BLIND SEQUENTIAL EXTRACTION OF POST-NONLINEARLY MIXED SOURCES USING KALMAN FILTERING

Wai Yie Leong and Danilo P. Mandic

Communications and Signal Processing Group
Department of Electronics and Electrical Engineering
Imperial College London, SW7 2AZ, UK

ABSTRACT

A novel approach which extends blind source separation (BSS) of one or group of sources to the case of post-nonlinear mixtures is proposed. This is achieved by an adaptive algorithm in which the cost function jointly estimates the kurtosis and a measure of nonlinearity. Next, Kalman filtering is applied to blindly extract the signal of interest. The analysis of the proposed approach is conducted for the case of smooth post-nonlinear mixing and simulations are provided to illustrate both the quantitative and qualitative performance of the proposed algorithm.

1. INTRODUCTION

We have recently witnessed a large research body dedicated to sequential state estimation. This approach is normally based on some sort of a state space model, and the subsequent application of Kalman filtering [9]. This type of estimation is optimal within the framework of second order statistics (SOS) and its applications are manifold. Extensions of the basic sequential state estimation problem include Extended Kalman Filter [4, 8], Unscented Kalman filter [15] and particle filtering [13, 5]. This has also been recognised in the recent special issue of the Proceedings of the IEEE on Nonlinear State Estimation [6]. Little is known, however, whether the concept of nonlinear state estimation can be successfully applied within blind source separation (BSS) [12, 2]. More specifically blind source extraction (BSE) [11], where we desire to extract only one or a few signals from their mixtures, is nothing else but a variant of nonlinear sequential estimation, whereby the sequential nature of the problem is represented by the so-called ”deflation” [11]. This is normally achieved within the framework of SOS [10].

Our aim in this paper is to investigate whether the BSE as a sequential estimation problem can be extended to the cases of nonlinear mixing, and whether we can make use the associated non-Gaussian nature of such mixtures. To that case, we propose a combination of post-nonlinear BSS followed by a deflation procedure, based on Kalman filter. The actual deflation is performed by an adjacent linear estimator, and we consider both standard Least Mean Square (LMS) based adaptive filters [7], and Kalman filter [9] in this context.

2. POST-NONLINEAR MIXTURES

Consider *n* unknown sources **s**(*k*) = $[s_1(k), \ldots, s_n(k)]^T$ with zero mean. Sources are observed through a nonlinear vector mapping **M**(·) and an ill-conditioned mixing matrix **A**, to give measurements **x**(*k*). This nonlinear mixing problem (from the unknown sources $\mathbf{s}(k)$ to the observation **x**(*k*)) can be modelled as a post-nonlinear system. We therefore assume the signals $\mathbf{x}(k)$ are nonlinear memoryless mixtures of n unknown statistically independent sources **s**(*k*), and the observation process can be expressed as

$$\mathbf{x}(k) = \mathbf{M}(\mathbf{A}\mathbf{s}(k)) \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is an unknown ill-conditioned mixing matrix which is assumed to be non-singular.

Our goal is to separate the sources of interest without any prior knowledge of their distributions and the nonlinear mixing mechanism. To that cause, we need to derive a separation structure which involves learning rule for the estimation of the unmixing (linear) matrix **W**, and a way to estimate the nonlinearity within. This unmixing operation can be expressed as

$$\tilde{\mathbf{y}}(k) = \mathbf{W}(\mathbf{M}^{-1}(\mathbf{x}(k))) \tag{2}$$

where $\tilde{\mathbf{y}}(k)$ denotes the separated output signals.

In order to extract $m \leq n$ sources, the observations (1) will be processed by an $(m \times n)$ separating matrix **W**, satisfying $\mathbf{W}\mathbf{W}^T = \mathbf{I}$, which yields the output vector (or estimated sources) $\tilde{\mathbf{y}}(k)$. The matrix **g** = ***tanh***(**WA**) denotes an $m \times n$ global demixing matrix from the sources to the outputs.

