Competitive Algorithm Blending for Enhanced Source Separation

Keith D. Gilbert and Karen L. Payton
Electrical and Computer Engineering Department
University of Massachusetts Dartmouth
285 Old Westport Road
North Dartmouth, MA 02747

Abstract – This paper proposes to enhance the blind source separation (BSS) solution by running multiple BSS algorithms in parallel and blending the outputs to produce a set of source estimates that is at least as good as any individual method, and potentially better. Although the method is applicable to more general BSS problems, the proposed blending method is described in the case of instantaneous mixtures of stationary, zero-mean, unit-variance, white sources. Experimental results show that the method is able to select a best set of sources with respect to minimum mutual information from an input consisting of source estimates.

I. INTRODUCTION

The blind source separation (BSS) problem occurs when we wish to recover some number of unknown signals (sources) from a set of observations that are comprised of mixtures of the unknown sources. The problem arises in a wide variety of fields including imaging, biomedical data processing, telecommunication systems, electromagnetic transmission systems, and audio signal processing, e.g. see [1 chs. 15-19].

Although the work in this paper is applicable to other types of BSS solutions, we will develop our work under the model that the BSS outputs are estimates of sources that are observed through linear, time-invariant, instantaneous mixtures, i.e.

\[ \mathbf{x}(t) = \mathbf{G}\mathbf{s}(t), \]  

(1)

where \( \mathbf{x}(t) = [x_1(t), \ldots, x_p(t)]^T \) is the vector of observations at time \( t \), \( \mathbf{G} \) is a \([p \times q]\) matrix of mixing coefficients, and \( \mathbf{s}(t) = [s_1(t), \ldots, s_q(t)]^T \) is the vector of source samples time \( t \).

At this point, source separation is a well-studied problem, and although numerous strategies exist, e.g. maximum likelihood, information-theoretic, non-negative matrix factorization, sparse component analysis, and Bayesian methodologies, BSS methods usually make some assumptions about the sources and/or mixing systems in order to gain traction on the problem, and at present there is no “one size fits all” approach. [1]-[8]

The work here aims to increase the blindness of BSS by allowing multiple BSS methods to run in parallel and compete to produce a “best” set of source estimates. In essence, we want to exploit the diversity of multiple competing BSS methods in order to possibly produce a set of source estimates that is better than any single method could provide.

Although the work here was inspired by the competitive algorithm/model framework that has been studied in the universal prediction and adaptive filtering communities, e.g. [9]-[15], the concept of competition is not new to the BSS problem. In [3], Pham derived an optimal method to blend score functions, in [7] cumulant blending was employed, and in [6] Amari used a convex combination of score-dual functions. As a generalization, Comon has shown in [5] that for any set of valid source separation contrast functions, any convex-constrained linear combination of the set of functions is also a valid contrast function. All of the aforementioned methods can be seen as cost function blending, while our approach is a blending of the outputs.

Although the method we present here is applicable to more general situations, we will make several simplifying assumptions. As stated earlier, the first assumption is that the source estimates are produced from a set of linear instantaneous mixtures of the form (1). The second assumption is that all the sources and source estimates are stationary, zero-mean, white processes. The third assumption is that the sources we wish to recover, or select, are mutually independent, so that the joint probability density function (pdf) of the sources decomposes to the product of the marginal pdfs of the individual sources. Another assumption that we make is that the BSS methods are operating in steady-state.

