Bayesian versus Constrained Structure Approaches for Source Separation in Post-Nonlinear Mixtures

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

Sophie Achard
Univ. of Grenoble
Lab. of Modeling and Computation,
IMAG, C.N.R.S.
B.P. 53X, 38041 Grenoble Cedex, France
E-mail: Sophie.Achard@imag.fr

Christian Jutten
Lab. des Images et des Signaux,
INPG
46 Avenue Félix Viallet,
38031 Grenoble Cedex, France
E-mail: Christian.Jutten@inpg.fr

***Abstract*— The paper presents experimental comparison of two approaches introduced for solving the nonlinear blind source separation (BSS) problem: the Bayesian methods developed at Helsinki University of Technology (HUT), and the BSS methods introduced for post-nonlinear (PNL) mixtures at Institut National Polytechnique de Grenoble (INPG). The comparison is performed on artificial test problems containing PNL mixtures. Both the standard case when the number of sources is equal to the number of observations and the case of overdetermined mixtures are considered. A new interesting result of the experiments is that globally invertible PNL mixtures, but with non-invertible component-wise nonlinearities, can be identified and sources can be separated, which shows the relevance of exploiting more observations than sources.**

I. Introduction

Different approaches proposed for nonlinear independent component analysis (ICA) and blind source separation have been recently reviewed in [1]. However, their limitations and domains of preferable application have been studied only a little, and there are hardly any comparisons of the proposed methods. We have experimentally compared two approaches for nonlinear BSS: the Bayesian methods developed at the Neural Network Research Centre at HUT and the BSS methods introduced for the special case of post-nonlinear mixtures at INPG in France. The two methods have been derived from rather different perspectives but they can both be applied to the same nonlinear ICA problems.

The Bayesian methods of HUT deal with a general form of the nonlinear mixing model:

$$\mathbf{x}(t) = \mathbf{f}(\mathbf{s}(t)) + \mathbf{n}(t) \tag{1}$$

where $\mathbf{f}$ is a nonlinear mapping from the source signals $\mathbf{s}$ to the observed signals $\mathbf{x}$. The dimensions of the vectors $\mathbf{s}(t)$ and $\mathbf{x}(t)$ are generally different, and often the dimensionality of $\mathbf{s}(t)$ is smaller. The BSS problem is solved by finding a compact nonlinear representation of the observations $\mathbf{x}$. The sources are modeled either by the Gaussian or the mixture-of-Gaussians model and the generative model (1) is learned according to the variational Bayesian principle.

The PNL methods of INPG consider a specific form of nonlinear mixtures called post-nonlinear mixtures:

$$x_i(t) = f_i(\sum_j a_{ij} s_j(t)) \qquad i = 1, \ldots, n \tag{2}$$

where a_{ij} are the elements of the unknown mixing matrix $\mathbf{A}$ and f_i are a set of scalar to scalar functions. It is usually assumed that the vectors $\mathbf{s}(t)$, $\mathbf{x}(t)$ are of the same dimensionality and the PNL distortions f_i are invertible. In this case, the BSS problem can be solved based upon the assumption of the statistical independence of the sources [1]: Under certain conditions on the distributions of the sources (at most one Gaussian source) and the mixing structure ($\mathbf{A}$ has at least 2 nonzero entries on each row or column), PNL mixtures are separable with the same well-known indeterminacies as in the linear mixtures [2].

In contrast to the Bayesian HUT methods, the PNL methods of INPG find the inverse of the generative model (2) using the separating structure

$$s_i(t) = \sum_j b_{ij} g_j(x_j(t), \boldsymbol{\theta}_j) \qquad i = 1, \ldots, n\,.$$

The purpose of the present work is to compare how the two alternative approaches perform in the same BSS problems. Since the Bayesian methods deal with a more general nonlinear mixing, we use PNL mixtures as test problems. The important question that is addressed in the paper is what are the limitations and preferable application domains of the two approaches.

