On information maximization and blind signal deconvolution

Axel Röbel

▶ **To cite this version:**

Axel Röbel. On information maximization and blind signal deconvolution. International Conference on Artificial Neural Networks (ICANN'99), Sep 1999, Edinborough, United Kingdom. 10.1049/cp:19991193 . hal-01253211

HAL Id: hal-01253211
https://hal.science/hal-01253211v1

Submitted on 8 Jan 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On Information Maximization and Blind Signal Deconvolution

Axel Röbel
Sekr. EN-8, Technical University of Berlin
Einsteinufer 17, 10587 Berlin, Germany
Tel: +49-30-314 25699, FAX: +49-30-314 21143,
email: roebel@kgw.tu-berlin.de

Abstract

In the following paper we investigate a recent algorithm for blind signal deconvolution and show that the algorithm is appropriate to solve the deconvolution problem only, if the deconvolution filter is constrained to be minimum phase. We improve the algorithm such that this constraint is removed and present experimental results that demonstrate the improved properties of the extended algorithm. Moreover, the experimental results show that the fixed signal density model implemented in the original algorithm has to be extended also, to achieve a proper objective function for the general deconvolution problem.

1 Introduction

Recently information theoretic formulation of blind signal separation and blind signal deconvolution criteria have received much interest [6; 8]. The goal of blind deconvolution [4] is to recover a source signal $x(n)$ given only the output $y(n)$ of an unknown filter with impulse response $\{a_k\}$. The problem is to find the inverse filter $\{b_k\}$ that yields

$$x(n) = \sum_{k=0}^{M} b_k y(n-k) \qquad (1)$$

given only $y(n)$. Because the knowledge of $y(n)$ is generally not sufficient to find the inverse filter we need to establish further constraints. In blind signal processing it is generally assumed that $x(n)$ is a white noise signal with non Gaussian density. Given this restriction the inverse filter has to remove all statistical dependencies across time that are introduced by the filter $\{a_k\}$. The infinitely many solutions of this problem differ only with respect to scaling and time shift. If we restrict the inverse filters to the class $\mathcal{B}$ of causal filters with $b_0 \neq 0$ and proper normalization the problem has a unique solution.

If $\{a_k\}$ and $\{b_k\}$ are restricted to be minimum phase with the normalization $b_0 = 1$, then the solution can be obtained by means of finding the filter $\{b_k\}$ that achieves the source signal $x(n)$ with minimum variance. This is the foundation of the well known and widely used linear prediction algorithm [7]. Without the restriction to minimum phase, however, there exist 2^M different filters $\{b_k\}$ with the same variance of the deconvolved signal. While one of these filters is the inverse of $\{a_k\}$, all the others include an additional all-pass component and, therefore, are indistinguishable by means of second order statistics.

It has been shown previously that many other objective functions may be used to find the inverse filter $\{b_k\}$, and that minimizing the entropy of $x(n)$

$$D(x) = -\int_x p(x) log(p(x)) dx, \qquad (2)$$

where $p(x)$ is the distribution of the samples of $x(n)$, yields asymptotically optimal results [2]. A deconvolution algorithm that properly minimizes the signal entropy as defined in eq. (2) is of special interest for data compression algorithms or source/filter signal models, which today use linear prediction to decorrelate the samples. Due to the restriction to minimum phase filtering and due to possible nonlinear dependencies in $x(n)$, however, the minimum variance objective of the linear prediction algorithm will generally fail to find the minimum entropy source signal, and, therefore, the results of the compression algorithms are suboptimal. The relation between the distribution $p(x)$ and the filter parameters $\{b_k\}$, however, is generally unknown, and, therefore, the use of the entropy as objective function has been rather crucial.

