Application of Mutual Information based Least dependent Component Analysis (MILCA) for Removal of Ocular Artifacts from Electroencephalogram

V Krishnaveni, S Jayaraman, and K Ramadoss

Abstract—The electrical potentials generated during eye movements and blinks are one of the main sources of artifacts in ElectroEncephaloGram (EEG) recording and can propagate much across the scalp, masking and distorting brain signals. In recent times, signal separation algorithms are used widely for removing artifacts from the observed EEG data. In this paper, a recently introduced signal separation algorithm Mutual Information based Least dependent Component Analysis (MILCA) is employed to separate ocular artifacts from EEG. The aim of MILCA is to minimize the Mutual Information (MI) between the independent components (estimated sources) under a pure rotation. Performance of this algorithm is compared with eleven popular algorithms (Infomax, Extended Infomax, Fast ICA, SOBI, TDSEP, JADE, OGWE, MS-ICA, SHIBBS, Kernel-ICA, and RADICAL) for the actual independence and uniqueness of the estimated source components obtained for different sets of EEG data with ocular artifacts by using a reliable MI Estimator. Results show that MILCA is best in separating the ocular artifacts and EEG and is recommended for further analysis.

Keywords—Electroencephalogram, Ocular Artifacts (OA), Independent Component Analysis (ICA), Mutual Information (MI), Mutual Information based Least dependent Component Analysis (MILCA)

I. INTRODUCTION

ELECTROENCEPHALOGRAM is a recording of electric fields of signals emerging from neural currents within the brain and is measured by placing electrodes on the scalp. The electrical dipoles of eyes change by eye movements and blinks, producing a signal known as electrooculogram (EOG). A fraction of EOGs contaminate the electrical activity of the brain and these contaminating potentials are commonly referred to as ocular artifacts (OA). In current data acquisition, these OA are often dominant over other electrophysiological contaminating signals (e.g. heart and muscle activity, head and body movements), as well as external interferences due to power sources. Hence, devising a method for successful removal of OA from EEG recordings is still is a major challenge. Fig 1 shows EEG signals corrupted with ocular artifacts. Since ocular artifacts decrease rapidly with the distance from the eyes, the most severe interference occurs in the electrodes placed on the patient's forehead. Notice the large dips on frontal channels FP1-F3, FP2-F4, FP1-FP7 and FP2-F8. Blink artifacts are so prominent on these channels because they are located nearest to the eyes.

A variety of methods have been proposed for correcting ocular artifacts and are reviewed in [1],[2]. One common strategy is artifact rejection. The rejection of epochs contaminated with OA is very laborious and time consuming and often result in considerable loss in the amount of data available for analysis. Eye fixation method in which the subject is asked to close their eyes or fix it on a target is often unrealistic. Widely used methods for removing OAs are based on regression in time domain [3] or frequency domain [4] techniques. All regression methods, whether in time or frequency domain depend on having one or more regressing (EOG) channels. Also both these methods share an inherent weakness, that spread of excitation from eye movements and EEG signal is bidirectional. Therefore regression based artifact removal eliminates the neural potentials common to reference electrodes and to other frontal electrodes.

Another class of methods is based on a linear decomposition of the EEG and EOG leads into source components, identifying artifactual components, and then reconstructing the EEG without the artifactual components. Lagerlund et.al [5] used Principal Component Analysis (PCA) [6] to remove the artifacts from EEG. It outperformed the regression based methods. However, PCA cannot completely separate OA from EEG, when both the waveforms have similar voltage magnitudes. PCA decomposes the leads into uncorrelated, but not necessarily independent components that are spatially orthogonal and thus it cannot deal with higher order statistical dependencies.

V Krishnaveni is with Department of Electronics & Communication Engineering, PSG College of Technology, Coimbatore – 641 004 India as a Senior Lecturer (corresponding author e-mail: venimurthy@hotmail.com).

S Jayaraman is with Department of Electronics &Communication Engineering, PSG College of Technology, and Coimbatore – 641 004 India as Professor and Head (e-mail: jayaramathreya@yahoo. Com)

K Ramadoss is with PSG Institute of Medical Sciences and Research, Coimbatore - 641 004 India as an Associate Professor / Consultant Neurologist.
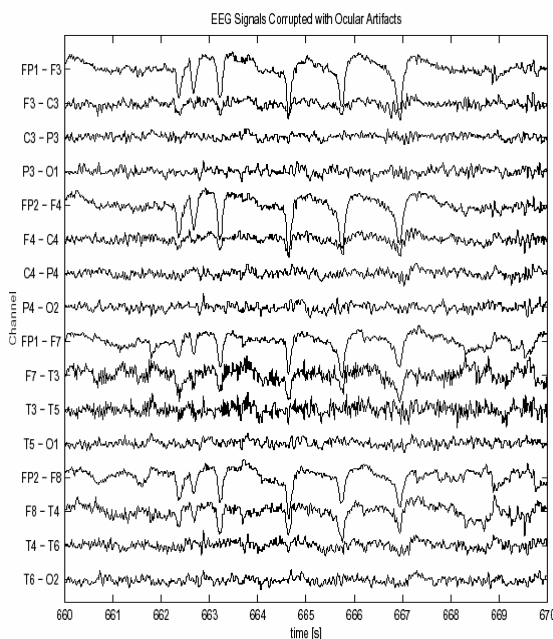
Fig.1 EEG recording corrupted by ocular-artifacts

An alternative approach is to use Independent Component Analysis (ICA), which was developed in the context of blind source separation problems to obtain components that are approximately independent [7]. ICA has been used to correct for ocular artifacts, as well as artifacts generated by other sources [8], [9], [10]. ICA is an extension of PCA which not only decorrelates but can also deal with higher order statistical dependencies. However, the ICA components lack the important variance maximization property possessed by the PCA components. ICA algorithms are superior to PCA, in removing a wide variety of artifacts from the EEG, even in the case of comparable amplitudes. The component based procedures used for artifact removal [5], [8], [9], [10] are not automated, and require visual inspection to select the artifactual components to decide their removal. An ICA based method for removing artifacts semi automatically was presented by Delorme et.al [11]. It is automated to flag trials as potentially contaminated, but these trials are still examined and rejected manually via a graphical interface. Carrie Joyce et.al [12] used SOBI algorithm along with correlation metrics and Nicolaou et.al [13] used TDSEP along with Support Vector Machine (SVM) for automatic removal of artifacts. The results of these studies does not imply that SOBI/TDSEP is the overall best approach for decomposing EEG sensor data into meaningful components, and has not been completely validated by the authors.

The estimated source signals (obtained from any ICA algorithm) should be as independent as possible (or least dependent on each other) for better removal of artifacts from EEG. Since, either by visual inspection, or by automated procedure, only the estimated sources are classified as EEG or artifacts. But, the actual independence of the components (estimated sources) obtained from ICA/BSS algorithms used in [8], [9], [10], [11], [12], [13] are not tested for their independence and uniqueness.

In this paper, a recently introduced signal separation algorithm Mutual Information based Least dependent Component Analysis (MILCA) [14] is employed to separate ocular artifacts from EEG. The aim of MILCA is to minimize the Mutual Information (MI) between the independent components (estimated sources) under a pure rotation. Performance of this algorithm is compared with eleven popular algorithms (Infomax [15], Extended Infomax [16], Fast ICA [17], SOBI [18], TDSEP [19], JADE [20], OGWE [21], MS-ICA [22], SHIBBS [20], Kernel-ICA [23], and RADICAL [24]) for the actual independence and uniqueness of the estimated source components obtained for different sets of EEG data with ocular artifacts by using a reliable MI Estimator [25]. Results show that MILCA is best in separating the ocular artifacts and EEG and is recommended for further analysis.

