Sequential fixed-point ICA based on mutual information minimization

Marc M. Van Hulle
K.U.Leuven, Laboratorium voor Neuro- en Psychofysiologie
Campus Gasthuisberg, O&N 2, Bus 1021, Herestraat 49
B-3000 Leuven, BELGIUM
E-mail: marc@neuro.kuleuven.be

July 17, 2007

Abstract

A new gradient technique is introduced for linear ICA based on the Edgeworth expansion of mutual information, for which the algorithm operates sequentially using fixed-point iterations. In order to address the adverse effect of outliers, a robust version of the Edgeworth expansion is adopted, in terms of robust cumulants, and robust derivatives of the Hermite polynomials are used. Also, a new constrained version of ICA is introduced, based on goal programming of mutual information objectives, and which is applied to the extraction of the antepartum fetal electrocardiogram (ECG) from multi-electrode cutaneous recordings on the mother’s thorax and abdomen.
1 Introduction

The central idea behind ICA is to transform the observed signals into statistically independent ones which, in turn, are expected to correspond to the usually assumed to be linearly admixed and independent, but otherwise unknown source signals (for a review, see Hyvärinen et al., 2001). A natural measure of independence is mutual information (MI) (Comon, 1994), a nonnegative scalar that equals zero when the signals are independent, so that ICA through MI minimization could be regarded as a direct method. However, MI estimation of continuous signals is notoriously difficult but, when performing prewhitening of the observed signal space, MI minimization becomes equivalent to finding the orthogonal directions for which the negentropy is maximized (Comon, 1994), so that much research has focused on developing one-dimensional approximations of negentropy and differential entropy.

A popular approximation of the one-dimensional differential entropy and negentropy is based on the one-dimensional Edgeworth expansion or Gram-Charlier expansion of a Gaussian (Comon, 1994; Amari et al., 1996). Other one-dimensional algebraic density approximations have been suggested, sometimes as alternatives to the Edgeworth expansion, such as the approximative maximum entropy approach (Hyvärinen, 1998), and those in the context of likelihood maximization and Infomax (Pham et al., 1992; Bell and Sejnowski, 1995) (for the connection between the latter two, see Cardoso, 1997), such as the Pearson distribution (Karvanen et al., 2000), the (Extended) Generalized Lambda Distribution (Eriksson et al., 2000), and the exponentially tilted Gaussian density (Hastie and Tibshirani, 2003). Another algebraic
density approximation, that in principle could also be used for ICA, is the normal inverse Gaussian density (Barndorff-Nielsen, 1978). However, most of these algebraic density approximations cannot be readily extended to the multivariate case.

Cardoso and Souloumiac (1993) used the multivariate Edgeworth expansion for obtaining a mutual information contrast function in terms of a sum of squared fourth-order cross-cumulants (JADE algorithm, see also Cardoso, 1999). Using the multilinearity property of the cumulants, since the mixing is linear, and using prewhitening, so as to ensure that the transformation matrix is orthonormal, the MI contrast is minimized with the Jacobi optimization technique (successive Givens rotations of pairs of coordinate axes using an analytical expression of the Givens angle) (for an extended approach, see Blaschke and Wiskott, 2002).

Recently, Max Welling (2006) introduced an orthogonal series expansion that is closely related to the Edgeworth expansion, but in terms of robust moments, $\langle (\gamma y)^n \exp^{\frac{1}{2}(1-\gamma^2)y^2} \rangle$, with $n$ the order of the moment and $1 < \gamma^2 < 2$, and robust cumulants. This offers new, yet unexplored possibilities to robustly estimate differential entropy and mutual information for cumulant-based ICA.

In this article, we will develop a gradient technique for minimizing the multivariate Edgeworth expansion of MI. The gradient consists of a sum of products of cumulants and derivatives of Hermite polynomial terms. Since the gradients of the marginal entropies cancel, the algorithm only relies on cross-cumulant terms. The independent components are found sequentially, with or without prewhitening. The algorithm uses fixed-point iterations.
We then consider Welling’s robust version of the Edgeworth expansion, and motivate how it fits in the gradient descent algorithm. We show that the accuracy of the robust version is comparable or better than that of the mainstream ICA algorithms, and that the robustness is now comparable to robust FastICA’s and Infomax’, albeit it is cumulant-based. As an application, we consider the problem of extracting the fetal electrocardiogram (ECG) from multi-electrode recordings on the mother’s skin, and for which we develop a new version of constrained ICA, based on goal programming of mutual information objectives.

2 MI minimization

Let \( s(t) = [s_1(t), \ldots, s_m(t)]^T \) be an \( m \)-dimensional component-wise zero-mean vector, \( s \in \mathbb{R}^m \), with \( t \) a time index, that corresponds to \( m \) mutually independent scalar-valued source signals \( s_i \). Instead of \( s \), a data vector \( x(t) = [x_1(t), \ldots, x_d(t)]^T \in \mathbb{R}^d \) is observed at each time instant \( t \), and which is the result of linearly admixing the sources:

\[
x(t) = As(t),
\]

where \( A \) is a full rank \( d \times m \) scalar matrix, called mixing matrix (thus, instantaneous linear mixing). We further assume that we have \( N \) such observations and that \( d = m \) (square mixture matrix) but, since the algorithm will build up the source estimates one by one, which is roughly equivalent to doing projection pursuit, it is sufficient that \( d \geq m \). Non-Gaussianity of the independent components is essential for the identifiability of the model (Comon, 1994). A noise vector may also be present in the model (noisy mixtures).
The goal of ICA is to find a matrix $W$ such that the signals $y$:

$$ y(t) = Wx(t) = WAs(t) $$

are statistically independent. The matrix $A$ is estimated as the inverse of
matrix $W$, and the sources are recovered up to scaling and permutation inde-
terminacies. The rows of $W$ are thus the independent components direction
estimates.

