Title: Subspace Independent Component Analysis using Vector Kurtosis

Author1 (correspondent author)**:**
Alok Sharma
Signal Processing Lab
Griffith University
Brisbane, Australia

University of the South Pacific
Suva, Fiji
Email: sharma_al@usp.ac.fj
Phone: +679 321 2870/ +617 3875 3754; Fax: +679 323 1538

Author2
Kuldip K. Paliwal
Signal Processing Lab
Griffith University
Brisbane, Australia

Abstract: This discussion presents a new perspective of subspace independent component analysis (ICA). The notion of a function of cumulants (kurtosis) is generalized to vector kurtosis. This vector kurtosis is utilized in the subspace ICA algorithm to estimate subspace independent components. One of the main advantages of the presented approach is its computational simplicity. The experiments have shown promising results in estimating subspace independent components.

1. **Introduction**: Independent component analysis is a widely accepted tool in solving blind source separation (BSS) problems. In BSS problem a set of observations is given but the underlying source information is hidden. The mixing weights of this underlying source information are also not known to the observer. The BSS problem is thus to identify the source signals and/or the mixing weights. The assumptions in the basic ICA model include the source signals being mutually independent and having nongaussian distributions. In the BSS problem an $M \times 1$ vector of observation **x** is modelled from statistically independent and nongaussian components **s** of size $M \times 1$:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

where **A** is a square and invertible mixing matrix of size $M \times M$. The elements of $\mathbf{s} = [s_1, \ldots, s_M]^T$ are linearly mixed with the mixing matrix **A** to give the observation **x**. The source signals could be obtained up to their permutation, sign and amplitude only, that is the order and variances of independent components cannot be determined. These indeterminacies are, however, insignificant in most of the applications.

Some techniques [1,2] have evolved in recent years that relax the assumptions of basic ICA model and generalize the problem. These techniques are a generalization of basic ICA model and are known as multidimensional ICA (MICA) [2] and subspace ICA [1] model. In MICA or subspace ICA it is not assumed that all the source signals are independent, instead it is assumed that some components that usually come in *n*-tuples or the elements of subspaces are mutually non-independent. However, the non-independencies among different *n*-tuples or subspaces are not allowed.

In this paper we present a new perspective of subspace ICA model. Unlike MICA [2] or subspace ICA [1] we have not applied an additive model. However, the multiplicative model as of basic ICA has been utilized except that it is partitioned into sub-matrices and sub-vectors. Then we generalize the notion of kurtosis [3] to vector kurtosis for our model and show the relationship of the optimized vector kurtosis to the subspace independent components. This approach would solve the BSS problem even when not all the components are independent i.e. it accounts for a generalized problem. One of the advantages of our subspace ICA algorithm is its computational simplicity due to the use of vector (generalized) kurtosis function.

2. **Evaluation of independent components by maximizing a quantitative measure of nongaussianity**: Independent components can be estimated by the maximization of nongaussianity. Two quantitative measures of nongaussianity readily used in ICA estimation are kurtosis and negentropy [3].

2.1 **Kurtosis**: Kurtosis or univariate kurtosis is a fourth order cumulant of a random variable. For zero-mean random variable, kurtosis is defined as:

$$\mathrm{kurt}(y) = E[y^4] - 3(E[y^2])^2 \tag{2}$$

Kurtosis value can be any real number. Random variables with $\mathrm{kurt}(y) > 0$ are considered supergaussian while with $\mathrm{kurt}(y) < 0$ are considered subgaussian. For gaussian random variables and a very few nongaussian variables $\mathrm{kurt}(y) = 0$. Thus nongaussianity can be measured by the absolute value of kurtosis. If the variance of random variables are kept constant (i.e. $E[y^2] = 1$) then kurtosis can be computed by the fourth moment of random variables. The main advantage of using kurtosis is its computational simplicity. One of the drawbacks of kurtosis inherited by the fourth order moments is its susceptibility (sensitivity) to outliers [3].

