Nonlinear Independent Component Analysis with Minimal Nonlinear Distortion

Kun Zhang KZHANG@CSE.CUHK.EDU.HK
Laiwan Chan LWCHAN@CSE.CUHK.EDU.HK
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong

Abstract

Nonlinear ICA may not result in nonlinear blind source separation, since solutions to nonlinear ICA are highly non-unique. In practice, the nonlinearity in the data generation procedure is usually not strong. Thus it is reasonable to select the solution with the mixing procedure close to linear. In this paper we propose to solve nonlinear ICA with the "minimal nonlinear distortion" principle. This is achieved by incorporating a regularization term to minimize the mean square error between the mixing mapping and the best-fitting linear one. As an application, the proposed method helps to identify linear, non-Gaussian, and acyclic causal models when mild nonlinearity exists in the data generation procedure. Using this method to separate daily returns of a set of stocks, we successfully identify their linear causal relations. The resulting causal relations give some interesting insights into the stock market.

1. Introduction

Independent component analysis (ICA) aims at recovering independent sources from their mixtures, without knowing the mixing procedure or any specific knowledge of the sources. If the sources are linearly mixed, under weak assumptions, ICA can recover the original sources with the trivial permutation and scaling indeterminacies. ICA is currently a popular method for blind source separation (BSS) of linear mixtures. However, nonlinear ICA does not necessarily lead to nonlinear BSS. Hyvärinen and Pajunen (1999) showed that solutions to nonlinear ICA always exist, and that they are highly non-unique. In fact, nonlinear BSS is impossible without additional prior knowledge on the mixing model, since the independence assumption is not strong enough in the general nonlinear mixing case (Jutten & Taleb, 2000).

If we constrain the nonlinear mixing mapping to have some particular forms, the indeterminacies in the results of nonlinear ICA can be reduced dramatically, and as a consequence, in these cases nonlinear ICA may lead to nonlinear BSS. But sometimes, the form of the nonlinear mixing procedure may be unknown. Consequently, in order to model arbitrary nonlinear mappings, one may need to resort to a flexible nonlinear function approximator, such as the multi-layer perceptron (MLP) or the radius basis function (RBF) network, to represent the nonlinear separation system. In this situation, in order to achieve BSS, nonlinear ICA requires extra constraints or regularization.

Almeida (2003) used the MLP to model the separation system and trains the MLP by information-maximization (Infomax). Moreover, smoothness provided by the MLP was believed to be a suitable regularization condition to achieve nonlinear BSS. But it seems not sufficient, as shown by the counterexample in Jutten and Karhunen (2003). In Tan et al. (2001), a RBF network is utilized to represent the separation system, and partial moments of the outputs are used for regularization. The matching between the relevant moments of the outputs and those of the original sources was expected to guarantee a unique solution. But the moments of the original sources may be unknown. In addition, if the transformation from the original sources to the recovered signals is non-trivial,[1] it seems that this regularization could not help to recover the original sources. Variational Bayesian nonlinear ICA (Valpola, 2000) utilizes the MLP to model the nonlinear mixing transformation. By resorting

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

[1] Roughly speaking, a trivial transformation of $\mathbf{y} = (y_1, y_2, \cdots, y_n)^T$ is a component-wise invertible transformation on a permuted version of y_i (Jutten & Taleb, 2000).

to the variational Bayesian inference technique, this method can do model selection and avoid overfitting. If we can find some additional knowledge about the nonlinear mixing transformation, the results of nonlinear ICA will be much more meaningful and reliable.

Although we may not know the form of the nonlinearity in the data generation procedure, fortunately, the nonlinearity in the generation procedure of natural signals is usually not strong. Hence, provided that the nonlinear ICA outputs are mutually independent, we would prefer the solution with the mixing transformation as close as possible to linear. This information can help to reduce the indeterminacies in nonlinear ICA greatly, and will be incorporated to solve the nonlinear ICA problem in this paper. In this paper we utilize the MLP to represent the separation system, and we should address that the analysis also applies if other flexible nonlinear models are adopted.

The minimal nonlinear distortion (MND) of the mixing system is achieved by the regularization technique. The objective function is the mutual information between outputs penalized by a regularization term measuring the level of "closeness to linear" of the mixing system. The mean square error (MSE) between the nonlinear mixing mapping and its best-fitting linear one provides such a term. A related regularizer is the smoothness regularizer exploiting second-order partial derivatives. We show that this regularizer actually indicates the local "closeness to linear" of the nonlinear function averaged at every point.

2. Nonlinear ICA with Minimum Nonlinear Distortion

2.1. Nonlinear ICA

Assume that the observed data $\mathbf{x} = (x_1, \cdots, x_n)^T$ are generated from an independent random vector $\mathbf{s} = (s_1, \cdots, s_n)^T$ by a nonlinear transformation $\mathbf{x} = \mathcal{F}(\mathbf{s})$, where $\mathcal{F}$ is an unknown real-valued n-component mixing function. Here for simplicity, we have assumed that the number of observed variables equals that of the original independent variables. The general nonlinear ICA problem is to find a mapping $\mathcal{G} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{y} = \mathcal{G}(\mathbf{x}; \boldsymbol{\theta})$ has statistically independent components. Nonlinear ICA is an ill-posed problem and its solutions are highly non-unique. In order to achieve nonlinear BSS, which aims at recovering the original independent sources s_i, we must have additional prior information or suitable regularization constraints.

