NONLINEAR INDEPENDENT COMPONENT ANALYSIS(ICA) USING POWER SERIES AND APPLICATION TO BLIND SOURCE SEPARATION

Ziyou Xiong and Thomas S. Huang

Beckman Institute for Advanced Science and Technology,
University of Illinois at Urbana-Champaign,
405 N. Mathews Av., Urbana, IL, 61081
E-mail: zxiong, huang@ifp.uiuc.edu

ABSTRACT

Contribution of this paper is the derivation of an algorithm that generalizes Bell & Sejnowski's classic ICA to tackle nonlinear ICA and the introduction of a new and efficient form of "natural gradient". This algorithm uses power series of non-linear mixtures to approximate the Taylor expansion of the inverse function mapping from sources to mixtures. The approximation enables derivation of learning rules for weight matrix associated with power series of any order. When applied to blind source separation, it successfully separated non-linear mixtures for which Bell & Sejnowski's algorithm could not due to its linear mixture model. In separating linear mixtures using this algorithm, the weight matrices for higher order mixtures converge to zero matrix. This is consistent with intuition, suggesting the validity of the generalization.

1. INTRODUCTION

Recently three classes of methods have been proposed for problems in which nonlinear mixtures of independent sources are to be separated, i.e, nonlinear ICA. The first class of methods adds nonlinear mixing model to the linear model [1] [2] [3] [4]. These methods resemble linear ICA methods very much with the only drastic difference being the introduction of unknown scaling and slope parameters to the nonlinear transfer function(recall that in ICA, this nonlinear transfer function is the same for all mixtures, e.g, logistic function $\frac{1}{1+e^{-u_i}}$). The drawback of introducing these parameters is that it limits the flexibility of the model. The second class of methods employ self-organizing maps(SOM)[5] to form a network structure equivalent to topology of the sources so that the SOM represents the inverse of the nonlinear transformation [6] [7]. SOM-based methods offer greater flexibility without introducing more parameters. One disadvantage of SOM-based approaches is its computational complexity. The third class of methods use traditional linear approximation to non-linear problems using Taylor expansion or other orthogonal expansion such as Fourier expansion and discrete wavelet expansion [8]. Some methods have been reported to use Taylor expansion around sources to approximate the nonlinear mixing process. The problem with these methods is that the sources are unknown beforehand hence the Taylor expansion is hard to do the around unknown sources. So far only an expansion to the second order has been shown in literature due to this difficulty. Our approach belongs to the third class of methods. Instead of expanding the nonlinear mapping from sources to mixtures, our expansion is done in the other direction, i.e, from mixtures to sources. This is justified by observations that expansion can be done around the given mixtures and that the mixing process can modeled by doing the Taylor expansion of the inverse of the un-mixing process.

This paper presents an algorithm that generalizes Bell & Sejnowski's [9] classic ICA to tackle nonlinear ICA. The generalization lies in that more weight matrices are introduced to higher order mixtures while in Bell & Sejnowski's work [9] there is only one weight matrix for the mixtures of first order and another one of zero-th order. The key idea of this algorithm is to use power series of non-linear mixtures to approximate the Taylor expansion of the inverse function mapping from independent sources to mixtures. This approximation enables the derivation of learning rules for weight matrix associated with power series of any order. Simulation results have shown success of this generalization. When applied to blind audio source separation, it successfully separated two non-linear mixtures for which Bell & Sejnowski's algorithm [9] could not, inherently due to its linear mixture model. Interestingly, in separating linear mixtures using this algorithm, the weight matrices for higher order mixtures converge to zero matrix. This is consistent with intuition, suggesting the validity of the generalization. And what's more, in contrast to the notion of "natural gradient" [10] by multiplying a symmetric matrix to

the right-hand side of the learning rules, another new and efficient notion of "natural gradient" is introduced by multiplying from the left-hand side of the rules.