II. Bayesian and constrained structure PNL methods

A. Bayesian BSS methods developed at HUT

The Bayesian algorithms presented in [3], [4], [5], [6] assume a very general model in which the observations are generated by a nonlinear mapping $\mathbf{f}$ from the sources as shown in (1). The nonlinear mapping $\mathbf{f}$ is modeled by a multilayer perceptron (MLP) network with two layers:

$$\mathbf{f}(\mathbf{s}(t)) = \mathbf{D}\tanh(\mathbf{C}\mathbf{s}(t) + \mathbf{c}) + \mathbf{d} \tag{3}$$

where $\mathbf{C}$, $\mathbf{c}$ and $\mathbf{D}$, $\mathbf{d}$ are the weight matrices and biases of the hidden and output layers respectively. Here and in the following, functions $\tanh$ and $\exp$ are applied component-wise to the elements of their vector arguments.

Implementing the Bayesian approach, all the unknown variables $\boldsymbol{\theta}$ in the model including the sources, the noise parameters and the parameters of the MLP network are assigned hierarchical priors. For example, the noise is assumed to be independent and Gaussian, which yields the likelihood

$$\mathbf{x}(t) \;\sim\; N\,(\,\mathbf{f}(\mathbf{s}(t)),\, \exp(2\mathbf{v}_n)\,)$$

where $N(\mu, \sigma^2)$ denotes a Gaussian density with mean μ and variance σ^2. All components of the noise parameter $\mathbf{v}_n$ share a common Gaussian prior:

$$v_{n,i} \;\sim\; N\,(\,m_{v_n},\, \exp(2v_{v_n})\,)\,,$$

and the hyperparameters m_{v_n}, v_{v_n} have very flat Gaussian priors.

The goal of the Bayesian methods is to estimate the posterior pdf of all the unknown variables $\boldsymbol{\theta}$. In the considered methods, this is done by variational Bayesian learning [7] which amounts to fitting a simple, parametric approximation $q(\boldsymbol{\theta})$ to the actual posterior pdf $p(\boldsymbol{\theta}|\mathbf{X})$[1]. The misfit between the approximation and the true posterior is measured by the Kullback-Leibler divergence $D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathbf{X}))$ which yields a cost function of the form

$$\begin{aligned}\mathcal{C} &= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\mathbf{X}, \boldsymbol{\theta})} d\boldsymbol{\theta} \\ &= D(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta}|\mathbf{X})) - \log p(\mathbf{X}) \geq -\log p(\mathbf{X})\,. \qquad (4)\end{aligned}$$

As follows from (4), the cost function gives a lower bound for the model evidence $p(\mathbf{X})$ which can be used for model selection [7].

In order for the cost function to be computable in practice, the maximally factorial Gaussian approximation is used

$$q(\boldsymbol{\theta}) = \prod_i q(\theta_i) = \prod_i N(\theta_i \mid \overline{\theta}_i, \widetilde{\theta}_i)\,.$$

The update rules for the posterior means $\overline{\theta}_i$ and variances $\widetilde{\theta}_i$ can be derived by differentiating (4), which yields a fixed-point iteration for $\widetilde{\theta}_i$ and an approximate Newton iteration for $\overline{\theta}_i$ [3], [4]. The optimal posterior means $\overline{\theta}_i$ of the unknown parameters usually define the solution of the BSS problem.

The basic Bayesian algorithm presented in [3] is called nonlinear factor analysis (NFA). It uses Gaussian prior for the sources:

$$\mathbf{s}(t) \;\sim\; N\,(\,0,\, \exp(2\mathbf{v}_s)\,)$$

and therefore does not provide independent source signals. The nonlinear ICA problem can then be solved by exploiting the idea somewhat similar to Gaussianization of factors in PNL mixtures [8]: The Gaussian factors found by NFA are further rotated by FastICA [9] to obtain independent source signals. We will further refer to this nonlinear ICA method as the NFA+FastICA approach.

[1]Here we denote by $\mathbf{X} = \{\mathbf{x}(t) \mid t = 1, \ldots, T\}$ the set of T available observations.

