Journal of Machine Learning Research 4 (2003) 1319-1338 Submitted 10/02; Published 12/03

Blind Separation of Post-nonlinear Mixtures using Linearizing Transformations and Temporal Decorrelation

Andreas Ziehe ZIEHE@FIRST.FHG.DE
Motoaki Kawanabe NABE@FIRST.FHG.DE
Stefan Harmeling HARMELI@FIRST.FHG.DE
Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, Germany

Klaus-Robert Müller KLAUS@FIRST.FHG.DE
Fraunhofer FIRST.IDA
Kekuléstr. 7, 12489 Berlin, Germany
and
Department of Computer Science
University of Potsdam
August-Bebel-Strasse 89, 14482 Potsdam, Germany

Editors: Te-Won Lee, Jean-François Cardoso, Erkki Oja and Shun-Ichi Amari

Abstract

We propose two methods that reduce the post-nonlinear blind source separation problem (PNL-BSS) to a linear BSS problem. The first method is based on the concept of *maximal correlation*: we apply the alternating conditional expectation (ACE) algorithm—a powerful technique from non-parametric statistics—to approximately invert the componentwise nonlinear functions. The second method is a Gaussianizing transformation, which is motivated by the fact that linearly mixed signals before nonlinear transformation are approximately Gaussian distributed. This heuristic, but simple and efficient procedure works as good as the ACE method. Using the framework provided by ACE, convergence can be proven. The optimal transformations obtained by ACE coincide with the sought-after inverse functions of the nonlinearities. After equalizing the nonlinearities, temporal decorrelation separation (TDSEP) allows us to recover the source signals. Numerical simulations testing "ACE-TD" and "Gauss-TD" on realistic examples are performed with excellent results.

Keywords: Blind Source Separation, Post-nonlinear Mixture, Gaussianization, Maximal Correlation, Temporal Decorrelation

1. Introduction

Blind methods for source separation had major success in many areas of science. This is mainly due to the simple, yet versatile model (Jutten and Hérault, 1991, Comon, 1994, Cardoso and Souloumiac, 1993, Deco and Obradovic, 1995, Amari et al., 1996, Bell and Sejnowski, 1995, Hyvärinen and Oja, 1997, Ziehe and Müller, 1998, Pajunen and Karhunen, 2000, Hyvärinen et al., 2001). For linear BSS we assume that at time t each component $x_i[t]$ of the observed n-dimensional data vector $x[t]$ is a linear combination of $m \leq n$ statistically independent signals, mathematically speaking

$$x_i[t] = \sum_{j=1}^{m} A_{ij} s_j[t].$$

Since the source signals $s_j[t]$ and the coefficients A_{ij} of the mixing matrix $\mathbf{A}$ are unknown, they have to be estimated from the observed signals $x[t]$. The goal is to find a separating matrix $\mathbf{B}$ and signals $u[t] = \mathbf{B}x[t]$ which are estimates for $s[t]$.

A challenging research topic is to extend this model to nonlinear transformations. In general, the nonlinear mixing model is of the form

$$x[t] = f(s[t]) \tag{1}$$

where f is an arbitrary nonlinear transformation which must be at least approximately invertible. An important special case is the so-called post-nonlinear (PNL) mixture,

$$x[t] = f(\mathbf{A}s[t]),$$

where f is an invertible nonlinear function that operates *componentwise* and $\mathbf{A}$ is a linear mixing matrix, more detailed:

$$x_i[t] = f_i\left(\sum_{j=1}^{m} A_{ij} s_j[t]\right), \qquad i = 1, \ldots, n. \tag{2}$$

The PNL model was introduced by Taleb and Jutten (1997). It represents an important subclass of the general nonlinear model and has therefore attracted the interest of several researchers (e.g. Lee et al., 1997, Yang et al., 1998, Taleb and Jutten, 1999, Ziehe et al., 2001, Achard et al., 2001, Taleb, 2002, Jutten and Karhunen, 2003). Applications are found, for example, in the fields of telecommunications, where power efficient wireless communication devices with nonlinear class C amplifiers are used (Larson, 1998) or in the field of biomedical data recording, where sensors can have nonlinear characteristics (Ziehe et al., 2000).

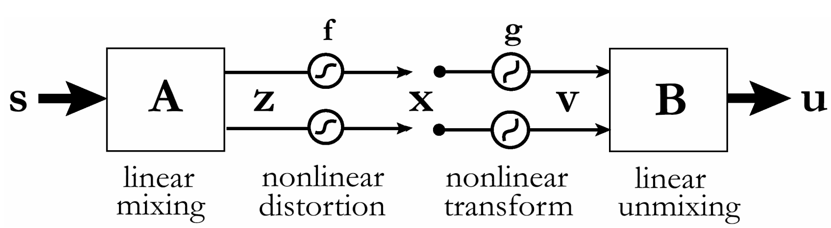


Figure 1: Building blocks of the PNL mixing model and the separating system.

Many algorithmic approaches to the nonlinear BSS problem have been proposed so far, including diverse methods such as self-organizing maps (Pajunen et al., 1996, Lin et al., 1997), extensions of generative topographic mapping (GTM, Pajunen and Karhunen, 1997), neural networks (Yang et al., 1998, Marques and Almeida, 1999), parametric sigmoidal functions (Lee et al., 1997) or Bayesian ensemble learning (Valpola et al., 2000, 2003). Recently, a very efficient temporal decorrelation method in approximated kernel feature spaces (see e.g. Schölkopf et al., 1998, Müller et al., 2001, Schölkopf and Smola, 2002, Vapnik, 1998) has been presented by Harmeling et al. (2002, 2003). However, these methods—except the method of Harmeling et al. (2003)—are often computationally very demanding and may face convergence problems in higher dimensions. In contrast, algorithms tailored to the PNL problem (Taleb and Jutten, 1999, Ziehe et al., 2001, Achard et al., 2001) are usually much simpler than those for the general problem, see Equation (1), since they only need to invert one-dimensional nonlinear functions, see Equation (2).

In this contribution, we propose a decoupled two-stage process to solve the PNL-BSS problem: first, we estimate the inverse functions g_i of the component-wise non-linearities f_i using the ACE algorithm (as described in Ziehe et al., 2001) or alternatively by applying a Gaussianizing transformation (as in Ziehe et al., 2003a). After having equalized the nonlinearities, we can apply in the second step standard linear BSS techniques to recover the underlying source signals (Figure 1).

Since the two stages—equalizing and unmixing—are independent from each other, the method can be easily generalized to handle also convolutive mixtures with PNL distortions, that is, mixed signals that are modeled as

$$x[t] = f\left(\sum_{\tau} \mathbf{A}[\tau] s[t-\tau]\right),$$

where $\mathbf{A}[\tau]$ is an $n \times m$ filter matrix for each τ. The sources $s[t]$ can be simply recovered by replacing the BSS algorithm for instantaneous mixture with a method for convolutive unmixing (e.g. Parra and Spence, 2000, Murata et al., 2001).

The paper is organized as follows: in Subsection 2.1, we briefly introduce the ACE framework and prove that the algorithm converges to the correct inverse nonlinearities—provided that they exist. After that we propose an alternative linearization procedure employing Gaussianizing transformations in Subsection 2.2. In Section 3 we discuss some implementation issues and describe numerical simulations that illustrate our method. Finally, we give a conclusion in Section 4.

