International Journal of Neural Systems, Vol. 0, No. 0 (April, 2000) 00–00

Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for nonlinear mixtures

Christian JUTTEN
Laboratory of Images and Signals (UMR CNRS 5083, INPG, UJF), 46 avenue Félix Viallet
38031 Grenoble Cedex, France
E-mail: Christian.Jutten@inpg.fr

Juha KARHUNEN
Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 5400
FIN-02015 HUT, Espoo, Finland
E-mail: Juha.Karhunen@hut.fi

Received (to be inserted
Revised by Publisher)

In this paper, we review recent advances in blind source separation (BSS) and independent component analysis (ICA) for nonlinear mixing models. After a general introduction to BSS and ICA, we discuss in more detail uniqueness and separability issues, presenting some new results. A fundamental difficulty in the nonlinear BSS problem and even more so in the nonlinear ICA problem is that they provide non-unique solutions without extra constraints, which are often implemented by using a suitable regularization. In this paper, we explore two possible approaches. The first one is based on structural constraints. Especially, post-nonlinear mixtures are an important special case, where a nonlinearity is applied to linear mixtures. For such mixtures, the ambiguities are essentially the same as for the linear ICA or BSS problems. The second approach uses Bayesian inference methods for estimating the best statistical parameters, under almost unconstrained models in which priors can be easily added. In the later part of this paper, various separation techniques proposed for post-nonlinear mixtures and general nonlinear mixtures are reviewed.

1. From linear to nonlinear BSS and ICA

1.1. *From BSS problem to ICA method*

Consider N samples of the observed data vector $\mathbf{x}$, modeled by

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \sum_{j=1}^{n} s_j \mathbf{a}_j \tag{1}$$

where $\mathbf{A}$ is the unknown mixing matrix with column vectors $\mathbf{a}_j$, $j = 1, 2, \ldots, n$, and $\mathbf{s}$ is an unknown n-dimensional source vector containing the source signals $s_1, s_2, \ldots, s_n$, which are assumed to be statistically independent. In general, the dimensionality m of the vectors $\mathbf{x}$ and $\mathbf{a}_j$ can be different from n. Usually it is assumed that there are at least as many mixtures as sources ($m \geq n$), the mixing matrix $\mathbf{A}$ has full rank, and that at most one of the sources s_j is Gaussian.

Such a model can be used in different situations, for example:

- In multidimensional signal processing, where each sensor receives an unknown superimposition $x_i(t)$ of unknown source signals $s_j(t)$ at time instants $t = 1, \ldots, N$ [1,2].
- In sparse coding, where one tries to code the N

data vectors as a (sparse) linear combination of linearly independent components $\mathbf{a}_j$ [3,4].

In the first problem, the goal is to recover the n unknown actual source signals $s_j(t)$ which have given rise to the observed mixtures. This is referred to as the blind source separation (BSS) problem; *blind* since no or very little prior information about the sources is required.

The second problem (which is not addressed further in this paper) is related to a hypothetical and approximate model which tries to approximate the data as well as possible using n suitably chosen linearly independent components. We content with providing just a few early references [1,3,4] to this large area here. In sparse coding, independence assumptions may sometimes also be replaced by other requirements such as positivity of the sources s_j and the elements of the mixing matrix $\mathbf{A}$.

Using the fundamental assumption that the source signals s_j, $j = 1, 2, \ldots, n$, are statistically independent, BSS can be achieved by estimating a separating matrix $\mathbf{B}$ such that $\mathbf{y} = \mathbf{B}\mathbf{x}$ becomes a random vector with mutually independent components. This method is called independent component analysis (ICA). For the linear model (1), it has been proved that ICA provides the independent sources up to well-known scaling and permutation indeterminacies on certain conditions [5,6,7].

Standard linear ICA is now understood quite well [8,1,2], and since 1985 (see the paper [9] for a historical review and early references), several well-performing BSS and ICA algorithms [10,11,12,13,14,15,16,2] have been developed and applied to an increasing number of applications [17,18,19,20,21,22,23,1,24,2]. Many more references on linear ICA and BSS can be found in the recent books [1,2,22,23,24] and proceedings of conferences devoted to ICA and BSS [17,18,19,20,21].

If the source signals are not plain random variables but have a temporal structure, linear blind source separation can be achieved by utilizing either temporal correlations [25,26] or nonstationarity [27,28] under somewhat different assumptions [1,2,24]. Under these priors, ICA is not required for solving BSS: algorithms based on fairly simple temporal and spatial decorrelation are then sufficient. We have again mentioned a few references only on these topics which are outside the main scope of this paper.

1.2. *Nonlinear BSS and ICA*

A natural extension of the linear model (1) is to consider nonlinear mixing models. For instantaneous mixtures, the nonlinear mixing model has the general form

$$\mathbf{x} = \mathcal{F}(\mathbf{s}) \qquad (2)$$

where $\mathbf{x}$ and $\mathbf{s}$ denote the data and source vectors as before, and $\mathcal{F}$ is an unknown real-valued m-component mixing function.

Assume now for simplicity that the number of independent components n equals the number of mixtures m. General nonlinear ICA then consists in finding a mapping $\mathcal{G} : \mathcal{R}^n \rightarrow \mathcal{R}^n$ that yields components

$$\mathbf{y} = \mathcal{G}(\mathbf{x}) \qquad (3)$$

which are statistically independent. Unfortunately, in the general nonlinear case, the ICA method provides many solutions, which generally do not lead to source separation. In fact, for any random variables x and y, one can find nonlinear mappings with a non-diagonal Jacobian, which provide statistically independent random variables (see Section 3). Moreover, if x and y are two independent random variables, any of their functions $f(x)$ and $g(y)$ are independent, too. Consequently, if ICA leads to separation, sources s_i may be recovered up to any nonlinear distortion $h_i(s_i)$!

Contrary to the linear case, the BSS problem for general nonlinear mixtures differs greatly from the nonlinear ICA problem defined above. If some arbitrary independent components are found for the data generated by (2), they may be quite different from the true source signals. Generally, solving the nonlinear BSS problem is not easy, and requires additional prior information or suitable regularizing constraints.

The remainder of this paper is organized as follows. First, we mention a few early works on nonlinear BSS and ICA. In Section 3, we recall the main results on the existence of solutions to the nonlinear BSS problem. In Section 4, we show that structural constraints can lead to a unique BSS solution. In Section 5, we detail the so-called post-nonlinear model and related methods and algorithms. In Section 6, we explain how nonlinear BSS can be solved in a Bayesian framework. The last parts of the paper contain a short description of other current research directions in nonlinear ICA and BSS in Section 7, followed by conclusions. To keep the length of the paper reasonable, we have not included any experimental results; they can be found in the given references.

2. Early works in nonlinear BSS and ICA

Already in 1987, Jutten [29] used soft nonlinear mixtures for assessing the robustness and performance of the seminal Hérault-Jutten algorithm. However, Burel [30] was probably the first to introduce an algorithm specifically for nonlinear ICA. His method, restricted to known nonlinearities with unknown parameters and based on back-propagation type neural learning, suffers from huge computational complexity and problems with local minima.

A few years later on, Deco and Parra with their co-authors developed in a series of papers [31,32,33,34,35] nonlinear ICA methods based on volume conserving symplectic transformations. In fact, the constraint of volume conservation is somewhat restrictive, and hence these methods are usually not able to recover the original sources.

One of the earliest ideas for achieving general nonlinear BSS (or ICA), introduced first by Pajunen in [36] and then together with co-authors in [37], is to use the well-known self-organizing map (SOM) (see for example [38,39]) to that end. SOM learns in an unsupervised manner a nonlinear mapping from the data to a usually 2-D rectangular grid. With suitable modifications [36,37], the mapping provided by SOM is roughly uniformly distributed on the grid. The marginal densities along the sides of the rectangular grid become then statistically independent. The SOM mapping also provides a regularization mechanism needed in nonlinear BSS, because it tries to preserve the structure of the data by using a nonlinear mapping as simple as possible [38,1].

Taleb and Jutten [40], considering the large indeterminacies of nonlinear BSS and ICA, then studied in 1997 constrained nonlinear models, especially the so-called post-nonlinear mixture, mainly for ensuring uniqueness of found solution.

3. Existence and uniqueness of nonlinear ICA and BSS

Several authors [41,42,9,43,44,6] have recently addressed the important issues on the existence and uniqueness of solutions for the nonlinear ICA and BSS problems. Their main results, which are direct consequences of Darmois's results on factorial analysis [45], are reported in this section. However, other results can be derived without requiring Darmois'results, assuming simple priors on the sources (*e.g.* distributions with bounded support, see section 1 and [46,47]).

3.1. *Indeterminacies*

Recall first a few definitions.

Definition 1 *A random vector* $\mathbf{x}$ *has mutually statistical independent components if its joint probability density function (pdf)* $p_{\mathbf{x}}(\mathbf{u})$ *satisfies* $p_{\mathbf{x}}(\mathbf{u}) = \prod_i p_{x_i}(u_i)$, *where* $p_{x_i}(u_i)$ *are the marginal pdfs of the random variables* x_i.

The product of a permutation matrix $\mathbf{P}$ by any diagonal mapping both preserves independence and ensures separability.

Definition 2 *A one-to-one mapping* $\mathcal{H}$ *is called* **trivial**, *if it transforms* **any** *random vector* $\mathbf{s}$ *with independent components into a random vector with independent components.*

The set of trivial transformations will be denoted by $\mathfrak{T}$. Trivial mappings preserve the independence property of *any* random vector. One can easily show that a one-to-one mapping $\mathcal{H}$ is trivial if and only if it satisfies

$$\mathcal{H}_i(u_1, u_2, \ldots, u_n) = h_i(u_{\sigma(i)}),\ i = 1, 2, \ldots, n \quad (4)$$

where the h_i are arbitrary functions and σ is any permutation over $\{1, 2, \ldots, n\}$. These mappings are also characterized by a diagonal (up to a permutation) Jacobian matrix.

This result establishes a link between the independence assumption and the objective of source separation. It will shortly become clear that the objective of source separation is to make the global transformation $\mathcal{H} = \mathcal{G} \circ \mathcal{F}$ trivial using the independence assumption.

However, from (4) it is clear that the sources can only be separated up to a permutation and a nonlinear function. For any invertible mapping

$$\mathcal{F}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})]^T$$

whose each component is a scalar nonlinear mapping $f_i(\mathbf{x}) = f_i(x_i), i = 1, \ldots, n$, it is evident that if $p_{\mathbf{x}}(\mathbf{u})$

$= \prod_i p_{x_i}(u_i)$, then

$$p_{\mathcal{F}(\mathbf{x})}(\mathbf{v}) = \prod_i p_{f_i(x_i)}(v_i).$$

3.2. *Results from factor analysis*

For general mappings $\mathcal{H}$, Darmois established already in 1947-1951 [45] that preserving independence is not a strong enough constraint for ensuring separability in the sense of equation (4). Using a method [48] similar to the well-known Gram-Schmidt orthogonalization procedure in linear algebra, it is easy to decompose any random vector into a non-trivial mapping of independent variables. In [41], Hyvärinen and Pajunen have applied Darmois's idea to construction of parameterized families of nonlinear ICA solutions.

Darmois's result shows that there exist non-trivial transformations $\mathcal{H}$ which "mix" the variables while preserving their statistical independence. Hence blind source separation is simply *impossible* for general nonlinear transformations by resorting to statistical independence only without additional constraints or prior information.

In the conclusion of [45], Darmois clearly states: *"These properties [...] clarify the general problem of factor analysis by showing the great indeterminacies it presents as soon as one leaves the already very wide field of linear diagrams."*

3.3. *A simple example*

We now present a simple example of mixing mappings preserving independence, derived from [42]. Suppose $s_1 \in \mathcal{R}^+$ is a Rayleigh distributed variable with a probability density function (pdf) $p_{s_1}(s_1) = s_1 \exp(-s_1^2/2)$, and s_2 is independent of s_1, having a uniform pdf $s_2 \in [0, 2\pi)$. Consider the nonlinear mapping

$$\begin{aligned} [y_1, y_2] &= \mathcal{H}(s_1, s_2) \\ &= [s_1 \cos(s_2), s_1 \sin(s_2)] \end{aligned} \qquad (5)$$

which has a non-diagonal Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \cos(s_2) & -s_1 \sin(s_2) \\ \sin(s_2) & s_1 \cos(s_2) \end{pmatrix}. \qquad (6)$$

The joint pdf of y_1 and y_2 is

$$p_{y_1,y_2}(y_1, y_2) = \qquad (7)$$

$$\left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y_1^2}{2}\right)\right] \times \left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-y_2^2}{2}\right)\right]$$

This shows that the random variables y_1 and y_2 are independent, although they are still nonlinear mixtures of the sources s_1 and s_2.

Other examples can be found in the literature (see for example [49,41]), or can be easily constructed.

3.4. *Constraints or priors*

Constraining the mapping $\mathcal{H}$ to lie in a certain set of transformations $\mathfrak{Q}$, or adding priors, can reduce the great indeterminacies predicted by the Darmois's result. In this subsection, we briefly discuss three ideas: (i) priors on the sources, (ii) soft constraints on mappings, and (iii) hard constraints on mappings.

