Interspeech 2022
18-22 September 2022, Incheon, Korea

Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings

Niko Brümmer[1], *Albert Swart*[1], *Ladislav Mošner*[2], *Anna Silnova*[2], *Oldřich Plchot*[2], *Themos Stafylakis*[3], *Lukáš Burget*[2]

[1]Phonexia, South Africa
[2]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czechia
[3]Omilia - Conversational Intelligence, Athens, Greece

niko.brummer@gmail.com, adswart@gmail.com

Abstract

In speaker recognition, where speech segments are mapped to embeddings on the unit hypersphere, two scoring backends are commonly used, namely cosine scoring or PLDA. Both have advantages and disadvantages, depending on the context. Cosine scoring follows naturally from the spherical geometry, but for PLDA the blessing is mixed—length normalization Gaussianizes the between-speaker distribution, but violates the assumption of a speaker-independent within-speaker distribution. We propose PSDA, an analogue to PLDA that uses Von Mises-Fisher distributions on the hypersphere for both within and between-class distributions. We show how the self-conjugacy of this distribution gives closed-form likelihood-ratio scores, making it a drop-in replacement for PLDA at scoring time. All kinds of trials can be scored, including single-enroll and multi-enroll verification, as well as more complex likelihood-ratios that could be used in clustering and diarization. Learning is done via an EM-algorithm with closed-form updates. We explain the model and present some first experiments.

Index Terms: speaker recognition, PSDA, Von Mises-Fisher

1. Introduction

Probabilistic *linear* discriminant analysis (PLDA) [1, 2], is a popular backend for scoring speaker recognition embeddings in $\mathbb{R}^d$, following [3, 4]. However, [5] showed that length-normalizing the embeddings onto the unit sphere, $\mathbb{S}^{d-1}$ has a Gaussianizing effect that improves accuracy and this has been standard practice ever since. One disadvantage of the length-normalization is that within-speaker variability is squashed in the radial direction, making it *speaker-dependent*, in violation of the PLDA assumption of a constant within-class distribution. Moreover, given a flexible, discriminatively trained embedding extractor, it is often found that cosine scoring (dot products between embeddings in $\mathbb{S}^{d-1}$) outperforms PLDA, especially when the test data is in domain, e.g. [6, 7]. We propose to enrich the field with a new backend that is intermediate between cosine scoring and PLDA: Probabilistic *spherical* discriminant analysis (PSDA) uses Von Mises-Fisher (VMF) distributions on $\mathbb{S}^{d-1}$ in place of Gaussians. A Python implementation is available here.[1]

To the best of our knowledge, the only related work in speaker recognition is [8], where a VMF mixture was used for speaker clustering of length-normalized i-vectors. For face recognition, VMF mixtures were explored in [9].

[1] https://github.com/bsxfan/PSDA

2. Von Mises-Fisher distribution

When embeddings in Euclidean space, $\mathbb{R}^d$ are length-normalized, they are projected onto the *unit hypersphere*:[2]

$$\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| = 1\} \tag{1}$$

When $\mathbf{x} \in \mathbb{S}^{d-1}$, it is *on* the sphere. If $\|\mathbf{x}\| < 1$, it is *inside*. To construct the PSDA model, we replace the Gaussians in PLDA with Von Mises-Fisher (VMF) distributions. The density for $\mathbf{x} \in \mathbb{S}^{d-1}$ is [10]:

$$\mathcal{V}(\mathbf{x} \mid \boldsymbol{\mu}, \kappa) = K_d C_\nu(\kappa) e^{\kappa \boldsymbol{\mu}'\mathbf{x}} \quad \text{where} \quad \nu = \frac{d}{2} - 1 \tag{2}$$

