Journal of Machine Learning Research 9 (2008) 2455-2487 Submitted 9/07; Revised 5/08; Published 11/08

Minimal Nonlinear Distortion Principle for Nonlinear Independent Component Analysis

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

Editor: Aapo Hyvärinen

Abstract

It is well known that solutions to the nonlinear independent component analysis (ICA) problem are highly non-unique. In this paper we propose the "minimal nonlinear distortion" (MND) principle for tackling the ill-posedness of nonlinear ICA problems. MND prefers the nonlinear ICA solution with the estimated mixing procedure as close as possible to linear, among all possible solutions. It also helps to avoid local optima in the solutions. To achieve MND, we exploit a regularization term to minimize the mean square error between the nonlinear mixing mapping and the best-fitting linear one. The effect of MND on the inherent trivial and non-trivial indeterminacies in nonlinear ICA solutions is investigated. Moreover, we show that local MND is closely related to the smoothness regularizer penalizing large curvature, which provides another useful regularization condition for nonlinear ICA. Experiments on synthetic data show the usefulness of the MND principle for separating various nonlinear mixtures. Finally, as an application, we use nonlinear ICA with MND to separate daily returns of a set of stocks in Hong Kong, and the linear causal relations among them are successfully discovered. The resulting causal relations give some interesting insights into the stock market. Such a result can not be achieved by linear ICA. Simulation studies also verify that when doing causality discovery, sometimes one should not ignore the nonlinear distortion in the data generation procedure, even if it is weak.

Keywords: nonlinear ICA, regularization, minimal nonlinear distortion, mean square error, best linear reconstruction

1. Introduction

Independent component analysis (ICA) is a popular statistical technique aiming to recover independent sources from their observed mixtures, without knowing the mixing procedure or any specific knowledge of the sources (Hyvärinen et al., 2001; Cardoso, 1998; Cichocki and Amari, 2003). In the case that the observed mixtures are a linear transformation of the sources, under weak assumptions, ICA can recover the original sources with the trivial permutation and scaling indeterminacies. Linear ICA is currently a popular method for blind source separation (BSS) of linear mixtures.

However, nonlinear ICA does not necessarily lead to nonlinear BSS. In Hyvärinen and Pajunen (1999), it was shown that solutions to nonlinear ICA always exist and that they are highly non-unique. Actually, one can easily construct a nonlinear transformation of some non-Gaussian independent variables to produce independent outputs. Below are a few examples. Let $y_1, ..., y_n$ be some independent variables. Their component-wise nonlinear functions are still mutually independent. If we use Gaussianization (Chen and Gopinath, 2001) to transform y_i into Gaussian variables u_i, any component-wise nonlinear function of $\mathbf{U} \cdot \mathbf{u}$, where $\mathbf{u} = (u_1, ..., u_n)^T$ and $\mathbf{U}$ is an orthogonal matrix, still has mutually independent components. Taleb and Jutten (1999) also gave an example in which nonlinear mixtures of independent variables are still independent. One can see that 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 and Taleb, 2000; Taleb, 2002).

If we constrain the nonlinear mixing mapping to have some specific forms, the indeterminacies in the results of nonlinear ICA can be reduced dramatically, and as a consequence, nonlinear ICA may lead to nonlinear BSS. For example, in Burel (1992), a parametric form of the mixing transformation is assumed known and one just needs to adjust the unknown parameters. The learning algorithms were improved in Yang et al. (1998). By exploiting the extensions of the Darmois-Skitovich theorem (Kagan et al., 1973) to nonlinear functions, a particular class of nonlinear mixing mappings, which satisfy an addition theorem in the sense of the theory of functional equations, were considered in Eriksson and Koivunen (2002). In particular, the post-nonlinear (PNL) mixing model (Taleb and Jutten, 1999), which assumes that the mixing mapping is a linear transformation followed by a component-wise nonlinear one, has drawn much attention.

In practice, the exact form of the nonlinear mixing procedure is probably 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 radial basis function (RBF) network, to represent the nonlinear separation system. Almeida (2003) uses the MLP to model the separation system and trains the MLP by information-maximization (Infomax). Moreover, the smoothness constraint,[1] which is implicitly provided by MLP's with small initial weights and with a relatively small number of hidden units, was believed to be a suitable regularization condition to achieve nonlinear BSS. In Tan et al. (2001), a RBF network is adopted to represent the separation system, and partial moments of the outputs of the separation system 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 sources is non-trivial,[2] this regularization could not help to recover the original sources. Variational Bayesian nonlinear ICA (Lappalainen and Honkela, 2000; Valpola, 2000) uses the MLP to model the nonlinear mixing transformation. By resorting to the variational Bayesian inference technique, this method can do model selection and avoid overfitting. If we can have some additional knowledge about the nonlinear mixing transformation and incorporate it efficiently, 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, in many cases the nonlinearity for generating natural signals we deal with is not strong. Hence, provided that the nonlinear ICA outputs are mutually independent, we would prefer the solution with the estimated data generation procedure of minimal nonlinear distortion (MND). This

[1] Following Tikhonov and Arsenin (1977), here we use the term "smoothness" in a very general sense. Often it means that that the function does not change abruptly and/or that it does not oscillate too much.

[2] For the definition of a trivial transformation, one may see Jutten and Taleb (2000). A one-to-one mapping $\mathcal{H}$ is trivial if and only if it satisfies $\mathcal{H}_i(y_1, y_2, ..., y_n) = h_i(y_{\sigma(i)})$, $i = 1, 2, ..., n$, where h_i are arbitrary functions and σ is any permutation over $\{1, .2, ..., n\}$. That is, a trivial mapping of $\mathbf{y}$ is a permutation of y_i followed by a component-wise transformation.

information can help to reduce the indeterminacies in nonlinear ICA greatly, and moreover, to avoid local optima in the solutions to nonlinear ICA. The minimal nonlinear distortion of the mixing system is achieved by the technique of regularization. The objective function of nonlinear ICA with MND is the mutual information between outputs penalized by some terms measuring the level of "closeness to linear" of the mixing system. The mean square error (MSE) between the nonlinear mixing system and its best-fitting linear one provides such a regularization term. To ensure that nonlinear ICA results in nonlinear BSS, one may also need to enforce the local MND of the nonlinear mapping averaged at every point, which turns out to be the smoothness regularizer exploiting second-order partial derivatives.

MND, as well as the smoothness regularizer, can be incorporated in various nonlinear ICA methods to improve the results. Here we consider two nonlinear ICA methods. The first one is the MISEP method (Almeida, 2003), where the MLP is used to represent the separation system. As regularization is powerful for complexity control in neural networks (Bishop, 1995), the structure of the MLP is not optimized during the learning process, that is, it is fixed. The second one is nonlinear ICA based on kernels (Zhang and Chan, 2007a), in which the nonlinear separation system is modeled using some kernel methods. We then explain why MND helps to alleviate the ill-posedness in nonlinear ICA solutions, by investigating the effect of MND on trivial and non-trivial indeterminacies in nonlinear ICA solutions systematically. Next, we conduct experiments using synthetic data to compare the performance of several nonlinear ICA methods. The results confirm the effectiveness of the proposed MND principle to avoid unwanted solutions and to improve the separation performance. Finally, nonlinear ICA with MND is used to discover linear causal relations in the Hong Kong stock market and give encouraging results. We also give experimental results on synthetic data, which illustrate that when performing ICA-based causality discovery on the data whose generation procedure involves nonlinear distortion, one should take into account the nonlinear effect in the ICA separation system, even if it is mild.[3]

2. Nonlinear ICA with Minimal Nonlinear Distortion

In this section we first briefly review the general nonlinear ICA problem, and then propose the minimal nonlinear distortion (MND) principle for regularization of nonlinear ICA.

2.1 Nonlinear ICA

In the nonlinear ICA model, the observed data $\mathbf{x} = (x_1, ..., x_n)^T$ are assumed to be generated from a vector of independent variables $\mathbf{s} = (s_1, ..., s_n)^T$ by a nonlinear transformation:

$$\mathbf{x} = \mathcal{F}(\mathbf{s}), \tag{1}$$

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 $\mathcal{G} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that

$$\mathbf{y} = \mathcal{G}(\mathbf{x})$$

has statistically independent components. As mentioned in Section 1, the results of nonlinear ICA are highly non-unique. In order to achieve nonlinear BSS, which aims at recovering the original sources s_i, we should resort to additional prior information or suitable regularization constraints.

[3] Some preliminary results of this paper were presented at ICML2007 (Zhang and Chan, 2007b).

2.2 With Minimum Nonlinear Distortion

We now propose the MND principle to restrict the space of nonlinear ICA solutions. As a consequence, the ill-posedness of the nonlinear ICA problem is alleviated. Under the condition that the separation outputs y_i are mutually independent, this principle prefers the solution with the estimated mixing transformation $\hat{\mathcal{F}}$ as close as possible to linear.