Under the assumption of mutual independence, the mutual information between each source is zero, where the mutual information between two random variables is defined as,
\[ I(u;v) = H(u) + H(v) - H(u,v) \]
where \( H(u) \) and \( H(v) \) are the respective marginal entropies of the individual variables \( u \) and \( v \), and \( H(u,v) \) is the joint entropy of \( u \) and \( v \). The mutual information is zero when the variables are independent and has a maximum equal to the minimum of \( H(u) \) and \( H(v) \). In this work we will use a normalized version of mutual information that is confined to the interval \([0,1]\) and is given as,

\[ K[u,v] = \left( H_G \right)^{-1} I\left( \sigma_u^{-1} u; \sigma_v^{-1} v \right), \]

where \( H_G \) is the entropy of the unit-variance Gaussian pdf, and \( \sigma_u \) and \( \sigma_v \) are the standard deviations of the zero-mean random variables, \( u \) and \( v \), respectively. By normalizing the variables to unit variance and noting that \( H(\sigma_u^{-1} u) \leq H_G \) and \( H(\sigma_v^{-1} v) \leq H_G \), we see that

\[ 0 \leq K[u,v] \leq H_G \min\{H(\sigma_u^{-1} u), H(\sigma_v^{-1} v)\} \leq 1. \]

Considering the pair of random vectors \( u = [u_1, ..., u_N]^T \) and \( v = [v_1, ..., v_N]^T \), we further define the extension to vectors as the matrix of the pairwise vector elements’ mutual information given by,

\[
K[u,v] = \begin{bmatrix}
K[u_1,v_1] & \cdots & K[u_1,v_M] \\
\vdots & \ddots & \vdots \\
K[u_N,v_1] & \cdots & K[u_N,v_M]
\end{bmatrix},
\]

and introduce the special shorthand notation \( K[u] = K[u,u] \).

[16][17]

II. COMPETITIVE ALGORITHM BLENDING FOR BSS

The focus of this paper is the blending of the outputs produced by multiple BSS algorithms running in parallel in order to produce a set of source estimates that is at least as good as the best set produced across all algorithms in the minimum mutual information sense, and we emphasize that the best set of sources may not be produced by a single method. Although we model the inputs as BSS algorithm outputs, we shall see that the method we propose here is more general, in that we do not need to know anything about the BSS methods themselves, nor do we need access to the mixtures.

If we consider the estimates from the \( m \)th algorithm at time \( t \) to be the \([Q_m,1]\) vector of samples \( \tilde{y}_{m,t}(t) = [\tilde{y}_{m,1}(t), ..., \tilde{y}_{m,Q_m}(t)]^T \), then the input to our blending system is the \([N,1]\) vector \( y(t) = [y_1(t), ..., y_N(t)]^T \), where \( N = \sum_{m=1}^{M} Q_m \) and \( y_m(t) \) is the vector of normalized zero-mean, unit-variance samples whose \( i \)th entry is \( y_{m,i}(t) = \sigma_m^{-1} (\tilde{y}_{m,i}(t) - \mu_m) \), for \( m = 1, ..., M \) and \( i = 1, ..., Q_m \).

![Fig. 2 - Competitive algorithm blending for BSS.](Image)

\( N \) normalized BSS outputs, \( y \), are blended together via \( B_q \) to produce \( q \) source estimates in \( \tilde{s}_q \) for \( q = 2, ..., N \). \( B_q \) is adapted via a performance-based loss matrix, \( R_q \), for \( q = 2, ..., N \). A set of performance-based weights, \( z \), select a best set of sources, \( \hat{s}_Q \) with \( Q \in \{2, ..., N\} \).

Assuming that we wish to estimate \( q \) number of sources, then our source estimates are given by,

\[
\hat{s}_q(t) = B_q^T(t)y(t),
\]

where \( \hat{s}_q(t) = [\hat{s}_{q,1}(t), ..., \hat{s}_{q,q}(t)]^T \) is the \([q,1]\) vector of source estimates and \( B_q(t) = [b_{q,1}(t), ..., b_{q,q}(t)]^T \) is an \([N,q]\) matrix of blending weights whose \( i \)th column is the \([N,1]\) vector \( b_{q,i}(t) \).

