ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Range-based non-orthogonal ICA using cross-entropy method

S. Easter Selvan[1], A. Chattopadhyay[2], U. Amato[3] and P.-A. Absil[1] *

1- Université catholique de Louvain - ICTEAM Institute
1348 Louvain-la-Neuve - Belgium

2- Université catholique de Louvain - iMMC

3- Consiglio Nazionale delle Ricerche - Istituto per le Applicazioni del Calcolo
Napoli 80131 - Italy

Abstract.
A derivative-free framework for optimizing a non-smooth range-based contrast function in order to estimate independent components is presented. The proposed algorithm employs the von-Mises Fisher (vMF) distribution to draw random samples in the cross-entropy (CE) method, thereby intrinsically maintaining the unit-norm constraint that removes the scaling indeterminacy in independent component analysis (ICA) problem. Empirical studies involving natural images show how this approach outperforms popular schemes [1] in terms of the separation performance.

1 Introduction

The separation performance of independent component analysis (ICA) relies on the choice of the contrast function as well as the optimization strategy. Even though a source adaptive contrast function yields more accurate source estimates than surrogate functions, the former suffers from computational overheads. Furthermore, even the typical criterion—the Shannon-entropy-based mutual information (MI)—combined with a local optimization strategy will not guarantee the recovery of sources due to the presence of mixing local optima. Several attempts to develop criteria free of mixing local optima encountered other implementation issues, e.g., the kurtosis-based contrast is not robust against outliers [2].

Amidst a myriad of contrast functions proposed, the range-based contrast introduced in [3] and investigated in [1] is endowed with the discriminacy property—being devoid of mixing local optima—meaning that each local optimum of the function corresponds to a satisfactory solution. Besides, it befits well with the boundedness of sources in signal/image applications and can also be efficiently estimated without much computational effort. Nevertheless, on the downside, the range-based contrast is non-smooth, and finding a good estimate of its derivative is difficult since its estimation is based on order statistics. Consequently, one cannot easily rely on a gradient-based algorithm in order to optimize the range-based contrast function. Moreover, the local optima of the contrast

*This paper presents research results of the Belgian Network DYSCO (Dynamical Systems, Control, and Optimization), funded by the Interuniversity Attraction Poles Programme, initiated by the Belgian State, Science Policy Office. The scientific responsibility rests with its authors.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

estimated from a finite sample set cannot be expected to be perfectly unmixing, hence favoring convergence to global optimum may still offer an advantage.

In keeping with these considerations, we propose to optimize the range-based contrast using a derivative-free probabilistic approach, namely, the cross-entropy (CE) optimizer. The crucial choice in the CE design is the probability distribution function (pdf) depending upon the nature of the optimization problem. Premised on the underlying geometry of the ICA problem stemming from the unit-norm constraint of the independent components, the von-Mises Fisher (vMF) distribution is used in conjunction with the CE algorithm in the present work to handle the constraint intrinsically. It is remarked that the vMF distribution has not yet been studied in the CE context, as far as we are aware. The paper concludes with experimental results wherein the solution quality in the CE method surpasses the outcome of the state-of-the-art approaches [1].

2 Preliminaries

Given a random vector $\mathbf{m} \in \mathbb{R}^n$, ICA estimates an unmixing matrix $\mathbf{X} \in \mathbb{R}^{n\times n}$ with unit-norm columns such that the n components of $\mathbf{b} = \mathbf{X}^{\mathrm{T}}\mathbf{m}$ are maximally independent as measured by some contrast function. We adopt the contrast function proposed in [3] and [8] for recovering the original sources from the observed data, since it possesses desirable contrast properties and is appropriate for the bounded sources assumption in natural images. For completeness sake, the contrast function based on a range estimation approach using order statistics is restated from [3]: $f(\mathbf{X}) := \log|\det \mathbf{X}| - \sum_{k=1}^{n} \log R(\mathbf{x}_k^{\mathrm{T}}\mathbf{m})$, where $R(\cdot)$ is the range function and $\mathbf{x}_k \in \mathbb{R}^n$ is the kth column vector of $\mathbf{X}$. An estimate of the range of a random variable Y [8], that remains insensitive to noise and outliers, based on an ordered finite sequence of observations $y_{(s)}$, $s = 1, 2, \ldots, S$, is $R(Y) := \frac{1}{m}\sum_{r=1}^{m} R_r(Y)$ with $R_r(Y) := y_{(S-r+1)} - y_{(r)}$. Given the value of S, m can be empirically determined [8] as $m(S) = \max(1, \overline{[\Re\{(\frac{S-18}{6.5})^{0.65}\} - 4.5]})$, where $\overline{\psi}$ denotes the nearest integer to ψ.