Comon (1994) argued that a natural measure of independence is mutual
information (MI), which we can define as the difference between the sum of
the marginal entropies and the joint entropy:

$$ MI(y_1, \ldots, y_d) = \sum_i H_i(y_i) - H(y_1, \ldots, y_d). $$

Since MI estimation is considered to be a hard problem (for review, see
Paninski, 2003), usually the link with negentropy is made, by constraining
the variables to be uncorrelated (prewhitening). Here, we will estimate MI
based on the multivariate Edgeworth expansion of a Gaussian.

We first need the Edgeworth expansion of the multivariate density $p(y)$. We
consider the Edgeworth expansion up to order four about its best normal
estimate $\phi_p$ (i.e., with the same mean and covariance matrix as $p$) (Barndorff-
Nielsen and Cox, 1989) (also called Gram-Charlier A series):

$$ p(y) \approx \phi_p(y) \left( 1 + \frac{1}{3!} \sum_{i,j,k} \kappa^{i,j,k} h_{ijk}(y) + \frac{1}{4!} \sum_{i,j,k,l} \kappa^{i,j,k,l} h_{ijkl}(y) \right) $$

with $h_{ijk}$ the $ijk$-th Hermite polynomial, with $i, j, k$ the corresponding input
dimensions, $i, j, k \in \{1, \ldots, d\}$, and $\kappa^{i,j,k}$ the corresponding standardized
cumulant, $\kappa^{i,j,k} = \frac{\kappa^{ijk}}{\sigma_i \sigma_j \sigma_k}$ (for the number of observations $N$ large), with $\kappa^{ijk}$
the third central moment over input dimensions $i, j, k$ and $\sigma_i$ the variance of
$y_i$, and $h_{ijkl}$ the $ijkl$-th Hermite polynomial over input dimensions $i, j, k, l$ and the corresponding standardized cumulant $κ_{ijkl}^{ij,k,l} = \frac{κ_{ijkl} - σ_{ij}σ_{kl}}{σ_{ij}σ_{kl}}[3]$, with $κ_{ijkl}$ the fourth central moment over input dimensions $i, j, k, l$ and $σ_{ij}σ_{kl}$ a compact representation of the sum over the 3 partitions of the subscripts $i, j, k, l$ and $σ_{ij}$ the covariance between $y_i$ and $y_j$. For the connection between moments and cumulants in the multivariate case, we refer to (McCullagh, 1987). We can rewrite eq. (4) more compactly as $p(y) \approx φ_p(y)(1 + z(y))$.

Now since $MI(p(y)) = \sum_i H_i(p(y_i)) - H(p(y)) \approx -\sum_i \left( E[log(φ_p(y_i))] - E[log(1+z_i(y_i))] \right) + E[log(φ_p(y))] + E[log(1+z(y))]$, and since we will perform gradient descent on it, we need the derivatives of the marginal and joint entropies. We will extract the independent components sequentially, one by one. Let $X = \{x^μ|μ = 1, \ldots, N\}$ be a sample of $N$ mixtures and $Y = \{y^μ = Wx^μ|μ = 1, \ldots, N\}$ and $Y_i = \{y_i^μ = W_i x^μ|μ = 1, \ldots, N\}$ the corresponding output samples, with $W_i$ the $ith$ row in $W$ (note that the dimensionality of $y$ equals $d ≥ m$). Assume that the $ith$ independent component direction is being optimized, the marginal entropy derivative then becomes, with the expectations replaced by averages:

$$\frac{\partial H_i(Y_i)}{\partial W_i} \approx -\frac{1}{N} \sum_μ \frac{κ_{i,i,i,i}^{ij,i,i} x^μ}{σ_i^3} [(y_i^μ - \bar{y}_i)^2 - σ_i^2]$$

$$+ α \frac{κ_{i,i,i,i}^{ij,i,i} x^μ}{6} (y_i^μ - \bar{y}_i)(y_i^μ - \bar{y}_i)^2 - 3σ_i^2]$$

with $\bar{y}_i$ the mean of $Y_i$ and $α \equiv \text{sign}(κ_{i,i,i,i}^{ij,i,i})$, and after some algebraic manipulations, and when taking for the common denominator $1 + z \approx 1$ to simplify the expression. The joint entropy derivative is then, given the already estimated independent component directions $j$:

$$\frac{\partial H(Y)}{\partial W_i} \approx -\frac{1}{N} \sum_μ \frac{1}{3!} \left[ κ_{i,i,j}^{ij,i,i} \left( \frac{3}{σ_i^3} (y_i^μ - \bar{y}_i)^2 x^μ - \frac{3}{σ_i} x^μ \right) + \right.$$
\[
\sum_{j,j < i} \left( \kappa_{i,j} \left( \frac{2}{\sigma_i \sigma_j} (y_i^\mu - \bar{y}_i) x^\mu (y_j^\mu - \bar{y}_j) - 2 \frac{\sigma_{ij}}{\sigma_i \sigma_j} x^\mu \right) + \kappa_{i,j,j} \left( \frac{1}{\sigma_i \sigma_j} x^\mu (y_j^\mu - \bar{y}_j)^2 - \frac{x^\mu}{\sigma_i} \right) \right)
\]

\[
+ \frac{1}{4!} \left[ \alpha \kappa_{i,i,i,j} \left( \frac{4}{\sigma_i^3} (y_i^\mu - \bar{y}_i)^3 x^\mu - \frac{12}{\sigma_i^2} (y_i^\mu - \bar{y}_i) x^\mu \right) + \beta \sum_{j,j < i} \left( \kappa_{i,i,i,j} \left( \frac{3}{\sigma_i^3 \sigma_j} (y_i^\mu - \bar{y}_i)^2 x^\mu (y_j^\mu - \bar{y}_j) - \frac{3}{\sigma_i \sigma_j} x^\mu (y_j^\mu - \bar{y}_j) \right) - \frac{6 \sigma_{ij}}{\sigma_i \sigma_j} (y_i^\mu - \bar{y}_i) x^\mu \right) + \kappa_{i,j,j,j} \left( \frac{2}{\sigma_i^2 \sigma_j^2} (y_i^\mu - \bar{y}_i) x^\mu (y_j^\mu - \bar{y}_j)^2 - \frac{2 \sigma_{ij}}{\sigma_i \sigma_j} x^\mu (y_j^\mu - \bar{y}_j) + \frac{1}{\sigma_i} \right) \right],
\]

with \( \beta \equiv \text{sign}(\kappa_{i,j,j,j}) \). Note that we have restricted ourselves to \((i, j)\) (pairwise, \(j < i\)) interactions since their signs \( \beta \) are uniquely specified, and for reasons of computational simplicity (see further). When \( i \) is the first independent component direction that is estimated, then we need only to perform gradient descent on \( H_i(Y_i) \) (the terms in \( \frac{\partial H(Y)}{\partial W_i} \) drop out).