The organization of the paper is as follows. In Section 2, an overview of the nonlinear ICA model is given, our approach is introduced. Generalization of Tony & Sejnowski's one-input-one-output network is shown in Section 3. Generalization of their N-input-N-output network is shown in Section 4. Section 5 details the experiment results. Some discussion on the weakness of the proposed algorithm is carried out in Section 6. Conclusions and future research can be found in Section 7.

2. NONLINEAR ICA MODEL & POWER SERIES APPROXIMATION

The N nonlinear mixed signals $x_1, ..., x_N$ are related to N independent source signals $s_1, ..., s_N$ through:

$$x_1 = f_1(s_1, ..., s_N)$$

$$x_2 = f_2(s_1, ..., s_N)$$

$$...$$

$$x_N = f_N(s_1, ..., s_N).$$

This can be denoted in vector form:

$$\mathbf{X} = \mathbf{F}(\mathbf{S}) \tag{1}$$

To reconstruct the original signals, another nonlinear transformation is applied to $x_1, ..., x_N$ to get $u_1, ..., u_N$ through:

$$u_1 = h_1(x_1, ..., x_N)$$

$$u_2 = h_2(x_1, ..., x_N)$$

$$...$$

$$u_N = h_N(x_1, ..., x_N).$$

Or equivalently:

$$\mathbf{U} = \mathbf{H}(\mathbf{X}) \tag{2}$$

Hopefully $u_1, ..., u_N$ can be a good approximation to $s_1, ..., s_N$ subject to permutation and scaling and statistically independent of each other.

Our proposed approach to this problem is using power series of $x_1, ..., x_N$ to approximate the Taylor expansion of functions $h_1, ..., h_N$. The approximated relationship between $u_1, ..., u_N$ and $x_1, ..., x_N$ is expressed as:

$$\mathbf{U} = \mathbf{W_0} + \mathbf{W_1} \times \mathbf{X} + \mathbf{W_2} \times \mathbf{X^2} + \ldots + \mathbf{W_n} \times \mathbf{X^n} + \ldots \tag{3}$$

where $\mathbf{W_0}$ is a bias-weight vector and $\mathbf{W_n}$ is the weight matrix associated with the n-th power of $\mathbf{X}$, i.e, $\mathbf{X}^n$. To clarify notations, $\mathbf{W_0}$, $\mathbf{W_n}$ and $\mathbf{X}^n$ are expressed in the following matrix and vector form:

$$\mathbf{W_0} = \begin{pmatrix} W_{01} \\ \vdots \\ W_{0N} \end{pmatrix}. \tag{4}$$

$$\mathbf{W_n} = \begin{pmatrix} (w_n)_{11} & \ldots & (w_n)_{1N} \\ \vdots & & \vdots \\ (w_n)_{N1} & \ldots & (w_n)_{NN} \end{pmatrix} \tag{5}$$

$$\mathbf{X^n} = \begin{pmatrix} x_1^n \\ \vdots \\ x_N^n \end{pmatrix}. \tag{6}$$

Then $\mathbf{U}$ is passed through a transformation function $\mathbf{G}$ to give an output vector $\mathbf{Y}$, i.e,

$$\begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} \frac{1}{1+e^{-u_1}} \\ \vdots \\ \frac{1}{1+e^{-u_N}} \end{pmatrix}. \tag{7}$$

Here the logistic function $\mathbf{Y} = \frac{1}{1+e^{-\mathbf{U}}}$ is used although other strictly increasing or decreasing functions can also be used such as tanh.

In order to maximize the mutual information between **X** and **U**, i.e, the sources and the mixtures, it is equivalent to maximize the mutual information between **X** and **Y** since the mapping from **U** to **Y** is deterministic. Bell & Sejnowski [9] have shown the above information maximization is also equivalent to maximizing the differential entropy [11] of the output **Y** alone. The generalization here is that weight matrices are introduced to higher power terms of **X** hence more learning needs to be done.