The paper is organized as follows: Theoretical review of the ICA/BSS algorithms used for removing ocular artifacts are discussed in section 2. In section 3, the algorithm for estimating mutual information is given. Results are discussed in Section 4 and the paper is concluded in Section 5.

II. INDEPENDENT COMPONENT ANALYSIS & BLIND SOURCE SEPARATION

Independent Component Analysis (ICA) [7] is a novel statistical technique that aims at finding linear projections of the data that maximize their mutual independence. ICA has received attention because of its potential applications in signal processing such as in feature extraction, and blind source separation (BSS) with special emphasis to physiological data analysis and audio signal processing. The goal of BSS is to recover the source signals given only sensor observations that are linear mixtures of independent source signals. ICA is a statistical technique for obtaining independent sources, S from their linear mixtures, X, when neither the original sources nor the actual mixing, A are known.

$$\begin{bmatrix} s_1(t) \\ \vdots \\ s_n(t) \end{bmatrix} = \mathbf{s} \rightarrow \boxed{A} \rightarrow \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} = \mathbf{x} \rightarrow \boxed{W} \rightarrow \begin{bmatrix} \hat{s}_1(t) \\ \vdots \\ \hat{s}_n(t) \end{bmatrix} = \hat{s}$$
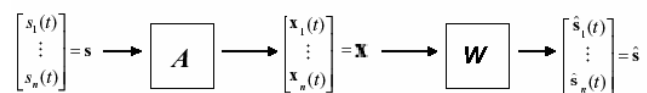
Fig. 2 Basic BSS model. Unobserved signals: S, Observations: X,

Estimated source signals: $\hat{S}$

This is achieved by exploiting higher order signal statistics and optimization techniques. The result of the separation process is a demixing matrix W, which can be used to obtain the estimated unknown sources, $\widehat{S}$ from their mixtures. This process is described by Equation 1 and a schematic illustration

of the mathematical model in shown in Fig 2.

$$X = AS \longrightarrow \widehat{S} = WX \qquad (1)$$

III. THEORETICAL REVIEW OF THE ICA/BSS ALGORITHMS

A brief description of various algorithms used in this paper is given below.

A. Infomax and Extended Infomax Algorithm

Bell and Sejnowski [15] have proposed an adaptive learning algorithm that blindly separates mixtures, X of independent sources, S using information maximization (infomax) and is described by the following steps:

i) The demixing matrix W is initialized to an identity matrix.

ii) The signal sources are estimated by equation (1) and then they are transformed by a nonlinear transfer function. For a sigmoidal transfer function, the resulting signals Y are expressed as

$$Y = g(\widehat{S}) = 1/1 + e^{-(\widehat{S}+w_o)} \qquad (2)$$

where w_o is a vector of bias weights which is initialized to a zero vector.

iii) The nonlinearly transformed signals Y are processed by a learning rule which maximizes their joint entropy that can approximately minimize their mutual information. This is achieved by changing the weight matrix by an amount ΔW, where

$$\Delta W = [W^T]^{-1} + (1 - 2y)x^T \qquad (3)$$

The change in the bias weight is expressed by

$$\Delta w_o = 1 - 2y \qquad (4)$$

iv) The ICA algorithm is trained by repeating steps (ii) and (iii). After each iteration, the demixing matrix W is updated by ΔW until convergence is achieved.

The algorithm stops training when the rate of change falls below a predefined small value, e.g 1.0×10^{-6}. The rate of change is computed by squaring the difference between corresponding elements of the demixing matrix before and after each iteration and then summing the values.

The algorithm of Bell and Sejnowski [15] which uses a sigmoidal activation function is specifically suited to separate signals with super-Gaussian distribution (i.e. positive kurtosis). Lee and Sejnowski [16] proposed an extension of ICA that is able to separate with sub and as well as super Gaussian distribution. This preserves the ICA architecture of Infomax algorithm [15], but it uses a learning rule derived by Girolami and Fyfe [26]. It determines the sign changes (positive to negative and vice versa) required by the algorithm to handle both sub and super Gaussian distributions. This is achieved by considering the normalized fourth-order kurtosis of the estimated signal sources. In extended ICA, the amount of change ΔW required to update the demixing weight matrix W is given by

$$\Delta W = (\partial H(Y)/\partial W) \times W^T W$$
$$= [I - sign(k_4)(1 - 2y)\widehat{s}^T - \widehat{s}\widehat{s}^T]W \qquad (5)$$

where $W^T W$ is the "natural gradient" of Amari et.al [27] used for speeding up the convergence.

B. Fast ICA Algorithm

Fast ICA is based on a fixed-point iteration scheme [17]. According to the central limit theorem sum of two independent random variables usually has a distribution that is closer to gaussian than of the two original random variables. Thus, maximizing the nongaussianity yields independent components. Approximation of negentropy is used for measuring the nongaussianity. The operation of Fast ICA is outlined as follows:

i) The mean of the mixed signal X is subtracted so as to make X as a zero mean signal $X = X - E[X]$, where $E[X]$ is the mean of the signal.

ii) Covariance matrix $R = E[XX^T]$ is obtained and eigen value decomposition is performed on it and is given by $R = EDE^T$ where E is the orthonormal matrix of eigen vectors of R and D is the diagonal matrix of eigen values. Find the whitening matrix WM which transforms the covariance matrix into an identity matrix.

$$WM = Inv(sqrt(D) \times E^T) \qquad (6)$$

iii) Choose an initial weight vector w. Find a direction, i.e. a unit vector w such that the projection $w^T x$ maximizes nongaussianity.

$$w^+ = E\{xg(w^T x)\} - E\{g'(w^T x)\}w \qquad (7)$$

where g is the derivative of the nonquadratic function.

$$g_1(u) = \tanh(a_1 u), \; g_2(u) = u \exp(-u^2 / 2) \qquad (8)$$

where $1 \le a_1 \le 2$

iv) The variance of $w^{+T} x$ must be made unity. Since x is already whitened it is sufficient to constrain the norm of w^+ to be unity.

$$w = w^+ / \|w^+\| \qquad (9)$$

If w not converges means (i.e. the old and new values of w does not point to the same direction) go back to step (iv).

v) The demixing matrix is given by $W = W^T \times WM$ and Independent components are obtained by $\widehat{S} = W \times X$

C. SOBI Algorithm

SOBI algorithm exploits joint-diagonalization of time delayed second order correlation matrices [18]. The operation of SOBI is given below:

i) Step (ii) discussed in FastICA algorithm is performed and the Whitening Matrix WM is obtained.

$$WM = Inv(sqrt(D) \times E^T) \qquad (10)$$

ii) Obtain the time delayed cross correlation matrix of the mixed signals.

$$R_\tau(x) = \langle x(\tau)x^T(t+\tau) \rangle \qquad (11)$$

where $\langle .... \rangle$ denotes the time average and τ is the certain time lag.

iii) The diagonal elements of this matrix are formed by the values of the autocorrelation functions and the off diagonal elements are the respective cross correlations.

$$R_\tau(x) = \begin{bmatrix} \phi_{x_1,x_2}(\tau)..............\phi_{x_1,x_n}(\tau) \\ \phi_{x_2,x_1}(\tau)..............\phi_{x_2,x_n}(\tau) \\ ... \\ ... \\ ... \\ \phi_{x_n,x_1}(\tau)..............\phi_{x_n,x_n}(\tau) \end{bmatrix} \qquad (12)$$

where ϕ denotes the correlation function.

iv) Symmetrize the time delayed cross correlation matrix $R_\tau(x)$

$$\widetilde{R}_\tau(x) = (R_\tau(x) + R^T_\tau(x))/2 \qquad (13)$$

v) It is obvious that, if the signals are independent over time then all time delayed correlation matrices should be diagonal because the cross correlations of independent signals will vanish. Hence obtain the matrix U by joint diagonalizing the set of p correlation matrices $\{\widetilde{R}_{\tau_i}(x)/_{i=1,2......p}\}$

$$\widetilde{R}_\tau(x) = U.R_\tau(s).U^T \qquad (14)$$

vi) Then the demixing matrix is given by $W = U^T \times WM$ where WM is the whitening matrix. Independent Components are obtained by $\widehat{S} = W \times X$.

