IV. Conclusion

In this letter, an enhancement of the NBLM algorithm is proposed. It is shown that, by locally adapting one learning coefficient of the NBLM for each neighborhood, significant improvements on the performance of the method are achieved. The suggested modification requires only minor changes in the original algorithm, and reinforces the local character of the NBLM.

With the proposed local adaptation, the modified NBLM achieves better performance than the LM method even for very small neighborhood sizes. This allows very large NNs to be efficiently trained in a fraction of the time LM would require, still reaching lower error rates. Moreover, it makes possible to retain the efficiency of that method in those situations where the application of the original algorithm was impractical.

Acknowledgment

The authors would like to thank the reviewers for their useful hints and constructive comments throughout the reviewing process.

References

[1] K. Levenberg, "A method for the solution of certain problems in least squares," *Q. Appl. Math.*, vol. 2, pp. 164–168, 1944.

[2] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, pp. 431–441, 1963.

[3] M. T. Hagan and M. Menhaj, "Training feedforward networks with the Marquardt algorithm," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 989–993, Nov. 1994.

[4] G. Lera and M. Pinzolas, "A quasilocal Levenberg-Marquardt algorithm for neural network training," in *Proc. Int. Joint Conf. Neural Networks (IJCNN'98)*, vol. 3, Anchorage, AK, May, pp. 2242–2246.

[5] ——, "Neighborhood-based Levenberg-Marquardt algorithm for neural network training," *IEEE Trans. Neural Netw.*, vol. 13, no. 5, pp. 1200–1203, Sep. 2002.

[6] M. Pinzolas, J. J. Astrain, J. R. González, and J. Villadangos, "Isolated hand-written digit recognition using a neurofuzzy scheme and multiple classification," *J. Intell. Fuzzy Syst.*, vol. 12, no. 2, pp. 97–105, Dec. 2002.

[7] J. M. Cano, M. Pinzolas, J. J. Ibarrola, and J. López-Coronado, "Identificación de funciones utilizando sistemas lógicos difusos," in *Proc. XXI Jornadas Automática*, Sevilla, Spain, 2000.

[8] M. Pinzolas, J. J. Ibarrola, and J. López-Coronado, "A neurofuzzy scheme for on-line identification of nonlinear dynamical systems with variable transfer function," in *Proc. Int. Joint Conf. Neural Networks (IJCNN'00)*, Como, Italy, pp. 215–221.

Sparse Component Analysis and Blind Source Separation of Underdetermined Mixtures

Pando Georgiev, Fabian Theis, and Andrzej Cichocki

***Abstract*—In this letter, we solve the problem of identifying matrices $\mathrm{S} \in \mathbb{R}^{n\times N}$ and $\mathrm{A} \in \mathbb{R}^{m\times n}$ knowing only their multiplication $\mathrm{X} = \mathrm{AS}$, under some conditions, expressed either in terms of A and sparsity of S (*identifiability* conditions), or in terms of X (sparse component analysis (SCA) conditions). We present algorithms for such identification and illustrate them by examples.**

***Index Terms*—Blind source separation (BSS), sparse component analysis (SCA), underdetermined mixtures.**

I. Introduction

One of the fundamental questions in data analysis, signal processing, data mining, neuroscience, etc. is how to represent a large data set $\mathbf{X}$ (given in form of a $(m \times N)$-matrix) in different ways. A simple approach is a linear matrix factorization

$$\mathbf{X} = \mathbf{AS} \quad \mathbf{A} \in \mathbb{R}^{m\times n}, \quad \mathbf{S} \in \mathbb{R}^{n\times N} \tag{1}$$

where the unknown matrices $\mathbf{A}$ (dictionary) and $\mathbf{S}$ (source signals) have some specific properties, for instance:

1) the rows of $\mathbf{S}$ are (discrete) random variables, which are statistically independent as much as possible—this is independent component analysis (ICA) problem; 2) $\mathbf{S}$ contains as many zeros as possible—this is the sparse representation or sparse component analysis (SCA) problem; 3) the elements of $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{S}$ are nonnegative—this is nonnegative matrix factorization (NMF) [8].

There is a large amount of papers devoted to ICA problems [2], [5] but mostly for the case $m \geq n$. We refer to [1], [6], [7], and [9]–[11] for some recent papers on SCA and underdetermined ICA $(m < n)$.

A related problem is the so called blind source separation (BSS) problem, in which we know *a priori* that a representation such as in (1) exists and the task is to recover the sources (and the mixing matrix) as accurately as possible. A fundamental property of the complete BSS problem is that such a recovery (under assumptions in 1) and non-Gaussianity of the sources) is possible up to permutation and scaling of the sources, which makes the BSS problem so attractive.