The parameters are the *mean direction*, $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ and the *concentration*, $\kappa \geq 0$. While K_d depends only on the dimension and is of no further interest here,[3] the other normalization factor is all-important for our purposes:

$$C_\nu(\kappa) = \frac{\kappa^\nu}{I_\nu(\kappa)} = \Big(\sum_{i=0}^{\infty} \frac{\kappa^{2i}}{2^{2i+\nu}\, i!\, \Gamma(i+\nu+1)}\Big)^{-1} \tag{3}$$

where I_ν is the modified Bessel function of the first kind (Bessel-I) of order ν. Note $I_\nu(\kappa) \geq 0$ and $I_\nu(0) = 0$ for $\nu > 0$, but $I_0(0) = 1$. The derivative can be expressed as [11]:

$$\frac{\partial}{\partial \kappa} I_\nu(\kappa) = \frac{\nu}{\kappa} I_\nu(\kappa) + I_{\nu+1}(\kappa) \tag{4}$$

which shows it is monotonic rising. As a function of κ, $C_\nu(\kappa)$ is positive and strictly monotonic decreasing,[4] and $\lim_{\kappa \to 0} C_\nu(\kappa) = 2^\nu \Gamma(\nu+1)$. The concentration parameter, κ is roughly analogous to precision in the normal distribution. For smaller κ, the distribution is more widely spread, until at $\kappa = 0$ it gives the uniform hypersphere distribution. For larger κ, the distribution concentrates more tightly around $\boldsymbol{\mu}$. It should be noted that $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ is *not the expected value*, which instead is at [10]:

$$\langle \mathbf{x} \rangle = \rho(\kappa)\boldsymbol{\mu}, \quad \text{where} \quad 0 \leq \rho(\kappa) = \frac{I_{\nu+1}(\kappa)}{I_\nu(\kappa)} < 1 \tag{5}$$

[2] Do not confuse *sphere* with *ball*: $\mathbb{S}^{d-1}$ is the surface of the ball. Euclidean norm is denoted $\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}}$.

[3] Different authors (e.g. [10] vs [11]) use different expressions for K_d, depending on the reference measure for the density. K_d is analogous to the usually irrelevant $(2\pi)^{d/2}$ factor in the multivariate normal density.

[4] On a log-log plot it is relatively flat for $0 \leq \kappa << \sqrt{\nu+1}$ and then plunges dramatically for large κ.

10.21437/Interspeech.2022-731

where $\langle \mathbf{x} \rangle$ is inside the sphere, not on it. The norm, $\left\| \langle \mathbf{x} \rangle \right\| = \rho(\kappa)$ is strictly increasing w.r.t. κ, where $\lim_{\kappa \to 0} \rho(\kappa) = 0$ and $\lim_{\kappa \to \infty} \rho(\kappa) = 1$. The empirical mean, say $\bar{\mathbf{x}}$, of a cluster of points on the hypersphere has the same behaviour: $\bar{\mathbf{x}}$ is inside the sphere and moves closer to it ($\|\bar{\mathbf{x}}\|$ increases towards 1), as the cluster becomes more concentrated.

2.1. Maximum likelihood parameter estimates

Given data set in $\mathbb{S}^{d-1}$, say $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^n$, assumed to have been sampled *iid* from $\mathcal{V}(\boldsymbol{\mu}, \kappa)$, the maximum-likelihood (ML) estimate of the parameters is obtained by maximizing the log-likelihood:

$$\log \prod_{i=1}^{n} \mathcal{V}(\mathbf{x}_i \mid \boldsymbol{\mu}, \kappa) = n \log C_\nu(\kappa) + n\kappa \boldsymbol{\mu}' \bar{\mathbf{x}} + \text{const} \quad (6)$$

where $\bar{\mathbf{x}} = \frac{1}{n} \sum_i \mathbf{x}_i$. The maximum w.r.t. $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ is at:

$$\boldsymbol{\mu}_{ML} = \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|} \quad (7)$$

