Independent vector analysis with multivariate student's *t*-distribution source prior for speech separation

Y. Liang, G. Chen, S.M.R. Naqvi and J.A. Chambers

The independent vector analysis algorithm can theoretically avoid the permutation problem in frequency domain blind source separation by using a multivariate source prior to retain the dependency between different frequency bins of each source. A super-Gaussian multivariate Student's *t*-distribution is adopted as the source prior to model the spectrum of speech signals and to mitigate imprecise variance knowledge as is commonplace in non-stationary signal processing. Moreover, the new multivariate source prior can be interpreted as a joint distribution constructed by a *t*-copula, which can describe the nonlinear inter-frequency dependency. Experimental results using 50 speech mixtures formed from the TIMIT database confirm the advantages of the proposed algorithm.

Introduction: Independent vector analysis (IVA) is a frequency domain method that solves the convolutive blind source separation problem (CBSS) [1]. The IVA method adopts a dependent multivariate super-Gaussian distribution as the source prior, instead of a univariate distribution used by traditional CBSS approaches. Thus, the IVA method can theoretically avoid the permutation ambiguity by exploiting certain statistical inter-dependency between frequency bins within each source vector, while removing the dependency between different sources. However, the form of the multivariate source prior should not always be fixed due to various types of dependency within the sources. In this Letter, we introduce a multivariate student's *t*-distribution as the source prior. It has a heavier tail than Gaussian distribution when the degree of freedom is small, which is required to model the spectrum of speech signals [2]. Moreover, it can be expressed as a scaled mixture of multivariate Gaussian distributions and thereby retain the variance dependency between different frequency bins as in the original IVA source prior without having precise knowledge about the variance, so important when modelling non-stationary signals [3]. The proposed multivariate source prior can also be interpreted as a joint distribution constructed by a *t*-copula with marginal univariate student's *t*-distribution. It is well known that copulas are used to describe nonlinear dependency [4]. Thus the multivariate student's *t* source prior can introduce exactly the *t*-copula to describe such dependency between different frequency bins.

IVA using multivariate student's t source prior: For the CBSS problem, the basic noise-free model in the frequency domain is described as

$$\boldsymbol{x}^{(k)} = \boldsymbol{H}^{(k)}\boldsymbol{s}^{(k)} \tag{1}$$

$$\hat{\boldsymbol{s}}^{(k)} = \boldsymbol{W}^{(k)}\boldsymbol{x}^{(k)} \tag{2}$$

where $\boldsymbol{x}^{(k)} = \left[x_1^{(k)}, x_2^{(k)}, \ldots, x_m^{(k)}\right]^{\mathrm{T}}$, $\boldsymbol{s}^{(k)} = \left[s_1^{(k)}, s_2^{(k)}, \ldots, s_n^{(k)}\right]^{\mathrm{T}}$ and $\hat{\boldsymbol{s}}^{(k)} = \left[\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \ldots, \hat{s}_n^{(k)}\right]^{\mathrm{T}}$ are the observed signal vector, the source signal vector and the estimated source vector, respectively, in the frequency domain, and $(\cdot)^{\mathrm{T}}$ denotes vector transpose. $\boldsymbol{H}^{(k)}$ is the mixing matrix with $m \times n$ dimensions, and $\boldsymbol{W}^{(k)}$ is the unmixing matrix with $n \times m$ dimensions. It is assumed that $m = n$ in this Letter. The index $k = 1, \ldots, K$ denotes the kth frequency bin.