In this letter, we consider SCA and BSS problems in the underdetermined case ($m < n$, i.e., more sources than sensors, which is more challenging problem), where the additional information compensating the limited number of sensors is the *sparseness* of the sources. It should be noted that this problem is quite general and fundamental, since the sources could be not necessarily sparse in time domain. It would be sufficient to find a linear transformation (e.g., wavelet packets), in which the sources are sufficiently sparse.

In the sequel, we present new algorithms for solving the BSS problem: matrix identification algorithm and source recovery algorithm under conditions that the source matrix $\mathbf{S}$ has at most $m - 1$ nonzero elements in each column and if the identifiability conditions

Manuscript received November 20, 2003; revised July 25, 2004.

P. Georgiev is with ECECS Department, University of Cincinnati, Cincinnati, OH 45221 USA (e-mail: pgeorgie@ececs.uc.edu).

F. Theis is with the Institute of Biophysics, University of Regensburg, D-93040 Regensburg, Germany (e-mail: fabian@theis.name).

A. Cichocki is with the Laboratory for Advanced Brain Signal Processing, Brain Science Institute, The Institute of Physical and Chemical Research (RIKEN), Saitama 351-0198, Japan (e-mail: cia@bsp.brain.riken.jp).

Digital Object Identifier 10.1109/TNN.2005.849840

are satisfied (see Theorem 1). When the sources are locally very sparse (see condition i) of Theorem 2) the matrix identification algorithm is much simpler. We used this simpler form for separation of mixtures of images. After sparsification transformation (which is in fact appropriate wavelet transformation) the algorithm works perfectly in the complete case. We demonstrate the effectiveness of our general matrix identification algorithm and the source recovery algorithm in the underdetermined case for 7 artificially created sparse source signals, such that the source matrix $\mathbf{S}$ has at most 2 nonzero elements in each column, mixed with a randomly generated (3×7) matrix. For a comparison, we present a recovery using l_1-norm minimization [3], [4], which gives signals that are far from the original ones. This implies that the conditions which ensure equivalence of l_1-norm and l_0-norm minimization [4], Theorem 7, are generally not satisfied for randomly generated matrices. Note that l_1-norm minimization gives solutions which have at most m nonzeros [3], [4]. Another connection with [4] is the fact that our algorithm for source recovery works "with probability one," i.e., for *almost all* data vectors $\mathbf{x}$ (in measure sense) such that the system $\mathbf{x} = \mathbf{A}\mathbf{s}$ has a sparse solution with less than m nonzero elements, this solution is unique, while in [4] the authors proved that for *all* data vectors $\mathbf{x}$ such that the system $\mathbf{x} = \mathbf{A}\mathbf{s}$ has a sparse solution with less than Spark$(\mathbf{A})/2$ nonzero elements, this solution is unique. Note that Spark$(\mathbf{A}) \leq m + 1$, where Spark$(\mathbf{A})$ is the smallest number of linearly dependent columns of $\mathbf{A}$.

II. Blind Source Separation

In this section, we develop a method for completely solving the BSS problem if the following assumptions are satisfied:

A1) the mixing matrix $\mathbf{A} \in IR^{m \times n}$ has the property that any square $m \times m$ submatrix of it is nonsingular;

A2) each column of the source matrix $\mathbf{S}$ has at most $m-1$ nonzero elements;

A3) the sources are sufficiently rich represented in the following sense: for any index set of $n - m + 1$ elements $I = \{i_1, \ldots, i_{n-m+1}\} \subset \{1, \ldots, n\}$ there exist at least m column vectors of the matrix $\mathbf{S}$ such that each of them has zero elements in places with indexes in I and each $m - 1$ of them are linearly independent.

A. Matrix Identification

We describe conditions in the sparse BSS problem under which we can identify the mixing matrix uniquely up to permutation and scaling of the columns. We give two type of such conditions. The first one corresponds to the least sparsest case in which such identification is possible. Further, we consider the most sparsest case (for small number of samples) as in this case the algorithm is much simpler.

1) General Case—Full Identifiability:

Theorem 1: (Identifiability Conditions—General Case): Assume that in the representation $\mathbf{X} = \mathbf{A}\mathbf{S}$ the matrix $\mathbf{A}$ satisfies condition A1), the matrix $\mathbf{S}$ satisfies conditions A2) and A3) and only the matrix $\mathbf{X}$ is known. Then the mixing matrix $\mathbf{A}$ is identifiable uniquely up to permutation and scaling of the columns.

Proof: It is clear that any column $\mathbf{a}_j$ of the mixing matrix lies in the intersection of all $\binom{n-1}{m-2}$ hyperplanes generated by those columns of $\mathbf{A}$ in which $\mathbf{a}_j$ participates.

