BLIND SEPARATION OF NON STATIONARY NON GAUSSIAN SOURCES

Dinh-Tuan Pham
Laboratory of Modeling and Computation, IMAG, C.N.R.S.
Univ. of Grenoble, B.P. 53X, 38041 Grenoble Cedex, France
e-mail: Dinh-Tuan.Pham@imag.fr

ABSTRACT

Most blind sources separation methods are based on the non Gaussianity or the coloration of the sources and only recently their non-stationarity. This work proposes new procedures which exploit both the first and last aspects. We adopt the quasi-maximum likelihood approach which provided a set of estimating equations involving the score functions, which are then estimated by a projection method and through the idea blocking or kernel smoothing. Efficient off-line and on-line algorithms are developed. A simpler and less costly procedure based on a simple contrast for sub Gaussian sources is also considered. Some simulation experiments are given illustrating the high performance of the method.

1 INTRODUCTION

In previous works [5], we have developed a blind source separation procedure adapted to source signals with time varying intensity. This procedure however ignores the possible non Gaussianity of the source and relies only on the diversity of the variation of their intensity for separation. Here we shall it to exploit both the non Gaussianity and the non stationarity of the sources. This should improve the performance of the method as we use more relevant information. Further our procedure would work both in the case of stationary but non Gaussian and Gaussian but non stationary signals.

For simplicity, we shall restrict ourselves to the model of instantaneous mixtures without noise:

$$\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t) \tag{1}$$

where $\mathbf{X}(t) = [X_1(t) \quad \cdots \quad X_K(t)]^{\mathrm{T}}$ is the vector of observations (at time t), $\mathbf{A}$ is an unknown $K \times K$ invertible matrix and $\mathbf{S}(t) = [S_1(t) \quad \cdots \quad S_K(t)]^{\mathrm{T}}$ is the vector of sources and $^{\mathrm{T}}$ denotes the transpose. The goal is to reconstruct the sources $S_k(t)$ based *only on the assumption of their mutual independence*. No particular knowledge about their distributions is assumed and we are mostly interested in the case when these distributions change in time.

2 SEPARATION METHODS

2.1 The Quasi Maximum Likelihood Approach

This approach, introduced in Pham and Garat [6], assumes first that the distributions of the sources are known in order to write down the likelihood and derive a system of estimating equations, then modify it to obtain a practical and usable system (as the sources distributions are actually unknown). Thus, let p_{it} be the density of $S_i(t)$, assumed known for the moment. As we are not interested in exploiting the temporal dependency of the source, we further assume that the $S_i(t)$ for different t are independent. This is only a *working assumption*, made in order to write down the likelihood, as it will be apparent later that the method still works without this assumption. The negative of the log likelihood function of the model (1) then writes

$$-T\Big[\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{K}\log p_{it}(\mathbf{A}^{-1}\mathbf{X})_i(t) + \log|\det\mathbf{A}|\Big]$$

where $(\mathbf{A}^{-1}\mathbf{X})_i$ denotes the i-th component of $\mathbf{A}^{-1}\mathbf{X}$ and T is the sample length. For clarity of notation, we now write $\mathbf{B}$ (the separating matrix) in place of $\mathbf{A}^{-1}$ and regard it as a generic parameter and reserve the notation $\mathbf{A}$ for the *true* mixing matrix. Equating to zero the (relative) gradient of the log likelihood with respect to $\mathbf{B}$, we get a system of estimating equations:

$$\frac{1}{T}\sum_{t=1}^{T}\psi_{it}[(\mathbf{B}\mathbf{X})_i(t)](\mathbf{B}\mathbf{X})_j(t) = 0, \quad 1 \le i \ne j \le K \tag{2}$$

$$\frac{1}{T}\sum_{t=1}^{T}\psi_{it}[(\mathbf{B}\mathbf{X})_i(t)](\mathbf{B}\mathbf{X})_i(t) = 1, \quad 1 \le i \le K \tag{3}$$

where ψ_{it} is the derivative of $-\log p_{it}$ (called the score function of $S_i(t)$). Equations (2) can be seen to be related to the independence of the sources, since they are satisfied when the $(\mathbf{B}\mathbf{X})_i(t)$ are independent (assuming, as it is customary, that the data have zero mean). The main point is that the unknown ψ_{it} may be replaced by some guess, as these equations with are still satisfied as soon as the $(\mathbf{B}\mathbf{X})_i$ are independent, while the equations

