Independent Component Analysis Minimizing the Mutual Information Criteria in Polar Coordinates

Carl Hoh
Department of Radiology, University of California, San Diego
200 W. Arbor Drive, Mail Code: 8758
San Diego, CA 92103-8758, USA

ABSTRACT

Determine the feasibility of an method of independent component analysis (ICA) on signal mixtures which directly minimizes the mutual information (MI) criteria between the signal components. The parameters in the optimization are performed in polar coordinates allowing 1) a plot of the MI criteria over each final parameter solution and to 2) visualize the behavior of the MI criteria in relation to the histogram bin size and data sample size. The algorithm was tested with dynamic images composed of time varying pixel intensities simulating concentrations of radiopharmaceutical activity. The method was able to converge to the global minimum and generate the underlying source signal and component images if the initial conditions of the optimization parameter are close ($\pm 10^o$) from the final solution.

Keywords: Independent Component Analysis, Dynamic images, Mutual Information, Polar coordinates, Optimization.

1. INTRODUCTION

Independent Component Analysis (ICA) is a method for temporal and spatial signal extraction [Herault 1986, Comon 1994]. The general form of the method is describe in Eq (1), where **X** is the signal mixture array, **W** is the unmixing matrix and **S** is the estimated source single array.

$$\mathbf{S} = \mathbf{W}\,\mathbf{X} \tag{1}$$

Many of the previously described methods involve a further transformation of **S** by g(**S**) based on a model probability density function (pdf). In these methods, there is a search for a unmixing matrix **W** is that maximizes the entropy of the system [Bell & Sejnowski 1995] or maximizes the likelihood estimate to a give model (see figure 1). The actual pdf of the source signal is not known; however, these techniques work if the model pdfs are an approximation to the source signal pdfs [Cardoso 2000, Amari 1998].

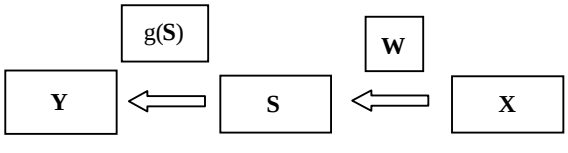

Figure 1. Independent component analysis where X is an m x n array of signal mixtures, W is a m x m weight matrix (unmixing matrix), S is the estimated source signals, and Y is the signal after transformation g(S) which is based on the model pdf of the source signal.

Other forms of ICA have been reported which do not rely on model pdfs and instead involve the direct minimization of the mutual information or other cost function between the estimated source signals in **S** [Hyvarinen 1997, Almedia 2003, Stogbauer 2004]. In some techniques, the requirement of true "independence" between the signals is replaced with the goal of separating mutually correlated signals, i.e. separating statistically "dependent" signals [Almedia 2003]. The mutual information (MI) criteria can be defined for two discrete signals $\boldsymbol{A}$ and $\boldsymbol{B}$ in Eq (2) [Clover & Thomas 1991], where $\boldsymbol{p_{AB}}$ represents the joint histogram distribution and $\boldsymbol{p_A}$ and $\boldsymbol{p_B}$ are the marginal distributions. In Eq 2, $\boldsymbol{A}$ and $\boldsymbol{B}$ would represent two different rows within matrix **S**. The total MI would then be the sum of MIs for all possible combinations of rows in **S**.

$$MI(A,B) = \sum_{a_i \in \Omega_A} \sum_{b_i \in \Omega_B} p_{AB}(a_i, b_i) \log\left(\frac{p_{AB}(a_i, b_i)}{p_A(a_i)\, p_B(b_i)} \right) \tag{2}$$

The goal of this paper was to see if the total mutual information between the estimated source signals could be directly used as a cost function to optimize the search for the appropriate unmixing matrix **W**. To evaluate the behavior of the MI as a cost function, the optimization was performed in polar coordinates rather than in the cartesian coordinates. The polar coordinates allow a plot of the MI over a range of angular rotations for each final parameter showing the convergence or non-convergence on a global minimum. In addition, the effect of histogram bin size on the behavior of the MI cost function could be revealed. It is known the MI function can be unsmooth causing problems in the search for either it's maximum or minimum [Maes 1997]. Known factors which may affect the behavior of the MI included the number of histogram bins, histogram interpolation methods, and data sample size [Maes 1997]. In this paper, the test data were dynamic images where each pixel intensity varied over time.