In a recent investigation on information theoretic objectives for blind signal processing it

has been shown that by means of a matrix formulation of the filter operation an approximate solution to the minimum entropy deconvolution can be obtained [1]. In that paper the inverse filter $\{b_k\}$ is assumed to be causal, which is no severe restriction. However, there exist two problems with this method. First, as we will argue later, the matrix expression of eq. (1) given in [1] is only suitable for minimum phase filters $\{b_k\}$ (which is considerably more restrictive then originally stated). Second, our experimental results demonstrate that the fixed density model used in [1] is not sufficient for deconvolution even for super Gaussian sources. In the following we will show, that the restriction to minimum phase filtering that is implicit in the algorithm can be relaxed with affordable increase in calculation complexity. To address the second problem we propose the use of an adaptive bimodal density model, which can be used to deconvolve super and sub Gaussian sources.

The following paper is organized as follows. In section 2 we shortly describe the blind deconvolution method introduced by Bell and Sejnowski. In section 3 we describe an alternative matrix formulation of the filtering process and argue that the methods differ only for non minimum phase problems, for which the new methods achieves correct results. In section 4 we describe our adaptive bimodal source distribution model. Section 5 shortly explains some experimental results we have obtained for white noise test signals and section 6 concludes with an outlook on further work.

2 Information maximization and minimum entropy

Bell and Sejnowski developed their deconvolution algorithm as an application of the minimum entropy blind signal separation algorithm they presented in the same paper [1]. In the following we give a short summary of the key idea of the algorithm, for detailed description see the original paper. Assume we are given an L-channel instantaneously mixed signal $\vec{y}(n)$ and are searching the original L source signals $x_i(n)$ that are assumed to be statistically independent. Formally, we are looking for the unmixing matrix B that achieves

$$\vec{x}(n) = B\vec{y}(n). \tag{3}$$

As Bell and Sejnowski has shown, the task can be addressed by maximizing the joint entropy of a nonlinearly transformed output signal $\vec{z}(n)$ with components

$$z_i(n) = f_i(x_i(n)),$$

with all f_i being constraint to be monotonically increasing with fixed range, i.e. $[-1, 1]$. Following [1] the joint entropy $D(\vec{z})$ can be approximately expressed as

$$\begin{aligned} D(\vec{z}) &= \log(|\det(B)|) \\ &+ \sum_{i=1}^{L} \frac{1}{N} \sum_{n=0}^{N-1} \log(\frac{\partial z_i(n)}{\partial x_i(n)}) + C, \end{aligned} \tag{4}$$

where N is the length of the respective signal vectors and C is constant and equal to the joint entropy $D(\vec{y})$. The approximation is due to the calculation of a sample mean instead of the entropy integral. This term yields the expectation of the logarithm of the derivative $f_i'(x)$. Due to the special structure of f_i this derivative has the properties of a density, and, therefore the expectation is maximized if $f_i'(x)$ equals the density of x. In this case and with the same approximation as in eq. (4) this second term equals the negative sum of the L entropies $D(x_i)$ and due to the basic relations between joint and scalar entropies we can rewrite eq. (4) as

$$D(\vec{z}) = \log(|\det(B)|) - D(\vec{x}) - T(\vec{x}) + C, \tag{5}$$

where $T(\vec{x})$ is the mutual information between the channels. From the basic laws of variable transformation it is known that the joint entropy $D(\vec{x})$ equals the sum of the joint entropy $D(\vec{y})$ $(= C)$, which is constant here, and a scaling term given by $\log(|\det(B)|)$. Therefore, we conclude that the first term in eq. (4) compensates any scaling that is produced by means of the linear transformation B. As long as the derivative of the nonlinearity f_i equals the density of the samples x_i we have

$$D(\vec{z}) = -T(\vec{x}), \tag{6}$$

and, therefore, under this constraint maximization of the joint entropy of $\vec{z}$ is equivalent to the minimization of the mutual information [8]. To simplify the algorithm Bell and Sejnowski proposed to use a fixed nonlinearity

$$f_i = \tanh(x_i)$$

which is equivalent to assume a fixed density model for the signals x_i. They conjecture that successful separation of super Gaussian sources is possible even if f_i is not equal to the source distribution. It has been shown that for zero mean signals the position of local maxima of eq. (4) is unchanged if f_i does not match the signal distribution [8]. However, the decrease in entropy due to the mismatch between the density $p(x_i)$ and $f_i'(x_i)$ depends on the matrix B, and our experimental results in section **5** show that there exist cases where the entropy decrease due to distribution mismatch is much larger for

the ideal unmixing matrix than for other matrices, and, therefore, the global optimum might become a local one.