(3) may be dropped. This is the approach of Pham and Garat [6]. But in their case $\psi_{it} \equiv \psi_i$ and the form of ψ_i is relatively simple while we are concerned here with the case where ψ_{it} varies with t and we cannot realistically know how it varies, except that it should vary *slowly*. Thus we need at least a crude estimate of ψ_{it}. To this end, we adopt the technique of projection in Pham and Garat [6], together with the idea of blocking and kernel smoothing. We approximate the score function $\psi_{(\mathbf{BX})_i(t)}$ of $(\mathbf{BX})_i(t)$ by projecting it onto a given linear space, spanned by a basis $\{\phi_1, \ldots, \phi_N\}$, says. Thus we minimize $\mathrm{E}\{\psi_{(\mathbf{BX})_i(t)} - \mathbf{c}_i^{\mathrm{T}}\phi[(\mathbf{BX})_i(t)]\}^2$ with respect to $\mathbf{c}_i$, where $\phi = [\phi_1 \ \cdots \ \phi_N]^{\mathrm{T}}$. By expanding the above square and using integration by parts, one is led to minimize $\mathrm{E}\{\mathbf{c}_i^{\mathrm{T}}\phi[(\mathbf{BX})_i(t)]\}^2 - 2\mathbf{c}_i^{\mathrm{T}}\mathrm{E}\phi'[(\mathbf{BX})_i(t)]$, $'$ denoting the derivative. The key point is that this expression no longer involves the unknown $\psi_{(\mathbf{BX})_i(t)}$ but only the expectation operator, which can be replaced by a time average in the stationary case. To handle non stationarity, we resort to two ideas introduced in [5].

1. Blocking: We subdivide the time indexes $\{1, \ldots, T\}$ into L consecutive blocks $T_1, \ldots T_L$, such that the distribution of $\mathbf{X}(t)$ may be regarded as the same for all t in each block. Then for any function φ, $\mathrm{E}\varphi[\mathbf{X}_i(t)]$ is estimated by
$$\hat{\mathrm{E}}\varphi[\mathbf{X}(t)] = \frac{1}{\#T_l}\sum_{t\in T_l}\varphi[\mathbf{X}(t)] \quad \text{for } t \in T_l, \qquad (4)$$
where $\#T_l$ denotes the number elements in T_l.
2. Kernel smoothing: We estimate $\mathrm{E}\varphi[\mathbf{X}_i(t)]$ by
$$\hat{\mathrm{E}}\varphi[\mathbf{X}(t)] = \frac{\sum_{\tau=1}^{T} k(\frac{\tau-t}{M})\varphi[\mathbf{X}(\tau)]}{\sum_{\tau=1}^{T} k(\frac{\tau-t}{M})} \qquad (5)$$
where k is a kernel function (with compact support in general) and M is a window width parameter.

The kernel method is very similar to the blocking method, but it is costlier and should be better, as the estimate is a local average *around* the time point of interest while the blocking method averages over blocks.