We will show that these hyperplanes can be obtained by the columns of the data $\mathbf{X}$ under the condition of the theorem. Let $\mathcal{J}$ be the set of all subsets of $\{1, \ldots, n\}$ containing $m - 1$ elements and let $J \in \mathcal{J}$. Note that $\mathcal{J}$ consists of $\binom{n}{m-1}$ elements. We will show that the hyperplane (denoted by H_J) generated by the columns of $\mathbf{A}$ with indexes from J can be obtained by some columns of $\mathbf{X}$. By A2) and A3), there exist m indexes $\{t_k\}_{k=1}^m \subset \{1, \ldots, N\}$ such that any $m - 1$ vector columns of $\{\mathbf{S}(:, t_k)\}_{k=1}^m$ form a basis of the $(m-1)$-dimensional coordinate subspace of $\mathbb{R}^n$ with zero coordinates given by $\{1, \ldots, n\} \backslash J$. Because of the mixing model, vectors of the form

$$\mathbf{v}_k = \sum_{j \in J} S(j, t_k)\mathbf{a}_j, \quad k = 1, \ldots, m$$

belong to the data matrix $\mathbf{X}$. Now, by condition A1) it follows that any $m-1$ of the vectors $\{\mathbf{v}_k\}_{k=1}^{m-1}$ are linearly independent, which implies that they will span the same hyperplane H_J. By A1) and the above, it follows that we can cluster the columns of $\mathbf{X}$ in $\binom{n}{m-1}$ groups $\mathcal{H}_k, k = 1, \ldots, \binom{n}{m-1}$ uniquely such that each group $\mathcal{H}_k$ contains at least m elements and they span one hyperplane H_{J_k} for some $J_k \in \mathcal{J}$. Now we cluster the hyperplanes obtained in such a way in the smallest number of groups such that the intersection of all hyperplanes in each group gives a single one-dimensional (1-D) subspace. It is clear that such 1-D subspace will contain one column of the mixing matrix, the number of these groups is n and each group consists of $\binom{n-1}{m-2}$ hyperplanes. ■

The proof of this theorem gives the idea for the matrix identification algorithm.

Algorithm for Identification of the Mixing Matrix:

1) Cluster the columns of $\mathbf{X}$ in $\binom{n}{m-1}$ groups $\mathcal{H}_k, k = 1, \ldots, \binom{n}{m-1}$ such that the span of the elements of each group $\mathcal{H}_k$ produces one hyperplane and these hyperplanes are different.

2) Cluster the normal vectors to these hyperplanes in the smallest number of groups $G_j, j = 1, \ldots, n$ (which gives the number of sources n) such that the normal vectors to the hyperplanes in each group G_j lie in a new hyperplane $\hat{H}_j$.

3) Calculate the normal vectors $\hat{\mathbf{a}}_j$ to each hyperplane $\hat{H}_j, j = 1, \ldots, n$. Note that the 1-D subspace spanned by $\hat{\mathbf{a}}_j$ is the intersection of all hyperplanes in G_j. The matrix $\hat{\mathbf{A}}$ with columns $\hat{\mathbf{a}}_j$ is an estimation of the mixing matrix (up to permutation and scaling of the columns).

2) Degenerate Case—Sparse Instances:

Theorem 2: (Identifiability Conditions—Locally Very Sparse Representation): Assume that the number of sources is unknown and the following:

i) for each index $i = 1, \ldots, n$ there are at least two columns of $\mathbf{S}$: $\mathbf{S}(:, j_1)$, and $\mathbf{S}(:, j_2)$ which have nonzero elements only in position i (so each source is uniquely present at least twice);

ii) $\mathbf{X}(:, k) \neq c\mathbf{X}(:, q)$ for any $c \in \mathbb{R}$, any $k = 1, \ldots, N$ and any $q = 1, \ldots, N, k \neq q$ for which $\mathbf{S}(:, k)$ has more that one nonzero element.

Then the number of sources and the matrix $\mathbf{A}$ are identifiable uniquely up to permutation and scaling.

Proof: We cluster in groups all nonzero normalized column vectors of $\mathbf{X}$ such that each group consists of vectors which differ only by sign. From conditions i) and ii), it follows that the number of the groups containing more that one element is precisely the number of sources n, and that each such group will represent a normalized column of $\mathbf{A}$ (up to sign). ■

In the following, we include an algorithm for identification of the mixing matrix based on Theorem 2.

Algorithm for Identification of the Mixing Matrix in the Very Sparse Case:

1) Remove all zero columns of $\mathbf{X}$ (if any) and obtain a matrix $\mathbf{X}_1 \in \mathbb{R}^{m \times N_1}$.