Our contribution is to estimate $\mathbf{X}$ using a CE algorithm which samples from the vMF distribution in order to generate candidate solutions. The rationale behind this is that the set of N samples represents feasible unmixing matrices $\mathbf{X}_i = \begin{bmatrix}\mathbf{x}_{1,i} & \mathbf{x}_{2,i} & \cdots & \mathbf{x}_{n,i}\end{bmatrix}$ with column vectors $\mathbf{x}_{k,i} \in \mathbb{R}^n$, $k = 1, 2, \ldots, n$ and $i = 1, 2, \ldots, N$, in the ICA problem such that $\mathrm{ddiag}(\mathbf{X}_i^{\mathrm{T}}\mathbf{X}_i) = \mathbf{I}_n$ and $\mathrm{rk}(\mathbf{X}_i) = n$, where n denotes the number of sources to be estimated, $\mathrm{ddiag}(\cdot)$ represents the diagonal matrix whose diagonal elements are those of the matrix in the argument, $\mathrm{rk}(\cdot)$ is the matrix rank and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Note that the $\mathbf{X}_i$'s are considered to be square matrices here, since the number of observed variables is assumed to be equal to the number of underlying sources. Following the definition of $\mathbf{X}_i$, the columns of the $\mathbf{X}_i$'s are constrained to be unit-norm vectors, i.e., $\|\mathbf{x}_{k,i}\|_2 = 1$; this means that $\mathbf{x}_{k,i}$ must lie on the unit hypersphere $\mathbb{S}^{n-1}$. It can readily be seen that the ICA unit-norm constraint is implicitly met when each column vector of the random samples, $\mathbf{x}_{k,i}$, is drawn from the vMF distribution in the n-dimensional space.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Formally the vMF distribution of an n-variate unit random vector $\mathbf{x}$ is defined as

$$\mathsf{MF}(\mathbf{x}; \boldsymbol{\mu}, \kappa) = c_n(\kappa) e^{\kappa \boldsymbol{\mu}^{\mathrm{T}} \mathbf{x}}, \tag{1}$$

which is parameterized by the unit-norm mean direction $\boldsymbol{\mu}$ and the concentration parameter $\kappa \geq 0$; $c_n(\kappa)$ is the normalizing constant for the density function given by $c_n(\kappa) = \frac{(\kappa/2)^{n/2-1}}{\Gamma(n/2) I_{n/2-1}(\kappa)}$, where $\Gamma(\cdot)$ is the well-known Gamma function and $I_s(\kappa)$ denotes the modified Bessel function of the first kind. κ characterizes the concentration of the unit vectors drawn according to the probability in (1) centered around $\boldsymbol{\mu}$. $\kappa = 0$ implies that (1) reduces to the uniform density on $\mathbb{S}^{n-1}$, and as $\kappa \to \infty$, the distribution degenerates to point mass at $\boldsymbol{\mu}$.

3 CE method employing vMF distribution

The CE[1] optimization is an iterative procedure consisting of the following two steps: (i) a random sample of candidate solutions is generated from a parameterized probability distribution and evaluated using the objective function (here termed as the contrast function); (ii) a subset of "elite samples", selected based on the objective function value, is used to update the parameters of the sampling distribution. This parameter update scheme will preserve the probability of producing "better" solutions in the subsequent iteration.