The quality of the NFA+FastICA solution can still be improved by using a mixture-of-Gaussians model for the sources. The resulting model is called nonlinear independent factor analysis (NIFA) [4] being a nonlinear counterpart of independent factor analysis [10]. We shall however use the computationally more efficient NFA+FastICA approach in the comparison.

The software implementation of the Bayesian BSS algorithms and the FastICA algorithm is available at [11] and [12] respectively.

B. PNL ICA methods developed at INPG

Assuming the PNL structure (2) of the observed signals, the BSS methods developed at INPG [2], [13] recover independent sources using a separation structure which consists of two subsequent parts:

1) A nonlinear stage, which should cancel the nonlinear distortions f_i. This part consists of nonlinear scalar functions $z_i = g_i(x_i, \boldsymbol{\theta}_i)$.
2) A linear stage that separates the approximately linear mixtures $\mathbf{z}$ obtained after the nonlinear stage. This is done as usual by learning an $n \times n$ demixing matrix $\mathbf{B}$ for which the components of the output vector $\mathbf{y} = \mathbf{B}\mathbf{z}$ of the separating system are statistically independent (or as independent as possible).

In both stages, the mutual information $I(\mathbf{y})$ between the components $y_1, \ldots, y_n$ of the output vector is used as the cost function and the independence criterion.

For the linear part, minimization of the mutual information leads to the same estimation equations as for linear mixtures:

$$\frac{\partial I(\mathbf{y})}{\partial \mathbf{B}} = -\,\mathrm{E}\left\{\boldsymbol{\psi}(\mathbf{y})\mathbf{z}^T\right\} - \left(\mathbf{B}^T\right)^{-1} \qquad (5)$$

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

$$\psi_i = \frac{d}{du} \log p_i(u) = \frac{p_i'(u)}{p_i(u)}\,.$$

Here $p_i(u)$ is the pdf and $p_i'(u)$ its derivative. The practical learning rule to estimate $\mathbf{B}$ is the natural gradient algorithm

$$\mathbf{B}(t+1) = (\mathbf{I} + \lambda\mathbf{H})\mathbf{B}(t) \qquad (6)$$

with $\mathbf{H} = \mathbf{I} + \mathrm{E}\left\{\boldsymbol{\psi}(\mathbf{y})\mathbf{y}^T\right\}$.

The gradient learning rule for the nonlinear stage can be derived from the estimating equations:

$$\begin{aligned}\frac{\partial I(\mathbf{y})}{\partial \boldsymbol{\theta}_k} = &-\,\mathrm{E}\left\{\frac{\partial \log |g_k'(x_k, \boldsymbol{\theta}_k)|}{\partial \boldsymbol{\theta}_k}\right\} \\ &-\,\mathrm{E}\left\{\sum_{i=1}^{n} \psi_i(y_i) b_{ik} \frac{\partial g_k(x_k, \boldsymbol{\theta}_k)}{\partial \boldsymbol{\theta}_k}\right\}\end{aligned} \qquad (7)$$

where x_k is the k-th component of the observation vector, b_{ik} is the element ik of the demixing matrix $\mathbf{B}$, and g_k' is the derivative of the k-th nonlinear function g_k. The exact computation algorithm depends on the specific parametric form of the nonlinear mapping $g_k(x_k, \boldsymbol{\theta}_k)$.

The equations (5), (7) for the derivatives of $I(\mathbf{y})$ contain the score functions ψ_i which are not known and must be estimated (adaptively) from the output vector $\mathbf{y}$. In the present work, we used the batch algorithm proposed in [13] with the Gaussian kernel density estimator for calculating ψ_i. We will further refer to this approach as the PNL ICA algorithm.

The speed of the algorithm can be increased by using a relevant initialization of the parameters based on the Gaussianization of $\mathbf{z}$ [14]. In the simplest approach, the k-th scalar function g_k is initialized by

$$g_k = \Phi^{-1} \circ F_k \tag{8}$$

where Φ is the Gaussian cumulative density function (cdf) and F_k is the cdf of the k-th observation x_k.

The software implementation of the PNL ICA algorithm is available at `http://www.lis.inpg.fr/demos/sep_sourc/ICAdemo/`.