3.4.1. *Time correlated sources in nonlinear mixtures*

Consider two independent and identically distributed (i.i.d.) random signals $s_1(t)$ and $s_2(t)$. Using the Darmois decomposition procedure [48], one can construct new signals $y_1(t)$ and $y_2(t)$ which are statistically independent although the underlying mapping is still a mixing mapping:

$$\begin{cases} y_1(t) &= F_{X_1}(s_1(t)), \\ y_2(t) &= F_{X_1|X_2}(s_1(t), s_2(t)). \end{cases}$$

Here F_X denotes the cumulative distribution function (cdf) of the random variable X. If the sources are temporally correlated, Hosseini and Jutten [50] proved that the above mapping no longer preserves independence. Of course, this partial theoretical result does not give any proof for the separability of nonlinear mixtures of temporally correlated sources. But it shows that even fairly weak prior information on the sources can reduce the typical indeterminacies of ICA encountered in nonlinear mixtures.

3.4.2. *Smooth mappings*

Recently, multi-layer perceptron (MLP) networks (see [51,38,39]) have been used in [52,53,54] for estimating the generic nonlinear mappings $\mathcal{H}$. It is conjectured that smooth mappings provided by MLP networks are usually sufficient for ensuring that nonlinear ICA leads to nonlinear BSS, too. However, the following example [55] shows that smoothness alone is not sufficient for separation.

Without a loss of generality, consider two independent random variables $\mathbf{x} = (x_1, x_2)^T$ which are both uniformly distributed in the interval $[-1, 1]$, and the nonlinear smooth mapping

$$\mathbf{R} = \begin{pmatrix} \cos(\theta(r)) & -\sin(\theta(r)) \\ \sin(\theta(r)) & \cos(\theta(r)) \end{pmatrix}. \tag{9}$$

This describes a rotation for which the rotation angle $\theta(r)$ depends on the radius $r = (x_1^2 + x_2^2)^{1/2}$ as follows:

$$\theta(r) = \begin{cases} \theta_0(1-r)^q, & 0 \le r \le 1 \\ 0, & r > 1 \end{cases} \tag{10}$$

where $q \ge 2$. The smooth mapping $\mathbf{R}$ preserves independence since the Jacobian of the transformation is equal to 1, but it is still mixing the sources x_1 and x_2, because the Jacobian matrix of $\mathbf{R}$ is not diagonal. This counterexample shows that restricting the mapping to be smooth is not yet sufficient for achieving separation for nonlinear mixtures.

This result implies that smoothness is too a vague condition, and one has to explore further for defining sufficient conditions. In the two-dimensional special case, Hyvärinen and Pajunen gave a partial answer to this question in [41], proving that a unique solution (up to a rotation) can be obtained if the mixing mapping $\mathcal{F}$ is constrained to be a conformal mapping.

3.4.3. *Structural constraints*

A natural way of regularizing the solution consists in looking for separating mappings belonging to a specific subspace $\mathfrak{Q}$. To characterize the indeterminacies for this specific model $\mathfrak{Q}$, one must solve the independence preservation equation

$$\forall E \in \mathfrak{M}_n, \qquad \int_E dF_{s_1} dF_{s_2} \cdots dF_{s_n} = \int_{\mathcal{H}(E)} dF_{y_1} dF_{y_2} \cdots dF_{y_n} \tag{11}$$

where $\mathfrak{M}_n$ is the set of all measurable compacts in $\mathbb{R}^n$.

Let $\mathfrak{P}$ denote the set*

$$\mathfrak{P} = \{(F_{s_1}, F_{s_2}, \dots, F_{s_n}), / \exists \mathcal{H} \in \mathfrak{Q} \setminus (\mathfrak{T} \cap \mathfrak{Q}) : \mathcal{H}(\boldsymbol{s}) \text{ has independent components}\} \tag{12}$$

of all source distributions $(F_{s_1}, F_{s_2}, \dots, F_{s_n})$ for which there exists a non-trivial mapping $\mathcal{H}$ belonging to the model $\mathfrak{Q}$ †and preserving the independence of the components of the source vector $\mathbf{s}$.

Ideally, $\mathfrak{P}$ should be empty and $\mathfrak{T} \cap \mathfrak{Q}$ should contain the identity as a unique element. However, in general this is not fulfilled. We then say that source separation is possible if the distributions of the sources belong to the set $\bar{\mathfrak{P}}$, the complement of $\mathfrak{P}$. The sources are then restored up to a trivial transformation belonging to the set $\mathfrak{T} \cap \mathfrak{Q}$.

3.4.4. *Example: Linear models.*

For regular linear models, the transformation $\mathcal{F}$ is linear and can be represented by (1), where $\mathbf{A}$ is a square invertible matrix. Then, it suffices to constrain the separating model $\mathcal{G}$ to lie in the subspace of invertible square matrices, and one has to estimate a matrix $\boldsymbol{B}$ such that $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} = \boldsymbol{H}\boldsymbol{s}$ has independent components. The global transform $\mathcal{H}$ is then restricted to the subspace $\mathfrak{Q}$ of invertible square matrices.

The set of *linear* trivial transformations $\mathfrak{T} \cap \mathfrak{Q}$ is the set of matrices equal to the product of a permutation and a diagonal matrix. Moreover, from the Darmois-Skitovich theorem [45], it is clear that the set $\mathfrak{P}$ contains the distributions $(F_{s_1}, F_{s_2}, \dots, F_{s_n})$ such that at least two sources, *e.g.* F_{s_i} and F_{s_j}, are Gaussian. As stated by Comon [5], blind source separation in linear mixtures is therefore possible if we have at most one Gaussian source, and the sources can be restored up to a permutation and a diagonal matrix.

4. Constrained Nonlinear Mixtures

In this section, we consider particular nonlinear mixtures. That is, we restrict the set of nonlinear mixing models $\mathcal{F}$ and separating models $\mathcal{G}$. We especially focus on structured nonlinear mixing models satisfying an *addition theorem*. We call these models "generalized linear models" (GLM), because they can easily be transformed into linear mixing models.

*In equation (12), $\setminus$ denotes the difference between two sets

†Remember that $\mathcal{T}$ is the set of trivial mappings defined after Definition 2, in subsection 3.1

Post-nonlinear (PNL) mixtures are a particular GLM model, which has been studied since 1997 [40].

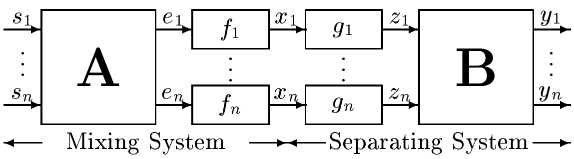

Fig. 1. The mixing-separating system for PNL mixtures.

4.1. *Separability of PNL mixtures*

In the post-nonlinear (PNL) model, the nonlinear observations have the following specific form (Fig. 1):

$$x_i(t) = f_i(\sum_{j=1}^{n} a_{ij} s_j(t)),\ i = 1, \ldots, n \qquad (13)$$

One can see that the PNL model consists of a linear mixture followed by a componentwise nonlinearity f_i acting on each output independently from the others. The nonlinear functions (distortions) f_i are assumed to be invertible. Such models appear in sensor array processing [56,57,58], satellite and microwave communications [59], and in many biological systems [60].

As discussed before, a key issue in dealing with nonlinear mixtures is their separability. First, the separation structure $\mathcal{G}$ must be constrained so that it will make possible inversion of the mixing system in the sense of Eq. (4). The separating system $\mathcal{G}$ takes then the form of the mirror structure of the mixing system $\mathcal{F}$ (see Fig. 1). The global transform $\mathcal{H}$ is then restricted to the subspace $\mathfrak{Q}$ of transforms, which consists of a cascade of an invertible linear mixture (regular matrix $\mathbf{A}$) followed by componentwise invertible distortions and again an invertible linear mixture (regular matrix $\mathbf{B}$); see Figure 1.

In [42], it has been shown that these mixtures are blindly separable for distributions having at most one Gaussian source (since the set $\mathfrak{P}$ contains the distributions having at least two Gaussian components), with the same indeterminacies as linear mixtures[‡]if $\mathbf{A}$ is regular and has (roughly speaking[§]) at least 2 nonzero entries on each row or column. Finally, in [42] it is assumed that the pdf's of the sources should be equal to zero on a small compact interval: this (technical) assumption can be removed, and the result holds if the joint density of the sources is differentiable and its derivative is continuous on the density support [61].

4.2. *Other separable nonlinear mixtures*

Due to the interesting Darmois's result for linear mixtures, it is clear that nonlinear mixtures which can be reduced to linear mixtures with a simple mapping should be separable.

4.2.1. *A simple example*

As an example, consider multiplicative mixtures:

$$x_j(t) = \prod_{i=1}^{n} s_i^{\alpha_i}(t),\ j = 1, \ldots, n \qquad (14)$$

where the $s_i(t)$ are positive independent sources. Taking the logarithm yields

$$\ln x_j(t) = \sum_{i=1}^{n} \alpha_i \ln s_i(t),\ j = 1, \ldots, n \qquad (15)$$

which is a linear model for the new independent random variables $\ln s_i(t)$. For instance, this type of mixtures can be used for modeling the dependency between the temperature and magnetic field in Hall silicon sensor [62], or gray-level images as a product of incident light and reflected light [44].

Considering in more detail the former example, the Hall voltage [63] is equal to

$$V_H = kBT^{\alpha} \qquad (16)$$

where α depends on the semiconductor type, since the temperature effect is related to the mobility of the majority carriers. Then, using two types (N and P) of sensors, we have

$$\left\{ \begin{array}{lcl} V_{H_N}(t) & = & k_N B(t) T^{\alpha_N}(t) \\ V_{H_P}(t) & = & k_P B(t) T^{\alpha_P}(t). \end{array} \right. \qquad (17)$$

For simplifying the equations, we now drop the variable t out. Because the temperature T is positive

‡The set of *linear* trivial transformations $\mathfrak{T} \cap \mathfrak{Q}$ is the set of matrices equal to the product of a permutation and a diagonal matrix.

§In fact, mixing is necessary for ensuring blind separation: if $\mathbf{A}$ is diagonal, the x_i's are already independent, and separation is already achieved but the nonlinear functions cannot be estimated. If one source s_k then appears unmixed in one sensor i, that is, $e_i = a_{ik} s_k$, there must exist an index $j \neq i$ for which $a_{jk} \neq 0$ and another index $l \neq k$ for which $a_{jl} \neq 0$.

but the sign of the magnetic field B can vary, taking the logarithm leads to the equations

$$\left\{ \begin{array}{lcl} \ln \mid V_{H_N} \mid & = & \ln k_N + \ln \mid B \mid + \alpha_N \ln T \\ \ln \mid V_{H_P} \mid & = & \ln k_P + \ln \mid B \mid + \alpha_P \ln T. \end{array} \right. \quad (18)$$

These equations describe a linear mixture of the two sources $\ln \mid B \mid$ and $\ln T$. They can be easily solved with a simple decorrelation approach since B appears with the same power in the two equations (17) and hence with the same coefficient in (18).

4.2.2. *Generalization to a class of mappings*

Extension of the Darmois-Skitovic theorem to nonlinear functions has been addressed by Kagan *et al.* in [64]. Their results have recently been revisited within the framework of BSS of nonlinear mixtures by Eriksson and Koivunen [44,7]. The main idea is to consider particular mappings $\mathcal{F}$ satisfying an *addition theorem* in the sense of the theory of functional equations. Mathematically, the addition theorem holds for mappings $\mathcal{F}$ for which there exists an invertible function f satisfying the condition

$$f(s_1 + s_2) = \mathcal{F}[f(s_1), f(s_2)]. \quad (19)$$

Example. As a simple example of such a mapping, consider the nonlinear mixtures

$$\left\{ \begin{array}{lcl} x_1 & = & (s_1 + s_2)(1 + s_1 s_2)^{-1} \\ x_2 & = & (s_1 - s_2)(1 - s_1 s_2)^{-1} \end{array} \right.$$

of two independent random variables s_1 and s_2. Using the transforms $u_1 = \tan^{-1}(s_1)$ and $u_2 = \tan^{-1}(s_2)$, the above nonlinear model becomes

$$\left\{ \begin{array}{lcl} x_1 & = & \tan(u_1 + u_2) \\ x_2 & = & \tan(u_1 - u_2). \end{array} \right.$$

Applying again the transformation $\tan^{-1}$ to x_1 and x_2 yields

$$\left\{ \begin{array}{lcl} v_1 & = & \tan^{-1}(x_1) = u_1 + u_2 \\ v_2 & = & \tan^{-1}(x_2) = u_1 - u_2 \end{array} \right.$$

which is now a linear mixture of the two independent variables u_1 and u_2. This nice result is due to the the fact that $\tan(a + b)$ is a mapping $\mathcal{F}(\tan a, \tan b)$ of $\tan a$ and $\tan b$. In fact, one has $f(.) = \tan(.)$ and $\mathcal{F}(u, v) = (u + v)(1 + uv)^{-1}$.

[¶]In the case of two variables, but extension is straightforward.

4.2.3. *Conditions for $\mathcal{F}$*

Let $u \in \mathfrak{S}$ be in the range $[a, b]$. The basic properties[¶] required for the mapping $\mathcal{F}$ are the following:

- $\mathcal{F}$ is continuous at least separately for the two variables;
- $\mathcal{F}$ is commutative: $\forall (u, v) \in \mathfrak{S}^2$, $\mathcal{F}(u, v) = \mathcal{F}(v, u)$;
- $\mathcal{F}$ is associative: $\forall (u, v, w) \in \mathfrak{S}^3$, $\mathcal{F}(\mathcal{F}(u, v), w) = \mathcal{F}(u, \mathcal{F}(v, w))$;
- There exists an identity element $e \in \mathfrak{S}$ such that $\forall u \in \mathfrak{S}$, $\mathcal{F}(u, e) = \mathcal{F}(e, u) = u$;
- $\forall u \in \mathfrak{S}$, there exists an inverse element $u^{-1} \in \mathfrak{S}$ such that $\mathcal{F}(u, u^{-1}) = \mathcal{F}(u^{-1}, u) = e$.