Now we need a measure of “closeness to linear” of a mapping. Given a nonlinear mapping $\hat{\mathcal{F}}$, its deviation from the affine mapping $\mathbf{A}^*$, which fits $\hat{\mathcal{F}}$ 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 $\hat{\mathcal{F}} = \mathcal{G}^{-1}$ can be measured by the MSE between $\mathcal{G}^{-1}$ and $\mathbf{A}^*$. We denote this measure by $R_{MSE}(\theta)$, where θ denotes the set of unknown parameters in the nonlinear ICA system. Let $\mathbf{x}^* = (x_1^*, \cdots, x_n^*)^T$ be the output of the affine transformation from $\mathbf{y}$ by $\mathbf{A}^*$. Let $\tilde{\mathbf{y}} = [\mathbf{y}; 1]$. $R_{MSE}(\theta)$ can then be written as the MSE between x_i and x_i^*:

$$\begin{aligned} R_{MSE}(\theta) &= E\{(\mathbf{x}-\mathbf{x}^*)^T(\mathbf{x}-\mathbf{x}^*)\}\text{ , where} \\ \mathbf{x}^* &= \mathbf{A}^*\tilde{\mathbf{y}},\text{ and } \mathbf{A}^* = \arg_{\mathbf{A}} \min E\{(\mathbf{x}-\mathbf{A}\mathbf{y})^T(\mathbf{x}-\mathbf{A}\mathbf{y})\}. \end{aligned} \tag{2}$$

Here $\mathbf{A}^*$ is a $n \times (n+1)$ matrix.[4] Figure 1 shows the separation system $\mathcal{G}$ together with the generation process of R_{MSE}.

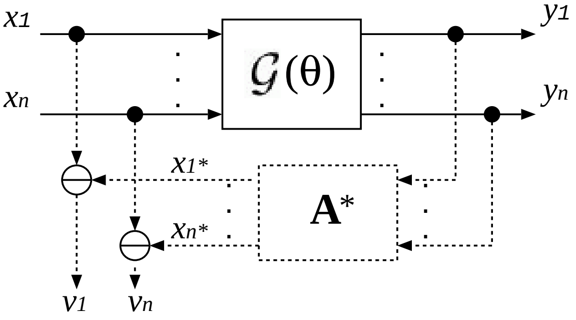

Figure 1: The separation system $\mathcal{G}$ (solid line) and the generation of the regularization term R_{MSE} (dashed line). $R_{MSE} = \sum_{i=1}^n v_i^2$, where $v_i = x_i - x_i^*$.

With R_{MSE} measuring the level of nonlinear distortion, nonlinear ICA with MND can be formulated as the following constrained optimization problem. It aims to minimize the mutual information between outputs, that is, $I(\mathbf{y})$, subject to $R_{MSE}(\theta) \le t$, where t is a pre-assigned parameter. The Lagrangian for this optimization problem is $L(\theta, \lambda) = I(\mathbf{y}) + \lambda[R_{MSE}(\theta) - t]$ with $\lambda \ge 0$. To find θ, we need to minimize

$$J = I(\mathbf{y}) + \lambda R_{MSE}(\theta). \tag{3}$$

The non-negative constant λ depends on the pre-assigned parameter t.

Another advantage of the MND principle is that it tends to make the mapping $\mathcal{G}$ invertible. In the general nonlinear ICA problem, it is assumed that both $\mathcal{F}$ and $\mathcal{G}$ are invertible. But in practice

[4] 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.

it is not easy to guarantee the invertibility of the mapping provided by a flexible nonlinear function approximator, like the MLP. MND pushes $\mathcal{G}$ to be close to a linear invertible transformation. Hence when nonlinearity in $\mathcal{F}$ is not too strong, MND helps to guarantee the invertibility of the nonlinear ICA separation system $\mathcal{G}$.

2.2.1 SIMPLIFICATION OF R_{MSE}

R_{MSE}, given in Eq. 2, can be further simplified. According to Eq. 2, the derivative of R_{MSE} w.r.t. $\mathbf{A}^*$ is $\frac{\partial R_{MSE}}{\partial \mathbf{A}^*} = -2E\{(\mathbf{x}-\mathbf{A}^*\tilde{\mathbf{y}})\tilde{\mathbf{y}}^T\}$. Setting the derivative to $\mathbf{0}$ gives $E\{(\mathbf{x}-\mathbf{A}^*\tilde{\mathbf{y}})\tilde{\mathbf{y}}^T\} = \mathbf{0}$, which implies

$$\mathbf{A}^* = E\{\mathbf{x}\tilde{\mathbf{y}}^T\}[E\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\}]^{-1}. \tag{4}$$

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

Due to Eq. 4, we have $E\{\mathbf{A}^*\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\mathbf{A}^{*T}\} = E\{\mathbf{x}\tilde{\mathbf{y}}^T\}\mathbf{A}^{*T}$,and R_{MSE} then becomes

$$\begin{aligned} R_{MSE} &= \mathrm{Tr}\big(E\{(\mathbf{x}-\mathbf{A}^*\tilde{\mathbf{y}})(\mathbf{x}-\mathbf{A}^*\tilde{\mathbf{y}})^T\}\big) \\ &= \mathrm{Tr}\big(E\{\mathbf{x}\mathbf{x}^T - \mathbf{A}^*\tilde{\mathbf{y}}\mathbf{x}^T - \mathbf{x}\tilde{\mathbf{y}}^T\mathbf{A}^{*T} + \mathbf{A}^*\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\mathbf{A}^{*T}\}\big) \\ &= \mathrm{Tr}\big(E\{\mathbf{x}\mathbf{x}^T - \mathbf{A}^*\tilde{\mathbf{y}}\mathbf{x}^T - \mathbf{x}\tilde{\mathbf{y}}^T\mathbf{A}^{*T} + \mathbf{x}\tilde{\mathbf{y}}^T\mathbf{A}^{*T}\}\big) \\ &= -\mathrm{Tr}\big(E\{\mathbf{A}^*\tilde{\mathbf{y}}\mathbf{x}^T\}\big) + \mathrm{Tr}\big(E\{\mathbf{x}\mathbf{x}^T\}\big) \\ &= -\mathrm{Tr}\big(E\{\mathbf{x}\tilde{\mathbf{y}}^T\}[E\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\}]^{-1}E\{\tilde{\mathbf{y}}\mathbf{x}^T\}\big) + \text{const.} \end{aligned} \tag{5}$$

Since y_i are independent from each other, they are uncorrelated. We can also easily make y_i zero-mean. Consequently, $E\{\tilde{\mathbf{y}}\tilde{\mathbf{y}}^T\} = \mathrm{diag}\{E(y_1^2), E(y_2^2), ..., E(y_n^2), 1\}$, and R_{MSE} becomes

$$\begin{aligned} R_{MSE} &= -\mathrm{Tr}\big(E\{\mathbf{x}\tilde{\mathbf{y}}^T\}\cdot[\mathrm{diag}\{E(y_1^2), ..., E(y_n^2), 1\}]^{-1}\cdot E\{\tilde{\mathbf{y}}\mathbf{x}^T\}\big) + \text{const} \\ &= -\sum_{j=1}^{n}\sum_{i=1}^{n}\frac{E^2(x_j y_i)}{E(y_i^2)} + \text{const.} \end{aligned} \tag{6}$$

R_{MSE} depends only on the inputs and the outputs of the nonlinear ICA system $\mathcal{G}(\theta)$. Given a form for $\mathcal{G}$, the learning rule for nonlinear ICA with MND is derived by minimizing Eq. 3. Note that R_{MSE}, given in Eq. 2, is inconsistent with certain scaling properties of the observations $\mathbf{x}$. To avoid this, one needs to normalize the variance of the observations x_i through preprocessing, if necessary.

2.2.2 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 reduces the possibility of getting into unwanted solutions, which may be non-trivial transformations of the original sources s_i or local optima. As training goes on, the influence of the regularization term is relaxed, and $\mathcal{G}$ gains more freedom. Hopefully, nonlinearity will be introduced, if necessary. The choice 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. In our experiments, we found that the separation performance of nonlinear ICA with MND is robust to the value of λ_c in a certain range. If the variance of the observations x_i is normalized, typical values used in our experiments are $\lambda_0 = 5$ and $\lambda_c = 0.01$.

2.3 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, 2003). Almeida (2003) claimed that the MLP itself may provide suitable regularization for nonlinear ICA. Some means were also used for regularization in the experiments there. For example, first, direct connections between inputs and output units were incorporated in the $\mathcal{G}$ network. Direct connections can quickly adapt the linear part of the mapping $\mathcal{G}$. 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 (by keeping the output weights of the hidden layer equal to zero). After that, the $\mathcal{G}$ network began learning the nonlinear distortion. $\mathcal{G}$ is therefore expected to be not far from linear, and MND is achieved to some extent. Accordingly, nice experimental results reported there could support the usefulness of the MND principle. We should mention that the MND principle formulated here, as well as the corresponding regularizer, provides a way to control the nonlinearity of the mixing mapping. It can be incorporated by any nonlinear ICA method, including MISEP. Later, we will investigate the effect of MND on nonlinear ICA solutions theoretically, and compare various related nonlinear ICA methods empirically.

