variational lower bound. However, it is not clear how the covariance or concentration parameter should be optimised, especially in case cluster sizes could be expected to vary. The current work thus focussed on more flexible solutions.

While evaluation focussed on comparison between estimated clusters and speaker classes, cluster solutions were also evaluated in a speaker verification task. These results indicated that when labelled data is not available, speaker models can be reasonably estimated based on the k-means or DPVMM solutions. However, we also observed that EER and DCF cannot be predicted based on cluster evaluation measures and that the same clustering solution is not guaranteed to optimise both measures. For example, we observed the best minimum DCF when the 650-speaker parameter estimation data was partitioned into $K = 5000$ clusters with k-means. This is potentially because DCF emphasises false acceptance rate which could be easier to minimise when the observations are oversegmented so that the estimated within-speaker variation is small. Oversegmentation does not occur with DPVMM since the number of clusters is inferred based on the observed data.

The current experiments focussed on i-vector data, but the variational DPVMM updates and MATLAB implementation presented in this work are expected to handle any high-dimensional data that can be unit normalised and clustered based on cosine distance. The variational method can also be easily adapted to work with very large datasets using SVI [17]. DPVMMs should therefore make a potential candidate also for other speech clustering tasks such as the zero-resource systems for under-resourced languages [4].

6. ACKNOWLEDGEMENTS

This research was funded by the Academy of Finland project titled Computational Modeling of Language Acquisition.

7. REFERENCES

[1] C. E. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Ann. Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[2] S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *J. Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.

[3] O. Räsänen, G. Doyle, and M. C. Frank, "Unsupervised word discovery from speech using automatic segmentation into syllable-like units," in *Proc. INTERSPEECH*, 2015, pp. 3204–3208.

[4] H. Kamper, A. Jansen, and S. Goldwater, "A segmental framework for fully-unsupervised large-vocabulary speech recognition," *CoRR*, vol. abs/1606.06950, 2016. [Online]. Available: http://arxiv.org/abs/1606.06950

[5] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Communication*, vol. 54, no. 10, pp. 1065–1103, 2012.

[6] D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[7] K. V. Mardia and P. E. Jupp, *Directional Statistics*. John Wiley & Sons, 2009, vol. 494.

[8] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using von Mises-Fisher distributions," *J. Machine Learning Research*, vol. 6, pp. 1345–1382, 2005.

[9] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney, "Spherical topic models," in *Proc. ICML*, 2010, pp. 903–910.

[10] S. Gopal and Y. Yang, "Von Mises-Fisher clustering models," in *Proc. ICML*, 2014, pp. 154–162.

[11] J. Taghia, Z. Ma, and A. Leijon, "Bayesian estimation of the von-Mises Fisher mixture model with variational inference," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1701–1715, 2014.

[12] H. Tang, S. M. Chu, and T. S. Huang, "Generative model-based speaker clustering via mixture of von Mises-Fisher distributions," in *Proc. ICASSP*, 2009, pp. 4101–4104.

[13] G. Nuñez-Antonio and E. Gutiérrez-Peña, "A Bayesian analysis of directional data using the von Mises–Fisher distribution," *Communications in Statistics - Simulation and Computation*, vol. 34, no. 4, pp. 989–999, 2005.

[14] M. Bangert, P. Hennig, and U. Oelfke, "Using an infinite von Mises-Fisher mixture model to cluster treatment beam directions in external radiation therapy," in *Proc. ICMLA*, 2010, pp. 746–751.

[15] J. Straub, T. Campbell, J. P. How, and J. W. Fisher, "Small-variance nonparametric clustering on the hypersphere," in *Proc. CVPR*, 2015, pp. 334–342.

[16] K. Batmanghelich, A. Saeedi, K. Narasimhan, and S. Gershman, "Nonparametric spherical topic modeling with word embeddings," in *Proc. ACL*, 2016, pp. 537–542.

[17] M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley, "Stochastic variational inference." *J. Machine Learning Research*, vol. 14, no. 1, pp. 1303–1347, 2013.

[18] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[19] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting inter-conversation variability for speaker diarization," in *Proc. INTERSPEECH*, 2011, pp. 945–948.

[20] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 217–227, 2014.

[21] C. S. Greenberg, D. Bansé, G. R. Doddington, D. Garcia-Romero, J. J. Godfrey, T. Kinnunen, A. F. Martin, A. McCree, M. Przybocki, and D. A. Reynolds, "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Proc. Odyssey*, 2014, pp. 224–230.

[22] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inderences about identity," in *Proc. ICCV*, 2007, pp. 1–8.

[23] E. Khoury, L. El Shafey, and S. Marcel, "Hierarchical speaker clustering methods for the NIST i-vector challenge," in *Proc. Odyssey*, 2014, pp. 254–259.

[24] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[25] A. Sizov, K. A. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Proc. S+SSPR*, 2014, Matlab code: https://sites.google.com/site/fastplda/.

[26] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[27] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, no. 2, pp. 639–650, 1994.

[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. United States Department of Commerce, National Bureau of Standards, 1972.

[29] S. Sra, "A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$," *Computational Statistics*, vol. 27, no. 1, pp. 177–190, 2012.

[30] K. Hornik and B. Grün, "movMF: an R package for fitting mixtures of von Mises-Fisher distributions," *J. Statistical Software*, vol. 58, no. 10, 2014.

[31] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

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

[33] U. Remes, S. Seshadri, and O. Räsänen, "Approximate inference in Dirichlet process von Mises-Fisher mixture models," 2016. [Online]. Available: https://github.com/shreyas253/

[34] L. Hubert and P. Arabie, "Comparing partitions," *J. Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[35] D. E. Amos, "Computation of modified Bessel functions and their ratios," *Math. Comp.*, vol. 28, no. 125, pp. 239–251, 1974.