The goal is to minimize the mutual information between the source estimates which is to minimize the off-diagonal terms of \( K[\hat{s}_q(t)] \), or equivalently minimize the cost function,

\[
J(\hat{s}_q(t)) = \frac{1}{2q(q-1)} E \left\{ \sum_{i=1}^{q} K^o \left( \hat{s}_i(t) \right) \hat{s}_i(t) \right\},
\]

where \( E\{\} \) is the expectation operator, \( I_q \) is a \([q,1]\) vector of ones, and \( K^o(u) = \text{OffDiag}(K[u]) \) is the matrix formed by setting the diagonal elements of \( K[u] \) to zero. For convenience, we will abbreviate \( J_q(t) = J(\hat{s}_q(t)) \) and \( K^o_q(t) = K^o(\hat{s}_q(t)) \).

In order to determine the vector of blending weights for the \( i \)th source estimate, i.e. \( b_{q,i}(t) \), we look at the \( i \)th element of the product \( K^o_q(t)I_q \), i.e.,

\[
\left[ K^o_q(t)I_q \right]_{i} = \sum_{j=1}^{q} K \left\{ b_{q,j}(t)y(t), \hat{s}_{q,j}(t) \right\},
\]

and consider the performance-based vector loss function,

\[
r_{q,i}(y(t), \hat{s}_{q,i}(t)) = \frac{1}{q-1} \sum_{j=1, j \neq i}^{q} K \left\{ y(t), \hat{s}_{q,j}(t) \right\}.
\]

Clearly, \( r_{q,i} \) assigns a lower loss to the elements of \( y(t) \) that have the least mutual information with the other source estimates, \( \hat{s}_{q,1}(t), ..., \hat{s}_{q,i-1}(t), \hat{s}_{q,i+1}(t), ..., \hat{s}_{q,q}(t) \), while assigning a
higher loss to the elements of $y(t)$ that share the most information with the other source estimates.

Although our goal is to minimize (6), in this paper we are focused on selecting the best set of source estimates from $y(t)$ and are not concerned with trying to improve upon the performance of the best set, which is a topic for future work. To this end, we will impose a convex constraint on our blending vectors such that $[b_{q,i}(t)]_{j} \geq 0$ and $b_{q,i}^{T}(t)1_q = 1$ for $q = 2, ..., N$, $i = 1, ..., q$, and $j = 1, ..., q$. Defining the accumulated loss up to time $n$ as,

$$r^{(n)}_{q} = \sum_{t=1}^{n} r_{q,i}(y(t), \hat{s}_{q}(t)),$$

and the element-wise exponential function of an $[P \times 1]$ vector as $exp\{u\} = [e^{u1}, e^{u2}, ..., e^{uP}]^{T}$, we construct the blending vector for the $i^{th}$ source estimate according to,

$$b_{q,i} = \frac{D_{q,i} \exp\{-[c_{q,i}(t)]^{-1}r_{q,i}^{(t)}\}}{\sum_{t=1}^{n} D_{q,i} \exp\{-[c_{q,i}(t)]^{-1}r_{q,i}^{(t)}\}},$$

where $c_{q,i}(t)$ is a scalar and $D_{q,i}$ is a diagonal $[N \times N]$ matrix with $a priori$ weights for the elements of $y(t)$ along the diagonal. Under the assumption that there is a minimum value in $r_{q,i}(y(t), \hat{s}_{q}(t))$, and defining the next largest value as $p_{q,i}(t)$, then

$$c_{q,i}(t) = (1 - a)p_{q,i}(t),$$

where $a$ is a small positive constant. For now, we include this value without justification and will explain the choice of $c_{q,i}(t)$ later. We include the general form of $D_{q,i}$ for completeness, but for the purpose of this presentation we only consider the uniform weighting given by $D_{q,i} = N^{-1}I_N$, where $I_N$ is the $[N \times N]$ identity matrix.