Using one of the above estimators, $\psi_{(\mathbf{BX})_i(t)}$ is estimated by $\hat{\psi}_{(\mathbf{BX})_i(t)} = \hat{\mathbf{c}}_i^{\mathrm{T}}\phi$ where $\hat{\mathbf{c}}_i$ is determined by

$$\hat{\mathrm{E}}\{\phi[(\mathbf{BX})_i(t)]\phi^T[(\mathbf{BX})_i(t)]\}\hat{\mathbf{c}}_i = \hat{\mathrm{E}}\{\phi'[(\mathbf{BX})_i(t)]\}. \quad (6)$$

The separating matrix $\mathbf{B}$, in turn, is determined by the estimating equations (2) with ψ_{it} replaced by $\hat{\psi}_{(\mathbf{BX})_i(t)}$.

We will take ϕ_1 to be the identity function to ensure the optimality of the projection method [6]. Hence

$$\hat{\mathrm{E}}\{\hat{\psi}_{(\mathbf{BX})_i(t)}[(\mathbf{BX})_i(t)](\mathbf{BX})_i(t)\} = 1 \qquad (7)$$

which proves to be useful later. As for $\phi_2, \ldots, \phi_N$ we will take them to be power functions, of the form $\phi_i(x) = \mathrm{sign}(x)|x|^{\alpha_i}$, $\alpha_i > 0$. This choice ensures the invariance of the estimating equations with respect to scale change, in the sense that they are still satisfied when a solution $\mathbf{B}$ is pre-multiplied by a diagonal matrix.

2.2 The Contrast Approach

A drawback of the use of estimating equations is that they generally admit several distinct solutions (not counting those differing only by permutation and scaling. Thus there is a risk of obtaining a spurious solution (although by using the above projection method and some clever algorithm, this might be avoided, see [3]). Therefore it is of interest to consider an alternative method which consists in minimizing a contrast function (see [1, 4] for a definition). A popular contrast is based on the mutual information, which is closely related to the expected log likelihood [5] and hence can be expected to yield efficient separation procedures. However, it is not practical in this context since it requires estimating the density of $(\mathbf{BX})_i(t)$ for each i and t. In [5], this difficulty is avoided by working with the Gaussian mutual information instead, which amounts to ignoring the possible non Gaussianity of the sources. Our approach here is to abandon the mutual information criterion and to consider a simpler and easier to implement contrast.

Pham [4] has laid down a general framework for constructing contrast functions. Let Q be a functional over distributions of random variables, which is superadditive of class II in the sense of [2], that is $Q(X + c) = Q(X)$ and $Q(cX) = |c|Q(X)$ for any random variable X and

$$Q^2(X + Y) \geq Q^2(X) + Q^2(Y) \qquad (8)$$

for any pair of independent random variables X, Y. Then it has been shown in [4] that in the stationary case the criterion $\sum_{i=1}^{K} \log Q[(\mathbf{BX})_i(t)] + \log\det\mathbf{B}$ is a contrast function, which is discriminating (in the sense of [1]) if the inequality (8) is strict unless X and Y are both Gaussian. Note that the time t index is irrelevant in this case but it does matter in the non stationarity case where it is easy to extend the above result to show that it still holds provided that $\log Q[(\mathbf{BX})_i(t)]$ is replaced by $T^{-1}\sum_{t=1}^{T}\log Q[(\mathbf{BX})_i(t)]$.

Direct calculations show that the functional $Q(X) = (\mathrm{E}X^4)^{1/4}$ is superadditive of class II over the set of *zero mean sub Gaussian* random variables. (A zero mean random variable X is said to be sub Gaussian if its fourth cumulant $\mathrm{E}X^4 - 3(\mathrm{E}X^2)^2$ is non positive). Therefore, *if the sources are sub Gaussian*, a very simple contrast function is

$$\frac{1}{4T}\sum_{i=1}^{K}\sum_{t=1}^{T}\log\mathrm{E}[(\mathbf{BX})_i(t)] + \log\det\mathbf{B} \qquad (9)$$

To implement this contrast, one only needs to replace $\mathrm{E}[(\mathbf{BX})_i(t)]$ by its estimate. As before one can either use the blocking method or the kernel smoothing method for this purpose. Taking the gradient the resulting criterion, one gets the system of estimating equations