3. **Subspace ICA and MICA**: Cardoso [2] introduced the notion of MICA by generalizing basic ICA model. MICA is an additive model which is derived from the multiplicative model. Its components $\mathbf{s}_i$ are vector-valued, instead of scalar-valued as of equation 1 and not all the elements of $\mathbf{s}_i$ are assumed to be independent. MICA was estimated by maximum likelihood (ML) estimation and illustrated on foetal ECG dataset [4]. The author argued that the dataset was well modelled by MICA decomposition into one bi-dimensional component (mother) and one mono-dimensional component (foetal).

Hyvärinen and Hoyer [1] combined the technique of MICA and the principle of invariant-feature subspaces[1] [5] to explain the emergence of phase- and shift-invariant features. The authors call the *n* elements of

---

[1] The principle of invariant-feature subspace is that invariant-feature can be considered as a linear subspace in a feature space and its value can be computed by taking the norm of the projection on that subspace.

$\mathbf{s}_i$ as the subspace spanned by a set of *n* basis vectors an independent subspace and referred the algorithm as independent subspace analysis (ISA) or subspace ICA and estimated subspace independent components by ML estimation. Thus different subspaces are mutually independent but the entries of each subspace are not independent. The probability density of each subspace is considered to be spherically symmetric, i.e. it depends only on the norm of the projection.

4 **Subspace ICA model: a new perspective**: We take the multiplicative model and partition the entries of matrix and vectors:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \qquad \text{or} \qquad \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_M \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \cdots & \mathbf{A}_{1M} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{M1} & \cdots & \mathbf{A}_{MM} \end{bmatrix} \begin{bmatrix} \mathbf{s}_1 \\ \vdots \\ \mathbf{s}_M \end{bmatrix} \tag{3}$$

where $\mathbf{x}_j$ and $\mathbf{s}_j$ of $\mathbf{x}$ and $\mathbf{s}$ respectively are vectors of dimension d and can be defined as $\mathbf{x}_j = [x_1^j, x_2^j, \ldots, x_d^j]^T$ and $\mathbf{s}_j = [s_1^j, s_2^j, \ldots, s_d^j]^T$ for $j = 1, \ldots, M$. Partitioned matrix $\mathbf{A}$ (equation 3) is of size $Md \times Md$ since its entries $\mathbf{A}_{ij}$ are matrices of size $d \times d$. We made the following assumptions for our model:

<u>*Assumption1*</u>: Components $\mathbf{s}_j$ are vector-valued, nongaussian, mutually independent and of identity covariance.

<u>*Assumption2*</u>: Entries of $\mathbf{s}_j$ are not independent and all are of equal dimension d.

<u>*Assumption3*</u>: Sample data is centered and whitened.

To estimate subspace independent components we take a d-dimensional vector $\mathbf{y}$ which is defined as:

$$\mathbf{y} = \mathbf{B}^T \mathbf{x} = \sum_{j=1}^{M} \mathbf{B}_j^T \mathbf{x}_j \tag{4}$$

where size of $\mathbf{B}$ and $\mathbf{B}_j$ are $Md \times d$ and $d \times d$ respectively. Given equation 4, now the problem is to identify and/or estimate subspace independent components from the observation $\mathbf{x}$ only. The problem is solved in section 4.2.

4.1 **Extension of univariate kurtosis to vector kurtosis**: Univariate kurtosis or simply kurtosis (section 2.1) is utilized when the variable y is a scalar quantity or one-dimensional vector. It does not accommodate for multidimensional features. To solve the multidimensional problem we first need to extend the basic kurtosis function. The natural generalization of basic kurtosis function for any vector $\mathbf{y}$ can be given as:

$$\text{kurt}(\mathbf{y}) = E[(\mathbf{y}^T\mathbf{y})^2] - 3(E[\mathbf{y}^T\mathbf{y}])^2 \tag{5}$$

which is a multidimensional equivalent of equation 2. There is no covariance term in equation 5. This is due to one of our assumptions that the sample data is whitened ($E[\mathbf{y}\mathbf{y}^T] = I_{d\times d}$). We refer to this generalized kurtosis as vector kurtosis. As of kurtosis function, vector kurtosis is computationally simple but sensitive to the outliers.