2. Methods

For the sake of simplicity, we introduce our method for the 2×2 case. The extension to the general case is briefly explained in the appendix. Let us consider the two-dimensional post-nonlinear mixing model:

$$\begin{aligned} x_1 &= f_1(A_{11}s_1 + A_{12}s_2) \\ x_2 &= f_2(A_{21}s_1 + A_{22}s_2) \end{aligned}$$

where s_1 and s_2 are independent source signals with time-structure, x_1 and x_2 are the observed signals, $\mathbf{A} = (A_{ij})$ is the mixing matrix and f_1 and f_2 are the componentwise nonlinear transformations which are invertible.[1]

2.1 Maximizing Correlations with the ACE Algorithm

In order to invert the nonlinear functions f_1 and f_2, we propose to maximize the correlation

$$\mathrm{corr}(g_1(x_1), g_2(x_2)) \tag{3}$$

with respect to some nonlinear functions g_1 and g_2. That means, we want to find transformations g_1 and g_2 of the observed signals such that the relationship between the transformed variables becomes linear. Intuitively speaking, the nonlinear transformation bends the data in a scatter plot away from a

1. Note that a one-dimensional function is invertible if and only if it is monotonic.

straight line. Thus we have to find a transformation that nonlinearly squeezes and stretches the data in the scatter plot until the transformed signals have *maximal correlation* in order to undo the effect of the distortion (see Figure 4 for an illustration of this process). Under certain conditions that we will state in detail later, this problem is solved by the ACE method that finds so-called optimal transformations g_1^* and g_2^* which maximize Equation (3). Breiman and Friedman (1985) showed existence and uniqueness of those optimal transformations and proved that the ACE algorithm, which is described in the following, converges to these solutions (see Figure 4).

The ACE algorithm is an iterative procedure for finding the optimal nonlinear functions g_1^* and g_2^*. Starting from the observation that for fixed g_1 the optimal g_2 is given by

$$g_2(x_2) = E\{g_1(x_1)|x_2\},$$

and conversely, for fixed g_2 the optimal g_1 is

$$g_1(x_1) = E\{g_2(x_2)|x_1\},$$

we are ready to understand the key idea of the ACE algorithm that is to compute these conditional expectations alternately until convergence has been reached. To avoid trivial solutions $g_1(x_1)$ is normalized in each step by using the function norm $||\cdot|| := \left(E\{\cdot\}^2\right)^{1/2}$. The algorithm for two variables is summarized below. It is also possible to extend the procedure to the multivariate case, which is explained in Appendix A (for further details see Hastie and Tibshirani, 1990, Breiman and Friedman, 1985).

Algorithm 1 The ACE algorithm for two variables

{initialize}
$g_1^{(0)}(x_1) \leftarrow x_1/||x_1||$
repeat
$\quad g_2^{(k+1)}(x_2) \leftarrow E\{g_1^{(k)}(x_1) \mid x_2\}$
$\quad \tilde{g}_1^{(k+1)}(x_1) \leftarrow E\{g_2^{(k+1)}(x_2) \mid x_1\}$
$\quad g_1^{(k+1)}(x_1) \leftarrow \tilde{g}_1^{(k+1)}(x_1)/||\tilde{g}_1^{(k+1)}(x_1)||$
until $E\{g_1(x_1) - g_2(x_2)\}^2$ fails to decrease

The ACE algorithm was originally introduced to solve a special kind of regression problem, where one wants to find nonlinear transformations $g_1(\cdot)$ and $g_2(\cdot)$ such that $g_2(x_2)$ explains $g_1(x_1)$ optimally. The optimality is defined by the least squares criterion, that is,

$$\min_{g_1,g_2} \frac{E\{g_1(x_1) - g_2(x_2)\}^2}{Eg_1^2(x_1)},$$

where the denominator prevents scale indeterminacy. Interestingly, the optimal transformation in the least squares sense is equivalent to the transformation with maximal correlation,

$$\max_{g_1,g_2} \operatorname{corr}(\,g_1(x_1),\, g_2(x_2)\,),$$

up to scale factors. This equivalence, which was proved by Breiman and Friedman (1985), also reveals how to extend the ACE algorithm to the multivariate case (Appendix A).

An important point in the implementation of the ACE algorithm is the estimation of the conditional expectations from the data. Usually, conditional expectations are computed by data smoothing for which numerous techniques exist (Breiman and Friedman, 1985, Härdle, 1990). Our implementation uses a nearest neighbor smoothing that applies a simple moving average filter to appropriately sorted data. Here the size of the moving average filter is a parameter to balance the trade-off between over- and under-smoothing.

In the following we want to verify that applying g_1^* and g_2^* to the mixed signals x_1 and x_2, removes the effect of the nonlinear functions f_1 and f_2. To this end we show for $z_1 = A_{11}s_1 + A_{12}s_2$ and $z_2 = A_{21}s_1 + A_{22}s_2$ that g_1^* and g_2^* obtained from the ACE procedure are the desired inverse functions given that z_1 and z_2 are jointly normally distributed, with other words we prove the following relationship:

$$\begin{aligned} h_1^*(z_1) &:= g_1^*(f_1(z_1)) \propto z_1 \\ h_2^*(z_2) &:= g_2^*(f_2(z_2)) \propto z_2. \end{aligned} \tag{4}$$

The proof proceeds exactly along the lines of Breiman and Friedman (1985). Noticing that the correlation of two signals does not change, if we scale one or both signals and following *Proposition 5.4.* and *Theorem 5.3.* of Breiman and Friedman (1985), we know that:

$$\begin{aligned} g_1^*(x_1) &\propto E\{g_2^*(x_2)|x_1\} \\ g_2^*(x_2) &\propto E\{g_1^*(x_1)|x_2\}. \end{aligned}$$

Here the conditional expectation $E\{g_2^*(x_2)|x_1\}$ is a function of x_1 and the expectation is taken with respect to x_2, analogously for the second expression.

Since $g_1^*(x_1) = h_1^*(z_1)$ and $g_2^*(x_2) = h_2^*(z_2)$, furthermore $x_1 = f_1(z_1)$ and $x_2 = f_2(z_2)$ we get:

$$\begin{aligned} h_1^*(z_1) &\propto E\{h_2^*(z_2)|f_1(z_1)\} \\ h_2^*(z_2) &\propto E\{h_1^*(z_1)|f_2(z_2)\}. \end{aligned}$$

Because f_1 and f_2 are invertible functions they can be omitted in the condition of the conditional expectation, leading us to:

$$\begin{aligned} h_1^*(z_1) &\propto E\{h_2^*(z_2)|z_1\} \\ h_2^*(z_2) &\propto E\{h_1^*(z_1)|z_2\}. \end{aligned} \tag{5}$$

Assuming that the vector $(z_1, z_2)^\top$ is normally distributed and the correlation coefficient $\mathrm{corr}(z_1, z_2)$ is non-zero, a straightforward calculation shows

$$\begin{aligned} E\{z_2|z_1\} &\propto z_1 \\ E\{z_1|z_2\} &\propto z_2. \end{aligned}$$

This means that z_1 and z_2 satisfy Equation (5), which then immediately implies our claim Equation (4).

It is favorable to our intended application that the above assumptions are usually fulfilled because mixtures of independent signals have a distribution that is closer to a Gaussian according to the central limit theorem and that they are more correlated than unmixed signals. In practice, we see that the ACE algorithm is able to equalize the nonlinearities rather well, even if the assumptions are not perfectly met, though there might be difficulties in the presence of strong noise or low correlations.

In summary, the essential idea of applying the ACE framework to the PNL problem uses the fact that we can approximate the inverses of the post-nonlinearities by searching for those nonlinear transformations which *maximize* the *linear* correlations between the *non-linearly* transformed observed variables.

2.2 Transformations to Gaussian Random Variables

In the previous section we have argued that linearly mixed signals are most likely subject to Gaussian distributions due to the central limit theorem. This fact gives rise to a simple alternative for the ACE procedure in the first stage of our PNL-BSS method by making use of this Gaussianity assumption directly.