W.Y. Leong is with the Communications and Signal Processing Group, Department of Electronics and Electrical Engineering, Imperial College London, SW7 2AZ, UK. email: waiyie@ieee.org, w.leong@imperial.ac.uk

D.P. Mandic is with the Communications and Signal Processing Group, Department of Electronics and Electrical Engineering, Imperial College London, SW7 2AZ, UK. email: d.mandic@imperial.ac.uk

3. THE PROPOSED SEPARATION ALGORITHM

3.1. Nonlinear Separation Algorithm

For the separation of post-nonlinear mixtures, we propose the following ”mixed norm” criterion:

$$\boldsymbol{J}(\tilde{\mathbf{y}}(k)) = \sum_{i=1}^{n} |cum[\tilde{y}_i^4(k)]| - E\{log \sum_{i=1}^{n}[f_i(\tilde{y}_i)(k)|\} \quad (3)$$

where $f_i(\cdot)$ is the nonlinearity. The left hand side part of (3) is responsible for standard BSS, whereas the right hand part of (3) estimates the nonlinearity within the mixing process. It is important to note that (3) holds only if the functions $f_i(\cdot)$ are invertible, a restriction that must be taken into account in the development of learning algorithms.

In order to derive a learning algorithm corresponding to (3), let us consider separately the minimisation of either part of cost function (3). Let J_K correspond to the first term in (3) (kurtosis) and J_N to the second term (nonlinearity). Observe that

$$\begin{aligned} J_N(\mathbf{W}(k), \tilde{\mathbf{y}}(k)) &= \frac{\partial \sum_{i=1}^{n} log f_i(\tilde{y}_i(k))}{\partial \mathbf{W}(k)} \\ &= \frac{\partial \sum_{i=1}^{n} log f_i(\tilde{y}_i(k))}{\partial \tilde{\mathbf{y}}(k)} \frac{\partial \tilde{\mathbf{y}}(k)}{\partial \mathbf{W}(k)} \quad (4) \end{aligned}$$

where $\mathbf{f}(\tilde{\mathbf{y}}(k)) = [f_1(\tilde{y}_1(k)), f_2(\tilde{y}_2(k)), \ldots, f_n(\tilde{y}_n(k))]^T$ is the column vector whose *i*th component is

$$\begin{aligned} f_i(\tilde{y}_i(k)) &= -\frac{\partial log q_i(\tilde{y}_i(k))}{\partial \tilde{y}_i(k)} \\ &= -\frac{\partial q_i(\tilde{y}_i(k))/\partial \tilde{y}_i(k)}{q_i(\tilde{y}_i(k))} \\ &= -\frac{q_i'(\tilde{y}_i(k))}{q_i(\tilde{y}_i(k))} \quad (5) \end{aligned}$$

where $q_i(\tilde{y}_i(k))$, i = 1, ..., n, are true probability density functions of the source signals. In fact, minimising the above cost function leads to the minimisation of the mutual information [14].

On the basis of the standard gradient descent, we obtain an approximate learning rule, given by

$$\begin{aligned} \triangle \mathbf{W}(k) &= -\eta_0(k) \frac{\partial J_N}{\partial \mathbf{W}(k)} \\ &= \eta_0(k)[\mathbf{I} + \mathbf{W}(k)]\mathbf{f}(\tilde{\mathbf{y}}(k))\tilde{\mathbf{y}}^T(k) \quad (6) \end{aligned}$$

which finally yields a sequential update in the form of

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta_0(k)[\mathbf{I} + \mathbf{W}(k)]\mathbf{f}(\tilde{\mathbf{y}}(k))\tilde{\mathbf{y}}^T(k) \quad (7)$$

A classical measure of non-Gaussianity is the kurtosis, which for zero-mean random variable $\tilde{\mathbf{y}}(k)$ is defined as [3]. Hence, we can represent the term $cum[\tilde{\mathbf{y}}^4(k)]$ from the left hand side of (3) as

$$\begin{aligned} cum[\tilde{\mathbf{y}}^4(k)] &= kurt(\tilde{\mathbf{y}}(k)) \\ &= E\{\tilde{\mathbf{y}}^4(k)\} - 3(E\{\tilde{\mathbf{y}}^2(k)\})^2. \quad (8) \end{aligned}$$

it has the same value for all the output signals $\tilde{\mathbf{y}}(k)$. The normalised kurtosis, K_{norm} [10] is obtained when the kurtosis *kurt*($\tilde{\mathbf{y}}(k)$) is divided by the square of the variance $E\{\tilde{\mathbf{y}}^2(k)\}$

$$K_{norm} = \frac{E\{|\tilde{\mathbf{y}}|^4(k)\}}{E^2\{|\tilde{\mathbf{y}}|^2(k)\}} - 3 \quad (9)$$