When looking carefully, in the computation of the mutual information derivative, we see that the terms in \( \frac{\partial H(Y)}{\partial W_i} \) are exactly canceled by the \( \kappa_{i,i,i} \) and \( \kappa_{i,i,i,i} \) terms in \( \frac{\partial H(Y)}{\partial W_i} \) (auto-cumulant terms). Since we only need to consider \( \kappa_{i,j}, \kappa_{i,j,j}, \kappa_{i,i,j,j}, \kappa_{i,i,j,j}, \kappa_{i,j,j,j}, \forall j < i \), the number of terms to compute increases linearly with the number of independent component directions. (Note that the Edgeworth expansion of the joint density has \( \mathcal{O}(d^3) \) terms in \( \kappa_{i,j,k} \) and \( \mathcal{O}(d^4) \) terms in \( \kappa_{i,j,k,l} \), but the complexity is less here since the independent components are computed sequentially and since we restrict ourselves to pairwise interactions.)
The complete incremental learning rule for the \( i \)th independent component direction is then \( \Delta W_i(Y) = -\eta \frac{\partial MI(Y)}{\partial W_i} \), with \( \eta \) the learning rate (a constant), and follows from plugging eq. (6) into the derivative of \( MI \), without the auto-cumulant terms. The complete algorithm is listed in Appendix 1. For a learning rate of 0.01 (as in the simulations), typically about 10 iterations are needed. One can do without the learning rate \( \eta \) by considering normalized gradient descent (adaptive step size, see Appendix 2). This speeds up the algorithm by a factor 5.

Since we use normalized cumulants, normalized by the second-order moments, prewhitening is, strictly speaking, not needed. For that case, we do not need to decorrelate the independent components (with the covariance matrix of \( y \)) to prevent the independent component directions from converging to the same direction, since this is taken care of by the cross-cumulant terms in the joint entropy derivative. However, as we will demonstrate, prewhitening leads to a better and more uniform performance, since less parameters have to be estimated.

\[ 2.1 \text{ Backfitting} \]

Since we use an iterative procedure, we can significantly improve the initial independent component estimates by using backfitting: we estimate the \( i \)th component direction again, given the \((d - 1)\) others (thus by keeping them fixed), using our update scheme (stage 4 in Appendix 1). This also enables us to improve the first estimated independent component direction. As an example, we consider two Laplacian-distributed source signals and consider them as observables (we thus mix with a unit matrix). We look at the largest
of the two difference angles (error angles) between the estimated and true independent component directions. We take the fixed step size version of our algorithm ($\eta$ constant) and consider sets of $N = 10000$ observations and 20 data sets, and no prewhitening. The result is shown in Fig. 1A (thin line). We clearly see that five backfitting runs suffice to get the error down from 10.2 to 0.80 deg, thus, a tenfold gain in accuracy.

2.2 Computational complexity

The computational complexity for the first run of our algorithm is on the order of $O(d^4N)$, but for each backfitting run it becomes $O(d^3N)$. The FastICA algorithm has a complexity on the order of $O(d^3N)$, but ours requires backfitting, so that the complexity additionally scales with the number of backfitting runs. Hence, our algorithm is computationally more intensive. The complexity of Infomax is similar to our algorithm: $O(d^3N)$ times the number of iterations performed (estimated from Bell’s Matlab code, ftp://ftp.cnl.salk.edu/pub/tony/sep96.public). The complexity of the JADE algorithm is on the order of $O(d^4N)$ (estimated from Cardoso’s Matlab code, http://www.tsi.enst.fr/~cardoso/guidesepsou.html).

2.3 Convergence speed

FastICA’s convergence is quadratic in general and cubic for symmetric distributions (Hyvärinen, 1999). The JADE algorithm uses Givens rotations, the convergence of which is known to be cubic (Golub and Van Loan, 1989). Note that, since both our algorithm (with $\eta$ constant) and Infomax are gradient algorithms with non-adaptive step sizes, they converge linearly (Bishop, 1995).
For the adaptive step size version of our algorithm (normalized gradient), it can be shown that we can employ fixed-point iteration (Appendix 3), and that its convergence is approximately quadratic for symmetric distributions and linear in general (Appendix 4).

3 Robust Edgeworth expansion

Recently, an orthogonal series expansion was introduced by Max Welling (2006) that is closely related to the Edgeworth expansion, but expressed in terms of robust moments, $\langle (\gamma y)^n \exp^{\frac{1}{2}(1-\gamma^2)y^2} \rangle$ (or correspondingly, robust cumulants), that contain a decreasing exponential that makes them robust against outliers, with $n$ the order of the moment, and $\gamma$ a parameter that can be tuned between the classical case $\gamma^2 = 1$ and the robust case $1 < \gamma^2 < 2$. The approximation of a one-dimensional density about its best normal estimate $\phi_p$, up to the fourth order, is then (Welling, 2006):

$$p(y) \approx \phi_p(y) \left( \kappa_0^\gamma + \kappa_1^\gamma h_1(\gamma y) + \frac{1}{2}(\kappa_2^\gamma - \kappa_0^\gamma) h_2(\gamma y) + \frac{1}{3!}(\kappa_3^\gamma - 3\kappa_1^\gamma) h_3(\gamma y) + \frac{1}{4!}(\kappa_4^\gamma - 6\kappa_2^\gamma + 3\kappa_0^\gamma) h_4(\gamma y) \right)$$

(7)

with $\kappa_n^\gamma$ the robust moments of order $n$. Note that this equation is actually the robust version of the Gram-Charlier expansion of a Gaussian; the robust version of the Edgeworth expansion is the same as in eq. (4) but with the cumulants replaced by their robust counterparts (for the link between robust cumulants and robust moments, see Appendix B of Welling, 2006).