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 made zero-mean and of unit variance. From Eq. 6 we can see that one can select y_i with large $\sum_{j=1}^{n} \frac{E^2(x_j y_i)}{E(y_i^2)} = \sum_{j=1}^{n} E^2(x_j y_i) = \sum_{j=1}^{n} \mathrm{var}(x_j) \cdot \mathrm{corr}^2(x_j, y_i)$.

It is worth mention that the principle of least mean square error reconstruction has been used for training a class of neural networks and gives some interesting results (Xu, 1993). For one-layer networks with linear/nonlinear units, this principle leads to principal component analysis (PCA)/ICA. We should address that the reconstruction in their work is quite different from that discussed in Section 2.2 in this paper. In their work, the forward process and the reconstruction process share the same weights; in this paper, reconstructed signals are an affine mapping of the outputs, and parameters in the affine mapping are determined by minimizing the reconstruction error.

Smoothness provides a constraint to prevent a neural network from overfitting noisy data. It is also useful to ensure nonlinear ICA to result in nonlinear BSS (Almeida, 2003). In fact, the smoothness regularizer exploiting second-order derivatives (Tikhonov and Arsenin, 1977; Poggio et al., 1985) is also related to the MND principle, as shown below.

2.4 Local Minimal Nonlinear Distortion: Smoothness

R_{MSE}, given in Eq. 2, indicates the deviation of the mapping $\hat{\mathcal{F}}$ from the affine mapping which fits $\hat{\mathcal{F}}$ *globally* best. In contrast, one may enforce the *local* MND of the nonlinear mapping averaged at every point. We will show that this regularization actually leads to the smoothness regularizer exploiting second-order partial derivatives (Tikhonov and Arsenin, 1977; Poggio et al., 1985; Bishop, 1993).

For a one-dimensional sufficiently smooth function $g(\mathbf{x})$, we can use the second-order Taylor expansion to approximate its function value in the vicinity of $\mathbf{x}$ in terms of $g(\mathbf{x})$:

$$g(\mathbf{x}+\varepsilon) \approx g(\mathbf{x}) + \Big(\frac{\partial g}{\partial \mathbf{x}}\Big)^T \cdot \varepsilon + \frac{1}{2}\varepsilon^T \mathbf{H}_{\mathbf{x}} \varepsilon,$$

where ε 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, which is a linear function, to approximate $g(\mathbf{x}+\varepsilon)$, the square error is

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

The above inequality holds due to the Cauchy's inequality. Now we can see that in order to achieve the local MND of g averaged in the domain of $\mathbf{x}$, we just need to minimize the following

$$\int_{\mathbb{D}_{\mathbf{x}}} \sum_{i,j=1}^{n} \bigtriangledown_{ij}^2 d\mathbf{x} = \int_{\mathbb{D}_{\mathbf{x}}} \Big(\sum_{i=1}^{n} \bigtriangledown_{ii}^2 + 2\sum_{\substack{i,j=1,\\ i<j}}^{n} \bigtriangledown_{ij}^2\Big) d\mathbf{x}. \qquad (7)$$

This regularizer has been used for achieving the smoothness constraint (see, e.g., Grimson 1982 for its application in computer vision). When the mapping is vector-valued, we need to apply the above regularizer to each component of the mapping.

Originally we intended to do regularization on the mixing mapping $\hat{\mathcal{F}}$, but it is difficult to do since it is hard to evaluate $\frac{\partial^2 x_l}{\partial y_i \partial y_j}$. Instead, we do regularization on $\mathcal{G}$, the inverse of $\hat{\mathcal{F}}$. The regularization term in Eq. 3 then becomes

$$R_{local}(\theta) = \int_{\mathbb{D}_{\mathbf{x}}} \sum_{l=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \Big(\frac{\partial^2 y_l}{\partial x_i \partial x_j}\Big)^2 d\mathbf{x} = \int_{\mathbb{D}_{\mathbf{x}}} \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij} d\mathbf{x}, \qquad (8)$$

where $P_{ij} \triangleq \sum_{l=1}^{n}\Big(\frac{\partial^2 y_l}{\partial x_i \partial x_j}\Big)^2$. Nonlinear ICA with a smooth de-mixing mapping can be achieved by minimizing the mutual information between y_i, with R_{local}, given by Eq. 8, as the regularization term. 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 of R_{local}. For simplicity and computational reasons, sometimes one may drop the cross derivatives in Eq. 8, that is, $\Big(\frac{\partial^2 y_l}{\partial x_i \partial x_j}\Big)^2$ with $i \neq j$, and consequently obtain the curvature-driven smoothing regularizer proposed in Bishop (1993), with the number of different terms in the integrand being n^2.

3. Incorporation of MND in Different Nonlinear ICA Methods

Now we should choose a model for the nonlinear ICA separation system $\mathcal{G}(\theta)$ and give the learning rule for nonlinear ICA with MND as well as nonlinear ICA with the smoothness constraint for $\mathcal{G}$. Two nonlinear ICA methods are considered here. They are MISEP (Almeida, 2003) and nonlinear ICA based on kernels (Zhang and Chan, 2007a).

3.1 MISEP with MND

Before incorporating MND into the MISEP method (Almeida, 2003) for nonlinear ICA, we give an overview of this method.

3.1.1 MISEP FOR NONLINEAR ICA

MISEP adopts the MLP to model the separation function $\mathcal{G}$ in the nonlinear ICA problem. Figure 2 shows the structure used in this method. This method extends the original Infomax method for linear ICA (Bell and Sejnowski, 1995) in two aspects. First, the separation system is a nonlinear transformation, which is modeled by the MLP. Second, the nonlinearities ψ_i are not fixed in advance, but tuned by the Infomax principle, together with $\mathcal{G}$.

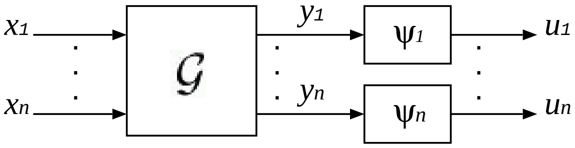

Figure 2: The network structure used in Infomax and MISEP. $\mathcal{G}$ is the separation system, and ψ_i are the nonlinearities applied to the separated signals. In MISEP, $\mathcal{G}$ is a nonlinear transformation, and both $\mathcal{G}$ and ψ_i are learned by the Infomax principle.

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, which can be written as $H(\mathbf{u}) = H(\mathbf{x}) + E\{\log|\det \mathbf{J}|\}$, where $\mathbf{J} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ is the Jacobian of the nonlinear transformation from $\mathbf{x}$ to $\mathbf{u}$. As $H(\mathbf{x})$ does not depend on the parameters in $\mathcal{G}$ and ψ_i, it can be considered as a constant. Maximizing $H(\mathbf{u})$ is thus equivalent to minimizing

$$J_1(\theta) = -E\{\log|\det \mathbf{J}|\}, \tag{9}$$

where θ denotes the set of unknown parameters. The learning rules for θ were derived by 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. For the hidden units, the activation function $l(\cdot)$ may be the logistic sigmoid function, the arctan function, etc. Direct connections between the inputs and output units are also allowed. Let $\mathbf{a} = [a_1, ..., a_M]^T$ be the inputs to the hidden units, $\mathbf{z} = [z_1, ..., z_M]^T$ be the output 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: $\mathbf{W}^{(d)}$ denotes the weights from the inputs to output units, $\mathbf{W}^{(1)}$ those from the inputs to the hidden layer, and $\mathbf{W}^{(2)}$ those from the hidden layer to the output units. $\mathbf{b}^{(1)}$ and $\mathbf{b}^{(2)}$ are the bias vectors in the hidden layer and in the output units, respectively. The output of the $\mathcal{G}$ network represented by this MLP takes the form:

$$\begin{aligned} \mathbf{y} &= \mathbf{W}^{(2)} \cdot \mathbf{z} + \mathbf{W}^{(d)}\mathbf{x} + \mathbf{b}^{(2)}, \text{ where} \\ z_i &= l(a_i), \text{ and } \mathbf{a} = \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}. \end{aligned} \tag{10}$$

3.1.2 MISEP WITH MND

For MISEP with MND, the objective function to be minimized is Eq. 9 regularized by R_{MSE} given in Eq. 6. The learning rule for θ to minimize Eq. 9 has been considered in Almeida (2003). Hence here we only give the gradient of R_{MSE} w.r.t. θ.

Using the chain rule, also noting Eq. 10, the gradient of $R_{MSE}(\theta)$ w.r.t. $\mathbf{W}^{(2)}$ can be obtained:

$$\frac{\partial R_{MSE}}{\partial \mathbf{W}^{(2)}} = E\Big\{\sum_{i=1}^{n} 2\Big[\frac{E^2(x_j y_i)}{E^2(y_i^2)} y_i - \frac{E(x_j y_i)}{E(y_i^2)} x_i\Big] \cdot \frac{\partial y_i}{\partial \mathbf{W}^{(2)}}\Big\} = E\{\mathbf{K} \cdot \mathbf{z}^T\}, \tag{11}$$

where $\mathbf{K} \triangleq [K_1, ..., K_n]^T$ with its i-th element being $K_i = 2\sum_j \Big[\frac{E^2(x_j y_i)}{E^2(y_i^2)} y_i - \frac{E(x_j y_i)}{E(y_i^2)} x_j\Big]$, and $\mathbf{z} = [z_1, z_2, ..., z_M]^T$ is the output of the hidden layer of the MLP. For the gradient of R_{MSE} w.r.t. $\mathbf{W}^{(1)}$, $\mathbf{W}^{(d)}$, $\mathbf{b}^{(2)}$, and $\mathbf{b}^{(1)}$, see Appendix A.