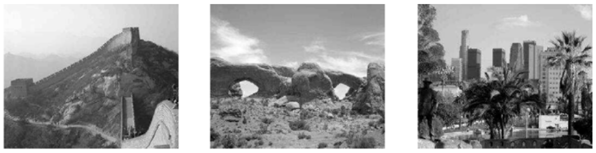

Fig. 1. Original images.

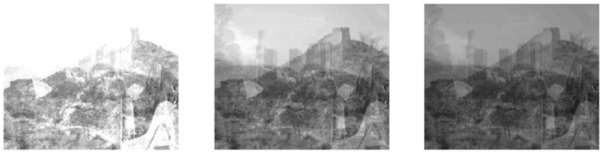

Fig. 2. Mixed (observed) images.

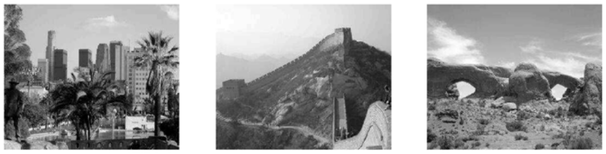

Fig. 3. Estimated normalized images using the estimated matrix. The signal-to-noise ratios with the sources from Fig. 1 are 232, 239, and 228 dB, respectively.

2) Normalize the columns $\mathbf{x}_i, i = 1,\dots,N_1$ of $\mathbf{X}_1 : \mathbf{y}_i = \mathbf{x}_i/\|\mathbf{x}_i\|$ and set $\varepsilon > 0$.
Multiply each column $\mathbf{y}_i$ by -1 if the first element of $\mathbf{y}_i$ is negative.
3) Cluster $\mathbf{y}_i, i = 1,\dots,N_1$ in $n-1$ groups $G_1,\dots,G_{n+1}$ such that for any $i = 1,\dots,n, \|\mathbf{x}-\mathbf{y}\| < \varepsilon, \forall \mathbf{x},\mathbf{y} \in G_i$, and $\|\mathbf{x}-\mathbf{y}\| \geq \varepsilon$ for any $\mathbf{x},\mathbf{y}$ belonging to different groups
4) Chose any $\mathbf{y}_i \in G_i$ and put $\mathbf{a}_i = \mathbf{y}_i$. The matrix $\mathbf{A}$ with columns $\{\mathbf{a}_i\}_{i=1}^n$ is an estimation of the mixing matrix, up to permutation and scaling.

We should mention that the very sparse case in different settings is already considered in the literature, but in more restrictive sense. In [6], the authors suppose that the supports of the Fourier transform of any two source signals are disjoint sets—a much more restrictive condition than our condition. In [1], the authors suppose that for any source there exists a time-frequency window where only this source is nonzero and that the time-frequency transform of each source is not constant on any time-frequency window. We would like to mention that their condition should include also the case when the the time-frequency transforms of any two sources are not proportional in any time-frequency window. Such a quantitative condition (without frequency representation) is presented in our Theorem 2, condition ii).

B. Identification of Sources

Theorem 3: (Uniqueness of Sparse Representation): Let $\mathcal{H}$ be the set of all $\mathbf{x} \in \mathbb{R}^m$ such that the linear system $\mathbf{A}\mathbf{s} = \mathbf{x}$ has a solution with at least $n-m+1$ zero components. If $\mathbf{A}$ fulfills A1), then there exists a subset $\mathcal{H}_0 \subset \mathcal{H}$ with measure zero with respect to $\mathcal{H}$, such that for every $\mathbf{x} \in \mathcal{H}\backslash\mathcal{H}_0$ this system has no other solution with this property.

Proof: Obviously $\mathcal{H}$ is the union of all $\begin{pmatrix} n \\ m-1 \end{pmatrix} = (n!)/((m-1)!(n-m+1)!$ hyperplanes, produced by taking the linear hull of every subsets of the columns of $\mathbf{A}$ with $m-1$ elements. Let $\mathcal{H}_0$ be the union of all intersections of any two such subspaces. Then $\mathcal{H}_0$ has a measure zero in $\mathcal{H}$ and satisfies the conclusion of the theorem. Indeed, assume that $\mathbf{x} \in \mathcal{H}\backslash\mathcal{H}_0$ and $\mathbf{A}\mathbf{s} = \mathbf{A}\bar{\mathbf{s}} = \mathbf{x}$, where $\mathbf{s}$ and $\bar{\mathbf{s}}$ have at least $n-m+1$ zeros. Since $\mathbf{x} \notin \mathcal{H}_0, \mathbf{x}$ belongs to only one hyperplane produced as a linear hull of some $m-1$ columns $\mathbf{a}_{i_1},\dots,\mathbf{a}_{i_{m-1}}$ of $\mathbf{A}$. It means that the vectors $\mathbf{s}$ and $\bar{\mathbf{s}}$ have $n-m+1$ zeros in places with indexes in $\{1,\dots,n\}\backslash\{i_1,\dots,i_{m-1}\}$. Now from the equation $\mathbf{A}(\mathbf{s}-\bar{\mathbf{s}}) = 0$ it follows that the $m-1$ vector columns $\mathbf{a}_{i_1},\dots,\mathbf{a}_{i_{m-1}}$ of $\mathbf{A}$ are linearly dependent, which is a contradiction with A1). ■