When we continue with this equation, and estimate the entropy from the
integral, we obtain, after some straightforward algebraic manipulations:

\[
H(y) \approx \frac{1}{2} \log(2\pi e) - \frac{1}{2} \left( \kappa^1 - \frac{1}{2} (\kappa^2 - \kappa^0) + \frac{1}{3!} (\kappa^3 - 3\kappa^1) + \frac{1}{4!} (\kappa^4 - 6\kappa^2 + 3\kappa^0) \right),
\]

where we have taken the entropy bias into account, so that the Gaussian case is correct, and where we have assumed that the location of the density has been subtracted from the data (e.g., the median) and the scale (e.g., the standardized median of the absolute deviations from the median, MAD) has been used to rescale the data to unit variance. For \( \gamma^2 = 1.3 \) the entropy estimate closely matches the theoretical entropy value of a unit variance Laplacian, so that we continue with this value (but any value \( 1 \ll \gamma^2 < 2 \) has proved to be sufficient in the ICA simulations).

In our MI descent algorithm, the derivatives of the Hermite polynomials are present, and they are sensitive to outliers. Hence, we suggest here to transform their arguments \( y \rightarrow \gamma y \exp^{\frac{1}{2}(1-\gamma^2)}y^2 \) so that the Hermite polynomial derivatives also become robust. We further note that \( \kappa^1 \approx 0, \kappa^0 \approx 1 \) (but we correct for the bias), \( \kappa^2 \approx 1 \) (due to prewhitening), and \( \kappa^0 \) drops out when taking the derivatives, so that our algorithm stays in essence the same, except that robust cumulant estimates are used, and that the arguments of the Hermite polynomial derivatives have been replaced by \( \gamma y \exp^{\frac{1}{2}(1-\gamma^2)}y^2 \).

(Note that the transformation causes an extra scale factor \( \frac{\partial}{\partial y} \gamma y \exp^{\frac{1}{2}(1-\gamma^2)}y^2 \) in the Hermite polynomial derivatives.)
4 Simulations

We first compare the performance of the algorithm based on the non-robust and the robust versions of the Edgeworth expansion. The result of backfitting for the robust algorithm, using otherwise the same settings, is shown in Fig. 1A (thick line). We observe a faster convergence and a better performance compared to the non-robust version (thin line). In order to further compare the performance of both versions, we consider a mixture of five Laplacian-distributed source signals (supergaussian case). We align the independent component directions along the coordinate axes in $y$-space except for the second independent component direction which makes an angle $\theta$ to the horizontal axis (X-axis, thus, the direction of the first independent component), which we vary between 5 and 90 deg in the XY-plane (for a comparison using random mixing matrices, see further). In order to quantify the performance, we determine the maximum error angle between the true and estimated independent component directions. We could also look at the mutual information achieved (see further) or, since the mixing is known to us, at the cross-talk error (Yang and Amari, 1997), but the error angle is more easy to interpret, also given that it can be as small as 1 deg. The median maximum error angles for $N = 10000$ observations and 20 data sets are plotted in Fig. 1B for the non-robust case without prewhitening (stippled line) and the robust case, also without prewhitening (thin solid line). We can also apply in our algorithm prewhitening using the Jacobi transform, and incremental Gram-Schmidt-like decorrelation at every iteration step. The result for the robust version (thick solid line) is a constant performance that is even better than the $\theta = 90$ deg case without prewhitening.
We will now compare our robust algorithm with FastICA (Hyvärinen, 1999, using the contrast function $G(y_i) = -\exp\left(-\frac{y_i^2}{2}\right)$), MaxKurt and JADE (Cardoso and Souloumiac, 1993; Cardoso, 1999), and Infomax (Bell and Sejnowski, 1995; or extended Infomax (Lee et al., 1999) when there are sub-gaussian sources), all assuming prewhitening. For the FastICA, JADE and (extended) Infomax algorithms, we use the publicly available Matlab codes; for MaxKurt we developed the code around the one mentioned in (Cardoso, 1999). We again consider the previous example (but the results are similar for mixtures of other types of distributions such as sub- and supergaussian and asymmetric ones). The median maximum error angles are plotted in Fig. 1C for $d = 2$ and $5$ (thin vs. thick lines), again for $N = 10000$ observations and 20 data sets. We observe that the performance in general degrades as $d$ increases, as expected, since there are more parameters to estimate. We also observe that for $d = 2$ there is no significant difference in average maximum error angle of the robust version of our algorithm compared to each one of the other algorithms (the Stouffer combined test applied to 18 t-tests for different means returned $p > 0.01$). Also, the average MI, estimated with Kraskov’s non-parametric method (Kraskov et al., 2004), before mixing and after demixing are statistically indistinguishable for the robust version of our algorithm (Stouffer combined test $p > 0.01$). This means that, statistically, and in MI terms, one cannot do better. More importantly, we observe that our algorithm for $d = 5$ is superior to any of the other algorithms since there is a significant difference in the average maximum error angle of our algorithm compared to each one of them (Stouffer combined test $p \ll 0.01$). Furthermore, the average MI before mixing and after demixing for our al-
algorithm are not significantly different ($p > 0.01$). On the other hand, for example for JADE, the average MI before mixing and after demixing are significantly different (Stouffer combined test $p \ll 0.01$).