3.1.3 MISEP WITH SMOOTHNESS CONSTRAINT ON $\mathcal{G}$

The mapping provided by a MLP may not be smooth enough to make nonlinear ICA result in nonlinear BSS. So here we also implement MISEP with the smoothness constraint on $\mathcal{G}$. The objective function to be minimized becomes Eq. 9 regularized by R_{local} given in Eq. 8. P_{ij} appears in the expression of R_{local}. We first derive its gradient w.r.t. θ in a way analogous to that in Bishop (1993); see Appendix B.

In calculation of $\frac{\partial R_{local}}{\partial \theta}$, the integral in Eq. 8 is difficult to evaluate. Below are two ways to tackle this problem. A very simple way to approximate Eq. 7 is to use the average of the integrand over all observations instead of the integral (ignoring a constant scaling factor), just as Bishop (1993) does:

$$R_{local}^{(1)}(\theta) = E\Big\{\sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\Big\}. \tag{12}$$

This approximation actually assumes that the distribution of $\mathbf{x}$ is close to uniform, as seen from below. Eq. 8 can be rewritten as

$$R_{local}(\theta) = \int_{\mathbb{D}_{\mathbf{x}}} p(\mathbf{x}) \cdot \frac{1}{p(\mathbf{x})} \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij} d\mathbf{x} = E\Big\{\frac{1}{p(\mathbf{x})} \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\Big\}. \tag{13}$$

If $p(\mathbf{x})$ is a constant in the domain $\mathbb{D}_{\mathbf{x}}$, Eq. 12 is equivalent to Eq. 13; otherwise, the approximation using Eq. 12 may result in large error, and we may need another way to approximate the integral in Eq. 8.

When the nonlinear ICA algorithm has run for a certain number of epochs, $\mathbf{u}$, the output of the system in Figure 2, has approximately independent components and is approximately uniformly distributed in $[0,1]^n$. This means that $p(\mathbf{u})$ is approximately 1. As $p(\mathbf{x}) = p(\mathbf{u}) \cdot |\det \mathbf{J}|$, one can see that $p(\mathbf{x})$ is approximately equal to $|\det \mathbf{J}|$. Consequently Eq. 13 becomes $R_{local}(\theta) \approx E\Big\{\frac{1}{|\det \mathbf{J}|} \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\Big\}$. The gradient of $R_{local}(\theta)$ is

$$\frac{\partial R_{local}(\theta)}{\partial \theta} \approx E\Big\{\frac{1}{|\det \mathbf{J}|} \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial P_{ij}}{\partial \theta}\Big\}. \tag{14}$$

As $\mathbf{J} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$ can be easily calculated according to the network structure in Figure 2, Eq. 14 is also easy to evaluate, using Eq. 20 of Appendix B.

3.2 MND for Nonlinear ICA Based on Kernels

Nonlinear ICA based on kernels (Zhang and Chan, 2007a) exploits kernel methods to construct the separation system $\mathcal{G}$, and unknown parameters are adjusted by minimizing the mutual information between outputs y_i.[5] We have applied the MND principle and the smoothness regularizer to nonlinear ICA based on kernels; for details, see Zhang and Chan (2007a). Note that unlike the mapping provided by a MLP, which is comparatively smooth, the mapping constructed by kernel methods may not be smooth. So it is quite necessary to explicitly enforce the smoothness constraint for nonlinear ICA based on kernels.

4. Investigation of the Effect of MND

In this section we intend to explain why the MND principle, including the smoothness regularization, helps to alleviate the ill-posedness of nonlinear ICA from a mathematical viewpoint. There are two types of indeterminacies in solutions to nonlinear ICA, namely trivial indeterminacies and non-trivial indeterminacies. Trivial indeterminacies mean that the estimate of s_j produced by nonlinear ICA may be any nonlinear function of s_j; non-trivial indeterminacies mean that the outputs of nonlinear ICA, although mutually independent, are still a mixing of the original sources. Let us begin with the effect of MND on trivial indeterminacies.

4.1 For Trivial Indeterminacies

Let us assume in this section that, in the solutions of nonlinear ICA, each component depends only on one of the sources. Before presenting the main result, let us first give the following lemma.

Lemma 1 *Suppose that we are given the random vector* $\mathbf{d} = (d_1, d_2, \cdots, d_n)^T$. *Let* R_y *be the mean square error of reconstructing* $\mathbf{d}$ *from the variable* y *with the best-fitting linear transformation, that is,* $R_y = \min_{\mathbf{a}} E\{||\mathbf{d} - \mathbf{a} \cdot y||^2\}$, *where* $\mathbf{a} = (a_1, a_2, \cdots, a_n)^T$. *The variable* y *which gives the minimum* R_y *is the first non-centered principal component of* $\mathbf{d}$ *multiplied by a constant, and if* y *is constrained to be zero-mean, it is the first principal component of* $\mathbf{d}$ *multiplied by a constant.*

See Appendix C for a proof. Now let us consider a particular kind of nonlinear mixtures, in which each observed nonlinear mixture x_i is assumed to be generated by

$$x_i = f_{i1}(s_1) + f_{i2}(s_2) + \cdots + f_{in}(s_n), \tag{15}$$

where f_{ij} are invertible functions. We call such nonlinear mixtures distorted source (DS) mixtures, since each observation is a linear mixture of nonlinearly distorted sources. For this nonlinear mixing model, we have the following theorem on the effect of MND on trivial indeterminacies in its nonlinear ICA solutions. Here the following assumptions are made:

A1: In the output of nonlinear ICA, each component depends only on one of the sources and is zero-mean.

A2: The nonlinear ICA system has enough flexibility to reach the minimum of the MND regularization term R_{MSE} defined by Eq. 2.

[5] The difference between nonlinear ICA based on kernels discussed here and the kernel-based nonlinear BSS method by Harmeling et al. (2003) should be made clear. Both of them use kernels. However, the former produces statistically independent outputs, while the latter exploits the temporal structure of the sources for separation.

Theorem 1 *Suppose that each observed nonlinear mixture x_i is generated according to Eq. 15. Under assumptions A1 & A2, the estimate of s_j produced by nonlinear ICA with MND is the first principal component of $\mathbf{f}_{*j}(s_j) = [f_{1j}(s_j), \cdots, f_{nj}(s_j)]^T$, multiplied by a constant.*

See Appendix D for a proof. The DS mixing model Eq. 15 may be restrictive. Now let us consider the case where nonlinearity in $\mathcal{F}$ is mild such that $\mathcal{F}$ can be well approximated by its Maclaurin expansion of degree 3. Let

$$\triangledown_{i,j} = \frac{\partial x_i}{\partial s_j}\Big|_{s_j=0}, \triangledown_{i,jk} = \frac{\partial^2 x_i}{\partial s_j \partial s_k}\Big|_{s_j,s_k=0}, \text{ and } \triangledown_{i,jkl} = \frac{\partial^3 x_i}{\partial s_j \partial s_k \partial s_l}\Big|_{s_j,s_k,s_l=0}.$$

The following theorem discusses the effect of MND on trivial indeterminacies of nonlinear ICA solutions in this case. In particular, it states that by incorporating MND into the nonlinear ICA system, trivial indeterminacies in nonlinear ICA solutions are overcome; it shows how the outputs of the nonlinear ICA system, as the estimate of the sources, are related to the original sources s_i and the mixing system $\mathcal{F}$.

Theorem 2 *Suppose that each component of the mixing mapping $\mathcal{F} = (f_1, \cdots, f_n)^T$ in Eq. 1 is generated by the following Maclaurin series of degree 3:*

$$x_i = f_i(\mathbf{s}) = f_i(\mathbf{0}) + \sum_j \triangledown_{i,j} s_j + \frac{1}{2}\sum_{j,k} \triangledown_{i,jk} \cdot s_j s_k + \frac{1}{6}\sum_{j,k,l} \triangledown_{i,jkl} \cdot s_j s_k s_l,$$

where $E\{s_j\} = 0$ and $E\{s_j^2\} = 1$, for $j = 1, \cdots, n$. Let

$$D_{ij}(s_j) \triangleq \Big(\triangledown_{i,j} + \frac{1}{2}\sum_{k \neq j} \triangledown_{i,jkk}\Big) \cdot s_j + \frac{1}{2}\triangledown_{i,jj} \cdot s_j^2 + \frac{1}{6}\triangledown_{i,jjj} \cdot s_j^3.$$

And let $\tilde{D}_{ij}(s_j)$ be the centered version of $D_i(s_j)$, that is, $\tilde{D}_{ij}(s_j) = D_{ij}(s_j) - E_i\{D_{ij}(s_j)\}$. Under assumptions A1 & A2, the estimate of s_j produced by nonlinear ICA with MND is the first principal component of $\tilde{\mathbf{D}}_{*j}(s_j) = [\tilde{D}_{1j}(s_j), \cdots, \tilde{D}_{nj}(s_j),]^T$, multiplied by a constant.