In other words, denoting $u \circ v = \mathcal{F}(u, v)$, these conditions imply that the set $(\mathfrak{S}, \circ)$ is an Abelian group. Under this condition, Aczel [65] proved that there exists a monotonic and continuous function $f : \mathbb{R} \to [a, b]$ such that

$$f(x + y) = \mathcal{F}(f(x), f(y)) = f(x) \circ f(y). \quad (23)$$

Clearly, applying f^{-1} (which exists since f is monotonic) to the above equation leads to

$$x + y = f^{-1}(\mathcal{F}(f(x), f(y))) = f^{-1}(f(x) \circ f(y)). \quad (24)$$

Using the above property (23), one can define a product $\star$ with integer and extend it to real variables:

$$f(cx) = c \star f(x). \quad (25)$$

Taking the inverse f^{-1} and denoting $f(x) = u$, this yields

$$c f^{-1}(u) = f^{-1}(c \star u). \quad (26)$$

Then for any constants $c_1, \ldots, c_n$ and random variables $u_1, \ldots, u_n$, the following relation holds:

$$c_1 f^{-1}(u_1) + \ldots c_n f^{-1}(u_n) = f^{-1}(c_1 \star u_1 \circ \ldots \circ c_n \star u_n) \quad (27)$$

Main result. Finally, Kagan *et al.* [64] stated the following theorem:

Theorem 1 *Let $u_1, \ldots, u_n$ be independent random variables such that*

$$\begin{cases} x_1 = a_1 \star u_1 \circ \ldots \circ a_n \star u_n \\ x_2 = b_1 \star u_1 \circ \ldots \circ b_n \star u_n \end{cases}$$

are independent, and the operators $\star$ and $\circ$ satisfy the above conditions. Denoting by f the function defined by the operator $\circ$, $f^{-1}(u_i)$ is Gaussian if $a_i b_i \neq 0$.

This theorem can be easily extended to source separation. For such mixtures, the separation algorithm consists of 3 practical steps [44,7]:

- Apply f^{-1} to the nonlinear observations for providing linear mixtures in $s_i = f^{-1}(u_i)$;
- Solve the linear mixtures in s_i by any BSS method;
- Restore the actual independent sources by applying $u_i = f(s_i)$.

4.2.4. *Practical issues*

Unfortunately, this algorithm is not blind since the function f must be known. If f is not known, a suitable separation structure is a cascade of identical nonlinear componentwise blocks (able to approximate f^{-1}) followed by a linear matrix $\mathbf{B}$ able to separate the sources in linear mixtures. This stage is further followed by identical nonlinear componentwise blocks (which approximate f) for restoring the actual sources. Note that the two first blocks of this structure are identical to the separation structure of PNL mixtures (in fact slightly simpler, since all the nonlinear blocks are similar). We can then estimate the independent distorted sources s_i with a PNL mixture separation algorithm. After computing f from the nonlinear block estimates (which approximate f^{-1}), one can then restore the actual sources. The problem is to know if one has to apply (for GLM) or not (for PNL) the restoration using f^{-1}: this can be done only if one has prior information either on the mixtures or on the distributions of the sources.

The PNL mixtures are close to these mappings. They are in fact more general since the nonlinear functions f_i can be different and unknown. Consequently, algorithms developed for separating sources in PNL mixtures (for example [42]) can be used for *blindly* separating these nonlinear mappings, avoiding the above step 1 used in [44]. Other examples of mappings satisfying the addition theorem are given in [64,44]. However, realistic mixtures belonging to this class seem unusual, except for the PNL mixtures (13) and the multiplicative mixtures (14).

Finally, the main idea in GLM or PNL mixtures consists in linearizing them by applying first a nonlinear transform. This idea is also used in the so-called kernel ICA [66,67], where the nonlinear mixture is linearized in the kernel space.

4.3. *Bounded sources in PNL mixtures*

Let us consider PNL mixtures of sources whose pdf has a bounded support, with nonzero values on the edges of the support. For example the uniform distribution or the distribution of a random sample of a sine wave satisfy this condition. For simplicity, we discuss only PNL mixtures of two sources, but the results can be easily extended to more sources. The joint distribution of the two independent sources $\mathbf{s}$ is then contained in a rectangle. The joint distribution of the linear mixtures $\mathbf{e} = \mathbf{A}\mathbf{s}$ lies inside a parallelogram. After the componentwise invertible nonlinear distortions, f_i, the joint distribution of $\mathbf{x}$ (the PNL mixtures) is contained in a "distorted" parallelogram.

Babaie-Zadeh *et al.* [46] proved that a distribution contained in a parallelogram (DCP) is transformed by componentwise invertible mappings into another DCP only if the mappings are linear. The nonlinearities g_i compensating f_i can then be estimated so that the borders of the distorted parallelogram associated with the joint distribution of the PNL mixtures become straight lines. Details of the algorithm and experimental results are given in [46]. This method proves that using a simple prior assumption, the nonlinear distortions can be estimated without using the independence assumption. In other words, bounded sources provide useful extra information for simplifying separation algorithms in PNL.

5. Separation methods for post-nonlinear mixtures

5.1. *Minimization of mutual information*

Consider now BSS methods proposed for post-nonlinear mixtures (13). Taleb and Jutten have studied this case in several papers [40,42,68], and we start with a brief discussion of their results.

The separation algorithm for the post-nonlinear mixtures (13) generally consists of two subsequent parts or stages (Fig. 1):

1. A *nonlinear stage*, which should cancel the nonlinear distortions f_i, $i = 1, \ldots, n$. To achieve this goal, nonlinear functions $g_i(\boldsymbol{\theta}_i, u)$ are applied.

2. A *linear stage* that separates the approximately linear mixtures $\mathbf{z}$ obtained after the nonlinear stage. This is carried out as usual by learning an $n \times n$ separating matrix $\mathbf{B}$ for which the components of the output vector $\mathbf{y} = \mathbf{Bz}$ of the separating system are statistically independent (or as independent as possible).

5.1.1. *Simplified equations*

As the cost function and independence criterion in both stages, Taleb and Jutten [42] use the mutual information $I(\mathbf{y})$ of the output vector:

$$I(\mathbf{y}) = \sum_i H(y_i) - H(\mathbf{y}) \tag{29}$$

Since $\mathbf{y} = \mathbf{BG}(\mathbf{x})$, where $\mathbf{B}$ is a regular matrix and $\mathbf{G}(\mathbf{x})$ is the vector $(g_1(x_1), \ldots, g_n(x_n))^T$,

$$\begin{aligned} I(\mathbf{y}) \quad = \quad & \sum_i H(y_i) - H(\mathbf{x}) - \log|\det \mathbf{B}| \\ & - \mathrm{E}\{\log|\prod_i g_i'(x_i)|\} \end{aligned} \tag{30}$$

Because $H(\mathbf{x})$ does not depend on either $\mathbf{B}$ or $\mathbf{G}(\mathbf{x})$, minimizing the mutual information is equivalent to the minimization of the cost function

$$J = \sum_i H(y_i) - \log|\det \mathbf{B}| - \mathrm{E}\{\log|\prod_i g_i'(x_i)|\} \tag{31}$$

For the linear part, minimization of the mutual information leads to the same estimation equations as for linear mixtures [1,2]

$$\frac{\partial I(\mathbf{y})}{\partial \mathbf{B}} = \mathrm{E}\{\boldsymbol{\psi}\mathbf{x}^T\} - (\mathbf{B}^T)^{-1} \tag{32}$$

where components ψ_i of the vector $\boldsymbol{\psi}$ are score functions of the components y_i of the output vector $\mathbf{y}$:

$$\psi_i(u) = -\frac{d}{du}\log p_i(u) = -\frac{p_i'(u)}{p_i(u)} \tag{33}$$

where $p_i(u)$ is the pdf of y_i. In practice, the natural gradient algorithm [12,14,15] is used for providing equivariant performance, which does not depend on the mixing matrix $\mathbf{A}$ provided that there is no noise present.

For the nonlinear stage, one can derive the gradient [42]

$$\begin{aligned} \frac{\partial I(\mathbf{y})}{\partial \boldsymbol{\theta}_k} = & -\mathrm{E}\left\{\frac{\partial \log|g_k'(\boldsymbol{\theta}_k, x_k)|}{\partial \boldsymbol{\theta}_k}\right\} \\ & - \mathrm{E}\left\{\sum_{i=1}^{n} \psi_i(y_i) b_{ik} \frac{\partial g_k(\boldsymbol{\theta}_k, x_k)}{\partial \boldsymbol{\theta}_k}\right\}. \end{aligned} \tag{34}$$

Here x_k is the kth component of the observation vector, b_{ik} is the element ik of the separating matrix $\mathbf{B}$, and g_k' is the derivative of the kth nonlinear function g_k. The exact computation algorithm naturally depends on the specific parametric form of the nonlinear mapping $g_k(\boldsymbol{\theta}_k, x_k)$, where $\boldsymbol{\theta}_k$ denotes the parameter vector. In [42], a multilayer perceptron (MLP) network is used for modeling the functions $g_k(\boldsymbol{\theta}_k, x_k)$, $k = 1, \ldots, n$, but other models, parametric [69] or not [68], can be used.

Contrary to BSS of linear mixtures, separation performance for nonlinear mixtures is strongly related to the estimation accuracy of the score functions (33) [42]. The first approach is to estimate the pdf, and then compute using differentiation the score function. Pdf estimation based on the Gram-Charlier expansion [5,1] performs appropriately only for mild post-nonlinear distortions. For hard nonlinearities, a simple pdf estimation based on kernel methods is preferable. Alternatively, the score functions can be estimated using the least-squares method [70,71], which provides very good results also for hard nonlinearities.

5.1.2. *Differential of mutual information*

Minimization of Eq. (30) is simple since it only requires estimates of marginal pdf's. However, it cannot be used for convolutive mixtures, and it may lead to biased estimates [72]. Babaie-Zadeh *et al.* have considered direct minimization of (29) in a series of papers [73,74]. In the following, we briefly present their main results.

5.1.3. *Definitions*

For a random vector $\mathbf{x}$, its marginal and joint score functions are defined by

Definition 3 (MSF) *The Marginal Score Function (MSF) of $\mathbf{x}$ is the vector of score functions of its component. It is denoted by $\psi_{\mathbf{x}}(\mathbf{x})$, and its i-th component is equal to*

$$\{\psi_{\mathbf{x}}(\mathbf{x})\}_i = -\frac{d \log p_{x_i}(x_i)}{dx_i}. \tag{35}$$

Definition 4 (JSF) *The Joint Score Function (JSF) of $\mathbf{x}$ is the gradient of $-\ln p_{\mathbf{x}}(\mathbf{x})$. It is denoted by $\phi_{\mathbf{x}}(\mathbf{x})$, and its i-th component is equal to*

$$\{\phi_{\mathbf{x}}(\mathbf{x})\}_i = -\frac{\partial \log p_{\mathbf{x}}(\mathbf{x})}{\partial x_i}. \tag{36}$$

Then, the score function difference of a random vector $\mathbf{x}$ is defined by

Definition 5 (SFD) *The Score Function Difference (SFD) of $\mathbf{x}$ is the difference between its marginal and joint score functions:*

$$\boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x}) \triangleq \psi_{\mathbf{x}}(\mathbf{x}) - \phi_{\mathbf{x}}(\mathbf{x}). \tag{37}$$

5.1.4. *Results*

It is easy to prove the following property [75], relating independence to SFD:

Property 1 *The components of a random vector $\mathbf{x} = (x_1, \ldots, x_N)^T$ are independent if and only if $\boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x}) \equiv \mathbf{0}$, that is, if and only if:*

$$\phi_{\mathbf{x}}(\mathbf{x}) = \psi_{\mathbf{x}}(\mathbf{x}). \tag{38}$$

More generally, one can compute the differential of the mutual information (MI) using the following theorem stated and proved in [75].

Theorem 2 (Differential of the MI) *Let $\mathbf{x}$ be a bounded random vector, and $\boldsymbol{\Delta}$ a 'small' random vector having the same dimensionality. Then*

$$I(\mathbf{x} + \boldsymbol{\Delta}) - I(\mathbf{x}) = E\left\{\boldsymbol{\Delta}^T \boldsymbol{\beta}_{\mathbf{x}}(\mathbf{x})\right\} + o(\boldsymbol{\Delta}) \tag{39}$$

where $o(\boldsymbol{\Delta})$ denotes higher order terms in $\boldsymbol{\Delta}$, and $\boldsymbol{\beta}_{\mathbf{x}}$ is the score function difference (SFD) of $\mathbf{x}$.

Recall that for a multivariate (differentiable) function $f(\mathbf{x})$ we have:

$$f(\mathbf{x} + \boldsymbol{\Delta}) - f(\mathbf{x}) = \boldsymbol{\Delta}^T \cdot (\nabla f(\mathbf{x})) + o(\boldsymbol{\Delta}). \tag{40}$$

A comparison of this equation with (39) shows that SFD can be understood and called as the "stochastic gradient" of the mutual information.

Finally, the following theorem (proved in [75]) points out that the mutual information does not have any local minima.