2. METHODS

In Eq (1), each column in **X** represent a random variable **x**(t) and each row in **W** represents a weight vector **w** [Stone 2004]. If **w** are unit length vectors, then each element of **S** can be conceptualized as the orthogonal projection of **X** on to each the unit length **w** vectors in **W**. The proposed method involves non-orthogonally rotating these unit length weight vectors with the goal of minimizing the mutual information (MI) criteria calculated between each source array row in **S**.

The general processing algorithm is shown in Fig 2. The

algorithm starts with a dimensionally reduced and whitened form of the data $\mathbf{X_w}$, which is a k x n sized matrix Eq (3), where k is the number of selected principal components and n is the number of data vectors. The preprocessing involves principal component analysis (PCA) to obtain the k largest eigenvalues (λ) and corresponding k eigenvectors ($\mathbf{\Phi}$) of $\mathbf{X}^T$. ICA requires a zero mean vector, so preprocessing with PCA was also performed assuming a zero mean as describe by (Naganawa 2005). In Eq (3), $\mathbf{\Lambda}^{-1/2}$ is a k x k diagonal matrix containing the inverse square roots of the eigenvalues to whiten the data. The value of k (number of principal component selected) is an operator determined value based on a visual assessment of where the plot of the cumulative variance has the largest change.

$$\mathbf{X_w} = \mathbf{\Lambda}^{-1/2}\,\mathbf{\Phi}\,\mathbf{X} \qquad (3)$$

The initial conditions are then determined by searching for the k largest vectors which have unique directions. This search can be rapidly accomplished by first reverse sorting the column vector lengths in $\mathbf{X}_w$ and then selecting the vectors which show a significant change in vector direction from the prior vector, correlation of less than 0.4. This subset of k selected column vectors is used to create $\mathbf{X_f}$ which approximates the apex positions that are sought after in factor analysis [DiPaola 1982] (Fig 2a). The matrix inverse of $\mathbf{X_f}$ provides the initial unmixing matrix $\mathbf{W_o}$ which consists of k vector rows $\mathbf{w_o}$ with unit length (Fig 2b). In the optimization portion of the algorithm, each $\mathbf{w_o}$ weight vector row is individually rotated with k parameter angles within $\boldsymbol{\theta}$ and then reformed to create a new unmixing matrix $\mathbf{W}$ (Fig 2c). The matrix $\boldsymbol{\theta}$ represents a k x k sized matrix consisting of k parameter angles for the k weight vector rows in $\mathbf{W_o}$.

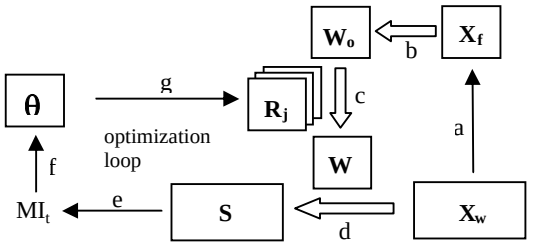

Figure 2. General processing algorithm of ICA using polar coordinates $\boldsymbol{\theta}$. a) find k column vectors in $\mathbf{X_w}$ which approximate the apices in factor analysis. b) matrix inversion to get initial weight vectors $\mathbf{w_o}$ (rows in $\mathbf{W_o}$). c) using $\mathbf{R}$, orthogonally transform each $\mathbf{w_o}$ separately to create $\mathbf{W}$. d) transform $\mathbf{X_w}$ with the unimixing matrix $\mathbf{W}$. e) calculate total MI_t, sum of MI between each row in $\mathbf{S}$. f) Obtain new $\boldsymbol{\theta}$. g) Generate Givens rotations $\mathbf{R}$ to reapply to each $\mathbf{w_o}$.