Unlike the conventional CE approach in [4], where the pdf's are considered to be Gaussians, we work with the vMF pdf's defined earlier, and the parameters of the distributions—$\boldsymbol{\mu}$'s and κ's—are updated based on the best performing subset of samples ((2) and (3)) at each iterative step t. What follows is the description of the algorithm (see Algorithm 1 below) with implementation subtleties. In the initialization step $t = 0$, $\hat{\boldsymbol{\mu}}_{0,k}$ and $\hat{\kappa}_{0,k}$ corresponding to the parameters of the vMF distribution for generating the kth column of the samples are taken as the canonical basis vector $\mathbf{e}_k$ and one, respectively; whereas for $t > 0$, the kth column of a random sample is drawn from the vMF, $\mathsf{MF}(\hat{\boldsymbol{\mu}}_{t,k}, \hat{\kappa}_{t,k})$, with $\hat{\boldsymbol{\mu}}_{t,k}$ and $\hat{\kappa}_{t,k}$ being the estimates of $\boldsymbol{\mu}$ and κ for column k at iteration t. In the sequel, the generation of random unit n-dimensional vectors following the vMF on the hypersphere $\mathbb{S}^{n-1}$ by a sampling scheme as suggested by Wood [5] is presented. This procedure generates a unit vector $\mathbf{v}$ sampled uniformly from the hypersphere in $\mathbb{R}^{n-1}$ and a scalar random variable w with range $(-1, 1)$, whose density function is proportional to $(1-w^2)^{(n-3)/2} e^{\kappa w}$, using rejection sampling. Consequently the unit vector, $\mathbf{z} = ((1-w^2)^{1/2} \mathbf{v}^{\mathrm{T}}, w)^{\mathrm{T}}$, follows the vMF distribution with modal direction $(0, 0, \ldots, 1)^{\mathrm{T}}$ and κ; then $\mathbf{x} = \mathbf{A}\mathbf{z}$ has the vMF distribution, $\mathsf{MF}(\boldsymbol{\mu}, \kappa)$, where $\mathbf{A}$ is any orthogonal matrix with the last column being $\boldsymbol{\mu}$. The samples from the desired vMF distributions are used to evaluate the objective function, and subsequently the resulting costs $f(\mathbf{X}_1), f(\mathbf{X}_2), \ldots, f(\mathbf{X}_N)$ are sorted to identify a smaller subset of $N_{\text{elite}} = \rho N$ elite samples with better

[1]The name "cross-entropy" originates from the fact that traditionally a metric based on the Kullback-Leibler divergence is employed to control the parameter update, and this should not be confused with entropy-based contrast functions.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

solutions. Notice that $\rho \in (0,1)$ determines the size of the elite population and their indices $\mathcal{I}$ are recorded as well. In the ensuing step, the parameters of the vMF distribution for sampling the kth column of the candidate solutions are updated to be the maximum-likelihood estimates (MLEs) from the elite samples, $\tilde{\boldsymbol{\mu}}_{t,k}$ and $\tilde{\kappa}_{t,k}$, as given in [6] and [7], respectively. By omitting the iteration and column indices for convenience, the MLEs of the distribution pertaining to any one of the columns of samples can be stated as

$$\tilde{\boldsymbol{\mu}} = \frac{\sum_{i\in\mathcal{I}} \mathbf{x}_i}{\|\sum_{i\in\mathcal{I}} \mathbf{x}_i\|}; \quad \tilde{\kappa} = \frac{\bar{R}(n-\bar{R}^2)}{1-\bar{R}^2}, \tag{2}$$

where $\bar{R} = \dfrac{\|\sum_{i\in\mathcal{I}} \mathbf{x}_i\|}{n}$. Note that $\tilde{\kappa}$ above is not the true MLE estimate, which is difficult to compute, but rather the empirically determined approximation in [7]. To avoid convergence to a wrong solution that is feared to happen sometimes in the early stages of optimization, the smoothed update expressions

$$\hat{\boldsymbol{\mu}}_{t,k} := \alpha\tilde{\boldsymbol{\mu}}_{t,k} + (1-\alpha)\tilde{\boldsymbol{\mu}}_{t-1,k}; \quad \hat{\kappa}_{t,k} := \beta_t\tilde{\kappa}_{t,k} + (1-\beta_t)\tilde{\kappa}_{t-1,k} \tag{3}$$

are applied. The fixed smoothing parameter α lies in the interval $[0.5, 0.9]$ in the conventional setting, whereas the dynamic smoothing parameter β_t is expressed as $\beta_t = \beta - \beta\left(1-\frac{1}{t}\right)^q$ with q being an integer typically between 5 and 10, and the smoothing constant β is chosen in the range $[0.8, 0.99]$. The dynamic smoothing of $\tilde{\kappa}_{t,k}$ is intended to avoid premature convergence to a degenerate distribution, which otherwise results in a sub-optimal solution. The iterative procedure is halted when all the $\hat{\kappa}_{t,k}$'s exceed a convergence threshold τ, implying that further progress is not feasible since all the pdf's are degenerate to point mass at $\begin{bmatrix}\hat{\boldsymbol{\mu}}_{t,1} & \hat{\boldsymbol{\mu}}_{t,2} & \cdots & \hat{\boldsymbol{\mu}}_{t,n}\end{bmatrix}$. The step-wise implementation procedure of the vMF-CE algorithm is concisely presented in Algorithm 1.