See Appendix E for a proof. Under the condition that nonlinear distortion in the mixing mapping $\mathcal{F}$ is not strong, $\tilde{D}_{ij}(s_j)$ would not be far from linear. Moreover, if the nonlinear part of $\tilde{D}_{ij}(s_j)$ varies for different i, the estimate of s_j is expected to be closer to linear than $\tilde{D}_{ij}(s_j)$, because it is the first principal component (PC) of $\tilde{\mathbf{D}}_{*j}(s_j)$. To summarize, Theorems 1 and 2 show that trivial indeterminacies in nonlinear ICA solutions can be overcome by the MND principle; and when the mixing mapping is not strong, the nonlinear distortion in the nonlinear ICA outputs w.r.t. the original sources is weak.

4.1.1 REMARK

In the proof of Theorems 1 and 2, we have made use of the fact that mutual information is invariant to any component-wise strictly monotonic nonlinear transformation of the variables. Consequently, trivial transformations do not affect the first term in Eq. 3, and they can be determined by minimizing R_{MSE} only, as claimed in the theorems. However, in practical implementations of nonlinear ICA algorithms, one needs to estimate the densities of y_i or their variations. Due to estimation error, the gradient of the mutual information $I(y_1,\cdots,y_n)$ may be sensitive to the distribution of y_i, or it may be slightly affected by trivial transformations. This may cause the results of Theorems 1 and 2 to be violated slightly.

Fortunately, this phenomenon can be avoided easily. To model the trivial transformations, we apply a separate nonlinear function approximator (such as a MLP) to each output of nonlinear ICA to generate the final nonlinear ICA result. These nonlinear function approximators are then learned by minimizing R_{MSE} (Eq. 6). This provides a way to tackle the trivial indeterminacies; after performing nonlinear ICA with any nonlinear ICA method, if we know that there only exist trivial indeterminacies, we can adopt the above technique to determine the trivial transformations.

4.2 For Non-Trivial Indeterminacies

Now let us investigate the effect of MND on non-trivial indeterminacies in nonlinear ICA solutions. Generally speaking, there exist an infinite number of ways in which non-trivial indeterminacies occur, and it is impossible to formulate all of them. Hyvärinen and Pajunen (1999) gave some families of non-trivial transformations preserving mutual independence.

4.2.1 A Particular Class of Non-Trivial Indeterminacies

For the convenience of analysis, here we consider the following manner to construct non-trivial transformations preserving mutual independence. First, using the Gaussianization technique (Chen and Gopinath, 2001), we transform each of the independent variables s_i to a standard Gaussian variable u_i with an strictly increasing function q_i, that is, $u_i = q_i(s_i)$. Clearly u_i are mutually independent. Second, we can apply an orthogonal transformation $\mathbf{U}$ to $\mathbf{u} = (u_1,\cdots,u_n)^T$. The components of $\mathbf{e} = \mathbf{U}\mathbf{u}$ are still jointly Gaussian and mutually independent.[6] Finally, let $\mathbf{y} = \mathbf{r}(\mathbf{e})$, where $\mathbf{r} = (r_1,\cdots,r_n)^T$ is a component-wise function with each r_i strictly increasing. Components of $\mathbf{y}$ are still mutually independent. That is, $\mathbf{y}$ *is always a solution to nonlinear ICA of the nonlinear mixture* $\mathbf{x} = \mathcal{F}(s)$. The procedure transforming $\mathbf{s}$ to $\mathbf{y}$ can be described as $\mathbf{r}\circ\mathbf{U}\circ\mathbf{q}$, as shown in Figure 3. When $\mathbf{U}$ is a permutation matrix, this transformation is trivial; otherwise it is not.

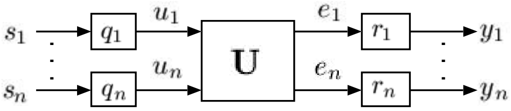

Figure 3: A non-trivial transformation from $\mathbf{s}$ to $\mathbf{y}$ preserving independence, that is, $\mathbf{r}\circ\mathbf{U}\circ\mathbf{q}$.

4.2.2 Effect of MND

To see the effect of MND on $\mathbf{y}$ in Figure 3 (recall that $\mathbf{y}$ is a solution to nonlinear ICA of $\mathbf{x} = \mathcal{F}(\mathbf{s})$), we need to find how MND affects $\mathbf{U}$, as well as r_i. First, let us consider the case where the outputs y_i are Gaussian, meaning that each component of $\tilde{\mathbf{r}}$ is a linear mapping. Without loss of generality, we further assume that y_i are zero-mean and of unit variance, that is, $E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{I}$. Consequently, r_i are identity mappings and $\mathbf{y} = \mathbf{e} = \mathbf{U}\mathbf{u}$. Assuming x_i are zero-mean, according to Eq. 5, We have $R_{MSE} = -\mathrm{Tr}\big(E\{\mathbf{x}\mathbf{y}^T\}E\{\mathbf{y}\mathbf{x}^T\}\big) + \text{const} = -\mathrm{Tr}\big(E\{\mathbf{x}\mathbf{u}^T\}\mathbf{U}^T\mathbf{U}E\{\mathbf{u}\mathbf{x}^T\}\big) + \text{const} = -\mathrm{Tr}\big(E\{\mathbf{x}\mathbf{u}^T\}E\{\mathbf{u}\mathbf{x}^T\}\big) + \text{const}$. In this case R_{MSE} is not affected by $\mathbf{U}$, and the MND principle

[6] $\mathbf{U}$ may depend on $||\mathbf{u}||$. In other words, $\mathbf{U}$ may be different for $\mathbf{u}$ of different norms.

could not help to avoid such non-trivial indeterminacies. We have empirically found that in general, when y_i are close to Gaussian, the separation performance tends to be bad. To make sure that the separation result is reliable, one should check the non-Gaussianity of y_i after the algorithm converges.

Next, suppose that that both s_l and y_j are non-Gaussian. r_i are then nonlinear. Consider the extreme case that the mixing mapping $\mathcal{F}$ is linear; in order to minimize R_{MSE} (Eq. 2), $\mathbf{U}$ in Figure 3 must be a permutation matrix. One can then image that if nonlinearity in $\mathcal{F}$ is weak enough, $\mathbf{U}$ in Figure 3 should be approximately a permutation matrix, meaning that the original sources $\mathbf{s}$ could be recovered.

However, if nonlinearity in $\mathcal{F}$ is strong, $\mathbf{U}$ may not be a permutation matrix, and non-trivial transformations from $\mathbf{s}$ to $\mathbf{y}$ may occur. This is actually quite natural. Consider the mixing mapping $\mathbf{x} = \mathcal{F}(\mathbf{s})$ which can be decomposed as a non-trivial transformation of $\mathbf{s}$ shown in Figure 3 (denote by $\mathbf{z}$ its output), followed by a nonlinear transformation $\mathbf{x} = \mathcal{F}_L(\mathbf{z})$ which is close enough to linear. In this situation, the output of nonlinear ICA with MND would be an estimate of $\mathbf{z}$, and if no additional knowledge of the mixing mapping is given, it is impossible to recover the original sources s_i.

Below we give an two-channel example to illustrate the relationship between R_{MSE} and the orthogonal matrix $\mathbf{U}$ when nonlinearity in $\mathbf{F}$ is strong. The two independent sources are a uniformly distributed signal and a super-Gaussian signal, and their scatter plot is given in Figure 9(a). The observations x_i, whose scatter plot is shown in Figure 4(a), are generated by applying a 2-3-2 MLP to the source signals. From this figure we can see that nonlinearity in the mixing procedure is comparatively strong. The orthogonal matrix $\mathbf{U}$ in Figure 3 is parameterized as $\mathbf{U} = [\cos(\alpha), -\sin(\alpha); \sin(\alpha), \cos(\alpha)]$. From Eq. 6 and Figure 3, one can see that R_{MSE} depends on α and r_i. For each value of α, r_i ($i = 1, 2$) are modelled by a 1-6-1 MLP and they are learned by minimizing R_{MSE}. Finally, $\min_{r_i} R_{MSE}$ is a function of α, with a period of 90 degrees, as plotted in Figure 4(b). In this example, α determined by the MND principle is about 11 degrees. It is not that close to zero, but it is still comparatively small and consequently the sources s_i are recovered approximately.