2.2. With Minimum Nonlinear Distortion

Inspired by the fact that in practice the nonlinearity in the data generation procedure is usually not very strong, we propose the "minimal nonlinear distortion" (MND) principle to alleviate the ill-posedness of the nonlinear ICA problem. That is, under the condition that the separation outputs y_i are mutually independent, we prefer the solution with the corresponding mixing transformation $\mathcal{F}$ as close as possible to linear.

Now we need a measure of "closeness to linear" of a mapping. Given a nonlinear mapping $\mathcal{F}$, its deviation from the affine mapping $\mathbf{A}^*$, which fits $\mathcal{F}$ best among all affine mappings $\mathbf{A}$, is an indicator of its "closeness to linear", or the level of its nonlinear distortion. The deviation can be measured in various ways. The MSE is adopted here, as it greatly facilitates subsequent analysis. Consequently, the "closeness to linear" of $\mathcal{F} = \mathcal{G}^{-1}$ can be measured by the MSE between $\mathcal{G}^{-1}$ and $\mathbf{A}^*$. We denote this measure by $R_{MSE}(\boldsymbol{\theta})$. Let $\mathbf{x}^* = (x_1^*, \cdots, x_n^*)^T$ be the output of the affine transformation from $\mathbf{y}$ by $\mathbf{A}^*$. Let $\tilde{\mathbf{y}} = [\mathbf{y}; 1]$. $R_{MSE}(\boldsymbol{\theta})$ can then be written as the MSE between x_i and x_i^*:

$$R_{MSE}(\boldsymbol{\theta}) = E\{(\mathbf{x} - \mathbf{x}^*)^T(\mathbf{x} - \mathbf{x}^*)\} \text{ , where} \quad (1)$$
$$\mathbf{x}^* = \mathbf{A}^*\tilde{\mathbf{y}}, \text{ and } \mathbf{A}^* = \arg_{\mathbf{A}} \min E\{(\mathbf{x} - \mathbf{A}\mathbf{y})^T(\mathbf{x} - \mathbf{A}\mathbf{y})\}$$

Here $\mathbf{A}^*$ is a $n \times (n+1)$ matrix.[2] Figure 1 shows the separation system $\mathcal{G}$ together with the generation process of R_{MSE}. With R_{MSE} measuring the level of nonlinear distortion, nonlinear ICA with MND is a constrained optimization problem; it aims to minimize the mutual information between outputs, i.e. $I(\mathbf{y})$, subject to $R_{MSE}(\boldsymbol{\theta}) \leq t$, where t is a pre-assigned parameter. This is equivalent to minimizing

$$J = I(\mathbf{y}) + \lambda R_{MSE}(\boldsymbol{\theta}) \quad (2)$$

where λ is the regularization parameter.

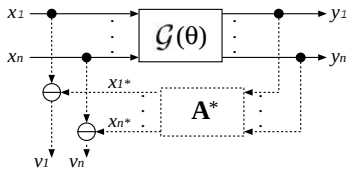


Figure 1. Separation system $\mathcal{G}$ (solid line) and generation of R_{MSE} (dashed line). $R_{MSE} = \sum_{i=1}^{n} v_i^2$, where $v_i = x_i - x_i^*$. Here it is assumed that $\mathbf{x}$ and $\mathbf{y}$ are zero-mean; consequently $\mathbf{x}^* = \mathbf{A}^*\mathbf{y}$ and $\mathbf{A}^*$ is $n \times n$ (see Footnote 2).

3. With $\mathcal{G}$ modelled by a MLP

MND can be incorporated in many nonlinear ICA methods to avoid unwanted solutions. In particular, here we adopt the MLP to represent the de-mixing

[2] If $E(\mathbf{y}) = E(\mathbf{x}) = \mathbf{0}$, $\mathbf{x}^*$ can be obtained as $\mathbf{x}^* = \mathbf{A}^*\mathbf{y}$ instead, and here $\mathbf{A}^*$ is a $n \times n$ matrix.

mapping $\mathcal{G}$, just as the MISEP method (Almeida, 2003) does. This method is an extension of the Infomax method for linear ICA (Bell & Sejnowski, 1995). Figure 2 shows the structure used in this method.

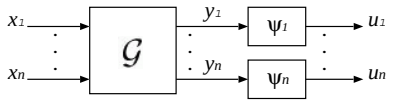


Figure 2. Network structure used in Infomax and MISEP. $\mathcal{G}$ is the separation system, and ψ_i are the nonlinearities applied to the separated signals.

With the Infomax principle, parameters in $\mathcal{G}$ and ψ_i are learned by maximizing the joint entropy of the outputs of the structure in Figure 2, i.e. $H(\mathbf{u}) = H(\mathbf{x}) + E\{\log|\det \mathbf{J}|\}$, where $\mathbf{J} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ is the Jacobian of the transformation from $\mathbf{x}$ to $\mathbf{u}$. As $H(\mathbf{x})$ does not depend on the parameters in $\mathcal{G}$ and ψ_i, maximizing $H(\mathbf{u})$ is equivalent to the minimization of $-E\{\log|\det \mathbf{J}|\}$. The resulting learning rule for $\boldsymbol{\theta}$, the parameters in $\mathcal{G}$, is the same as that obtained by minimizing $I(\mathbf{y})$. It was derived in Almeida (2003), in a manner similar to the back-propagation algorithm.