4.2 **Relation of optimized vector kurtosis to the subspace independent components**: In this section we discuss how the subspace independent components are related to the optimization of vector kurtosis. Let us consider the subspace independent component (equation 4) again. From equations 3 and 4, the component can be written as:

$$\mathbf{y} = \mathbf{B}^{\mathrm{T}}\,\mathbf{A}\mathbf{s} = \mathbf{Q}^{\mathrm{T}}\,\mathbf{s} = \sum_{i=1}^{\mathrm{M}} \mathbf{Q}_i^{\mathrm{T}}\,\mathbf{s}_i \tag{6}$$

where size of $\mathbf{Q}$ and $\mathbf{Q}_i$ are $\mathrm{Md}\times\mathrm{d}$ and $\mathrm{d}\times\mathrm{d}$ respectively. Equation 6 is a linear combination of vectors $\mathbf{s}_i$. To show the relationship, consider two observations ($\mathrm{M}=2$) s_1 and s_2 each of dimension d . This would simplify equation 6 as:

$$\mathbf{y} = \mathbf{Q}_1^{\mathrm{T}}\,\mathbf{s}_1 + \mathbf{Q}_2^{\mathrm{T}}\,\mathbf{s}_2 \tag{7}$$

Using the additive property of kurtosis (which can be shown for vector kurtosis as well) we can say:

$$\begin{aligned} f(\mathbf{Q}_1,\mathbf{Q}_2) &= \mathrm{kurt}(\mathbf{y}) = \mathrm{kurt}(\mathbf{Q}^{\mathrm{T}}\mathbf{s}) \\ &= \mathrm{kurt}(\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1) + \mathrm{kurt}(\mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2) \end{aligned} \tag{8}$$

where $\mathrm{kurt}(\mathbf{Q}_j^{\mathrm{T}}\mathbf{s}_j) = \mathrm{E}[(\mathbf{s}_j^{\mathrm{T}}\mathbf{Q}_j\mathbf{Q}_j^{\mathrm{T}}\mathbf{s}_j)^2] - 3(\mathrm{E}[(\mathbf{s}_j^{\mathrm{T}}\mathbf{Q}_j\mathbf{Q}_j^{\mathrm{T}}\mathbf{s}_j)])^2$. Now we put a constraint g on $\mathbf{Q}$ (since $\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$):

$$\begin{aligned} \mathrm{E}[\mathbf{y}^{\mathrm{T}}\mathbf{y}] &= \mathrm{E}[(\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1 + \mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2)^{\mathrm{T}}(\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1 + \mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2)] \\ &= \mathrm{E}[\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1] + \mathrm{E}[\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2\mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2] \quad \{\because \mathbf{s}_1 \text{ and } \mathbf{s}_2 \text{ are independent and } \mathrm{E}[\mathbf{s}_1] = \mathrm{E}[\mathbf{s}_2] = 0_{\mathrm{d}\times 1}\} \end{aligned}$$

again $\mathrm{E}[\mathbf{y}^{\mathrm{T}}\,\mathbf{y}] = \mathrm{E}[\mathit{trace}(\mathbf{y}^{\mathrm{T}}\,\mathbf{y})] = \mathrm{E}[\mathit{trace}(\mathbf{y}\mathbf{y}^{\mathrm{T}})] = \mathit{trace}(\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}]) = \mathit{trace}(\mathrm{I}_{\mathrm{d}\times\mathrm{d}}) = \mathrm{d}$ and using equation 7

$$\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \mathbf{Q}_1^{\mathrm{T}}\mathbf{Q}_1 + \mathbf{Q}_2^{\mathrm{T}}\mathbf{Q}_2 = \mathrm{I}_{\mathrm{d}\times\mathrm{d}} \tag{9}$$

therefore, constraint g can be written as:

$$g(\mathbf{Q}_1,\mathbf{Q}_2) = \mathrm{E}[\mathbf{y}^{\mathrm{T}}\mathbf{y}] - \mathrm{d} = \mathrm{E}[\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1] + \mathrm{E}[\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2\mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2] - \mathrm{d} = 0 \tag{10}$$

