SOURCE COUNTING IN SPEECH MIXTURES USING A VARIATIONAL EM APPROACH FOR COMPLEX WATSON MIXTURE MODELS

Lukas Drude, Aleksej Chinaev, Dang Hai Tran Vu, Reinhold Haeb-Umbach

Department of Communications Engineering, University of Paderborn, 33098 Paderborn, Germany

ABSTRACT

In this contribution we derive a variational EM (VEM) algorithm for model selection in complex Watson mixture models, which have been recently proposed as a model of the distribution of normalized microphone array signals in the short-time Fourier transform domain. The VEM algorithm is applied to count the number of active sources in a speech mixture by iteratively estimating the mode vectors of the Watson distributions and suppressing the signals from the corresponding directions. A key theoretical contribution is the derivation of the MMSE estimate of a quadratic form involving the mode vector of the Watson distribution. The experimental results demonstrate the effectiveness of the source counting approach at moderately low SNR. It is further shown that the VEM algorithm is more robust with respect to used threshold values.

Index Terms— Blind source separation, Bayes methods, Directional statistics, Number of speakers

1. INTRODUCTION

Blind source separation (BSS) approaches that exploit the sparseness of speech in the short-time Fourier transform (STFT) domain have recently become very popular [1, 2]. With this approach it is assumed that at most one source is active in a time-frequency slot. An advantage of sparseness-based source separation is that it is applicable to both over- and underdetermined cases. This may be particularly interesting if the number of sources is not known in advance, which is often the case in practical situations. Model selection algorithms, that compare models of different complexity with respect to criteria, such as penalized likelihood or Bayesian information [3], are usually computationally expensive. An alternative are finite mixture models, where the model order is determined in the course of the mixture parameter estimation, such as the variational Expectation Maximization (EM) algorithm for Gaussian mixture models [4, 5].

Araki *et al.* modeled the histogram of directions of arrivals (DOAs) with a Gaussian mixture model (GMM) [5]. Imposing a sparse Dirichlet prior on the mixture weights caused the weights of those mixture components to tend to zero, that did not correspond to a speaker. After sufficiently many iterations there remained as many non-zero mixture weights as there were active speakers.

While the GMM had to be adjusted to be an appropriate model for DOAs, i.e., angles, a perhaps more elegant approach is to directly model the microphone signals in the STFT domain. In [6] we proposed to model the unit-length normalized vectors of microphone signals to be draws from a complex Watson mixture model (cWMM) since it allows to model uncertainties about directions of complex unit-norm vectors. The Watson distribution is a probability distribution defined on the unit hypersphere in the D-dimensional complex vector space [7].

Rather than imposing a Dirichlet prior on the mixture weights of the cWMM, we propose a quite different approach to estimate the number of simultaneously active speakers: the posterior distribution of the mode vectors of a two-component Watson mixture is estimated via a variational EM algorithm. Here, one mixture component is meant to represent the most dominant speaker while the other captures the remaining speech and noise. Then signals from the direction indicated by the mode vector corresponding to the dominant speaker are suppressed and the VEM algorithm is restarted with this modified input signal. This procedure is continued until a maximum number of expected speakers is reached. The sources are then counted by thresholding concentration parameters, weights and scalar products between mode vectors. The benefit of a fully Bayesian treatment becomes apparent, if the VEM algorithm is compared to an EM algorithm with point estimates, which is more sensitive with respect to the choice of a threshold.

2. VARIATIONAL EM ALGORITHM

2.1. Modeling and feature extraction

Consider a convolutive mixture of K independent source signals $S_k(t, f)$, $k = 1, \ldots, K$, captured by D microphones yielding the sensor signals $X_d(t, f)$, $d = 1, \ldots, D$ in STFT domain [2]. Using vector notation we have

$$\mathbf{X}(t, f) = \sum_{k=1}^{K} \mathbf{H}_k(f) S_k(t, f) + \mathbf{N}(t, f), \qquad (1)$$

The work was in part supported by Deutsche Forschungsgemeinschaft under contract no. Ha3455/8-1.

where $\mathbf{X} = (X_1, \ldots, X_D)^{\mathrm{T}}$ is the vector of sensor signals, $\mathbf{H}_k = (H_{1,k}, \ldots, H_{D,k})^{\mathrm{T}}$ is the vector of multiplicative transfer functions associated to source k, and $\mathbf{N} = (N_1, \ldots, N_D)^{\mathrm{T}}$ is the noise vector. Here $t = 1, \ldots T$ denotes the time frame and $f = 1, \ldots F$ the frequency bin index while ignoring the constant component $f = 0$.