III. Comparison experiments

The nonlinear BSS problems are not well known and there are interesting phenomena to uncover even in low-dimensional mixtures. Furthermore, the nonlinear methods are usually computationally quite expensive and often suffer from the local minima problem, which usually requires several runs of an algorithm using different initializations of the estimated parameters. Therefore, we used PNL mixtures of only two independent sources in most of the presented experiments. In the experiments with two sources, the sources were a sine wave and uniformly distributed white noise.

Additionally, the two BSS approaches were compared on PNL mixtures of four sub-Gaussian sources which were uniformly distributed white noise.

The test data $\mathbf{x}$ were generated by mixing the simulated source signals using a randomly chosen mixing matrix $\mathbf{A}$ and applying component-wise nonlinear distortions f_i to the linear mixture as in (2). The data were then centered and normalized to unit variance. In the experiments with noisy mixtures, the observation noise with standard deviation σ_n was added. The number of samples was 400 in all experiments.

The generated test data were processed by the two alternative methods. The number of hidden sources was always set to the known value ($m = 2$ or $m = 4$) for both algorithms.

In the PNL ICA algorithm, different values of the spread parameter σ of the Gaussian kernel density estimator were used. The adaptation steps were $\mu = 0.1$ for the outputs of the nonlinear stage and $\lambda = 0.1$ for the demixing matrix $\mathbf{B}$ (see [13] for more details). The results with the maximum signal-to-noise (SNR) ratio of the recovered sources were used in the comparison.

The NFA solution was found by trying different model structures, i.e. different number of neurons in the hidden layer of the MLP (3), and several random initializations of the parameters $\overline{\theta}_i$, $\widetilde{\theta}_i$ to be optimized. The model with the smallest value of the cost function (4) was chosen as the NFA solution. The sources of the best NFA model were further rotated by FastICA as we explained earlier.

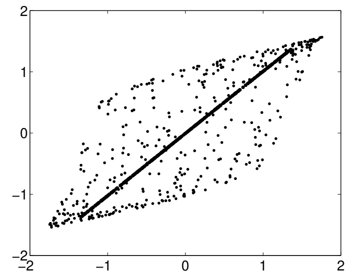


Fig. 1. Experiment A. The distribution of the observations. The thick points represent the noiseless reconstruction $\mathbf{f}(\mathbf{s})$ of the data $\mathbf{x}$ found by the NFA algorithm.

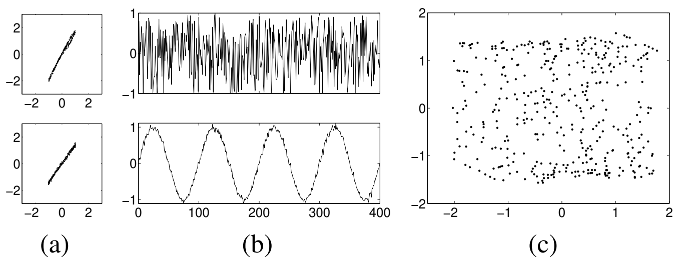


Fig. 2. Experiment A. The sources found by the PNL ICA method: (a) – the scatter plots; (b) – the estimated time series; (c) – the distribution of the sources. Signal-to-noise ratio is 20.78 dB.

A. PNL mixtures of the same dimensionality

The most common assumption of the existing linear ICA methods as well as the considered PNL ICA algorithm is that the number of observed mixtures n is equal to the number of hidden sources m. Therefore, we started the comparison experiments by applying the two compared methods to this type of mixtures.

Experiment A. The two independent sources were mixed using the following PNL mapping:

$$\mathbf{x} = \begin{bmatrix} \tanh(y_1) \\ \tanh(y_2) \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 1.0 & 0.4 \\ 0.7 & 1.0 \end{bmatrix} \mathbf{s}\,.$$

Note that all the post-nonlinear distortions were invertible and no observation noise was added to the mixtures. The distribution of the observed signals $\mathbf{x}$ is shown in Fig. 1.