D. TDSEP Algorithm

Temporal Decorrelation source SEParation (TDSEP) [19] is proposed by Andreas Ziehe and Klaus miller and it employs first whitening step and then an approximate simultaneous diagonization of several time delayed second order correlation matrices. The operation of TDSEP is outlined as follows:

i) Step (ii) discussed in FastICA algorithm is performed and the Whitening Matrix WM is obtained.

$$WM = Inv(sqrt(D) \times E^T) \qquad (15)$$

ii) Minimize the generalized cost function.

$$l(c_{i,j}) = \sum_{i \neq j} \langle x_i(\tau)x_j(t+\tau) \rangle^2 + \sum_{k=1}^{N} \sum_{i \neq j} \langle x_i(\tau)x_j(t+\tau_k) \rangle^2 \qquad (16)$$

After whitening the first term in the cost function becomes zero explicitly. The cost function then can be minimized by approximate simultaneous diagonalization of several correlation matrices through several JACOBI rotations.

$$R_{\tau_k}(x) = \langle x(\tau)x^T(t+\tau_k) \rangle \qquad (17)$$

iii) Obtain the rotation matrix Q by a sequence of elementary rotations each trying to minimize the off diagonal elements of the respective correlation matrices $R_{\tau_k}(x)$.

iv) Demixing matrix is given by $W = Q^T \times WM$ and Independent Components are given by $\widehat{S} = W \times X$

E. JADE Algorithm

Yet another signal source separation technique is the Joint Approximation Diagonalisation of Eigen matrices (JADE) algorithm [20]. This exploits the fourth order moments in order to separate the source signals from mixed signals. The operation of JADE follows as given below:

i) Step (ii) discussed in FastICA algorithm is performed and the Whitening Matrix WM is obtained.

$$WM = Inv(sqrt(D) \times E^T) \qquad (18)$$

ii) The fourth cumulants of the whitened mixtures are computed. Their n most significant eigen values λ_i and their corresponding eigen matrices M_i are determined. An estimate of the unitary matrix $\hat{V}$ is obtained by maximizing the criteria $N = \lambda_i M_i$ by means of joint diagonalisation. If N cannot be exactly jointly diagonalised, the maximization of the criteria defines a joint approximate diagonalisation.

iii) An estimate of the demixing matrix is obtained by $W = WM \times \hat{V}$ and Independent Components are given by $\widehat{S} = W \times X$.

F. OGWE Algorithm:

In OGWE (Optimized Generalized Weighted Estimator) [21], the marginal entropy contrast function (Φ^{ME}) is written in terms of second-order and fourth-order cumulants, and then it is minimized for all possible distributions for the sources S [28], it follows that

$$\varphi^{ME}(\mathbf{Y}) \approx \frac{1}{48}\varphi_{24}^{ME}(\mathbf{Y}) = -\frac{1}{48}\sum_i (C_{iiii}^{\mathbf{Y}})^2 \qquad (19)$$

where, for zero-mean signals, $C_{iiii}^Y = E[Y_i^4] - 3E[Y_i^2]$ are the marginal cumulants or autocumulants [29].

In the two dimensional case, the pair of normalized sources

$s_l = [\bar{s}_p(t)\bar{s}_q(t)]^T$ in polar coordinates may be written as $(r(t), \alpha(t))$ so that the outputs yield

$$\begin{pmatrix} Y_p(t) \\ Y_q(t) \end{pmatrix} = \mathbf{R}(\theta) \begin{pmatrix} r(t)\cos(\beta(t)) \\ r(t)\sin(\beta(t)) \end{pmatrix} = \mathbf{R}(\theta)Z_t \qquad (20)$$

where $Z_t = [Z_p(t)\ Z_q(t)]^T$ are the whitened mixtures, and matrix $\mathbf{E}$ performs a rotation of θ so that $\rho(t) = \theta + \beta(t)$ is the angle of vector $\mathbf{y}$. Note that ideally, at separation $\theta + \beta(t) = \alpha(t)$.

(i) The whitening matrix WM is computed as given in step (ii) of FastICA Algorithm to whiten the vector X and the vector Y = WM x X is formed.

(ii) One Sweep. For all g = m (m-1)/2 pairs, i.e., for $1 \leq p \leq q \leq m$, the following steps have to be done:

(a) The Given angle $\theta_{pq} = \theta_{GWE}$ is computed, (with $[z_p\ z_q]^T = [y_p\ y_q]^T$) as follows:

$$\hat{\theta}_{GWE}(\omega_r, \omega_\xi) = \frac{1}{4} \angle (\omega_\xi \omega_r \xi_r + (1 - \omega_\xi)\xi_\eta), \\ 0 < \omega_\xi < 1, \omega_r \pm 1, \gamma \qquad (21)$$

$$\hat{\theta}_{SICA} = \hat{\theta}_{GWE}(\gamma, \frac{3}{7}) \qquad (22)$$

where $\angle(.)$ supplies the principal value of the argument.

$$\xi_r = E[r^4(t)e^{j4\beta(t)}]$$
$$\xi_\eta = E^2[r^4(t)e^{j2\beta(t)}] \qquad (23)$$
$$\gamma = E[r^4(t)] - 8$$

(b) If $\theta_{pq} > \theta_{min}$, the pair (Z_p, Z_q) is rotated by θ_{pq} according to Equation (19) and also the rotation matrix $\mathbf{R}$ is updated. The value of θ_{min} is selected in such a way that rotations by a smaller angle are not statistically significant. Typically $\theta_{min} = 10^{-2}/\sqrt{N}$ where N is the number of samples.

(iii) End if the number of iterations n_{it} satisfies $n_{it} \geq 1 + \sqrt{M}$ or no angle θ_{pq} has been updated, stop. Otherwise go to step(ii) for another sweep.

(iv) Then the demixing matrix $\mathbf{W} = \mathbf{R}$ x WM and the independent sources are estimated as $\hat{S} = W \times X$

G. ICA-MS Algorithm

Molgedey and Schuster [22] proposed an approach based on dynamic decorrelation which can be used if the independent source signals have different autocorrelation functions. The main advantage of this approach is that the solution is simple and constructive, and can be implemented in a fashion that requires the minimal user intervention (parameter tuning).