The IVA method adopts the Kullback-Leibler divergence between the joint probability density function $p(\hat{\boldsymbol{s}}_1 \cdots \hat{\boldsymbol{s}}_n)$ and the product of marginal probability density functions of the individual source vectors $\prod_q(\hat{\boldsymbol{s}}_i)$ as the cost function:

$$\begin{aligned} J &= KL\Big(p(\hat{\boldsymbol{s}}_1 \cdots \hat{\boldsymbol{s}}_n) || \prod q(\hat{\boldsymbol{s}}_i)\Big) \\ &= \int p(\hat{\boldsymbol{s}}_1 \cdots \hat{\boldsymbol{s}}_n) \log \frac{p(\hat{\boldsymbol{s}}_1 \cdots \hat{\boldsymbol{s}}_n)}{\prod q(\hat{\boldsymbol{s}}_i)} d\hat{\boldsymbol{s}}_1 \cdots d\hat{\boldsymbol{s}}_n \\ &= \text{const} - \sum_{k=1}^{K} \log\left|\det\left(\boldsymbol{W}^{(k)}\right)\right| - \sum_{i=1}^{n} E\left[\log q(\hat{\boldsymbol{s}}_i)\right] \end{aligned} \tag{3}$$

where $E[\cdot]$ denotes the statistical expectation operator, $\det(\cdot)$ is the matrix determinant operator, K is the number of frequency bins and const denotes a constant number. The dependency between different source vectors should be removed but the dependency between the components of each vector can be retained, when the cost function is minimised.

The gradient descent method is used to minimise the cost function. By differentiating the cost function J with respect to the coefficients of the separating matrices $w_{ij}^{(k)}$, the gradients for the coefficients can be obtained as follows:

$$\begin{aligned} \Delta w_{ij}^{(k)} &= -\frac{\partial J}{\partial w_{ij}^{(k)}} = \left(w_{ij}^{(k)}\right)^{-\dagger} \\ &- E\left[\varphi^{(k)}\left(\hat{\boldsymbol{s}}_i^{(1)} \cdots \hat{\boldsymbol{s}}_i^{(k)}\right) x_j^{*(k)}\right] \end{aligned} \tag{4}$$

where $(\cdot)^{\dagger}$ and $(\cdot)^{*}$ denote the Hermitian transpose and the conjugate operators, respectively, and $\varphi^{(k)}(\cdot)$ is a nonlinear score function, which is given as follows:

$$\varphi^{(k)}\left(\hat{\boldsymbol{s}}_i^{(1)} \cdots \hat{\boldsymbol{s}}_i^{(k)}\right) = -\frac{\partial \log q\left(\hat{\boldsymbol{s}}_i^{(1)} \cdots \hat{\boldsymbol{s}}_i^{(k)}\right)}{\partial \hat{\boldsymbol{s}}_i^{(k)}} \tag{5}$$

For traditional CBSS approaches, the scalar Laplacian distribution is widely used for the source prior. However, the resultant nonlinear score function is a univariate function, which cannot keep the dependency between different frequency bins for each source. Therefore, a multivariate score function that is derived from the multivariate source prior is needed to retain the dependency between different frequency bins.

Our contribution is to propose the multivariate student's *t*-distribution as the source prior that takes the form

$$q(\boldsymbol{s}_i) \propto \left(1 + \frac{(\boldsymbol{s}_i - \boldsymbol{\mu}_i)^{\dagger} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{s}_i - \boldsymbol{\mu}_i)}{v}\right)^{-(v+K/2)} \tag{6}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are, respectively, the mean vector and a positive definite matrix of scale parameters, and v represents the degrees of freedom. We assume that the mean vector is a zero vector and $\boldsymbol{\Sigma}_i$ is an identity matrix. As such, with appropriate normalisation the nonlinear score function in (5) becomes

$$\varphi^{(k)}\left(\hat{\boldsymbol{s}}_i^{(1)} \cdots \hat{\boldsymbol{s}}_i^{(k)}\right) = \frac{\hat{\boldsymbol{s}}_i^{(k)}}{1 + (1/v)\sum\left|\hat{\boldsymbol{s}}_i^{(k)}\right|^2} \tag{7}$$

which is also a multivariate function as in the original IVA score function. All the frequency bins are accounted for during the learning process. Thus it can retain the inter-frequency dependency and provide the v parameter to tune the variance and leptokurtic nature of the model as in (6). With decreasing v, the tails become heavier and a suitable value can be estimated by the tail-index estimation method [5].