To be able to apply the algorithm for blind deconvolution, Bell and Sejnowski formulate the deconvolution in eq. (1) by means of matrix multiplication between an $N \times N$ matrix B and an N-dimensional vector $\vec{y}(n) = (y(n), y(n+1), \ldots, y(n+N-1))^T$. To construct B they set the matrix elements on the k-th diagonal to b_k, where the main diagonal is identified with $k = 0$ and the diagonals are counted from right to left. As an example we construct the matrix B for a causal filter of order $M = 3$

$$B = \begin{pmatrix} b_0 & 0 & 0 & 0 & 0 & \cdots \\ b_1 & b_0 & 0 & 0 & 0 & \cdots \\ b_2 & b_1 & b_0 & 0 & 0 & \cdots \\ 0 & b_2 & b_1 & b_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}. \quad (7)$$

Multiplication of B with $\vec{y}$ from the right yields a vector representation of the output of the filter $x(n)$. Based on this matrix representation the blind separation algorithm summarized above may be applied. For causal filters B is lower triangular. Using the same nonlinearity f_i for all channels i and employing the additional assumption that channel and time averages are equivalent eq. (4) becomes simply

$$D(\vec{z}) = L \log(|b_0|) + \sum_{i=0}^{L-1} \log(\frac{\partial z(i)}{\partial x(i)}) + C. \quad (8)$$

Using this equation the gradient of $D(\vec{z})$ with respect to the filter parameters is easy to calculate and can be employed for an adaptive algorithm for blind signal deconvolution.

While Bell and Sejnowski has successfully applied their algorithm to a number of blind deconvolution tasks, there exists a weak point in the above argumentation that restricts the usage of the algorithm to the case of minimum phase filters $\{b_k\}$. The assumption of equal sample distributions for all channels i that leads to the simple form of eq. (8) is generally violated for the first $(M-1)$ channels. Given a sequence of vectors $\vec{y}(n)$ that are constructed from different segments of a signal $y\prime(n)$ of length $N' > N$ we find that the first $(M-1)$ channels always contain transients of the filter response and, therefore, obey different distributions. The impact of this deviation seems to be small for $M \ll L$, however, for non minimum phase filters the transient channels change the scaling behavior of the matrix B compared to the filter output such that an application of eq. (8) yields incorrect results.

[1] In the following denoted as CM.

As explained above the first term in the above entropy equations has to compensate for the increase in entropy that is due to scaling. While the first term in eq. (8) indicates that the scaling due to linear transformation eq. (7) depends solely on b_0, this is not true for the output $x(n)$ of a non minimum phase filter. However, from the above reasoning it is difficult to develop the correct scaling compensation that has to be applied in eq. (8).

3 Circular Filtering

The above formulation of the FIR-filtering as a matrix multiplication is not the only one possible. While the above formulation leads to a somewhat unclear relation between the filter operation and the matrix representation, we will now develop along a different line and will show that eq. (8) is correct only for minimum phase filters. Motivated by the FIR matrix algebra of Lambert [5] we consider the use of so called quadratic *circular matrices*[1] $\hat{B}$ of size L. In contrast to Lambert, however, we use a slightly different rule to construct the CM for a given FIR filter. As a consequence the analysis of the relation between matrix algebra and FFT FIR filter algebra is simplified, because the variable time shift that Lambert has to obey is fixed to zero.