The simulation results presented in Fig. 2 show that the PNL ICA method of INPG performs perfectly in this problem. The scatter plot (Fig. 2a) shows how well the original sources were reconstructed. Each point corresponds to one source $s_i(t)$. The abscissa of a point is the original source which was used for generating the data and the ordinate is the estimated source. The optimal result would be a straight line which would mean that the estimated values of the sources coincide with the true values.

In contrast, the NFA+FastICA fails to recover the original sources in this problem. The noiseless reconstruction of the data shown in Fig. 1 indicates that the best NFA model uses only one out of two sources: One of the two sources was set to zero by the algorithm, which yielded the one-dimensional

TABLE I

SIGNAL-TO-NOISE RATIO OF FOUR SUB-GAUSSIAN SOURCES RECOVERED FROM n NOISY PNL MIXTURES.

Noise variance	PNL ICA		NFA+FastICA
σ_n^2	$n=4$	$n=8$	$n=8$
0	18.73	19.18	14.53
0.01	10.47	15.42	13.27
0.1	1.76	8.05	6.9

manifold shown by the thick line in Fig. 1. The deviations from that line were explained as the observation noise $\mathbf{n}(t)$.

This result suggests that the ICA problems with the number of mixtures equal to the number of sources is difficult for the presented NFA+FastICA approach: The variational NFA algorithm usually tends to find a manifold of a lower dimension than the dimension of the observation space. This problem could be overcome by requiring the observation noise to be small (e.g. by fixing the noise variance $\mathbf{v}_s$ to small values) but this would significantly slow down the learning process.

We also applied the PNL ICA algorithm to test problems with four noisy PNL mixtures of four sub-Gaussian sources. The following post-nonlinear distortions were used for generating the data:

$$
\begin{aligned}
&f_1(y) = \tanh(y/4) \qquad && f_2(y) = 0.3y + \tanh(y) \\
&f_3(y) = 10y + y^3 \qquad && f_4(y) = -0.3y - \tanh(y)
\end{aligned} \tag{9}
$$

and the observation noise with different values of the noise variance σ_n^2 was added. The results of these experiments are presented in the second column of Table I. They demonstrate that the PNL ICA algorithm can be applied to noisy mixtures as well but its performance naturally deteriorates when the noise level increases.

B. Overdetermined PNL mixtures with invertible post-nonlinear distortions

In the following experiments, we consider PNL mixtures with the number of observations n greater than the number of sources m. This case is usually referred as the overdetermined BSS (or undercomplete bases) problem and it is particularly suitable for applying the Bayesian BSS methods. On the contrary, the standard PNL ICA algorithm requires some adjustment.

The nonlinear stage does not need any changes but the demixing matrix $\mathbf{B}$ in the linear stage should now be rectangular. However, Zhang *et al.* showed in [15] that the natural algorithm for overdetermined mixtures can be simplified to the standard learning rule (6) which can be used for estimating the rectangular $\mathbf{B}$.

In the linear stage, this approach amounts to projecting the data $\mathbf{y}$ onto the subspace defined by the initialization of $\mathbf{B}$ and seeking for independent sources within this subspace. This can be shown using the QR decomposition of $\mathbf{B}^\mathsf{T} = \mathbf{QR}$ where $\mathbf{Q}$ is a rectangular matrix such that $\mathbf{Q}^\mathsf{T}\mathbf{Q} = \mathbf{I}$. Rewriting the learning rule (6) yields

$$
\mathbf{B}(t+1) = \mathbf{R}(t+1)^\mathsf{T}\mathbf{Q}^\mathsf{T} = (\mathbf{I} + \lambda\mathbf{H})\mathbf{R}(t)^\mathsf{T}\mathbf{Q}^\mathsf{T} \, ,
$$

which shows that only the part $\mathbf{R}$ rotating the projected data $\mathbf{Q}^\mathsf{T}\mathbf{z}$ is updated.

Therefore, the initialization of $\mathbf{B}$ is of great importance for the natural gradient algorithm for overdetermined mixtures. For example, using the standard initialization of [13] $\mathbf{B}(0) = [\mathbf{I} \;\; \mathbf{0}]$ would be equivalent to using only the first m observations.