The actual orthogonal rotation of each $\mathbf{w_o}$ in k dimensional space was accomplished by generating an orthogonal rotation matrix $\mathbf{R}_j$ for each $\mathbf{w_{oj}}$, where j indexes a row in $\mathbf{W_o}$ (Fig 2g). Each $\mathbf{R}_j$ matrix itself is the product of k Givens rotations $\mathbf{G_j}$ which are the counter-clockwise orthogonal rotations around each of the k principal component axis [Golub 1996], Eq (4).

$$\mathbf{R}_j = \mathbf{G}_{j1}\mathbf{G}_{j2}\ldots\mathbf{G}_{jk} \qquad \text{where } j = 1 \text{ to } k \qquad (4)$$

The elements in each $\mathbf{G}_{ji}$ matrix consist of only $\cos(\theta_i)$, $\sin(\theta_i)$, $-\sin(\theta_i)$, with the remaining diagonals set to one, and the remaining elements set to zero. The index i ranges from 1 to k to cover the k rotation parameters in $\boldsymbol{\theta}$ associated with each $\mathbf{w_o}$. In the algorithm, once the number of principal components (k) was defined by the user, the elements of all the Givens rotations were coded as look up tables of sine and cosine functions or constants (0 or 1) to improve computational speed. The new unmixing matrix $\mathbf{W}$ can now perform a non-orthogonal transformation of the data $\mathbf{X_w}$ to obtain $\mathbf{S}$ (Fig 2d).

The MI cost function was calculated by Eq (2), where the MI between all k axes (i.e. rows) in S were totaled (Fig 2e). The search for the minimum MI was performed by the Nealder Mead downhill simplex algorithm [Neadler 1965] where the parameters of the optimization are the k x k angular rotations in $\boldsymbol{\theta}$ (Fig 2f). The initial set of parameter consisted of random values ranging from -2 to +2 angular degrees from the initial directions of $\mathbf{w_o}$. For each iteration, there is the generation of a new parameter matrix $\boldsymbol{\theta}$ which is applied to the rows of $\mathbf{W_o}$ to generate a new $\mathbf{W}$, which then transform $\mathbf{X_w}$ to a new $\mathbf{S}$, which then provides a new MI value cost function (Fig 2e). The algorithm iterates to a minimum MI with the stopping criteria where the MI values and each parameter value in $\boldsymbol{\theta}$ changes by less than 0.01 percent. At convergence, a final parameter matrix $\boldsymbol{\theta_f}$ and its corresponding unmixing matrix $\mathbf{W_f}$ were saved. $\mathbf{W_f}$ allows the generation of properly scaled signals $\mathbf{P}$ and parametric ICA images by performing the inverse of the transformation matrices Eq (5), where $\mathbf{S_m}$ is a k x k sized diagonal matrix containing the maximum values of each row of $\mathbf{S}$ from $\mathbf{W_f}$ x $\mathbf{X_w}$. The k rows in $\mathbf{P}$ are the independent or "pure" vectors in the units of the original data $\mathbf{X}$. The entire algorithm was implemented in IDL 6.3 (ITT, Boulder, CO).

$$\mathbf{P} = (\mathbf{W_f^{-1}}\,\mathbf{S_m})^T\mathbf{\Lambda}^{1/2}\,\mathbf{\Phi} \qquad (5)$$

Dynamic images (128 x 128 x 30 time frames) containing time varying signal and random noise were created to test the software. Dynamic test image #1 (Fig 3) consists of three regions. In the region 1, the image pixel intensity in of counts/pixel/sec is proportional to the measured concentration of radioactivity in blood after intravenous bolus injection of the radiopharmaceutical 18F-fluorodeoxyglucose (FDG) in a human. In regions 2 and 3, the image pixel intensities are simulated to be proportional to the time varying tissue concentration of radioactivity in the normal human liver and malignant tumor, respectively, by tracer kinetic modeling [Hoh 1996]. The vectors $\mathbf{x}(t)$ were generated by taking the time varying pixel intensities for each spatial location within a user drawn rectangular region of interest. The number of data vectors analyzed, $\mathbf{x}(t)$, could be increased by bilinear interpolation within the rectangular region of interest on the image in each time frame. Random Gaussian type noise was added to each element in $\mathbf{x}(t)$, where E is the output of the random Gaussian IDL function RANDOMN, x_{max} is the maximum value in the $\mathbf{X}$, and b is the user set percent noise level, Eq (5). The value b was set to 5% in both dynamic test images.

$$\text{noise} = (E\,x_{max}\,b)/100 \qquad (5)$$

3. RESULTS

The effect of the number of joint histogram bin sizes (32 x

32, 64 x 64, 128 x 128, 256 x 256) on the total mutual information (MI_t) from a data set consisting of 4484 data samples (i.e. 4484 $\mathbf{x}(t)$ vectors) is shown in (Fig 4). The MI_t is plotted varying all parameter angles $\boldsymbol{\theta}$ simultaneously from -90° to 90°. The plot shows that there are local minimum which are relatively close ($\pm$ 20 ° -30°) to the global minimum. These minimum appear to be less deep when the number of histogram bin sizes increases. Even with the 256 x 256 bin size there are still persistent local minima. Another interesting finding is that near the global minimum ($\pm$10°), the MI cost function 'appears' relatively well behaved.