The MLP adopted in this paper has linear output units and a single hidden layer with activation function h. Direct connections between the inputs and output units are also permitted. Let $\mathbf{a} = (a_1, \cdots, a_M)^T$ and $\mathbf{z} = (z_1, \cdots, z_M)^T$ be the inputs and outputs of the hidden units, and $\mathbf{W}$ and $\mathbf{b}$ denote the weights and biases, respectively. We use superscripts to distinguish the locations of these parameters. The output of the $\mathcal{G}$ network in Figure 2 takes the form:

$$\begin{aligned} \mathbf{y} &= \mathbf{W}^{(d)} \cdot \mathbf{x} + \mathbf{W}^{(2)} \cdot \mathbf{z} + \mathbf{b}^{(2)}, \text{ where} \\ z_i &= h(a_i), \text{ and } \mathbf{a} = \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \end{aligned} \tag{3}$$

3.1. Learning Rule

Now we incorporate MND into the above nonlinear ICA method. The objective function becomes Eq. 2. The learning rule for $\boldsymbol{\theta}$ to minimize the first term in Eq. 2 has been considered in Almeida (2003); hence here we just give the gradient of R_{MSE} w.r.t $\boldsymbol{\theta}$.

According to Eq. 1, we have

$$\frac{\partial R_{MSE}}{\partial \mathbf{A}^*} = -2E\{(\mathbf{x} - \mathbf{A}^*\tilde{\mathbf{y}})\tilde{\mathbf{y}}^T\} \tag{4}$$

Setting the derivative to $\mathbf{0}$ gives $\mathbf{A}^*$:

$$\begin{aligned} & E\{(\mathbf{x} - \mathbf{A}^*\tilde{\mathbf{y}})\tilde{\mathbf{y}}^T\} = \mathbf{0} \\ \Longleftrightarrow \quad & \mathbf{A}^* = E\{\mathbf{x}\tilde{\mathbf{y}}^T\}[E\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\}]^{-1} \end{aligned} \tag{5}$$

We can see that due to the adoption of MSE, $\mathbf{A}^*$ can be obtained in closed form, which greatly simplifies the derivation.

R_{MSE} can then be written as

$$\begin{aligned} R_{MSE} &= \mathrm{Tr}\big(E\{(\mathbf{x} - \mathbf{A}^*\tilde{\mathbf{y}})(\mathbf{x} - \mathbf{A}^*\tilde{\mathbf{y}})^T\}\big) \\ &= -\mathrm{Tr}\big(E\{\mathbf{A}^*\tilde{\mathbf{y}}\mathbf{x}^T\}\big) + \text{const} \\ &= -\mathrm{Tr}\big(E\{\mathbf{x}\tilde{\mathbf{y}}^T\}[E\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\}]^{-1}E\{\tilde{\mathbf{y}}\mathbf{x}^T\}\big) + \text{const} \end{aligned}$$

Since y_i are independent from each other, they are uncorrelated. Moreover, we can easily make y_i zero-mean by adjusting $\mathbf{b}^{(2)}$, the biases in the output layer. Consequently, $E\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\} = \mathrm{diag}\{E(y_1^2), E(y_2^2), \cdots, E(y_n^2), 1\}$, and R_{MSE} becomes

$$R_{MSE} = -\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{E^2(x_j y_i)}{E(y_i^2)} + \text{const} \tag{6}$$

Define $\mathbf{K} = (K_1, \cdots, K_n)^T$ with its i-th element being

$$K_i = 2\sum_{j=1}^{n}\Big[\frac{E^2(x_j y_i)}{E^2(y_i^2)}y_i - \frac{E(x_j y_i)}{E(y_i^2)}x_j\Big] \tag{7}$$

We then have $\frac{\partial R_{MSE}}{\partial y_i} = E\{K_i\}$. Using the chain rule, also noting Eq. 3, the gradient of R_{MSE} w.r.t. $\mathbf{W}^{(2)}$ can be obtained:

$$\frac{\partial R_{MSE}}{\partial \mathbf{W}^{(2)}} = E\Big\{\sum_{i=1}^{n} K_i \cdot \frac{\partial y_i}{\partial \mathbf{W}^{(2)}}\Big\} = E\{\mathbf{K} \cdot \mathbf{z}^T\} \tag{8}$$

where $\mathbf{z} = (z_1, \cdots, z_M)^T$ is the output of the hidden layer of the MLP. Let $\mathbf{H} = \mathrm{diag}\{h'(a_1), \cdots, h'(a_M)\}$. After tedious derivation, we have

$$\frac{\partial R_{MSE}}{\partial \mathbf{W}^{(1)}} = E\{\mathbf{H} \cdot \mathbf{W}^{(2)T} \cdot \mathbf{K} \cdot \mathbf{x}^T\} \tag{9}$$

$$\frac{\partial R_{MSE}}{\partial \mathbf{W}^{(d)}} = E\{\mathbf{K} \cdot \mathbf{x}^T\} \tag{10}$$

$$\frac{\partial R_{MSE}}{\partial \mathbf{b}^{(2)}} = E\{\mathbf{K}\} \tag{11}$$

$$\frac{\partial R_{MSE}}{\partial \mathbf{b}^{(1)}} = E\{\mathbf{H} \cdot \mathbf{W}^{(2)T} \cdot \mathbf{K}\}. \tag{12}$$

R_{MSE}, given in Eq. 1, is inconsistent with certain scaling properties of the observations $\mathbf{x}$. To avoid this, one may need to normalize the variance of the observations x_i as preprocessing, if necessary.

3.2. Determination of the Regularization Parameter λ

We suggest initializing λ with a large value λ_0 at the beginning of training and decreasing it to a small constant λ_c during the learning process. A large value for λ at the beginning helps to reduce the possibility of getting into unwanted solutions. As training goes on, the influence of the regularization term is reduced, and

$\mathcal{G}$ has more freedom. Theoretically, the determination of λ_c depends on the level of nonlinear distortion in the mixing procedure. If the nonlinear distortion is considerable, we should use a very small value for λ_c to give the $\mathcal{G}$ network enough flexibility. If the variance of the observations x_i is normalized, typical values used in our experiments are $\lambda_0 = 5$ and $\lambda_c = 0.01$.