Thus far, we have described how to produce the set of $q$ source estimates that minimizes (6), so the final step in our method is to find the set over $q = 2, ..., N$ that produces the best source estimates. We would like to construct a set of weights, $z(t)$, similar to the $q$-set methods, and select the method that has the highest weight according to an $([N - 1] \times 1)$ vector loss function, $\hat{z}(t)$. Here we will use the instantaneous loss of each $q$-set method defined for an $[N \times 1]$ vector as,

$$l(u) = \frac{1}{2N(N - 1)} K^C(u) 1_N.$$

Letting $[\hat{z}(t)]_{q,i} = l(\hat{s}_{q}(t))$, we construct the weights as,

$$z(t) = \frac{D_{q,i} \exp\{-[c_{q,i}(t)]^{-1}\hat{z}_{q,i}(t)\}}{\sum_{t=1}^{n} D_{q,i} \exp\{-[c_{q,i}(t)]^{-1}\hat{z}_{q,i}(t)\}},$$

where, similar to the individual $q$-set case, $c_{q,i}(t)$ is a scalar and $D_{q,i}$ is a diagonal matrix with $a priori$ weights for the elements of $z(t)$ along the diagonal, i.e. the $q$-set loss priors. The scalar $c_{q,i}(t)$ is determined the same way as in the $q$-set case, i.e. $c_{q,i} = (1 - a)p_{q,i}(t)$ where $p_{q,i}(t)$ is the second smallest value in $\hat{z}(t)$. If we let $d_{q,i}$ denote the diagonal values of $D_{q,i}$, we find that

$$[d_{q,i}] = \log(q)\sum_{q=2}^{Q} \log(q^{q})^{-1}$$

for $q = 2, ..., N$ produces suitable results. We need to give higher order $q$-set methods a higher prior due to a source segregation phenomenon (discussed below) of $q$-set methods when $q < Q$, where $Q$ is the actual number of sources. Thus, the set of source estimates at time $t$ is $\hat{s}_{q(t)}(t)$ where,

$$\hat{Q}(t) = 1 + \arg\max_{i=1, ..., N} z_{q,i}(t).$$

In order to increase the convergence rate, we use a modified version of the system just described. In the spirit of the universal method in [9], we create our $q$-set source estimates according to $\hat{s}_{q}(t) = B_{q}(t)y(t)$, where $B_{q}(t)$ is a matrix with columns $\hat{b}_{q}(t)$. If any weight in $\hat{b}_{q}(t)$ is greater than some threshold, $T (> 0.5)$, then $\hat{b}_{q}(t)$ is a vector of all zeros except for a single 1 at the position where $\hat{b}_{q}(t)$ exceeded the threshold. If no weights in $\hat{b}_{q}(t)$ exceed the threshold, then $\hat{b}_{q}(t) = b_{q}(t)$. Note that $B_{q}(t)$ remains unmodified, but the source estimates, $\hat{s}_{q}(t)$, will be affected.

III. CHOICE OF $c_{q,i}(t)$ AND THE GAUSSIAN APPROXIMATION

Before we begin a justification of the choice of $c_{q,i}(t)$ in (11), we need to point out that there are two “convergences” associated with the system we have just outlined. The first convergence is associated with source segregation, and the second is associated with producing a best set of sources. For a $q$-set estimator, segregation is characterized by orthogonality of the columns of $B_{q}$, so that at “segregation convergence,” the off-diagonal elements of $B_{q}B_{q}$ are zero. The “best set” convergence is characterized by the columns of $B_{q}$ being all zeros except for a single one. In other words, the system presented here first splits the signals into groups that minimize (6), before proceeding to select the individual signals that further minimize (6). The performance analysis of the blending system is an ongoing research topic which consists of determining under what conditions source estimates will be segregated, and once segregated, how source estimates will converge to the best set. For the argument
below we will make two assumptions; the blending system has segregated the sources, and the blending system has converged to all members of the best set except for the \(i^{th}\) element.