Derivation of the learning rules for the weight matrices $\mathbf{W_0}, \mathbf{W_1}, ..., \mathbf{W_n}, ...$ is by a similar method to Bell & Sejnowski's. Some key differences are also worth discussing. The derivation starts now.

3. GENERALIZATION OF TONY & SEJNOWSKI'S ONE-INPUT-ONE-OUTPUT NETWORK

In the case of one-input-one-output network from x to u and transfer function from u to y being logistic, their relationships can be expressed as $u = w_0 + w_1 x + w_2 x^2 + ... + w_n x^n + ...$ and $y = \frac{1}{1+e^{-\mathbf{u}}}$. The stochastic gradient ascent rule

$$\Delta w_i \propto \frac{\partial H(y)}{\partial w_i} = \frac{\partial}{\partial w_i}\left(ln\left|\frac{\partial y}{\partial x}\right|\right) = \left(\frac{\partial y}{\partial x}\right)^{-1} \frac{\partial}{\partial w_i}\left(\frac{\partial y}{\partial x}\right) \tag{8}$$

gives us the following learning rules:

$$\Delta w_0 \propto (1 - 2y) \tag{9}$$

$$\Delta w_1 \propto \frac{x}{w_1 + 2w_2x + ... + nw_nx^{n-1} + ...} + (1-2y)x \quad (10)$$

$$\Delta w_2 \propto \frac{2x}{w_1 + 2w_2x + ... + nw_nx^{n-1} + ...} + (1-2y)x^2 \quad (11)$$

...

$$\Delta w_n \propto \frac{nx^{n-1}}{w_1 + 2w_2x + ... + nw_nx^{n-1} + ...} + (1-2y)x^n \quad (12)$$

where the following useful equation is used:

$$\frac{\partial y}{\partial x} = y(1-y)\frac{\partial u}{\partial x} = y(1-y)(w_1+2w_2x+...+nw_nx^{n-1}+...). \quad (13)$$

In one-input-one-output network case, the above results are the direct generalization of Bell & Sejnowski's [9] in that if only w_0, w_1 are used to form u then the derivation becomes identical.

The above derivation serves as good guideline for the N-input-N-output network problem, which is much more mathematically involved.

4. GENERALIZATION ON N-INPUT-N-OUTPUT NETWORK

In the case of N-input-N-output network from **X** to **U** and transfer function from **U** to **Y** again being logistic, the stochastic gradient ascent rule becomes:

$$\Delta \mathbf{W_i} \propto \frac{\partial}{\partial \mathbf{W_i}} (ln\,|\mathbf{J}|) \quad (14)$$

The Jacobian **J** of the transformation [12] from **X** to **Y** can be shown to have the following relation with $\mathbf{W_1}, ..., \mathbf{W_n}, ...$:

$$\mathbf{J} = (det\mathbf{K}) \times \prod_{n=1}^{\infty} y_i^{'} \quad (15)$$

where

$$y_i^{'} = \frac{\partial y_i}{\partial u_i} \quad (16)$$

$$\mathbf{K} = \mathbf{W_1} + 2\mathbf{W_2}diag(\mathbf{X}) + ... + n\mathbf{W_n}diag(\mathbf{X^{n-1}}) + ... \quad (17)$$

and diag(**X**) represents forming a diagonal matrix using the elements of vector **X**, i.e,

$$diag(\mathbf{X}) = \begin{pmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \vdots & & & \vdots \\ 0 & \dots & \dots & x_N \end{pmatrix} \quad (18)$$

The learning rules for $\mathbf{W_0}, \mathbf{W_1}, ..., \mathbf{W_n}, ...$ can be shown to be the following:

$$\Delta \mathbf{W_0} \propto \mathbf{1} - \mathbf{2Y} \quad (19)$$

$$\Delta \mathbf{W_1} \propto (\mathbf{K}^T)^{-1} + (\mathbf{1} - \mathbf{2Y})\mathbf{X^T} \quad (20)$$

$$\Delta \mathbf{W_2} \propto 2(\mathbf{K}^T)^{-1}diag(\mathbf{X}) + (\mathbf{1} - \mathbf{2Y})(\mathbf{X^2})^{\mathbf{T}} \quad (21)$$