As a cost function for kurtosis based BSS, we may employ

$$\begin{aligned} J_K(\mathbf{W}(k)) &= -\frac{1}{4}|(E\{\tilde{\mathbf{y}}^2(k)\})^2| \\ &= -\frac{\beta}{4}|(E\{\tilde{\mathbf{y}}^2(k)\})^2| \quad (10) \end{aligned}$$

and the paramter β determines the sign of the kurtosis of the signal, where

$$\beta = \begin{cases} -1, & \text{for source signal with negative kurtosis,} \\ +1, & \text{for source signal with positive kurtosis.} \end{cases} \quad (11)$$

Applying standard gradient descent to minimise the cost function, we have

$$\begin{aligned} \triangle \mathbf{W}(k) &= -\eta_0(k) \frac{\partial J_K(\mathbf{W}(k))}{\partial \mathbf{W}(k)} \\ &= \eta_0(k)\beta \frac{m_4(\tilde{\mathbf{y}}(k))}{m_2^3(\tilde{\mathbf{y}}(k))} \\ &\quad \times \left[\frac{m_2(\tilde{\mathbf{y}}(k))}{m_4(\tilde{\mathbf{y}}(k))} E\{\tilde{\mathbf{y}}^3(k)\mathbf{x}(k)\} - E\{\tilde{\mathbf{y}}(k)\mathbf{x}(k)\}\right] \quad (12) \end{aligned}$$

where $\eta_0(k) > 0$. It should be noted that the term

$$E\{|\tilde{\mathbf{y}}(k)|^4\}/E^3\{|\tilde{\mathbf{y}}(k)|^2\} = m_4(\tilde{\mathbf{y}}(k))/m_2^3(\tilde{\mathbf{y}}(k)) \quad (13)$$

is always positive, and can be absorbed by the learning rate $\widetilde{\eta}_0(k) = \frac{m_4(\tilde{\mathbf{y}}(k))}{m_2^3(\tilde{\mathbf{y}}(k))}\eta_0(k) > 0$.

As a special case, applying a simple Euler approximation to the update (12), yields the discrete-time learning rule

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta_0(k)\mathbf{f}(\tilde{\mathbf{y}}(k))\mathbf{x}(k) \quad (14)$$

where $\mathbf{x}(k)$ is a vector of sensor signals and $\mathbf{f}(\cdot)$ the nonlinearity.

Our proposed algorithm for BSS of ill-conditioned post-nonlinear mixtures can be derived as

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta_0(k)\mathbf{f}(\tilde{\mathbf{y}}(k))[\mathbf{x}(k) - (\mathbf{I} + \mathbf{W}(k))\tilde{\mathbf{y}}^T(k)] \quad (15)$$

where the separated outputs, $\tilde{\mathbf{y}}(k) = \mathbf{W}(\mathbf{M}^{-1}(k)\mathbf{x}(k))$. We propose to subsequently apply a sequential deflation procedure based on Kalman filter in order to refine these estimates.

4. DEFLATION METHOD

After the separation by means of cost function (3), we perform source extraction, via a post-processing stage in order to remove any remaining effects of post-nonlinear mixing. This is achieved by a combination of a Kalman filter and the so-called BSE based on linear predictor [10].

4.1. The Use of Kalman Filter

To cope with signals that are both nonlinear and nonstationary, where the dynamical range of the signal is not known beforehand, we propose to use Kalman filter. In deriving the equations for the Kalman filter, we begin with finding an equation that computes a deflated output $\mathbf{y}(k+1)$ as a combination of an *a priori* estimate $\tilde{\mathbf{y}}(k)$ and a weighted difference between an actual measurement $\mathbf{ds}(k)$ and a measurement prediction $\boldsymbol{H}\tilde{\mathbf{y}}(k)$, where $\boldsymbol{H}$ is an identity matrix, as shown below.

$$\mathbf{y}(k+1) = \mathbf{y}(k) - \mathbf{G}(\mathbf{ds}(k) - \boldsymbol{H}\tilde{\mathbf{y}}(k)) \qquad (16)$$

The difference $(\mathbf{ds}(k) - \boldsymbol{H}\tilde{\mathbf{y}}(k))$ in (16) is called the measurement innovation, or the residual. The residual reflects the discrepancy between the predicted measurement $\boldsymbol{H}\tilde{\mathbf{y}}(k)$ and the actual measurement $\mathbf{ds}(k)$. A residual of zero means that the two are in complete agreement.