Let $\mathbf{X}_\tau$ be the time shifted version of the mixed vector $\mathbf{X}$. The delayed correlation approach is based on solving the simultaneous eigenvalue problem for the correlation matrices $\mathbf{X}_\tau \mathbf{X}^T$ and $\mathbf{X}\mathbf{X}^T$ [21]. This is implemented by solving the

eigenvalue problem for the quotient matrix $\mathbf{Q} \equiv \mathbf{X}_\tau \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}$. From, $\mathbf{X}\mathbf{X}^T = \mathbf{A}\mathbf{S}\mathbf{S}^T\mathbf{A}^T$ and $\mathbf{X}_\tau \mathbf{X}^T = \mathbf{A}\mathbf{S}_\tau \mathbf{S}^T\mathbf{A}^T$ are obtained.

If the sources furthermore are independent, the diagonal source cross-correlation matrix is obtained at lag zero in the limit $\lim_{N\to\infty} N^{-1}\mathbf{S}\mathbf{S}^T = \mathbf{C(0)}$. Similarly, $\lim_{N\to\infty} N^{-1}\mathbf{S}_\tau \mathbf{S}^T = \mathbf{C(\tau)}$ produces the diagonal cross correlation matrix at lag τ. Hence, to zero[th] order in 1/N,

$$\mathbf{X}_\tau \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \approx \mathbf{A}\mathbf{C(\tau)}\mathbf{A}^T(\mathbf{A}^T)^{-1}\mathbf{C(0)}^{-1}\mathbf{A}^{-1} \qquad (24)$$

with $\mathbf{C(\tau)}\mathbf{C(0)}^{-1}$ being a diagonal matrix. If the eigenvalue problem is solved for the quotient matrix [30].

$$\mathbf{Q} \equiv \mathbf{X}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1} \approx \mathbf{A}\mathbf{C(\tau)}\mathbf{C(0)}^{-1}\mathbf{A}^{-1} \qquad (25)$$

then the direct scheme is obtained to estimate $\mathbf{A}, \mathbf{S}$. Let

$$\mathbf{Q}\Phi = \Lambda \Phi \qquad (26)$$

and $\Phi = \mathbf{A}$ and $\Lambda = \mathbf{C(\tau)}\mathbf{C(0)}^{-1}$ up to scaling factors are identified.

Then the demixing matrix $\mathbf{W}$ is the inverse of the mixing matrix $\mathbf{A}$. The sources can be estimated as $\hat{S} = W \times X$

H. SHIBBS Algorithm:

Another signal separation technique is Shifted Block Blind Seperation (SHIBBS) [20] to estimate the demixing matrix $\mathbf{W}$.

(i) A fixed set X = {X_1, . . . ,X_m} of $m \times n$ matrices is selected. A Whitening matrix $\mathbf{WM}$ and set $Z = WM \times X$ are estimated.

(ii) The set $\{\hat{\mathbf{Q}}^Z(\mathbf{X}_m)|1 \leq p \leq M\}$ of M cumulant matrices is estimated and a joint diagonalizer $\mathbf{R}$ of it is found.

(iii) If $\mathbf{R}$ is close enough to the identity transform, stop. Otherwise, the data is rotated using the equation $\mathbf{Z} = \mathbf{R}^T\mathbf{Z}$ and step (ii) is repeated.

(iv) Then the demixing matrix $W = R \times WM$ is used to estimate the independent components $\hat{S} = W \times X$

The SHIBBS algorithm is implemented in the same way as JADE is done. But the joint diagonalization of the significant eigen matrices is done without going through the estimation of the whole cumulant set and through the computation of its eigen-matrices.

I. Kernel-ICA Algorithm:

The Kernel-ICA algorithm [23] uses the contrast functions based on Canonical Correlation Analysis (CCA) [31] in a Reproducing Hilbert Kernel Space (RKHS). The outline of Kernel Canonical Correlation Analysis (KCCA) is given as follows:

(i) Let $x_1, x_2, \ldots, x_m$ be the data vectors and $\mathbf{K}(x_i, x_j)$ be the kernel.

(ii) Data is whitened and the whitening matrix WM is obtained.

(iii) The contrast function C(**W**) is minimized with respect to **W**.

(iv) The contract function is minimized in the following way:

a) The centered Gram matrices [23] $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_m$ of the estimated sources $\{y_1, y_2, \ldots, y_m\}$, where $y^i = Wx^i$ are computed.

b) The minimal eigenvalue of the generalized eigenvector equation $\hat{\lambda}_F^\kappa(K_1, \ldots, K_m)$ is defined as

$$K_\kappa \alpha = \lambda D_\kappa \alpha \qquad (27)$$

c) Then $C(\mathbf{W}) = \hat{I}_{\lambda_F}(\mathbf{K}_1, \ldots, \mathbf{K}_m) = -\frac{1}{2}\log \hat{\lambda}_F^\kappa(\mathbf{K}_1, \ldots, \mathbf{K}_m)$

(v) The demixing matrix **W**, is then formed $W = W \times WM$ and the independent components are estimated by

$$\hat{S} = W \times X$$

J. RADICAL Algorithm:

The RADICAL (Robust, Accurate, Direct Independent Component Analysis Algorithm) [24] estimates the independent sources using differential entropy estimator based on 'm'-spacing estimator. The contrast function in equation (28) is to be minimized by RADICAL is almost equivalent to Vasicek estimator [32],

$$\hat{H}_{RADICAL}(Z^1, \ldots, Z^N) \equiv \frac{1}{N-m}\sum_{i=1}^{N-m}\log(\frac{N+1}{m}(Z^{(i+m)} - Z^{(i)})) \qquad (28)$$

The data vectors X1, X2, ….., XM are whitened. Let m be the size of spacing. The value of 'm' is taken as $\sqrt{N}$ where N is the number of samples in each source.

Let $\Re$ be the number of replicated points per original data point to eliminate the local minima problem [24]. Let σ be the standard deviation of replicated points. For N < 1000, σ = 0.35 and for N ≥ 1000, σ = 0.175, where N is the number of samples in each source before replication. Let K be the number of angles for which cost function has to be evaluated. The optimum value of K here is 350.

(i) For each of M-1 sweeps (or until convergence), where M is the number of sources.

(ii) For each of M(M-1)/2 jacobi rotations for dimensions (p,q).

(a) A pair of whitened mixture is taken (Z_p, Z_q).

(b) Create Z' by replicating $\Re$ points with Gaussian noise for each original point.

(c) For each θ in K number of angles, the augmented data are rotated to this angle

$$\mathbf{Y} = \mathbf{R}(\theta) \times \mathbf{Z'} \qquad (29)$$

and the contrast function is evaluated.

(d) The Jacobian matrix for the optimal θ is formed and it is incorporated into the rotation matrix **R**. The optimal θ is one which yields the minimum vasicek estimator value [32] for the rotated pair.

(iii) The final rotational matrix **R** is the accumulation of all the jacobi rotations of optimal θ.

(iv) The demixing matrix $W = R \times WM$ and the estimated sources $\hat{S} = W \times X$ are obtained.