From equation 9 it can be stated that column vectors of rectangular matrix $\mathbf{Q}$ are orthonormalized. The optimization problem can now be solved by finding $\mathbf{Q}_1$ and $\mathbf{Q}_2$ that occur at constrained relative-extremum of $f(\mathbf{Q}_1,\mathbf{Q}_2)$ (equation 8) under the constrained curve $g(\mathbf{Q}_1,\mathbf{Q}_2)$ (equation 10) using the method of Lagrange multipliers:

$$\nabla_{(\mathbf{Q}_1,\mathbf{Q}_2)} f(\mathbf{Q}_1,\mathbf{Q}_2) = \lambda\,\nabla_{(\mathbf{Q}_1,\mathbf{Q}_2)} g(\mathbf{Q}_1,\mathbf{Q}_2) \quad \text{where } \lambda \neq 0 \tag{11}$$

Solving for the derivatives of functions f and g (partial proof of equations 12 and 13 can be viewed in appendix

1), we get

$$\nabla_{(\mathbf{Q}_1,\mathbf{Q}_2)} f(\mathbf{Q}_1,\mathbf{Q}_2) = \{4\mathrm{E}[(\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1)(\mathbf{s}_1\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1)] - 12\mathrm{E}[\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1]\mathrm{E}[\mathbf{s}_1\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1]\}\,\hat{i}_{\mathrm{Q}_1} + \{4\mathrm{E}[(\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2\mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2)(\mathbf{s}_2\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2)] - 12\mathrm{E}[\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2\mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2]\mathrm{E}[\mathbf{s}_2\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2]\}\,\hat{i}_{\mathrm{Q}_2} \quad (12)$$

and $$\nabla_{(\mathbf{Q}_1,\mathbf{Q}_2)} g(\mathbf{Q}_1,\mathbf{Q}_2) = 2\mathrm{E}[\mathbf{s}_1\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1]\,\hat{i}_{\mathrm{Q}_1} + 2\mathrm{E}[\mathbf{s}_2\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2]\,\hat{i}_{\mathrm{Q}_2} \quad (13)$$

substituting equations 12 and 13 in equation 11 and comparing $\hat{i}_{\mathrm{Q}_1}$ terms, we get:

$$4\mathrm{E}[(\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1)(\mathbf{s}_1\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1)] - 12\mathrm{E}[\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1\mathbf{Q}_1^{\mathrm{T}}\mathbf{s}_1]\mathrm{E}[\mathbf{s}_1\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1] = \lambda 2\mathrm{E}[\mathbf{s}_1\mathbf{s}_1^{\mathrm{T}}\mathbf{Q}_1] \quad (14)$$

It is evident from equation 14 that $\mathbf{Q}_1 = 0_{\mathrm{d}\times\mathrm{d}}$ is one of the solutions. The corresponding value of $\mathbf{Q}_2$ for this value of $\mathbf{Q}_1$ can be obtained by substituting $\mathbf{Q}_1 = 0_{\mathrm{d}\times\mathrm{d}}$ in constraint curve (equation 10), which yields: $g(0_{\mathrm{d}\times\mathrm{d}},\mathbf{Q}_2) = \mathrm{E}[\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2\mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2] - \mathrm{d} = 0$

or $\mathrm{E}[\mathbf{s}_2^{\mathrm{T}}\mathbf{Q}_2\mathbf{Q}_2^{\mathrm{T}}\mathbf{s}_2] = \mathrm{d}$; or $trace(\mathbf{Q}_2^{\mathrm{T}}\,\mathbf{Q}_2) = \mathrm{d}$ (15)