From Theorem 3 it follows that the sources are identifiable generically, i.e., up to a set with a measure zero, if they have level of sparseness grater than or equal to $n-m+1$, and the mixing matrix is known. In the following, we present an algorithm, based on the observation in Theorem 3.

Source Recovery Algorithm:

1) Identify the the set of k-codimensional subspaces $\mathcal{H}$ produced by taking the linear hull of every subsets of the columns of $\mathbf{A}$ with $m-1$ elements.
2) Repeat for $i = 1$ to N:
2.1) Identify the space $H \in \mathcal{H}$ containing $\mathbf{x}_i := \mathbf{X}(:,i)$, or, in practical situation with presence of noise, identify the one to which the distance from $\mathbf{x}_i$ is minimal and project $\mathbf{x}_i$ onto H to $\tilde{\mathbf{x}}_i$.
2.2) if H is produced by the linear hull of column vectors $\mathbf{a}_{i_1},\dots,\mathbf{a}_{i_{m-1}}$, then find coefficients $\lambda_{i,j}$ such that

$$\tilde{\mathbf{x}}_i = \sum_{j=1}^{m-1} \lambda_{i,j}\mathbf{a}_{i_j}.$$

These coefficients are uniquely determined if $\tilde{\mathbf{x}}_i$ does not belong to the set $\mathcal{H}_0$ with measure zero with respect to $\mathcal{H}$ (see Theorem 3);
2.3) Construct the solution $\mathbf{s}_i = \mathbf{S}(:,i)$: it contains $\lambda_{i,j}$ in the place i_j for $j = 1,\dots,m-1$, the other its components are zero.

III. SCA

In this section, we develop a method for the complete solution of the SCA problem. Now the conditions are formulated only in terms of the data matrix $\mathbf{X}$.

Theorem 4: (SCA Conditions): Assume that $m \leq n \leq N$ and the matrix $\mathbf{X} \in \mathbb{R}^{m\times N}$ satisfies the following conditions:

i) the columns of $\mathbf{X}$ lie in the union $\mathcal{H}$ of $\begin{pmatrix} n \\ m-1 \end{pmatrix}$ different hyperplanes, each column lies in only one such hyperplane, each hyperplane contains at least m columns of $\mathbf{X}$ such that each $m-1$ of them are linearly independent;
ii) for each $i \in \{1,\dots,n\}$ there exist $p = \begin{pmatrix} n-1 \\ m-2 \end{pmatrix}$ different hyperplanes $\{H_{i,j}\}_{j=1}^p$ in $\mathcal{H}$ such that their intersection $L_i = \cap_{k=1}^p H_{i,j}$ is 1-D subspace;
iii) any m different L_i span the whole $\mathbb{R}^m$.

Then the matrix $\mathbf{X}$ is representable uniquely (up to permutation and scaling of the columns of $\mathbf{A}$ and rows of $\mathbf{S}$) in the form $\mathbf{X} = \mathbf{A}\mathbf{S}$, where the matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{S} \in \mathbb{R}^{n\times N}$ satisfy the conditions A1) and A2), A3), respectively.

Proof: Let L_i be spanned by $\mathbf{a}_i$ and set $\mathcal{A} = \{\mathbf{a}_i\}_{i=1}^n$. Condition iii) implies that any hyperplane from $\mathcal{H}$ contains at most $m-1$ vectors from $\mathcal{A}$. By i) and ii), it follows that these vectors are exactly $m-1$: only in this case the calculation of the number of all hyperplanes by ii) will give the number in i): $n\begin{pmatrix} n-1 \\ m-2 \end{pmatrix}/(m-1) = \begin{pmatrix} n \\ m-1 \end{pmatrix}$. Let $\mathbf{A}$ be a matrix whose column vectors are all vectors from $\mathcal{A}$ (taken in an arbitrary order). Since every column vector $\mathbf{x}$ of $\mathbf{X}$ lies only in one hyperplane from $\mathcal{H}$, the linear system $\mathbf{A}\mathbf{s} = \mathbf{x}$ has unique solution, which has at least $n-m+1$ zeros (see the Proof of Theorem 3). Let $\{\mathbf{x}_i\}_{i=1}^m$ be m column vectors from $\mathbf{X}$, which span one hyperplane from $\mathcal{H}$, and $m-1$ of them are linearly independent (such vectors exist by i)). Then we have: $\mathbf{A}\mathbf{s}_i = \mathbf{x}_i$, for some uniquely determined vectors $\mathbf{s}_i, i = 1,\dots,m-1$, which are linearly independent and have