K. MILCA Algorithm:

As described in section II Independent component analysis (ICA) is a statistical method for transforming an observed multi-component data set $X(t) = (x_1(t), x_2(t), \ldots, x_n(t))$ into components that are statistically as independent from each other as possible. In theoretical analyses, certain model for the data is assumed, for which the decomposition into completely independent components is possible, but in real life applications the latter will not be true i.e. least dependent components are only possible. Depending on the assumed structure of the data, a parameterized guess is made about how they can be decomposed (linearly or not, using only equal times or using also delayed superpositions, etc.) and then fixes the parameters by minimizing some similarity measure between the output components. Using mutual information (MI) would be the most natural way to solve this problem. This idea leads to the new signal separation algorithm Mutual Information based Least Dependent Component Analysis (MILCA) [14] based on k^{th} closest neighbour statistics. The aim of MILCA algorithm is to minimize $I(X_1 \ldots X_N)$ under a pure rotation R. Any rotation can be represented as a product of rotations which act only in some 2 × 2 subspace,

$$R = \prod_{ij} R_{ij}(\phi) \qquad (30)$$

where $R_{ij}(\phi)(x_1 \ldots x_i \ldots x_j \ldots x_n) = (x_1 \ldots x_i' \ldots x_j' \ldots x_n')$

with $x_i' = \cos\phi x_i + \sin\phi x_j$ and $x_i' = \cos\phi x_i + \sin\phi x_j$.

For such a rotation using grouping property of MI

$$I(R_{ij}(\phi)X) - I(X) = I(X_i', X_j') - I(X_i, X_j) \qquad (31)$$

i.e., the change of $I(X_1 \ldots X_N)$ under any rotation can be computed by adding up changes of two-variable MIs. To find the optimal angle ϕ in a given (i, j) plane, $\hat{I}_{ij}(\phi) = \hat{I}(X_i', X_j')$ is calculated for typically 150 different angles in the interval $[0, \pi/2]$ and these values are fitted by typically 3-15 Fourier components, and then the minimum of the fit is taken.

The resulting MILCA-algorithm can be summarized:

i) Whitening Matrix WM is obtained.

ii) For each pair (i, j) with $i, j = 1 \ldots n$ find the angle ϕ which minimizes a smooth fit to $\hat{I}_{ij}(\phi) = \hat{I}(X_i', X_j')$.

iii) If $\hat{I}(X_1^{'}....X_n^{'})$ has not yet converged, go back to step (ii), else an estimate of the demixing matrix is obtained by $W = R \times WM$ and Independent Components are given by $\hat{S} = W \times X$.

In this algorithm, k^{th} closest neighbour should be selected and the value of k value must be chosen properly. Depending on the k value the algorithm minimizes either the statistical errors or the systematic errors. The higher the value of k , the lower is the statistical error of $\hat{I}$ .The systematic error shows exactly the opposite behavior. Thus, to keep the balance between these two errors, the best choice for k would lie in the middle range. But for some cases this may deviate, e.g. finding the most independent signal sources. In this case the true values of the MI are small and thus also the systematic errors for all k so it is better to use large k in order to reduce statistical errors, but too large values of k should be avoided since then the increase of systematic errors outweighs the decrease of statistical ones. On the other hand, when the data files are very long there is no need to worry about statistical errors so it is better to choose small k . In this application k value is taken as 14.

IV. ESTIMATING MUTUAL INFORMATION

If X and Y are two random variables with joint distribution $\mu(x, y)$ and marginal distributions $\mu_x(x)$ and $\mu_y(y)$, then Mutual Information $I(X,Y)$ between X and Y is defined as

$$I(X,Y) = \iint \mu(x, y) \times \log(\mu(x, y) / \mu_x(x)\mu_y(y))dxdy \qquad (32)$$

The algorithm proposed by Kraskov et.al [25] estimates $I(X,Y)$ from the set $\{Z_i\}$ alone without explicit estimation of the unknown densities and is outlined below.

For any set of N bivariate measurements $z_i = (x_i, y_i)$, the kth closest neighbour of each Z_i is found according to the metric

$$\|z - z'\| = \max\{\|x - x'\|, \|y - y'\|\} \qquad (33)$$

The kth nearest neighbor is then projected onto the X and Y axes giving the distances $\varepsilon_x(i)/2$ and $\varepsilon_y(i)/2$. The estimate for MI is given by

$$\hat{I}(x, y) = \psi(k) - (1/k) - \langle \psi(n_x) + \psi(n_y) \rangle + \psi(N) \quad (34)$$

Where $n_x(i)$ and $n_y(i)$ be the number of points with $\|x_i - x_j\| \le \varepsilon_x(i)/2$ and $\|y_i - y_j\| \le \varepsilon_y(i)/2$ and $\psi(.)$ is

the digamma function $\psi(x) = \Gamma(x^{-1}) \times (d\Gamma(x)/dx)$ and $\langle .... \rangle = N^{-1} \sum_{i=1}^{N} E[.......(i)]$.

V. RESULTS AND DISCUSSIONS

ICA is a statistical method for transforming an observed multi-component data set into independent components that are statistically as independent as possible. The components estimated by an ICA algorithm should be least dependent on each other, for better removal of artifacts from EEG and so the actual dependencies between the obtained components is to be estimated and it is most often ignored. Hence it becomes necessary to estimate the actual dependencies between the components and to find the best ICA algorithm that transforms the observed data set into components that are least dependent on each other. There are various measures to evaluate the independence among the estimated sources. Some of the measures are kurtosis, negentropy, Mutual Information, etc [33]. Kurtosis is the fourth-order cumulant. In terms of robustness and asymptotic variance, the cumulant based estimator tend to be far from optimal. Intuitively there are two main reasons for this. Firstly, higher order cumulants measure mainly the tails of the distributions, and are largely unaffected by structure in the middle of the distribution. Secondly, estimators of the higher order cumulants are highly sensitive to outliers [33]. Their value may depend on only a few observations in the tails of distribution which may be outliers. Negentropy involves estimation of probability density function which is very difficult. Cumulant-based approximations of negentropy are inaccurate and in many cases too sensitive to outliers. Among these measures, Mutual Information (MI) is the best choice to measure the independence of the estimated sources. However, MI was not extensively used for measuring interdependence because estimating MI from statistical samples is not easy. In the ICA literature very crude approximations to MI based on cumulant expansions are popular because of their ease of use. In this paper, an efficient methodology to estimate MI [25] based on k-nearest neighbour distances without estimating the probability densities is used to assess the actual independence of the components obtained from MILCA Algorithm and its performance is compared with the popular ICA algorithms Infomax, Extended Infomax, Fast ICA, SOBI, TDSEP, JADE, OGWE, MS-ICA, SHIBBS, Kernel-ICA, and RADICAL.

EEG data with ocular artifacts are taken from [36], for testing various algorithms and to find the best ICA algorithm, for removal of ocular artifacts from EEG. The contaminated mixed EEG signals and the independent components obtained using MILCA algorithm is shown in Figure 3.

To evaluate the reliability of various algorithms,
  ➤ The pairwise MI estimates of the independent components and

> The difference between the overall MI of the contaminated Raw EEG and the overall MI of the independent components

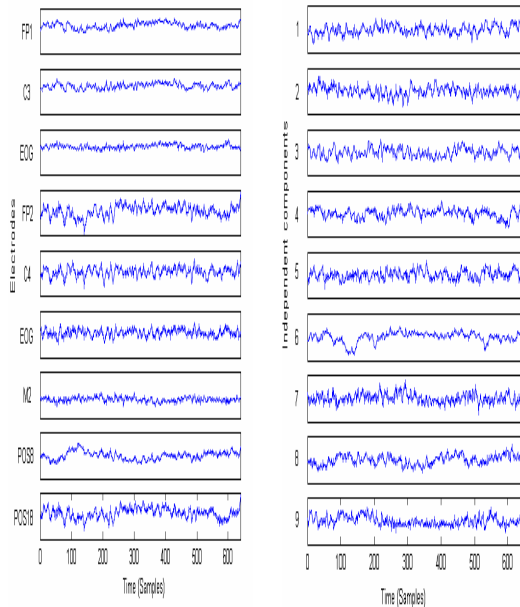obtained by various signal separation algorithms are compared.

