SOURCE SEPARATION IN STRUCTURED NONLINEAR MODELS

Anisse Taleb

Australian Telecommunications Research Institute
School of Electrical and Computer Engineering
Curtin University of Technology
GPO Box U1987, Perth WA 6845. Australia
email: anisse@atri.curtin.edu.au

ABSTRACT

This paper discusses several issues related to blind source separation in nonlinear models. Specifically, separability results show that separation in the general case is impossible, however, for specific nonlinear models the problem does have a solution. A specific set of parametric nonlinear mixtures is considered, this set has the Lie group structure. In the parameter set, a group operation is defined and a relative gradient is defined. The latter is applied to design stochastic algorithms for which the equivariance property is shown.

1. INTRODUCTION AND PROBLEM STATEMENT

Blind source separation in instantaneous and convolutive linear mixtures has been intensively studied over the last decade. However, in a general real world situation the observed signals do not fit this model. Nonlinear models, because of their better approximation capabilities, appear to be powerful tools for modeling such practical situations. Unfortunately, little work has been dedicated to the problem of blind source separation in nonlinear mixtures.

Several contributions have been made to the problem [3, 9, 10, 13, 11] however these lacked theoretical justification, relying more on analogies with linear models, rather than providing new theory for the nonlinear blind source separation problem. In [12], Taleb and Jutten propose the separation of a particular class of nonlinear mixtures, known as the post nonlinear mixtures. The observations are a component-wise nonlinear function of a linear instantaneous mixture.

The problem of source separation in general nonlinear instantaneous mixtures consists of retrieving unobserved sources $\boldsymbol{s}(t) = (s_1(t), s_2(t), \ldots, s_n(t))^T$, by only observing a nonlinear mixture $\boldsymbol{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T$ which writes as:

$$\boldsymbol{x}(t) = \mathcal{F}(\boldsymbol{s}(t)) \tag{1}$$

where $\mathcal{F}$ is an unknown nonlinear one-to-one mapping. This is done by constructing a nonlinear transform $\mathcal{G}$ (separation structure) in order to isolate in each component of the output vector $\boldsymbol{y}(t) = \mathcal{G}(\boldsymbol{e}(t))$ the image of one source:

$$y_i(t) = k_i(s_{\sigma(i)}(t)),\ i = 1, 2, \ldots, n \tag{2}$$

where σ is a permutation on $\{1, 2, \ldots, n\}$, and k_i is a function which represents a probable residual distortion.

The problem in this form is ill-posed, without additional assumptions on the sources signals there is an infinite number of solutions. By assuming statistical independence of the sources, the problem becomes quasi well-posed in the case of linear source separation.

In our case, given this assumption, the role of the separation structure is to transform the observations vector $\boldsymbol{e}(t)$ to a vector $\boldsymbol{y}(t)$ with independent components. Is the statistical independence assumption sufficient to obtain a separation in the sense of equation (2)? The next section is concerned with this question.

2. SEPARABILITY AND INDETERMINACIES

Provided the independence assumption holds, separability then consists of determining the form of the transformations $\mathcal{H}$ which leave the components of $\boldsymbol{s}$ independent.

2.1. Trivial transformations

Trivial transformations are those which map a random vector with independent components into a random vector with independent components thus conserving the independence property. It is shown that a one-to-one mapping $\mathcal{H}$ is trivial if and only if it can be written:

$$\mathcal{H}_i(u_1, u_2, \ldots, u_n) = h_i(u_{\sigma(i)}),\ i = 1, 2, \ldots, n \tag{3}$$

where h_i are arbitrary functions and σ is any permutation over $\{1, 2, \ldots, n\}$. From this result, it appears clearly that the objective of source separation is to force $\mathcal{H} = \mathcal{G} \circ \mathcal{F}$ to be trivial.

2.2. Darmois Results

In the general case, if the global transformation $\mathcal{H}$ is not restricted to belong to a certain set, it has been established by Darmois [8] that any random vector can, using a simple constructive approach, be mapped to a random vector with independent components by an infinite number of ways, .

As a corollary of this result, there exist an infinite number of transformations $\mathcal{H}$, without particular structure, which algebraically mix the variables while conserving their statistical independence. Source separation is then impossible by only using the statistical independence.