4. Relation to Previous Works

The MISEP method has been reported to solve some nonlinear BSS problems successfully, including separating a real-life nonlinear image mixture (Almeida, 2005; Almeida, 2003). Actually, in these experiments, MND seems to be utilized implicitly, though not exactly. First, direct connections between inputs and output units were incorporated in the $\mathcal{G}$ network. They can quickly adapt the linear part of $\mathcal{G}$. Second, in Almeida (2005), the $\mathcal{G}$ network was initialized with an identity mapping, and during the first 100 epochs, it was constrained to be linear. Early stopping was also applied, and hence $\mathcal{G}$ is expected to be not far from linear. We would like to mention that in this paper we formulate MND as a general principle, claim its validity and usefulness for solving nonlinear ICA problems, and give the corresponding regularizer.

In the kernel-based nonlinear BSS method (Harmeling et al., 2003), the data are first mapped to a high-dimensional kernel feature space. Next, a BSS method based on second order temporal decorrelation is performed. In this way a large number of components are extracted. When the nonlinearity in data generation is not too strong, the MND principle provides a way to select a subset of output components corresponding to the original sources. Assume that the outputs y_i are zero-mean and of unit variance. From Eq. 6 we can see that one can select y_i with large $\sum_{j=1}^{n} \frac{E^2(x_j y_i)}{E(y_i^2)} = \sum_{j=1}^{n} E^2(x_j y_i) = \sum_{j=1}^{n} \mathrm{var}(x_j) \cdot \mathrm{corr}^2(x_j, y_i)$.

The smoothness regularizer exploiting second-order derivatives (Tikhonov & Arsenin, 1977; Poggio et al., 1985) is also related to the MND principle, as shown below.

4.1. Smoothness: Local Closeness to Linear

R_{MSE}, given in Eq. 1, indicates the deviation of the mapping $\mathcal{F}$ from the affine mapping which fits $\mathcal{F}$ *globally* best. In contrast, one may enforce the *local* "closeness to linear" of the mapping averaged at every point. This actually leads to the smoothness regularizer exploiting second-order derivatives, as shown below.

For a one-dimensional sufficiently smooth function $g(\mathbf{x})$, its second-order Taylor expansion in the vicinity of $\mathbf{x}$ is $g(\mathbf{x}+\boldsymbol{\varepsilon}) \approx g(\mathbf{x}) + \left(\frac{\partial g}{\partial \mathbf{x}}\right)^T \cdot \boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\varepsilon}^T \mathbf{H}_{\mathbf{x}} \boldsymbol{\varepsilon}$. Here $\boldsymbol{\varepsilon}$ is a small variation of $\mathbf{x}$ and $\mathbf{H}_{\mathbf{x}}$ denotes the Hessian matrix of g. Let $\bigtriangledown_{ij} = \frac{\partial^2 g}{\partial x_i \partial x_j}$. If we use the first-order Taylor expansion of g at $\mathbf{x}$ to approximate $g(\mathbf{x}+\boldsymbol{\varepsilon})$, the square error is

$$
\begin{aligned}
& \Big|\Big| g(\mathbf{x}+\boldsymbol{\varepsilon}) - g(\mathbf{x}) - \Big(\frac{\partial g}{\partial \mathbf{x}}\Big)^T \cdot \boldsymbol{\varepsilon} \Big|\Big|^2 \\
\approx\ & \frac{1}{4}\big|\big|\boldsymbol{\varepsilon}^T \mathbf{H}_{\mathbf{x}} \boldsymbol{\varepsilon}\big|\big|^2 = \frac{1}{4}\Big(\sum_{i,j=1}^{n} \bigtriangledown_{ij} \varepsilon_i \varepsilon_j\Big)^2 \\
=\ & \frac{1}{4}\Big(\sum_{i=1}^{n} \bigtriangledown_{ii} \varepsilon_i^2 + \sum_{\substack{i,j=1,\\ i\neq j}}^{n} (\sqrt{2}\bigtriangledown_{ij})(\sqrt{2}\varepsilon_i \varepsilon_j)\Big)^2 \\
\leq\ & \frac{1}{4}\Big(\sum_{i=1}^{n} \bigtriangledown_{ii}^2 + 2\sum_{\substack{i,j=1,\\ i\neq j}}^{n} \bigtriangledown_{ij}^2\Big)\Big(\sum_{i=1}^{n} \varepsilon_i^4 + 2\sum_{\substack{i,j=1,\\ i\neq j}}^{n} \varepsilon_i^2 \varepsilon_j^2\Big) \\
=\ & \frac{1}{4}||\boldsymbol{\varepsilon}||^4 \cdot \Big(\sum_{i=1}^{n} \bigtriangledown_{ii}^2 + 2\sum_{\substack{i,j=1,\\ i\neq j}}^{n} \bigtriangledown_{ij}^2\Big) \\
=\ & \frac{1}{4}||\boldsymbol{\varepsilon}||^4 \cdot \sum_{i,j=1}^{n} \bigtriangledown_{ij}^2 \qquad (13)
\end{aligned}
$$

The above inequality holds due to the Cauchy's inequality. We can see that in order to make g locally close to linear at every point in the domain of $\mathbf{x}$, we just need to minimize $\int_{D_{\mathbf{x}}} \sum_{i,j=1}^{n} \bigtriangledown_{ij}^2 d\mathbf{x}$. This regularizer has been widely used for achieving the smoothness constraint in many problems (Tikhonov & Arsenin, 1977; Poggio et al., 1985). When the mapping is vector-valued, we need to apply this regularizer to each output of the mapping.

