Theoretical Method for Solving BSS-ICA Using SVM

Carlos G. Puntonet[2], Juan Manuel Górriz[1], Moisés Salmerón[2], and Susana Hornillo-Mellado[3]

[1] E.P.S. Algeciras, Universidad de Cádiz, Avda. Ramón Puyol s/n, 11202 Algeciras Cádiz, Spain
juanmanuel.gorriz@uca.de
[2] E.S.I., Informática, Universidad de Granada C/ Periodista Daniel Saucedo, 18071 Granada, Spain
{carlos,moises}@atc.ugr.es
[3] Escuela Superior de Ingenieros, Universidad de Sevilla Avda. de los Descubrimientos s/n 41092 Sevilla , Spain
susanah@us.es

Abstract. In this work we propose a new method for solving the blind source separation (BSS) problem using a support vector machine (SVM) workbench. Thus, we provide an introduction to SVM-ICA, a theoretical approach to unsupervised learning based on learning machines, which has frequently been proposed for classification and regression tasks. The key idea is to construct a Lagrange function from both the objective function and the corresponding constraints, by introducing a dual set of variables and solving the optimization problem. For this purpose we define a specific cost function and its derivative in terms of independence, i.e. inner products between the output and the objective function, transforming an unsupervised learning problem into a supervised learning machine task where optimization theory can be applied to develop effective algorithms.

1 Introduction

Independent Component Analysis (ICA) is a recently developed method in which the goal is to find a suitable representation of non-gaussian sources so that the components are as independent as possible [1]. ICA has been applied successfully to fields such as biomedicine, speech, sonar and radar, signal processing and, more recently, to time series forecasting [2].

There exists a wide range of ICA algorithms for solving blind source separation (BSS) problems, consisting of the minimization (or maximization) of a contrast function [3–7]. In practice, thus ICA, is an algorithm for maximizing the selected statistical principle, i.e. the stochastic gradient descent method can be used to minimize mutual information. The heuristics (learning rates, starting parameters) used in this kind of methods, however, damage the convergence rates. The gradient-based method fails to obtain the correct parameters of the separation system from different initializations due to its limited local search

C.G. Puntonet and A. Prieto (Eds.): ICA 2004, LNCS 3195, pp. 256–262, 2004.

ability and to the complex nonlinear characteristics of the problem (nonlinear or high dimensional ICA)[8].

Optimization Theory is the branch of mathematics concerned with characterizing the solutions to such problems and with developing efficient algorithms for finding such solutions. Any optimization problem can be described using an objective function and equality or inequality constraints (functions defined in a domain $\Omega \subset \mathcal{R}^n$). Depending on the nature of these functions, the problem is called a linear, quadratic, etc. programme.

In this paper, support vector machine (SVM) methodology is applied to ICA in the search for the separation matrix, in order to make use of feature space learning and the numerous regression algorithms developed in this context. The paper is organized as follows; in Section 2 we give a brief overview of basic ICA theory and introduce the notation used in the rest of the paper. The new method is presented in Sections 3 and 4 and some conclusions are drawn in section 5.

2 Definition of ICA

We define ICA using a statistical latent variables model (Jutten & Herault, 1991). Assuming the number of sources n is equal to the number of mixtures, the linear model can be expressed as:

$$x_j(t) = b_{j1}s_1 + b_{j2}s_2 + \ldots + b_{jn}s_n \quad \forall j = 1 \ldots n \ , \tag{1}$$

where we explicitly emphasize the time dependence of the samples of the random variables and assume that both the mixture variables and the original sources have zero mean without loss of generality. Using matrix notation instead of sums and including additive noise, the latter mixing model can be written as:

$$\mathbf{x}(t) = \mathbf{B} \cdot \mathbf{s}(t) + \mathbf{b}(t) \ , \ or \tag{2}$$

$$\mathbf{s}(t) = \mathbf{A} \cdot \mathbf{x}(t) + \mathbf{c}(t) \ , \ where \ \mathbf{A} = \mathbf{B}^{-1}, \ \mathbf{c}(t) = -\mathbf{B}^{-1} \cdot \mathbf{b}(t) \ . \tag{3}$$

The conditions that must be satisfied to guarantee the separation are given by Darmois' Theorem in [9]. In brief, the components s_i must be non-gaussian statistically independent. For simplicity, we assume that the unknown matrix is square and that the mixing can be characterized by a linear scenario. Noise is included in the model for two reasons: because the classical statistical linear model is used and because in many applications there is some noise in the measurements (the 'cocktail party' effect).