2.3. Specific model

The constructive method used by Darmois was mainly based on the free-of-constraints specification of the transformation. In fact,

if one restricts the transformation to belong to a specific set, this supplementary constraint associated with the independence assumption reduces dramatically the indeterminacies. The characterization of the indeterminacies for a specific model $\mathcal{Q}$ follows from solving an independence conservation functional equation:

$$\forall E \in \mathfrak{M}_n \quad \int_E dF_{s_1} dF_{s_2} \cdots dF_{s_n} = \int_{\mathcal{H}(E)} dF_{y_1} dF_{y_2} \cdots dF_{y_n}, \quad (4)$$

The set of distributions for which the solutions of this equations are not trivial contains those distributions for which separation is impossible. By determining this set, one identifies the sources for which separation is possible. For these sources, they are restored up to a trivial transformation of the model $\mathcal{Q}$.

3. STRUCTURED NONLINEAR MODELS

3.1. The set of separating transformations

Let us consider the generic nonlinear mixture of independent sources which writes as (1). As mentioned above, the global transformation $\mathcal{H} = \mathcal{G} \circ \mathcal{F}$ must not model any nonlinear transformation, in fact this will generate strong indeterminacies which provide solutions without interest. We then suppose that $\mathcal{G}$ has a particular form, formally $\mathcal{G}$ is constrained to belong to a specific set $\mathfrak{T}$. Each element $\mathcal{G}$ of the set of separating transformations $\mathfrak{T}$ is uniquely determined by a vector of parameters θ belonging to $\Theta \subset \mathbb{R}^d$. We then have:

$$\mathfrak{T} = \{\mathcal{G}_\theta, \quad \theta \in \Theta \subset \mathbb{R}^d\} \quad (5)$$

More over, this set is assumed to be structured: The couple $(\mathfrak{T}, \circ)$ forms a group. This group structure imposed on $\mathfrak{T}$ induces a group structure on the parameter set Θ with a certain law $\star$ satisfying

$$\forall\, \theta_1, \theta_2 \in \Theta, : \mathcal{G}_{\theta_1} \circ \mathcal{G}_{\theta_2} = \mathcal{G}_{\theta_1 \star \theta_2} \quad (6)$$

The neutral element of $(\Theta, \star)$ is denoted ζ, and corresponds to the neutral element of $(\mathfrak{T}, \circ)$: $\mathcal{G}_\zeta = \mathcal{I}$. The inverse of θ, denoted $\tilde{\theta}$, is defined by the relations $\theta \star \tilde{\theta} = \tilde{\theta} \star \theta = \zeta$.

It is interesting to notice that since, at least in theory, $\mathcal{G}$ is designed to invert the mixing transformation $\mathcal{F}$, the latter also belongs to $\mathfrak{T}$, hence there exists $\tau \in \Theta$ such that $\mathcal{F} = \mathcal{G}_\tau$. The global transformation writes then as $\mathcal{H} = \mathcal{G}_\theta \circ \mathcal{G}_\tau = \mathcal{G}_{\theta \star \tau}$ and obviously belongs to $\mathfrak{T}$.

3.2. A relative gradient in $(\Theta, \star)$

The previous algebraic properties are the basis for the definition of a relative gradient on $\mathfrak{T}$. The relative gradient for linear mixtures was introduced by Cardoso *et al.* [6], independently identified by Cichoki *et al.* [7] as being efficient and finally known as natural gradient by Amari [1]. The definition of the relative gradient has been extended to location-scale models by Cardoso [5]. A comparison of the natural and relative gradient has been studied in [4].

The idea behind the relative gradient is to *objectively* compare the variations of a function at different points. The objectivity depends on the structure of the function's parameter set. In our case, we want to have a measure of variation with respect to relative deviations from the identity transformation.

To put this idea into execution, let $\boldsymbol{y} = \mathcal{G}_\theta(\boldsymbol{x})$, and let a *small* variation $\delta\theta$ of θ, we then have:

$$\boldsymbol{y} + \delta\boldsymbol{y} = \mathcal{G}_{\theta+\delta\theta} \circ \mathcal{G}_\theta^{-1}(\boldsymbol{y}) = \mathcal{G}_{(\theta+\delta\theta)\star\tilde{\theta}}(\boldsymbol{y}) \quad (7)$$

The transformations in the neighborhood of the identity can be parameterized by $\zeta + \mu\varepsilon$, where ε is any normed vector, and μ is a small scalar. Hence, the right translation $\mathcal{R}_\theta$ maps any neighborhood of the identity ζ to a neighborhood of θ, which can be written as:

$$\theta + \delta\theta = \mathcal{R}_\theta(\zeta + \mu\varepsilon) \quad (8)$$

and which means that we have a relative deviation of $\mu\varepsilon$ from θ.