...

$$\Delta \mathbf{W_n} \propto n(\mathbf{K}^T)^{-1}diag(\mathbf{X^{n-1}}) + (\mathbf{1} - \mathbf{2Y})(\mathbf{X^n})^{\mathbf{T}} \quad (22)$$

...

where **1** is a vector of ones. A skeleton of the derivation is in the Appendix. The extreme similarity of the above results to the ones in one-input-one-output network can help to understand these results.

These results resemble Bell & Sejnowski's [9] when only $\mathbf{W_0}, \mathbf{W_1}$ are taken into account. However a major difference between these results and theirs is that weight matrices $\mathbf{W_0}, \mathbf{W_1}, ..., \mathbf{W_n}, ...$ appear in the update rules for $\mathbf{W_0}, \mathbf{W_1}, ..., \mathbf{W_n}, ...$ due to the Jacobian term **J**. Another big difference is the appearance of multiplication from right-hand side of the diagonal matrix terms. This makes it inappropriate to use the ordinary notion of "natural gradient" [10] for the learning rules to avoid matrix inversion and speed up convergence. The introduction of "natural gradient" by Amari et al has been a big step in the research of ICA where a term $\mathbf{W^TW}$ is multiplied to the learning rules from the right-hand side of the equations. Instead we propose that a term $\mathbf{KK^T}$ be multiplied to the above rules from the left-hand side to avoid matrix inversion. Theoretical impact of this difference is worth further investigation. Doing this generates another set of learning rules:

$$\Delta \mathbf{W_0} \propto \mathbf{KK^T}(\mathbf{1} - \mathbf{2Y}) \quad (23)$$

$$\Delta \mathbf{W_1} \propto \mathbf{K} + \mathbf{KK^T}(\mathbf{1} - \mathbf{2Y})\mathbf{X^T} \quad (24)$$

$$\Delta \mathbf{W_2} \propto 2\mathbf{K}diag(\mathbf{X}) + \mathbf{KK^T}(\mathbf{1} - \mathbf{2Y})(\mathbf{X^2})^{\mathbf{T}} \quad (25)$$

...

$$\Delta \mathbf{W_n} \propto n\mathbf{K}diag(\mathbf{X^{n-1}}) + \mathbf{KK^T}(\mathbf{1} - \mathbf{2Y})(\mathbf{X^n})^{\mathbf{T}} \quad (26)$$

Experiment results validate this multiplication from the left-hand side. Details are in Section 5.

5. EXPERIMENT RESULTS ON BLIND SOURCE SEPARATION

5.1. Separation with Linear Mixture Using Nonlinear ICA

4 speech clips from a set of test data clips provided by Prof. Dan Ellis at Columbia University are mixed by a randomly generated linear matrix **M**. These clips are available at http://www.ee.columbia.edu/∼dpwe. The pre-whitened mixtures are fed into the nonlinear ICA algorithms developed above. The pre-whitening is done using linear Principle Component Analysis(PCA).

In one experiment, only power series up to second order without bias-weight term, i.e, only $\mathbf{W_1}, \mathbf{W_2}$ are used. First, the algorithm using the Euclidean gradient is tested, i.e, the weight update uses Equation (19), ..., (22). Intuitively, if the original statically independent sources are to be separated, the matrix $\mathbf{W_2}$ should converge to zero matrix and $\mathbf{W_1}$ should converge to contribute mostly to the reconstruction of the original sources. Experiment results confirm this intuition. An example of the final values of $\mathbf{W_1}, \mathbf{W_2}$ after sweeping the speech samples once are:

$$\mathbf{W_1} = \begin{pmatrix} 0.8903 & 0.0496 & -0.0051 & 0.3162 \\ 0.1888 & -0.8532 & 0.1240 & -0.4627 \\ 0.2025 & 0.3794 & -0.1472 & -0.7502 \\ 0.0709 & 0.1607 & 1.2688 & -0.1179 \end{pmatrix} \quad (27)$$

$$\mathbf{W_2} = \begin{pmatrix} 0.0566 & 0.0123 & 0.0292 & -0.0346 \\ 0.0247 & -0.0030 & 0.0118 & -0.0281 \\ 0.0038 & 0.0030 & 0.0113 & -0.0478 \\ 0.0388 & -0.0097 & 0.0910 & -0.0019 \end{pmatrix} \quad (28)$$