Originally we intend to do regularization on the mixing mapping $\mathcal{F}$, but it is difficult to evaluate $\frac{\partial^2 x_l}{\partial y_i \partial y_j}$. Instead, we do regularization on $\mathcal{G}$, the inverse of $\mathcal{F}$. The regularization term in Eq. 2 then becomes $R_{local}(\boldsymbol{\theta}) = \int_{D_{\mathbf{x}}} \sum_{i=1,j=1}^{n} P_{ij} d\mathbf{x}$, where $P_{ij} \triangleq \sum_{l=1}^{n} \big(\frac{\partial^2 y_l}{\partial x_i \partial x_j}\big)^2$. There are totally $\frac{n^2(n+1)}{2}$ different terms $\big(\frac{\partial^2 y_l}{\partial x_i \partial x_j}\big)^2$ in the integrand. For simplicity and computational reasons, sometimes one may drop the cross derivatives in the integrand, i.e. $\big(\frac{\partial^2 y_l}{\partial x_i \partial x_j}\big)^2$ with $i \neq j$, and consequently obtain the curvature-driven smoothing regularizer proposed in Bishop (1993).

5. Experiments with Synthetic Data

5.1. Methods and Performance Evaluation

In this section we investigate the performance of the proposed nonlinear ICA method using synthetic data. The following six methods (schemes) were used to separate various nonlinear mixtures. *1.* MISEP: The

MISEP method (Almeida, 2003) with $\boldsymbol{\theta}$ randomly initialized. *2.* Linear init.: MISEP with $\mathcal{G}$ initialized as a linear mapping. This was achieved by adopting the regularization term Eq. 1 with $\lambda = 7$ (which is very large) in the first 50 epochs. *3.* MND: MISEP incorporating the MND principle (Section 3). The regularization parameter λ decayed from $\lambda_0 = 5$ to $\lambda_c = 0.01$ in the first 350 epochs. After that λ was fixed to λ_c. *4.* Smooth (I): MISEP incorporating the smoothness regularizer (Section 4.1). λ decayed from 1 to 0.004 in the first 350 epochs. *5.* Smooth (II): Same as Smooth (I), but λ was fixed to 0.007. *6.* VB-NICA: Variational Bayesian nonlinear ICA (Valpola, 2000). PCA was used for initialization. After obtaining nonlinear factor analysis solutions using the package, we applied linear ICA, performed by FastICA (Hyvärinen, 1999), to achieve nonlinear BSS. In addition, in order to show the necessity of exploiting nonlinear ICA methods for separating nonlinear mixtures, linear ICA (FastICA was adopted) was also used to separate the nonlinear mixtures. To reduce the random effect, all the methods were repeated for 40 runs, and in each run the MLP was randomly initialized.

In this section, we just consider the 2-source-2-mixture case. For comparison, the MLP without direct connections and that with direct connections were both adopted to represent $\mathcal{G}$. Like in Almeida (2003), the MLP has 20 "arctan" hidden units, 10 of which are connected to each of y_i. We used the signal to noise ratio (SNR) of y_i relative to s_i, denoted by $SNR(y_i)$, to assess the separation performance of s_i. In addition, to take into account possible trivial transformations between s_i and y_i (for the definition of trivial transformations, see Footnote 1), we applied a flexible nonlinear transformation h to y_i to minimize the MSE between $h(y_i)$ and s_i. The SNR of $h(y_i)$ relative to s_i was used as another performance measure. In our experiments h was implemented by a two-layer MLP with eight "tansig" hidden units and a linear output unit.

5.2. Experimental Results

In experiments both super-Gaussian and sub-Gaussian sources were used. Three kinds of nonlinear mixtures were investigated. They are distorted source (DS) mixtures, post-nonlinear (PNL) mixtures (Taleb & Jutten, 1999), and generic nonlinear (GN) mixtures generated by a MLP. The DS mixtures x_i were generated according to $x_1 = a_{11}s_1 + f_1(s_2), x_2 = f_2(s_1) + a_{22}s_2$, where f_i are invertible nonlinear functions. We call them DS mixtures since each observation is a linear mixture of nonlinearly distorted sources. Each signal has 1000 samples. Figure 3(a) shows the scatter plot of the DS mixtures x_i used here. To see the level of nonlinear distortion in the mixing transformation, we give the scatter plot of the affine transformation of s_i which fits x_i best, shown by gray points. PNL mixtures were generated by a linear mixing procedure of s_i followed by a mild component-wise invertible nonlinear transformation. We used a 2-2-2 MLP with "arctan" hidden units to generate the GN mixtures. The hidden units also have biases. All weights in the MLP were randomly generated. They are not large such that the resulting nonlinear distortion is not very strong. The scatter plot of the PNL mixtures and that of the GN mixtures used in our experiments are given in Figure 3(b) and (c), respectively.

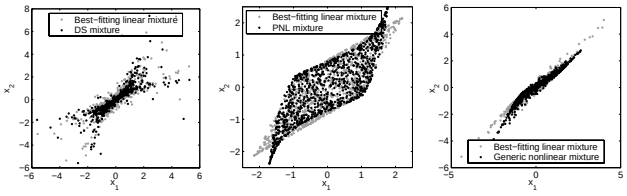
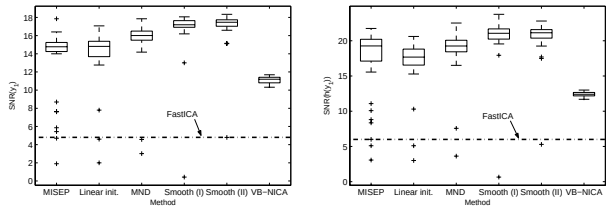


Figure 3. (a) Scatter plot of the DS mixtures, *generated from two super-Gaussian sources.* (b) That of the PNL mixtures. *The sources are a uniformly distributed white signal and a sinusoid waveform.* (c) That of the GN mixtures, *generated from the first sources in (a) and (b).* Points in gray are linear mixtures of s_i which fit x_i best.