Equations 9 and 15 imply that $\mathbf{Q}_2^{\mathrm{T}}\,\mathbf{Q}_2 = \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. These values suggest that the norm of y is equal to the norm of one of the subspace independent components $||\mathbf{y}||^2 = \mathbf{y}^{\mathrm{T}}\mathbf{y} = (\mathbf{Q}_i^{\mathrm{T}}\mathbf{s}_i)^{\mathrm{T}}(\mathbf{Q}_i^{\mathrm{T}}\mathbf{s}_i) = \mathbf{s}_i^{\mathrm{T}}\mathbf{s}_i = ||\mathbf{s}_i||^2$ or $||\mathbf{y}|| = ||\mathbf{s}_i||$. Therefore, for any whitened data **z** (which can be achieved for example by eigenvalue decomposition procedure of covariance of sample data **x**), we search for $\mathbf{W}^{\mathrm{T}}\,\mathbf{z}$ (where **W** is a rectangular matrix of the same size as $\mathbf{Q}$) that maximizes vector kurtosis. We see that $\mathrm{Q} = (\mathbf{VA})^{\mathrm{T}}\,\mathbf{W}$ and $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = (\mathbf{W}^{\mathrm{T}}\mathbf{VA})(\mathbf{A}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{W}) = \mathbf{W}^{\mathrm{T}}\mathbf{W}$. It can also be observed from equation 9 that $\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{Q}_1^{\mathrm{T}}\mathbf{Q}_1 + \mathbf{Q}_2^{\mathrm{T}}\mathbf{Q}_2 = \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. Thus we maximize $\mathbf{W}^{\mathrm{T}}\,\mathbf{z}$ under the constraint $\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$. This **W** will give first subspace independent component and second subspace IC will be mutually orthogonal to the first one. Altogether there are M subspace ICs. The p^{th} subspace IC is orthogonal to all the previous $1\ldots p-1$ subspace ICs. The same algorithm needs to be run M times to get all the subspace ICs. It is therefore rather appropriate to define a square matrix $\mathbf{\Lambda}$ of size $\mathrm{Md}\times\mathrm{Md}$ that consists of M rectangular matrices **W** such that $\mathbf{\Lambda} = [\mathbf{W}_1\ldots\mathbf{W}_{\mathrm{M}}]$. Therefore the objective is to find all **W** to get projection $\mathbf{\Lambda}^{\mathrm{T}}\,\mathbf{z}$.

4.3 **Fixed point algorithm using vector kurtosis:** In this section we discuss the fixed-point algorithm [6] for finding the projection matrix $\mathbf{W}\in\mathbf{\Lambda}$ which would enable us to find subspace independent components. Let the whitened data **z** be a set of vectors defined as $\mathbf{z} = [\mathbf{z}_1^{\mathrm{T}},\ldots,\mathbf{z}_{\mathrm{M}}^{\mathrm{T}}]^{\mathrm{T}}$, where $\mathbf{z}_j$ is a vector of d dimension and given by $[\mathrm{z}_1^j,\ldots,\mathrm{z}_{\mathrm{d}}^j]^{\mathrm{T}}$ (z_i^j are scalar quantities). For a projection matrix **W** of size $\mathrm{Md}\times\mathrm{d}$ the gradient of absolute value of vector kurtosis can be computed as (see appendix 1 for the proof)

$$\frac{\partial\,|\,\mathrm{kurt}(\mathbf{W}^{\mathrm{T}}\mathbf{z})\,|}{\partial\,\mathbf{W}} = 4\mathrm{sign}(\mathrm{kurt}(\mathbf{W}^{\mathrm{T}}\mathbf{z}))\left[\mathrm{E}[(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})(\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W})] - 3\mathrm{E}[\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z}]E[\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W}]\right]$$

For whitened data **z** and normalized[2] **W**, the fixed-point algorithm for subspace ICA model (see appendix 1) would be $\mathbf{W} \leftarrow \mathrm{E}[(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})(\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W})] - 3\mathrm{d}\,\mathbf{W}$. The algorithm will converge when the norm of new and old values of **W** point in the same direction, i.e. $||(\mathbf{W}^{+})^{\mathrm{T}}\,\mathbf{W}|| \approx ||\mathrm{I}_{\mathrm{d}\times\mathrm{d}}||$(where $\mathbf{W}^{+}$ is the new value of **W** and $||\bullet||$ is Frobenius norm). The iterative process can also be terminated when the vector kurtosis stops increasing.

4.4 **Orthonormalization of a rectangular matrix**: This procedure used in subspace ICA is briefly explained here since it is slightly different from the regular vector orthonormalization procedure.