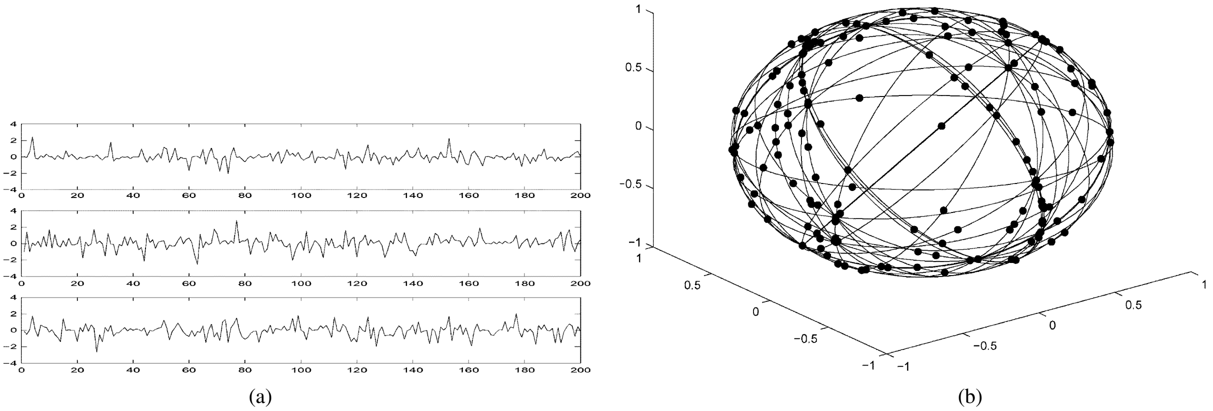

Fig. 4. (a) Mixed signals and (b) normalized scatter plot (density) of the mixtures together with the 21 data set hyperplanes, visualized by their intersection with the unit sphere in $\mathbb{R}^3$.

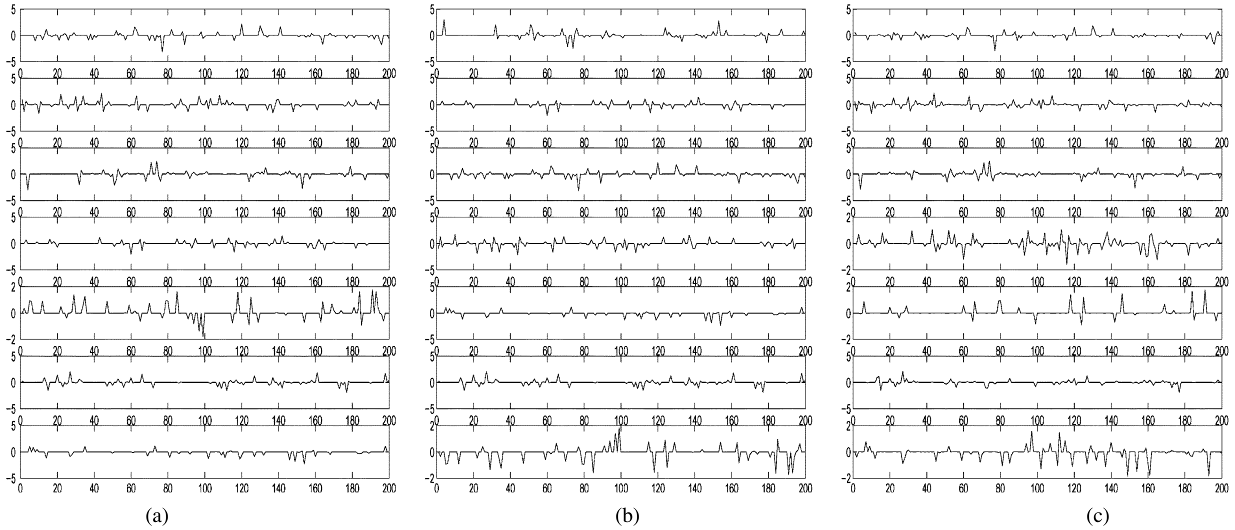

Fig. 5.(a) Original source signals.(b) Recovered source signals—the signal-to-noise ratio between the original sources and the recoveries is very high (above 278 dB after permutation and normalization). Note that only 200 samples are enough for excellent separation. (c) Recovered source signals using l_1-norm minimization and known mixing matrix. Simple comparison confirms that the recovered signals are far from the original ones, and the signal-to-noise ratio is only around 4 dB.

at least $n - m + 1$ zeros in the same coordinates. In such a way, we can write: $\mathbf{X} = \mathbf{AS}$ for some uniquely determined matrix $\mathbf{S}$, which satisfies A2) and A3). ■

We should mention that our algorithms are robust with respect to small additive noise and big outliers, since the algorithms cluster the data on hyperplanes approximately, up to a threshold $\varepsilon > 0$, which could accumulate a noise with amplitude less than ε. The big outliers will not be clustered to any hyperplane.