Fig. 3 Contaminated EEG Data
and the Independent Components obtained
from Contaminated EEG using MILCA Algorithm

First, the pairwise MI estimate is given by $\hat{I}(\hat{s}_i, \hat{s}_j)$ for all $i, j = 1, ..., n$ and $i \neq j$ if $i = j$ then $\hat{I}(\hat{s}_i, \hat{s}_j)$ set as zero. i.e. the MI between independent components (1,2), (1,3), (1,4)………(2,1),(2,3)………. (9,1),(9,2)…………. and the MI between the same independent components i.e. (1,1),(2,2)…..is set as zero. Since MI is low for most independent components the values of the pairwise MI must be low for the best separation algorithm. Fig 4 shows the pairwise MI estimates of all the independent components obtained by various signal separation algorithms. The pairwise MI values are low for MILCA when compared to other algorithms, and hence it is clear that the inter dependencies are less in case of MILCA.

To evaluate the reliability, next the difference between the overall MI of the contaminated EEG and the overall MI of the independent components are calculated. The independent components for various ICA algorithms are obtained and MI for the mixed signals $\hat{I}(X_1, X_2,....X_N)$ and for the components $\hat{I}(\hat{S}_1, \hat{S}_2,....\hat{S}_N)$ are estimated by using Eqn

$$\hat{I}(x, y) = \psi(k) - (1/k) - \left\langle \psi(n_x) + \psi(n_y) \right\rangle + \psi(N) \qquad ()$$

choosing k = 6. Practical considerations for selecting k are discussed in [25]. Consider the first data set used in the analysis,

Overall MI of the Contaminated EEG
$\hat{I}(X_1, X_2,....X_N)$ - 7.0348
Overall MI of the estimated Independent
Components obtained by MILCA
$\hat{I}(\hat{S}_1, \hat{S}_2,....\hat{S}_N)$ - 0.9652
Difference between the two estimates   - 6.0696

Ideally, MI is zero, if two random variables are strictly independent. Hence it is expected that $\hat{I}(\hat{S}_1, \hat{S}_2,....\hat{S}_N)$ will decrease when compared to $\hat{I}(X_1, X_2,....X_N)$. So the difference between the two estimates is maximum, when the overall MI of the independent components is minimum, in other words, the obtained components are more independent. This difference is expected to be high for best separation algorithm. The difference between the overall MI of the contaminated EEG and the overall MI of the Independent Components obtained by various Signal Separation Algorithms for 10 data sets are tabulated in Table 1. Results show that for all EEG data sets this difference is high for MILCA algorithm.

To compare the uniqueness of the independent components obtained by various algorithms,

> Measure the pairwise MI of the obtained independent components under rotations in the two dimensional plane $\hat{I}_{ij}(\phi)$

> Calculate the square root of variability (σ_{ij}) of $\hat{I}_{ij}(\phi)$.

Since for unique solutions all the independent components are dissimilar, the pair wise MI will change significantly i.e. large variability σ_{ij}. But for ambiguous outputs, pair wise MI will stay almost constant so σ_{ij} is small. Square root of variability σ_{ij} of $(\hat{I}_{ij}(\phi))$ the independent components obtained by various signal separation are shown in Fig 5. From Fig 5 it can be observed that in case of MILCA the square root of variability is high hence it gives unique components when compared with other signal separation algorithms.

VI. CONCLUSION AND FUTURE SCOPE

Ocular artifact correction is a challenging task. A variety of techniques have been proposed in the literature for the same. However there is no general consensus amongst researchers upon the selection of the best, appropriate and feasible technique which enables the satisfactory removal of ocular artifacts and preservance of EEG information intact. In this paper, a recently introduced signal separation algorithm

Mutual Information based Least Dependent Component Analysis (MILCA) is used for separation of ocular artifacts from EEG. The performance of this algorithm is compared with eleven popular signal separation algorithms for the reliability and uniqueness of the decomposition using a reliable MI estimator. Results show that, MILCA algorithm performs best at separating the original sources from the observed signals. In [34], it is shown that JADE outperforms the well-known ICA/BSS algorithms such as Infomax, Extended Infomax, FastICA, SOBI, TDSEP. In [35], the authors has shown that RADICAL is superior compared to JADE, OGWE, SHIBBS, ICA-MS and Kernal-ICA. But in this paper, it is shown that MILCA has emerged superior when compared with JADE and RADICAL on the basis of Mutual Information Estimation. Once the components are as independent as possible, then the components can be

classified either as artifacts or EEG and can be removed from the mixed signals to obtain the artifact free EEG data. In this paper, the inspection of ocular artifact channel is identified visually, once the independent components are separated. However as an improvement over the current process, this inspection of artifact channel can be automated by using Adaptive Thresholding of Wavelet coefficients. The advantage of automated correction procedure is that it eliminates the subjectivity associated with non-automated correction procedures and can be used during on-line EEG monitoring for clinical purposes. Further it is our considered opinion that the usefulness of the best separation algorithm for removing ocular artifacts from EEG can also be justified quantitatively by proposing a suitable performance metric for validating the de-noised EEG signals.

REFERENCES

[1]   Croft RJ, Barry RJ (2000) "Removal of ocular artifact from the EEG: a review" *Clinical Neurophysiology*, 30(1), pp 5-19.

[2]   A.Kandaswamy, V Krishnaveni, .S. Jayaraman, N.Malmurugan and.K.Ramadoss (2005), "Removal of Ocula Artifacts from EEG - A Survey" *IETE Journal of Research*, Vol 52, No.2, March-April-2005

[3]   Gratton. G, Coles MG, Donchin E (1983) "A new method for off-line removal of ocular artifact", *Electroencephalography and Clinical Neurophysiology*, 55(4), pp 468-484.

[4]   Woestengurg JC, Verbaten MN, Slangen JL (1982), "The removal of the eye movement artifact from the EEG by regression analysis in the frequency domain" *Biological Physiology*, 16, pp 127-147.

[5]   Lagerlund TD, Sharbrough FW, Busacker NE (1997), "Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition", *Clinical Neurophysiology,* 14(1), pp 73 – 82.

[6]   Joliffe I T (1986), "Principal Component Analysis", Springer Verlag, New York,.

[7]   Comon P. (1994), "Independent Component Analysis, A new concept", *Signal Processing* 36(3), pp 287-314.

[8]   Scott Makeig, Tzyy-Ping Jung, Anthony J Bell, Terrence J Sejnowski (1996), "Independent Component Analysis of Electroencephalographic data", Advances in *Neural Information Processing Systems* 8 MIT Press, Cambridge MA, Vol (8), pp 145-151.

[9]   Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-won Lee, Martin J Mckeown, Vincent Iragui and Terrence J Sejnowski (1998), "Extended ICA removes Artifacts from Electroencephalographic recordings", Advances in Neural Information Processing Systems 10, MIT Press, Cambridge, MA, pp 894-900.

[10]   Vigario R, Jaakko Sarela, Veikko Jousmaki, Matti Hamalainen, Erkki Oja (2000), "Independent Component Approach to the Analysis of EEG and MEG Recordings", *IEEE Transactions on Biomedical Engineering*, Vol 47, No.5, pp 589-593.