In the experiments, we used the initialization of $\mathbf{B}$ using principal component analysis (PCA): The rows of $\mathbf{B}$ were initialized with the first m principal vectors found from the normalized (and optionally Gaussianized with (8)) observations. Such initialization worked well in practice.

A possible alternative approach for the case of ICA for overdetermined PNL mixtures could be using the approximation of the entropy of the outputs $\mathbf{y}$ proposed by Stone and Porrill [16]. This would give a gradient based update rule for $\mathbf{B}$ based on the pseudo-inverse of $\mathbf{B}$ with respect to the covariance matrix of $\mathbf{z}$. However, according to our experiments, the natural gradient algorithm with the PCA initialization seemed to outperform this approach in the cases where the exact number of mixed signals is known a priori.

Experiment B. The following PNL mixtures of two independent sources were used for generating the data:

$$
\mathbf{x} = \begin{bmatrix} \tanh(y_1) \\ \tanh(y_2) \\ \tanh(y_3) \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 1.0 & 0.4 \\ 0.7 & 1.0 \\ 1.1 & 0.6 \end{bmatrix} \mathbf{s} \, .
$$

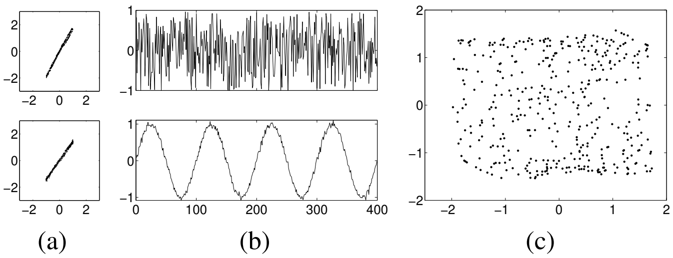
No observation noise was added.

The results provided by the PNL ICA algorithm and the Bayesian approach for this mixture are presented in Fig. 3. The PNL ICA algorithm works nicely and the quality of the source restoration is now slightly better than in Experiment A.

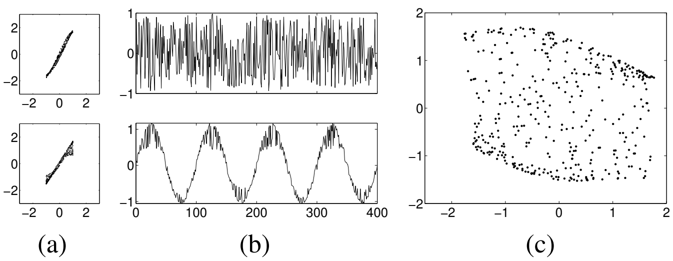
The NFA+FastICA approach is also able to retrieve the original sources but with a smaller signal-to-noise ratio than the PNL ICA method. This is a natural result since the NFA algorithm does not take into account the PNL structure of the data, and the source estimation by NFA is not based on the statistical independence of the sources.

The results of the experiments with overdetermined PNL mixtures ($n = 8$) of four sub-Gaussian sources are presented in the last two columns of Table I. The data were generated using the four post-nonlinear distortions from (9) and other four scalar functions which were same as in (9) but with the opposite sign. The results show that both methods recovered the original sources with SNR depending on the noise level. The performance of the NFA+FastICA approach is again slightly worse compared with the PNL ICA algorithm.

Note that adding more observations to noisy PNL mixtures increased the SNR obtained with PNL ICA (compare the results for $n = 4$ and $n = 8$). The results suggests that using more observations in noisy mixtures improves the performance of the PNL ICA algorithm while it is not necessarily true for noiseless mixtures.



PNL ICA: SNR is 22.14 dB



NLFA+FastICA: SNR is 15.27 dB

Fig. 3. Experiment B. The sources found by the PNL ICA (above) and NLFA+FastICA (below) approaches: (a) – the scatter plots; (b) – the estimated time series; (c) – the distribution of the sources.