Algorithm 1 : The vMF-CE algorithm for the range-based ICA estimation.

Input: range-based contrast function f, elite sample size $\rho \in (0,1)$, sample size N and convergence threshold τ.
Output: minimum of f
Initialization: $t := 0$; $\hat{\boldsymbol{\mu}}_{t,k} \leftarrow \mathbf{e}_k$ and $\hat{\kappa}_{t,k} \leftarrow 1$ for $k = 1, 2, \ldots, n$.
while $\min_k \hat{\kappa}_{t,k} \le \tau$ **do**
  $t := t + 1$.
  Draw random samples $\mathbf{X}_i := \begin{bmatrix}\mathbf{x}_{1,i} & \mathbf{x}_{2,i} & \cdots & \mathbf{x}_{n,i}\end{bmatrix}$, where $i = 1, 2, \ldots, N$, and each $\mathbf{x}_{k,i} \sim \mathsf{MF}(\hat{\boldsymbol{\mu}}_{t-1,k}, \hat{\kappa}_{t-1,k})$ for $k = 1, 2, \ldots, n$.
  Evaluate f at $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$, and record the indices $\mathcal{I}$ of ρN samples for which $f(\mathbf{X}_i) \le f(\mathbf{X}_j)$, $\forall i \in \mathcal{I}$, $j \notin \mathcal{I}$.
  Set $\mathbf{X}^* \leftarrow \mathbf{X}_{i^*}$, where $i^* \in \mathcal{I}$ satisfies $f(\mathbf{X}_{i^*}) \le f(\mathbf{X}_i)$, $\forall i \in \mathcal{I}$.
  Update $\hat{\boldsymbol{\mu}}_{t,k}$ and $\hat{\kappa}_{t,k}$ for $k = 1, 2, \ldots, n$ following (2) and (3).
end while
return minimum $\mathbf{X}^*$.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4 Simulation using natural images

Dimension (n)	PI (mean±std.dev.)		
	SWICA	NOSWICA	vMF-CE
4	-9.58 ± 5.52	-9.54 ± 6.60	$\mathbf{-19.96 \pm 8.95}$
5	-5.55 ± 4.18	-3.38 ± 4.47	$\mathbf{-15.94 \pm 7.72}$
6	-1.47 ± 3.21	-0.02 ± 3.31	$\mathbf{-12.03 \pm 6.69}$
7	0.54 ± 2.80	2.45 ± 2.61	$\mathbf{-11.42 \pm 5.60}$
8	2.49 ± 2.31	4.49 ± 2.02	$\mathbf{-9.02 \pm 4.58}$
9	4.31 ± 1.72	5.54 ± 1.91	$\mathbf{-8.09 \pm 3.66}$

Table 1: Mean PI values of the SWICA, NOSWICA and vMF-CE. The values in bold face represent the minimum obtained among the experimented schemes.

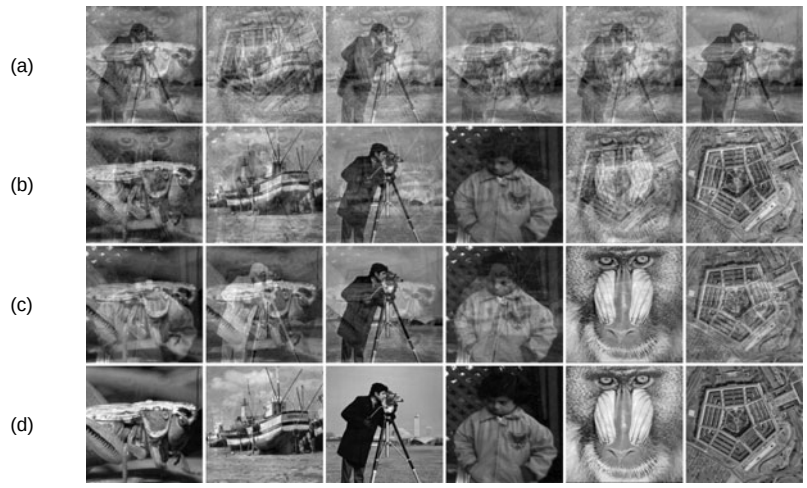

Fig. 1: (a) Mixed images. (b)-(d) Reconstructed images with the unmixing matrices having the PI values, 1.34, -2.07 and -16.08, estimated by the SWICA, NOSWICA and vMF-CE, respectively, followed by reordering.