From the previous discussion, one can define the relative gradient with respect to θ of a function $Q(\theta)$ with scalar values by the d-dimensional vector function, $\overline{\nabla} Q$, by the following first order expansion :

$$Q((\zeta + \mu\varepsilon) \star \theta) = Q(\theta) + \mu\varepsilon^T \overline{\nabla} Q(\theta) + o(\mu) \quad (9)$$

where o is a function such that $\lim_{\mu \to 0} \frac{o(\mu)}{\mu} = 0$.

Equation (9) means that we compare, at the point θ, the value of Q at the point $\theta + \delta\theta$, obtained by a relative modification $\mu\varepsilon$ of θ, and the value of Q at the point θ. From the first order Taylor expansion:

$$\mathcal{R}_\theta(\zeta + \mu\varepsilon) = (\zeta + \mu\varepsilon) \star \theta = \theta + \mu J_{\mathcal{R}_\theta}(\zeta)\varepsilon + o(\mu) \quad (10)$$

and given that the *classical* gradient of Q, ∇Q is defined by:

$$Q(\theta + \mu\varepsilon) = Q(\theta) + \mu\varepsilon^T \nabla Q + o(\mu) \quad (11)$$

A simple expansion of (9) using (10) and identifying with (11) leads to a relation between the relative gradient and the classical gradient:

$$\overline{\nabla} Q(\theta) = J_{\mathcal{R}_\theta}(\zeta)^T \nabla Q(\theta) \quad (12)$$

The relative gradient as defined by (9) naturally provides Θ with a Riemannian structure. The metric tensor is given by the matrix $\chi(\theta) = J_{\mathcal{R}_\theta}(\zeta)^{-T} J_{\mathcal{R}_\theta}(\zeta)^{-1}$ which is positive definite since the right translation is one-to-one. This metric is the only one which leaves vector lengths invariant under the group translation operation *i.e.* $\mathcal{R}_\theta$. For instance, let $\phi_t(\theta) = (\zeta + t\epsilon) \star \theta$, when t varies the point $\phi_t(\theta)$ defines a curve $\mathcal{C}_\theta$ in the manifold Θ. The tangent vector v_θ to the curve $\mathcal{C}_\theta$ at the point θ, is defined by:

$$v_\theta = \left.\frac{d\phi_t(\theta)}{dt}\right|_{t=0} = J_{\mathcal{R}_\theta}(\zeta)\varepsilon \quad (13)$$

now consider the translation of this curve by $\mathcal{R}_{\tilde{\theta}}$, hence θ is mapped to ζ, and the curve $\mathcal{C}_\theta$ to the curve $\mathcal{C}_\zeta$. The tangent vector of the curve $\mathcal{C}_\zeta$ at the point ζ is now simply $v_\zeta = \varepsilon$. Consider now the inner product in the Riemannian manifold (Θ, χ): $< U, V >_\theta = U^T \chi(\theta) V$. It is straightforward to check that $< v_\zeta, v_\zeta >_\zeta = < v_\theta, v_\theta >_\theta$ and hence that the norm of the tangent vector at time $t = 0$ is conserved under translation.

The *natural* gradient in the Riemannian manifold (Θ, χ) is given by the relation:

$$\tilde{\nabla} Q = \chi(\theta)^{-1} \nabla Q \quad (14)$$

We shall see in the subsequent sections that the use of the relative gradient is equivalent to the use of the natural gradient.

4. ESTIMATING EQUATIONS

Statistical independence can be measured by different manners. Mutual information is one independence measure. It is defined by:

$$I(\boldsymbol{y}) = \sum_{i=1}^{n} H(y_i) - H(\boldsymbol{y}), \tag{15}$$

where H denote the differential entropy, $H(y_i) = -\int p_{y_i}(u) \log p_{y_i}(u) du$. Mutual information is always positive, and vanishes if and only if the components of the random vector $\boldsymbol{y}$ are statistically independent. The separation structure can then be estimated by minimizing the output mutual information $I(\boldsymbol{y})$:

$$\hat{\mathcal{G}} = \underset{\mathcal{G} \in \mathfrak{T}}{\operatorname{argmin}}\, I(\boldsymbol{y}) \iff \hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, I(\boldsymbol{y}) \tag{16}$$

It is clear that, in practice, the optimization of (16) cannot be done directly since one does not have access to the output distributions, these are either replaced by a guess or by 'plugging' at the output $\boldsymbol{y}(t)$ a density estimation procedure.