3 ICA and Convex Optimization Under Discrepancy Constraints

In order to solve ICA problems using the SVM paradigm, we use an approach based on reformulating the determination of the unknown demixing matrix $\mathbf{A} = \mathbf{B}^{-1}$ in the model (3) as a convex optimization problem. The optimization

program we formulate is solved using the Lagrange multiplier method combined with an approximation to a given derivative of a convenient discrepancy function based on cumulants or on the characteristic function of the original sources. Note that our approach could easily be modified to take into account other paradigms in ICA research such as density estimation-based approximation methods.

We first restrict the range of possible solutions to the problem, by what is usually a reasonable normalizing constraint: that the Frobenius norm of the matrix $\mathbf{A}$ that we wish to find is minimum. We take the following, however, to be our explicit objective function:

$$minimize \quad \frac{1}{2} \cdot \|\mathbf{A}\|_2^2, \tag{4}$$

because this makes our program a convex one (at least with the Frobenius norm). The discrepancy between the model and what is iteratively observed is contained in the restrictions:

$$-\epsilon < \tilde{L}(\mathbf{a}_i) < \epsilon \ , (i = 1, 2, \ldots, n) \ . \tag{5}$$

where, for each time instant t, we have $\tilde{L}(\mathbf{a}_i) \approx< \mathbf{a}_i, \mathbf{x} > -c_i - s_i$, with $\mathbf{a}_i$ denoting the i-th row of the demixing matrix $\mathbf{A}$, and c_i being the i-th component on vector $\mathbf{c}$. Note that for simplicity we have not written the dependency on the time instant t, but of course this must be taken into account when implementing.

We define the Lagrangian corresponding to (5) as (introducing a soft margin in equation 5)

$$\begin{array}{c} \mathcal{L}_i = \frac{1}{2} \cdot \|\mathbf{a}_i\|_2^2 + C \cdot \sum_{j=1}^{l} (\xi_j + \xi_j^*) - \sum_{j=1}^{l} \alpha_j (\epsilon + \xi_j + \tilde{L}(\mathbf{a}_i)) \\ - \sum_{j=1}^{l} \alpha_j^* (\epsilon + \xi_j^* - \tilde{L}(\mathbf{a}_i)) - \sum_{j=1}^{l} (\eta_j \xi_j + \eta_j^* \xi_j^*) \ . \end{array} \tag{6}$$

where l is the number of samples and $\xi_j, \xi_j^*, \alpha_j, \alpha_j^*, \eta_j, \eta_j^*$ are the slack variables introduced in Lagrangian optimization problems.

Now we take the corresponding partial derivatives (according to the Lagrangian method) and equal them to 0, as follows

$$\partial_{c_i} \mathcal{L}_i = \sum_{j=1}^{l} (\alpha_j^* + \alpha_j) = 0 \ . \tag{7}$$

$$\partial_{\xi_i^{(*)}} \mathcal{L}_i = C - \alpha_j^{(*)} - \eta_j^{(*)} = 0 \ . \tag{8}$$

$$\partial_{\mathbf{a}_i} \mathcal{L}_i = \mathbf{a}_i - \sum_{j=1}^{l} (\alpha_j - \alpha_j^*) \cdot \partial_{\mathbf{a}_i} \tilde{L}(\mathbf{a}_i) = 0 \ . \tag{9}$$

From equation 9 we see how the algorithm is able to extract independent components one by one, just working with the maximization of the selected Lagrangian function $\mathcal{L}_i$. The selection of a suitable function $\tilde{L}(\mathbf{a}_i, \mathbf{x})$ determines the current algorithm or strategy used in the process, i.e. if we describe it in terms

of neg-entropy we obtain a generalization of FastICA [7]. After some algebraic manipulation, we obtain

$$\mathcal{L}_i = \frac{1}{2} \cdot \left\| \sum_{j=1}^{l} (\alpha_j - \alpha_j^*) \partial_{\mathbf{a}_i} \tilde{L}(\mathbf{a}_i) \right\|^2 - \epsilon \sum_{j=1}^{l} (\alpha_j + \alpha_j^*) - \sum_{j=1}^{l} (\alpha_j - \alpha_j^*) \tilde{L}(\mathbf{a}_i). \quad (10)$$

Finally, ICA is transformed into a multidimensional maximization of the Lagrangian function defined as:

$$\mathcal{L} = \begin{pmatrix} \mathcal{L}_1 \\ \mathcal{L}_2 \\ \vdots \\ \mathcal{L}_n \end{pmatrix} . \quad (11)$$