Implying sparseness of the source signals, each vector $\mathbf{X}(t, f)$ is associated to either one prevalent source or considered to be noise. This is expressed by the latent binary $(K + 1)$-dimensional random vector $\mathbf{c}(t, f)$, for which component $c_k(t, f) = 1$, if the k-th source is dominant, while all other components are zero. $c_{K+1}(t, f) = 1$ indicates that only noise is present in the given time-frequency slot.

According to the normalized observation vector approach which has been proposed in [8] vectors $\mathbf{X}$ are phase normalized, frequency normalized and normalized to unit-norm. Arbitrarily selecting the signal of the first microphone as the reference we have

$$\begin{aligned} \tilde{X}_d(t,f) &= |X_d(t,f)| \exp\!\left(j \frac{\arg\left(X_d(t,f) X_1^*(t,f)\right)}{4f/F f_{\mathrm{s}} c^{-1} d_{\max}}\right), \\ \mathbf{Y}(t,f) &= \tilde{\mathbf{X}}(t,f) / \left\| \tilde{\mathbf{X}}(t,f) \right\|, \end{aligned} \tag{2}$$

where f_{s} is the sampling rate, c is the speed of sound and $d_{\max}$ is the maximum distance between the sensors. The frequency normalization assumes that the microphone spacing is small enough such that no spatial aliasing occurs and that there is a linear frequency dependency of the phase of $\mathbf{H}(f)$.

2.2. Statistical modeling

The normalized observation vectors $\mathcal{Y} = \{\mathbf{Y}(t, f) | \forall t, f\}$ form clusters on a D-dimensional complex hypersphere [6]. This distribution is modeled by a cWMM with the set of mode vectors $\mathcal{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_{K+1}\}$, considered to be random variables to infer, while the Watson concentrations $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{K+1})^{\mathrm{T}}$ and mixture weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{K+1})^{\mathrm{T}}$ are considered parameters to estimate. The concentration for the spatially uncorrelated or diffuse noise component is fixed to $\kappa_{K+1} = 0$ which corresponds to a uniform distribution on the complex hypersphere. Thus, the likelihood is given by:

$$\begin{aligned} &p(\mathcal{Y} | \mathcal{C}, \mathcal{W}; \boldsymbol{\kappa}) \\ &\quad = \prod_{t=1}^{T} \prod_{f=1}^{F} \prod_{k=1}^{K+1} \left(\frac{1}{c_{\mathrm{W}}(\kappa_k)} \mathrm{e}^{\kappa_k |\mathbf{W}_k^{\mathrm{H}} \mathbf{Y}(t,f)|^2} \right)^{c_k(t,f)}. \end{aligned} \tag{3}$$

The set of the class responsibilities $\mathcal{C} = \{c_k(t, f) | \forall t, f\}$ has a categorical distribution with mixture weights $\boldsymbol{\pi}$ [3], whereas the set of Watson mode vectors $\mathcal{W}$ follows a complex Bingham distribution depending on the set of complex Hermitian positive-semidefinite Bingham parameter matrices $\mathcal{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_{K+1}\}$:

$$p(\mathcal{C}; \boldsymbol{\pi}) = \prod_{t=1}^{T} \prod_{f=1}^{F} \prod_{k=1}^{K+1} \pi_k^{c_k(t,f)}, \tag{4}$$

$$p(\mathcal{W}; \mathcal{B}) = \prod_{k=1}^{K+1} \frac{1}{c_{\mathrm{B}}(\mathbf{B}_k)} \mathrm{e}^{\mathbf{W}_k^{\mathrm{H}} \mathbf{B}_k \mathbf{W}_k}. \tag{5}$$