Theorem 3 *Let $\mathbf{x}_0$ be a random vector with a continuously differentiable pdf. If for any "small" random vector $\boldsymbol{\Delta}$ the condition $I(\mathbf{x}_0) \leq I(\mathbf{x}_0 + \boldsymbol{\Delta})$ holds, then $I(\mathbf{x}_0) = 0$.*

5.1.5. *Practical issues*

Using the differential of MI requires estimation of both MSF (like for minimizing simplified MI) and JSF. Of course, JSF is a multivariate function (whose dimension is equal to the source number n), and consequently its estimation may be very costly. Fast methods for computing joint entropies and score functions have been proposed by Pham [76]. The computational complexity is then multiplied by 3^{n-1}, where n is the dimension.

It is also attractive to use the fact that MI has no local minima. For this purpose, Babaie-Zadeh *et al.* [73] proposed a new class of algorithms, called Minimization-Projection (MP), which estimate $\mathbf{y} = \mathcal{F}(\mathbf{x})$ using two successive steps:

- Minimization step, where $I(\mathbf{y})$ is minimized without considering the model $\mathcal{F}$;
- Projection step, where the model $\mathcal{F}$ minimizing the least-square error $\mathrm{E}\{(\mathcal{F}(\mathbf{x}) - \mathbf{y})^2\}$ is estimated.

MP algorithms can be designed for any mixture model, including convolutive, PNL, or convolutive PNL models.

5.2. *Other methods for PNL mixtures*

Several other authors have studied methods for blind separation of post-nonlinear mixtures starting from different viewpoints. An early method proposed by Lee, Koehler, and Orglmeister in [77] is based on an extension of the natural gradient algorithm, and uses either parametric sigmoidal nonlinearities or more flexible higher-order polynomials. Achard, Pham and Jutten [69] proposed another parameterization of the mutual information criterion involving the derivatives of the nonlinearities. As these derivatives are parameterized by piecewise constant functions, the resulting algorithms are very simple.

5.2.1. *Fast estimation of nonlinearities*

The main difficulty in blind separation of post-nonlinear mixtures is the estimation of the nonlinearities. In fact, from the Central Limit Theorem of basic probability theory, one can assume that the linear mixtures $\mathbf{e}_i$ (before the nonlinear functions f_i) are approximately Gaussian [1]. This assumption holds surprisingly well in practice even for a fairly small number of sources which can have quite different pdf's. One can now utilize this result by estimating the inverses g_i of f_i so that $(g_i \circ f_i)(\mathbf{e})$ becomes again Gaussian. This idea, developed independently by Solé *et al.* [78] and Ziehe *et al.* [79], leads to very simple and fast estimation of the invertible nonlinearities g_i according to the formula

$$\hat{g}_i = \Phi^{-1} \circ F_{X_i}. \tag{41}$$

Here F_{X_i} is the cumulative distribution function (cdf) of the random variable X_i, and Φ is the cdf of Gaussian distribution. Of course, the Gaussianity assumption[‖] holds as an approximation only. But the method is very robust to this assumption, and the idea provides a simple, fast, and good starting point [80] for the PNL algorithms described earlier.

5.2.2. *Using prior information on sources*

Like for linear mixtures, one can exploit prior information on the sources for simplifying the blind separation algorithm, or improving the performance. When the sources are temporally correlated, the TDSEP method [81] or a Markov model [82] can efficiently perform BSS of PNL's.

5.2.3. *Other approaches*

Peng, Chi, and Siu [83] have introduced a semi-parametric hybrid neural network model based on the MLP network [38] for separating post-nonlinear mixtures. The main advantage of their method is that it is able to consider cross-channel post-nonlinearities, too, but experimental results have been presented on separation of two sources only. A similar generalized post-nonlinear mixture model has been addressed also in [84] by using adaptive spline neural networks.

In [85], Puntonet *et al.* have proposed combining a geometric approach with neural network learning for separating a special class of post-nonlinear mixtures, which are assumed to be powers of linear mixtures of the source signals. This geometric approach has been extended to more general nonlinear mixtures in [86]. Another geometric approach for separating bounded sources in PNL mixtures has been proposed by Babaie-Zadeh *et al.* in [46]. In this method, as well as in the recent work by Theis and Lang [87], the novelty is based on a linearization of the non linear functions f_i which does not require independence criterion. Hence, the g_i's and $\mathbf{B}$ can be estimated independently, with two different criteria.

5.3. *Extension of PNL mixtures*

Blind source separation of PNL mixtures can be generalized to convolutive PNL (CPNL) mixtures, in which the instantaneous mixtures (matrix $\mathbf{A}$) are replaced by linear filters (matrix of filters $\mathbf{A}(z)$). CPNL mixtures have been introduced by Babaie-Zadeh, Jutten and Nayebi [88] for taking into account propagation and time delays which are commonplace in many realistic situations. The observation vector for CPNL mixture is

$$x_i(k) = f_i([\mathbf{A}(z)]\mathbf{s}(k)),\ i = 1, \ldots, n. \tag{42}$$

A few separation algorithms based on the generalization of mutual information minimization for random processes have been proposed in [88,73,74].

Theis *et al.* have studied in [89] separability of a cascade of PNL stages, forming a structure similar to multi-layer perceptron networks.

A Wiener system consists of the cascade (Fig. 2) of a linear filter $[H(z)]$ followed by a memory-

[‖]It is easy to construct a counterexample, having nonlinearities fitted to the mixtures so that $f_i(e_i)$ become Gaussian. Since the sources and mixtures are not dependent, this situation is somewhat unrealistic in practice. Moreover, the estimation leads to the result $\hat{g}_i(u) = u$ which is equivalent to the starting point we use without estimating $\hat{g}_i$.

less nonlinearity f, whose input is an independent and identically distributed signal $s(k)$. The output is then $x(k) = f([H(z)]s(k))$. Using a suitably chosen parameterization, Taleb, Solé and Jutten [90] proved that Wiener systems can be expressed as PNL mixtures. Consequently, separability of PNL mixtures ensures blind invertibility of Wiener systems, and non-parametric [90] as well as parametric [91] algorithms based on minimization of mutual information rate [92] can be easily derived. A similar problem appearing in satellite communications has been solved using Monte Carlo Markov Chain (MCMC) simulation methods [93].

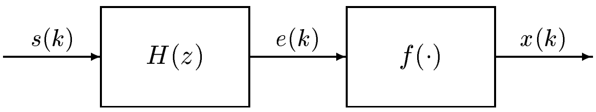

Fig. 2. A Wiener system consisting of a filter followed by a nonlinear distortion.

6. Variational Bayesian methods for general nonlinear blind source separation

6.1. *Variational Bayesian learning*

Advanced Bayesian inference methods are becoming increasingly popular both in neural networks and statistical signal processing, because one can often obtain excellent results using them provided that the assumed model is of correct type. They allow utilization of the available prior information by modeling them using suitable prior distributions, and a fully Bayesian treatment makes it possible to select an optimal model order, making such methods robust against overfitting. The main disadvantages of fully Bayesian estimation methods have been their often quite high computational load and intractable computations without approximations. These obstacles have prevented their application to realistic unsupervised or blind learning problems where the number of unknown parameters to be estimated grows easily large.

Variational Bayesian learning, also called Bayesian ensemble learning [94,95], utilizes an approximation which is fitted to the posterior distribution of the parameter(s) to be estimated. The approximative distribution is often chosen to be Gaussian because of its simplicity and computational efficiency. In our setting, the method tries to estimate the sources $\mathbf{s}$ and the mixing mapping $\mathbf{f}(\mathbf{s})$ which have most probably generated the observed data $\mathbf{x}$. Strictly speaking, according to the Bayesian philosophy, all the models of different complexity belonging to the selected model family should be taken into account and weighted according to their probabilities. But in practice it usually suffices to select the best matching model and use it. The best matching model means here the particular model which has the highest evidence of having generated the observed data [51,96].

Variational Bayesian learning can be justified using information theoretic arguments as well. The interpretation is then to look for the sources $\mathbf{s}$ and the mixing mapping $\mathbf{f}(\mathbf{s})$ which can generate the observed data $\mathbf{x}$, and have together the minimum complexity (measured in bits). The method was originally intoduced using such arguments in [94]. Further information theoretic considerations and experiments with a hierarchical nonlinear factor (source) model can be found in [97].

These probabilistic and information-theoretic principles, on which variational Bayesian (ensemble) learning relies, lay a solid and sound foundation to the method. They also provide in practice the regularization needed in nonlinear ICA and BSS. However, proving that the method provides a correct solution on some conditions would be extremely difficult or even impossible. Local minima and proper initialization require always care when iterative algorithms are applied to this kind of nonlinear problems, as will be discussed later on.

Variational Bayesian methods were first applied to standard linear ICA and BSS in [98,99]. Since then, several research groups have used Bayesian approaches to handle various blind problems for linear models; see [22,1,24,96] and the references therein. Valpola (earlier Lappalainen) and his co-authors have introduced several methods based on variational Bayesian learning for blind estimation and separation in nonlinear mixture (data) models. In the following, we consider their approach in more detail. We explain their basic method called Nonlinear Factor Analysis (NFA) mathematically, and then consider its various extensions and improvements related to nonlinear ICA and BSS.

6.2. *The nonlinear factor analysis (NFA) method*

6.2.1. *Model and cost function*

The NFA method assumes that the data can be adequately described by the noisy nonlinear mixing model

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t) \tag{43}$$

Here $\mathbf{x}(t)$ and $\mathbf{s}(t)$ denote the data and source vectors at time or index value t, $\mathbf{n}(t)$ is the noise or error vector at time t, and $\mathbf{f}(\cdot)$ is a nonlinear mixing function. The dimensions of $\mathbf{x}(t)$ and $\mathbf{s}(t)$ are in general different, and the components of the mapping $\mathbf{f}(\cdot)$ are some smooth enough scalar functions of the source vector $\mathbf{s}(t)$.

Denote by $\mathbf{X} = \{\mathbf{x}(t)|t\}$ the set of available data (mixture) vectors, by $\mathbf{S} = \{\mathbf{s}(t)|t\}$ the respective source vectors, and by $\boldsymbol{\theta}$ all the unknown parameters of the data model. Furthermore, $p(\mathbf{S},\boldsymbol{\theta}|\mathbf{X})$ denotes the exact posterior pdf and $q(\mathbf{S},\boldsymbol{\theta})$ its parametric approximation. The misfit between the approximation and the true posterior is measured by the Kullback-Leibler divergence

$$C_{\mathrm{KL}} = \int_{\mathbf{S}}\int_{\boldsymbol{\theta}} q(\mathbf{S},\boldsymbol{\theta}) \ln \frac{q(\mathbf{S},\boldsymbol{\theta})}{p(\mathbf{S},\boldsymbol{\theta}|\mathbf{X})} d\boldsymbol{\theta} d\mathbf{S} \tag{44}$$

which measures the difference in the probability mass between the pdf's $p(\cdot)$ and $q(\cdot)$. The minimum value 0 of the Kullback-Leibler divergence is achieved if and only if the two densities are the same.

The posterior distribution can be written as $p(\mathbf{S},\boldsymbol{\theta}|\mathbf{X}) = p(\mathbf{S},\boldsymbol{\theta},\mathbf{X})/p(\mathbf{X})$. The normalizing term (evidence of the data) $p(\mathbf{X})$ cannot usually be evaluated, and therefore the actual cost function used in variational Bayesian learning is

$$\begin{aligned} C &= C_{\mathrm{KL}} - \ln p(\mathbf{X}) \\ &= \int_{\mathbf{S}}\int_{\boldsymbol{\theta}} q(\mathbf{S},\boldsymbol{\theta}) \ln \frac{q(\mathbf{S},\boldsymbol{\theta})}{p(\mathbf{S},\boldsymbol{\theta},\mathbf{X})} d\boldsymbol{\theta} d\mathbf{S} \end{aligned} \tag{45}$$

The cost function (45) can be used also for model selection as explained in [95]. In the nonlinear BSS problem, it provides the necessary regularization. For each source signal and parameter, its posterior pdf is estimated instead of some point estimate. In many cases, an appropriate point estimate is given by the mean of the posterior pdf of the desired quantity, and the respective variance provides at least a rough measure of the confidence of this estimate.

6.2.2. *MLP network model*

We use the well-known multilayer perceptron (MLP) networks [51,38,39] for modeling the nonlinear mapping $\mathbf{f}$ in (43). These networks suit well to modeling both strong and mild nonlinearities. The MLP network model for $\mathbf{f}$ is

$$\mathbf{f}(\mathbf{s}(t)) = \mathbf{B}\tanh[\mathbf{A}\mathbf{s}(t) + \mathbf{a}] + \mathbf{b} \tag{46}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the weight matrices of the hidden and output layers, and $\mathbf{a}$ and $\mathbf{b}$ the corresponding bias vectors. The sigmoidal tanh nonlinearity is applied componentwise to its argument vector.

The parameters of the network are: (1) the weight matrices $\mathbf{A}$ and $\mathbf{B}$ and the bias vectors $\mathbf{a}$ and $\mathbf{b}$; (2) the parameters of the distributions of the noise, source signals and column vectors of the weight matrices; (3) the hyperparameters used for defining the distributions of the biases and the parameters in the group (2).