4 Statistical Independence Criterion

The Statistical Independence of a set of random variables can be described in terms of their joint and individual probability distribution. The independence condition for the independent components of the output vector $\mathbf{y}$ is given by the following definition of independence random variables:

$$p_{\mathbf{y}}(\mathbf{y}) = \prod_{i=1}^{n} p_{\mathbf{y}_i}(y_i) \quad (12)$$

where $p_{\mathbf{y}}$ is the joint pdf of the random vector (observed signals) $\mathbf{y}$ and $p_{\mathbf{y}_i}$ is the marginal PDF of y_i. In order to measure the independence of the outputs, equation 12 is expressed in terms of higher order statistics (cumulants) using the characteristic function (or moment generating function) $\phi(\mathbf{k})$, where $\mathbf{k}$ is a vector of variables in the Fourier transform domain, and considering its natural logarithm $\Phi = log(\phi(\mathbf{k}))$. We first evaluate the difference between the terms in equation 12 to obtain:

$$\pi(\mathbf{y}) = \left\| p_{\mathbf{y}}(\mathbf{y}) - \prod_{i=1}^{n} p_{\mathbf{y}}(y_i) \right\|^2 \quad (13)$$

where the norm $||\dots||^2$ can be defined using the convolution operator with different window functions according to the specific application [8] as follows:

$$\|F(y)\|^2 = \int \{F(\mathbf{y}) * v(\mathbf{y})\}^2 d\mathbf{y} \quad (14)$$

and $v(\mathbf{y}) = \prod_{i=1}^{n} w(y_i)$. In the Fourier domain and taking natural log (in order to use higher order statistics, i.e. cumulants) this equation is transformed into:

$$\Pi(\mathbf{k}) = \int \left\| \mathbf{\Psi}_{\mathbf{y}}(\mathbf{k}) - \sum_{i=1}^{n} \mathbf{\Psi}_{\mathbf{y_i}}(k_i) \right\|^2 \mathbf{V}(\mathbf{k}) d\mathbf{k} \quad (15)$$

where $\mathbf{\Psi}$ is the cumulant generating or characteristic function (the natural log of the moment generating function) and V is the Fourier transform of the selected window function $v(\mathbf{y})$. If we take the Taylor expansion around the origin of the characteristic function, we obtain:

$$\mathbf{\Psi_y}(\mathbf{k}) = \sum_{\lambda} \frac{1}{\lambda!} \frac{\partial^{|\lambda|} \mathbf{\Psi_y}}{\partial k_1^{\lambda_1} \dots \partial k_n^{\lambda_n}} (\mathbf{0}) k_1^{\lambda_1} \dots k_n^{\lambda_n} \tag{16}$$

where we define $|\lambda| \equiv \lambda_1 + \ldots + \lambda_n$, $\lambda \equiv \{\lambda_1 \dots \lambda_n\}$, $\lambda! \equiv \lambda_1! \dots \lambda_n!$ and:

$$\mathbf{\Psi_{y_i}}(\mathbf{k_i}) = \sum_{\lambda_i} \frac{1}{\lambda_i!} \frac{\partial^{\lambda_i} \mathbf{\Psi_{y_i}}}{\partial k_i^{\lambda_i}} (0) k_i^{\lambda_i} \tag{17}$$

where the factors in the latter expansions are the cumulants of the outputs (cross and non-cross cumulants):

$$C_{y_1 \dots y_n}^{\lambda_1 \dots \lambda_n} = (-j)^{|\lambda|} \frac{\partial^{\lambda_1 + \ldots + \lambda_n} \mathbf{\Psi_y}}{\partial k_1^{\lambda_1} \dots \partial k_n^{\lambda_n}} (\mathbf{0}) \qquad C_{y_i}^{\lambda_i} = (-j)^{\lambda_i} \frac{\partial^{\lambda_i} \mathbf{\Psi_{y_i}}}{\partial k_i^{\lambda_i}} (0) \tag{18}$$

Thus, we define the difference between the terms in equation 15 as

$$\beta_\lambda = \frac{1}{\lambda!} (j)^{|\lambda|} C_{\mathbf{y}}^{\lambda} \tag{19}$$

which contains the infinite set of cumulants of the output vector $\mathbf{y}$. By substituting 19 into 15 we obtain

$$\mathit{\Pi}(\mathbf{k}) = \int \left\| \sum_{\lambda} \beta_\lambda k_1^{\lambda_1} \dots k_n^{\lambda_n} \right\|^2 \mathbf{V}(\mathbf{k}) d\mathbf{k} \tag{20}$$