Inserting this back into (6), we need to maximize:

$$n \log C_\nu(\kappa) + n\kappa \|\bar{\mathbf{x}}\| \quad (8)$$

to find κ. Setting the derivative to zero, using (4), gives [10]:

$$\kappa_{ML} = \rho^{-1}(\|\bar{\mathbf{x}}\|) \quad (9)$$

We used a numerical (derivative-free) rootfinding algorithm[5] to invert ρ. At the ML estimate $\langle \mathbf{x} \rangle = \bar{\mathbf{x}}$. The ML parameters depend *solely* on the sufficient statistic $\bar{\mathbf{x}}$. When $\bar{\mathbf{x}} = \mathbf{0}$, $\boldsymbol{\mu}$ is irrelevant and the likelihood is maximized at $\kappa = 0$, which gives the uniform distribution.

3. The PSDA model

PSDA is constructed much like PLDA [4]. For every speaker we posit a hidden speaker identity variable, $\mathbf{z} \in \mathbb{S}^{d-1}$, having a VMF prior, $\mathcal{V}(\mathbf{z} \mid \boldsymbol{\mu}, b)$, where $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ is the *speaker mean direction* and $b \geq 0$ is the *between-speaker concentration*. Embeddings with low speaker concentration (as spread out as possible), ideally $b = 0$, is required for good accuracy. The simplest variant of PSDA has a uniform between-speaker distribution ($b = 0$ and $\boldsymbol{\mu}$ irrelevant). If however, the speaker distribution is believed to be non-uniform, b and $\boldsymbol{\mu}$ can be learnt from labelled data, as we are accustomed to do with PLDA.

The observed embeddings in $\mathbb{S}^{d-1}$, are supposed to have been generated from speaker-dependent VMF distributions: embeddings from different speakers are independent and those from the same speaker are conditionally independent, given $\mathbf{z}$. If $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^n$ are embeddings of a common speaker, then:

$$P(\mathbf{X} \mid \mathbf{z}) = \prod_t \mathcal{V}(\mathbf{x}_t \mid \mathbf{z}, w) \propto \exp\left[w \mathbf{z}' \bar{\mathbf{x}}\right] \quad (10)$$

where $\bar{\mathbf{x}} = \frac{1}{n} \sum_t \mathbf{x}_t$ and where $w > 0$ is the *within-speaker concentration*. Note the conjugacy: the product of VMF distributions for the observed data doubles as a likelihood function for $\mathbf{z}$, which is also in VMF form.

In summary, the learnable PSDA model parameters are $w, b \in \mathbb{R}$ and $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$. Next we show how to do inference and learning. We start with inference for $\mathbf{z}$, followed by inference for the speaker hypothesis (scoring). Finally, maximum-likelihood learning can be done with an EM-algorithm.

[5] `scipy.optimize.toms748`

3.1. The hidden variable posterior

Given one or more observations, $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^n$, assumed to be of the same speaker, the identity variable posterior is still VMF:

$$\begin{aligned} P(\mathbf{z} \mid \mathbf{X}) &\propto \mathcal{V}(\mathbf{z} \mid \boldsymbol{\mu}, b) \prod_t \mathcal{V}(\mathbf{x}_t \mid \mathbf{z}, w) \\ &\propto \exp\Big[\Big(b\boldsymbol{\mu} + w \sum_t \mathbf{x}_t\Big)' \mathbf{z}\Big] \\ &\propto \mathcal{V}\Big(\mathbf{z} \;\Big|\; \frac{\tilde{\mathbf{z}}}{\|\tilde{\mathbf{z}}\|}, \|\tilde{\mathbf{z}}\|\Big) \end{aligned} \quad (11)$$

where $\tilde{\mathbf{z}} = b\boldsymbol{\mu} + w \sum_t \mathbf{x}_t$. The concentrations, b and w behave in much the same way as the precisions in Gaussian PLDA. But, in Gaussian PLDA [4], the posterior precision is dependent only on the number of observations, while here, the posterior concentration, $\|\tilde{\mathbf{z}}\|$ is data-dependent. If the data all lie in the same quadrant then the more data we have, the more the concentration will grow. But if the data are spread with angles wider than 90 degrees, they can (partially) cancel and the posterior concentration can become arbitrarily small (e.g. if $b = 0$ and there are two antipodal observations, then $\tilde{\mathbf{z}} = \mathbf{0}$). This stands in contrast to the heavy-tailed PLDA model of [12], where larger norms are associated with larger within-speaker variation.