IV. Computer Simulation Examples

A. Complete Case

In this example for the complete case $(m = n)$ of instantaneous mixtures, we demonstrate the effectiveness of our algorithm for identification of the mixing matrix in the special case considered in Theorem 2. We mixed three images of landscapes (shown in Fig. 1) with a three-dimensional (3-D) Hilbert matrix $\mathbf{A}$ and transformed them by a two-dimensional (2-D) discrete Haar wavelet transform. As a result, since this transformation is linear, the high frequency components of the source signals become very sparse and they satisfy the conditions of Theorem 2. We use only one row (320 points) from the diagonal coefficients of the wavelet transformed mixture, which is enough to recover very precisely the ill conditioned mixing matrix $\mathbf{A}$. Fig. 3 shows the recovered mixtures.

B. Underdetermined Case

We consider a mixture of seven artificially created sources (see Fig. 5)—sparsified randomly generated signals with at least 5 zeros in each column—with a randomly generated mixing matrix with dimension 3×7. Fig. 4 gives the mixed signals together with a normalized scatterplot of the mixtures—the data lies in $21 = \binom{7}{2}$ hyperplanes. Applying the underdetermined matrix recovery algorithm to the mixtures gives the recovered mixing matrix perfectly well, up to permutation and scaling (not shown because of lack of space). Applying the source recovery algorithm, we recover the source signals up to permutation and scaling (see Fig. 5). This figure also shows that the recovery by l_1-norm minimization does not perform well, even if the mixing matrix is perfectly known.

V. Conclusion

We defined rigorously the SCA and BSS problems of sparse signals and presented sufficient conditions for their solving. We developed three algorithms: for identification of the mixing matrix (two types: for the sparse and the very sparse cases) and for source recovery. We presented two experiments: the first one concerns separation of a mixture of images, after wavelet sparsification (producing very sparse sources), which performs very well in the complete case. The second one shows the excellent performance of the another two our algorithms in the underdetermined BSS problem, for separation of artificially created signals with sufficient level of sparseness.

References

[1] F. Abrard, Y. Deville, and P. White, "From blind source separation to blind source cancellation in the underdetermined case: A new approach based on time-frequency analysis," in *Proc. 3rd Int. Conf. Independent Component Analysis and Signal Separation (ICA'2001)*, San Diego, CA, Dec. 9–13, 2001, pp. 734–739.

[2] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*. New York: Wiley, 2002.

[3] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[4] D. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.

[5] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley, 2001.

[6] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Proc. 2000 IEEE Conf. Acoustics, Speech, and Signal Processing (ICASSP'00)*, vol. 5, Istanbul, Turkey, Jun. 2000, pp. 2985–2988.

[7] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind sourse separation of more sources than mixtures using overcomplete represeintaitons," *IEEE Signal Process. Lett.*, vol. 6, no. 4, pp. 87–90, 1999.

[8] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 40, pp. 788–791, 1999.

[9] F. J. Theis, E. W. Lang, and C. G. Puntonet, "A geometric algorithm for overcomplete linear ICA," *Neurocomput.*, vol. 56, pp. 381–398, 2004.

[10] K. Waheed and F. Salem, "Algebraic overcomplete independent component analysis," in *Proc. Int. Conf. Independent Component Analysis (ICA'03)*, Nara, Japan, pp. 1077–1082.

[11] M. Zibulevsky and B. A. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Comput.*, vol. 13, no. 4, pp. 863–882, 2001.