In order to test the robustness of our approach, we apply the procedure used by Hyvärinen (1999) to test FastICA. We mix $d$ Laplacian signals using several random matrices of which the elements are drawn from the uniform distribution $[-1,1]$. The condition number of the matrix is in the interval $[1,d]$. To test the robustness, we replace $d$ data points by outliers with magnitudes 10 (thus, by increasing their vector lengths to 10); the distribution of the other data points approximately has a unit variance. We insert the outliers after determining the condition number and after prewhitening. We determine the median maximum error angles and quartiles for 20 data sets of $N = 10000$ observations and for $d = 2,\ldots,7$. The results are shown for our algorithm in Fig. 1D. We observe that our robust approach (solid lines) is again superior to our non-robust approach (stippled lines), not only because of the better absolute performances but also that the outliers do not affect the performance (compare thin and thick solid lines). This means that notwithstanding our method is based on the Edgeworth expansion, it is robust against outliers. We can also compare our robust algorithm against the standard algorithms (Fig. 1E). The results show that our robust algorithm is indeed as robust as FastICA and Infomax (for both algorithms, there is no significant difference in average maximum error angles with our algorithm: the Stouffer combined test applied to 5 t-tests returns $p > 0.01$, excluding $d = 2$), and several times more robust than JADE and MaxKurt.

Finally, we want to show the dependency on the number of data points
$N$, and compare it with that of the standard algorithms. We reconsider the example of random mixing matrices (but without outliers) for $d = 2$ and 5. The results are shown in Fig. 1F. We observe that our algorithm is again among the best performing ones.

5 Constrained ICA

When performing subspace ICA, several solutions can be found depending on initialization. Indeed, all combinations of independent components are valid subspaces. Hence, without setting some preference, by using prior knowledge, any combination could be taken. What we want to focus on here is the extraction of a specific combination of independent components (i.e., subspace extraction) that are “closest” to a set of reference signals. We use these reference signals as constraints for the independent component analysis of the original signal set. Subspace extraction in ICA started with the work of Luo and co-workers (Luo et al., 1999), after which several alternatives were introduced (Lu and Rajapakse, 2001; Liao and Carin, 2002; James and Gibson, 2003).

It is clear that, since we perform ICA, we should use independent components as constraints. However, in defining the reference signals, there is the issue whether or not they are coming from the same signal set. We will therefore consider two cases in this section: 1) we assume that the reference signals come from the same signal set on which ICA is performed, in order to verify that we indeed can extract the desired independent components subspace, but 2) we relax that in the second case where we will take another part of the signal set as the reference so that there is no longer a one-to-one relationship
(training/testing configuration). As an example, we consider the subspace that separates the fetal electrocardiogram (ECG) from the mother’s. As a “closeness” measure, we use mutual information.

Before we show how we can perform constrained ICA, we first assume that we are only interested in the extraction of a signal $y_i$ that is “closest” to a reference signal $z$. Taking MI as a closeness measure, the formulation is straightforward:

$$\max_{W_i} (MI(y_i, z)) = \max_{W_i} (H(y_i) + H(z) - H(y_i, z))$$

$$\equiv \min_{W_i} (-H(y_i) + H(y_i, z)), \quad (9)$$

and which is obtained by performing gradient descent incrementally. One can verify that it exactly fits our algorithm with the difference that we reverse the signs of the marginal and joint entropy derivatives, and that the direction of $z$ is not updated. We can easily extend this by using two or more constraints, $z_1, z_2, \ldots$, that are mutually independent, and with which we want to constrain extraction of two (or more) source estimates $y_1, y_2, \ldots$ (thus, a constrained ICA problem) (e.g., eq. (10)). For the latter, we will consider both the case where the reference signals come from the same data set on which ICA is performed, and from another one.

As an application, consider the problem of extracting the antepartum fetal ECG from multi-electrode recordings on the mother’s skin (cutaneous recordings). Hence, the purpose of ICA is to extract the subspace that separates the fetal ECG from the mother’s. The advantage over a parametric formulation of the (quasi-)periodicity of the heart rate pattern is that extrasystoles and missing heartbeats are better detected. We use the data set described in (De Lathauwer et al., 2000) (foetal_ecg.dat data set is available at
http://www.esat.kuleuven.be/sista/members/biomed/data006.htm). It consists of 10 s of an 8 channel set of cutaneous recordings (sampled @ 250 Hz) for which the first 5 channels were recorded on the mother’s abdomen and the last 3 on the thorax (Fig. 1 in De Lathauwer et al., 2000). Note that, due to the large amplitude of the mother’s ECG in the thoracic signals, the fetal’s ECG is much less visible.

We first consider the case of eq. (9) where we use as a reference signal the one that corresponds to the 6th source estimate in Fig. 3 of (De Lathauwer et al., 2000), thus, also obtained by applying an 8 channel JADE ICA in order to be able to compare our results with those found by De Lathauwer and co-workers (note that the reference signal is an independent component signal taken from the same data set eq. (9) is applied on: in this way, we can verify whether the algorithm finds the correct independent component signal). We only scale the constraint signal $z$ so that it has unit variance (since we apply prewhitening on the 8 channel set). The result is shown in Fig. 2A. The MSE between the extracted signal and the true one is 0.018 (and is therefore indistinguishable visually).

We now consider the case of two reference signals, $z_1$ and $z_2$, that are mutually independent, and we wish to extract both the fetal and the mother ECG (thus, a constrained ICA problem where the reference signals are taken from the same data set as on which ICA is applied, but we will relax that assumption later). Let $y_i$ and $y_j$ be the extracted source estimates for which we want $y_i$ to correspond to the fetal’s ECG, $z_1$, and $y_j$ to the mother’s, $z_2$. The problem formulation then becomes:

$$
\min_{W_i, W_j} (-MI(y_i, z_1) - MI(y_j, z_2) + MI(y_i, y_j)), \quad (10)
$$
with the last term added since we still want the source estimates $y_i$ and $y_j$ to be independent. (Notice the signs of the MI components.) The resulting mother’s ECG estimate is shown in Fig. 2B and is quite similar to the sources estimated with JADE; the fetal ECG estimate is visually indistinguishable from Fig. 2A and is not shown.

Note that our constrained ICA formulation is different from Lu and Rajapakse’s (which was also used by James and Gibson) who use correlation as a closeness measure with the constraint signal, and also from Liao and Carin (2002) who use quadratic constraints, instead of the more natural MI measure, as suggested here. Lu and Rajapakse solve the resulting constrained optimization problem using Lagrange optimization; Liao and Carin use a constrained maximization of the JADE criterion.