4.4.1 *Orthonormalization*: The orthonormalization of p rectangular matrix $\mathbf{W}_p \in \mathbf{\Lambda}$ can be computed by Gram-Schmidt process:

1. $\mathbf{W}_p \leftarrow \mathbf{W}_p - \sum_{j=1}^{p-1}\mathbf{W}_j\mathbf{W}_j^{\mathrm{T}}\mathbf{W}_p$ (orthogonalize **W**)
2. $\mathbf{W}_p \leftarrow \mathbf{W}_p(\mathbf{W}_p^{\mathrm{T}}\mathbf{W}_p)^{-1/2}$ (normalize **W**)

For orthonormalization of $\mathbf{W}_p$ check if the following two conditions are satisfied:

1. $\mathbf{W}_p^{\mathrm{T}}\mathbf{W}_p = \mathrm{I}_{\mathrm{d}\times\mathrm{d}}$
2. $(\mathbf{W}_i + \mathbf{W}_j)^{\mathrm{T}}(\mathbf{W}_i + \mathbf{W}_j) = \mathbf{W}_i^{\mathrm{T}}\mathbf{W}_i + \mathbf{W}_j^{\mathrm{T}}\mathbf{W}_j$ (from Pythagorean Theorem) or $\mathbf{W}_i^{\mathrm{T}}\,\mathbf{W}_j + \mathbf{W}_j^{\mathrm{T}}\,\mathbf{W}_i = 0_{\mathrm{d}\times\mathrm{d}}$

where $i = p$ and $j = p-1$ for $p \geq 2$. If the above two conditions are not satisfied then the Gram-Schmidt orthonormalization procedure should be repeated until both the conditions are satisfied or the values of $\mathbf{W}_p^{\mathrm{T}}\mathbf{W}_p$ and $\mathbf{W}_i^{\mathrm{T}}\mathbf{W}_j + \mathbf{W}_j^{\mathrm{T}}\,\mathbf{W}_i$ meet some predefined thresholds.

5. **Deflationary orthogonalization procedure for subspace ICA**: Deflationary orthogonalization procedure can be used to estimate subspace independent components one by one. We first estimate p matrices and then orthonormalize the obtained matrices prior to running the algorithm for $(p+1)^{\mathrm{th}}$ matrix. The size of matrix $\mathbf{W}_p$ is $\mathrm{Md}\times\mathrm{d}$. The procedure is illustrated as follows:

   1. Center data **x**.
   2. Whiten data **x** to give **z**.

[2] The term normalization for **W** is meant orthonormalization of the column vectors of **W**. Here we used this term to make distinction between the orthonormalization process of one **W** (say $\mathbf{W}_j$) with another (say $\mathbf{W}_k$) and to that of orthonormalization of column vectors within **W**.

3. Select M, the number of subspace independent components and dimension d for each of the subspaces. Set counter $p \leftarrow 1$.

4. Select an initial value of identity norm for $\mathbf{W}_p$, e.g. randomly.

5. Let $\mathbf{W}_p \leftarrow \mathrm{E}[(\mathbf{z}^{\mathrm{T}}\mathbf{W}_p\mathbf{W}_p^{\mathrm{T}}\mathbf{z})(\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W}_p)] - 3\mathrm{d}\,\mathbf{W}_p$.

6. Do orthonormalization for $\mathbf{W}_p$ (see section 4.4.1).

7. If $\mathbf{W}_p$ has not converged, go back to step 5.

8. Set $p \leftarrow p+1$ and go to step 4 until $p = \mathrm{M}$.

For special case, when $\mathrm{d} = 1$ (one-dimensional vector or scalar quantity) then the subspace ICA procedure will be reduced to the basic ICA procedure.

6. **Illustration using foetal ECG**: The subspace ICA model is illustrated on foetal ECG dataset [4]. The dataset consists of 2500 ECG points sampled at 500 Hz. We considered samples of four electrodes located on the abdomen of a pregnant woman. These observed samples are the mixtures of the cardiac rhythms of the mother and her foetus. The starting second of signals taken by each electrode are depicted in figure 1. In our model we assume two independent observations ($\mathrm{M} = 2$) and the dimension of each observation vector to be two as well (i.e. each observation vector has 2 non-independent components). From figure 1, row 1 and row 2 are assumed to be the 'first-subspace' and row 3 and row 4 are assumed to be the 'second-subspace'. Therefore row 1 and row 2 are dependent components; similarly row 3 and row 4 are dependent components. But dependencies between the two different subspaces are not allowed, i.e. they are considered as mutually independent.