Equivalence Between RAM-Based Neural Networks and Probabilistic Automata

Marcilio C. P. de Souto, Teresa B. Ludermir, and Wilson R. de Oliveira

***Abstract*—In this letter, the computational power of a class of random access memory (RAM)-based neural networks, called general single-layer sequential weightless neural networks (GSSWNNs), is analyzed. The theoretical results presented, besides helping the understanding of the temporal behavior of these networks, could also provide useful insights for the developing of new learning algorithms.**

***Index Terms*—Automata theory, computability, p random access memory (RAM) node, probabilistic automata, RAM-based neural networks, weightless neural networks (WNNs).**

I. Introduction

The neuron model used in the great majority of work involving neural networks is related to variations of the McCulloch–Pitts neuron, which will be called the *weighted neuron*. A typical weighted neuron can be described by a linear weighted sum of the inputs, followed by some nonlinear transfer function [1], [2]. In this letter, however, the neural computing models studied are based on artificial neurons which often have binary inputs and outputs, and no adjustable weight between nodes. Neuron functions are stored in lookup tables, which can be implemented using commercially available random access memories (RAMs). These systems and the nodes that they are composed of will be described, respectively, as weightless neural networks (WNNs) and weightless nodes [1], [2]. They differ from other models, such as the weighted neural networks, whose training is accomplished by means of adjustments of weights. In the literature, the terms "RAM-based" and "N-tuple based" have been used to refer to WNNs.

In this letter, the computability (computational power) of a class of WNNs, called general single-layer sequential weightless neural networks (GSSWNNs), is investigated. Such a class is an important representative of the research on temporal pattern processing in (WNNs) [3]–[8]. As one of the contributions, an algorithm (constructive proof) to map any probabilistic automaton (PA) into a GSSWNN is presented. In fact, the proposed method not only allows the construction of any PA, but also increases the class of functions that can be computed by such networks. For instance, at a theoretical level, these networks are not restricted to finite-state languages (regular languages) and can now deal with some context-free languages. Practical motivations for investigating probabilistic automata and GSSWNNs are found in their possible application to, among others things, syntactic pattern recognition, multimodal search, and learning control [9].

II. Definitions

A. Probabilistic Automata

Probabilistic automata are a generalization of ordinary deterministic finite state automata (DFA) for which an input symbol could take the automaton into any of its states with a certain probability [9].

Definition 2.1: A PA is a 5-tuple $\mathbf{A}_P = (\Sigma, Q, H, q_I, F)$, where

- $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_{|\Sigma|}\}$ is a finite set of ordered symbols called the *input alphabet*;
- $Q = \{q_0, q_2, \ldots, q_{|Q|}\}$ is a finite set of states;
- H is a mapping of $Q \times \Sigma$ into the set of $n \times n$ stochastic state transition matrices (where n is the number of states in Q). The interpretation of $H(a_m), a_m \in \Sigma$, can be stated as follows. $H(a_m) = [p_{ij}(a_m)]$, where $p_{ij}(a_m) \geq 0$ is the probability of entering state q_j from state q_i under input a_m, and $\sum_{j=1}^{n} p_{ij} = 1$, for all $i = 1, \ldots, n$. The domain of H can be extended from Σ to Σ^* by defining the following:
  1) $H(\epsilon) = I_n$, where ϵ is the empty string and I_n is an $n \times n$ identity matrix;
  2) $H(a_{m_1}, a_{m_2}, \ldots, a_{m_k}) = H(a_{m_1})H(a_{m_2}), \ldots, H(a_{m_k})$, where $k \geq 2$ and $a_{m_j} \in \Sigma, j = 1, \ldots, k$.
- $q_I \in Q$ is the initial state in which the machine is found before the first symbol of the input string is processed;
- F is the set of final states ($F \subseteq Q$).

The language accepted by a PA $\mathbf{A}_P$ is $T(\mathbf{A}_P) = \{(\omega, p(\omega)) | \omega \in \Sigma^*, p(\omega) = \pi_0 H(\omega)\pi_F > 0\}$ where: 1) π_0 is a n-dimensional row vector, in which the ith component is equal to one if $q_i = q_I$, and 0 otherwise and 2) π_F is an n-dimensional column vector, in which the jth component is equal to 1 if $q_j \in F$ and 0 otherwise.

The language accepted by $\mathbf{A}_P$ with *cut-point(threshold)* λ, such that $0 \leq \lambda < 1$, is $L(\mathbf{A}_P, \lambda) = \{\omega | \omega \in \Sigma^*$ and $\pi_0 H(\omega)\pi_F > \lambda\}$. ■

Probabilistic automata recognize exactly the class of weighted regular languages (WRLs) [9]. Such a class of languages includes properly

Manuscript received July 6, 2003; revised October 26, 2004.

M. C. P. de Souto is with the Department of Informatics and Applied Mathematics, Federal University of Rio Grande do Norte, Campus Universitario, Natal-RN 59072-970, Brazil (e-mail: marcilio@dimap.ufrn.br).

T. B. Ludermir is with the Center of Informatics, Federal University of Pernambuco, Recife 50732-970, Brazil.

W. R. de Oliveira is with the Department of Physics and Mathematics, Rural Federal University of Pernambuco, Recife 52171-900, Brazil.

Digital Object Identifier 10.1109/TNN.2005.849838