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\mathrm{E}}[(\mathbf{BX})_i^3(t)(\mathbf{BX})_j(t)]}{\hat{\mathrm{E}}[(\mathbf{BX})_i^4(t)]} = 0, \quad 1 \leq i \neq j \leq K$$

It can be seen that, *if the blocking method is used*, the above system is the same as in the quasi maximum likelihood approach, with $\hat{\psi}_{(\mathbf{BX})_i(t)}$ now being $1/\hat{\mathrm{E}}[(\mathbf{BX})_i^4(t)]$ times the cubic function. If the kernel method is used, it is not exactly the same but is still very close. Thus we end up with the same system of estimating equation as before, but the functions $\hat{\psi}_{(\mathbf{BX})_i(t)}$ are now much simpler and less costly to compute. The downside is that this method is designed for sub Gaussian sources only.

3 ALGORITHMS

3.1 Off-Line Algorithm

The goal is to solve for (2) with $\psi_{it} = \hat{\psi}_{(\mathbf{BX})_i(t)} = \hat{\mathbf{c}}_i^{\mathrm{T}}\phi$ as defined by (6). To this end we shall make use of the quasi Newton algorithm. Starting at a current value $\mathbf{B}$, one makes a small change $-\mathbf{HB}$ to it such that the estimating equations is satisfied up to the first order. Direct calculation shows that the corresponding change of the right hand side of (2) is, up to the first order,

$$-\frac{1}{T}\sum_{t=1}^{T}\{\psi'_{it}[(\mathbf{BX})_i(t)](\mathbf{HBX})_i(t)(\mathbf{BX})_j(t) + \psi_{it}[(\mathbf{BX})_i(t)](\mathbf{HBX})_j(t)\}$$

In the quasi-Newton algorithm, the above expression is further approximated, by regarding the $(\mathbf{BX})_i(t)$ for different i as independent (which is justified if we are close to the separating solution) and replacing the time average by expectation and vice-versa. Note that we haven't take into account of the fact that the ψ_{it} are set to $\hat{\psi}_{(\mathbf{BX})_i}$ which would change when $\mathbf{B}$ changes to $\mathbf{B} - \mathbf{HB}$. But using the above kind of approximation and the equations (6) one can show that the effect of this change can be neglected. Finally, the elements h_{ij} of $\mathbf{H}$ are determined by the equations

$$\begin{bmatrix} \omega_{ij} & 1 \\ 1 & \omega_{ji} \end{bmatrix} \begin{bmatrix} h_{ij} \\ h_{ij} \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T} \begin{bmatrix} \hat{\psi}_{(\mathbf{BX})_i(t)}[(\mathbf{BX})_i(t)](\mathbf{BX})_j(t) \\ \hat{\psi}_{(\mathbf{BX})_j(t)}[(\mathbf{BX})_j(t)](\mathbf{BX})_i(t) \end{bmatrix} \tag{10}$$

where

$$\omega_{ij} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathrm{E}}[\hat{\psi}'_{(\mathbf{BX})_i(t)}[(\mathbf{BX})_i(t)]\hat{\mathrm{E}}[(\mathbf{BX})_j^2(t)]. \tag{11}$$

Note that these equations do not determine h_{ii}. They can in fact be arbitrary as long as they are small, since they essentially induce scale changes. Thus we take $\mathbf{B} - \mathbf{HB}$ as new value of $\mathbf{B}$, where the $\mathbf{H}$ is the matrix with zero diagonal and off diagonal elements h_{ij}.