The denominators $c_{\mathrm{W}}(\kappa_k)$ and $c_{\mathrm{B}}(\mathbf{B}_k)$ are the normalization factors for the Watson and Bingham distributions, respectively:

$$c_{\mathrm{W}}(\kappa_k) = \frac{(D-1)!}{2\pi^D M(1, D, \kappa_k)}, \tag{6}$$

$$c_{\mathrm{B}}(\mathbf{B}_k) = \frac{1}{4\pi M(\frac{1}{2}, \frac{3}{2}, \mathbf{B}_{0,k})}, \tag{7}$$

where $M(\cdot)$ is the confluent hypergeometric function for scalar argument defined in [9], Equation 13.2.2, or for matrix argument defined in [10], correspondingly.

Figure 1 illustrates the statistical dependencies between the random variables in circles, the observable random variable $\mathcal{Y}$ doubly-circled and hyper-parameters depicted in squares.

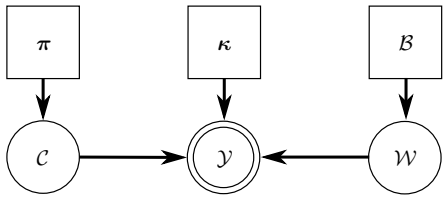

Fig. 1: Directed graphical diagram of statistical modeling

2.3. Derivation of variational EM algorithm

In the variational approach it is assumed that the posterior factorizes between the variables to be inferred [3]:

$$p(\mathcal{C}, \mathcal{W} | \mathcal{Y}) \approx q(\mathcal{C}) \cdot q(\mathcal{W}) \tag{8}$$

where

$$\ln q(\mathcal{C}) = \mathbb{E}_{\mathcal{W}} \left\{ \ln p(\mathcal{Y}, \mathcal{C}, \mathcal{W}) \right\} + \text{const.} \tag{9}$$

$$\ln q(\mathcal{W}) = \mathbb{E}_{\mathcal{C}} \left\{ \ln p(\mathcal{Y}, \mathcal{C}, \mathcal{W}) \right\} + \text{const.} \tag{10}$$

With Equation (3) - (7) the following update equations hold for the i-th iteration of the E-step ($i = 1 \ldots I$):

$$\begin{aligned} \ln \gamma_k^{(i)}(t,f) &= \kappa_k^{(i-1)} \mathbb{E}_{\mathbf{W}_k} \{ \mathbf{W}_k^{\mathrm{H}} \mathbf{Y}(t,f) \mathbf{Y}^{\mathrm{H}}(t,f) \mathbf{W}_k \} \\ &\quad - \ln M(1, D, \kappa_k^{(i-1)}) + \ln(\pi_k^{(i-1)}) + \text{const.}, \end{aligned} \tag{11}$$

$$\mathbf{B}_k^{(i)} = \kappa_k^{(i-1)} N_k^{(i)} \boldsymbol{\Phi}_{YY,k}^{(i)} + \mathbf{B}_{0,k}, \tag{12}$$

where

$$N_k^{(i)} = \sum_{t=1}^{T} \sum_{f=1}^{F} \gamma_k^{(i)}(t,f), \tag{13}$$

$$\boldsymbol{\Phi}_{YY,k}^{(i)} = \frac{1}{N_k^{(i)}} \sum_{t=1}^{T} \sum_{f=1}^{F} \gamma_k^{(i)}(t,f) \mathbf{Y}(t,f) \mathbf{Y}^{\mathrm{H}}(t,f). \tag{14}$$

Here, $\gamma_k^{(i)}(t,f) = P(c_k(t,f)=1|\mathcal{Y})$ is the a posteriori class probability, and $\mathbf{B}_k^{(i)}$ is the update of the Bingham parameter matrix from the a priori knowledge $\mathbf{B}_{0,k}$ after observing $\mathcal{Y}$.

In the M-step the parameters are estimated as follows:

$$\pi_k^{(i)} = N_k^{(i)} / \sum_{k=1}^{K+1} N_k^{(i)}, \tag{15}$$

$$\frac{M(2, D+1, \kappa_k^{(i)})}{D \cdot M(1, D, \kappa_k^{(i)})} = \mathbb{E}_{\mathbf{W}_k}\left\{ \mathbf{W}_k^{\mathrm{H}} \mathbf{\Phi}_{YY,k}^{(i)} \mathbf{W}_k \right\}, \tag{16}$$

where the latter is an implicit equation to determine $\kappa_k^{(i)}$.