We propose to estimate the inverse functions of the true nonlinearities by finding transformations which convert each component into a Gaussian random variable (Gaussianization). Our Gaussianization technique is inspired by algorithms commonly used in random number surrogate generators and adaptive bin size methods (Pompe, 1993, Kantz and Schreiber, 1997). The goal is to find functions $g_i(\cdot)$ such that

$$g_i(x_i) \sim N(0, \sigma_i^2).$$

where σ_i^2 is set to a constant (e.g. $\sigma_i^2 = 1$) which can be done without loss of generality because scaling factors do not matter. The method, which is based on the ranks of samples and percentiles of the Gaussian distribution, proceeds as follows: suppose that the observed samples $x[1], \ldots, x[T]$ fulfill an ergodicity assumption and T is sufficiently large. Let $r[t]$ be the rank of the t-th sample $x[t]$ and $\Phi(\cdot)$ be the cumulative distribution function of the standard Gaussian $N(0,1)$. The ranks r are assumed to be ordered from the smallest value to the largest, e.g. $r[t] = 1$ for $t = \operatorname{argmin}_t x[t]$ and $r[t] = T$ for $t = \operatorname{argmax}_t x[t]$. Then, the transformed samples

$$v[t] = \Phi^{-1}\left(\frac{r[t]}{T+1}\right), \qquad t = 1, \ldots, T$$

that is, the real values $v[t]$ such that $\Phi(v[t]) = \frac{r[t]}{T+1}$ can be regarded as realizations from the standard Gaussian distribution (c.f. Billingsley, 1995). Note that, adding 1 to the denominator is needed to avoid $v = \pm\infty$. Furthermore, to limit the influence of outliers, we recommend to compute modified versions of $v[t]$ according to

$$v[t] = \Phi^{-1}\left\{c\,\frac{r[t]}{T+1} + (1-c)\,\frac{1}{2}\right\}, \qquad \text{with } 0 < c < 1.$$

The choice of c depends on the application, but is not critical. We used $c = 0.97$ in our numerical experiments.

2.3 Source Separation

For a separation of the signals we could in principle apply any BSS technique that can solve linear problems. However, experiments show that second-order methods which use temporal information are more robust than higher-order based methods and allow to recover the sources more reliably in most natural applications.

Therefore we propose to use second-order BSS techniques that are based on the different time structure of the source signals. These methods (see Molgedey and Schuster, 1994, Belouchrani et al., 1997, Kawamoto et al., 1997, Pham and Garrat, 1997, Ziehe and Müller, 1998, Yeredor, 2000) rely on the special structure of time-shifted covariance matrices of the mapped signals $v[t] = g(x[t])$. The key insight is that the pairwise cross-correlation functions of the components of the original signal s are zero for certain time lags τ, that is,

$$E\left\{\, s_i[t]\, s_j[t+\tau] \,\right\} = 0, \qquad i \neq j.$$

Let $u[t] = \mathbf{B}v[t]$ be an estimator of the original signal $s[t]$ and

$$\mathbf{R}(\tau;\mathbf{B}) := \frac{1}{2(T-\tau)} \sum_{t=1}^{T-\tau} \left\{ u[t]\, u^\top[t+\tau] + u[t+\tau]\, u^\top[t] \right\},$$

be the symmetrized lagged covariance matrix of the estimator u, provided that u is already centered at $\mathbf{0}$. When we derive an estimator u close to the signal s, every lagged covariance matrix $\mathbf{R}(\tau;\mathbf{B})$ becomes approximately diagonal. Therefore, the task is to identify the demixing matrix $\mathbf{B}$ such that the covariance matrices $\mathbf{R}(0;\mathbf{B})$ and $\mathbf{R}(\tau_k;\mathbf{B}),\ k = 1,\ldots,K$ are diagonal. This can be done by a joint diagonalization method, that is, to minimize the following criterion w.r.t. $\mathbf{B}$ under certain restrictions

$$L(\mathbf{B}) = \sum_{i\neq j} R_{ij}^2(0;\mathbf{B}) + \sum_{k=1}^{K}\sum_{i\neq j} R_{ij}^2(\tau_k;\mathbf{B}),$$

where R_{ij}'s are the (i,j)-th components of the matrix $\mathbf{R}$.

In the literature several methods to solve this constrained optimization problem exist (Cardoso and Souloumiac, 1996, Yeredor, 2000, Pham, 2001, Ziehe et al., 2003b). Here we use TDSEP, an implementation based on the efficient extended Jacobi method by Cardoso and Souloumiac (1996) for the simultaneous diagonalization of several time-delayed correlation matrices (Belouchrani et al., 1997, Ziehe and Müller, 1998, Müller et al., 1999).

3. Numerical Simulations

¿From now on we call the procedure based on ACE "ACE-TD" and analogously, the procedure based on Gaussianization "Gauss-TD". To demonstrate the performance of the proposed approaches we apply these two methods to several post-nonlinear mixtures, both instantaneous and convolutive.

3.1 Toy Data: AR Processes

The first data set consists of Gaussian AR-processes of the form:

$$s_i[t] = \sum_{m=1}^{M} \alpha_m s_i[t-m] + \xi_i[t], \qquad i = 1,\ldots,n$$

where $\xi_i[t]$ is white Gaussian noise with mean zero and variance σ^2. For the experiment we choose $\sigma^2 = 1$, $M = 3$, $n = 2$ and generate 2000 data points. To the linearly mixed signals z we apply strong nonlinear distortions as in Taleb and Jutten (1999):

$$\begin{aligned} x_1[t] &= f_1(z_1[t]) = z_1^3[t], \\ x_2[t] &= f_2(z_2[t]) = \tanh(10 z_2[t]). \end{aligned} \tag{6}$$

The distribution of these mixed signals has a highly nonlinear structure which is clearly visible in the scatter plot in Figure 2.

The application of the linearization procedures– the ACE method or the Gaussianizing transformation– yields the estimated nonlinear functions g_1 and g_2 shown in Figure 3 (b) and (c). We see that the true inverses of the nonlinearities f_1 and f_2 are approximated well by the ACE algorithm and almost perfectly estimated by Gaussianization. In the ACE algorithm a local nearest neighbor

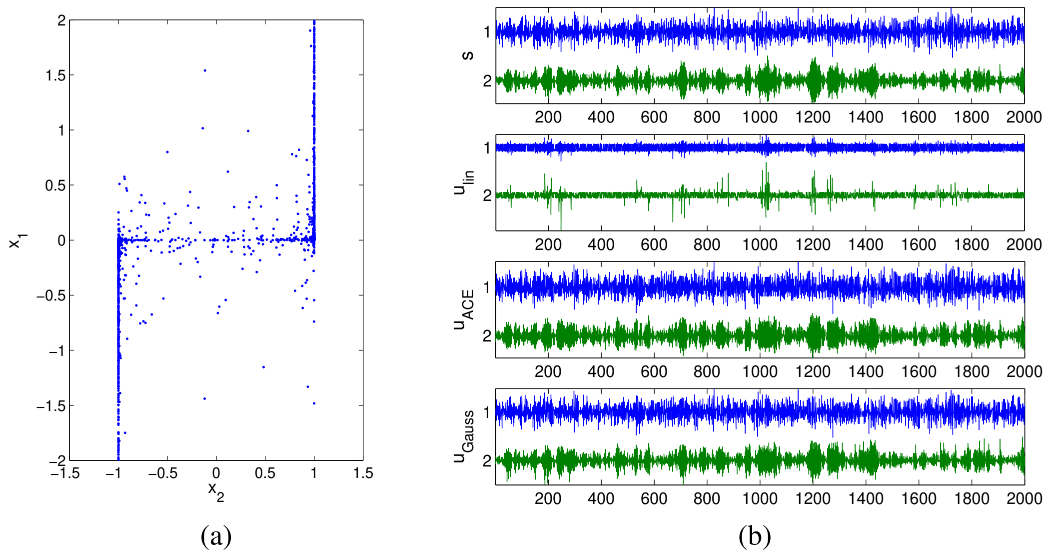


Figure 2: (a) Scatter plot of the mixed AR-processes ($x_1[t]$ vs $x_2[t]$) and (b) waveforms of the original sources (first row), the linearly unmixed signals (second), recovered sources by ACE-TD (third) and those by Gauss-TD (fourth). The smoother for ACE has length 101. For TDSEP we used time lags $\tau = 0 \ldots 20$.