Most of the elements in $\mathbf{W_1}$ are 10 times greater in magnitude than those in $\mathbf{W_2}$. $\mathbf{W_2}$ shows the trend of converging to zero matrix. The product of $\mathbf{W_1}$ and the above mixing matrix after undoing the pre-whitening is shown to be:

$$\mathbf{W_1 M} = \begin{pmatrix} -0.2487 & 21.5954 & 0.2877 & 2.2212 \\ 0.1901 & -0.5208 & -23.3314 & 0.8790 \\ -1.0601 & 1.4622 & 0.0815 & -25.3910 \\ -20.4493 & -0.1686 & 0.7512 & 1.7344 \end{pmatrix} \quad (29)$$

Listening test shows clearly the separation of the mixtures and almost perfect reconstruction of the originals.

When the weight matrices are updated by Equation (23), ..., (26), same conclusion about $\mathbf{W_1}, \mathbf{W_2}$ can be drawn and moreover the convergence is greatly sped up plus the quality of the reconstruction of the original speeches is even better. To illustrate this, the product of $\mathbf{W_1 M}$ in this case is shown below:

$$\mathbf{W_1 M} = \begin{pmatrix} -0.2288 & -20.3313 & -0.1306 & 0.9836 \\ 0.4531 & -0.9162 & 0.0764 & -30.4944 \\ 0.4756 & -0.8545 & 23.3622 & 0.4823 \\ 19.3270 & -0.1847 & 0.0156 & -1.1254 \end{pmatrix} \quad (30)$$

The masking of the target speech over interference speeches can be seen to be more dramatic in Equation (30) than Equation (29). This can explain why the quality of the reconstruction of the original speeches is better using the "natural gradient". Once again, notice that this "natural gradient" is implemented by multiplying $\mathbf{KK^T}$ from the left-hand side.

5.2. Separation with Nonlinear Mixtures

Two audio clips from the same collection of test data as above are non-linearly mixed manually. One example of the nonlinear relationship used is as follows:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} f_1(s_1, s_2) \\ f_2(s_1, s_2) \end{pmatrix} = \begin{pmatrix} 2 \times s_1 \\ 3 \times s_1^2 + s_2 \end{pmatrix} \quad (31)$$

This examples gives a rather simple reconstruction formula:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} h_1(x_1, x_2) \\ h_2(x_1, x_2) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} x_1 \\ x_2 - \frac{3}{4} x_1^2 \end{pmatrix} \quad (32)$$

That is,

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ -\frac{3}{4} & 0 \end{pmatrix} \begin{pmatrix} x_1^2 \\ x_2^2 \end{pmatrix} \quad (33)$$

The mixtures were not able to be separated by Bell & Sejnowski's [9] algorithm due to the inherent nonlinearity of the problem and the nature of their linear algorithm. Our nonlinear algorithm with our version of "natural gradient", however successfully separates out the independent sources and the listening tests show almost perfect masking of the interfering speech. $\mathbf{W_1}, \mathbf{W_2}$ are shown to be of the following after convergence one sweep through all the speech samples:

$$\mathbf{W_1} = \begin{pmatrix} 0.6320 & -0.0749 \\ 0.1681 & 0.7261 \end{pmatrix} \quad (34)$$

$$\mathbf{W_2} = \begin{pmatrix} 0.1156 & -0.1055 \\ 0.2310 & -0.0178 \end{pmatrix} \quad (35)$$

The above $\mathbf{W_1}, \mathbf{W_2}$ resemble the ideal weight matrices subject to scaling and sign change.