We found that the separated results in the two channels have a similar SNR, due to space limitation, here we just give the SNR in the first channel. Figures 4 and 5 compare the boxplot of $SNR(y_1)$ and $SNR(h(y_1))$ for the DS mixtures with different methods. In Figure 4, the MLP has no direct connections between inputs and output units, while in Figure 5 the MLP has direct connections. We can see that in this case the methods MND, Smooth(I), and Smooth(II) give very high SNR, and at the same time, produce very few unwanted results. Moreover, the MLP with direct connections behaves better than that without direct connections. The performance of VB-NICA is not good. But It should be noted that VB-NICA may not exhibit its potential powerfulness in the experiments, since the source number is given and no noise is considered.

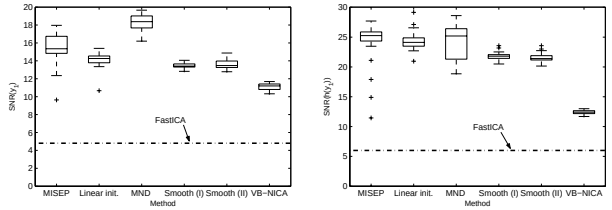
In separating the PNL mixtures, we found that the MLP with direct corrections also behaves better. But for the GN mixtures, the MLP without direct connections produces slightly better results. The separation results of these mixtures with the MLP without direct connections are given in Figures 6 and 7. Obviously, in these two cases MND gives the best performance, and the smoothness regularizer behaves poorly. Linear

initialization seems not helpful for separating the PNL mixtures, while it helps to separate the GN mixtures to some degree. Among all these three kinds of nonlinear mixtures, the PNL mixtures are most difficult to be separated by the MLP structure.



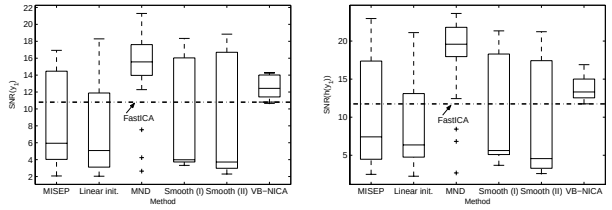
(a) (b)

Figure 4. Boxplot of the SNR of separating the DS mixtures by the MLP *without* direct connections between inputs and output units. (a) $SNR(y_1)$. (b) $SNR(h(y_1))$.



(a) (b)

Figure 5. Separating the DS mixtures by the MLP *with* direct connections. (a) $SNR(y_1)$. (b) $SNR(h(y_1))$.



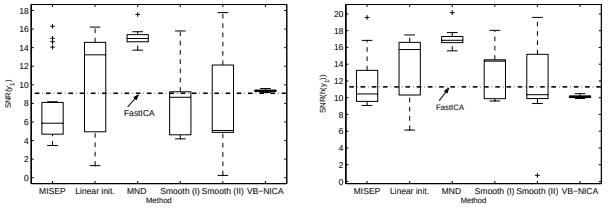
(a) (b)

Figure 6. Separating the PNL mixtures by the MLP *without* direct connections. (a) $SNR(y_1)$. (b) $SNR(h(y_1))$.

6. Application to Causality Discovery in the Stock Market

6.1. Introduction

It is well known that financial returns are not independent of each other. Their relations can be described in different ways. For example, in risk management, correlations are used to describe the relations and help to construct portfolios. The business group, which is a collection of firms bound together in some formal or informal ways, focuses on ties between financial assets. Here we are interested in how stock returns are affected by each other. The return of a particular stock may be influenced by those of other stocks, for many



(a) (b)

Figure 7. Separating the GN mixtures by the MLP *without* direct connections. (a) $SNR(y_1)$. (b) $SNR(h(y_1))$.

reasons, such as the ownership relations and financial interlinkages. According to the efficient market hypothesis, such influence should be reflected in stock returns immediately. In this part we aim to discover the causal relations among selected stocks by analyzing their daily returns.

Traditionally, causality discovery algorithms for continuous variables usually assume the Gaussianity of the variables. Under this assumption, only the correlation structure of variables is considered and all higher-order information is neglected. As a consequence, one would obtain some possible causal diagrams which are equivalent in their correlation structure, and could not find the true causal directions. Recently, it has been shown that the non-Gaussianity distribution of the variables allows us to distinguish the explanatory variable from the response variable, and consequently, to identify the full causal model. In particular, Shimizu et al. (2006) proposed an elegant and efficient method for identifying the *linear, non-Gaussian, acyclic causal model* (LiNGAM) by exploiting ICA.

6.2. Causality Discovery by ICA: Basic Idea

The LiNGAM model assumes that the causal relations among observed variables x_i can be written in matrix form: $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$, where $\mathbf{x} = (x_1, \cdots, x_n)^T$, $\mathbf{e} = (e_1, \cdots, e_n)^T$, and $\mathbf{B}$ can be permuted (by simultaneous equal row and column permutations) to strict lower triangularity if one knows the causal order of x_i. e_i are independent disturbances, and at most one of them is Gaussian. Let $\mathbf{W} = \mathbf{I} - \mathbf{B}$, we then have $\mathbf{e} = \mathbf{W}\mathbf{x}$, this is exactly the ICA separation procedure. As $\mathbf{B}$ can be permuted to strict lower triangularity, it is required that $\mathbf{W}$ can be permuted to lower triangularity. For details, see Shimizu et al. (2006).