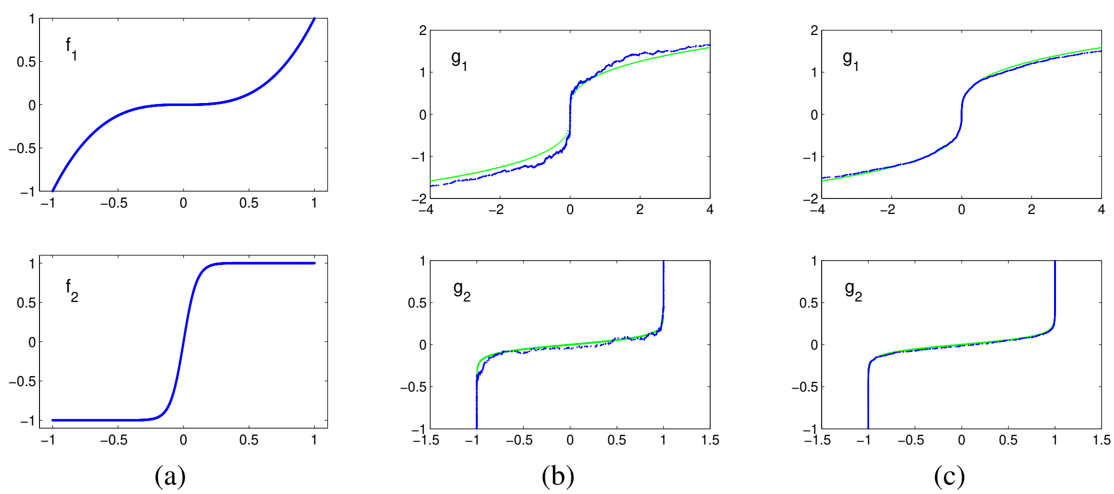


Figure 3: (a) Nonlinear functions f_1 and f_2. (b) True (thin line) and estimated (bold line) inverse functions g_1 and g_2 by the ACE method. (c) The same plots by the Gaussianizing transformation.

smoother (window length 101) has been used to compute the conditional expectation. We note that the remaining mismatch observed in Figure 3 (b) could be improved further by better smoothers.

After successful linearization, it becomes possible to separate the signals. We use the TDSEP algorithm (see Ziehe and Müller, 1998), where 20 time-delayed correlation matrices are simultaneously diagonalized (time lags $\tau = 0 \ldots 20$).

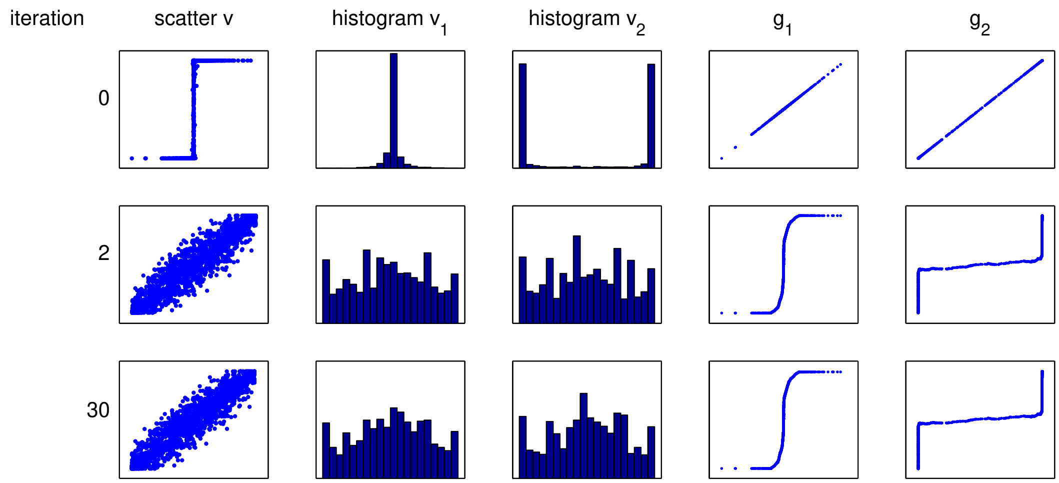


Figure 4: Monitoring of the iterative ACE procedure. The rows correspond to snapshots of the iteration at initialization and after two resp. thirty steps. The columns show a scatter plot of transformed data v (first column), histograms of v_1 (second) and v_2 (third), estimated inverse functions g_1 (fourth) and g_2 (fifth).

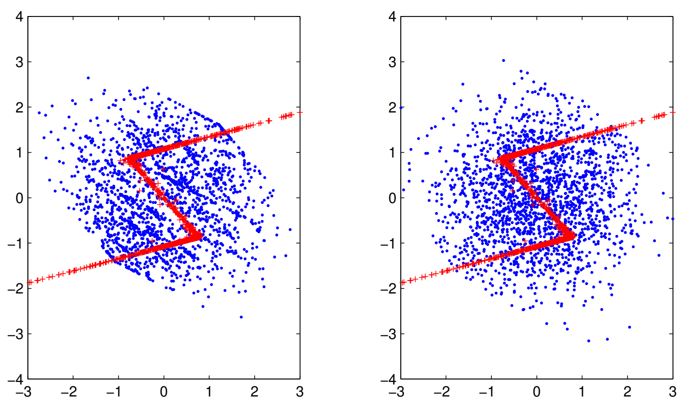


Figure 5: Left panel: scatter plot of the output distribution of a linear ('+') and the proposed nonlinear ACE-TD algorithm ('.'). Right panel: similar plot with the nonlinear Gauss-TD algorithm ('.').

Figure 2 (b) shows that the waveforms of the recovered sources by ACE-TD (the third row) and by Gauss-TD (the fourth row) closely resemble the original ones, while the result of the linear unmixing of the PNL mixture (the second row) can not recover the sources as is expected. This is also confirmed by comparing the output distributions that are shown in Figure 5. The left panel shows the scatter plots of the distributions obtained by TDSEP and ACE-TD, respectively. In the right panel we show the corresponding scatter plots for TDSEP and Gauss-TD, which indeed looks much more Gaussian.

3.2 Instantaneously Mixed Speech Data

Our method scales well both with increasing number of sources and with increasing number of data points. To demonstrate this, we apply the algorithm to mixtures of natural audio sources, where the input data set consists of 4 sound signals with 20,000 data points each.

The source signals are linearly mixed by a 4×4 random matrix $\mathbf{A}$, i.e. $z[t] = \mathbf{A}s[t]$. After the linear mixing the following nonlinearities were applied:

$$\begin{aligned} f_1(z_1) &= 0.1z_1 + 0.1z_1^3 \\ f_2(z_2) &= 0.3z_2 + \tanh(3z_2) \\ f_3(z_3) &= \tanh(2z_3) \\ f_4(z_4) &= z_4^3. \end{aligned}$$

In order to apply the ACE-TD method to this dataset, we have to employ the multivariate version of the ACE algorithm which is explained in Appendix A. As in the two-variables case, the extended version of ACE guarantees the identifiability of the inverses $g_1(\cdot), \ldots, g_n(\cdot)$ under the Gaussianity assumption, although there is an asymmetry in the treatment of these functions (for details see Appendix A). On the other hand, the extension of Gauss-TD to the multivariate case is straightforward.