Both, Equations (11) and (16) require the computation of a quadratic moment of a Bingham-distributed mode vector. Its solution is given in the appendix.

3. SOURCE COUNTING

If the VEM algorithm with $K = \nu_{\max}$ as the maximum number of expected mixture components is used directly for source counting, its performance strongly depends on the initial placement of the mode vectors of the Bingham distributions. Thus, the following approach is proposed to avoid this dependency:

In the first iteration ($\nu = 1$) observations $\mathbf{Y}(t,f)$ with highest energy, i.e., $A^{(\nu)}(t,f) := \|\tilde{\mathbf{X}}(t,f)\| > \mathrm{quantile}(q)$, are chosen to form a subset $\mathcal{Y}^{(\nu)}$, thus yielding a very basic speech activity detection. This criterion is similar to the amplitude criterion presented in [11] and is meant to select observations where the sparseness assumption is well fulfilled.

The VEM algorithm with one mixture component for a speaker and one mixture component for noise executed on this subset yields an updated a posteriori Bingham parameter matrix $\mathbf{B}_\nu$, and an estimate for the Watson concentration parameter κ_ν for the first speaker. The mode vector $\hat{\mathbf{W}}_\nu$ is the principal component of $\mathbf{B}_\nu$.

Inspired by [12], the energy of each time-frequency slot is weighted in the next iteration ($\nu \leftarrow \nu + 1$) to suppress observations from already found source directions:

$$A^{(\nu)}(t,f) = A^{(\nu-1)}(t,f) \cdot \left(1 - \mathrm{e}^{|\hat{\mathbf{W}}_{\nu-1}^{\mathrm{H}} \mathbf{Y}(t,f)|^{-1}}\right), \tag{17}$$

and the VEM algorithm is applied on this modified signal.

Algorithm 1 Source counting algorithm

1: Calculate $A^{(1)}(t,f) = \|\tilde{\mathbf{X}}(t,f)\|$
2: **for** $\nu = 1 \ldots \nu_{\max}$ **do**
3: **if** $\nu > 1$: **then** Use Equation (17) **end if**
4: Select observations $\mathcal{Y}^{(\nu)}$ with $A^{(\nu)} > \mathrm{quantile}(q)$
5: Use VEM algorithm with $\hat{\mathcal{Y}}^{(\nu)}$ to calculate $\mathbf{B}_\nu$, κ_ν
6: Calculate principal component $\mathbf{W}_\nu = \mathcal{P}(\mathbf{B}_\nu)$
7: **end for**
8: Calculate $s_\nu = \max_{\nu'=1\ldots\nu-1} |\mathbf{W}_\nu^{\mathrm{H}} \mathbf{W}_{\nu'}| \; \forall \nu = 2 \ldots \nu_{\max}$
9: Count iterations where $\kappa_\nu > \kappa_{\mathrm{Th}} \wedge \pi_\nu > \pi_{\mathrm{Th}} \wedge s_\nu < s_{\mathrm{Th}}$

This procedure is repeated until a maximum number of iterations $\nu_{\max}$ is reached. The number of speakers is then given by the number of iterations in which the VEM delivers a concentration parameter above a threshold κ_{Th}, the weight is above a threshold π_{Th} and a mode vector that indicates a spatial direction that is sufficiently different from those of the already found mode vectors. The latter is tested by comparing the absolute value of the scalar product between mode vectors to a threshold s_{Th}. Algorithm 1 summarizes the iterative procedure.

4. SIMULATIONS

Simulations have been performed with $K = 1 \ldots 6$ sources placed uniformly on 6, 8 or 10 fixed positions on a circle of radius 1 m around an array consisting of $D = 4$ sensors in a tetrahedral shape with 2 cm edge length in a non-reverberant room of dimensions 4 m × 4 m × 3 m. This yields a minimal angular distance of $\theta_{\min} = 60°$, 45° and 36°, respectively. The sources and the sensor array share the same height of 1.5 m although our sensor configuration allows arbitrary heights. The speech signals are 5 s signals from the TIMIT database with a sampling rate of 16 kHz to which white Gaussian noise at 10 dB SNR is added at each sensor. The STFT frame size is set to 1024 with a frame shift of 256.