The relative gradient of mutual information can be computed by using the first order expansion of $I(\boldsymbol{y})$ when θ varies as $\theta + \delta\theta = (\zeta + \mu\varepsilon) \star \theta$, and it is found that

$$\Delta I(\boldsymbol{y}) = -\mu E[\psi(\boldsymbol{y})^T M(\boldsymbol{y}) + V(\boldsymbol{y})^T]\varepsilon + o(\mu), \tag{17}$$

where $\psi(\boldsymbol{y})$ is the vector $(\psi_{y_1}(y_1), \ldots, \psi_{y_n}(y_n))^T$ of score functions, defined by $\psi_{y_i}(y_i) = \frac{p'_{y_i}(y_i)}{p_{y_i}(y_i)}$. $M(\boldsymbol{y})$ and $V(\boldsymbol{y})$ depend only on the parametric model, formally:

$$M(\boldsymbol{y}) = \nabla_\theta \mathcal{G}_\theta(\boldsymbol{y})|_{\theta=\zeta} \tag{18}$$

$$V(\boldsymbol{y}) = \left(\operatorname{trace}(J_{\frac{\partial \mathcal{G}_\theta}{\partial \theta_1}|_{\theta=\zeta}}), \ldots, \operatorname{trace}(J_{\frac{\partial \mathcal{G}_\theta}{\partial \theta_d}|_{\theta=\zeta}})\right)^T \tag{19}$$

From (17), the relative gradient is equal to $E[\psi(\boldsymbol{y})^T M(\boldsymbol{y}) + V(\boldsymbol{y})^T]$.

In the neighborhood of the optimal solution $\hat{\theta}$ of (16), small relative deviations from $\hat{\theta}$ must lead to a null variation of $I(\boldsymbol{y})$ at the first order, *i.e.* null relative gradient:

$$E[\psi(\boldsymbol{y})^T M(\boldsymbol{y}) + V(\boldsymbol{y})^T] = 0 \tag{20}$$

One may notice that this equation depends indirectly on θ by means of $\boldsymbol{y}$. One readily verifies that if $\boldsymbol{y}$ has independent components then (20) is verified. We will denote in the following $H(\boldsymbol{y}) = \psi(\boldsymbol{y})^T M(\boldsymbol{y}) + V(\boldsymbol{y})^T$.

In a practical situation, the expectation operator in the equation (20) is replaced by an empirical expectation over a sample $\mathcal{X} = \{\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(T)\}$. The empirical expectation will be denoted $\hat{E}_T$, and leads finally to the estimating equation:

$$\hat{E}_T[H(\boldsymbol{y})] = \frac{1}{T} \sum_{t=1}^{T} H(\boldsymbol{y}) = 0 \tag{21}$$

where for all $t = 1, 2, \ldots, T$, $\boldsymbol{y}(t) = \mathcal{G}_\theta(\boldsymbol{x}(t))$. This equation, when solved, provide a batch estimate of θ.

5. GENERIC ADAPTIVE EQUIVARIANT ALGORITHMS

Most stochastic adaptive algorithms use an additive update rule, for instance the estimation of a parameter θ writes as [2]:

$$\theta^{t+1} = \theta^t + \gamma_t Q(\boldsymbol{x}(t), \theta^t) + \gamma_t^2 \rho(\boldsymbol{x}(t), \theta^t), \tag{22}$$

where $\boldsymbol{x}(t)$ represent the on-line observations of the system, and θ^t is the sequence of vectors to be recursively updated. The function $Q(\boldsymbol{x}, \theta)$ defines how the parameter θ is updated as a function of the new observations, while $\rho(\boldsymbol{x}, \theta)$ defines a small perturbation on the algorithm.

The most common form of adaptive algorithms correspond to $\rho = 0$. Convergence of the adaptive algorithm (22) is governed by the associated Ordinary Differential Equation (ODE) which is defined as:

$$\dot{\theta} = q(\theta), \quad \theta(0) = \theta^0 \tag{23}$$

where, for stationary on-line observations $\boldsymbol{x}(t)$, the mean vector field $q(\theta)$ is given by $q(\theta) = E[Q(\theta, \boldsymbol{x})]$.