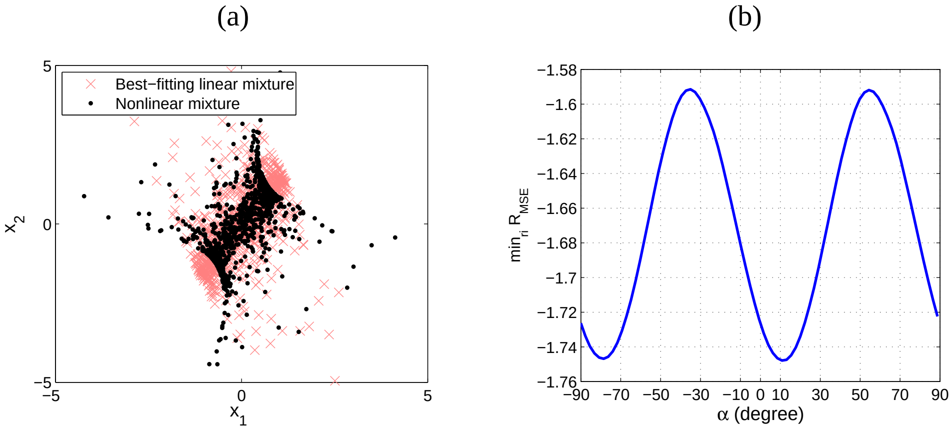

Figure 4: (a) Nonlinear mixtures of a sinusoid source signal and a super-Gaussian source signal (whose scatter plot is given in Figure 9.a) generated by a 2-3-2 MLP. x-mark points show linear mixtures of the sources which fit the nonlinear mixtures best. (b) $\min_{r_i} R_{MSE}$ as a function of α, whose minimum is achieved at $\alpha \approx 11$ degrees.

5. Simulations

In this section we investigate the performance of the proposed principle for solving nonlinear ICA using synthetic data. The experiments in Zhang and Chan (2007a) have empirically shown that both MND and the smoothness constraint are useful to ensure nonlinear ICA based on kernels to result in nonlinear BSS, when nonlinear distortion in the mixing procedure is not very strong. As its performance depends somewhat crucially on the choice of the kernel function, nonlinear ICA based on kernels is not used for comparison here. The following six methods (schemes) were used to separate various nonlinear mixtures:

1. MISEP: The MISEP method (Almeida, 2003) with parameters θ randomly initialized.[7] Note that in this method, the smoothness constraint has been implicitly incorporated to some extent, due to the property of the adopted MLP.

2. Linear init.: The MISEP method with $\mathcal{G}$ initialized as a linear mapping. This was achieved by adopting the regularization term Eq. 2 with $\lambda = 5$ (which is very large) in the first 50 epochs.

3. MND: The MISEP method incorporating MND, with R_{MSE}, the mean square error of the best linear reconstruction, as the regularization term (Section 2.2). The regularization parameter λ decayed from $\lambda_0 = 5$ to $\lambda_c = 0.01$ in the first 350 epochs. After that λ was fixed as λ_c.

4. Smooth (I): The MISEP method with the smoothness regularizer (Section 2.4) explicitly incorporated. 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: Bayesian variational nonlinear ICA (Lappalainen and Honkela, 2000; Valpola, 2000).[8] PCA was used for initialization. After obtaining nonlinear factor analysis solutions using the package, we applied linear ICA (FastICA by Hyvärinen 1999 was used) to achieve nonlinear BSS.

In addition, in order to show the necessity of nonlinear ICA methods for separating nonlinear mixtures, linear ICA (FastICA was adopted) was also used to separate the nonlinear mixtures.

It was addressed in Section 2.3 that the incorporation of direct connections between inputs and output units in the MLP representing $\mathcal{G}$ implicitly and roughly implements the MND principle. To check that, in our experiments, the MLP without direct connections and that with direct connections were both adopted to represent $\mathcal{G}$, for comparison reasons. Like in Almeida (2003), the MLP has 20 arctan hidden units, 10 of which are connected to each of the output units of $\mathcal{G}$.

We use 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. Besides, we apply a flexible nonlinear transformation h to y_i to minimize the MSE between $h(y_i)$ and s_i, and use the SNR of $h(y_i)$ relative to s_i as another performance measure. In this way possible trivial transformations between s_i and y_i are eliminated. In our experiments h was implemented by a two-layer MLP with eight hidden units with the hyperbolic tangent activation function and a linear output unit. This MLP was trained using the MATLAB neural network toolbox.

[7] Source code is available at http : //www.lx.it.pt/ ∼ lbalmeida/ica/mitoolbox.html.

[8] Source code is available at http : //www.cis.hut.fi/projects/bayes/. The following MATLAT commands were used to produce the ouput **y**: [nlfa_sources, net, params, status, fs] = nlfa(x, 'searchsources', 2, 'hidneurons', 15, 'iters', 2000); y = fastica(nlfa_sources.e, 'approach', 'symm', 'g', 'tanh');

Three kinds of nonlinear mixtures were investigated. They are distorted source (DS) mixtures, post-nonlinear (PNL) mixtures, and generic nonlinear (GN) mixtures which are generated by a MLP. Both super-Gaussian and sub-Gaussian sources were used.

5.1 For Distorted Source Mixtures

We first considered the DS mixtures defined in Eq. 15. Specifically, in the experiments the two-channel mixtures x_i were generated according to $x_1 = a_{11}s_1 + f_{12}(s_2)$, $x_2 = f_{21}(s_1) + a_{22}s_2$, where $a_{11} = a_{22} = 1$, and $f_{12}(s_i) = f_{21}(s_i) = 3\tanh(s_i/4) + 0.1s_i$. We used two super-Gaussian source signals, which are generated by $s_i = \frac{3}{5}n_i + \frac{2}{5}n_i^3$, where n_i are independent Gaussian signals. Each signal has 1000 samples. Figure 5 shows the scatter plot of the sources s_i and that of the observations x_i. To see the level of nonlinear distortion in the mixing transformation, we also give the scatter plot of the affine transformation of s_i which fits x_i the best.

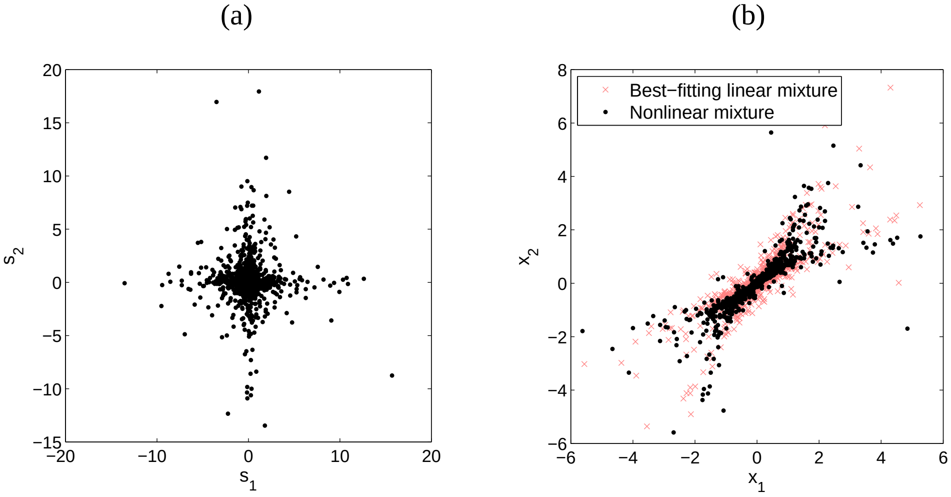

Figure 5: (a) Scatter plot of the sources s_i generating the DS mixtures. (b) Scatter plot of the DS mixtures x_i. x-mark points are linear mixtures of s_i which fit x_i best.

To reduce the random effect, all methods were repeated for 40 runs, and in each run the MLP was randomly initialized. We found that the separated results in the two channels have a similar SNR, so for saving space, here we just give the SNR in the first channel. Figure 6 compares the boxplot of SNR(y_1) and SNR$(h(y_1))$ for different methods. In Figure 6 (a, b), the MLP has no direct connections between inputs and output units, while in (c, d) 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 fewest unwanted results. Moreover, the MLP with direct connections behaves better than that without direct connections. The performance of VB-NICA is not very good. The reason may be that this method does not take into account the very useful information that nonlinearity in the mixing mapping is not very strong. 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.

