DIRICHLET PROCESS MIXTURE MODELS FOR CLUSTERING I-VECTOR DATA

Aalto University
School of Electrical Engineering

ACADEMY OF FINLAND

Shreyas Seshadri, Ulpu Remes and Okko Räsänen

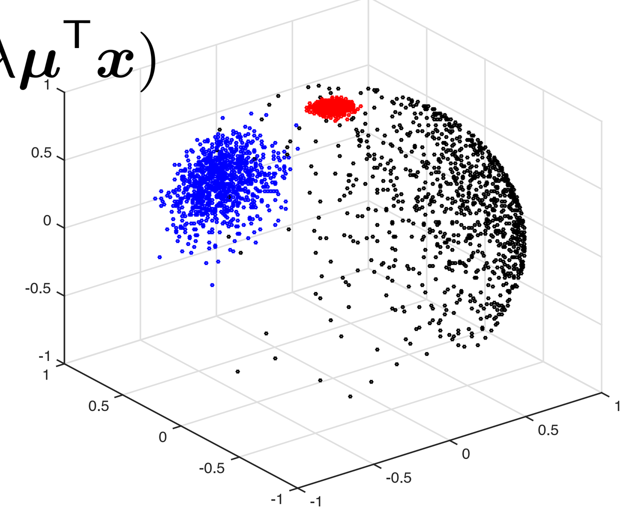
Department of Signal Processing and Acoustics, Aalto University, Finland

Goal of the study

Non-parametric Bayesian methods have recently gained popularity in unsupervised learning. They 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 Dirichlet process Gaussian mixture models (DPGMMs). Recently, von Mises-Fisher mixture models (VMMs) have also gained popularity in modelling high-dimensional unit-normalized features such as text documents and gene expression data. VMMs are potentially more efficient than GMMs in modeling certain speech representations such as i-vector data as they work with unit-normalized features based on cosine distance. We investigate the applicability of DPVMMs for i-vector-based speaker clustering and verification.

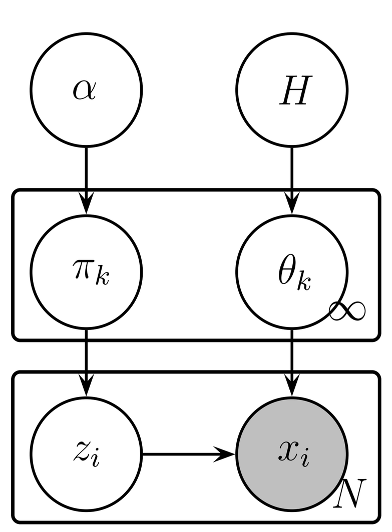
Von Mises-Fisher Distribution

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



- x is unit normalized D dimensional data and φ is the VMF model.
- Parameters – μ is the mean and λ is the concentration parameter
- $I_\nu(u)$ is the Bessel function of the first kind with order ν

Dirichlet Process Mixture Models



$$\pi \mid \alpha \sim GEM(\alpha)$$

$$\theta_k \mid H \sim H$$

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta = \theta_k)$$

$$z_i \mid \pi \sim \pi$$

$$x_i \mid \{\theta_k\}_{k=1}^{\infty}, z_i \sim F(\theta_{z_i})$$

- A Dirichlet process is a distribution over probability measures on a measurable space Θ
- Uniquely defined by base distribution H and concentration parameter α, written as $G \sim DP(\alpha, H)$.

Gaussian

When the mixture component is Gaussian i.e. $\theta_k = \{\mu_k, \Sigma_k\}$ representing the mean and variance parameters. We consider the conjugate prior i.e. The Normal Inverse Wishart distribution.

Von Mises-Fisher model

VMF models i.e. $\theta_k = \{\mu_k, \lambda_k\}$ represent the mean and concentration parameters. The prior $p(\theta_k) = p(\mu_k|\lambda_k)p(\lambda_k)$, where $p(\mu_k|\lambda_k)$ is modelled by a VMF distribution and $p(\lambda_k)$ modelled by a Gamma distribution

Variational Inference

- Approximate the analytically intractable posterior with a tractable distribution called the variational distribution.
- 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.
- Typically done by making some independence assumptions.
- Similar to the expectation–maximisation (EM) algorithm that iterates between finding the probabilities of z_i (called responsibilities) based on current the model and updating model parameters based the current responsibilities

Experiments

- The clustering experiments were conducted using DPVMM, DPGMM and k-means with cosine distance on the NIST SRE 2014 development partition that contains 600-dimensional i-vector features extracted from 4958 speakers.
- DPVMM and k-means used observations that were normalised to unit length; and 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.
- Speaker verification experiments were conducted on the complete dataset using using FASTPLDA, with PLDA parameters determined on the 650 speaker dataset.