6.3. By Nonlinear ICA with MND

The above method may fail to do causality discovery when nonlinear distortion or noise exists in the data generation procedure. Let us consider the general case of nonlinear distortion often encountered in observed

data, provided that it is smooth and mild. We use the MLP structure described in Section 3 to model the nonlinear transformation from the observed variables x_i to the disturbance variables e_i. This structure is a linear transformation $\mathbf{W}^{(d)}$ coupled with an ordinary MLP, denoted by $\boldsymbol{\phi}(\mathbf{x})$.

Due to the structure of the transformation from $\mathbf{x}$ to $\mathbf{e}$, we have $\mathbf{e} = \mathbf{W}^{(d)}\mathbf{x} + \boldsymbol{\phi}(\mathbf{x})$, and consequently $\mathbf{x} = (\mathbf{I} - \mathbf{W}^{(d)})\mathbf{x} - \boldsymbol{\phi}(\mathbf{x}) + \mathbf{e}$. As it is difficult to analyze the relations among x_i implied by the nonlinear transformation $\boldsymbol{\phi}(\mathbf{x})$, we expect that $\boldsymbol{\phi}(\mathbf{x})$ is weak enough such that its effect can be neglected. The *linear* causal relations among x_i can then by discovered by analyzing $\mathbf{W}^{(d)}$.

In order to do causality discovery, the separation system $\mathbf{e} = \mathbf{W}^{(d)}\mathbf{x} + \boldsymbol{\phi}(\mathbf{x})$ is expected to exhibit the following properties. *1.* The outputs e_i are mutually independent, since independence of e_i is a crucial assumption in LiNGAM. This can be achieved since nonlinear ICA always has solutions. *2.* $\mathbf{W}^{(d)}$ is sparse enough such that it can be permuted to lower triangularity. This can be enforced by incorporating the L_1 (Hyvärinen & Karthikesh, 2000) or smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) on the entries of $\mathbf{W}^{(d)}$. *3.* The nonlinear mapping $\boldsymbol{\phi}(\mathbf{x})$ is weak enough such that we just care about the linear causal relations indicated by $\mathbf{W}^{(d)}$. To achieve this, we adopt nonlinear ICA with MND presented in Sections 2 and 3. In addition, we initialize the system with linear ICA results and use early stopping: $\mathbf{W}^{(d)}$ is initialized to the linear ICA separation matrix, and the initial values for weights in $\boldsymbol{\phi}(\mathbf{x})$ are around 0; early stopping means that we stop the training process once the LiNGAM property holds for $\mathbf{W}^{(d)}$. After the algorithm terminates, $\frac{\mathrm{var}(\phi_i(\mathbf{x}))}{\mathrm{var}(e_i)}$ can be used to measure the level of nonlinear distortion in each channel, if needed.

6.4. Data

The Hong Kong stock market has some structural features different from the US and UK markets. One typical feature is that the concentration of market activities and equity ownership in relatively small group of stocks, which probably makes causal relations in the Hong Kong stock market more obvious. However, we should be aware that it is probably very hard to discover the causal relations among the selected stocks, since financial data are somewhat non-stationary, the data generation mechanism is not clear, and there may exist some confounder variables.

We aim at discovering the causality network among 14 stocks selected from the Hong Kong stock market.[3] The selected 14 stocks are constituents of Hang Seng Index (HSI).[4] They are almost the largest companies in this stock market. We use the daily dividend/split adjusted closing prices from Jan. 4, 2000 to Jun. 17, 2005, obtained from the Yahoo finance database. For the few days when the stock price is not available, the simple linear interpolation is used to estimate the price. Denoting the closing price of the ith stock on day t by P_{it}, the corresponding return is calculated by $x_{it} = \frac{P_{it} - P_{i,t-1}}{P_{i,t-1}}$. The observed data are $\mathbf{x}_t = (x_{1t}, \cdots, x_{14,t})^T$. Each return series contains 1331 samples.

6.5. Empirical Results

We first tried to do causality discovery on $\mathbf{x}_t$ by applying standard ICA. Both FastICA and the natural gradient ICA algorithm were adopted. We used the LiNGAM software[5] to permutate $\mathbf{W}$ and to obtain the matrix $\mathbf{B}$. $\mathbf{B}$ seems unlikely to be lower-triangular; in fact, the ratio of the sum of squares of its upper-triangular entries to that of all entries is at least 0.24, which is very large. We may conclude that the data $\mathbf{x}$ do not satisfy the LiNGAM model.

We then adopted the method discussed in Section 6.3. The SCAD penalty (Fan & Li, 2001) is applied to entries of $\mathbf{W}^{(d)}$ with $\lambda_{SCAD} = 0.04$. The regularization parameter for nonlinear ICA with MND (Eq. 8∼12) is $\lambda = 0.14$. After 195 epoches, $\mathbf{W}^{(d)}$ satisfies the LiNGAM assumption and the training process was terminated. The nonlinear distortion level $\frac{\mathrm{var}(\phi_i(\mathbf{x}))}{\mathrm{var}(e_i)}$ is 0.0485, 0.0145, 0.0287, 0.2075, 0.0180, 0.0753, 0, 0.0001, 0.0193, 0.0652, 0.0146, 0.0419, 0.0544, and 0.0492, respectively, for the 14 outputs e_i. From them we can see that nonlinear distortion is very weak. By inspection of their kurtoses, we found that all e_i are non-Gaussian. By analyzing the learned $\mathbf{W}^{(d)}$, we obtained the linear causal relations among these stocks, shown in Figure 8.