3.2. Scoring

Given a trained PSDA model, with parameters $(w, b, \boldsymbol{\mu})$, we derive a general recipe for computing likelihood-ratio scores. As with PLDA [4], PSDA provides closed-form scores for a variety of verification and clustering trials. Let $\mathbf{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_m\}$ denote a collection of $m \geq 1$ enrollment observations hypothesized to be from a common speaker. $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_n\}$ denotes a collection of $n \geq 1$ test observations from a common (but possibly different) speaker. We want to compute a *likelihood-ratio* (LR) for hypothesis H_1, that all observations come from one common speaker, against hypothesis H_2, that they come from two different speakers:

$$\frac{P(\mathbf{E}, \mathbf{T} \mid H_1)}{P(\mathbf{E}, \mathbf{T} \mid H_2)} = \frac{P(\mathbf{E}, \mathbf{T} \mid H_1)}{P(\mathbf{E} \mid H_1) P(\mathbf{T} \mid H_1)} \quad (12)$$

All RHS factors are marginals, where $\mathbf{z}$ has been integrated out. We can use the conjugacy and the availability of the VMF normalizer (3) to solve these integrals in closed form. Following the derivation in [4], the LR can be rewritten as:

$$\frac{P(\mathbf{E}, \mathbf{T} \mid H_1)}{P(\mathbf{E}, \mathbf{T} \mid H_2)} = \frac{P(\mathbf{z}_0 \mid \mathbf{E}) P(\mathbf{z}_0 \mid \mathbf{T})}{P(\mathbf{z}_0 \mid \mathbf{E}, \mathbf{T}) P(\mathbf{z}_0)} \quad (13)$$

Since the LHS is independent of $\mathbf{z}_0 \in \mathbb{S}^{d-1}$, so is the RHS: all factors of the form $e^{\mathbf{z}_0' \cdots}$ cancel, leaving only the VMF normalization constants to yield our general *closed-form scoring function*:

$$\frac{P(\mathbf{E}, \mathbf{T} \mid H_1)}{P(\mathbf{E}, \mathbf{T} \mid H_2)} = \frac{C_\nu(\|b\boldsymbol{\mu} + w\tilde{\mathbf{e}}\|) C_\nu(\|b\boldsymbol{\mu} + w\tilde{\mathbf{t}}\|)}{C_\nu(\|b\boldsymbol{\mu} + w\tilde{\mathbf{e}} + w\tilde{\mathbf{t}}\|) C_\nu(b)} \quad (14)$$

where $\tilde{\mathbf{e}} = \sum_{t=1}^m \mathbf{e}_t$, and $\tilde{\mathbf{t}} = \sum_{t=1}^n \mathbf{t}_t$.

3.2.1. Relationship with cosine scoring

In the special case when we set the model parameter $b = 0$ and when the enroll and test sets are singletons, $m = n = 1$, there is a close relationship between the PSDA score (14) and

the ubiquitous cosine score. When $b = 0$, the score simplifies to:

$$\frac{P(\mathbf{E}, \mathbf{T} \mid H_1)}{P(\mathbf{E}, \mathbf{T} \mid H_2)} = \frac{C_\nu(w)^2}{C_\nu(w \left\|\tilde{\mathbf{e}} + \tilde{\mathbf{t}}\right\|) \lim_{b\to 0} C_\nu(b)} \tag{15}$$