Figure 6 shows the results of the separation using ACE-TD (with smoothing window length 101 and time delays $\tau = 0 \ldots 30$) and Gauss-TD (with time delays $\tau = 0 \ldots 30$) in the panels (c) and (d).

We observe again very good separation performances that are quantified by calculating the correlation coefficients between the source signals and the extracted components (see Table 1). This is also confirmed by listening to the separated audio signals, where we perceive almost no crosstalk, although the noise level is slightly increased (c.f. the silent parts of signal 2 in Figure 6 (a), (c) and (d)).

	TDSEP			
	u_1	u_2	u_3	u_4
s_1	0.89	0.03	-0.02	-0.06
s_2	0.12	-0.05	0.11	0.45
s_3	0.04	0.03	0.77	0.09
s_4	0.18	-0.05	0.06	0.69

	ACE-TD			
	u_1	u_2	u_3	u_4
s_1	0.99	0.00	-0.02	-0.03
s_2	-0.01	0.95	-0.02	0.03
s_3	0.01	0.03	0.99	0.02
s_4	0.05	0.01	0.00	0.95

	Gauss-TD			
	u_1	u_2	u_3	u_4
s_1	0.99	-0.00	-0.02	-0.02
s_2	0.00	0.96	-0.03	0.02
s_3	0.00	0.03	0.99	0.02
s_4	0.04	0.01	0.00	0.95

Table 1: Correlation coefficients for the signals shown in Figure 6 .

In order to compare these methods and to collect statistics about the behavior of the proposed algorithms, we replicated the experiment with 100 different mixing matrices leaving all other parameters unchanged. The results are summarized in Table 2, where we can see that ACE-TD found signals very similar to the true sources in more than 70% of the trials, while Gauss-TD found them in most cases. Thus, Gauss-TD is also preferable on this data set. Changing the window length of the smoother in ACE-TD to a smaller value (=17) yields worse results, because the estimators of the nonlinear functions are then under-smoothed. This shows that the choice of smoother is important for the ACE algorithm and it might be possible to further improve ACE-TD by optimizing the

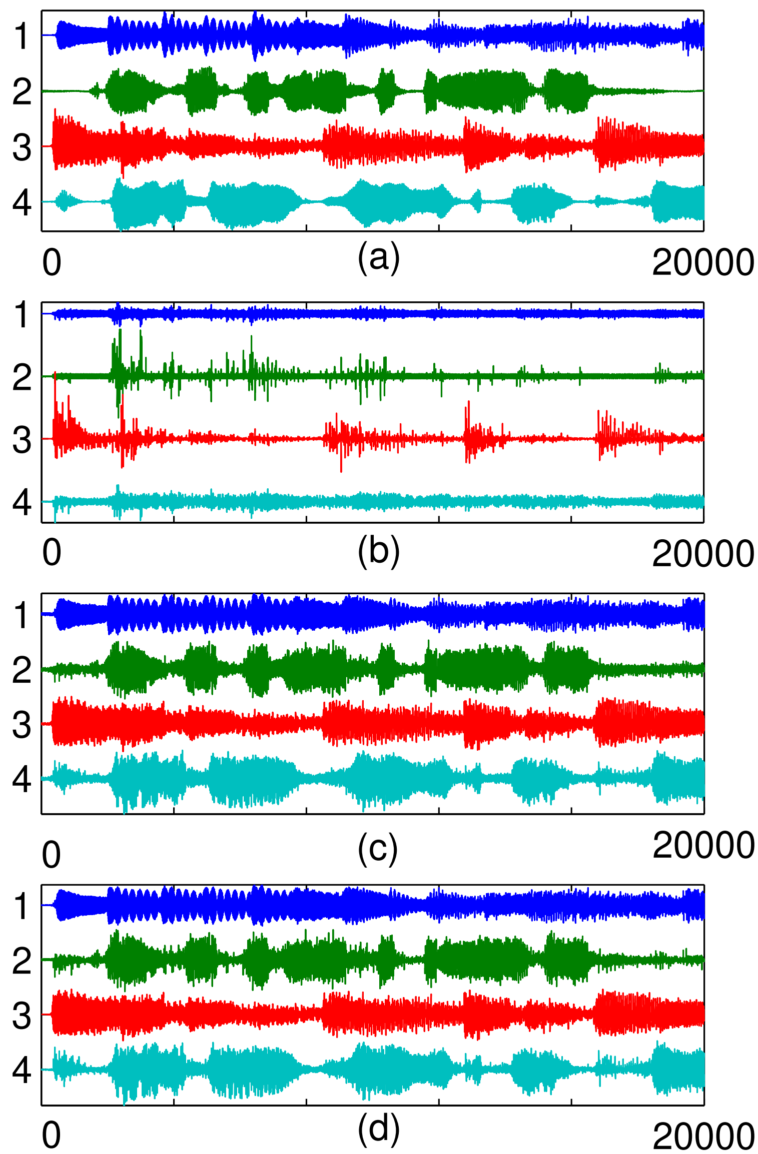


Figure 6: Four channel audio dataset: (a) waveforms of the original sources, (b) linearly unmixed signals with TDSEP, (c) recovered sources using ACE-TD and (d) recovered sources using Gauss-TD. A smoother with window length 101 was used in ACE and time lags $\tau = 0 \ldots 30$ were employed for the separation. The audio files corresponding to these experiments are also available on the web page `http://www.first.fhg.de/~ziehe/research.html`.

algorithms	statistics	s_1	s_2	s_3	s_4
TDSEP	median	0.7285	0.6341	0.6710	0.6448
	#{corr > 0.9}	3	1	1	1
ACE-TD	median	0.9585	0.9480	0.9527	0.9533
(window length 101)	#{corr > 0.9}	77	66	73	75
ACE-TD	median	0.9084	0.8918	0.8946	0.9125
(window length 17)	#{corr > 0.9}	56	45	45	57
Gauss-TD	median	0.9831	0.9800	0.9790	0.9748
	#{corr > 0.9}	94	93	89	86

Table 2: Maximum correlation between sources and recovered signals. 'median' denotes the median of the 100 trials and '#{corr > 0.9}' is the number of trials which gave correlations larger than 0.9. In ACE, besides the smoother with window length 101, we also used a shorter window length (= 17), which however resulted in under-smoothed nonlinearity estimators. In the separation with TDSEP time lags $\tau = 0 \ldots 30$ have been used.

window length (Härdle, 1990); however, the detailed analysis of this issue exceeds the scope of this contribution.

3.3 Speech Data: Convolutive Mixtures

The third experiment gives an example of applying our method to convolutive mixtures with a PNL distortion. With nonlinear transfer functions as in our first example (c.f. Equation (6)), we deliberately distorted real-room recordings[2] of speech and background music (from Lee et al., 1998). For the separation we apply a convolutive BSS algorithm of Parra and Spence (2000) that requires only second-order statistics and exploits the non-stationarity of the signals. While an unmixing of the distorted recordings obviously fails, we could achieve a good separation after the unsupervised linearization by the ACE procedure or the Gaussianization (c.f. Figure 7).

3.4 Comparison to Taleb and Jutten (1999)

The standard method for PNL-BSS (due to Taleb and Jutten, 1999), abbr. TJ, estimates simultaneously the linear and the nonlinear part of the mixing model. This is done by minimizing a contrast function based on mutual information. For this the densities have to be estimated which is a difficult problem on its own. In contrast, Gauss-TD decouples the linear and the nonlinear stages and solves each one separately by simple methods (Gaussianization and TDSEP).