Since noise vectors $\mathbf{n}(t)$ are assumed to be Gaussian and white, their values at different time instants t and for different components at the same time t are independent. The likelihood of the observations $\mathbf{s}(t)$ defined by the model (43) can then be written as

$$\begin{aligned} p(\mathbf{X}|\mathbf{S},\boldsymbol{\theta}) &= \prod_{i,t} p(x_i(t)|\mathbf{s}(t),\boldsymbol{\theta}) \\ &= \prod_{i,t} N(x_i(t);\ f_i(\mathbf{s}(t)), \exp(2v_i)) \end{aligned} \tag{47}$$

where $N(x;\ \mu,\sigma^2)$ denotes a Gaussian distribution over x with mean μ and variance σ^2, $f_i(\mathbf{s}(t))$ denotes the ith component of the output of the nonlinearity $\mathbf{f}$, and v_i is a hyperparameter specifying the noise variance. The variances are parameterized using the exponential function $\exp(2v)$ where $v \sim N(\alpha,\beta)$, because this makes it easier to build hierarchical models for the prior distributions.

All the parameters of the model have hierarchical Gaussian priors. For example the noise parameters v_i of different components of the data share a common prior. The pdf's of the parameters of the model (46) are all Gaussian, and they have been specified in [100,96].

6.2.3. *Computation of the cost function*

In variational Bayesian learning each variable has its own associated posterior pdf. In our case, the variables consist of all the source signals $s_i(t)$ for different t and i as well as of the other parameters θ_j of the MLP network. These posterior pdf's are in turn characterized by their means $\bar{s}_i(t)$, $\bar{\theta}_i$ and variances $\widetilde{s}_i(t)$, $\widetilde{\theta}_i$. The total cost C is a function of all these means and variances. This is because instead of finding a point estimate, the joint posterior pdf of the sources and parameters is estimated.

Denote the two parts of the cost function (45) arising from the denominator and numerator of the logarithm by

$$C_q = \int_{\mathbf{S}} \int_{\boldsymbol{\theta}} q(\mathbf{S}, \boldsymbol{\theta}) \ln q(\mathbf{S}, \boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{S}, \tag{48}$$

$$C_p = -\int_{\mathbf{S}} \int_{\boldsymbol{\theta}} q(\mathbf{S}, \boldsymbol{\theta}) \ln p(\mathbf{S}, \boldsymbol{\theta}, \mathbf{X}) d\boldsymbol{\theta} d\mathbf{S} \tag{49}$$

For evaluating the cost function $C = C_q + C_p$, we need two things: the exact formulation of the joint probability density $p(\mathbf{S}, \boldsymbol{\theta}, \mathbf{X})$, and its parametric approximation $q(\mathbf{S}, \boldsymbol{\theta})$.

Usually the joint pdf $p(\mathbf{S}, \boldsymbol{\theta}, \mathbf{X})$ is a product of simple terms due to the definition of the model. The joint pdf can be written

$$p(\mathbf{S}, \boldsymbol{\theta}, \mathbf{X}) = p(\mathbf{X}|\mathbf{S}, \boldsymbol{\theta}) p(\mathbf{S}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{50}$$

The pdf $p(\mathbf{X}|\mathbf{S}, \boldsymbol{\theta})$ has already been evaluated in (47), and the pdf's $p(\mathbf{S}|\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are also products of univariate Gaussian distributions. They can be obtained directly from the model structure [100,96].

The cost function can be minimized efficiently if a suitably simple factorial form for the approximation is chosen. We use $q(\boldsymbol{\theta}, \mathbf{S}) = q(\boldsymbol{\theta}) q(\mathbf{S})$, where

$$q(\boldsymbol{\theta}) = \prod_i q(\theta_i) = \prod_i N(\theta_i; \overline{\theta}_i, \widetilde{\theta}_i) \tag{51}$$

is the product of the Gaussian posterior pdf approximations $N(\theta_i; \overline{\theta}_i, \widetilde{\theta}_i)$ of each parameter θ_i. The distribution $q(\mathbf{S})$ has a similar form.

The part C_q is a sum of negative entropies of Gaussians, and it has the exact form [100,96]

$$C_q = \sum_i -\frac{1}{2}[1 + \ln(2\pi\widetilde{\theta}_i)] + \sum_{t,i} -\frac{1}{2}[1 + \ln(2\pi\widetilde{s}_i(t))] \tag{52}$$

Evaluation of the cost C_p arising from the data is somewhat more complicated. However, the cost C_p also splits into a sum of simple terms, and most of them can be evaluated analytically. Only the terms involving the outputs $\mathbf{f}(\mathbf{s}(t))$ of the MLP network (46) cannot be computed exactly. These terms can be approximated by using a truncated Taylor series for $\mathbf{f}(\mathbf{s}(t))$. This is explained in detail in [100,96].

6.2.4. *Updating the parameters and sources*

The parameters of the approximating distribution are optimized using gradient based iterative algorithms. During one sweep of the algorithm all the parameters are updated once, using all the available data. One sweep consists of two different phases. The order of the computations in these two phases is the same as in standard back-propagation algorithm for MLP networks [51,38,39] but otherwise the learning procedure is quite different. The most important differences are that in the NFA method learning is unsupervised, the cost function is different, and unknown variables are characterized by distributions instead of a point estimates.

In the forward phase, the distributions of the outputs of the MLP networks are computed from the current values of the inputs. The value of the cost function is also evaluated as explained in the previous subsection. In the backward phase, the partial derivatives of the cost function with respect to all the parameters are fed back through the MLP and the parameters are updated using this information.

An update rule for the posterior variances $\widetilde{\theta}_i$ is obtained by differentiating (45) with respect to $\widetilde{\theta}_i$, yielding [100,96]

$$\frac{\partial C}{\partial \widetilde{\theta}_i} = \frac{\partial C_p}{\partial \widetilde{\theta}_i} + \frac{\partial C_q}{\partial \widetilde{\theta}_i} = \frac{\partial C_p}{\partial \widetilde{\theta}_i} - \frac{1}{2\widetilde{\theta}_i} \tag{53}$$

Equating this to zero yields a fixed-point iteration:

$$\widetilde{\theta}_i = \left[2 \frac{\partial C_p}{\partial \widetilde{\theta}_i} \right]^{-1} \tag{54}$$

The posterior means $\overline{\theta}_i$ can be estimated from the approximate Newton iteration [100,96]

$$\overline{\theta}_i \leftarrow \overline{\theta}_i - \frac{\partial C_p}{\partial \overline{\theta}_i} \left[\frac{\partial^2 C}{\partial \overline{\theta}_i^2} \right]^{-1} \approx \overline{\theta}_i - \frac{\partial C_p}{\partial \overline{\theta}_i} \widetilde{\theta}_i \tag{55}$$

It is possible that these iterative algorithms converge to a local minimum only, depending on the initialization and the successfulness of the general learning procedure, which is discussed in the next subsection. Additional stabilization is sometimes needed to ensure the convergence of the iterative algorithms (54) and (55); see [100]. Some computational improvements are given later on in this section.

6.2.5. *Overall learning procedure*

In general learning proceeds in batches. After each sweep through the data the posterior pdf approximations $q(\mathbf{S})$ and $q(\boldsymbol{\theta})$ are updated as described in the previous subsection. Because we try to learn a nonlinear model (43) with many parameters and unknowns in a completely blind manner from the data $\mathbf{X}$ only, care is required especially in the beginning of learning. Otherwise the whole procedure could converge to some false solution which is far from the correct one.

The learning procedure is therefore somewhat different in the beginning of training. The hyperparameters governing the distributions of other parameters are not updated to avoid pruning away parts of the model that do not seem useful at the moment. The posterior means of most of the parameters are initialized to random values. The posterior variances are initialized to small constant values. The posterior means of the sources $\mathbf{s}(t)$ are initialized using a suitable number of principal components of the entire data set $\mathbf{X}$. They are kept fixed to these values for the first 50 sweeps when only the MLP network $\mathbf{f}$ is updated. Updating of the hyperparameters begins after the first 100 sweeps.

The learning is continued until convergence. In variational Bayesian learning, there is no danger of overlearning or overfitting. Therefore the "optimal" number of learning sweeps is in principle infinite. Sometimes the actual number of learning sweeps can be quite large.

Assuming that the source signals $\mathbf{s}$ at the input layer of the MLP network have simple Gaussian distributions, one obtains a nonlinear principal component analysis (PCA) solution called nonlinear factor analysis (NFA) [100,101,102,1]. The NFA solution can usually model quite well the nonlinear mixtures (observed data), but it does not yet provide estimates of the independent source signals, because the sources have plain Gaussian distributions in the NFA method.

The simplest way to achieve nonlinear BSS is to apply standard linear ICA to the found NFA solution. The quality of this nonlinear BSS solution can be improved still somewhat by continuing Bayesian ensemble learning, but using now a more sophisticated mixture-of-Gaussians model for the sources. It is well-known [51] that suitable mixtures of Gaussian distributions are able to model with sufficient accuracy any smooth enough probability distribution functions. This improved method is called Nonlinear Independent Factor Analysis (NIFA).

The NFA and NIFA methods were first introduced in [100], and a more theoretical derivation has been presented in [101]. Experimental results with artificially generated data, showing that the NFA method followed by linear ICA and the NIFA method are able to approximate pretty well the true sources, have been presented in [1,100,102]. These methods have been applied also to real-world data sets, including 30-dimensional pulp data [1,100,102] and speech data [103,104], but interpretation of the results is somewhat difficult, requiring problem specific expertise.

6.3. *Extension to a dynamic model*

Somewhat later on, the NFA method was extended to include a nonlinear dynamic model for the sources in [105]. The developed NDFA (Nonlinear Dynamic Factor Analysis) method is presented thoroughly in [96], and the results obtained thus far have been summarized in [104]. The MATLAB codes for the NFA and NDFA methods are available at the www site [106].

More specifically, the data model (43) used in the NDFA method is the same as in the NFA method, but in addition the source vectors $\mathbf{s}(t)$ are assumed to obey the dynamic model

$$\mathbf{s}(t) = \mathbf{g}(\mathbf{s}(t-1)) + \mathbf{m}(t) \tag{56}$$

In (56), $\mathbf{g}(t)$ is another unknown nonlinear function which controls the dynamics of the sources $\mathbf{s}(t)$, and $\mathbf{m}(t)$ is a similar additive noise term as $\mathbf{n}(t)$ in the static nonlinear data model (43). Just as in the NFA method, the function $\mathbf{g}(t)$ is modeled by another MLP network as follows:

$$\mathbf{g}(\mathbf{s}(t-1)) = \mathbf{s}(t-1) + \mathbf{D}\tanh[\mathbf{C}\mathbf{s}(t-1) + \mathbf{c}] + \mathbf{d} \tag{57}$$

The matrices $\mathbf{C}$ and $\mathbf{D}$ contain the weights of the hidden and output layers, respectively, and the vectors $\mathbf{c}$ and $\mathbf{d}$ their biases. Because the sources $\mathbf{s}(t)$ do usually not change much from their previous values $\mathbf{s}(t-1)$ during a single time step, only their change is modeled in (57).

The unknown mappings $\mathbf{f}(t)$ and $\mathbf{g}(t)$ as well as the sources can be learned using variational Bayesian learning. The model (43), (56) and the associated NDFA learning method are discussed in detail in [96]. A special problem in NDFA is that the components of subsequent source vectors $\mathbf{s}(t)$ and $\mathbf{s}(t-1)$ are no longer completely independent due to the model (57). This is handled by using a linear dependency model for their same components $s_i(t)$ and $s_i(t-1)$, while different components are still assumed to be independent. This does not increase the computational load too much. The NDFA method is initialized using PCA of concatenated $2d+1$ subsequent data vectors $\mathbf{z}^T(t) = [\mathbf{x}^T(t+d), \ldots, \mathbf{x}^T(t-d)]$. Such embedding methods are standard tools in the analysis of nonlinear dynamic systems.

Many real-world data sets can be appropriately described as nonlinear dynamical systems such as (43), (56), and therefore nonlinear BSS for dynamical systems may in fact have more practical applications than static nonlinear BSS. The first paper about nonlinear dynamical ICA (to our knowledge) is [107], where the authors have considered state-space models and a hyper radial-basis function (RBF) network [51,39,38] for nonlinear mixtures. However, the method introduced in [107] is not completely blind, because it partly resorts to supervised learning.

In experiments with difficult chaotic data [105,96], the NDFA method performed excellently, outperforming for example the prediction results given by nonlinear autoregressive modeling learned by standard back-propagation (see [38], Chapter 15) by an order of magnitude. The NDFA method has been applied also to BSS of biomedical MEG data in [108], and it provided clearly better results than standard linear ICA. In this application, the static data model (43) was a standard linear ICA model, but the dynamic model (56) was nonlinear.

The NDFA method has been used not only to blind estimation of the dynamic system and its source signals, but also to detection of changes in the states (sources) of the process. First results on this application appeared in [109]. The journal papers [104] and especially [110] contain a more thorough exposition. The method performed again much better than the compared state-of-the-art techniques of change detection.

6.4. *Computational improvements and other extensions*

A problem in particular in the NDFA method but also in the NFA and NIFA methods is that their computational load is still high in problems of realistic size in spite of the efficient Gaussian posterior pdf approximation used. Another problem is that the Bayesian ensemble learning procedure may get stuck to a local minimum and requires careful initialization. To combat these problems, a simpler block approach which neglects all posterior dependencies has been introduced in [111]. The block approach allows straightforward construction and Bayesian ensemble learning of different types of models, and it is computationally clearly more efficient and robust against local minima.