The Bingham prior is uninformative: $\mathbf{B}_0^\nu = \mathbf{0}$. The threshold for the scalar products is set to $s_{\mathrm{Th}} = 0.7$, which corresponds to a minimum angle between two mode vectors of 30°. The performance is sensitive with respect to this parameter since it is related to spacial diversity of the sources. The threshold for the Watson concentration parameters is set to $\kappa_{\mathrm{Th}} = 1$, representing a nearly uniform distribution which is most likely not caused by a speaker.

The threshold π_{Th} is set to 10^{-3} since this is the optimal value for the EM algorithm. The differences to the VEM are that the expectation operators are omitted in Equations (11) and (16) and that the Update Equation (12) is not present. Fig-

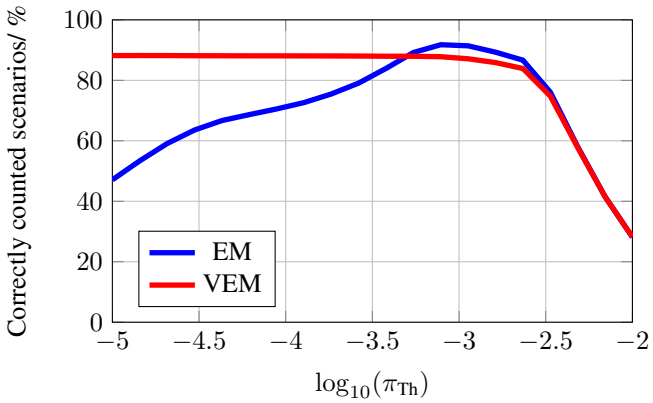

Fig. 2: Sensitivity with respect to threshold value for mixture weight averaged over all simulated scenarios

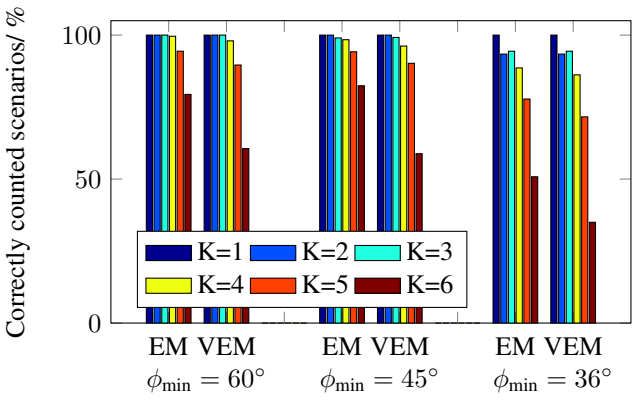

Fig. 3: Percentage of correctly counted number of active speakers for EM and VEM algorithms

ure 2 points out that the threshold for the mixture weights has a great influence on the EM performance whereas the VEM is less sensitive. This renders the VEM algorithm more robust since this threshold does not have to be tuned to a specific scenario. Each EM/VEM algorithm has converged sufficiently after $I = 10$ iterations.

Setting $q = 90\,\%$, the subset $\mathcal{Y}^{(\nu)}$ contains all observations with $A^{(\nu)}(t,f) > \text{quantile}(q)$. The simulation results are averages over 500 simulations for each number of speakers and each minimal angular distance.

Figure 3 shows the percentage of correctly counted scenarios, i.e., the percentage of simulations where the estimated number was equal to the true number of active speakers. The results obtained with the VEM algorithm are compared to an EM algorithm without Bingham priors.

In Figure 3 it can be seen that the counting accuracy decreases with decreasing θ_{min}, as expected. With $\theta_{\text{min}} = 60°$, thus meaning $K = 1 \ldots 6$ sources are randomly positioned on 6 places, the average counting accuracy achieved by the VEM algorithm is 91 % and it drops with a minimum angular distance of $45°$ and $36°$ to 91 % and 80 %, respectively. For the optimal choice of the threshold π_{Th} the performance of the EM algorithm is slightly better: 96 %, 96 % and 84 %.