As an example with more than 2 reference signals, and where we do not assume that the reference signals are independent components of the same data set, we take a 60 s long 8 channel set of cutaneous recordings sampled at 500 Hz (haesen1.dat data set, courtesy of Lieven De Lathauwer), which we downsample to 250 Hz. This data set is much more noisy and time-variant than the previous one. We apply JADE on the first half of the downsampled signal ($N = 7500$), and take from the JADE solution three reference signal estimates, one fetal heart rate signal estimate, one mother’s, and one low frequent, noisy signal (a respiratory signal). We now apply these constraints on the second half of the data set (training/testing configuration, so no longer assuming a one-to-one relationship between source estimate and reference signal). The extra constraint has proved to be essential to extract a cleaner mother ECG estimate. (Note that the ICs of the training set still
lead to a mixture for the mother ECG signal in the test set.) The problem
to be solved is thus:

$$
\min_{W_i, W_j} \left( -MI(y_i, z_1) - MI(y_j, z_2) + MI(y_j, z_3) + MI(y_i, y_j) \right). \quad (11)
$$

The fetal signal estimate used as a constraint is shown in Fig. 2C, thus from
the first half of the data set; the extracted fetal signal estimate, thus for the
second half of the data set (compare the time scales), is shown in Fig. 2D.
We observe the correspondence.

It is interesting to note that the problem formulation of eq. (11) is actually
that of a goal programming problem: we have multiple goals to which we
strive simultaneously (Hillier and Lieberman, 2001). (Actually, a case of
nonpreemptive goal programming with upper and lower, one-sided goals.)
The difference is that the goals are now formulated in terms of min(±MI),
instead of non-negative or non-positive but otherwise linear terms, which
has, to the best of our knowledge, never been done before.

Finally, the problem formulation of eq. (10) is different from the one
recently introduced by Karhunen and Ukkonen (2006) for finding the mu-
tually corresponding dependent components of two related data sets (one is
a transformation of the other). In the latter, two ICA problems are solved,
for determining y and z, and then a singular value decomposition (SVD)
is performed to find the mutually dependent components between the two
related data sets. In our case, only one ICA problem is solved since the z are
the constraint signals, and they need not be a transformation of the source
signals. Furthermore, the constraints need not be statistically independent
either, which addresses the theoretical issue that two source estimates could
depend on the same constraint (for a discussion, see Karhunen and Ukkonen.
2006). Also, we can freely choose which constraints need to be maximally dependent, and which ones maximally independent (cf., eq. (11)).

6 Conclusion

We have introduced a new sequential, fixed-point algorithm for linear, subspace ICA based on the gradient of a robust version of the Edgeworth expansion of mutual information, and also developed a new constrained ICA version based on goal programming of mutual information objectives. We have shown the accuracy of our algorithm by comparing it with the most popular ICA algorithms. We have also tested its sensitivity to outliers, and showed that it performs comparably to robust FastICA and Infomax, and much better than JADE and MaxKurt.
Acknowledgments

The author wishes to express his gratitude to Dr. Danilo Mandic, Department of Electrical and Electronic Engineering, Imperial College of Science, Technology and Medicine, London (UK), for helpful discussions. M.M.V.H. is supported by research grants received from the Excellence Financing program of the K.U.Leuven (EF 2005), the Belgian Fund for Scientific Research – Flanders (G.0248.03 and G.0234.04), the Interuniversity Attraction Poles Programme – Belgian Science Policy (IUAP P5/04), the Flemish Regional Ministry of Education (Belgium) (GOA 2000/11), and the European Commission (NEST-2003-012963, STREP-2002-016276, and IST-2004-027017).
References


James, C.J., and Gibson, O.J. (2003). Temporally constrained ICA: an ap-
plication to artifact rejection in electromagnetic brain signal analysis.


Appendix 1: Algorithm

\[ \Delta W_i(\mathcal{Y}) = -\frac{\eta}{N} \sum_{\mu} \frac{1}{3!} \left[ \sum_{j,j<i} \left( \kappa^{i,j} \left( \frac{2}{\sigma^2 i \sigma_j} (y_i^\mu - \bar{y}_i)(y_j^\mu - \bar{y}_j) - 2 \frac{\sigma_{ij}}{\sigma^2_i \sigma_j} x^\mu \right) + \kappa^{i,j} \frac{1}{\sigma_i \sigma_j} x^\mu (y_j^\mu - \bar{y}_j)^2 - \frac{x^\mu}{\sigma_i} \right) \right] \]

\[ + \frac{1}{4!} \left[ \beta \sum_{j,j<i} \left( \kappa^{i,j,j} \frac{3}{\sigma^3 i \sigma_j} (y_i^\mu - \bar{y}_i)^3 x^\mu (y_j^\mu - \bar{y}_j) - 3 \frac{\sigma_{ij}}{\sigma_i \sigma_j} x^\mu (y_j^\mu - \bar{y}_j) - \frac{2}{\sigma^2 i} (y_i^\mu - \bar{y}_i)x^\mu \right) + \kappa^{i,j,j} \frac{1}{\sigma_i \sigma_j} x^\mu (y_j^\mu - \bar{y}_j)^3 - 3 \frac{\sigma_{ij}}{\sigma_i \sigma_j} x^\mu (y_j^\mu - \bar{y}_j) \right] \]

(12)

1 Choose \( W_i \) randomly with unit-length, \( \forall i \) /* initialization */

2 Do until convergence /* estimate first IC direction */
   2.1 calculate \( \kappa^{1,1,1} \) and \( \kappa^{1,1,1,1} \)
   2.2 \( \Delta W_1(\mathcal{Y}_1) = -\eta \frac{\partial H_1(\mathcal{Y}_1)}{\partial W_1} \)