To construct a circular matrix for a periodic sequence of length L we use as first row of the matrix the period of the sequence with time origin positioned at the first column. All following rows are built by circularly shifting the previous row to the right. The relation between a FIR filter of order $M < L$ and a CM is given by the filter response to a unit impulse train with period L. As an example we construct a CM $\hat{B}$ of size $L = 5$ for a FIR causal filter of order $M = 3$

$$\hat{B}(b_k) = \begin{pmatrix} b_0 & b_1 & b_2 & 0 & 0 \\ 0 & b_0 & b_1 & b_2 & 0 \\ 0 & 0 & b_0 & b_1 & b_2 \\ b_2 & 0 & 0 & b_0 & b_1 \\ b_1 & b_2 & 0 & 0 & b_0 \end{pmatrix}. \quad (9)$$

Using the circular matrices we are able to interpret all matrix operations required for blind signal separation in terms of operations on the periodic sequences used to construct the CM. The multiplication of two CM, $\hat{B}$ and $\hat{Y}$, that are constructed from sequences, $b(n)$ and $y(n)$, yields a CM, $\hat{X}$, that can equivalently be constructed from the result $x(n)$ of the L-point circular convolution of $y(n)$ and $b(n)$. Transposition of a CM is equivalent to reflecting the periodic sequence at time $n = 0$. Inversion of a CM

yields again a CM and the result is equivalently obtained by means of L-point discrete Fourier transformation (DFT) of the related sequence, invert the elements of the result, apply the inverse transformation and construct a CM from the result. The determinant of a CM is equal to the product of all elements of the L point DFT of the elements of the related signal.

With the circular matrices obtained from the filter coefficients $\{b_k\}$ and an N-periodic signal $y(n)$ we can express blind deconvolution for circular filtering without any error. Using the above relations between circular matrices and periodic sequences operations and neglecting the constant term C we formulate the joint entropy eq. (4) using the CM of size $L = N$ as follows

$$D(\vec{z}) = \sum_{i=0}^{L-1} \log(|H_b(i)|) + \log(\frac{\partial z(i)}{\partial x(i)}). \quad (10)$$

Here $H_b(i)$ is the L-point DFT of the filter impulse response $\{b_k\}$. Note that due to the symmetry of the CM in case of circular filtering and provided we use the same nonlinearity in all channels the assumption of equal channel distributions is correct. Therefore, for circular filtering the calculation of the mean in eq. (4) can be neglected without error. However, we are interested in non circular deconvolution, and, therefore, we have to apply some corrections to the above equation.

First we consider the scaling compensation. If we consider $y(n)$ to be of finite length we may apply the circular deconvolution to its L-periodic continuation. If we increase L to infinity then the results of circular and non circular convolution agree. For increasing L the scaling term in eq. (10) yields an increasingly dense sampling of the transfer function of the FIR filter, and, besides a factor L, the sum over $H_b(i)$ achieves an continually improved approximation of the integral of the log magnitude of the transfer function $H_b(jw)$

$$S_b = \frac{1}{2\pi} \int_w \log(|H_b(jw)|) dw. \quad (11)$$

It is well known that eq. (11) is related to the scaling properties of the filter $\{b_k\}$ [7, p. 130], and we conclude that eq. (11) is a normalized (with respect to block length L) measure of the appropriate scaling compensation for FIR filtering operation. To be able to apply eq. (10) to non circular deconvolution the mean of the first term in eq. (10) should accurately approximate eq. (11), and, therefore we shall choose L as large as possible. For large L and minimum phase filtering the scaling compensation obtained by the two formulars eq. (10) and eq. (8) agree. For non minimum phase filters, however, the scaling effect is under estimated in eq. (8) such that the entropy is systematically to small, and, therefore, we expect that non minimum phase solutions can not be found. Consider now the second term in eq. (10). From the previous section we know that this term approximates the negative channel entropy $D(x_i)$. As long as we achieve a sufficient sampling of $p(x)$ we may choose to sum over a sample subset of size K, and out weight the sub sampling by a additional factor L/K. Moreover, if we want to neglect the transients at the borders of the circular filtered $x(n)$ from the density adaptation, we may delete them from the summation with the same correction applied as above. Due to the possibility to use less than L samples to approximate the entropy $D(\vec{x_i})$ we are free to select L as large as we need to achieve sufficient accuracy for the approximation of eq. (11) by the first term in eq. (10).
Due to the algebraic relations stated above, the gradient of eq. (10) can be caluclated efficiently without any matrix operations.