To keep the computational efforts for TJ within a reasonable range, we use a dataset with two channels and 500 samples for the simulation. The sources are random uniform noise and a sinusoidal signal as shown in the upper panel of Figure 9. The sources were mixed with a 2×2 mixing matrix and componentwise distorted using the nonlinear transfer functions:

$$\begin{aligned} x_1[t] &= f_1(z_1[t]) = z_1^3[t], \\ x_2[t] &= f_2(z_2[t]) = \tanh(2z_2[t]). \end{aligned}$$

2. Available on the internet: `http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html`.

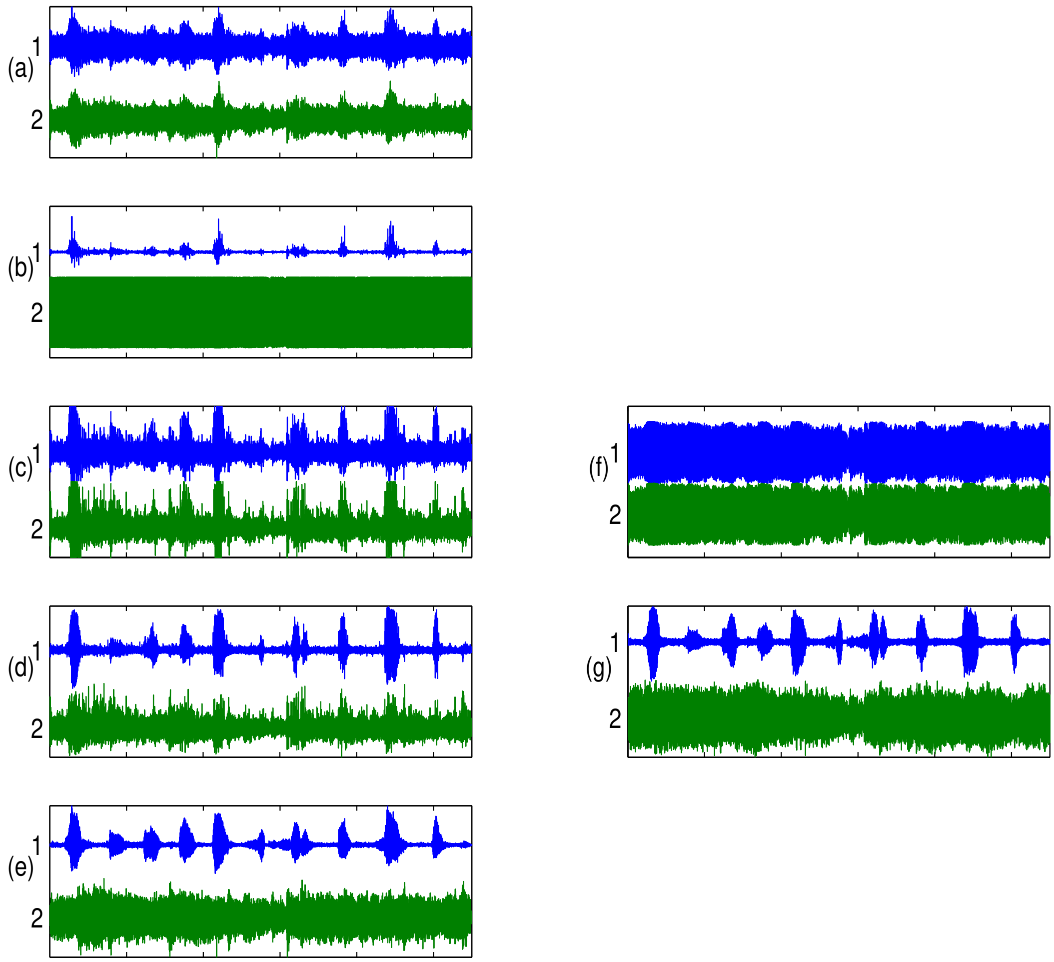
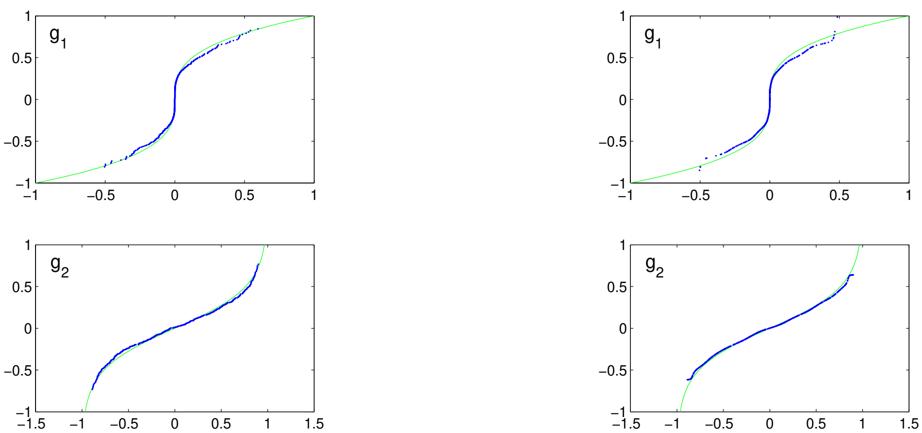


Figure 7: Two channel audio dataset: (a) waveforms of the recorded (undistorted) microphone signals, (b) observed PNL distorted signals, (c) result of ACE (smoothing window length 101), (d) recovered sources using ACE and a convolutive BSS algorithm and (e) for comparison convolutive BSS separation result for undistorted signals from (a). The right panels: (f) result of Gaussianizing transformation and (g) recovered sources using Gaussianization and a convolutive BSS algorithm. For listening to these audio data we refer to `http://www.first.fhg.de/˜ziehe/research.html`.

This data is shown in the second row of Figure 9. A linear separation of these signals obviously fails as can be seen in the third row of Figure 9. Only after the non-linear distortions have been equalized by the estimated inverse functions (shown in Figure 8), blind separation is successful and yields signals that closely resemble the underlying sources. The last row of Figure 9 shows the results of Taleb and Jutten's method. We observe a good separation except for a few points at the edges of the scatter plot. These problems are possibly due to the fixed bin size in the kernel density estimators used by this method. This emphasizes again the need for better smoothers and adaptive bin size methods. However, since Gauss-TD results are comparable (c.f. 4th and 5th row of Figure 9), Gauss-TD is often due to its simplicity the method of choice.



(a) (b)

Figure 8: (a) True (thin line) and estimated (bold line) inverse functions g_1 and g_2 obtained by the Gaussianization method. (b) The same plots for the Taleb-Jutten (TJ) algorithm.

4. Discussion and Conclusion

In this work we proposed a simple technique for blind separation of linear mixtures with post-nonlinear distortions based on ACE resp. Gaussianization with a subsequent application of BSS: ACE-TD and Gauss-TD. The algorithms proceed in two steps: first, ACE-TD searches nonlinear transformations that *maximize* the linear correlations between the transformed variables. Hereby it approximates the inverses of the nonlinearities. This search can be done efficiently by applying the ACE technique (Breiman and Friedman, 1985) from non-parametric statistics that performs an alternating estimation of conditional expectations by the smoothing of scatter plots. The bottom line is that the nonlinear re-transformation reduces the PNL mixing problem to a linear one. This allows us to apply a linear BSS algorithm as a second step. While in principle any ICA/BSS method could be used, second-order techniques (e.g. Belouchrani et al., 1997, Ziehe and Müller, 1998) are more consistent with the Gaussianity assumptions, because in higher-order ICA methods the source signals are required to be non-Gaussian. Second-order methods are less influenced by remaining, small non-linear distortions.