Roughly speaking, the behavior of the algorithm (22) will be approximated by that of the ODE (23), the reader is invited to consult [2] and the technical details therein.

In our context, due to the use of the relative gradient, the proposed adaptive algorithm will use the *serial update rule* which has been used by Cardoso and Laheld [6] for adaptive linear source separation case. The serial updating rule relies heavily on the fact that we use a transformation model belonging to a group of transformations, for instance the group $GL(n)$ in linear source separation. It consists of plugging a transformation close to the identity $\mathcal{G}_{\zeta+\gamma_t H(\boldsymbol{y}(t))}$ at the output of the current estimate $\mathcal{G}_{\theta^t}$ to get the new estimate $\mathcal{G}_{\theta^{t+1}}$ (Fig. 1). This suggests the following non-additive update rule:

$$\theta^{t+1} = (\zeta + \gamma_t H(\boldsymbol{y}(t))) \star \theta^t, \tag{24}$$

where $\boldsymbol{y}(t)$ is the output at time t of the system: $\boldsymbol{y}(t) = \mathcal{G}_{\theta^t}(\boldsymbol{x}_t)$.

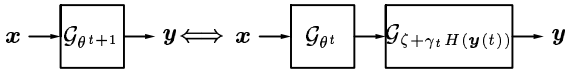

Fig. 1. Serial update rule

Clearly, the proposed algorithm (24) does not follow, in the general case, the general form of an adaptive algorithm (22). This occurs because the binary operation $\star$ is, in general, non-distributive with respect to vector addition. Linear source separation is an exception since matrix multiplication is distributive with respect to addition.

For sufficiently small γ_t, the convergence of (24) will only be affected by first order terms in γ_t. In fact, assuming that $\mathcal{R}_\theta$ is twice continuously differentiable, a second order expansion of (24) can be written as:

$$\theta^{t+1} = \theta^t + \gamma_t J_{\mathcal{R}_{\theta^t}}(\zeta) H(\boldsymbol{y}(t)) + \gamma_t^2 \rho(\gamma_t, \theta^t, \boldsymbol{y}(t)) \tag{25}$$

with $\rho(\gamma_t, \theta^t, \boldsymbol{y}(t))$ bounded for θ and $\boldsymbol{y}$. The mean vector field associated with the algorithm is then

$$\begin{aligned} h(\theta) &= J_{\mathcal{R}_\theta}(\zeta) E[H(\mathcal{G}_\theta(\boldsymbol{x}))] \\ &= J_{\mathcal{R}_\theta}(\zeta) J_{\mathcal{R}_\theta}(\zeta)^T E[J_{\mathcal{R}_\theta}(\zeta)^{-T} H(\boldsymbol{y})] \end{aligned} \tag{26}$$

the term inside the expectation operator is the classical gradient of the cost function. Here, we see that the use of the relative gradient is equivalent to a simple stochastic gradient descent in the Riemannian manifold Θ possessing the metric tensor $\chi(\theta)$.

Let $\theta(t)$ be a solution of the ODE with an initial point $\theta(0) = \theta^0$, and let $\phi(t) = \theta(t) \star \tau$, then:

$$\begin{aligned}\dot{\phi} = J_{\mathcal{R}_\tau}(\theta)h(\theta) &= J_{\mathcal{R}_\tau}(\theta)J_{\mathcal{R}_\theta}(\zeta)E[H(\mathcal{G}_\theta(\boldsymbol{x}))] \\ &= J_{\mathcal{R}_\phi}(\zeta)E[H(\mathcal{G}_\phi(\mathcal{G}_{\tilde{\tau}}(\boldsymbol{x})))] \quad (27)\end{aligned}$$

where we used that $\mathcal{R}_{\theta\star\tau} = \mathcal{R}_\tau \circ \mathcal{R}_\theta$. If τ is the parameter of the mixing system, then $\phi(t)$, the global parameter of $\mathcal{G}_\theta \circ \mathcal{F}$, will be solution of:

$$\dot{\phi} = J_{\mathcal{R}_\phi}(\zeta)E[H(\mathcal{G}_\phi(\boldsymbol{s}))], \quad \phi(0) = \theta^0 \star \tau \qquad (28)$$

It is obvious that the trajectory $\{\phi(t),\ t \geq 0\}$ depends only on $\theta^0 \star \tau$, hence changing the parameter of the mixing system would have the same effect on the algorithm global trajectory as changing the algorithm initial condition. The key point is that, as pointed out by Cardoso *et al.* [6] in the linear case, since the objective is the recovery of the initial source signals $\boldsymbol{s}(t)$, the performances of the adaptive algorithm are essentially determined by the closeness of the global system parameter ϕ to the neutral element ζ, and does not depend on the individual values of θ or τ. Hence, by the use of the relative gradient, the adaptive algorithm inherits the equivariant property from batch algorithms.