C. Overdetermined PNL mixtures with non-invertible post-nonlinear distortions

The standard BSS methods for PNL mixtures reported in [13], [2] usually assume the same dimensionality of the vectors $\mathbf{x}$ and $\mathbf{s}$ and that all post-nonlinear distortions are modeled by invertible functions. Under such conditions, the entire PNL mapping is invertible and the source separation is therefore possible. However, the following experiment shows that overdetermined PNL mixtures can be invertible even if some of the post-nonlinearities are modeled by non-invertible functions. The higher dimensionality of the observed mixtures is a necessary but not sufficient condition for that.

Experiment C. The PNL mapping used for generating the data now contains one non-invertible post-nonlinearity:

$$\mathbf{x} = \begin{bmatrix} y_1^2 \\ \tanh(y_2) \\ \tanh(y_3) \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 1.0 & 0.4 \\ 0.7 & 1.0 \\ 1.1 & 0.6 \end{bmatrix} \mathbf{s}\,. \tag{10}$$

The two sources were mixed using (10) and no observation noise was added to the data. The distribution of the observed signals is presented in Fig. 4. It indicates that there exists a bijection from the two-dimensional source space to the data manifold in the three-dimensional observation space, i.e. the entire PNL mapping is invertible.

The simulation result in Fig. 4b states that the PNL ICA method could not cope with the non-invertible post-nonlinearity. If the non-invertible (first) channel is taken into consideration, the algorithm cannot provide a good solution. Note that if only the last two observations are used, the PNL ICA algorithm can recover the original sources (like in Experiment A).

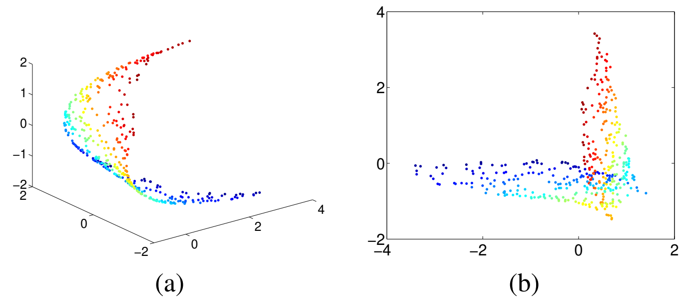


Fig. 4. Experiment C. (a) – The distribution of the observations. (b) – The distribution of the hidden sources found by PNL ICA. The PNL ICA algorithm was not able to unfold the data manifold.

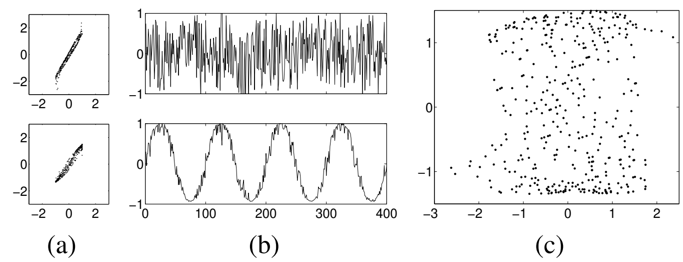


Fig. 5. Experiment C. The sources found by the NLFA+FastICA approach: (a) – the scatter plots; (b) – the estimated time series; (c) – the distribution of the sources. Signal-to-noise ratio is 13.67 dB.

In this PNL problem, the NFA+FastICA approach clearly outperforms the PNL ICA method (see the simulation results in Fig. 5) because it does not invert the post-nonlinearities but learns the entire generative model.

It is also possible to show that the entire PNL mapping can be invertible even if several post-nonlinear distortions are modeled with non-invertible functions. For example, we tested three-dimensional PNL mixtures with two quadratic distortions and the NFA+FastICA approach was able to recover the two original sources.

IV. Conclusions

Based on the experimental results from Section III, we draw the following conclusions on the applicability of the PNL ICA and Bayesian NFA+FastICA approaches to post-nonlinear ICA problems.