The $n \times m$ matrix $\mathbf{G}$ in (16) is the gain or blending factor that minimizes the *a posteriori* error covariance, given by

$$\boldsymbol{P}(k) = E[e(k)e(k)^T] \qquad (17)$$

where the estimation error

$$e(k) \equiv \mathbf{ds}(k) - \tilde{\mathbf{y}}(k) \qquad (18)$$

This minimization can be accomplished by first substituting (16) into the above definition for $e(k)$, substituting that into (17), performing the indicated expectations, taking the derivative of the trace of the result with respect to $\mathbf{G}$, and setting that result equal to zero. This way, resulting $\mathbf{G}$ that minimizes (17) is given by

$$\begin{aligned}\mathbf{G}(k+1) &= \boldsymbol{P}(k)\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}(k)\boldsymbol{H}^T + \boldsymbol{R})^{-1} && (19)\\ &= \frac{\boldsymbol{P}(k)\boldsymbol{H}^T}{\boldsymbol{H}\boldsymbol{P}(k)\boldsymbol{H}^T + \boldsymbol{R}} && (20)\end{aligned}$$

From (19) we see that as the measurement error covariance $\boldsymbol{R}$ approaches zero, the gain $\mathbf{G}$ weighs the residual more heavily. More specifically,

$$\lim_{\boldsymbol{R}(k)\to 0} \mathbf{G}(k) = \boldsymbol{H}^{-1} \qquad (21)$$

On the other hand, as the a priori estimate error covariance $\boldsymbol{P}(k)$ approaches zero, the gain $\mathbf{G}$ weights the residual less heavily. Specifically,

$$\lim_{\boldsymbol{R}(k)\to 0} \mathbf{G}(k) = 0 \qquad (22)$$

Another way of approaching the weighting by $\mathbf{G}$ is that as the measurement error covariance $\boldsymbol{R}$ approaches zero, the actual measurement $\mathbf{ds}(k)$ is ”trusted” more and more, while the predicated measurement $\boldsymbol{H}\tilde{\mathbf{y}}(k)$ is trusted less and less. On the other hand, as the a priori estimate error covariance $\boldsymbol{P}(k)$ approaches zero the actual measurement $\mathbf{ds}(k)$ is trusted less and less, while the predicted measurement $\boldsymbol{H}\tilde{\mathbf{y}}(k)$ is trusted more and more.

5. EXPERIMENTS

In the experiments, the simulations were based on three source signals: s_1 with binary distribution, s_2 with sine waveform, and s_3 with Gaussian distribution. The input for all signals was scaled to range [0, 2], with positive kurtosis ($\beta = 1$), learning rates $\eta_0(k)$. Monte Carlo simulations with 5000 iterations of independent trials were performed. The initial values of the predictor weights and the demixing matrix vector $\mathbf{W}(k)$ were randomly generated for each run. The simulations were conducted without prewhitening. In theory, by minimisation of the normalised kurtosis of the extracted signal, we will recover the first source, since it has the smallest kurtosis value (binary signal).

A 3×3 mixing matrix was randomly generated, and is given by

$$\mathbf{A} = \begin{vmatrix} -0.8623 & -0.5502 & -0.0542 \\ 0.1812 & 0.3532 & -0.2561 \\ -0.5511 & -0.4358 & 0.9759 \end{vmatrix} \qquad (23)$$

Hence, post-nonlinear mixtures can be modelled as

$$\mathbf{x}(k) = \boldsymbol{tanh}(\mathbf{As}(k)) \qquad (24)$$

To measure the quantitative performance of the proposed algorithm, we employ the performance index (PI) defined by [1]

$$\text{PI} = 10log_{10}\left(\frac{1}{L-1}\left(\sum_{l=0}^{L-1}\frac{g_l^2}{max\{g_0^2, g_1^2, \ldots, g_{L-1}^2\}} - 1\right)\right) \qquad (25)$$

hence, the smaller the value of PI, the better the quality of extraction.

Initial simulation results presented in Fig.1 show that the proposed method has the potential to separate the post-nonlinear mixtures for which the output scatter plots are closely matched with the original sources (Fig.1). The proposed adaptive method is also likely to exhibit faster convergence and better performance index than the normalised method [10] as shown in Fig.2.

6. CONCLUSIONS

We have proposed an approach for post-nonlinear blind source extraction whereby Kalman filter is used on the deflation stage.