Clearly, ACE is not limited to the 2×2 case, since it scales naturally to the $n \times n$ case for which an algorithmic description can be found in the appendix (or more detailed in Breiman and Friedman,

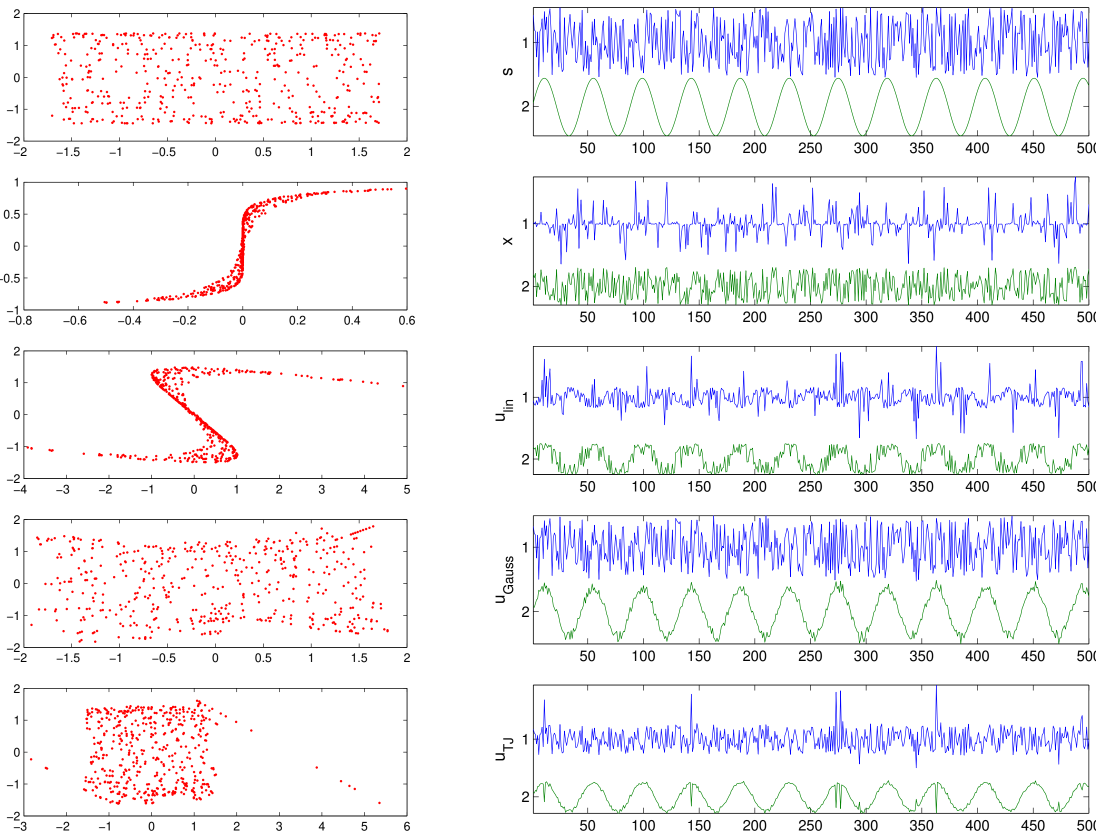


Figure 9: Scatter plots and waveforms for the sources, mixed and unmixed signals (c.f. Figure 1). From top to bottom row we see true source signals, observed post-nonlinear mixtures, linearly unmixed signals, separation results of Gauss-TD, separation results of Taleb and Jutten's algorithm.

1985, Härdle, 1990). Furthermore, Gauss-TD is a computationally simpler alternative of ACE-TD, since the former estimates the inverses of the nonlinearities directly by Gaussianizing the individual components. Gaussianization was also employed as an initialization of a more sophisticated PNL algorithm and turned out to improve computational speed substantially (Solé et al., 2003). Both algorithms have theoretical guarantees to yield the correct results, if the linearly mixed signals before the PNL distortion are Gaussian distributed and provided that there is none or only low additive noise present after the post-nonlinearity. Because of the central limit theorem, the Gaussianity assumption usually holds approximately in practice, especially if the number of sources are large and the mixing matrix is not sparse. While the Gaussianization gives satisfactory results under these assumptions, the ACE method may run into difficulties if the correlations of the linearly mixed signals are low (see also the discussion in Breiman and Friedman, 1985).

Throughout the paper we have assumed that the linearly mixed sources are subject to a Gaussian distribution, which is well justified due to the central limit theorem. Note however that one can – in a straight forward manner – also incorporate other assumptions on the marginal distribution of the mixed signals. So, in other words we could very well, for example, 'Laplacianize' once we have such prior knowledge available.

In contrast to methods that simultaneously estimate the mixing matrix, the nonlinear equalization functions and the distributions of sources (like e.g. Taleb and Jutten, 1999, Achard et al., 2001, S. Achard and Jutten, 2003), our two-stage procedures are computational more attractive. They work well with large data sets and scale linearly with respect to the number of sources. Interestingly, our methods can be easily applied to convolutive mixtures, which is important for solving the real-room "cocktail-party" problem with BSS (see also Koutras, 2002). Here occasionally nonlinear transfer functions of the sensors (microphones) or the use of class C amplifiers can hinder a proper separation.

Concluding, the proposed framework gives simple algorithms of high efficiency with a solid theoretical background for signal separation in applications with a PNL distortion that are of importance, for example, in real-world sensor technology. In our simulations the simpler Gauss-TD algorithm compares favorably to ACE-TD and to mutual information based methods as proposed by Taleb and Jutten (1999). However, it is conceivable that further improvements of the smoothing procedure will lead to increased performance of the ACE and other methods relying on non-parametric density estimation. Future research will therefore be concerned with incorporating domain knowledge for a better tuning of the smoothers to improve and robustify the separation performance.

Acknowledgements

This work is the extended version of Ziehe et al. (2001, 2003a). This work has been partially supported by the European Community under the Information Society Technology (IST) RTD programme, contract IST-1999-14190-BLISS. The authors are solely responsible for the content of this article. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

Appendix A. The Multi-dimensional ACE Algorithm

Whenever there are more than two sensors, we apply the multivariate ACE algorithm (Breiman and Friedman, 1985, Hastie and Tibshirani, 1990, Härdle, 1990):

Algorithm 2 The ACE algorithm for multi-dimensional cases

{initialize}
$g_1^{(0)}(x_1) \leftarrow x_1/||x_1||$ and $g_2^{(0)}(x_2), \ldots, g_n^{(0)}(x_2) = 0$
repeat
  for $i = 2$ to n **do**:
    $g_i^{(k+1)}(x_i) \leftarrow E\{g_1^{(k)}(x_1) - \sum_{j<i} g_j^{(k+1)}(x_j) - \sum_{j>i} g_j^{(k)}(x_j) \mid x_i\}$
  end for loop
  $g_1^{(k+1)}(x_1) \leftarrow E\{\sum_{j=2}^n g_j^{(k+1)}(x_j) \mid x_1\}/\|E\{\sum_{j=2}^n g_j^{(k+1)}(x_j) \mid x_1\}\|$
until $E\{g_1(x_1) - \sum_{j=2}^n g_j(x_j)\}^2$ fails to decrease

This general ACE algorithm yields transformations $g_1(\cdot), \ldots, g_n(\cdot)$ which minimize the least squares error

$$\min_{g_1,\ldots,g_n} \frac{E\{g_1(x_1) - \sum_{j=2}^n g_j(x_j)\}^2}{E g_1^2(x_1)}.$$

As in the case of two variables, it has been proven in Breiman and Friedman (1985) that the optimal transformations in the least squares sense are equivalent to the transformations with maximal correlation,

$$\max_{g_1,\dots,g_n} \operatorname{corr}(\, g_1(x_1),\, \sum_{j=2}^{n} g_j(x_j)\,),$$

up to scaling factors.