The absolute of vector kurtosis ($|\mathrm{kurt}|$) for both 'first-subspace' and 'second-subspace' attains some finite value and converge after a few iterations. The figure of $|\mathrm{kurt}|$ versus iteration counts is not displayed here due to space limitations.

The subspace independent components estimated by subspace ICA method using vector kurtosis are depicted in figure 2. The first two rows of the figure show the cardiac rhythms of the mother and the last row shows the cardiac rhythms of the foetus. The third row of the figure does not precisely follow any cardiac rhythm and is thus considered as noise being emitted from the electrodes. It can be seen that subspace ICA is well modelled on ECG dataset and is able to extract hidden cardiac rhythms.

The subspace ICA model using vector kurtosis has estimated the rhythms in a similar fashion as MICA model [2] has on the same foetal ECG database. This proves the validity of our approach. Although some finer points remain unanswered at this stage (which we have included in the 'conclusion and future work' section), the prime objective of introducing the concept of vector kurtosis for subspace ICA model is achieved.

7. **Conclusion and future work**: We have presented a new perspective of subspace ICA algorithm. The subspace ICA model is derived by partitioning the multiplicative model of basic ICA. The idea of kurtosis is extended to vector kurtosis to solve generalized version of BSS problem, i.e. when dependent components are involved. The relationship between the optimization of vector kurtosis and subspace independent components, which enabled us to estimate subspace independent components by maximizing vector kurtosis is established. It is seen that the approach works well on ECG dataset. Some essential questions are included here under to be answered in future:

- How to appropriately select the value of d?
- If two or more signals are linearly dependent then it is possible to have reduced rank covariance matrix $\mathrm{E}[\mathbf{zz}^{\mathrm{T}}]$. How to apply the algorithm on reduced rank cases?
- How to select the value of M if the number of sources is completely unknown to the observer?

Acknowledgement: The authors gratefully acknowledge helpful consultations with Prof. Erkki Oja of Helsinki University of Technology, Finland.

Appendix 1:

Lemma: Let vector kurtosis $\mathrm{kurt}(\mathbf{W}^{\mathrm{T}}\mathbf{z}) = \mathrm{E}[(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})^2] - 3(\mathrm{E}[\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z}])^2$ be a differentiable function of an $m \times n$ rectangular matrix $\mathbf{W}$ for $m \geq n$; z be any vector of size $m \times 1$. The gradient of $\mathrm{kurt}(\mathbf{W}^{\mathrm{T}}\mathbf{z})$ is defined as $\nabla_{\mathbf{W}}\,\mathrm{kurt}(\mathbf{W}^{\mathrm{T}}\mathbf{z}) = 4\mathrm{E}[(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})(\mathbf{zz}^{\mathrm{T}}\mathbf{W})] - 12\mathrm{E}[\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z}]\mathrm{E}[\mathbf{zz}^{\mathrm{T}}\mathbf{W}]$. In the case of whitened $\mathbf{z}$ and normalized $\mathbf{W}$, the second term of the equation will be $12n\mathbf{W}$.

Proof: Let the scalar function be defined as $h(\mathbf{W}) = (\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})$. The derivative of h with respect to $\mathbf{W}$ will then be given as:

$$\frac{\partial(h(\mathbf{W}))}{\partial \mathbf{W}} = \frac{\partial}{\partial \mathbf{W}}(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z}) \quad \text{or} \quad \partial(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z}) = \partial(trace(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})) = trace(\mathbf{z}^{\mathrm{T}}\partial(\mathbf{W}\mathbf{W}^{\mathrm{T}})\mathbf{z})$$

$$= 2trace(\mathbf{zz}^{\mathrm{T}}\mathbf{W}(\partial\mathbf{W})^{\mathrm{T}})\ \{\text{since } tr(\mathbf{A}^{\mathrm{T}}) = tr(\mathbf{A}) \text{ and } tr(\mathbf{AB}) = tr(\mathbf{BA})\}$$

or $$h(\mathbf{W})' = 2(\mathbf{zz}^{\mathrm{T}}\mathbf{W}) \qquad \text{A1}$$

Therefore the derivative of vector kurtosis (from equation A1) can be written as:

$$\nabla_{\mathrm{W}}\,\mathrm{kurt}(\mathbf{W}^{\mathrm{T}}\mathbf{z}) = 2\mathrm{E}[h(\mathbf{W})h(\mathbf{W})'] - 6\mathrm{E}[h(\mathbf{W})]\mathrm{E}[h(\mathbf{W})']$$

$$= 4\mathrm{E}[(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})(\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W})] - 12\mathrm{E}[\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z}]\mathrm{E}[\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W}] \qquad \text{A2}$$