Evaluation

Speaker Clustering

Accuracy as geometric mean of average cluster and speaker purities:

$$ACP = \frac{1}{N} \sum_i \frac{(\sum_j n_{ij}^2)}{n_i} \qquad ASP = \frac{1}{N} \sum_j \frac{(\sum_i n_{ij}^2)}{n_j}$$

where n_{ij}, n_i, n_j, N are the number of utterances in cluster i spoken by speaker j, the number of utterances in cluster I, number of utterances spoken by speaker j and the total number of speakers respectively.

Speaker Verification

- *Equal Error Rate:* calculated at an operation point t where false acceptance and false rejection errors occur at equal rate
- *Minimum Decision Cost Function:* Calculated at point t where DCF(t) is minimum

$$\mathrm{DCF(t) = FRR(t) + 100 \times FAR(t)}$$

Results

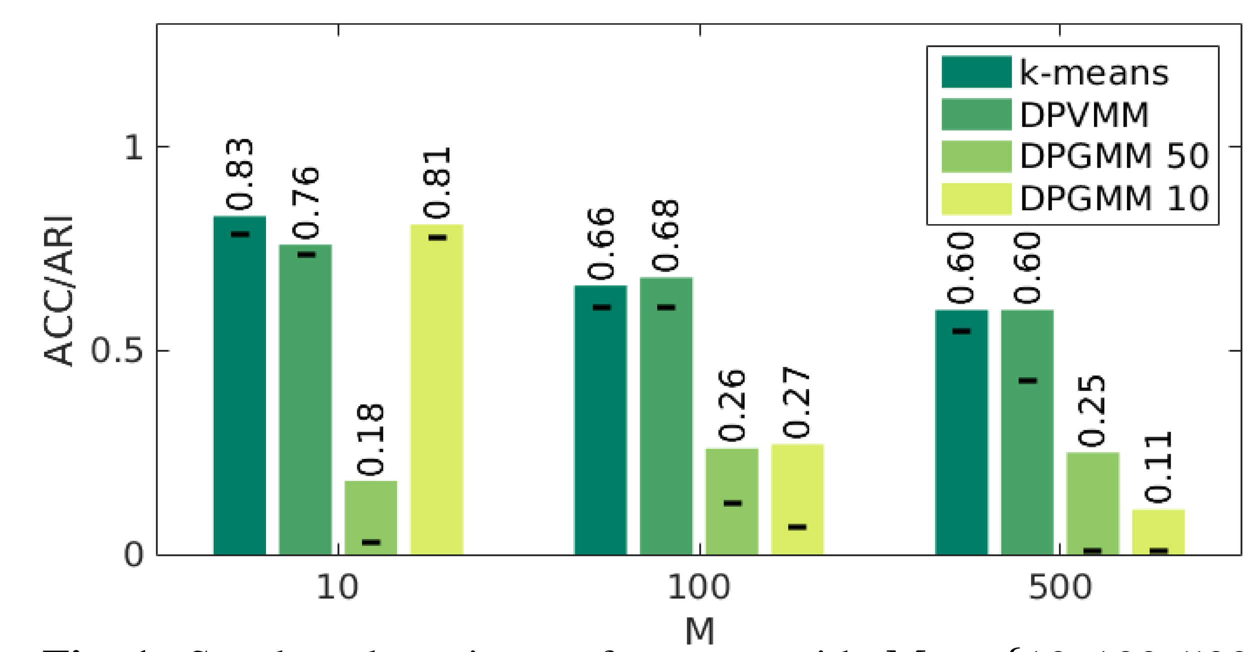


Fig. 1. Speaker clustering performance with $M = \{10, 100, 500\}$ speakers, based on accuracy (bars) and ARI (-).

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

Conclusions

- The comparison indicates that DPVMM can produce more accurate speaker clusters than DPGMM.
- Even though the Bayesian methods had no knowledge of the correct number of speakers, DPVMM solutions were still able to compete with the k-means with K corresponding to the correct number of speakers.
- DPGMM results are comparable to DPVMM results in the 10-speaker case where low-dimensional features can be used. This indicates that DPGMMs can model speaker clusters when i-vectors can be mapped to low-dimensional features without too much information loss.
- For speaker verification, speaker models can be reasonably estimated based on the k-means or DPVMM solutions. Also note that EER and DCF are not monotonically related to the clustering measures.
- The implementation of the DPVMM/DPGMM is available on http://github.com/shreyas253/variational_NP_BMM/

References

1. D. M. Blei and M. I. Jordan, "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.
2. 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.
3. 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.
4. 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.