1) The PNL ICA method definitely performs better in classical PNL mixtures with the same number of sources and observations when all PNL distortions are invertible.
2) The performance of both methods can be improved by exploiting more mixtures than the number of sources especially in the case of noisy mixtures.
3) The performance of both methods in overdetermined mixtures largely depends on good initialization of the model parameters.
4) The advantage of the Bayesian methods in PNL problems is that they can separate PNL mixtures with non-invertible post-nonlinearities provided that the PNL

mapping is globally invertible. The existing PNL ICA methods cannot do this due to its constrained separation structure.

5) The variational Bayesian methods are computationally more expensive and usually require several runs with different initializations as they often suffer from the problem of local minima.

More generally, this preliminary study shows the relevance of exploiting more observations than sources, especially in the nonlinear mixtures. In that case, globally invertible PNL mixtures, but with non-invertible component-wise nonlinearities, can be identified and sources can be separated, which is a new and interesting result. Independently, similar results were reported by J. Lee in his PhD dissertation [17] where he applied successfully a nonlinear mapping (for reducing the dimensions) before ICA in such a model (globally but not locally invertible).

These experimental results can be further investigated, for improving understanding of the capabilities and the limitations of the two alternative approaches. In the present work, we did not try to cover all aspects of the compared methods. The important questions that can be investigated are how the algorithms scale to higher-dimensional problems with a larger number of sources, what is the performance of the methods for PNL mixtures of both sub-Gaussian and super-Gaussian sources and others.

ACKNOWLEDGMENT

The authors would like to thank their colleagues for the very fruitful discussions. This research has been funded by the European Commission project BLISS (IST-1999-14190), and the Finnish Center of Excellence Programme (2000–2005) under the project New Information Processing Principles.

REFERENCES

[1] C. Jutten and J. Karhunen, "Advances in nonlinear blind source separation," in *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, 2003, pp. 245–256, invited paper in the special session on nonlinear ICA and BSS.

[2] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2807–2820, 1999.

[3] H. Lappalainen and A. Honkela, "Bayesian nonlinear independent component analysis by multi-layer perceptrons," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Berlin: Springer-Verlag, 2000, pp. 93–121.

[4] H. Valpola, "Nonlinear independent component analysis using ensemble learning: theory," in *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, Helsinki, Finland, 2000, pp. 251–256.

[5] H. Valpola, E. Oja, A. Ilin, A. Honkela, and J. Karhunen, "Nonlinear blind source separation by variational Bayesian learning," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, no. 3, pp. 532–541, 2003.

[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. J. Wiley, 2001.

[7] H. Lappalainen and J. Miskin, "Ensemble learning," in *Advances in Independent Component Analysis*, M. Girolami, Ed. Berlin: Springer-Verlag, 2000, pp. 75–92.

[8] A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller, "Blind separation of post-nonlinear mixtures using gaussianizing transformations and temporal decorrelation," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 269–274.

[9] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[10] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, no. 4, pp. 803–851, 1999.

[11] H. Valpola, A. Honkela, and X. Giannakopoulos, "Matlab codes for the NFA and NDFA algorithms," *http://www.cis.hut.fi/projects/bayes/*, 2002.

[12] "The FastICA MATLAB package," Available at http://www.cis.hut.fi/projects/ica/fastica/, 1998.

[13] A. Taleb and C. Jutten, "Batch algorithm for source separation in post-nonlinear mixtures," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, Aussois, France, 1999, pp. 155–160.

[14] J. Solé, M. Babaie-Zadeh, C. Jutten, and D.-T. Pham, "Improving algorithm speed in PNL mixture separation and Wiener system inversion," in *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA2003)*, Nara (Japan), April 2003, pp. 639–644.

[15] L.-Q. Zhang, A. Cichocki, and S. Amari, "Natural gradient algorithm for blind separation of overdetermined mixture with additive noise," *IEEE Signal Processing Letters*, vol. 6, no. 11, pp. 293–295, 1999.

[16] J. V. Stone and J. Porrill, "Undercomplete independent component analysis for signal separation and dimension reduction," Dept. Psych. Univ. Sheffield, Sheffield, England, Tech. Rep., October 1997.

[17] J. Lee, "From principal component analysis to non-linear dimensionality reduction and blind source separation," PhD thesis, Université Catholique de Louvain-La-Neuve, December 2003.