On the other hand, we can interpret this source prior in terms of copulas. Copulas are widely used for modelling the dependency between the marginal distributions of a joint distribution. The IVA algorithm requires a multivariate source prior that can retain the dependency between different frequency bins, so the copula is appropriate to model such dependency. We assume that the marginal distribution obeys a univariate student's *t*-distribution

$$q\left(\boldsymbol{s}_i^{(k)}\right) = \frac{\Gamma((v+K)/2)}{\sqrt{v\pi}\Gamma(v/2)} \left(1 + \frac{\left|\boldsymbol{s}_i^{(k)}\right|^2}{v}\right)^{(v+1/(2))} \tag{8}$$

where $\Gamma(\cdot)$ is the Gamma function. It is a super-Gaussian distribution, and is appropriate to model the spectrum of a speech signal.

According to [4], when using a copula to model the dependency, the joint distribution is established by

$$q\left(\boldsymbol{s}_i^{(1)}, \ldots, \boldsymbol{s}_i^{(K)}\right) = c(u_1, \ldots, u_K) \prod_{k=1}^{K} q\left(\boldsymbol{s}_i^{(k)}\right) \tag{9}$$

where $c(u_1, \ldots, u_K)$ is the copula density function and u_k are the marginal distribution functions.

In this Letter, we use a *t*-copula to model the dependency between different frequency bins, and the correspondent *t*-copula density function is [6]

$$c(u_1,\ \ldots,\ u_K) = \frac{\Gamma((v+K)/2)\Gamma(v/2)^{K-1}}{|\Sigma|^{(1/2)}\Gamma((v+1)/2)^K} \frac{\prod_{k=1}^{K}\left(1+\left(|y_k|^2/v\right)\right)^{(v+1/(2))}}{\left(1+\left(\boldsymbol{y}^{\dagger}\Sigma^{-1}\boldsymbol{y}/(v)\right)\right)^{(v+K/(2))}} \tag{10}$$

where y_k is the inverse distribution function of u_k. Thus by combining (8), (9) and (10), we can obtain the proposed multivariate student's t-distribution as the source prior for IVA.

Experimental results: In this simulation, we chose different speech signals from the TIMIT dataset [7]. Each speech signal was approximately 7 s long. The image method [8] was used to generate the room impulse responses, and the size of the room was $7 \times 5 \times 3$ m^3. The DFT length was 1024 and RT60 = 200 ms. We used a 2×2 mixing case, for which the microphone positions are [3.48, 2.50, 1.50] m and [3.52, 2.50, 1.50] m, respectively. The sampling frequency was 8 kHz. The separation performance was evaluated objectively by the signal-to-distortion ratio (SDR) and the signal-to-interference ratio (SIR) [9]. We found empirically that $v = 4$ is the appropriate value for the degrees of freedom parameter for the speech signals under test. Fig 1 shows the experimental setting.