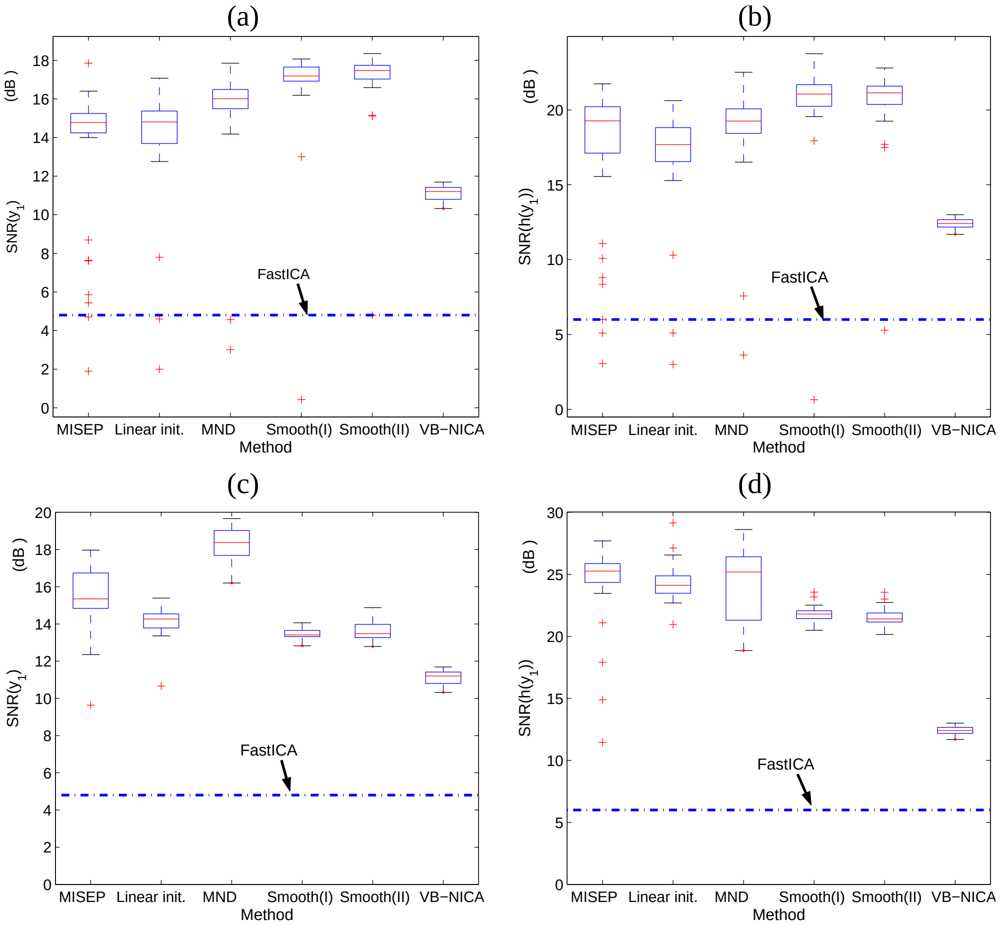

Figure 6: Boxplot of the SNR of separating the DS mixtures by the MLP without or with direct connections between inputs and output units. Top: Without direct connections. Bottom: With direct connections. (a, c) SNR(y_1). (b, d) SNR$(h(y_1))$.

5.2 For Post-Nonlinear Mixtures

The second experiment is to separate PNL mixtures. We used two sub-Gaussian source signals, which are a uniformly distributed white signal and a sinusoid waveform. The sources were first mixed with the mixing matrix **A** = [-0.2261, -0.1189; -0.1706, -0.2836], producing linear mixtures **z**. The observations were then generated as $x_1 = z_1/2.5 + \tanh(3z_1)$ and $x_2 = z_2 + z_2^3/1.5$. Figure 7 shows the scatter plot of the sources and that of the PNL mixtures (after standardization). Figure 8 gives the separation performance of s_1 by various methods.[9] In this case, the proposed nonlinear ICA with MND (labelled by MND) also gives almost the best results; especially for the MLP without direct connections, the result of nonlinear ICA with MND is clearly the best. Again, the MLP with direct connections produces better results. Moreover, one can see that compared to the DS

[9] If we use the PNL mixing model (Taleb and Jutten, 1999) to separate such mixtures, theoretically the sources could be well recovered. But in this paper we assume that the form of the mixing procedure is unknown, and treat it as a general nonlinear ICA problem.

mixtures in Section 5.1, the PNL mixtures considered here are comparatively hard to be separated by the MLP structure.

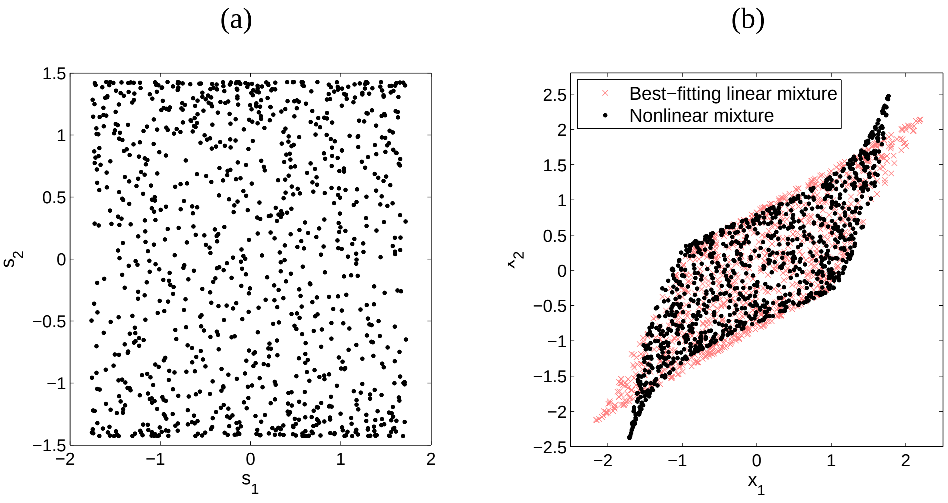

Figure 7: (a) Scatter plot of the sources s_i generating PNL mixtures. (b) Scatter plot of the PNL mixtures x_i.

5.3 For Generic Nonlinear Mixtures

We used a 2-2-2 MLP to generate nonlinear mixtures from sources. Hidden units have the arctan activation function. The weights between the input layer and the hidden layer are random numbers between -1 and 1. They are not large such that the mixing mapping is invertible and the nonlinear distortion produced by the MLP would not be very strong. The sources used here were the first source in Experiment 1 (super-Gaussian) and the second one in Experiment 2 (sub-Gaussian). Figure 9 shows the scatter plot of the sources and that of the GN mixtures. The performance of various methods for separating such mixtures is given in Figures 10. Apparently nonlinear ICA with MND gives the best separation results in this case.

Summed over all the three cases discussed above, we can see that MISEP with MND produces promising results for the general nonlinear ICA problem, provided that nonlinearity in the mixing mapping is not very strong. Specifically, it gives the fewest unwanted solutions, and its separation performance is very good. Moreover, the MLP with direct connections usually performs better than that without direct connections, but we also found that in some cases it got stuck into unwanted solutions more easily.

5.4 On Trivial Indeterminacies

In Section 4.1 we have discussed the effect of the MND principle on trivial indeterminacies of nonlinear ICA solutions. In particular, Theorem 1 states that for DS mixtures, if there are only trivial indeterminacies, each output of nonlinear ICA with MND is the PC of the contributions of the corresponding source to all mixtures. Now let us illustrate this with the help of the DS mixtures used in Section 5.1.

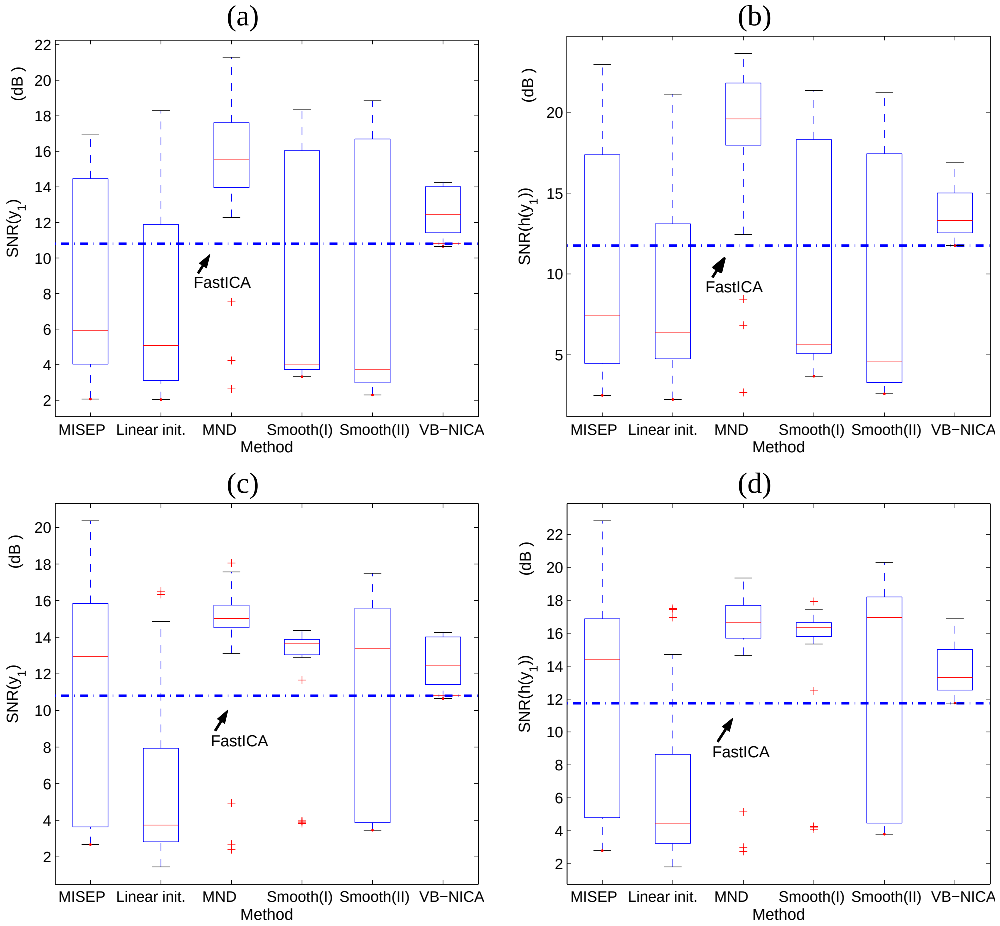

Figure 8: Boxplot of the SNR of separating the PNL mixtures by the MLP without or with direct connections between inputs and output units. Top: Without direct connections. Bottom: With direct connections. (a, c) SNR(y_1). (b, d) SNR$(h(y_1))$.