In the contrast approach, similar arguments lead to the same equation (10) but now (11) simplifies to

$$\omega_{ij} = \frac{1}{T}\sum_{t=1}^{T} \frac{3\hat{\mathrm{E}}[(\mathbf{BX})_i^2(t)]\hat{\mathrm{E}}[(\mathbf{BX})_j^2(t)]}{\hat{\mathrm{E}}[(\mathbf{BX})_i^4(t)]} \tag{12}$$

It is important that the matrix in equation (10) be positive definite. Indeed, the right hand side of this equation can be viewed as an approximate gradient of the negative of the log likelihood (in the quasi-maximum likelihood approach) or of the estimated contrast (in the contrast approach). Thus the positive definiteness of this matrix means that the algorithm moves the point $\mathbf{B}$ in a direction which decrease the criterion. Looking at (10), this is equivalent to $\omega_{ij}\omega_{ji} > 1$. But from (6), $\hat{\mathrm{E}}\hat{\psi}'_{(\mathbf{BX})_i(t)}[(\mathbf{BX})_i(t)] = \hat{\mathrm{E}}\hat{\psi}^2_{(\mathbf{BX})_i(t)}[(\mathbf{BX})_i(t)]$ which, by (7) and the Schwartz inequality, is bounded below by $1/\hat{\mathrm{E}}[(\mathbf{BX})_i^2(t)]$. Thus by the same arguments as in [5], $\omega_{ij}\omega_{ji} > 1$ unless $\hat{\mathrm{E}}[(\mathbf{BX})_j^2(t)]/\hat{\mathrm{E}}[(\mathbf{BX})_i^2(t)]$ is constant in t and $\hat{\mathrm{E}}\hat{\psi}^2_{(\mathbf{BX})_i(t)}[(\mathbf{BX})_i(t)] = 1/\hat{\mathrm{E}}[(\mathbf{BX})_i^2(t)$ for all t. In the contrast approach, since the random variable $(\mathbf{BX})_i(t)$ is sub Gaussian, one may expect that $3\{\hat{\mathrm{E}}[(\mathbf{BX})_i^2(t)]\}^2 \geq \hat{\mathrm{E}}[(\mathbf{BX})_i^4(t)$ (although this may not always be true) which would again result in $\omega_{ij}\omega_{ij} > 1$.

3.2 On-line algorithm

The on-line analogue to the above algorithm is based on an updating rule for the functions $\hat{\psi}_{(\mathbf{BX})_i(t)}$. To simplify the notation, we denote by $\hat{S}_i(t)$ the i-th components of $\mathbf{B}(t)\mathbf{X}(t)$, $\mathbf{B}(t)$ being the estimated separating matrix at time t. In the quasi maximum likelihood approach, we introduce the matrix $\mathbf{G}_i(t)$ and the vector $\mathbf{g}_i(t)$ which are updated as

$$\mathbf{G}_i(t) = (1-\rho)\mathbf{G}_i(t-1) + \rho\phi[\hat{S}_i(t)]\phi^T[\hat{S}_i(t)]$$
$$\mathbf{g}_i(t) = (1-\rho)\mathbf{g}_i(t-1) + \rho\phi'[\hat{S}_i(t)]$$

where ρ is a small learning step. We then compute $\hat{\mathbf{c}}_i(t) = \mathbf{G}_i^{-1}(t)\mathbf{g}_i(t)$ and take $\hat{\psi}_{\hat{S}_i(t)}$ to be $\hat{\mathbf{c}}_i^{\mathrm{T}}(t)\phi$ and obtain $\hat{\mathrm{E}}\hat{\psi}'_{\hat{S}_i(t)}[\hat{S}_i(t)]$ on-line as $\hat{\mathbf{c}}_i^{\mathrm{T}}(t)\mathbf{g}_i(t)$. In the contrast approach, one simply takes $\hat{\psi}_{\hat{S}_i(t)}$ to be $1/\hat{\mathrm{E}}\hat{S}_i^4(t)$ times the cubic function, $\hat{\mathrm{E}}\hat{S}_i^4(t)$ being updated similarly to the $\mathbf{G}_i(t)$ or $\mathbf{g}_i(t)$:

$$\hat{\mathrm{E}}\hat{S}_i^4(t) = (1-\rho)\hat{\mathrm{E}}\hat{S}_i^4(t-1) + \rho\hat{S}_i^4(t).$$

One then takes $\hat{\mathrm{E}}\hat{\psi}'_{\hat{S}_i(t)}[\hat{S}_i(t)]$ to be $3\hat{\mathrm{E}}\hat{S}_i^2(t)/\hat{\mathrm{E}}\hat{S}_i^4(t)$.