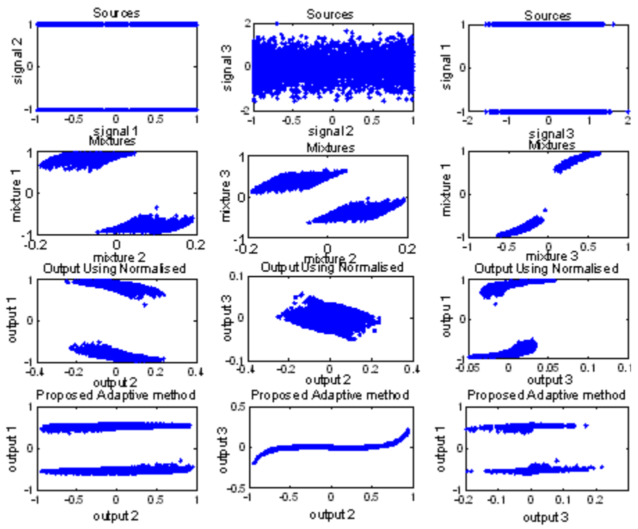

Fig. 1. Scatter plot comparing the independence of the output signals; Column 1: signal 1 vs signal 2; Column 2: signal 2 vs signal 3; Column 3: signal 1 vs signal 3.

The proposed adaptive algorithm which does not require any prepocessing (prewhitening), it is particularly suitable for blind source extraction with post-nonlinear mixing matrices. Simulation results have confirmed the validity of the theoretical results and demonstrated the performance of the algorithm.

7. REFERENCES

[1] A.Cichocki and S.I.Amari. *Adaptive Blind Signal and Image Processing*. Locally Adaptive Algorithms for ICA and Their Implementations. John Wiley Sons, Ltd, England, 2002.

[2] A.Hyvarinen and E.Oja. Independent Component Analysis: A tutorial. Technical report, Helsinki University of Technology, April 1999.

[3] A.Hyvarinen, J.Karhunen, and E.Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, Canada, 2001.

[4] A. Gelb. *Applied Optimal Estimation*. Cambridge, MA: MIT Press, 1974.

[5] S. J. Godsill and P. J. W. Rayner. *Digital Audio Restoration - a statistical model-based approach*. Berlin: Springer-Verlag, 1998.

[6] S. Haykin and N. de Freitas. *Special Issue on Sequential State Estimation*. Proceedings of the IEEE, Sept 2003.

[7] S. Haykin and B. Widrow. *Least-Mean-Square Adaptive Filters*. John Wiley Sons, New York, 2003.

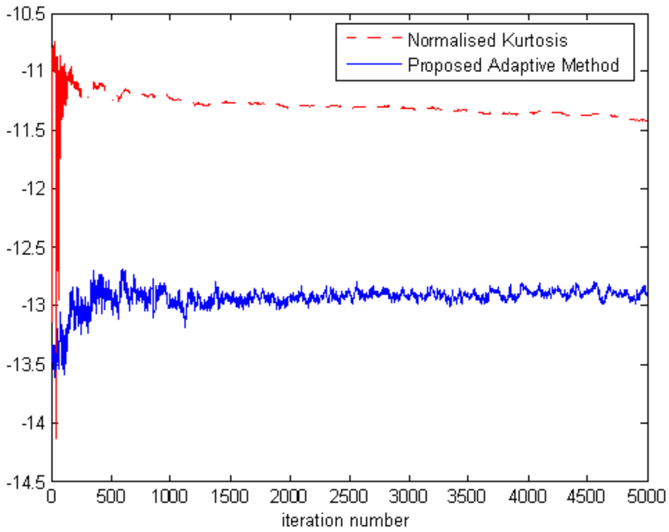

Fig. 2. Learning curve of the extraction algorithms (without prewhitening).

[8] K. Ide and M. Ghil. Extended Kalman filtering for vortex systems: I. methodology and point vortices. *Journal of Turbulence*.

[9] F. L. Lewis. *Optimal Estimation*. John Wiley Sons, New York, 1st edition, 1986.

[10] W. Liu and D. P. Mandic. A normalised kurtosis based algorithm for blind source extraction from noisy measurements. *Signal Processing (In press)*.

[11] N.Delfosse and P.Loubaton. Adaptive blind separation of independent sources: A deflation approach. *Signal Processing*, 49:59–83, 1995.

[12] P.Comon. Independent Component Analysis, a new concept? *Special Issue on Higher-Order Statistics, Signal Processing*, 36(3):287–314, April 1994.

[13] M. K. Pitt and N. Shephard. Filtering via simulation: auxiliary particle filter. *Journal Amer. Statist. Assoc.*

[14] S.Amari. Natural gradient works efficiently in learning. In *Neural Computation*, volume 10, pages 251–276, Jan 1998.

[15] E. Wan and R. van der Merwe. The unscented Kalman filter. Wiley Publishing, 2001.