From Figure 8 we have some interesting findings. *1.* The ownership relation tends to cause a causal relation. If A is a holding company of B, there tends to be a causal relation from B to A. There are two significant relations $x_8 \to x_5$ and $x_{10} \to x_1$. In fact, x_5 owns some 60% of x_8, and x_1 holds about 50% of x_{10}. *2.* Stocks belonging to the same subindex tend to be connected together. For example, x_2, x_3, and x_6, which are linked together, are the only three constituents of

[3] They are not listed here; see the legend in Figure 8.

[4] except that Hang Lung Development Co. Ltd (0010.hk) was deleted from HSI on Dec. 2, 2002.

[5] It is available at *http://www.cs.helsinki.fi/group/neuroinf/lingam/*.

Hang Seng Utilities Index. x_1, x_9, and x_{11} are constituents of Hang Seng Property Index. *3.* Large bank companies are the cause of many stocks, meaning that the international impact to the Hong Kong stock market is probably reflected through large banks. Here x_5 and x_8 are the two largest banks in Hong Kong. *4.* Stocks in Hang Seng Property Index tend to depend on many other stocks, while they hardly influence others. Here x_1, x_9, and x_{11} are in Hang Seng Property Index. These findings also indicate that the independent factor model may provide a reasonable way to explain the generation of stock returns.

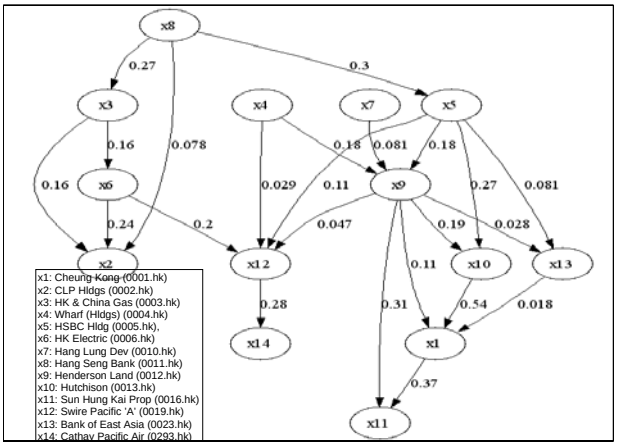


Figure 8. Casual diagram of the 14 stocks.

7. Conclusion

We have proposed the "minimal nonlinear distortion" principle for solving the nonlinear ICA problem. This principle helps to reduce the indeterminacies in solutions of nonlinear ICA and to overcome the ill-posedness of nonlinear ICA. With this principle, the solution whose nonlinear mixing system is close to linear is preferred. Experimental results with synthetic data show that when the data are generated with mild nonlinear distortion, the proposed method produces good and reliable results for separating various nonlinear mixtures. The successful application of the proposed nonlinear ICA method to causality discovery in the Hong Kong stock market illustrates the applicability of the method and the validity of the "minimal nonlinear distortion" principle for some real-life problems. The result also supports the validity of the independent factor model in finance.

Acknowledgement

This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administration Region, China.

References

Almeida, L. B. (2003). MISEP — linear and nonlinear ICA based on mutual information. *Journal of Machine Learning Research*, *4*, 1297–1318.

Almeida, L. B. (2005). Separating a real-life nonlinear image mixture. *Journal of Machine Learning Research*, *6*, 1199–1229.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Bishop, C. (1993). Curvature-driven smoothing: a learning algorithm for feedforward networks. *IEEE Trans. on Neural Networks*, *4*, 882–884.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, *96*, 1348–1360.

Harmeling, S., Ziehe, A., Kawanabe, M., & Müller, K. (2003). Kernel-based nonlinear blind source separation. *Neural Computation*, *15*, 1089–1124.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, *10(3)*, 626–634.

Hyvärinen, A., & Karthikesh, R. (2000). Sparse priors on the mixing matrix in independent component analysis. *Proc. 2nd Int. Workshop on ICA and BSS (ICA2000)* (pp. 477–452). Helsinki, Finland.

Hyvärinen, A., & Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, *12*, 429–439.

Jutten, C., & Karhunen, J. (2003). Advances in nonlinear blind source separation. *Proc. 4th Int. Symp. on ICA and BSS (ICA2003)* (pp. 245–256). Invited paper in the special session on nonlinear ICA and BSS.

Jutten, C., & Taleb, A. (2000). Source separation: From dusk till dawn. *2nd Int. Workshop on ICA and BSS (ICA 2000)* (pp. 15–26). Helsinki, Finland.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature*, *317*, 314–319.

Shimizu, S., Hoyer, P., Hyvärinen, A., & Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*, 2003–2030.

Taleb, A., & Jutten, C. (1999). Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, *47*, 2807–2820.

Tan, Y., Wang, J., & Zurada, J. M. (2001). Nonlinear blind source separation using a radial basis function network. *IEEE Trans. on Neural Networks*, *12*, 124–134.

Tikhonov, A. N., & Arsenin, V. A. (1977). *Solutions of ill-posed problems.* Washington: Winston & Sons.

Valpola, H. (2000). Nonlinear independent component analysis using ensemble learning: Theory. *Proc. 2nd Int. Workshop on ICA and BSS (ICA2000)* (pp. 251–256). Helsinki, Finland.