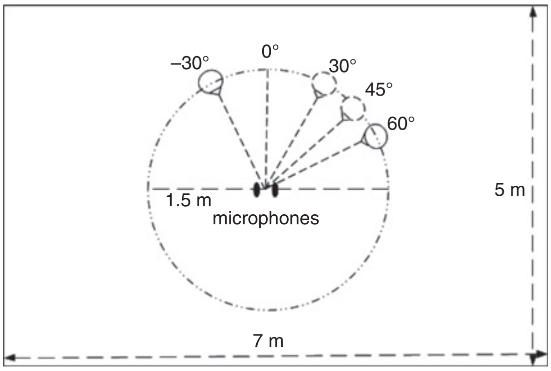

Fig. 1 *Plan view of source and microphone positions in room environment*

We chose two different speech signals randomly from the TIMIT dataset and convolved them into two mixtures. Then the original IVA method and the proposed IVA method with the new source prior were used to separate the mixtures, respectively. We changed the source positions to repeat the simulation. For every pair of speech signals, three different azimuth angles for the sources relative to the normal to the microphone array were set for testing; these angles were selected from 30°, 45°, 60° and $-30°$ as shown in Fig 1. After that, we chose another pair of speech signals to repeat the above simulations. In total, we used 10 different pairs of speech signals, and repeated the simulation 30 times at different positions. Table 1 shows the average separation performance for each pair of speech signals in terms of SDR and SIR, respectively.

Table 1: Separation performance comparison in SIR (dB)

Mixtures	Original (SDR)	Proposed (SDR)	Original (SIR)	Proposed (SIR)
Mixture 1	12.27	18.64	14.08	20.83
Mixture 2	8.88	12.59	10.72	14.27
Mixture 3	15.57	17.09	16.98	18.77
Mixture 4	18.10	19.50	20.14	20.78
Mixture 5	16.84	19.53	19.53	21.45
Mixture 6	18.81	20.17	20.30	21.47
Mixture 7	15.94	17.28	17.88	18.97
Mixture 8	9.97	11.73	12.08	12.77
Mixture 9	11.68	12.40	14.42	14.97
Mixture 10	18.80	19.91	20.28	20.95

The results shown in Table 1 confirm the advantage of the proposed IVA method that adopts the new multivariate source prior. We formed 50 different mixtures in total from the TIMIT database to test the separation performance, and the average SDR and SIR improvements were 1.3 and 1.1 dB, respectively.

Conclusion: In this Letter, we have proposed a new IVA method by choosing a multivariate student's t-distribution as the source prior. This new super-Gaussian source prior can model the spectrum of a speech signal even when the knowledge of its variance is limited. The experimental results confirm that the proposed IVA method can improve separation performance significantly. Future work will consider schemes for the estimation of the degrees of freedom parameter v.

14 June 2013
doi: 10.1049/el.2013.1999

Y. Liang, G. Chen, S.M.R. Naqvi and J.A. Chambers (*School of Electronic, Electrical and System Engineering, Loughborough University, Leicestershire, LE11 3TU, United Kingdom*)

E-mail: Y.Liang2@lboro.ac.uk

References

1 Kim, T., Attias, H.T., Lee, S.-Y., and Lee, T.-W.: 'Blind source separation exploiting higher order frequency dependencies', *IEEE Trans. Audio, Speech Lang. Process.*, 2007, **15**, (1), pp. 70–79
2 Cohen, I.: 'Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation', *Speech Commun.*, 2005, **47**, pp. 336–350
3 Peel, D., and McLachlan, G.J.: 'Robust mixture modelling using the *t* distribution', *Stat. Comput.*, 2000, **10**, pp. 39–348
4 Nelsen, R.B.: 'An introduction to copulas' (Springer, 2006)
5 Huisman, R., Koedijk, K.G., Kool, J.M.C., and Palm, F.: 'Tail–index estimate in small samples', *J. Bus. Econ. Stat.*, 2001, **19**, pp. 208–216
6 Daul, S., De Giorgi, E., Lindskog, F., and McNeil, A.: 'The grouped *t*-copula with an application to credit risk', *RISK*, 1970, **16**, pp. 73–76
7 Garofolo, J.S. *et al.*: 'TIMIT Acoustic-Phonetic Continuous Speech Corpus' (Linguistic Data Consortium, 1993)
8 Allen, J.B., and Berkley, D.A.: 'Image method for efficiently simulating small-room acoustic', *J. Acoust. Soc. Am.*, 1979, **65**, (4), pp. 943–950
9 Vincent, E., Fevotte, C., and Gribonval, R.: 'Performance measurement in blind audio source separation', *IEEE Trans. Audio Speech Lang. Process.*, 2006, **14**, pp. 1462–1469