To compute the new value $\mathbf{B}(t+1)$, we first update the ω_{ij}, now depending on t, as

$$\omega_{ij}(t) = (1-\lambda)\omega_{ij}(t-1) + \lambda\hat{\mathrm{E}}\hat{\psi}'_{\hat{S}_i(t)}[\hat{S}_i(t)]\hat{\mathrm{E}}\hat{S}_j^2(t)$$

where λ is another learning step. It is important that λ be *much less* than ρ so that $\omega_{ij}(t)$ appears as a local average of $\hat{\mathrm{E}}\hat{\psi}'_{\hat{S}_i(t)}[\hat{S}_i(t)]\hat{\mathrm{E}}\hat{S}_i^2(t)$ over a time period much longer than that used to calculate $\hat{\mathrm{E}}\hat{\psi}'_{\hat{S}_i(t)}[\hat{S}_i(t)]$ and $\hat{\mathrm{E}}\hat{S}_i^2(t)$. As before, in the quasi maximum likelihood approach, one has $\hat{\omega}_{ij}(t)\hat{\omega}_{ji}(t) > 1$. This is because, from the above equation for $\hat{\mathbf{c}}_i(t)$ and the fact that $\phi_1(s) = s$, $\hat{\mathrm{E}}\hat{\psi}'_{\hat{S}_i(t)}[\hat{S}_i(t)]$ is bounded below by $1/G_{i,11}(t)$ where $G_{i,11}(t)$ denotes the upper left element of $\mathbf{G}_i(t)$, which is no other than $\hat{\mathrm{E}}\hat{S}_i^2(t)$. This bound also generally holds in the contrast approach provided that the sources are sub Gaussian.

Finally, $\mathbf{B}(t+1) = \mathbf{B}(t) - \lambda \mathbf{H}(t)\mathbf{B}(t)$ where the matrix $\mathbf{H}(t)$ has off diagonal elements solution of the equations

$$\begin{bmatrix} \omega_{ij}(t) & 1 \\ 1 & \omega_{ji}(t) \end{bmatrix} \begin{bmatrix} h_{ij}(t) \\ h_{ij}(t) \end{bmatrix} = \begin{bmatrix} \hat{\psi}_{\hat{S}_i(t)}[\hat{S}_i(t)]\hat{S}_j(t) \\ \hat{\psi}_{\hat{S}_j(t)}[\hat{S}_j(t)]\hat{S}_i(t) \end{bmatrix} \tag{13}$$

The diagonal element of $\mathbf{H}$ may be set to zero. But this could lead to a slow continuous drift in the scale of the reconstructed sources. To avoid this, one may take, as in [5], $h_{ii}(t) = \alpha[\hat{S}_i^2(t) - 1]$, α being a small number, in order to drive gently to 1 the long term average variance of each reconstructed source.

3.3 Other on-line algorithms

As in [5], one may consider the stochastic gradient algorithm. This amounts to setting $[h_{ij}(t) \quad h_{ji}(t)]^{\mathrm{T}}$ to the right hand side of (13) and thus the $\omega_{ij}(t)$ are no longer needed. But this algorithm is much slower than the Newton like algorithm, while the gain in computational cost is small. A more important reduction of this cost is to update directly $\hat{\mathbf{c}}_i(t)$ by a stochastic gradient technique as in [3]. Explicitly

$$\hat{\mathbf{c}}_i(t) = \hat{\mathbf{c}}_i(t-1) + \rho \mathbf{\Gamma}\{\phi'[\hat{S}_i(t)] - \phi[\hat{S}_i(t)]\hat{\psi}_{\hat{S}_i(t-1)}[\hat{S}_i(t)]\}$$

where $\mathbf{\Gamma}$ is a given matrix designed to accelerate the convergence. The difficulty is to find a good matrix $\mathbf{\Gamma}$, otherwise the convergence can be too slow.