Hence, vanishing cross-cumulants are a necessary condition for $y_1, \dots, y_n$ to be independent[1]. Equation 20 can be transformed into:

$$\mathit{\Pi}(\mathbf{k}) = \int \sum_{\lambda, \lambda^*} \beta_\lambda \beta_{\lambda^*}^* k_1^{\lambda_1 + \lambda_1^*} \dots k_n^{\lambda_n + \lambda_n^*} \mathbf{V}(\mathbf{k}) d\mathbf{k} \tag{21}$$

Finally, by interchanging the sequence of summation and integral equation 21 can be rewritten as:

$$\mathit{\Pi} = \sum_{\lambda, \lambda^*} \beta_\lambda \beta_{\lambda^*}^* \mathbf{\Gamma}_{\lambda, \lambda^*} \tag{22}$$

where $\mathbf{\Gamma} = \int k_1^{\lambda_1 + \lambda_1^*} \dots k_n^{\lambda_n + \lambda_n^*} \mathbf{V}(\mathbf{k}) d\mathbf{k}$. In this way, we describe the generic function $\tilde{\mathbf{L}}$ in the Lagrangian function $\mathcal{L}$. We must impose some additional restrictions on $\tilde{\mathbf{L}}$, which is a version of the previous one but limiting the set λ. That is, we only consider a finite set of cumulants $\{\lambda, \lambda^*\}$ such as $|\lambda| + |\lambda^*| < \tilde{\lambda}$

[1] In practice, we need independence between sources two against two.

and include only the cumulants affecting the current Lagrangian component. Mathematically, these two restrictions are expressed as:

$$\tilde{\mathbf{L}}_i \equiv \mathit{\Pi} = \sum_{\{\lambda,\lambda^*\}} \beta_\lambda \beta^*_{\lambda^*} \mathbf{\Gamma}_{\lambda,\lambda^*} \qquad \setminus \left\{ \begin{array}{c} \{\lambda,\lambda^*\} \bigcap \{\lambda_i\} \neq 0 \\ |\lambda| + |\lambda^*| < \tilde{\lambda} \end{array} \right\} \tag{23}$$

In order to evaluate the most relevant term in the Lagrangian $\frac{\partial \tilde{\mathbf{L}}}{\partial \mathbf{a_i}}$ the above equations must be rewritten in terms of the output vector as $y_i = \mathbf{a_i}\mathbf{x}$, and we must use the connection between cumulants and moments shown in [10]:

$$\frac{\partial \tilde{\mathbf{L}}_i}{\partial \mathbf{a_i}} \propto \frac{\partial C^\lambda_{\mathbf{y}}}{\partial \mathbf{a_i}} \propto \frac{\partial \mathbf{a_i} \cdot \mathbf{x}}{\partial \mathbf{a_i}} \tag{24}$$

4.1 Using the Connection Between Moments and Cumulants

The connection between moments and cumulants can be expressed as:

$$C^\lambda_{\mathbf{y}} = \sum_{p_1,\ldots,p_m} (-1)^{m-1}(m-1)! \cdot E[\prod_{j\in p_1} Y_j] \ldots E[\prod_{j\in p_m} Y_j] \tag{25}$$

where $\{p_1,\ldots,p_m\}$ are all the possible partitions with $m = 1,\ldots,\lambda$ included in the set of integers $\{1,\ldots,\lambda\}$. In SVM methodology, we work with instantaneous values (sample by sample) and thus we have to approximate expected values to instantaneous ones. Finally, by evaluating the derivative term in equation 25 and using the above-mentioned approximations, we obtain

$$\frac{\partial C^\lambda_{\mathbf{y}}}{\partial \mathbf{a_i}} = \sum_{p_1,\ldots,p_m} (-1)^{m-1}(m-1)! \cdot \sum_{k=1}^{m} \left(\frac{s_k (A^{-1} \cdot \mathbf{y})^{s_k - 1}}{y_i^{s_k}} \prod_{j\in p_1} y_j \cdots \prod_{j\in p_m} y_j \right) \tag{26}$$

where λ satisfies the conditions shown in equation 23 and s_k is an integer in the set $\{1,\ldots,\tilde{\lambda}\}$. In practice, the order of the statistics used never exceeds four or five, and so the latter expression can be simplified significantly, rewriting the cumulants in terms of dot products between the output signals y_i. Expressions of cumulants in terms of moments are well-known and thus equations 26 and 9 allow us to iteratively obtain the coefficients α_j, α^*_j and then the support vector parameters $\mathbf{a_i}$ of the separation matrix $\mathbf{A}$:

$$\begin{array}{c} \mathbf{a}_i = \sum_{j=1}^{l} (\alpha_j - \alpha^*_j) \cdot \partial_{\mathbf{a}_i} \tilde{L}(\mathbf{a}_i) = \sum_{j=1}^{l} (\alpha_j - \alpha^*_j) \cdot \sum_{\{\lambda,\lambda^*\}} \partial_{\mathbf{a}_i} (\beta_\lambda \beta^*_{\lambda^*}) \mathbf{\Gamma}_{\lambda,\lambda^*} \\ = \sum_{j=1}^{l} (\alpha_j - \alpha^*_j) \cdot \sum_{\{\lambda,\lambda^*\}} \frac{(j)^{|\lambda|+|\lambda^*|}}{\lambda!\lambda^*!} \partial_{\mathbf{a}_i} \left(\mathbf{C}^\lambda_{\mathbf{y}} \mathbf{C}^{\lambda^*}_{\mathbf{y}} \right) \mathbf{\Gamma}_{\lambda,\lambda^*} \end{array} \tag{27}$$

5 Conclusions

A support vector-based BSS-ICA method has been developed to solve the BSS problem from linear mixtures of independent sources. The generalization to non-linear ICA is straightforward considering nonlinear maps to feature spaces. The

proposed method obtains a good performance (this statement is back up by the extensive work in the workbench of SVM algorithms), and benefits from the Theoretical Optimization Theory, which consists of solving a uniquely solvable (with order n) optimization problem instead of Newton or gradient descent methods, which require suitable nonlinear optimization, with the consequent risk of getting stuck in local minima.

The tacit assumption in equation 5 avoids cases such as in noisy environments where the separation matrix does not actually exist as a linear function between independent components and observed signals, i.e. the convex optimization problem is not feasible. That is, in cases where the separation is not possible, we use a "soft margin" by introducing slack variables to cope with the otherwise unfeasible constraints of the optimization problem [11]. The main disadvantage of this kind of methods is that Quadratic programs are computationally quite expensive as they scale between quadratic and cubic in the number of patterns although there exists a unique solution, but this is also true for algebraic algorithms like e.g. Cardoso's JADE [12].

References

1. Hyvarynen, A., Oja, E., Independent Component Analysis: Algorithms and Applications Neural Networks Vol 13 411-430 Elsevier (2000)
2. Górriz, J.M., Puntonet, C.G., Salmerón, M., Ortega, J., New method for filtered ICA signals applied to volatile time series 7th International Work Conference on Artificial and Natural Neural Networks IWANN 2003 Lecture Notes in Computer Science Vol 2687 / 2003, Springer pp. 433-440 ISSN: 0302-9743. Menorca, Balearic Islands, Spain. Jun. 2003.
3. Barlow, H.B, Possible principles underlying transformation of Sensory messages. Sensory Communication, W.A. Rosenblith, MIT Press, New York, U.S.A. (1961).
4. Bell,A.J., Sejnowski, T.J. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Computation, vol 7, 1129-1159 (1995).
5. Cardoso, J.F., Infomax and maximum likelihood for source separation. IEEE Letters on signal processing, 4, 112-114 (1997).
6. Cichoki, A., Unbehauen, R., Robust neural networks with on-line learning for blind identification and blind separation of sources. IEEE Transactions on Circuits and Systems, 43 (11), 894-906 (1996).
7. Hyvärinen, A., Oja, E., A fast fixed point algorithm for independent component analysis. Neural Computation, 9: 1483-1492
8. Tan, Y., Wang, J., Nonlinear Blind Source Separation Using Higher order Statistics and a Genetic Algorithm. IEEE Transactions on Evolutionary Computation, vol. 5, num 6 (2001)
9. Darmois, G., Analyse Générale des Liaisons Stochastiques Rev. Inst. Internat. Stat 21, 2-8 (1953)
10. Nikias, C.L., Mendel, J.M., Signal Processing with Higher order Spectra IEEE Signal Processing Magazine pp 10–37 Jul (1993)
11. Smola, A.J., Schölkopf, B.: A tutorial on Support Vector Regression. NeuroCOLT2. Technical Report Series. NC2-TR-1998-030, October (1998)
12. High-order Contrasts for Independent Component Analysis. Jean-François Cardoso. Neural Computation, vol.11, no1, pp.157-192, Jan 1999