The building block approach can be used for developing and learning hierarchical extensions of linear and nonlinear ICA and BSS models containing variance sources. First results on hierarchical variance models appeared in [112], and they are described more thoroughly in [113]. In the latter journal paper, existence of variance sources in real-world signals has been demonstrated using biomedical MEG data. In [114], we have tested the block approach in nonlinear BSS. The results for artificial data are slightly worse than for the NFA method followed by linear ICA or for the NIFA method, but preliminary experiments with real-world speech data are quite encouraging.

Occasionally, the approximation used in the block method which neglects all posterior dependencies may be too simple for providing the true ICA or BSS solution, leading to inferior performance. This problem and a solution to it, using a full covariance matrix in the Gaussian approximation $q(\mathbf{s})$ of the posterior distribution of the sources, has been discussed in [115].

Ensemble learning can be accelerated also by applying an improved updating scheme for the parameters to be estimated. First results have been given in [116]. The journal paper [117] presents more conclusive comparison experiments, showing that the improved updating scheme can reduce the required convergence time by 60-85% in realistic variational Bayesian learning problems.

The truncated Taylor series expansion used to approximate the nonlinear transformation $\mathbf{f}(\mathbf{s}(t))$ in (46) may cause instability in learning with large source dimensionalities if the input variance is large. An improved approximation employing Gauss-Hermite quadrature has been introduced in [118], resulting in a new learning algorithm which is more stable with high source dimensionalities.

Instead of using the general nonlinear mixing model (43), one can invoke a specific post-nonlinear mixture model (13) with an added noise term $n_i(t)$, and develop a variational Bayesian method for such mixtures. This has been carried out in [119].

Experiments comparing this post-nonlinear NFA approach with the post-nonlinear BSS method presented in subsection 5.1 (which tries to minimize the mutual information) have been presented in [120]. They show that the latter method usually performs somewhat better, but on the other hand it is not able to handle locally non-invertible mixtures. On the contrary, the post-nonlinear NFA approach [119] can separate even such mixtures provided that they are globally invertible [120].

It is straightforward to handle missing values in variational Bayesian learning. This topic has been discussed in [121], where hierarchical nonlinear factor (source) analysis has been compared with linear factor analysis (FA) and nonlinear factor analysis (NFA) in this respect.

Instead of standard PCA, one can use so-called kernel PCA (see for example [38] for a brief description) for initializing the variational Bayesian nonlinear BSS method(s) described before. With a suitable choice of a kernel, kernel PCA can improve significantly convergence speed and signal-to-ratio of the found sources [122]. Kernel methods proposed for the nonlinear BSS problem will be briefly discussed in the next section.

Applications of the variational Bayesian methods discussed above to various real-world data sets are currently under development, in addition to the already existing applications to pulp data, speech data, and biomedical MEG data.

7. Other separation methods for general nonlinear mixtures

In this section, we briefly review current methods proposed for nonlinear ICA or BSS.

7.1. *SOM-based approaches*

The SOM (Self-Organizing Map [38,39]) based nonlinear BSS method [36,37] mentioned earlier has been successfully applied to denoising of images corrupted by multiplicative noise in a recent journal paper [123]. A comparison between standard SOM and its modified version which is more suitable for dealing with multiplicative noise is presented together with experimental separation results on test and real images.

In general, the SOM-based nonlinear BSS method is able to approximately separate the sources if their distributions are close to the uniform one, but the results become the poorer the farther away the distributions of the sources are from the uniform one [1]. Another difficulty in using SOM for nonlinear BSS or ICA is that computational complexity increases very rapidly with the number of the sources, limiting the potential application of this method to small-scale problems [1,123]. Some further results on the applicability of the SOM-based method to linear and nonlinear BSS in simple cases have been given also in [124]. Lin, Grier, and Cowan [125] have independently proposed using SOM for nonlinear ICA and BSS in a different manner by treating ICA as a local computational geometry problem. A similar approach is proposed by Lee *et al.* [126] using Curvilinear Distance Analysis for unfolding the data, before applying linear ICA.

The restriction of uniform distributions can be alleviated by using instead of SOM so-called generative topographic mapping (GTM) method, which was introduced in [127] as a principled and theoretically well founded alternative to the somewhat heuristic SOM method. A nonlinear BSS method relying on a slightly modified version of GTM was introduced in [128], and it is discussed somewhat more thoroughly in Section 17.4 of the book [1]. The method requires knowledge of the distributions of the sources which no longer need to be close to the uniform one, but the problem with the curse of the dimensionality remains.

7.2. *MLP-based approaches*

In addition to the variational Bayesian approaches of the previous subsection, MLP (Multilayer perceptron) networks [51,38,39] have been employed in several other nonlinear BSS or ICA methods as flexible models for the nonlinear mixing mapping (2). Autoassociative MLP networks [38] in which the desired output vector of the network is the same as the input mixture vector $\mathbf{x}(t)$ have been tried for this task. Both the generative model (2) and its inversion (3) are learned simultaneously, but separately without utilizing the fact that the models are connected. Autoassociative MLPs have shown some success in nonlinear data representation [38], but generally they suffer from slow learning prone to local minima.

Most works on autoassociative MLPs use point estimates for weights and sources obtained by minimizing the mean-square representation error for the data. It is then impossible to reliably choose the structure of the model, and problems with over- or underfitting can be severe. Hecht-Nielsen [129,130] proposed so-called replicator networks for universal optimal nonlinear coding of input data. Replicator networks are autoassociate MLP networks, where the data vectors are mapped onto a unit hypercube so that the mapped data is uniformly distributed inside the hypercube. The coordinates of the mapped data on the axes of the hypercube, called natural coordinates, form then in fact a nonlinear ICA solution, even though this has not been noticed in the original papers [129,130].

Hochreiter and Schmidhuber [131] have used in context with MLP networks a method based on minimum description length, called LOCOCODE. This method does estimate the distribution of the weights, but it has no model for the sources. It is then impossible to measure the description length of the sources. Anyway, the experimental results yielded by the LOCOCODE method show interesting connections with ICA; sometimes the method provides a nonlinear ICA solution, sometimes it does not [131].

Several authors have applied in various ways a well-known information theoretic criterion, the mutual information (29), to measuring statistical independence in context with MLP networks. In these works, various MLP network based methods have been introduced for nonlinear blind separation.

In particular, Almeida has in several papers, starting from [52] and presenting more complete results in his recent journal papers [54,53], introduced so-called MISEP method. It applies a smoothness condition in an MLP network for achieving regularization in the ill-posed nonlinear ICA and BSS problems. The method seems to perform appropriately in most cases provided that the mixtures are not too nonlinear, even though theoretically the smoothness condition is not alone sufficient for guaranteeing separation, as explained before. However, the MISEP method has been applied to blind separation of up to 10 sources from their nonlinear mixtures in [54], as well as to a real-world problem of separating two images from a paper document in [132]. There the image from the back page shows through, superimposed on the image in the front page, with the mixtures of these two images being markedly nonlinear.

Yang, Amari, and Cichocki [133] deal with extensions of the basic natural gradient method for nonlinear BSS, and furthermore present another extension based on entropy maximization and experiments with post-nonlinear mixtures. This technique has been generalized to mixtures of sigmoidal nonlinearities, allowing an improved fitting to complicated nonlinear mixing functions, in [134]. Yet another paper suggesting a maximum entropy method for nonlinear ICA using a MLP network structure is [135].

7.3. *Kernel-based approaches*

Kernel-based methods, in particular kernel ICA [136] and kernel PCA which is a nonlinear extension of standard PCA (see again [38] for a brief description) can also be used as a starting point for developing algorithms for nonlinear ICA and BSS. A first paper on this line of research to our knowledge is [137], where kernel canonical correlation analysis has been suggested for nonlinear ICA and certain other extensions of standard linear ICA. A more efficient algorithm with a successful example of blind separation of nonlinearly mixed speech signals has been introduced in [66]. An open problem with these somewhat heuristic kernel methods is how to choose the nonlinear transformation into a higher dimensional space so that the mixtures become roughly linearly separable there. With a poor choice of a kernel, the separation results may become poor, too.

Tan, Wang, and Zurada [138] have proposed a radial basis function (RBF) neural network [51,38,39] structure for approximating the separating mapping (3). Their contrast function consists of the mutual information as well as of partial moments of the estimated separated sources, which are used to provide the regularization needed in nonlinear BSS. Simulation results are presented for several artificially generated nonlinear mixture sets, confirming the validity of the method introduced in [138].

7.4. *Miscellaneous approaches*

Marques and Almeida [139] have introduced a pattern repulsion method based on the maximum entropy principle for solving the nonlinear ICA or BSS problems. Their model, whose origins lie in statistical physics, has been studied theoretically in [140]. Xiong and Huang [141] propose a generalization of the classic Bell-Sejnowski algorithm [13] which uses power series of the nonlinear mixtures to approximate the Taylor expansion of the separating mapping (3). Genetic algorithms have been considered for improving the estimation of parameters of the separating mapping in [142].

A technique inspired by properties of electric fields, applicable for nonlinear ICA and some other extensions of ICA, has been suggested in [143]. It can be viewed also as a method for constructing local density approximations to the joint and factorial distributions, providing a more rigorous theoretical foundation to this method [144].

Several researchers have considered the possibility of applying their methods to nonlinear ICA or BSS without presenting a complete study including experimental results. Xu [145] has developed a general Bayesian Ying-Yang framework which could be applied to nonlinear ICA. Hinton *et al.* [146] have interpreted ICA in a novel way as an energy-based probability density model, mentioning that it is easy to extend the approach to nonlinear ICA models. A related paper is [147]. Chen and Gopinath propose an iterative gaussianization technique in [148] which could according to them provide a computationally efficient solution to nonlinear ICA. The idea of applying gaussianization methods for separating post-nonlinear mixtures has been introduced independently in [79,78], see Section 5.2.

7.5. *Local ICA and BSS*

Nonlinear independent component analysis or blind source separation are generally both computationally and conceptually difficult problems. Therefore, local linear ICA and/or BSS methods have received some attention recently as a practical compromise between linear ICA and completely nonlinear ICA or BSS. These methods are more general than standard linear ICA in that several different linear ICA models are used to describe the observed data. The local linear ICA models can be either overlapping, as in the promising Bayesian mixture-of-ICA methods introduced in [149,150], or nonoverlapping, as in the clustering-based methods proposed in [151,152].

In [153,154], variational Bayesian approach has been used for determining the number of local independent components in high-dimensional data sets, with a successful application to a difficult real-world medical data. Another recent variational Bayesian approach [155] can be used to construct hierarchical systems where ICA forms one part, as well as for efficient unsupervised partitioning of data and for sparse coding of signals using a hybrid wavelet-ICA model.

Feature detection tools from image analysis have been used for estimating local ICA coordinate systems in [156]. In [157], Lin presents interesting theoretical considerations based on local geometric structure which can be used for BSS of nonlinear mixtures. In general, only some preliminary studies have been presented on local ICA models and methods, and many issues require much more investigation.

8. Concluding remarks

In this paper, we have considered ICA and BSS problems for nonlinear data models. In that case, ICA is characterized by huge indeterminacies, and extra constraints or regularization are necessary for actually achieving solutions which coincide to source separation.

Two main results can be stated. First, solving the nonlinear BSS problem appropriately using the independence assumption alone is possible only if an appropriate regularization procedure is applied. It can consist of suitable prior information on sources, for example bounded or temporally correlated sources. In this paper, we have discussed in more detail two approaches: the first is based on structural constraints, while the second is based on variational Bayesian learning.

The first approach is the most efficient if the structural constraint is very close to the correct mixing model. The variational Bayesian ensemble learning approach is more flexible, since it uses a more general mixture model. Consequently, it can provide good source estimates even for more complex data models.

A lot of work remains to be done in studying the nonlinear ICA and BSS problems. First, further studies are needed on the separability issue. Second, these nonlinear problems are difficult and ill-posed without a suitable regularization, and we feel that all the available information should be used whenever possible. For example, it has turned out that incorporation of temporal statistics is quite helpful. Moreover, a better modeling of the relationship between the independent components or sources and the observations is essential for choosing a suitable separation structure and subsequent studies of separability. Finally, up to now, the research has addressed mainly theoretical problems. The results will become more widely interesting only if they can be validated on realistic problems using real-world data.

Acknowledgments. This work has been supported by the European Commission project BLISS (IST-1999-14190) and the Academy of Finland. The authors thank their colleagues, in particular A. Taleb, M. Babaie-Zadeh, H. Valpola, A. Honkela, and A. Ilin, for contributing many of the results presented in this paper and for useful co-operation and comments.

1. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. J. Wiley, 2001.
2. A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. J. Wiley, 2002.
3. A. Hyvärinen and P. Hoyer, “Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces,” *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.
4. M. Zibulevsky and B. Pearlmutter, “Blind source separation by sparse decomposition in a signal dictionary,” *Neural Computation*, vol. 13, pp. 863–882, 2001.
5. P. Comon, “Independent component analysis, a new concept?” *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.
6. F. Theis, “A new concept for separability problems in blind source separation,” *Neural Computation*, vol. 16, pp. 1827–1850, 2004.
7. J. Eriksson, “Contributions to theory and algorithms of independent component analysis and signal separation,” Ph.D. dissertation, Helsinki Univ. of Technology, Espoo, Finland, 2004, signal Processing Lab. Report 47.
8. J.-F. Cardoso, “Blind signal separation: statistical principles,” *Proc. of the IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.
9. C. Jutten and A. Taleb, “Source separation: from dusk till dawn,” in *Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA2000)*, Helsinki, Finland, 2000, pp. 15–26.
10. C. Jutten and J. Hérault, “Blind separation of sources, Part I: an adaptive algorithm based on a neuromimetic architecture,” *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.
11. J.-F. Cardoso and A. Souloumiac, “Blind beamforming for non gaussian signals,” *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.
12. A. Cichocki, R. Unbehauen, and E. Rummert, “Robust learning algorithm for blind separation of signals,” *Electronics Letters*, vol. 30, no. 17, pp. 1386–1387, 1994.
13. A. Bell and T. Sejnowski, “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.
14. J.-F. Cardoso and B. Laheld, “Equivariant adaptive source separation,” *IEEE Trans. on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.
15. S. Amari, A. Cichocki, and H. Yang, “A new learning algorithm for blind signal separation,” in *Advances in Neural Information Processing Systems*, Denver (Colorado), December 1996, pp. 757–763.
16. A. Hyvärinen and E. Oja, “A fast fixed-point algorithm for independent component analysis,” *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.
17. J.-F. Cardoso, C. Jutten, and P. Loubaton, Eds., *Proc. of the First Int. Workshop on Independent Component Analysis and Signal Separation (ICA99)*, Aussois, France, January 1999.
18. P. Pajunen and J. Karhunen, Eds., *Proc. of the 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, Helsinki, Finland, June 2000.
19. T.-W. Lee, S. Makeig, and T. Sejnowski, Eds., *Proc. of the 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, December 2001.
20. S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003.
21. C. Puntonet and A. Prieto, Eds., *Proc. of the 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, Granada, Spain, September 2004, *Lecture Notes in Computer Science*, vol. 3195. Springer-Verlag.
22. M. Girolami, Ed., *Advances in Independent Component Analysis*. Springer-Verlag, 2000.
23. S. Haykin, Ed., *Unsupervised Adaptive Filtering, Vol. I: Blind Source Separation*. Wiley, 2000.
24. S. Roberts and R. Everson, Eds., *Independent Component Analysis: Principles and Practice*. Cambridge

Univ. Press, 2001.

25. L. Tong, V. Soon, Y. Huang, and R. Liu, "AMUSE: a new blind identification algorithm," in *Proc. of IEEE Int. Symp. on Circuits and Systems (ISCAS'90)*, New Orleans, LA, USA, 1990.
26. A. Belouchrani, K. A. Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique based on second order statistics," *IEEE Trans. on Signal Processing*, vol. 45, no. 2, pp. 434–444, 1997.
27. K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.
28. D. T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.
29. C. Jutten, "Calcul neuromimétique et traitement du signal, analyse en composantes indépendantes," Ph.D. dissertation, INPG, Univ. Grenoble, France, 1987, (in French).
30. G. Burel, "Blind separation of sources: A nonlinear neural algorithm," *Neural Networks*, vol. 5, no. 6, pp. 937–947, 1992.
31. G. Deco and W. Brauer, "Nonlinear higher-order statistical decorrelation by volume-conserving neural architectures," *Neural Networks*, vol. 8, pp. 525–535, 1995.
32. L. Parra, G. Deco, and S. Miesbach, "Redundancy reduction with information-preserving nonlinear maps," *Network*, vol. 6, pp. 61–72, 1995.
33. L. Parra, "Symplectic nonlinear component analysis," in *Advances in Neural Inform. Proc. Systems*. Cambridge, Mass.: MIT Press, 1996, vol. 8, pp. 437–443.
34. G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing*. Springer Verlag, 1996.
35. L. Parra, G. Deco, and S. Miesbach, "Statistical independence and novelty detection with information-preserving nonlinear maps," *Neural Computation*, vol. 8, pp. 260–269, 1996.
36. P. Pajunen, "Nonlinear independent component analysis by self-organizing maps," in *Proc. of the 1996 Int. Conf. on Artificial Neural Networks (ICANN'96)*, C. von der Malsburg et al., Ed. Bochum, Germany: Springer-Verlag, 1996, pp. 815–820.
37. P. Pajunen, A. Hyvärinen, and J. Karhunen, "Nonlinear blind source separation by self-organizing maps," in *Proc. Int. Conf. on Neural Information Processing*, Hong Kong, 1996, pp. 1207–1210.
38. S. Haykin, *Neural Networks - A Comprehensive Foundation*, 2nd ed. Prentice Hall, 1998.
39. F. Ham and I. Kostanic, *Principles of Neurocomputing for Science and Engineering*. McGraw-Hill, 2001.
40. A. Taleb and C. Jutten, "Nonlinear source separation: The post-nonlinear mixtures." in *Proc. of the 1997 European Symp. on Artificial Neural Networks (ESANN1997)*, Bruges, Belgium, April 1997, pp. 279–284.
41. A. Hyvärinen and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *Neural Networks*, vol. 12, no. 3, pp. 429–439, 1999.
42. A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2807–2820, 1999.
43. A. Taleb, "A generic framework for blind source separation in structured nonlinear models," *IEEE Trans. on Signal Processing*, vol. 50, no. 8, pp. 1819–1830, 2002.
44. J. Eriksson and V. Koivunen, "Blind identifiability of class of nonlinear instantaneous ICA models," in *Proc. of the XI European Signal Proc. Conf. (EUSIPCO 2002)*, vol. 2, Toulouse, France, September 2002, pp. 7–10.
45. G. Darmois, "Analyse des liaisons de probabilité," in *Proc. of Int. Statistics Conferences 1947*, vol. III A, Washington (D.C.), 1951, p. 231.
46. M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "A geometric approach for separating post nonlinear mixtures," in *Proc. of the XI European Signal Processing Conf. (EUSIPCO 2002)*, vol. II, Toulouse, France, 2002, pp. 11–14.
47. F. Theis and P. Gruber, "Separability of analytic postnonlinear blind source separation with bounded sources," in *Proc. European Symposium on Artificial Neural Network (ESANN2004)*, Bruges, Belgium, April 2004, pp. 217–222.
48. C. Jutten, M. Babaie-Zadeh, and S. Hosseini, "Three easy ways for separating nonlinear mixtures?" *Signal Processing*, vol. 84, no. 2, pp. 217–229, February 2004.
49. E. Lukacs, "A characterization of the Gamma distribution," *Annals of Mathematical Statistics*, no. 26, pp. 319–324, 1955.
50. S. Hosseini and C. Jutten, "On the separability of nonlinear mixtures of temporally correlated sources," *IEEE Signal Processing Letters*, vol. 10, no. 2, pp. 43–46, 2003.
51. C. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.
52. L. Almeida, "Linear and nonlinear ICA based on mutual information," in *Proc. IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (AS-SPCC)*, Lake Louise, Canada, October 2000, pp. 117–122.
53. ——, "Linear and nonlinear ICA based on mutual information - the MISEP method," *Signal Processing*, vol. 84, no. 2, pp. 231–245, 2004.
54. ——, "MISEP - linear and nonlinear ICA based on mutual information," *Journal of Machine Learning Research*, vol. 4, pp. 1297–1318, 2003.
55. M. Babaie-Zadeh, "On blind source separation in convolutive and nonlinear mixtures," Ph.D. dissertation, INPG, Grenoble, France, September 2002. [Online]. Available:

http://www.lis.inpg.fr/stages_dea_theses/theses/manuscript/babaie-zadeh.pdf

56. A. Parashiv-Ionescu, C. Jutten, A. Ionescu, A. Chovet, and A. Rusu, "High performance magnetic field smart sensor arrays with source separation," in *Proc. of the 1st Int. Conf. on Modeling and Simulation of Microsystems, MSM 98*, Santa Clara, CA, USA, April 1998, pp. 666–671.
57. A. Paraschiv-Ionescu, C. Jutten, and G. Bouvier, "Source separation based processing for smart sensor arrays," *IEEE Sensors Journal*, vol. 2, no. 6, pp. 663–673, December 2002.
58. G. Bedoya, C. Jutten, S. Bermejo, and J. Cabestany, "Improving semiconductor-based chemical sensor arrays using advanced algorithms for blind source separation," in *Proc. of the ISA/IEEE Sensors for Industry Conference (SIcon 04)*, New Orleans (USA), 27-29 January 2004, pp. 149–154.
59. S. Prakriya and D. Hatzinakos, "Blind identification of LTI-ZMNL-LTI nonlinear channel models," *IEEE Trans. on Signal Processing*, vol. 43, no. 12, pp. 3007–3013, December 1995.
60. M. Korenberg and I. Hunter, "The identification of nonlinear biological systems: LNL cascade models," *Biol. Cybernetics*, vol. 43, no. 12, pp. 125–134, December 1995.
61. S. Achard and C. Jutten, "Identifiability of post nonlinear mixtures," *IEEE Signal Processing Letters*, 2004, accepted.
62. L. Almeida, C. Jutten, and H. Valpola, "Realistic models of nonlinear mixtures," HUT, INESC, INPG, BLISS (IST1999-14190) Project report D5, May 2001.
63. R. Popovic, *Hall-Effect Devices*. Bristol: Adam Hilger, 1991.
64. A. Kagan, Y. Linnik, and C. Rao, "Extension of Darmois-Skitovic theorem to functions of random variables satisfying an addition theorem," *Communications in Statistics*, vol. 1, no. 5, pp. 471–474, 1973.
65. J. Aczel, *Lectures on Functional Equations and Their Applications*. New-York: Academic Press, 1966.
66. S. Harmeling, A. Ziehe, B. Blankertz, and K.-R. Müller, "Nonlinear blind source separation using kernel feature spaces," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, 2001, pp. 102–107.
67. ——, "Kernel-based nonlinear blind source separation," *Neural Computation*, vol. 15, pp. 1089–1124, 2003.
68. A. Taleb and C. Jutten, "Batch algorithm for source separation in post-nonlinear mixtures," in *Proc. First Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 155–160.
69. S. Achard, D. T. Pham, and C. Jutten, "Blind source separation in post nonlinear mixtures," in *Proc. of the 3rd Workshop on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego (California, USA), 2001, pp. 295–300.
70. D.-T. Pham, P. Garat, and C. Jutten, "Separation of mixtures of independent sources through a maximum likelihood approach," in *Signal Processing VI, Theories and Applications*, J. Vandewalle, R. Boite, M. Moonen, and A. Oosterlinck, Eds., vol. 2. Brussels, Belgium: Elsevier, August 1992, pp. 771–774.
71. A. Taleb and C. Jutten, "Entropy optimization, application to blind source separation," in *Proc. of the 1997 Int. Conf. on Artificial Neural Networks (ICANN97)*, Lausanne, Switzerland, October 1997, pp. 529–534.
72. S. Achard, D.-T. Pham, and C. Jutten, "Criteria for blind source separation in post nonlinear mixtures with mutual information," *Signal Processing*, 2004, accepted.
73. M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "Minimization-projection (MP) approach for blind source separation in different mixing models," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 915–920.
74. ——, "A minimization-projection (MP) approach for blind separating convolutive mixtures," in *Proc. of the IEEE 2004 Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, vol. 5, Montreal (Canada), May 2004, pp. 533–536.
75. ——, "Differential of mutual information," *IEEE Signal Processing Letters*, vol. 11, no. 1, pp. 48–51, 2004.
76. D.-T. Pham, "Fast algorithms for estimating mutual information, entropies and score functions," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 17–22.
77. T.-W. Lee, B. Koehler, and R. Orglmeister, "Blind source separation of nonlinear mixing models," in *Neural Networks for Signal Processing VII*. IEEE Press, 1997, pp. 406–415.
78. J. Solé, M. Babaie-Zadeh, C. Jutten, and D.-T. Pham, "Improving algorithm speed in PNL mixture separation and Wiener system inversion," in *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 639–644.
79. A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller, "Blind separation of post-nonlinear mixtures using gaussianizing transformations and temporal decorrelation," in *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 269–274.
80. J. Solé, C. Jutten, and D. T. Pham, "Fast approximation of nonlinearities for improving inversion algorithms of PNL mixtures and Wiener systems," *Signal*