4 Adaptive distribution model

With a fixed nonlinearity f_i the above algorithm is only valid if the distribution of the deconvolved signal $x(n)$ is close to the derivative of the nonlinearity. As will be demonstrated in the next section this is a severe problem even for super Gaussian distributions. Therefore, we propose to use an adaptive nonlinearity as follows

$$\begin{aligned} f(x) &= \frac{1}{2}(w \tanh(a_1 x + b_1) \\ &+ (1-w)\tanh(a_2 x + b_2)) \\ w &= \frac{1}{1+\exp(-w_h)}. \end{aligned} \quad (12)$$

Because the related density is bimodal it can be used to model sub and super Gaussian densities [3]. The weighting parameter w_h is transformed such that w is always in the interval $[0; 1]$ The nonlinearity is equivalent to a neural network with two hidden units, which can be adapted by gradient ascend of eq. (10) with respect to the network parameters. The model distribution consists of a mixture of two distributions of the form $\frac{a}{2\cosh(ax)^2}$. Using the Fourier Transform

$$\int_{-\infty}^{\infty} \frac{a}{2\cosh(ax)^2} e^{-jwx} dx = \frac{w\pi}{2a\sinh(\frac{w\pi}{2a})}$$

we have been able to calculate the moment generating function of this distribution and have

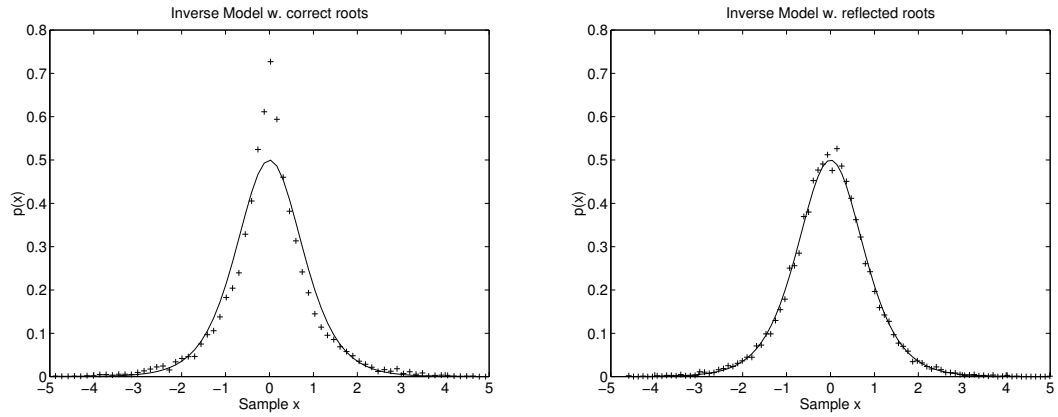

Figure 1: The sample histograms (marked $+$) of the deconvolved signals $p(x)$ in case of unknown filter $H_1(z)$ for the correct inverse model (left) and the inverse model with reflected roots (right) compared with the fixed model distribution $\frac{1}{\cosh^2(x)}$ (solid line). Due to distribution mismatch the incorrect inverse filter yields the global maximum of the joint entropy $D(\vec{z})$.

found that its variance is

$$\sigma^2 = \frac{\pi^2}{12a^2}.$$

This result is used to initialize the distribution parameters, such that the model distribution matches the variance of x(n) for the initial filter matrix $\hat{B}$. We initialize the model distribution as follows

$$\begin{aligned} 1.1a_1 &= 0.9a_2 = \frac{12\sigma_x}{pi} \\ b_1 &= -b2 = 0.001 \\ w_h &= 0.0 \end{aligned}$$

such that the model is slightly non symmetric, however, with a variance that is close to the variance σ_x^2 of the signal x obtained from the initial CM $\hat{B}$.