5. CONCLUSIONS

We proposed a variational EM algorithm for model selection within complex Watson mixture models. We successfully applied this algorithm to the estimation of the number of active sources in a speech mixture captured by a microphone array. A simplified version using ML instead of MMSE estimates was shown by simulations to be less robust.

The variational EM algorithm derived in this contribution paves the way for online blind source separation based on complex Watson mixture models since a priori knowledge about the Watson mode vectors is essential for recursive estimators within this framework.

6. APPENDIX

We wish to compute the expectation $\mathbb{E}_{\mathbf{W}}\left\{\mathbf{W}^{\text{H}}\mathbf{\Phi}\mathbf{W}\right\}$ over a Bingham distributed random variable $\mathbf{W}$. Exploiting the invariance of the trace operator with respect to circular permutations we obtain

$$\mathbb{E}_{\mathbf{W}}\left\{\mathbf{W}^{\text{H}}\mathbf{\Phi}\mathbf{W}\right\} = \text{tr}\left(\mathbf{\Phi}\mathbb{E}_{\mathbf{W}}\left\{\mathbf{W}\mathbf{W}^{\text{H}}\right\}\right). \quad (18)$$

Since the Bingham parameter matrix $\mathbf{B}$ is Hermitian it can be decomposed into two unitary matrices and a diagonal matrix $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\text{H}}$. The integration variable $\mathbf{W}$ can now be substituted by $\mathbf{W} = \mathbf{U}\widetilde{\mathbf{W}}$ where the Jacobian determinant $\det\mathbf{J} = 1$. This aligns the Bingham distribution with its eigenvectors and thus decouples the components of $\widetilde{\mathbf{W}}$. The normalization constant only depends on the eigenvalues of the parameter matrix and thus $\mathbf{B}$ is substituted by $\mathbf{\Lambda}$:

$$\begin{aligned}\mathbb{E}_{\mathbf{W}}\left\{\mathbf{W}\mathbf{W}^{\text{H}}\right\} &= \int_{\mathbf{W}^{\text{H}}\mathbf{W}=1} \mathbf{W}\mathbf{W}^{\text{H}} c_{\text{B}}^{-1}(\mathbf{B}) \text{e}^{\mathbf{W}^{\text{H}}\mathbf{B}\mathbf{W}} \text{d}\mathbf{W} \\ &= \mathbf{U}\int_{\widetilde{\mathbf{W}}^{\text{H}}\widetilde{\mathbf{W}}=1} c_{\text{B}}^{-1}(\mathbf{\Lambda})\widetilde{\mathbf{W}}\widetilde{\mathbf{W}}^{\text{H}}\text{e}^{\widetilde{\mathbf{W}}^{\text{H}}\mathbf{\Lambda}\widetilde{\mathbf{W}}}\text{d}\widetilde{\mathbf{W}}\mathbf{U}^{\text{H}}.\end{aligned} \quad (19)$$

Each element of the above integral may now be treated separately:

$$e_{i,j} = c_{\text{B}}^{-1}(\mathbf{\Lambda})\int_{\widetilde{\mathbf{W}}^{\text{H}}\widetilde{\mathbf{W}}=1} \tilde{w}_i\tilde{w}_j^{*}\text{e}^{\sum_{d=1}^{D}\lambda_d|\tilde{w}_d|^2}\text{d}\widetilde{\mathbf{W}}. \quad (20)$$

Kent's polar coordinate transformation with $s_d = |\tilde{w}_d|$ and $\theta_d = \arg(\tilde{w}_d)$ yields an integral region for $\mathbf{s}$ defined by the standard $(D-1)$-simplex

$$\Delta^{D-1} = \left\{\mathbf{s}\in\mathbb{R}^D : \sum_{d=1}^{D} s_d = 1 \wedge s_d \geq 0\right\}. \quad (21)$$