The curious reader may ask why use the update rule (24), when we can use the following:

$$\theta^{t+1} = \theta^t + \gamma_t J_{\mathcal{R}_{\theta^t}}(\zeta)H(\boldsymbol{y}(t)) \qquad (29)$$

which corresponds to the standard stochastic discretization of the ODE?

In fact, in the design of a stochastic adaptive algorithm, one starts by specifying the algorithm's ODE and then derives a stochastic Euler-like discretization of the latter. For *infinitely small* step sizes γ_t, the two rules are equivalent, however, the first update rule (24) respects the uniform performance property in discrete time while the latter does not. In fact, applying τ to the two members of (24), and denoting $\phi^t = \theta^t \star \tau$ leads to:

$$\phi^{t+1} = (\zeta + \gamma_t H(\mathcal{G}_{\phi^t}(\boldsymbol{s}(t)))) \star \phi^t. \qquad (30)$$

which presents the same equivariance property as (28) but in discrete-time. The complete study of the algorithm global performance follows the classical of the Lyapunov equation [2], solution of which provides the covariance matrix of ϕ^t: $E[(\phi - \zeta)(\phi - \zeta)^T]$. This issue is problem dependent and will not be addressed in this paper.

To summarize, the use of the relative gradient, whose definition is consistent with the parameter group operation, allows the design of non-additive adaptive equivariant algorithms. This fact, already known for the linear source separation problem, is here extended to more complex and general source separation models.

6. CONCLUSION

The blind source separation problem in its full generality is a very challenging statistical problem. Nonlinear instantaneous mixtures provide one level of this generality. It is shown that the generalization is not straightforward, the underlying indeterminacies when dealing with general nonlinear models are very strong. This makes the problem ill-posed and suggests its recasting for specific models.

Structured nonlinear models are one specific model of nonlinear mixtures. Taking into account the Lie group structure they posses, one can define a relative gradient whose use in an adaptive context leads to adaptive equivariant algorithms. This constitutes a generalization of equivariant linear source separation algorithms.

7. REFERENCES

[1] S-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

[2] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, 1990.

[3] G. Burel. Blind separation of sources: a nonlinear neural algorithme. *Neural Networks*, 5:937–947, 1992.

[4] J.-F. Cardoso. Learning in manifolds: the case of source separation. In *Proc. SSAP '98*, 1998.

[5] J.-F. Cardoso and S.-I. Amari. Maximum likelihood source separation: equivariance and adaptivity. In *Proc. of SYSID'97, 11th IFAC symposium on system identification, Fukuoka, Japan*, pages 1063–1068, 1997.

[6] J-F. Cardoso and B. Laheld. Equivariant adaptive source separation. *IEEE Trans. on S.P.*, 44(12):3017–3030, December 1996.

[7] A. Cichoki and Unbehauen R. New neural networks with online learning for blind identification and blind separation of sources. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, 43:894–906, November 1996.

[8] G. Darmois. Analyse des liaisons de probabilité. In *Proceedings Int. Stat. Conferences 1947*, volume III A, page 231, Washington (D.C.), 1951.

[9] P. Pajunen, A. Hyvarinen, and J. Karhunen. Nonlinear source separation by self-organizing maps. In *ICONIP 96*, volume 2, pages 1207–1210, Hong-Kong, September 1996.

[10] P. Pajunen and J. Karhunen. A maximum likelihood approach to nonlinear blind source separation. In *ICANN'97*, pages 541–546, Lausanne (Switzerland), October 1997.

[11] C.G. Puntonet, M.R. Alvarez, A. Prieto, and B. Prieto. Separation of sources in a class of post-nonlinear mixtures. In *ESANN 98*, pages 321–326, Bruges (Belgium), April 1998.

[12] A. Taleb and C. Jutten. Source separation in post nonlinear mixtures. *IEEE Trans. on SP*, 47(10):2807–2820, 1999.

[13] H. H. Yang, S. Amari, and A. Cichocki. Information-theoretic approach to blind separation of sources in nonlinear mixture. *Signal Processing*, 64(3):291–300, 1998.