In order to choose the value of \(c_{qj}(t)\), we can make an argument that is almost identical to that in [14], but for the sake of brevity, we will only briefly describe the major points leading to our choice of \(c_{qj}(t)\), and the reader is invited to explore the full argument in [14] (Theorem 1 proof). To begin, we denote \(s^*(t)=[s^*_1(t),...,s^*_q(t)]\) as the \(Q\)-set in \(y(t)\) that minimizes the model-order compensated \(\mathcal{F}(\cdot)\), i.e. \(\mathbf{D}_s\) in (13), and for the purposes of this paper we will assume that \(s^*(t)\) is unique (beyond permutation), i.e. there is only one \(Q\)-set of sources that minimizes (6). Next, we construct “probabilities” for the source estimates in \(y(t)\) for the \(i^{th}\) source estimate in the \(q=Q\) estimator operating at segregation convergence as,

\[
P_{qj}(t) = \alpha \exp\left[-c_{qj}(t)\right]_{Q,1}^*, j = 1,...,N
\]  

(15)

where \(\alpha\) is a scalar constant and \(r_{qj}^* = [r_{qj1}^*,...,r_{qjn}^*]\), and

\[
r_{qj}^* = \sum_{t=1}^T r_{qj}(y(n),s^*(n))
\]

With these probabilities we construct a universal probability measure via a uniformly weighted Bayes mixture as,

\[
P_{qj}(t) = \frac{1}{N} \sum_{j=1}^N P_{qj}(t)
\]

(16)

which, by definition,

\[
P_{qj}(t) \geq \max_j \{N^{-1} P_{qj}(t)\}
\]

(17)

Next we define the “probability” of our \(q=Q\) estimator as,

\[
\hat{P}_{qj}(t) = \alpha \exp\left[-c_{qj}(t)\right]_{Q,1}^* r_{qj}^*\]

(18)

The cumulative nature of our loss function allows us to write the Bayes mixture as,

\[
P_{qj}(t) = \prod_{n=1}^T b_{qj}^*(n) \exp\left[-c_{qj}(n)\right] r_{qj}(y(n),s^*(n))
\]

(19)

and our \(q=Q\) estimator as,

\[
\hat{P}_{qj}(t) = \prod_{n=1}^T \exp\left[-c_{qj}(n)\right] b_{qj}^*(n) r_{qj}(y(n),s^*(n))
\]

(20)

Note that the multiplicands in (19) are a convex combination of a function of \(r_{qj}\) and the multiplicands in (20) are the same function evaluated at a convex combination of \(r_{qj}\).

Thus, if the function \(\exp\left[-c_{qj}(n)\right] r_{qj}(\cdot)\) is concave, then,

\[
\hat{P}_{qj}(t) \geq P_{qj}(t) \geq \max_j \{N^{-1} P_{qj}(t)\}
\]

(21)

where the first inequality is a consequence of Jensen’s inequality and the second is from (17). Evaluating the negative logarithm of (21) gives a relationship between the performance of our blending method and the performance of the best estimates in \(y(t)\).

We will refrain from evaluating (21) at this point, since a major question remains; is the exponential of the scaled mutual information concave? If we define the joint distribution between two random variables, \(u\) and \(v\), as \(h(u,v)=f(u)g(v|u)\), i.e. the product of the marginal of \(u\) and the pdf of \(v\) conditioned on \(u\), then mutual information is concave in \(f\) and convex in \(g\). Similarly, if we look at an alternate form of mutual information, \(I(u;v) = H(u) - H(u|v)\), we see that minimization of \(I\) is a maximization of \(H(u|v)\). Intuitively, this tells us that movement towards minimal \(I\) is orthogonal to the concavity of \(I\), i.e. (generally) in the direction in which \(I\) is convex, and thus in the direction in which the exponentiated \(I\) is concave. A detailed analysis for the general case is required and might involve testing positive-definiteness of the mutual information Hessian in the direction of \(I = 0\).