When $\tilde{\mathbf{e}} = \mathbf{e}_1 \in \mathbb{S}^{d-1}$ and $\tilde{\mathbf{t}} = \mathbf{t}_1 \in \mathbb{S}^{d-1}$, the *cosine score* is the dot product $\langle \tilde{\mathbf{e}}, \tilde{\mathbf{t}} \rangle$, which can be rewritten as:

$$\langle \tilde{\mathbf{e}}, \tilde{\mathbf{t}} \rangle = \frac{\left\|\tilde{\mathbf{e}} + \tilde{\mathbf{t}}\right\|^2 - 2}{2} \tag{16}$$

Since C_ν is monotonic decreasing, there is a monotonic rising functional relationship between the PSDA score and the cosine score. This means the EER and minDCF and in general the whole DET-curve will be identical for cosine scoring and PSDA. However, for $b > 0$ and also all other kinds of trials, this scoring formula (14) has a more complex and possibly more useful behaviour.

The proposed model can be seen as intermediate between cosine scoring and PLDA. Cosine scoring has no learnable parameters, while PLDA has a substantial set of parameters, on the order of d^2. The PSDA model has parameters, but only about d of them. When b is small the EER for PSDA can get arbitrarily close to cosine scoring, but some extra flexibility is obtained when $w, b, \boldsymbol{\mu}$ are trained. For multiple enrollments, this model also gives a more interesting (arguably more principled) computation compared to simple averaging of the enrollment embeddings.

3.2.2. *Scoring implementation*

To implement (14) we note some details. Bessel-I functions are tricky. $I_\nu(\kappa)$ is available in `scipy.special`, but when ν is large (as here), both overflow and underflow occur. Overflow can be managed by using `scipy.special.ive`, which implements $I_\nu(\kappa)e^{-\kappa}$, which we used for:

$$\log I_\nu(\kappa) = \log\big(I_\nu(\kappa)e^{-\kappa}\big) + \kappa \tag{17}$$

This still underflows for small κ. Whenever $\kappa < \sqrt{\nu + 1}$, we used the first few terms (say 5) of the series expansion:

$$\log I_\nu(\kappa) = \log \sum_{i=0}^{\infty} \frac{(\kappa/2)^{2i+\nu}}{\Gamma(i+1)\Gamma(i+\nu+1)} \tag{18}$$

using `logsumexp`, $\log \kappa$ and `gammaln`.

Since $\left\|b\boldsymbol{\mu} + w\tilde{\mathbf{e}} + w\tilde{\mathbf{t}}\right\|$ is required in the denominator of (14), for *every* trial, a fast implementation is desirable. The numerator norms are of lesser concern. We can rewrite the norm as:

$$\begin{aligned} &\left\|b\boldsymbol{\mu} + w\tilde{\mathbf{e}} + w\tilde{\mathbf{t}}\right\|^2 \\ &= \left\|b\boldsymbol{\mu} + w\tilde{\mathbf{e}}\right\|^2 + \left\|w\tilde{\mathbf{t}}\right\|^2 + 2\left\langle b\boldsymbol{\mu} + w\tilde{\mathbf{e}}, w\tilde{\mathbf{t}}\right\rangle \end{aligned} \tag{19}$$

where the dot product for an m-by-n block of scores can be implemented with a fast matrix multiplication.

3.3. Learning

Maximum-likelihood-learning can be done via an EM algorithm with closed-form updates, given labelled observations for a number of speakers. For each speaker, i, let there be n_i observations with mean, $\bar{\mathbf{x}}_i$. The required statistics are just zero and first-order stats. In Gaussian PLDA, we need second-order statistics of the data too, but here, the speaker cluster spread is effectively contained in $\|\bar{\mathbf{x}}_i\|$. The total number of observations is $N = \sum_i n_i$ and the number of training speakers is S. The E-step is the computation of the *EM auxiliary*:

$$\begin{aligned} &Q(w, b, \boldsymbol{\mu}) = \text{const} + \\ &\sum_i \left\langle \log P(\mathbf{X}_i \mid \mathbf{z}, w) + \log P(\mathbf{z} \mid \boldsymbol{\mu}, b) \right\rangle_{P(\mathbf{z}|\mathbf{X}_i)} \\ &= \sum_i \left\langle n_i \log C_\nu(w) + n_i w \bar{\mathbf{x}}_i' \mathbf{z} \right\rangle_{P(\mathbf{z}|\mathbf{X}_i)} + \\ &\qquad \left\langle \log C_\nu(b) + b\boldsymbol{\mu}'\mathbf{z} \right\rangle_{P(\mathbf{z}|\mathbf{X}_i)} \\ &= N \log C_\nu(w) + Nw \frac{1}{N} \sum_i n_i \bar{\mathbf{x}}_i' \left\langle \mathbf{z} \right\rangle_i + \\ &\qquad S \log C_\nu(b) + Sb\boldsymbol{\mu}' \frac{1}{S} \sum_i \left\langle \mathbf{z} \right\rangle_i \end{aligned} \tag{20}$$

The expectations are taken w.r.t. the posteriors (11), where $\langle \mathbf{z} \rangle_i$ is the posterior expectation for speaker i, as given by (5). The M-step maximizes Q w.r.t. the parameters. For $b, \boldsymbol{\mu}$ this can be done by identifying the last line above with (6). For w, we identify the second last line with (8). This gives the updates:

$$\boldsymbol{\mu} \leftarrow \frac{\bar{\mathbf{z}}}{\|\bar{\mathbf{z}}\|}, \qquad b \leftarrow \rho^{-1}(\|\bar{\mathbf{z}}\|), \qquad w \leftarrow \rho^{-1}(\|\bar{\mathbf{r}}\|) \tag{21}$$

where:

$$\bar{\mathbf{z}} = \frac{1}{S} \sum_i \left\langle \mathbf{z} \right\rangle_i, \quad \text{and} \quad \|\bar{\mathbf{r}}\| = \frac{1}{N} \sum_i n_i \bar{\mathbf{x}}_i' \left\langle \mathbf{z} \right\rangle_i \tag{22}$$

Since ρ^{-1} is monotonic rising, we see:

- When $\bar{\mathbf{z}}$ is closer to the origin, the estimated between-speaker concentration, b is smaller (good for accuracy).
- The better the observations align with the hidden variables, the higher the estimated within-speaker concentration, w (also good).

4. The VMF classification head

The VMF-based PSDA model provides some interesting insights into classifier-style embedding extractor training. The functional form of the standard discriminatively trained multiclass linear classifier, with inputs in R^d, is obtained by letting the logits (softmax inputs) be the class log-likelihoods of a Gaussian model with a common within-class covariance. The logit for class i is:

$$\log \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_i, \mathbf{P}^{-1}) = \boldsymbol{\mu}_i' \mathbf{P} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i' \mathbf{P} \boldsymbol{\mu}_i + \text{const} \tag{23}$$

where $\mathbf{x}, \boldsymbol{\mu}_i \in \mathbb{R}^d$. Note the *class-dependent offsets*. If instead, we restrict $\mathbf{x}, \boldsymbol{\mu}_i \in \mathbb{S}^{d-1}$, we can derive a similar classifier using a VMF model, with common within-class concentration, so that the logits become:

$$\log \mathcal{V}(\mathbf{x} \mid \boldsymbol{\mu}_i, \kappa) = \kappa \boldsymbol{\mu}_i' \mathbf{x} + \text{const} \tag{24}$$

now *without* the offsets.[6] Although logits of the form (24) are now almost ubiquitous in machine learning,[7] and the connection

[6] If there is a known class imbalance, fixed, non-trainable offsets $\log p_i$, where p_i is class proportion, can be added to the logits.