5 Experimental results

To verify our reasoning we have applied the above algorithms to two deconvolution problems, with $\{a_k\}$ being minimum phase in the first and maximum phase in the second experiment. Due to the fixed density model, we selected super Gaussian source signal $x(n)$ with exponential distribution and variance 1. For the (unknown) filter $\{a_k\}$ we use the IIR filter transfer functions

$$H_1(z) = \frac{1}{1 + 0.5z^{-1} + 0.2z^{-2}}$$
$$H_2(z) = \frac{1}{1 + 2z^{-1} + 1.5z^{-2}}.$$

We realize the maximum phase filter $H_2(z)$ using a non causal filter. The inverse filter $\{b_k\}$ is provided with five coefficients, while the ideal deconvolution filter needs only three. We initialized the filter coefficients randomly with normal distribution and variance 1 and adapted the filters in batch mode with an epoch size of 10000 using the gradient calculated from the entropy equations explained above.

As expected the Bell and Sejnowski algorithm always converges to a minimum phase solution. In case of the minimum phase filter the solution is close to the inverse of $\{a_k\}$, however, for the maximum phase problem the algorithm have found solutions with roots of the transfer function that are reflected at the unit circle. The circular matrix algorithm with fixed density model finds the same results if the initial random filter is minimum phase, because for minimum phase filters both algorithms agree in their entropy estimation. With initial filters that have at least two roots of the transfer function on the proper side of the unit circle our new algorithm converges to a filter with the correct deconvolution filter.

Note, that in all cases the global maximum of the joint entropy eq. (4) is not obtained for the correct inverse filter, but, for the filter with reflected roots. This is due to the fixed density model. While the all pass component that remains in the signal introduce slight statistical dependencies the entropy is maximal in this case because the density of the source (exponential distribution) is further apart from the fixed density model than the density of the all pass filtered signal (figure 1). We conclude that for general case blind signal deconvolution the nonlinearity f_i has to be adapted even in case of a super Gaussian signal. Otherwise the global

maximum of the joint entropy does not indicate proper deconvolution. Using the adaptive nonlinearity proposed in section **4** the signal density can be modeled more accurately and with this algorithm the global maximum of the objective function is reached for the correct deconvolution filter.

6 Outlook and summary

In the present paper we have criticized a recent blind deconvolution algorithm and have shown, that the algorithm fails to solve the deconvolution problem if the unknown filter is not minimum phase. Motivated by the work of Lambert, we have presented an extended algorithm that is appropriate for the general deconvolution problem. Moreover, the experimental results demonstrate that the fixed density model has to be extended to an adaptive bimodal distribution to be able to properly solve the deconvolution problem, even if the source distribution is super Gaussian. Forthcoming investigations will consider applications of the algorithm to data compression of audio signals. Due to the explicit minimization of the entropy of the signal significant improvements of the actual algorithms based on linear prediction are expected. Initial investigations leads to the conclusion that the optimal deconvolution filter for audio signals in many cases requires maximum phase filtering. The compression improvements that are achieved with the new method are currently investigated.

References

[1] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1004–1034, 1995.

[2] David L. Donoho. On mimimum entropy deconvolution. In D. F. Findley, editor, *Proceedings of the Second Applied Time Series Symposium, 1980*, pages 565–608, 1981.

[3] M. Girolami. An alternative perspective on adaptive independent component analysis algorithms. *Neural Computation*, 10(8):2103–2114, 1998.

[4] S. Haykin, editor. *Blind Deconvolution*. Prentice-Hall, New Jersey, 1994.

[5] R. H. Lambert. *Multichannel Blind Deconvolution: FIR Matrix Algebra and Separation of Multipath Mixtures*. PhD thesis, University of Southern California, Department of Electrical Engineering, 1996.

[6] T-W. Lee, M. Girolami, A. J. Bell, and T. J. Sejnowski. A unifying information-theoretic framework for independent component analysis. *International Journal on Mathematical and Computer Modeling*, 1998. In press.

[7] J. D. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer Verlag, 1976.

[8] H. Yang and S. Amari. Adaptive online learning algorithms for blind separation: Maximum entropy and minimum mutual information. *Neural Computation*, 9:1457–1482, 1997.