The former integral bound $\widetilde{\mathbf{W}}^{\text{H}}\widetilde{\mathbf{W}} = 1$ does not constrain the phases θ_d and thus $\theta_d \in [0, 2\pi[$. The Jacobian determinant is given by $\det\mathbf{J} = 2^{-D}$ [13]:

$$e_{i,j} = c_{\text{B}}^{-1}\int_{\Delta^{D-1}}\int_{[0,\,2\pi[^D}\sqrt{s_i s_j}\text{e}^{\text{j}\theta_i}\text{e}^{\text{j}\theta_j}\text{e}^{\sum_{d=1}^{D}\lambda_d s_d}2^{-D}\text{d}\boldsymbol{\theta}\text{d}\mathbf{s}. \quad (22)$$

From $\int_0^{2\pi}\text{e}^{\text{j}\theta_i}\text{d}\theta_i = 0$ it follows that all cross terms vanish: $e_{i,j} = 0$ for $i \neq j$. In the case of $i = j$ the phase terms cancel each other and the integral for $\boldsymbol{\theta}$ reduces to the volume $(2\pi)^D$:

$$e_{i,i} = \frac{(2\pi)^D}{2^D}c_{\text{B}}^{-1}(\mathbf{\Lambda})\int_{\Delta^{D-1}} s_i\text{e}^{\sum_{d=1}^{D}\lambda_d s_d}\text{d}\mathbf{s}. \quad (23)$$

A further substitution with $s_1 = \widetilde{s}_1, \ldots s_{D-1} = \widetilde{s}_{D-1}$, $s_D = 1 - \sum_{d=1}^{D-1}\widetilde{s}_d$ with the Jacobian determinant $\det\mathbf{J} = 2$ yields an integration region equal to the set of vectors $\mathcal{L}_{D-2}$ between the standard $(D-2)$-simplex and the coordinate axis where the integral is basically the mean of a truncated multivariate exponentially distributed random variable $\tilde{\mathbf{s}}$:

$$e_{i,i} = \int_{\mathcal{L}_{D-2}}\tilde{s}_i\underbrace{2\pi^D c^{-1}(\mathbf{\Lambda})\text{e}^{\sum_{d=1}^{D}\lambda_d\tilde{s}_d}}_{p(\tilde{\mathbf{s}};\mathbf{\Lambda})}\text{d}\tilde{\mathbf{s}} = \mathbb{E}_{\tilde{s}_i}\left\{\tilde{s}_i\right\}. \quad (24)$$

The mean of a truncated multivariate exponential distribution is given by [13]:

$$\mathbb{E}_{\tilde{s}_i}\left\{\tilde{s}_i\right\} = c_{\text{B}}^{-1}(\mathbf{\Lambda})\frac{\partial c_{\text{B}}(\mathbf{\Lambda})}{\partial\lambda_i}. \quad (25)$$

7. REFERENCES

[1] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[2] S. Makino, T. W. Lee, and H. Sawada, Eds., *Blind Speech Separation*, Signals and communication technology. Springer, 2007.

[3] C. M. Bishop, *Pattern recognition and machine learning*, Springer, New York, 2006.

[4] H. Attias, "A variational Bayesian framework for graphical models," *Advances in neural information processing systems*, vol. 12, no. 1-2, pp. 209–215, 2000.

[5] S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Blind sparse source separation for unknown number of sources using Gaussian mixture model fitting with Dirichlet prior," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 33–36.

[6] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 2010, pp. 241–244.

[7] K. V. Mardia and I. L. Dryden, "The complex Watson distribution and shape analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 4, pp. 913–926, 1999.

[8] S: Araki, H. Sawada, R. Mukai, and S. Makino, "Normalized observation vector clustering approach for sparse source separation," *European Signal Processing Conference, Florence, Italy*, 2006.

[9] F. W. J. Olver and National Institute of Standards and Technology (U.S.), *NIST Handbook of Mathematical Functions*, Cambridge University Press, 2010.

[10] P. Koev and A. Edelman, "The efficient evaluation of the hypergeometric function of a matrix argument," *Mathematics of Computation*, vol. 75, no. 254, pp. 833–846, 2006.

[11] B. Lösch, *Complex Blind Source Separation with Audio Applications*, Ph.D. thesis, Stuttgart University, 2013.

[12] D. H. Tran Vu and R. Haeb-Umbach, "On initial seed selection for frequency domain blind speech separation.," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 1757–1760.

[13] J. T. Kent, "The complex Bingham distribution and shape analysis," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 285–299, 1994.