A recalculation and replot of the total MI with a finer resolution, from -10° to +10°, shows that the higher bin size (256 x 256) reduces the "smoothness" of the MI function (Fig 5a). The true global minimum can be in error and the lower bin sizes (128, 64 and 32) shows that it may be 1/2° off. Reanalysis with a distance weighted interpolation method as describe by Maez to smooth out the MI function is plotted over the -10° to +10° and shows no visible improvement in the MI function (Fig 5b).

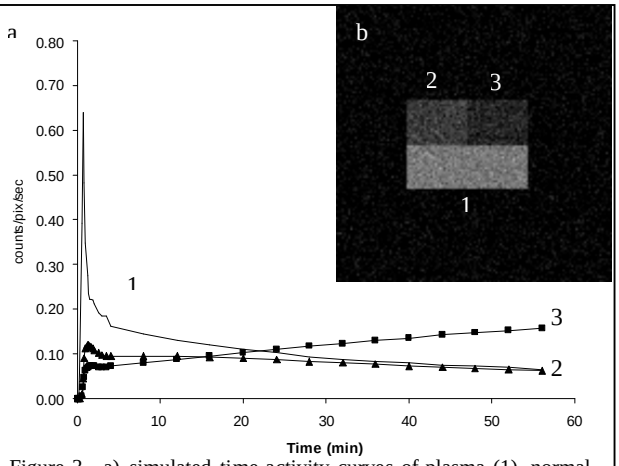

Figure 3. a) simulated time activity curves of plasma (1), normal liver (2), and tumor (3) after bolus injection of FDG. b) dynamic test image #1 with added noise composed of the three region where the pixel intensity values vary defined by the curves in (a).

Increasing the number of data samples analyzed by a factor of 4 (N=17936) by bilinear interpolation of the original image pixels, provides a smoother MI function when calculated with the 256 x 256 histogram bins (Fig 6). Also, all bin sizes appear to show the same global minimum at zero.

The complex behavior of the MI function is shown in Fig 7. The total MI is plotted for each parameter angle as it is individually changed from -90° to +90° while the other parameter angles are held constant.

The array of independent source vectors (or more appropriately, the least dependent source vectors) **P**, calculated in Eq (5), is shown in (Fig 8). Images of the "independent" components were generated by rescaling each value within a row of matrix **S** to a fraction of its row maximum value (Fig 8.1-3) and then assigning this value as the spatial intensity value. This value will range from 0 to 1 and is analogous to the method of creating factor images in factor analysis.

Dynamic test image #2 contains a linear mixture of pixel time activity curves and is shown in (Fig 9). Only the bottom edge, top left corner, and top right corners have pure source vectors of plasma, liver, and tumor, respectively.

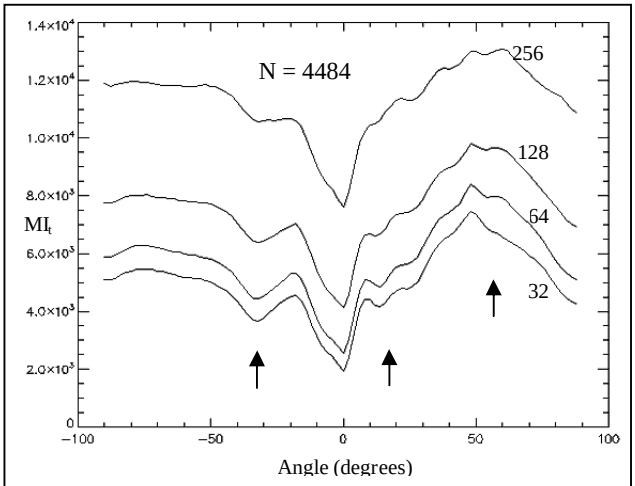

Figure 4. Effect of histogram bin sizes (32 x 32), (64 x 64), (128 x 128), and (256 x 256) on the total mutual information (MI_t) where all parameter angles are simultaneously rotate from -90° to +90°. Note: more local minimums (arrows) when bin size is less than 256 x 256.