[11]   Delorme.A, Makeig.. S & Sejnowski T (2001), "Automatic artifact rejection for EEG data using high-order statistics and independent component analysis", Proceedings of the *Third International ICA Conference,* pp 9-12.

[12]   Carrie A.Joyce, Irina F Gorodnitsky and Marta Kutas (2004), "Automatic removal of eye movement and blink artifacts from EEG data using blind component separation", *Psychophysiology*, Volume 41: Issue 2 , pp 313-325.

[13]   N.Nicolaou and S.J.Nasuto (2004), "Temporal Independent Component Analysis for automatic artefact removal from EEG", *2nd International Conference on Medical Signal and Information Processing*, Malta, pp 5-8.

[14]   Alexander Kraskov, Harald Stogbauer and Peter Grassberger, "Least Dependent Component Analysis Based on Mutual Information", ArXiv: physics/0405044 vol.2 28 Sep 2004.

[15]   Bell AJ, Sejnowski TJ, An information maximization approach to blind separation and blind deconvolution, Neural Computation, 7, 1995, pp 1004-1034.

[16]   Lee TW and Sejnowski T, " Independent Component Analysis for Sub Gaussian and Super-Gaussian Mixtures, *Proceedings of. 4th Joint Symposium on. Neural Computation*, 7, 1996, pp 132-139.

[17]   A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, 1997, pp. 1483–1492.

[18]   Beloucharani, K Meriam, J F Cardoso and E Moulines, "A blind source separation technique using second order statistics", *IEEE Transactions on Signal Processing*, 45, Feb 1997, pp 434-444.

[19]   A Ziehe and K R Muller, "TDSEP – an efficient algorithm for blind separation using time structure" in *Proceedings of ICANN '98*, December 1998, pp 675-680.

[20]   Jean-François Cardoso," High-order contrasts for independent component analysis", *Neural Computation*, vol. 11, no 1, Jan. 1999, pp. 157—192

[21]   Juan J. Murillo-Fuentes and Rafael Boloix-Tortosa, Francisco J. Gonz´alez-Serrano, "Initialized Jacobi Optimization in Independent Component Analysis".

[22]   L. Molgedey and H. Schuster, "Separation of independent signals using time-delayed correlations," *Physical Review Letters*, vol. 72, no. 23, pp. 3634–3637, 1994.

[23]   F. R. Bach and M. I. Jordan, "Kernel independent component analysis." J. of *Machine Learning Research*, 3:1–48, 2002.

[24]   Erik G. Miller and John W. Fisher III, "Independent components analysis by direct entropy minimization," Tech. Rep. UCB/CSD-03-1221, University of California at Berkeley, January 2003.

[25]   Alexander Kraskov, Harald Stogbauer and Peter Grassberger, "Estimating Mutual Information",. *ArXiv:cond-mat*/ 0305641 v1 28th May 2003

[26]   Girolami.M and Fyfe C, "Extraction of independent signal sources using a deflationary exploratory projection pursuit network with lateral inhibition", *IEEE proceedings on Vision Image Signal Processing, 1997*, 144, (5), pp 299-306

[27]   Amari.S, Cichocki.A and Yang H, "A new learning algorithm for blind signal separation" Advanced Neural Information Process. Syst. 1996, 8, pp 757-763.

[28]   Cardoso, J.-F., Bose, S., & Friedlander, B. (1996), "On optimal source separation based on second and fourth order cumulants," *Proc.IEEEWorkshoponSSAP*. Corfu, Greece.

[29]   DiMatteo, I., Genovese, C.R., Kass, R.E., 2001, "Bayesian curvefitting with free-knot splines," *Biometrika* 88, 1055– 1073.

[30]   J. Larsen L.K. Hansen and T. Kolenda, "On independent component analysis for multimedia signals," *Multimedia Image and Video Processing, CRC Press*, vol. Chapter 7, pp. 175–200, 2000.

[31]   B. Schölkopf and A. J. Smola, "Learning with Kernels." MIT Press, 2001.

[32]   Oldrich Vasicek, "A test for normality based on sample entropy,"*Journal of the Royal Statistical Society, Series B*, vol. 38, no. 1, pp. 54–   59, 1976.

[33] A. Hyvärinen and E. Oja, "A survey on independent component analysis," Helsinki University of Technology.

[34] V Krishnaveni, S Jayaraman, N Malmurugan, Chaitanya Mathi, K Ramadoss, "Quantitative Evaluation of Signal Seperation Algortihms for the removal of ocular artifacts from EEG" National Journal of Technology, No:2, Volume 1, June 2005 pp 47-53

[35] V Krishnaveni, S Jayaraman, P M Manoj Kumar, K Shivakumar, K Ramadoss, "Comparison of Independent Component Analysis Algorithms for removal of ocular artifacts from Electroencephalogram" Measurement Science Review Journal, Volume 5, Section 2, 2005 pp 67-79.

[36] http://www.sccn.ucsd.edu/~arno/famzdata/publicly_available_EEG_data .html

TABLE I
DIFFERENCE BETWEEN THE OVERALL MI OF THE CONTAMINATED EEG AND
OVERALL MI OF THE INDEPENDENT COMPONENTS OBTAINED
BY VARIOUS SIGNAL SEPARATION ALGORITHMS FOR 10 DATA SETS