Processing, 2004, accepted.
81. A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller, "Separation of post-nonlinear mixtures using ACE and temporal decorrelation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, 2001, pp. 433–438.
82. A. Larue, C. Jutten, and S. Hosseini, "Markovian source separation in non-linear mixtures," in *Proc. of the 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, C. Puntonet and A. Prieto, Eds., Granada, Spain, September 22-24 2004, pp. 702–709, *Lecture Notes in Computer Science*, vol. 3195. Springer-Verlag.
83. H. Peng, Z. Chi, and W. Siu, "A semi-parametric hybrid neural model for nonlinear blind signal separation," *Int. J. of Neural Systems*, vol. 10, no. 2, pp. 79–94, 2000.
84. M. Solazzi, R. Parisi, and A. Uncini, "Blind source separation in nonlinear mixtures by adaptive spline neural networks," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, 2001, pp. 254–259.
85. C. Puntonet, M. Alvarez, A. Prieto, and B. Prieto, "Separation of sources in a class of post-nonlinear mixtures," in *Proc. 6th European Symp. on Artificial Neural Networks (ESANN'98)*, Bruges, Belgium, April 1998, pp. 321–326.
86. F. Theis, C. Puntonet, and E. Lang, "Nonlinear geometrical ICA," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 275–280.
87. F. Theis and E. Lang, "Postnonlinear blind source separation via linearization identification," in *Proc. of the 2004 Int. J. Conf. on Neural Networks (IJCNN2004)*, Budapest, Hungary, July 25-29 2004, pp. 2199–2204.
88. M. Babaie-Zadeh, C. Jutten, and K. Nayebi, "Separating convolutive post non-linear mixtures," in *Proc. of the 3rd Workshop on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego (California, USA), 2001, pp. 138–143.
89. F. Theis and E. Lang, "Maximum entropy and minimal mutual information in a nonlinear model," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, 2001, pp. 669–674.
90. A. Taleb, J. Sole, and C. Jutten, "Quasi-nonparametric blind inversion of Wiener systems," *IEEE Trans. on Signal Processing*, vol. 49, no. 5, pp. 917–924, 2001.
91. J. Solé, C. Jutten, and A. Taleb, "Parametric approach to blind deconvolution of nonlinear channels," *Neurocomputing*, vol. 48, pp. 339–355, 2002.
92. T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.
93. S. Sénécal and P.-O. Amblard, "Blind equalization of a nonlinear satellite system using MCMC simulation methods," *Applied Signal Processing*, vol. 9, pp. 46–57, 2002.
94. G. Hinton and D. van Camp, "Keeping neural networks simple by minimizing the description length of the weights," in *Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, Santa Cruz, CA, USA, 1993, pp. 5–13.
95. H. Lappalainen and J. Miskin, "Ensemble learning," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Berlin: Springer-Verlag, 2000, pp. 75–92.
96. H. Valpola and J. Karhunen, "An unsupervised ensemble learning method for nonlinear dynamic state-space models," *Neural Computation*, vol. 14, no. 11, pp. 2647–2692, 2002.
97. A. Honkela and H. Valpola, "Variational learning and bits-back coding: an information theoretic view to Bayesian learning," *IEEE Trans. on Neural Networks*, vol. 15, no. 4, pp. 800–810, 2004.
98. H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.
99. H. Lappalainen, "Ensemble learning for independent component analysis," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 7–12.
100. H. Lappalainen and A. Honkela, "Bayesian nonlinear independent component analysis by multi-layer perceptrons," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Springer-Verlag, 2000, pp. 93–121.
101. H. Valpola, "Nonlinear independent ccmponent analysis using ensemble learning: Theory," in *Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, Helsinki, Finland, 2000, pp. 251–256.
102. H. Valpola, X. Giannakopoulos, A. Honkela, and J. Karhunen, "Nonlinear independent component analysis using ensemble learning: Experiments and discussion," in *Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, Helsinki, Finland, 2000, pp. 351–356.
103. A. Honkela and J. Karhunen, "An ensemble learning approach to nonlinear independent component analysis," in *Proc. of the European Conf. on Circuit Theory and Design (ECCTD'01), Vol. 1*, Espoo, Finland, August 28-31 2001, pp. 41–44.
104. H. Valpola, E. Oja, A. Ilin, A. Honkela, and J. Karhunen, "Nonlinear blind source separation by variational Bayesian learning," *IEICE Transactions (Japan)*, vol. E86-A, no. 3, pp. 532–541, March 2003.
105. H. Valpola, A. Honkela, and J. Karhunen, "An ensemble learning approach to nonlinear dynamic blind source separation using state-space models," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN'02)*,

Honolulu, HI, USA, 2002, pp. 460–465.
106. H. Valpola, A. Honkela, and X. Giannakopoulos, "MATLAB codes for the NFA and NDFA algorithms," *http://www.cis.hut.fi/projects/ica/bayes/*, 2002.
107. A. Cichocki, L. Zhang, S. Choi, and S.-I. Amari, "Nonlinear dynamic independent component analysis using state-space and neural network models," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 99–104.
108. J. Särelä, H. Valpola, R. Vigário, and E. Oja, "Dynamical factor analysis of rhythmic magnetoencephalographic activity," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, 2001, pp. 451–456.
109. A. Ilin, H. Valpola, and E. Oja, "Detecting process state changes by nonlinear blind source separation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, December 2001, pp. 704–709.
110. ——, "Nonlinear dynamical factor analysis for state change detection," *IEEE Trans. on Neural Networks*, vol. 15, no. 3, pp. 559–575, May 2004.
111. H. Valpola, T. Raiko, and J. Karhunen, "Building blocks for hierarchical latent variable models," in *Proc. 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, 2001, pp. 710–715.
112. H. Valpola, M. Harva, and J. Karhunen, "Hierarchical models of variance sources," in *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 83–88.
113. ——, "Hierarchical models of variance sources," *Signal Processing*, vol. 84, no. 2, pp. 267–282, 2004.
114. H. Valpola, T. Östman, and J. Karhunen, "Nonlinear independent factor analysis by hierarchical models," in *Proc. of 4th Int. Symp. on Independent Component Analysis and Blind Source Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 257–262.
115. A. Ilin and H. Valpola, "On the effect of the form of the posterior approximation in variational learning of ICA models," in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds., Nara, Japan, April 2003, pp. 915–920.
116. A. Honkela, "Speeding up cyclic update schemes by pattern searches," in *Proc. of the 9th Int. Conf. on Neural Inform. Processing (ICONIP'02)*, Singapore, 2002, pp. 512–516.
117. A. Honkela, H. Valpola, and J. Karhunen, "Accelerating cyclic update algorithms for parameter estimation by pattern searches," *Neural Processing Letters*, vol. 17, no. 2, pp. 191–203, 2003.
118. A. Honkela, "Approximating nonlinear transformations of probability distributions for nonlinear independent component analysis," in *Proc. of the 2004 Int. J. Conf. on Neural Networks (IJCNN2004)*, Budapest, Hungary, July 25-29 2004, pp. 2169–2174.
119. A. Ilin and A. Honkela, "Post-nonlinear independent component analysis by variational Bayesian learning," in *Proc. of the 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, C. Puntonet and A. Prieto, Eds., Granada, Spain, September 22-24 2004, pp. 766–773, *Lecture Notes in Computer Science*, vol. 3195. Springer-Verlag.
120. A. Ilin, S. Achard, and C. Jutten, "Bayesian versus constrained structure approaches for source separation in post-nonlinear mixtures," in *Proc. of the 2004 Int. J. Conf. on Neural Networks (IJCNN2004)*, Budapest, Hungary, July 25-29 2004, pp. 2188–2186.
121. T. Raiko, H. Valpola, T. Östman, and J. Karhunen, "Missing values in hierarchical nonlinear factor analysis," in *Proc. the Joint 13th Int. Conf. on Artificial Neural Networks and 10th Int. Conf. on Neural Information Processing*, O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, Eds., Turkey, Istambul, June 2003, pp. 185–188.
122. A. Honkela, S. Harmeling, L. Lundqvist, and H. Valpola, "Using kernel PCA for initialisation of variational Bayesian nonlinear blind source separation method," in *Proc. of the 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, C. Puntonet and A. Prieto, Eds., Granada, Spain, September 22-24 2004, pp. 790–797, *Lecture Notes in Computer Science*, vol. 3195. Springer-Verlag.
123. M. Haritopoulos, H. Yin, and N. Allison, "Image denoising using self-organizing map-based nonlinear independent component analysis," *Neural Networks*, vol. 15, pp. 1085–1098, 2002.
124. M. Herrmann and H. Yang, "Perspectives and limitations of self-organizing maps," in *Proc. Int. Conf. on Neural Information Processing (ICONIP'96)*, Hong Kong, 1996, pp. 1211–1216.
125. J. Lin, D. Grier, and J. Cowan, "Faithful representation of separable distributions," *Neural Computation*, vol. 9, no. 6, pp. 1305–1320, 1997.
126. J. Lee, C. Jutten, and M. Verleysen, "Non-linear ICA by using isometric dimensionality reduction," in *Proc. of the 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, C. Puntonet and A. Prieto, Eds., Granada, Spain, September 22-24 2004, pp. 710–715, *Lecture Notes in Computer Science*, vol. 3195. Springer-Verlag.
127. C. Bishop, M. Svensen, and C. Williams, "GTM: The generative topographic mapping," *Neural Computation*, vol. 10, pp. 215–234, 1998.
128. P. Pajunen and J. Karhunen, "A maximum likelihood

approach to nonlinear blind source separation," in *Proc. of the 1997 Int. Conf. on Artificial Neural Networks (ICANN'97)*, Lausanne, Switzerland, 1997, pp. 541–546.

129. R. Hecht-Nielsen, "Replicator neural networks for universal optimal source coding," *Science*, vol. 269, pp. 1860–1863, 1995.
130. ——, "Data manifolds, natural coordinates, replicator neural networks, and optimal source coding," in *Proc. Int. Conf. on Neural Information Processing (ICONIP'96)*, Hong Kong, 1996, pp. 41–45.
131. S. Hochreiter and J. Schmidhuber, "Feature extraction through LOCOCODE," *Neural Computation*, vol. 11, no. 3, pp. 679–714, 1999.
132. L. Almeida and M. Faria, "Separating a real-life nonlinear mixture of images," in *Proc. of the 5th Int. Conf. on Independent Component Analysis and Blind Signal Separation (ICA2004)*, C. Puntonet and A. Prieto, Eds., Granada, Spain, September 22-24 2004, pp. 734–741, *Lecture Notes in Computer Science*, vol. 3195. Springer-Verlag.
133. H. Yang, S.-I. Amari, and A. Cichocki, "Information-theoretic approach to blind separation of sources in non-linear mixture," *Signal Processing*, vol. 64, no. 3, pp. 291–300, 1998.
134. A. Koutras, E. Dermatas, and G. Kokkinakis, "Neural network based blind source separation of non-linear mixtures," in *Proc. of the 2001 Int. Conf. on Artificial Neural Networks (ICANN'01)*. Wien, Austria: Springer-Verlag, 2001, pp. 561–567.
135. J. Fisher and J. Principe, "Entropy manipulation of arbitrary nonlinear mappings," in *Neural Networks for Signal Processing VII*, J. P. et al. Ed. New York: IEEE Press, 1997, pp. 14–23.
136. F. Bach and M. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2002.
137. C. Fyfe and P. Lai, "ICA using kernel canonical correlation analysis," in *Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, Helsinki, Finland, 2000, pp. 279–284.
138. Y. Tan, J. Wang, and J. Zurada, "Nonlinear blind source separation using a radial basis function network," *IEEE Trans. on Neural Networks*, vol. 12, no. 1, pp. 124–134, January 2001.
139. G. Marques and L. Almeida, "Separation of nonlinear mixtures using pattern repulsion," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 277–282.
140. F. Theis, C. Bauer, C. Puntonet, and E. Lang, "Pattern repulsion revisited," in *Proc. Int. Work-Conference on Artificial and Natural Neural Networks (IWANN2001)*, J. Mira and A. Prieto, Eds. Granada, Spain: Springer-Verlag, June 2001, pp. 778–785.
141. Z. Ziong and T. Huang, "Nonlinear independent component analysis (ICA) using power series and application to blind source separation," in *Proc. 3rd Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, 2001, pp. 680–685.
142. F. Rojas, I. Rojas, R. Clemente, and C. Puntonet, "Nonlinear blind source separation using genetic algorithms," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, 2001, pp. 400–405.
143. S. Hochreiter and M. Mozer, "An electric field approach to independent component analysis," in *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, Helsinki, Finland, 2000, pp. 45–50.
144. ——, "Beyond maximum likelihood and density estimation: A sample-based criterion for unsupervised learning of complex models," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. The MIT Press, 2001, pp. 535–541.
145. L. Xu, "Temporal BYY learning for state space approach, hidden Markov model, and blind source separation," *IEEE Trans. on Signal Processing*, vol. 48, no. 7, pp. 2132–2144, 2000.
146. G. Hinton, M. Welling, Y. Teh, and S. Osindero, "A new view of ICA," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, 2001, pp. 746–751.
147. T. Marks and J. Movellan, "Diffusion networks, product of experts, and factor analysis," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, 2001, pp. 481–485.
148. S. Chen and R. Gopinath, "Gaussianization," in *Advances in Neural Inform. Proc. Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. The MIT Press, 2001, pp. 423–429.
149. T.-W. Lee, M. Lewicki, and T. Sejnowski, "ICA mixture models for unsupervised classification of non-gaussian sources and automatic context switching in blind signal separation," *IEEE Trans. Pattern Recognition and Machine Intelligence*, vol. 22, no. 10, pp. 1–12, 2000.
150. T.-W. Lee and M. Lewicki, "Image processing methods using ICA mixture models," in *Independent Component Analysis: Principles and Practice*, S. Roberts and R. Everson, Eds. Cambridge University Press, 2001, pp. 234–253.
151. J. Karhunen, S. Malaroiou, and M. Ilmoniemi, "Local linear independent component analysis based on clustering," *Int. J. of Neural Systems*, vol. 10, no. 6, pp. 439–451, 2000.
152. F. Palmieri and A. Budillon, "Multi-class independent component analysis (MUCICA)," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Springer-Verlag, 2000, pp. 145–160.

153. K. Chan, T.-W. Lee, and T. Sejnowski, "Variational learning of clusters of undercomplete nonsymmetric independent components," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, CA, USA, 2001, pp. 492–497.
154. ——, "Variational learning of clusters of undercomplete nonsymmetric independent components," *J. of Machine Learning Research*, vol. 3, pp. 99–114, August 2002.
155. S. Roberts, E. Roussos, and R. Choudrey, "Hierarcy, priors, and wavelets: structure and signal modelling using ICA," *Signal Processing*, vol. 84, no. 2, pp. 283–297, 2004.
156. J. Lin, D. Grier, and J. Cowan, "Feature extraction approach to blind source separation," in *Neural Networks for Signal Processing VII*, J. P. et al., Ed. New York: IEEE Press, 1997, pp. 398–405.
157. J. K. Lin, "Factorizing probability density functions: Generalizing ICA," in *Proc. First Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 313–318.