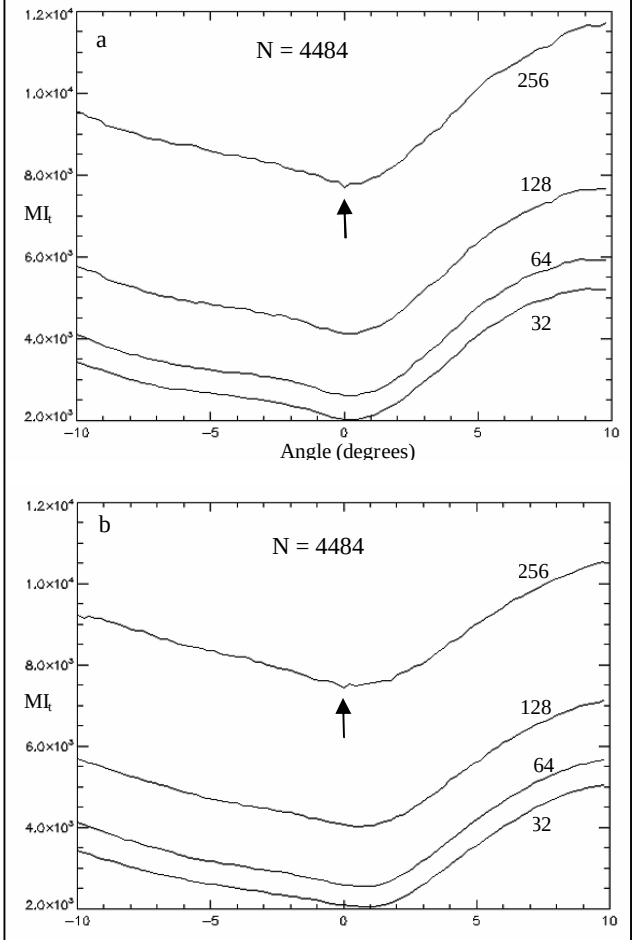

Figure 5. Effect of histogram bin sizes (32 x 32), (64 x 64), (128 x 128), and (256 x 256) on the total mutual information (MI_t) where all parameter angles are simultaneously rotate from -10° to +10°. a) without interpolation, b) with distance weighted histogram interpolation.

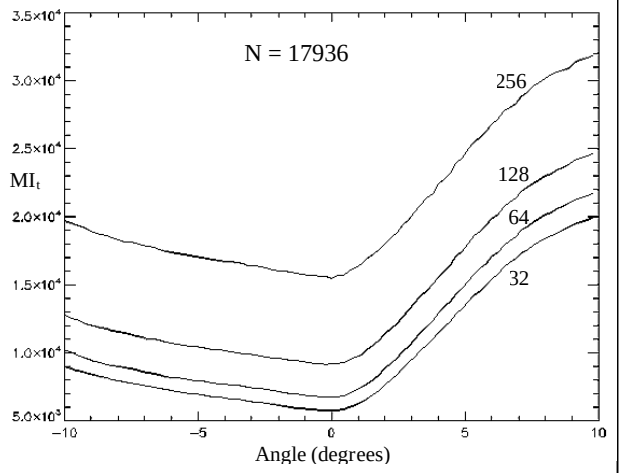

Figure 6. The cost function becomes more smooth with higher number of data samples N= 17936 as compared to 4484 samples in Figures 4 and 5.