Interestingly, when an order 3 algorithm is applied to this problem, the weight matrix associated with power series of the third power also converges to zero matrix as it is the case when a nonlinear ICA is used in linear mixture case. This shows that the algorithm successfully captures the quadratic relationship between **U** and **X**. The experiment gives essentially the same $\mathbf{W_1}, \mathbf{W_2}$ and the following values for element of matrix $\mathbf{W_3}$:

$$\mathbf{W_3} = \begin{pmatrix} -0.0306 & -0.0199 \\ -0.0091 & -0.0392 \end{pmatrix} \quad (36)$$

The proposed "natural gradient" by multiplying $\mathbf{KK^T}$ is also shown to be effective.

6. DISCUSSION

The proposed algorithm using power series for nonlinear ICA has shown some success in certain applications. However the algorithm is still quite limited for real world nonlinear ICA problems. This is due to at least the following two reasons.

For one, it is difficult to decide the accuracy of the approximation when power series are used to approximate the

Taylor expansion. Ideally Equation (3) should involve many cross production terms such as x_1x_2, x_1x_n and x_mx_n. One example is used to better illustrate this. A second order Taylor expansion takes the following form:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} a_1 + b_{11}x_1 + b_{12}x_2 + c_1x_1x_2 \\ +d_{11}x_1^2 + d_{12}x_2^2 \\ a_2 + b_{21}x_1 + b_{22}x_2 + c_2x_1x_2 \\ +d_{21}x_1^2 + d_{22}x_2^2 \end{pmatrix} \quad (37)$$

or another matrix form:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & c_1 & d_{11} & d_{12} \\ b_{21} & b_{22} & c_2 & d_{21} & d_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_1x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix} \quad (38)$$

Altogether there are 12 weight elements that need to be updated. But the power series expansion is Equation (3) only update 10 of them, ignoring the ones(c_1, c_2) for cross production term x_1x_2. The approximation can be worse when even higher power series are used. The justification on using power series as in Equation (3) may be its elegant derivation of the learning rules for all square matrices as shown in Equation (19), . . ., (22) or Equation (23), . . ., (26) rather than learning a rectangular matrix as in Equation (38), which has been shown to be of much higher complexity[13]. Our experiment results do show the degradation of performance when higher and higher order power series are used for Taylor expansion, partly due to this reason.

For another, the introduction of higher order power series causes a problem that is not in accordance with the assumption that statistically independent un-mixed outputs are the original sources, i.e, the independence criteria is not enough to recover the original sources. For example, $u_1 = s_1 + s_1^2, u_2 = s_2^3$ are statistically independent and possibly be the recovered signals, but obviously they are not the original sources. This is also one problem for the Taylor expansion based approaches, as explained on Page 134-135 in [8]. More research is needed to find other criteria to solve this problem.

7. CONCLUSION AND FUTURE WORK

Contribution of this paper is the derivation of an algorithm that generalizes Bell & Sejnowski's [9] classic ICA to tackle nonlinear ICA and the introduction of a new and efficient form of "natural gradient". The proposed algorithm has shown some success in certain applications. However the algorithm is still quite limited for real world nonlinear ICA problems due to degradation of the approximation accuracy when higher and higher power series are used and the fact that the independence criteria is not enough to recover the original sources when power series are used without the introduction of other criteria.

Several directions are ahead. One is the theoretic study on the introduction of new criteria to solve the inherent problem pointed out in Section 6 and [8]. Another is the application of the derived algorithm to the eigen-face derivation and face recognition project at University of Illinois at Urbana-Champaign(see http://www.ifp.uiuc.edu for more detail), at least using the lower order power series. The third one is the application of nonlinear ICA to over-complete representations in which fewer mixtures than sources are available [13]. Our approach will rely on making as many mixtures as sources by decomposing a mixtures into two using digital wavelet transformation. The manually-made mixtures, having the same number of mixtures as sources, are complex nonlinear mixtures.