3 Do \( i = 2 \) to \( d \) /* estimate remaining \( (d-1) \) ICs */
   Do until convergence
   3.1 calculate \( \kappa^{i,i,j}, \kappa^{i,j,j}, \kappa^{i,i,i,j}, \kappa^{i,i,j,j} \), \( \forall j < i \)
   3.2 \( \Delta W_i(\mathcal{Y}) = -\eta \frac{\partial M_i(\mathcal{Y})}{\partial W_i}, \forall j < i \) in eq. (12)

4 Do until convergence \( \forall i \) /* backfitting */
   Do until convergence
   4.1 calculate \( \kappa^{i,i,j}, \kappa^{i,j,j}, \kappa^{i,i,i,j}, \kappa^{i,i,j,j} \), \( \forall j < i \)
   4.2 \( \Delta W_i(\mathcal{Y}) = -\eta \frac{\partial M_i(\mathcal{Y})}{\partial W_i}, \) eq. (12)

In the implementation of the algorithm, to avoid local minima, we update
$W_i$ in stages 3.2 and 4.2 component-wise, for each set of $j$ terms in eq. (12).

Appendix 2: Normalized gradient

In order to speed up our algorithm, we perform normalized gradient descent, thus, by making our learning rate $\eta$ optimally adaptive.

Consider normalized gradient descent on mutual information. We start by expressing the future value of mutual information, $MI(t+1)$, by a first-order series expansion around its current value, $MI(t)$:

$$MI(t + 1) = MI(t) + \sum_i \frac{\partial MI(t)}{\partial W_i} \Delta W_i = MI(t) - \eta \sum_i \left(\frac{\partial MI(t)}{\partial W_i}\right)^2,$$

(13)

and choose $\eta$ so that the $MI(t+1)$ becomes zero, which is the desired solution:

$$\eta = \frac{MI(t)}{\sum_i (\frac{\partial MI(t)}{\partial W_i})^2} = \frac{\sum_i H_i - H}{\sum_i (\frac{\partial MI(t)}{\partial W_i})^2} \approx \frac{-\sum_i E[\log(1 + z_i(y_i))] + E[\log(1 + z(y))] \sum_i (\frac{\partial MI(t)}{\partial W_i})^2}{\sum_i (\frac{\partial MI(t)}{\partial W_i})^2},$$

(14)

since $H_i = -E[\log \phi(y_i)] - E[\log(1 + z_i(y_i))]$ and $H = -E[\log \phi(y)] - E[\log(1 + z(y))]$, and where we have assumed that $E[\log \phi(y)] = \sum_i E[\log \phi(y_i)]$ (which implies that the normalized gradient version assumes prewhitening). We then use this adaptive $\eta$ value in our gradient descent learning rule.

Furthermore, since our algorithm is a sequential one, we estimate the first independent component by minimizing its entropy, say $H_1$. We express the difference between the posterior entropy value and the maximal entropy value it can reach (i.e., the best normal estimate), as a first-order series expansion around the current entropy difference:

$$H_1(t + 1) - H_{1\text{max}} = H_1(t) - H_{1\text{max}} + \frac{\partial (H_1(t) - H_{1\text{max}})}{\partial W_i} \Delta W_i$$

$$H_1(t) - H_{1\text{max}} - \eta \left(\frac{\partial (H_1(t) - H_{1\text{max}})}{\partial W_i}\right)^2,$$

(15)
and choose \( \eta \) so that the left hand side becomes zero:

\[
\eta = - \frac{H_1(t) - H_{1\text{max}}}{(\frac{\partial (H_1(t) - H_{1\text{max}})}{\partial W_i})^2} = \frac{E[\log(1 + z_1(y_1))]}{(\frac{\partial (H_1(t) - H_{1\text{max}})}{\partial W_i})^2},
\]

(16)

since \( H_{1\text{max}} = -E[\log \phi(y_1)] \) and \( H_1(t) = -E[\log \phi(y_1)] - E[\log(1 + z_1(y_1))] \).

Evidently, the best normal estimate assumption makes this rule a heuristic one, but we remind that backfitting will improve the initial estimate.

### Appendix 3: Fixed-point iteration

The mutual information surface near the minimum value \( (MI = 0) \) is convex, so that the posterior value of mutual information, \( MI(t+1) \), can be expressed by a Taylor series expansion around \( MI(t) \) as:

\[
MI(t + 1) = MI(t) + \sum_i \frac{\partial MI(t)}{\partial W_i} \Delta W_i + \sum_i \frac{\partial^2 MI(t)}{\partial W_i^2} (\Delta W_i)^2 + \epsilon,
\]

(17)

where \( \epsilon \) denotes the contribution from the higher-order terms. Since we perform gradient descent, \( \Delta W_i = -\eta \frac{\partial MI(t)}{\partial W_i} \) – the optimization paradigm is to minimize the a posteriori estimate \( MI(t+1) \), and since we choose \( \eta \) as in eq. (14) (normalized gradient), eq. (17) becomes (ignoring \( \epsilon \)):

\[
MI(t + 1) = \sum_i \frac{\partial^2 MI(t)}{\partial W_i^2} (\Delta W_i)^2
\]

\[
= \sum_i \frac{\partial^2 MI(t)}{\partial W_i^2} \left( \frac{MI(t)}{\sum_i \frac{\partial^2 MI(t)}{\partial W_i^2}} \frac{\partial MI(t)}{\partial W_i} \right)^2 = f(MI, t).
\]

(18)

When taking the tangent to the \( MI \) function at coordinate \( (W_i(t), MI(t)) \) (Fig. 3), it crosses the \( MI = 0 \) line at the coordinate \( (W_i(t+1), 0) \), hence, since \( \Delta W_i < 0 \), \( MI(t+1) \leq MI(t) \) (note that MI is nonnegative). Furthermore \( MI(t+1) \) is bounded (it is smaller than the sum of the marginal en-
tropies of a multivariate Gaussian with unit covariance matrix, cf., prewhitening), \( f(MI, t) \) is also bounded. Hence, \(|f(MI, t + 1) - f(MI, t)| \leq L|MI(t + 1) - MI(t)|, 0 \leq L < 1\), which means that it satisfies Lipschitz continuity, and since \( f(MI, t) \) is continuous and differentiable, it is also a contractive mapping in the sense of the Contraction Mapping theorem (Mandic and Chambers, 2001). This means that for the convergence (uniform convergence for noise free signals) we may employ fixed-point iteration, which will, due to the contractivity of \( f(MI, t) \), lead to the optimal solution \( W_i^* \), as shown in Fig. 3.