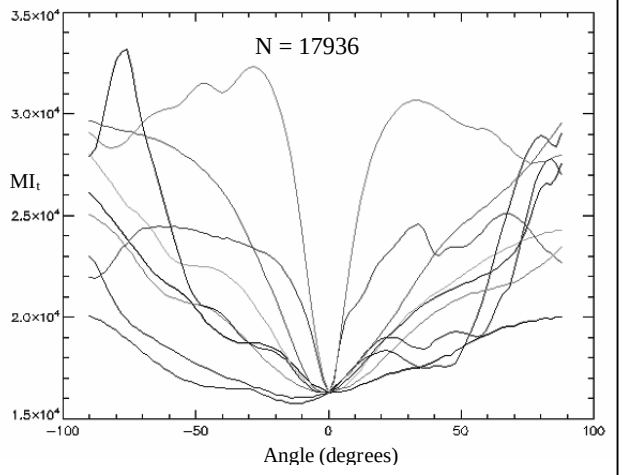

Figure 7. Total mutual information (MI$_t$) where each parameter angle is orthogonally rotated individually from -90^o to +90^o. Histogram bin size was set to 256 x 256 bins.

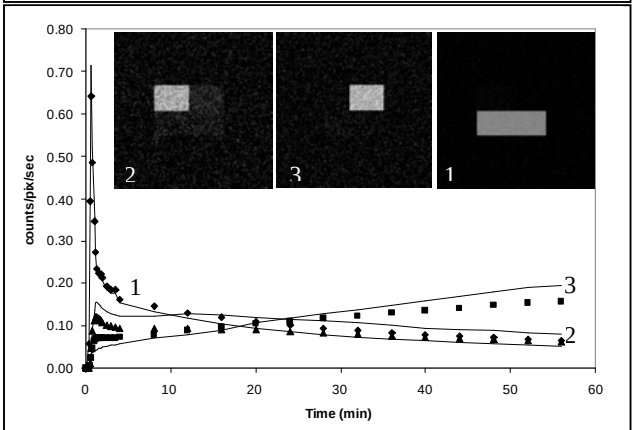

Figure 8. Comparison of ICA extracted curves (solid lines) to original curves (data points) and th eir associated component images generated from dynamic test image #1.

For the second dynamic test image, there are large and deep local minima in the MI$_t$ plot when the parameter angles are simultaneously changed from -90^o to + 90^o (Fig 10). The minima persist even with a high number of data samples (N=40356). The effect of rotating each parameter angle individually is shown in (Fig 11). Both plots show that the correct global minimum is achievable if the initial parameters are within 10^o of the global minimum.

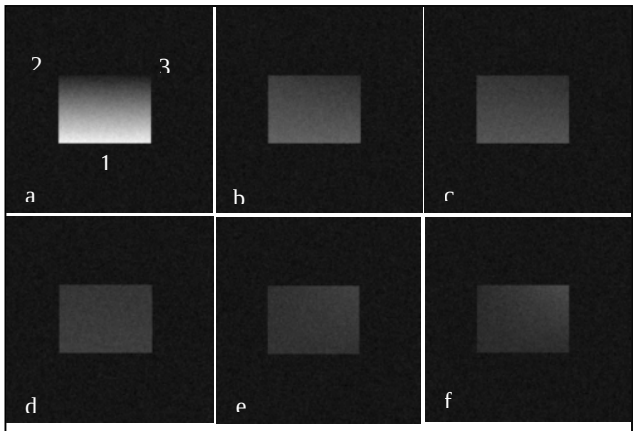

Figure 9. Dynamic test image #2, showing image frames at times 0.7, 1.5, 3.0, 16, 36, and 60 minutes (a thru f, respectively) containing plasma curves along the bottom row of pixels (1), normal liver in top left corner pixel (2), and tumor in top right corner pixel (3). Pixel within the phantom are a summed combination of intensities based on their inverse linear distant from their bottom edge or corners.

The three "independent" component images generated from the analysis of dynamic test image #2 show the original mixing pattern used to create the dynamic image (Fig 12).

The IDL6.3 software was run on a dual processor Athlon 1.79 GHz PC with 4 GBs of memory running Microsoft Windows XP. A multi-dimensional optimization involving three independent components (i.e. 9 rotation parameters) with 17936 data vectors was completed within 2 minutes. The Nealder Mead simplex method was implemented for the cost function minimization since it did not require function derivatives and the behavior of the mutual information cost function was known to be erratic. A more efficient gradient based optimizer could be used if the initial conditions are close to the global minimum.