Hence, if the linearly mixed signals z are jointly Gaussian distributed, we can recover the inverse functions $g_1(\cdot),\dots,g_n(\cdot)$ of the true nonlinearities with the multi-dimensional ACE algorithm. For details we refer to the proof presented in Breiman and Friedman (1985).

As a final remark, we note that in the original ACE algorithm one component—e.g. x_1—is treated somewhat special as the "response" variable, while the BSS problem is intrinsically symmetric with respect to the components $x_1,\dots,x_n$. Therefore, the development of a symmetric version of this algorithm is left as a topic for further research.

References

S. Achard, D.-T. Pham, and C. Jutten. Blind source separation in post nonlinear mixtures. In T.-W. Lee, editor, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, pages 295–300, San Diego, California, 2001.

S. Amari, A. Cichocki, and H.-H. Yang. A new learning algorithm for blind source separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 757–763. MIT Press, Cambridge, MA, 1996.

A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

A. Belouchrani, K. Abed Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique based on second order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

P. Billingsley. *Probability and Measure*. John Wiley & Sons, New York, 1995.

L. Breiman and J.H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598, 1985.

J.-F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings-F*, 140(6):362–370, 1993.

J.-F. Cardoso and A. Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM J.Mat.Anal.Appl.*, 17(1):161–164, 1996.

P. Comon. Independent component analysis—a new concept? *Signal Processing*, 36:287–314, 1994.

G. Deco and D. Obradovic. Linear redundancy reduction learning. *Neural Networks*, 8(5):751–755, 1995.

W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1990.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel feature spaces and nonlinear blind source separation. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 761–768. MIT Press, 2002.

S. Harmeling, A. Ziehe, M. Kawanabe, and K.-R. Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15:1089–1124, 2003.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1990.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

C. Jutten and J. Hérault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24:1–10, 1991.

C. Jutten and J. Karhunen. Advances in nonlinear blind source separation. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 245–256, Nara, Japan, April 1-4 2003. Invited paper in the special session on nonlinear ICA and BSS.

H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge University Press, Cambridge, UK, 1997.

M. Kawamoto, K. Matsuoka, and M. Oya. Blind separation of sources using temporal correlation of the observed signals. *IEICE Trans. Fundamentals*, E80-A(4):695–704, 1997.

A. Koutras. Blind separation of nonlinear convolved speech mixtures. In *Proc. ICASSP 2002*, pages 913–919, Orlando, Florida, 2002.

L. E. Larson. Radio frequency integrated circuit technology low-power wireless communications. *IEEE Personal Communications*, 5(3):11–19, 1998.

T.-W. Lee, B.U. Koehler, and R. Orglmeister. Blind source separation of nonlinear mixing models. *IEEE International Workshop on Neural Networks for Signal Processing*, pages 406–415, 1997.

T.-W. Lee, A. Ziehe, R.Orglmeister, and T.J. Sejnowski. Combining time-delayed decorrelation and ICA: Towards solving the cocktail party problem. In *Proc. ICASSP98*, volume 2, pages 1249–1252, Seattle, 1998.

J. K. Lin, D. G. Grier, and J. D. Cowan. Faithful representation of separable distributions. *Neural Computation*, 9(6):1305–1320, 1997.

G. Marques and Luis B. Almeida. Separation of nonlinear mixtures using pattern repulsion. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 277–282, Aussois, France, 1999.

L. Molgedey and H. G. Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72(23):3634–3637, 1994.

K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, and B. Schölkopf. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural Networks*, 12(2):181–201, 2001.

K.-R. Müller, P. Philips, and A. Ziehe. $JADE_{TD}$: Combining higher-order statistics and temporal information for blind source separation (with noise). In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 87–92, Aussois, France, 1999.

N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

P. Pajunen, A. Hyvärinen, and J. Karhunen. Nonlinear blind source separation by self-organizing maps. In *Proc. Int. Conf. on Neural Information Processing*, pages 1207–1210, Hong Kong, 1996.

P. Pajunen and J. Karhunen. A maximum likelihood approach to nonlinear blind source separation. In *Proceedings of the 1997 Int. Conf. on Artificial Neural Networks (ICANN'97)*, pages 541–546, Lausanne, Switzerland, 1997.

P. Pajunen and J. Karhunen, editors. *Proc. of the 2nd Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA200)*. Otamedia, Helsinki, Finland, 2000.

L.C. Parra and C. Spence. Convolutive blind source separation of non-stationary sources. *IEEE Trans. on Speech and Audio Processing*, 8(3):320–327, 2000. US Patent US6167417.

D.-T. Pham. Joint approximate diagonalization of positive definite matrices. *SIAM J. on Matrix Anal. and Appl.*, 22(4):1136–1152, 2001.

D.-T. Pham and P. Garrat. Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing*, 45(7):1712–1725, 1997.

B. Pompe. Measuring statistical dependences in a time series. *J. Stat. Phys.*, 73:587–610, 1993.

D.-T. Pham S. Achard and C. Jutten. Quadratic dependence measure for nonlinear blind sources separation. In *Proc. 4th Intern. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 263–268, Nara, Japan, 2003.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

B. Schölkopf, A.J. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel Eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

J. Solé, M. Babaie-Zadeh, C. Jutten, and D.-T. Pham. Improving algorithm speed in PNL mixture separation and Wiener system inversion. In *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2003)*, Nara, Japan, April 2003.

A. Taleb. A generic framework for blind source separation in structured nonlinear models. *IEEE Trans. on Signal Processing*, 50(8):1819–1830, 2002.

A. Taleb and C. Jutten. Nonlinear source separation: The post-nonlinear mixtures. In *Proc. European Symposium on Artificial Neural Networks*, pages 279–284, Bruges, Belgium, 1997.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820, 1999.

H. Valpola, X. Giannakopoulos, A. Honkela, and J. Karhunen. Nonlinear independent component analysis using ensemble learning: Experiments and discussion. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 351–356, Helsinki, Finland, 2000.

H. Valpola, E. Oja, A. Ilin, A. Honkela, and J. Karhunen. Nonlinear blind source separation by variational Bayesian learning. *IEICE Transactions (Japan)*, E86-A(3), March 2003. To appear in a Special Section on Blind Signal Processing (Invited paper).

V. N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

H.-H. Yang, S. Amari, and A. Cichocki. Information-theoretic approach to blind separation of sources in non-linear mixture. *Signal Processing*, 64(3):291–300, 1998.

A. Yeredor. Blind separation of gaussian sources via second-order statistics with asymptotically optimal weighting. *IEEE Signal Processing Letters*, 7(7):197–200, 2000.

A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller. Separation of post-nonlinear mixtures using ACE and temporal decorrelation. In T.-W. Lee, editor, *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2001)*, pages 433–438, San Diego, California, 2001.

A. Ziehe, M. Kawanabe, S. Harmeling, and K.-R. Müller. Blind separation of post-nonlinear mixtures using gaussianizing transformations and temporal decorrelation. In *Proc. 4th Intern. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 269–274, Nara, Japan, April 2003a.

A. Ziehe, P. Laskov, K.-R. Müller, and G. Nolte. A linear least-squares algorithm for joint diagonalization. In *Proc. 4th Intern. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 469–474, Nara, Japan, 2003b.

A. Ziehe and K.-R. Müller. TDSEP–an efficient algorithm for blind separation using time structure. In *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, pages 675–680, Skövde, Sweden, 1998.

A. Ziehe, K.-R. Müller, G. Nolte, B.-M. Mackert, and G. Curio. Artifact reduction in magnetoneurography based on time-delayed second-order correlations. *IEEE Trans. Biomed. Eng.*, 47(1): 75–87, 2000.