We performed a simulation study involving 12 natural images taken from the MATLAB Image Processing Toolbox, wherein all the pixel images were resized to have 200×200 pixels each. By concatenating the pixels in each image column-wise, 12 S-dimensional data vectors are obtained, with $S = 40,000$. During every trial, n images from the pool of 12 images were randomly chosen, where n varies from four to nine. The n S-dimensional data vectors were first mixed by a randomly generated non-orthogonal mixing matrix in $\mathbb{R}^{n \times n}$, and then whitened as this ICA preprocessing provides a good initialization for the ensuing optimization process. Subsequently, the whitened mixture was supplied to the following optimizers to estimate the unmixing matrix: (i) SWICA [8], (ii) NOSWICA [9] and (iii) vMF-CE algorithm. The separation performance of the aforementioned ICA algorithms was assessed over

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

100 trials for each value of n using the performance index (PI) [10] defined as $\mathrm{PI} = 20\log_{10}\left(\frac{1}{n}\left(\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{|q_{ij}|}{\max_l |q_{il}|} - 1\right)\right)\right)$, where q_{ij} is the (i,j)th element of the global system matrix $\mathbf{Q} = \mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{W}$, with $\mathbf{V}$ and $\mathbf{W}$ being the whitening and mixing matrices, respectively. The parameter values used in the simulation study are listed below: $N = 10 \times n$, $N_{\mathrm{elite}} = 10$, $\alpha = 0.8$, $\beta = 0.7$, $q = 5$ and $\tau = 10^4$. The experimental results recorded in Table 1 indicate that the vMF-CE yields lower PI values compared to the SWICA and NOSWICA for all the dimensions ($n = 4, 5, \ldots, 9$). An important observation is that the separation performance of vMF-CE suffers less from the "curse of dimensionality" [1] than the rest, though it incurs high computational overheads. It is noteworthy to mention that in 83, 93 and 98 percent of the total test cases for $n = 4, 5$ and 6, respectively, and in all the test cases for $n = 7, 8$ and 9, the PI is significantly less for the vMF-CE than the methods reported in the literature. To demonstrate a substantial improvement in the ICA estimation with the proposed approach subjectively, the reconstructed sources from the investigated schemes for a specific instance ($n = 6$) are shown in Fig. 1; the PI between the true and the estimated sources are provided alongside to bear evidence. Finally, in agreement with the findings in [9], we observed in experiments (omitted for brevity) that NOSWICA—which itself is outperformed by the proposed vMF-CE—outclasses the popular FastICA and Joint Approximated Diagonalization of Eigenmatrices (JADE) algorithm in terms of solution quality.

References

[1] F. Vrins, Contrast properties of entropic criteria for blind source separation: a unifying framework based on information-theoretic inequalities. PhD Thesis, Faculté des Sciences Appliquées, Université catholique de Louvain, Louvain-la-Neuve, Belgium, March 2007.

[2] P. J. Huber, Projection pursuit, *The Annals of Statistics*, 13(2):435-475, 1985.

[3] D.-T. Pham, Blind separation of instantaneous mixtures of sources based on order statistics, *IEEE Transactions on Signal Processing*, 48(2):363-375, 2000.

[4] D. P. Kroese, S. Porotsky and R. Y. Rubinstein, The cross-entropy method for continuous multi-extremal optimization, *Methodology and Computing in Applied Probability*, 8(3):383-407, 2006.

[5] A. T. A. Wood, Simulation of the von Mises-Fisher distribution, *Communications in Statistics: Simulation and Computation*, 23(1):157-164, 1994.

[6] S. Sra, A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$, *Computational Statistics*, 1-14, 2011.

[7] 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*, 6:1345-1382, 2005.

[8] F. Vrins, J. A. Lee and M. Verleysen, A minimum-range approach to blind extraction of bounded sources, *IEEE Transactions on Neural Networks*, 18(3):809-822, 2007.

[9] J. A. Lee, F. Vrins and M. Verleysen, Non-orthogonal support-width ICA. In M. Verleysen, editor, *proceedings of the* 14^{th} *European Symposium on Artificial Neural Networks* (ESANN 2006), d-side pub., pages 351-358, April 26-28, Bruges (Belgium), 2006.

[10] P.-A. Absil and K. A. Gallivan, Joint diagonalization on the oblique manifold for independent component analysis, *proceedings of the* 31^{st} *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 945-948, May 14-19, Toulouse (France), 2006.