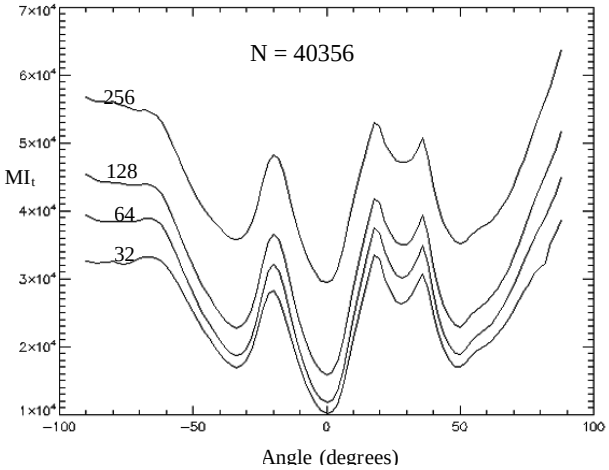

Figure 10. Effect of histogram bin size on total mutual information (MI$_t$) where all parameter angles are simultaneously changed from -90^o to +90^o for dynamic test image #2.

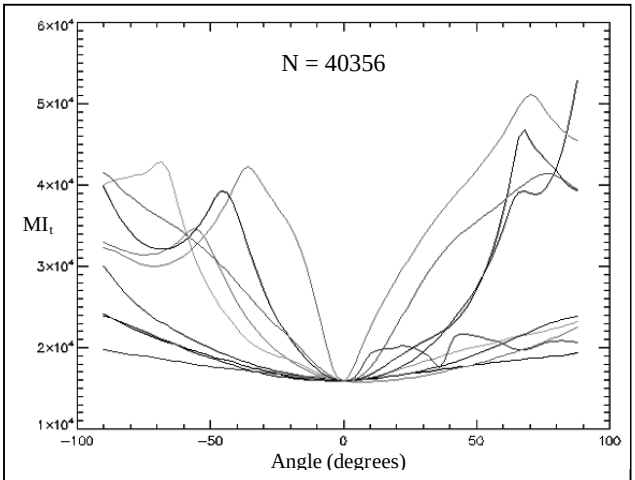

Figure 11. Plot of total mutual information (MI_t) where each parameter angle is orthogonally rotated individually from -90° to +90° for dynamic test image #2. Histogram bin size was set to 128 x 128 bins.

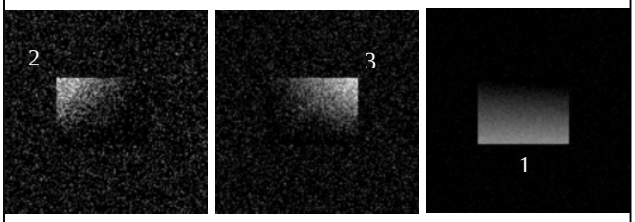

Figure 12. ICA component images show the original mixing pattern used to create dynamic test image #2.

4. DISCUSSION

With the dynamic image data tested, it appears that the direct minimization of the mutual information criteria for independent component analysis is feasible in polar coordinated but is highly dependent on the selection of the initial optimization parameters. These initial parameters, which represent the direction of the ICA weight vectors in multidimensional space, need to be within $\pm 10^o$ of the global minima due to the existence of deep minima just outside of this range. Using a high number of histogram bins in calculating the mutual information criteria reduces some of these minima but may not eliminate all deep local minima as was seen in the second dynamic test image with a high level of spatial mixing.

On the other hand, a high number of histogram bins (N=256) reduced the smoothness of the MI cost function when there was insufficient data samples and did not improve with a technique of distance weighted interpolation [Maes 1997]. Unlike the MI function applied in image co-registration, small changes in the angular parameters do not create large changes in the joint histogram of the MI calculation. The total MI function became smoother when the number of data vectors was increased by a factor of 4 to 17936 by bilinear interpolation of the original image pixel data. Interestingly, the lower bin sizes could be used if the initial parameters are close to the final solution, i.e. within $\pm 10^o$ of the global minimum.

The method for finding the initial conditions was based on the assumption that the source signals have higher signal amplitude compared to the corrupting noise. This allows the search for "unique" signal vectors in the PCA space that have "new" directions. The matrix inversion of these unique signal vectors then provides the initial ICA weight vector directions for further optimization. This method obviously will not work if the amplitude of the noise is equivalent or higher than that of the desired source signals. An alternative approach, to find the initial weight vectors, would be a polar "grid" search from -180° to +180°; however, this becomes computational expensive when dealing with higher dimensions.