Although we leave question of concavity of the exponentiated mutual information for future work, we will now look at the special case of a Gaussian assumption. Under the assumption that all our signals are Gaussian, then our definition of the normalized mutual information in (2) becomes \(K[u,v] = -(2H_q)^{-1} \ln(1-\rho_{uv}^2)\) where \(\rho_{uv}^2\) is the correlation coefficient between \(u\) and \(v\) [16]. Furthermore, \(\exp\{-cK[u,v]\} = (1-\rho_{uv}^2)^{-2c/|H_q|}\) is concave for \(c \geq (2H_q)^{-1} \rho_{uv}^2\) and convex for \(c < (2H_q)^{-1} \rho_{uv}^2\). If we treat the entries in \(r_{qj}(\cdot)\) as the mutual information between two random variables (and not the average over multiple variables, as in (8)), then the choice in (11) is made to include the best performer in the concave portion of \(\exp\{-c[c_{qj}(n)] r_{qj}(\cdot)\}\) so that (21) will hold, while putting all the other estimates in \(y(t)\) in the convex portion of the function.

IV. EXPERIMENTAL RESULTS

We define \(s_i(t)\) as a set of three mutually independent sources, and each element of \(s_i(t)\) is a stationary, zero-mean, unit variance, and white realization drawn from a distribution randomly selected from among the Laplace, Gaussian, and Uniform distributions. We will simulate two algorithms’ source estimates as \(y_i(t) = G_i s_i(t)\) and \(y_j(t) = G_j s_j(t)\) where \(G_1\) and \(G_2\) are given by,

\[
G_1 = \begin{bmatrix}
0.9996 & 0.0499 & 0.02 \\
-0.02 & 0.9986 & -0.02 \\
0.02 & -0.02 & 0.9996
\end{bmatrix}
\]

and

\[
G_2 = \begin{bmatrix}
0.9975 & 0.0499 & 0.0499 \\
-0.0499 & 0.9975 & -0.0499 \\
0.0499 & -0.0499 & 0.9975
\end{bmatrix}
\]

© 2014 SS&C & IEEE. Published in the Proceedings of the Asilomar Conference on Computers, Signals and Systems, Nov. 2-5, 2014, Pacific Grove, USA
We also define $s_*(t)$ as a set of two mutually independent sources that are each mutually independent of the sources in $s_1(t)$, and each element of $s_2(t)$ is a stationary, zero-mean, unit variance, and white realization drawn from a distribution randomly selected from among the Laplace, Gaussian, and Uniform distributions. For the experiment, the matrices $B_q$ for $q = 2, ..., 8$ are initialized with realizations of a Uniform $[0,1]$ distribution, and then normalized so the columns sum to one. Defining $s'(t) = [y_1'(t) \ y_2'(t) \ s_1'(t)]^T$, we use the instantaneous loss in (12) to define the regret at time $t$ for the $q$-set estimator as,

$$F(s_q(t), s_*(t)) = l(s_q(t)) - l(s'(t)).$$

Figures 3 and 4 show the results when $y(t) = [y_1'(t) \ y_2'(t) \ s_1'(t)]^T$ is used as the input to the blending system averaged over 25 realizations of $y(t)$.

V. CONCLUSION

Figure 3 shows that the proposed blending method selects the $q = 5$ source estimator, and Fig. 4 shows that the $q = 5$ source estimator attains the loss of the best set, $s'(t)$. We note that the $q = 2, ..., 4$ source estimators achieve a lower loss than $s'(t)$ since the loss associated with estimating $q$ sources which have the minimum mutual information is at most the loss associated with estimating $q + 1$ sources. The fact that the regret goes negative is purely a consequence of defining the best set as $s'(t)$.

The proposed method leaves plenty of room for future work. Evaluating the concavity of the exponentiated mutual information in a general setting could allow a proof that the proposed method is universal, in some sense. A hierarchical variant of the proposed method based upon branches emanating from the segregated $B_2$ case or the as yet undefined $B_1$ case will be explored. Under certain scenarios, the $B_1$ case also shows promise of selecting the best set of sources by itself. And, of course, the method should be extended to convolutive BSS with more general sources.

VI. REFERENCES