8. APPENDIX: PROOF OF EQUATION (19), . . ., (22)

The proof is similar to that by Bell & Sejnowski [9] with some subtle difference. First notice **K** in Equation (17) is the derivative of **U** in Equation (3). Next, the stochastic gradient can be shown to be:

$$\Delta \mathbf{W_i} \propto \frac{\partial H(\mathbf{Y})}{\partial \mathbf{W_i}} = \frac{\partial}{\partial \mathbf{W_i}} ln\left|\mathbf{J}\right| \quad (39)$$

or,

$$\Delta \mathbf{W_i} \propto \frac{\partial}{\partial \mathbf{W_i}} ln\left|det\mathbf{K}\right| + \frac{\partial}{\partial \mathbf{W_i}} ln \prod_i \left|y_i^{'}\right|. \quad (40)$$

Assume $\mathbf{W_i}$'s are pair-wise independent variables for all i and all $(w_i)_{jk}$'s are also pair-wise independent for any fixed i and all j, k. The following equation can be obtained:

$$\frac{\partial}{\partial (w_n)_{ij}} ln\left|det\mathbf{K}\right| = \frac{cofactor((w_n)_{ij}) n x_i^{n-1}}{det\mathbf{K_i}}. \quad (41)$$

In vector form, it is:

$$\frac{\partial}{\partial \mathbf{W_n}} ln\left|det\mathbf{K}\right| = n\frac{(adj\mathbf{K})^T}{det\mathbf{K}}\mathbf{X}^{n-1} = n\left(\mathbf{K}^T\right)^{-1}\mathbf{X}^{n-1}. \quad (42)$$

which gives the first term in Equation (22). Notice the term nx_i^{n-1} in Equation (41) leads into the difference between the final result here and Bell & Sejnowski's.

The second term in Equation (22) follows the same argument as in Bell & Sejnowski [9]. The derivation is thus omitted.

9. REFERENCES

[1] G. Burel, "A non-linear neural algorithm," *Neural networks*, vol. 5, pp. 937–947, 1992.

[2] B. Koehler T.-W. Lee and R. Orglmeister, "Blind separation of nonlinear mixing models," *IEEE NNSP*, pp. 406–415, 1997.

[3] A. Taleb and C. Jutten, "Nonlinear source separation: The post-nonlinear mixtures," *ESANN*, pp. 279–284, 1997.

[4] S. Amari H. Yang and A. Cichocki, "Information back-propagation for blind separation of sources from non-linear mixtures," *Proc. of ICNN*, pp. 2141–2146, 1997.

[5] O. Simula T. Kohonen, E. Oja and A. Visa, "Engineering application of the self-organizing maps," *Proceedings of the IEEE*, vol. 84, no. 10, 1996.

[6] M. Hermann and H. Yang, "Perspectives and limitations of self-organizing maps," *ICONIP'96*, 1996.

[7] P. Pajunen, "Nonlinear independent component analysis by self-organizing maps," *Technical report, Proc. ICANN*, 1996.

[8] T.-W. Lee, *Independent Component Analysis: Thoery and Application*, Kluwer Academic Publishers, 1998.

[9] A.J. Bell and T.J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, pp. 1129–1159, 1995.

[10] A. Cichocki S. Arari and H.H. Yang, "A new learning algorithm for blind source separation," *Advances in Neural Information Processing*, vol. 8, pp. 767–763, 1996.

[11] T. Cover and J. Thomas, *Elements of Information Thoery*, vol. 1, John Wiley and Sons, 1990.

[12] A. Papoulis, *Probability and Statistics*, vol. 1, Prentice Hall, 1990.

[13] M.S. Lewichi and T.j. Sejnowski, "Learning overcomplete representations," *Neural Computation*, vol. 12, pp. 337–365, 2000.