By setting all the weight vectors in **W** to the unit length of 1, there is a reduction in the degrees of freedom for each weight vector by one, i.e. each weight vector's direction is fully defined by k-1 angle parameters, (where k is the number of principal components). For the two dynamic test images used in this analysis, the over parameterization (k parameters x k weight vector rows), did not lead to problems with parameter convergence.

The algorithm was implemented to handle any number of data dimensions within the limits of IDL's access of computer memory. A faster numerical optimization other than the Nealder Mead simplex could be used if the initial conditions are close to the global minimum. In addition, the Givens rotations can be parallelize once the number of principal components of the analysis have been defined.

5. CONCLUSIONS

The method described in this paper shows that minimization of the mutual information can be performed using polar coordinates. No a prior information or objective function was needed, except for an estimate of the number of underlying component signals in the mixture. The mutual information as a cost function can be unpredictable except near the global minimum (within $\pm 10^o$). The function can have local minima which are dependent on the number of histogram bin sizes used in the calculation of the MI criteria and in the number of data samples used in the analysis. Near the global minimum, lower histogram bin sizes may be use.

5. REFERENCES

[1] J. Herault, C. Jutten, "Space or time adaptive signal processing by neural network models", **Neural Networks for Computing: AIP Conference Proceedings** 151, 1986, J. S. Denker, ed. American Institute for Physics, New York.

[2] P. Comon, "Independent Component Analysis - a new concept"? **Signal Processing**, Vol. 36, 1994, pp. 287-314.

[3] A.J. Bell, T.J. Segnoski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution", **Neural Computation**, Vol. 7, 1995, pp. 1129-1159.

[4] J.-F. Cardoso, "Infomax and maximum likelihood for source separation". **IEEE Letters on Signal Processing**, Vol. 4, 1997, pp. 112–114.

[5] J. Cardoso, "On the stability of source separation algorithms". **Journal of VLSI signal processing systems**, 2000, pp.7-14.

[6] S. Amari, “Natural gradient works efficiently in learning”. **Neural Computation**, Vol. 10, 1998, pp.251-276.

[7] J.V. Stone, **Independent Component Analysis, A Tutorial Introduction**, Cambridge: The MIT Press. 2004.

[8] R. Di Paola, J.P. Bazin, F. Aubry, et al. “Handling of dynamic sequences in nuclear medicine”. **IEEE Trans Nucl Sci**. NS29, 1982, pp. 1310-1321.

[9] A Hyvarinen, E. Oja, “Independent Component Analysis by Minimization of the Mutual Information”, **Independent Component Analysis**, Ed. Simon Haykin, John Wiley & Sons, Pub. 2001.

[10] L.B. Almeid, “MISEP – Linear and Nonlinear ICA Based on Mutual Information”, **Journal of Machine Learning Research**, Vol. 4, 2003, pp. 1297-1318.

[11] H. Stögbauer, A. Kraskov, S.A. Astakhov, P. Grassberger, “Least-dependent-component analysis based on mutual information”, **Physical Review E**, Vol. 70, 2004, 066123

[12] T.M. Cover, J.A. Thomas, **Elements of information theory**, John Wiley & Sons, New York, NY, Pub., 1991.

[13] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, P. Suetens, Multimodality Image Registration by Maximization of Mutual Information, **IEEE Trans on Med Imaging**, Vol. 16, No. 2, 1997, pp. 187-198.

[14] M. Naganawa, Y. Kimura, K. Ishii, K. Oda, K. Ishiwata, A Matani, “Extraction of a Plasma Time-Activity Curve From Dynamic Brain PET Images Based on Independent Component Analysis”. **IEEE Trans on Biomed Engineering**, Vol. 52, No. 2, 2005, pp. 201-210.

[15] G.H. Golub, C.F. van Loan, **Matrix Computations**, 3nd ed., The Hopkins Univ Press, 1996.

[16] J.A. Nelder, R. Mead. A simplex method for function minimization. **Computer Journal**, Vol. 7, 1965, pp 308–313.

[17] C.K. Hoh, M. Dahlbom, S.S. Gambhir, J. Yang, M.E. Phelps. State space modeler (SSM), a general software package for dynamic systems modeling. **J Nucl Med**, vol 37, 1996, 303P