| SET	INFOMAX	EXTENDED INFOMAX	FAST ICA	SOBI	TDSEP	JADE	RADICAL	ICA-MS	OGWE	Kernel _ICA	SHIBBS	MILCA
1	5.6882	5.7603	5.7967	4.8796	5.6123	5.9614	6.0695	4.9271	5.9328	5.9341	5.8848	**6.0696**
2	4.8316	5.3871	5.4945	5.1941	5.1311	5.5925	5.7057	5.0475	5.5388	5.5214	5.5292	**5.7774**
3	5.9039	6.0473	6.0987	5.3827	5.7605	6.0056	6.2622	5.2881	6.0751	6.0642	6.0652	**6.2689**
4	6.0471	6.0044	6.3819	5.8603	5.8967	6.4160	6.4445	5.9397	6.3642	6.3630	6.41434	**6.4790**
5	5.5765	5.6004	5.6041	5.4350	5.5170	5.6371	5.9231	5.4295	5.6572	5.7086	5.6737	**5.9532**
6	4.4646	4.5002	4.5048	3.9906	4.2827	4.6565	4.7250	3.8436	4.6621	4.6151	4.6704	**4.7259**
7	6.8702	6.7526	6.7607	5.5482	6.2028	6.7526	6.9988	5.6799	6.7835	6.9762	6.8149	**7.0546**
8	9.1995	8.9427	9.0852	8.1404	8.7504	9.1859	9.2331	8.1769	9.2062	8.9918	9.1184	**9.2615**
9	4.0303	4.3542	4.1389	3.5875	4.1999	4.3447	4.5797	3.5951	4.2867	4.5564	4.3775	**4.6991**
10	4.8420	4.8986	4.9295	4.2696	4.7905	4.9831	5.1721	4.4020	4.9305	5.0897	4.9752	**5.2013**
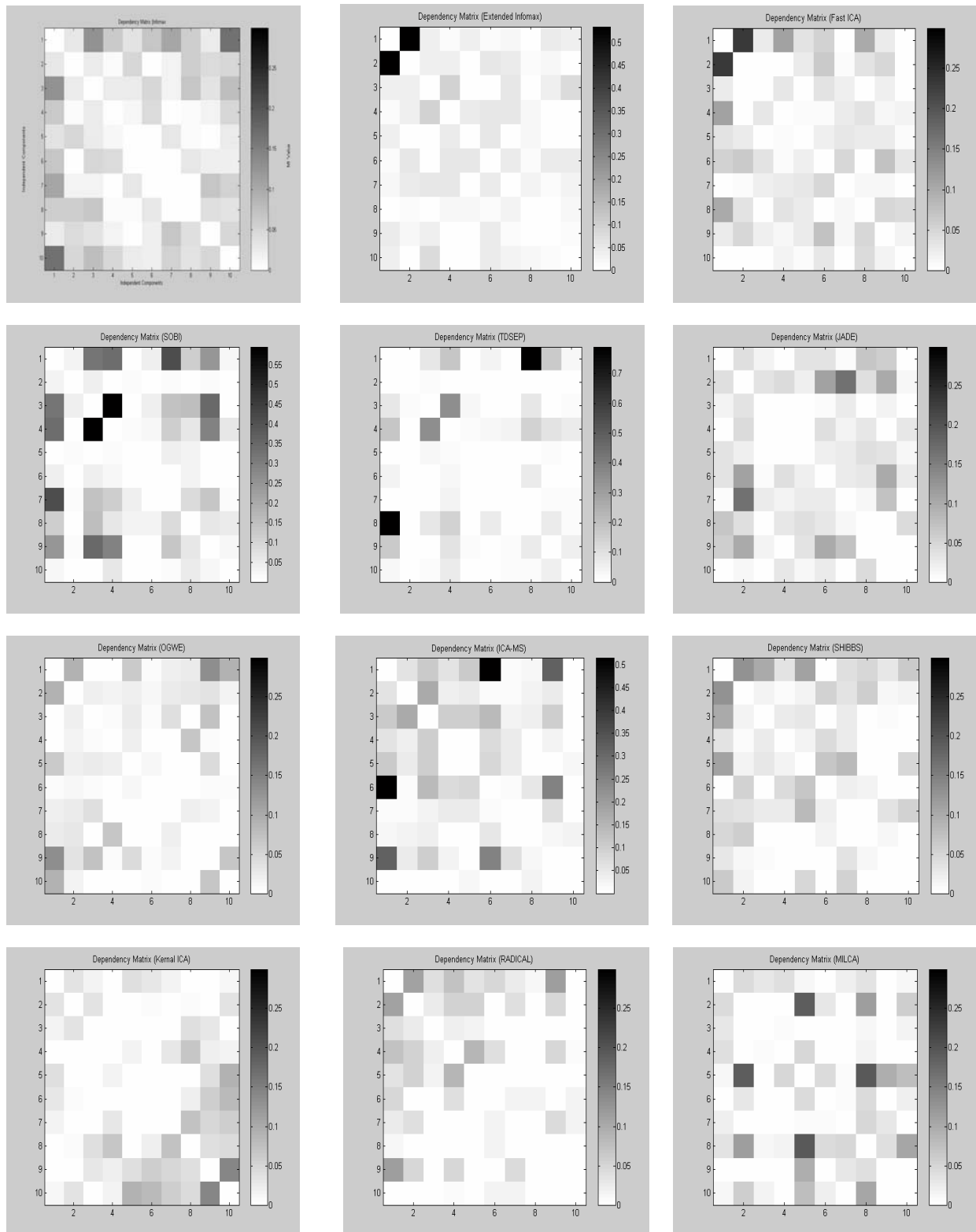
Fig. 4 Pairwise MI of all the Independent Components Obtained by Various
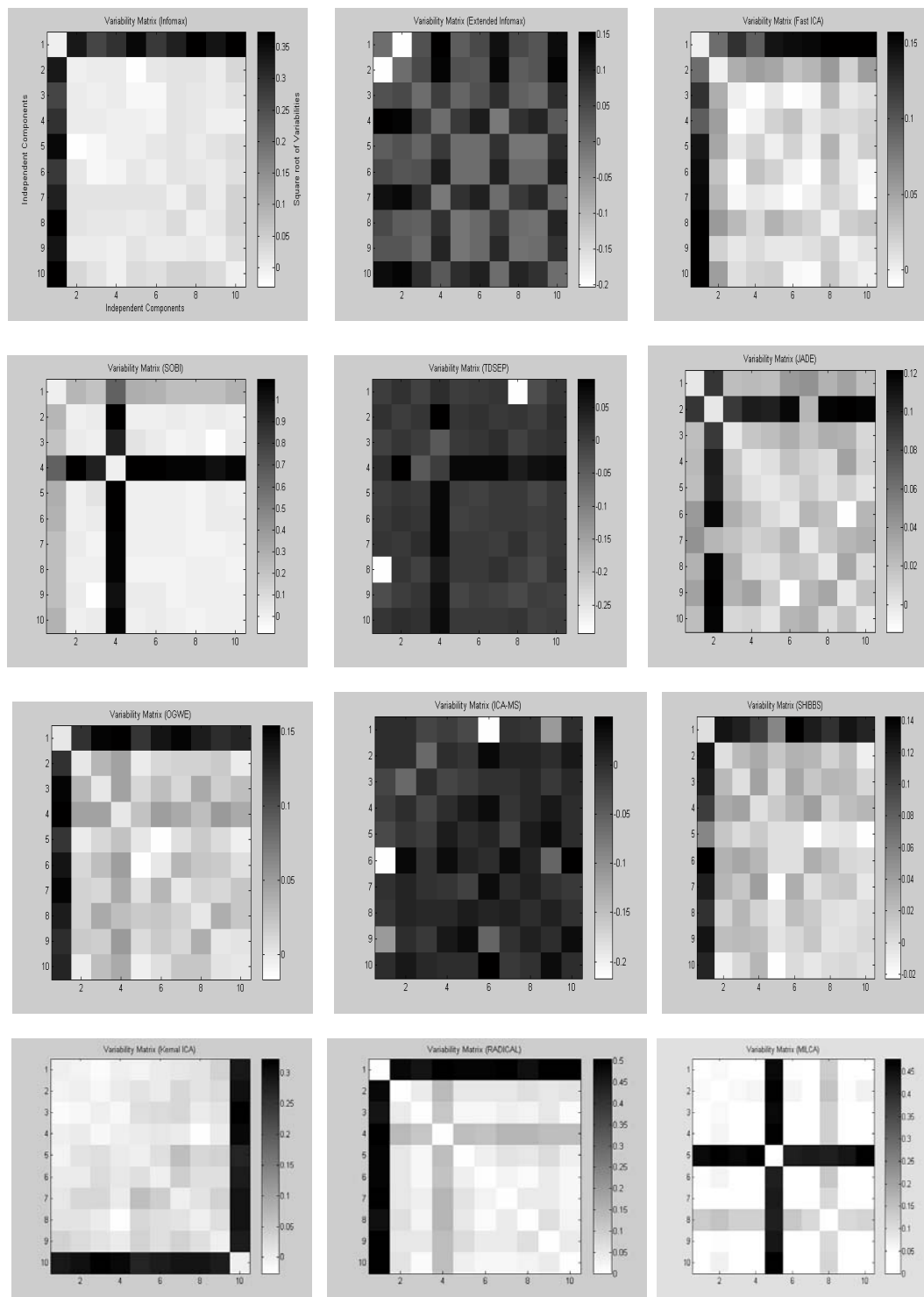Signal Separation algorithms for One Data Set

Fig .5 Square root of variability of all the EEG and OA components obtained by
various signal separation algorithms for one data set