[7] κ^{-1} is usually termed *temperature*

Table 1: *Comparison of the speaker verification performance for cosine scoring, PLDA, and PSDA. The performance metrics are EER (%) and minDCF for* $p_{\text{tar}} = 0.05$

Back-End	Dim.reduction	VoxCeleb1-O		VoxCeleb1-H	
		EER	MinDCF	EER	MinDCF
cos	-	1.10	0.071	2.13	0.126
cos	PCA 100	1.12	0.073	2.21	0.130
cos	LDA 100	2.05	0.157	5.28	0.340
PLDA	-	4.47	0.254	6.80	0.325
PLDA	PCA 100	1.34	0.094	2.69	0.152
PLDA	LDA 100	2.16	0.157	4.61	0.273
PSDA	-	1.10	0.071	2.13	0.126
PSDA	PCA 100	1.12	0.073	2.21	0.130
PSDA	LDA 100	2.04	0.155	5.20	0.333

with VMF likelihood has been explored in [9], it appears to be relatively unknown.

In speaker recognition (24) is used when embedding extractors are trained classifier-style (with one class per training speaker), although the problem may be made artificially harder by modifying the target logits with a margin, as in AM-softmax [13] and AAM-softmax [14].

In a variety of embedding extractors trained with AM-softmax and AAM-softmax, we found [15] that the length-normalized embeddings tend to collapse almost to a subspace, with very little variability in at least half of the dimensions. This is termed *dimensional collapse* in [16]. Curiously, if we train 512-dimensional embeddings, they collapse to less than 256 dimensions. If we train 256-dimensional embeddings, they collapse to less than 128 dimensions. Since dimensional collapse is not modelled by PSDA, we tried using PCA and LDA dimensionality reduction to better fit the data to the model, but with mixed results. For future work, we are interested in modifying the embedding extractor training criterion to combat such collapse and instead encourage a uniform hypersphere between-speaker distribution. The ideas in [17, 16, 9] may be helpful.

5. Experiments

We experimentally compared PSDA with two baseline back-ends, PLDA and cosine scoring, on a variety of data sets, with similar outcomes. We report only the VoxCeleb results here. Embeddings of dimension 256 were extracted with a ResNet34 that was trained on the development part of VoxCeleb2 [18] to optimize AAM loss. We reused the same dataset to train the back-end, with the difference that we concatenated all segments belonging to the same session before extracting the embedding, while for training the extractor, the original VoxCeleb segments were used. Also, augmentation was not done for back-end training. The performance was tested on two conditions: the "original" VoxCeleb1-O set and the "hard" trial list VoxCeleb1-H. We report equal error-rate (EER) and minimum detection cost function (DCF) at a target prior of 0.05.

Because of the above-mentioned dimensionality collapse, we tried dimensionality reduction before applying the various back-ends. We compared PCA, LDA, and no dimensionality reduction for each of the three back-ends. The results are presented in Table 1. On this data, cosine scoring provides better performance than PLDA, regardless of preprocessing, in agreement with previous works, e.g. [7]. For PLDA trained on the embeddings without dimensionality reduction we had to set the size of the speaker and channel subspaces to 100 to avoid singular covariance matrices. In other cases, we use two-covariance PLDA i.e. speaker and channel hidden variables have the same dimensionality as the observed data. Second, we observe that for both cosine scoring and PSDA, the raw embeddings without dimensionality reduction provide better performance than using low-dimensional embeddings. For PLDA we see the opposite trend: for good performance PLDA has to be applied after dimensionality reduction. Finally, the results of cosine scoring and PSDA are very similar, with the best performance achieved without preprocessing.