4 Simulation

In a simulation study, we have generated 3 independent sources by multiplying independent uniform random variables in $[-\sqrt{3}, \sqrt{3}]$ with the square root their variance profiles (shown in figure 1). They are then mixed to produce the observations.

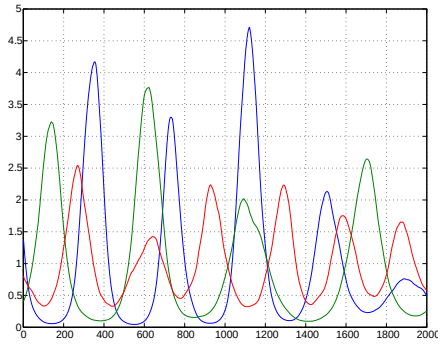

Figure 1: Variance profiles of the sources

We first apply our off-line algorithm with three basis power functions of exponent .5, 1 and 3, to the first 500 observations (which was divided into 3 blocs by the algorithm). The evolution of the global matrix $\mathbf{BA}$ (9 elements) with the iteration step is shown in figure 2. One can see that the algorithm converges in less than 15 steps and the final separating matrix is quite good.

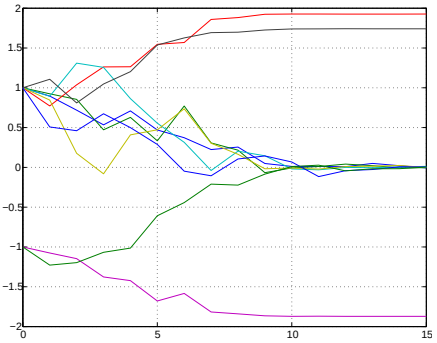

Figure 2: Convergence of the off-line algorithm

Finally, figure 3 shows the results of the application of the adaptative algorithm with the same basis function. The learning step are $\rho = .03$ and $\lambda = .01$. Figure 3 shows the evolution of the 9 elements of the global matrix $\mathbf{BA}$ in time. One can see that the algorithm performs quite well.

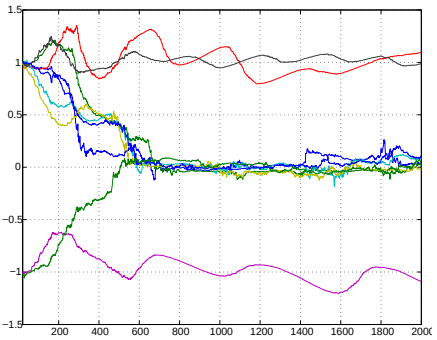

Figure 3: Convergence of the on-line algorithm

References

[1] P. Comon, Independent components analysis, a new concept ? *Signal Processing*, 36(3): 287–314, 1994.

[2] P.-J. Huber Projection pursuit. *Ann. Statist.* 13(2): 435–475, 1985.

[3] D.-T. Pham. Séparation aveugle de mélange instantanée de sources à l'aide de fonctions séparatrices ajustées. In *Proc. GRETSI, Grenoble, France*, pages 969–972, October, 1997

[4] D.-T. Pham. Contrast for blind separation and deconvolution of sources. In *Proc. ICA 2001*, San Diego, USA, December 2001

[5] D.-T. Pham and J.-F. Cardoso. Blind Separation of Instantaneous Mixtures of Non Stationary Sources. *IEEE Tr. SP*, 49(9): 1837–1848, September 2001.

[6] D.-T. Pham and P. Garat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Tr. SP*, 45(7): 1712–1725, July 1997.