**Appendix 4: Convergence speed**

To analyze the speed of convergence for the normalized gradient version of our algorithm, we reconsider the fixed-point iteration (FPI) type equation for \( MI \) eq. (18), \( MI(t + 1) = f(MI, t) \). It clearly satisfies the FPI form of \( x(t + 1) = G(x(t)) \), and will converge to a unique solution if \(|MI(t + 1) - MI(t)| < L|MI(t) - MI(t - 1)|, 0 \leq L < 1\) (see Appendix 3). This implies a uniform type of convergence, which for real world data is very difficult to achieve, but nevertheless this settling provides a rigorous and powerful framework to address convergence of the proposed algorithm. Notice also that we had initially truncated the Taylor series expansion to include only up to the second-order terms. In addition, it is actually the independent component direction estimates \( W_i, \forall i \), that are updated, and which ought to converge to the final solution. From the above equation, since \( MI \) is quadratic around \( MI = 0 \) in \( W_i \), and therefore convex, we have shown that the sequence of \( W_i \) iterates converges, and \( MI \) fill follow by continuity.
This shows that the convergence can be analyzed using FPI, however, the closed analytical expression, although possible to derive, is very difficult to interpret. Consider the Hermite polynomials in the Edgeworth expansion approximation of $MI$. We can substitute $y_i$, $\forall i$, in the Hermite polynomials and its derivatives, in the derivatives of $MI$, by $y_i = W_i^T X$, with $W_i$ the $i$th row in $W$. We can then verify that the smallest power in $MI$ is quadratic in $W_i$, for a symmetric mixture distribution, and linear in general, when assuming prewhitening, so that $\sigma_{ij} = 0$, $\forall i < j$, and given that the mixture distribution has zero-mean, so that $x_k = 0$, $\forall k$. Likewise, we conclude that the smallest power in the term $\frac{\partial^2 MI(t)}{\partial W_i^2}$ is quadratic in $W_i$, for symmetric distributions and linear in general. Furthermore, the smallest power in $W_i$ within the term $\frac{\partial MI(t)}{\partial W_i}$ is always linear. Hence, intuitively, convergence of $MI$ is approximately quadruple in $W_i$ for symmetric distributions and linear in general. It is therefore clear that the convergence of $W_i$ is approximately quadratic for symmetric distributions and linear in general.

As an illustrative example, we consider two Laplacian-distributed source signals (thus, with unit mixing matrix), and initialize the independent component directions so that they have an error angle of 30 deg. The median error angle as a function of the number of iterations (backfitting runs, thus, for an update of both independent component directions) is shown in Fig. 4 (solid line) for 20 data sets of size $N = 10000$. We also show the result for two exponentially-distributed source signals (dashed line). We observe that the convergence of the symmetric case is about twice as fast as that of the asymmetric case, and that the latter converges roughly linearly (the slope of the first update is 1.16, on a log-scale).
Figure 1: (A) The effect of the number of backfitting runs in the Edgeworth expansion-based algorithm (thin solid line) and its robust version (thick solid line) on the maximum difference angle (error angle) between the theoretical and obtained independent component directions for the case of two...
Figure 1, Cont’d: orthogonal independent component directions with Laplacian source amplitude distributions (supergaussian case). Plotted are the median error angles for 20 data sets of size $N = 10000$; the vertical bars denote the first and third quartiles. (B) Median error angles (and quartiles) plotted as a function of the angle $\theta$ between two independent component directions in the XY-plane (from $\theta = 5$ to 90 deg in steps of 5 deg), for square mixtures of 5 Laplacian source distributions. Shown are the results for the (non-robust) Edgeworth expansion-based algorithm (stippled line), without prewhitening, the robust version of the Edgeworth expansion-based algorithm without prewhitening (thin solid line) and with prewhitening (thick solid line). (C) Idem to (B) but now plotted for $d = 2$ and 5 Laplacian source distributions (thin and thick lines, respectively) and all assuming prewhitening. Shown are the results for FastICA (dashed lines), JADE (dotted lines), MaxKurt (dot-dot-dashed lines), Infomax (stippled lines) and the robust version of the Edgeworth expansion-based algorithm (solid lines). (D) Median error angles (and quartiles) plotted as a function of $d$ random mixings of Laplacian source amplitude distributions, with and without outliers (thick vs. thin lines), for the (non-robust) Edgeworth expansion-based algorithm (stippled lines) and its robust version (solid lines). (E) Idem to (D) but shown for FastICA, JADE, MaxKurt, Infomax and the robust version of the Edgeworth expansion-based algorithm. Same line conventions as in (C). (F) Dependency on the data set size $N$ for $d = 2$ (thin lines) and 5 (thick lines) using random mixings of Laplacians for the algorithms listed under (E). Same line conventions as in (C).
Figure 2: (A) Fetal ECG estimate obtained using as a constraint a single fetal source estimate of the foetal_ecg.dat data set. (B) Mother ECG estimate obtained using fetal and mother source estimates as the two constraints; the fetal ECG estimate is very similar to (A). (C) Fetal source estimate for the first half of the haesen1.dat data set. (D) Fetal ECG estimate for the second half of the data set and obtained using three constraints (see text).
Figure 3: Graphical display of $MI$ minimization with normalized gradient. Stippled line indicates the tangent, $W_i^*$ represents the optimal weight value.
Figure 4: Speed of convergence. Error angle as a function of the number of runs for 2 Laplacian- (solid line) and 2 exponentially-distributed source signals (dashed line) with unit mixing matrix. The initial error angle is 30 deg. Plotted are the median error angles for 20 data sets sized $N = 10000$; the vertical bars denote the first and third quartiles.