6. Conclusion

In speaker recognition state of the art (as in many other machine learning problems), length-normalized embeddings empirically perform better. Cosine scoring follows naturally, but is merely geometrically motivated (within-class distances should be small, between-class large). The PLDA model provides a beautiful, rich probabilistic scoring recipe, but makes use of Gaussians to model densities in $\mathbb{R}^d$. However, distributions restricted to $\mathbb{S}^{d-1}$ do not even possess densities in $\mathbb{R}^d$. Gaussians can only approximately fit length-normalized data. This is compounded with the above-mentioned speaker-dependent within-class distribution problem. We have shown that by using VMF distributions instead, PSDA can model distributions directly in $\mathbb{S}^{d-1}$, while still enjoying the scoring and training benefits of PLDA. We have shown theoretically and empirically, that (up to calibration) PSDA can be equivalent to cosine scoring. We hope this new tool can provide new theoretical insights and practical tools for both embedding extraction and scoring algorithms. In future we aim to explore calibration properties of PSDA, mixtures of PSDA, and PSDA scoring for clustering and diarization.

7. Acknowledgements

The work was supported by Czech Ministry of Interior project No. VJ01010108 "ROZKAZ", Czech National Science Foundation (GACR) project NEUREM3 No. 19-26934X and Horizon 2020 Marie Sklodowska-Curie grant ESPERANTO, No. 101007666. Computing on IT4I supercomputer was supported by the Czech Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "e-Infrastructure CZ – LM2018140".

8. References

[1] S. Ioffe, "Probabilistic linear discriminant analysis," in *9th European Conference on Computer Vision*, Graz, Austria, 2006.

[2] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision*, 2007.

[3] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey Speaker and Language Recognition Workshop*, Brno, Czech Republic, June 2010, keynote presentation.

[4] N. Brümmer and E. de Villiers, "The speaker partitioning problem," in *Odyssey 2010: The speaker and Language Recongnition Workshop, Brno*, 2010.

[5] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, Florence, Italy, 2011.

[6] M. Zhao, Y. Ma, M. Liu, and M. Xu, "The SpeakIn system for VoxCeleb Speaker Recognition Challenge 2021," *CoRR*, vol. abs/2109.01989, 2021. [Online]. Available: https://arxiv.org/abs/2109.01989

[7] H. Zeinali, S. Wang, A. Silnova, P. Matějka, and O. Plchot, "BUT system description to VoxCeleb Speaker Recognition Challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.

[8] H. Dubey, A. Sangwan, and J. H. Hansen, "Robust speaker clustering using mixtures of von mises-fisher distributions for naturalistic audio streams," *Proc. Interspeech 2018*, pp. 3603–3607, 2018.

[9] M. A. Hasnat, J. Bohné, J. Milgram, S. Gentric, and L. Chen, "Von Mises-Fisher mixture model-based deep learning: Application to face verification," *CoRR*, vol. abs/1706.04264, 2017. [Online]. Available: http://arxiv.org/abs/1706.04264

[10] K. V. Mardia and P. E. Jupp, *Directional Statistics*. Wiley, 2000.

[11] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra, "Clustering on the unit hypersphere using Von Mises-Fisher distributions," *Journal of Machine Learning Research*, 2005.

[12] A. Silnova, N. Brümmer, D. Garcia-Romero, D. Snyder, and L. Burget, "Fast variational Bayes for heavy-tailed PLDA applied to i-vectors and x-vectors," in *Interspeech, Hyderabad*, 2018.

[13] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.

[14] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[15] A. Silnova, T. Stafylakis, L. Mošner, O. Plchot, J. Rohdin, P. Matějka, L. Burget, O. Glembek, and N. Brümmer, "Analyzing speaker verification embedding extractors and back-ends under language and channel mismatch," in *Odyssey 2022: The speaker and Language Recongnition Workshop, Beijing*, submitted, 2022. [Online]. Available: https://arxiv.org/abs/2203.10300

[16] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *ICLR*, 2022. [Online]. Available: https://openreview.net/pdf?id=YevsQ05DEN7

[17] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," *CoRR*, 2020. [Online]. Available: https://arxiv.org/abs/2005.10242

[18] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech and Language*, 2019.