However, if data $\mathbf{z}$ is whitened ($\mathrm{E}[\mathbf{z}\mathbf{z}^{\mathrm{T}}] = \mathrm{I}_{m\times m}$) and rectangular matrix $\mathbf{W}$ is normalized ($\mathbf{W}^{\mathrm{T}}\mathbf{W} = \mathrm{I}_{n\times n}$) then equation A2 can be rewritten as:

$$\nabla_{\mathbf{W}}\,\mathrm{kurt}(\mathbf{W}^{\mathrm{T}}\mathbf{z}) = 4\mathrm{E}[(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})(\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W})] - 12n\mathbf{W} \qquad \text{A3}$$

$$\because\ \mathrm{E}[\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z}] = \mathrm{E}[\mathit{trace}(\mathbf{z}^{\mathrm{T}}\mathbf{W}\mathbf{W}^{\mathrm{T}}\mathbf{z})] = \mathit{trace}(\mathbf{W}^{\mathrm{T}}\mathrm{E}[\mathbf{z}\mathbf{z}^{\mathrm{T}}]\mathbf{W}) = \mathit{trace}(\mathbf{W}^{\mathrm{T}}\mathbf{W}) = \mathit{trace}(\mathrm{I}_{n\times n}) = n$$

and $\mathrm{E}[\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{W}] = \mathrm{E}[\mathbf{z}\mathbf{z}^{\mathrm{T}}]\mathbf{W} = \mathbf{W}$

References

[1]. A. Hyvärinen, P. Hoyer, "Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces", *Neural Comp.*, vol. 12, pp. 1705-1720, 2000.

[2]. J.-F. Cardoso, "Multidimensional independent component analysis", Proc. *ICASSP*, vol. 37, 1998.

[3]. A. Hyvärinen, J. Karhunen, E. Oja. Independent Component Analysis, *Wiley-Intersci. Pub.*, 2001.

[4]. B.L.R. De Moor, P.D. Gersem, B.D. Schutter, W. Favoreel (eds.), "Daisy: Database for the identification of systems", http://www.esat.kuleuven.ac.be/sista/daisy, 1997.

[5]. T. Kohonen, "Emergence of invariant-feature detectors in the adaptive-subspace self-organizing map", *Bio. Cyber.*, vol. 75, pp. 281-291, 1996.

[6]. A. Hyvärinen, E. Oja, "A fast fixed-point algorithm for independent component analysis", *Neural Comp.*, vol. 9, no. 7, pp. 1483-1492, 1997.

Figure 1: Observed ECG from 4 electrodes located on the abdomen of a pregnant woman

Figure 2: The estimated cardiac rhythms of the mother and her foetus using subspace ICA algorithm

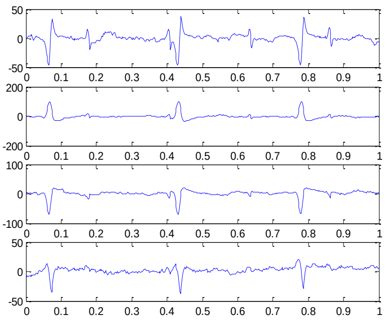

Figure 1

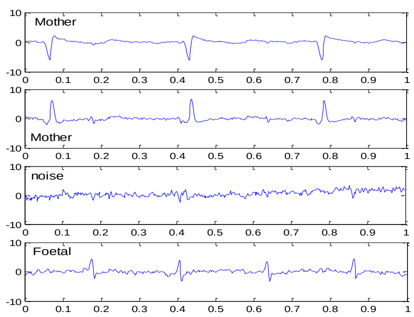

Figure 2