Figure 11 shows the relationship between y_i obtained by MISEP with MND in one run and the PC of $\mathbf{f}_{*i}(s_i) = [f_{1i}(s_i), f_{2i}(s_i)]^T$. We can see that each y_i is actually not very close to the corresponding PC, which may be caused by two reasons. First, there may exist some weak non-trivial transformation in the solution, as seen from the points close to the origin in Figure 11(b); y_2 is not solely dependent on s_2, but also slightly affected by s_1. The other reason is the error in estimating the density of y_i or its variation involved in the MISEP method, as explained in Section 4.1.1. We use the method proposed there to avoid the effect of the estimation error: a 1-8-1 MLP, denoted by τ_i, is applied to each y_i, and $\tau_i(y_i)$ is taken as the final nonlinear ICA output. Each τ_i is learned by minimizing R_{MSE} (Eq. 6). The resulting $\tau_i(y_i)$ is almost identical to the corresponding PC of $\mathbf{f}_{*i}(s_i)$, as seen from Figure 12. This has confirmed Theorem 1 and the validity of the method for tackling trivial indeterminacies proposed in Section 4.1.1.

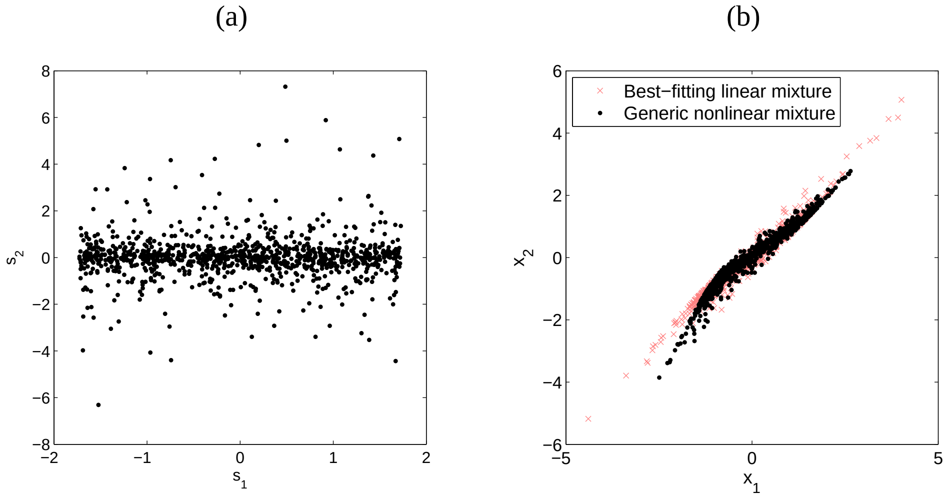

Figure 9: (a) Scatter plot of the sources s_i. (b) Scatter plot of the GN mixtures x_i.

6. Application to Causality Discovery in the Hong Kong Stock Market

In this section we give a real-life application of nonlinear ICA with MND. Specifically, we use this method to discover linear causal relations among the daily returns of a set of stocks. The empirical results were ever reported in Zhang and Chan (2006), without much detail of the method.

6.1 Introduction

It is well known that financial assets are not independent of each other, and that there may be some relations among them. Such relations can be described in different ways. In risk management, correlations are used to describe them and help to construct portfolios. The business group, which is a collection of firms bound together in some formal and/or informal ways, focuses on ties between financial assets and has attracted a lot of interest (Khanna and Rivkin, 2006). But these descriptions cannot tell us the causal relations among the financial assets.

The return of a particular stock may be influenced by those of other stocks, for many reasons, such as the ownership relations and financial interlinkages (Khanna and Rivkin, 2006). According to the efficient market hypothesis, such influence should be reflected in the stock returns immediately. In this part we aim to discover the causal relations among selected stocks by analyzing their daily returns.[10]

Traditionally, causality discovery algorithms for continuous variables usually assume that the dependencies are of a linear form and that the variables are Gaussian distributed (Pearl, 2000). Under the Gaussianity assumption, only the correlation structure of variables is considered and all higher-order information is neglected. As a consequence, one obtains some possible causal dia-

[10] In other words, here we aim to find the “instantaneous” causality in the stock market. In contrast, Granger causality (Granger, 1980) analysis has become an important tool to find the “lagged” causality between time series. A time series x_1 “Granger causes” another series x_2 if by incorporating the past history of x_1 can improve a prediction of x_2 over a prediction based only on the history of x_2 alone. The efficient market hypothesis implies no significant Granger causality between stock returns. In fact, we have applied the approach by Reale and Tunnicliffe Wilson (2001) and partial directed coherence (Baccala and Sameshima, 2001) to find the Granger causality among the selected stocks, and very few Granger causal relations were found.

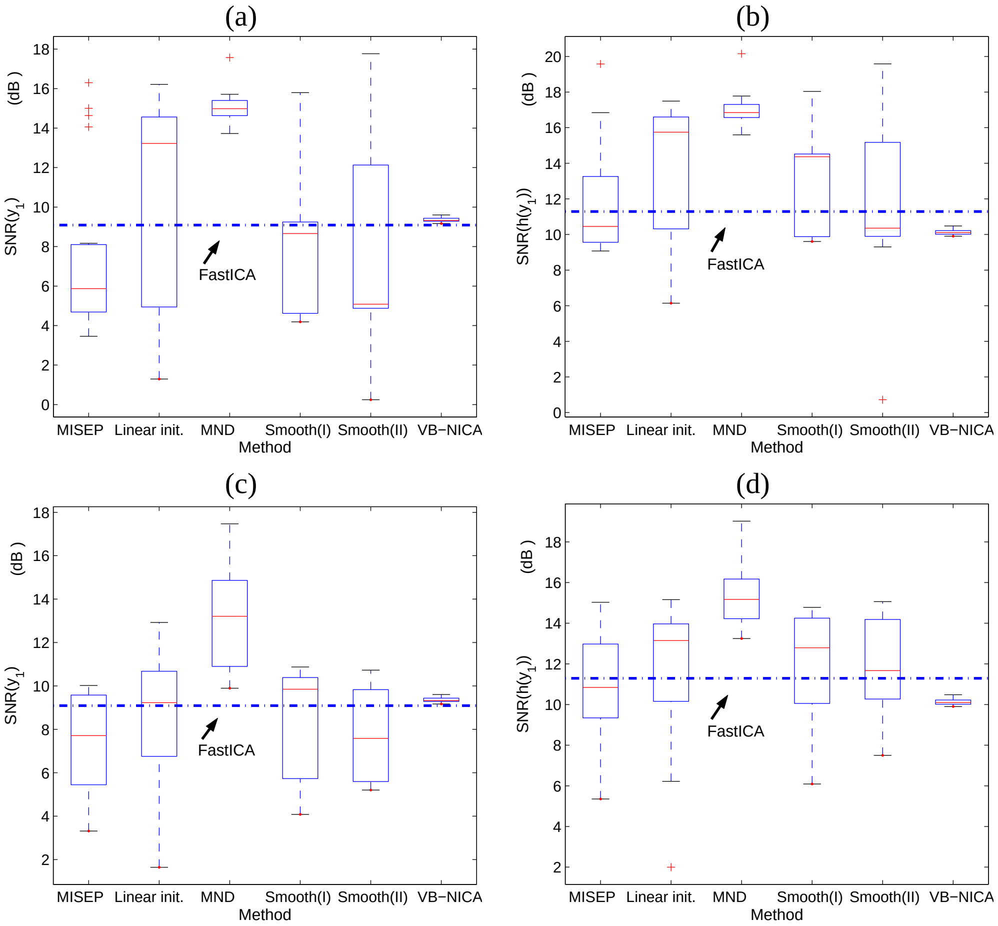

Figure 10: Boxplot of the SNR of separating the GN mixtures by the MLP without or with direct connections between inputs and output units. Top: Without direct connections. Bottom: With direct connections. (a, c) SNR(y_1). (b, d) SNR$(h(y_1))$.

grams which are equivalent in their correlation structure, and cannot 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 (Dodge and Rousson, 2001; Shimizu et al., 2006).

In particular, in Shimizu et al. (2006) an elegant and efficient method was proposed for identifying the *linear, non-Gaussian, acyclic causal model* (abbreviated LiNGAM) by exploiting ICA. If the data are generated according to the LiNGAM model, theoretically, the ICA de-mixing matrix $\mathbf{W}$ can be permuted to lower triangularity. However, in practice, this may not hold, due to the finite sample effect, the existence of unobserved confounder variables (Pearl, 2000), or mild nonlinearity and noise that are often encountered in the data generation procedure. To tackle possible mild nonlinearity in the data generation procedure, we use nonlinear ICA with MND, instead of linear ICA, to separate the observed data. As the nonlinear distortion is mild, it can be neglected and consequently, linear causal relations among the observed data can be discovered.

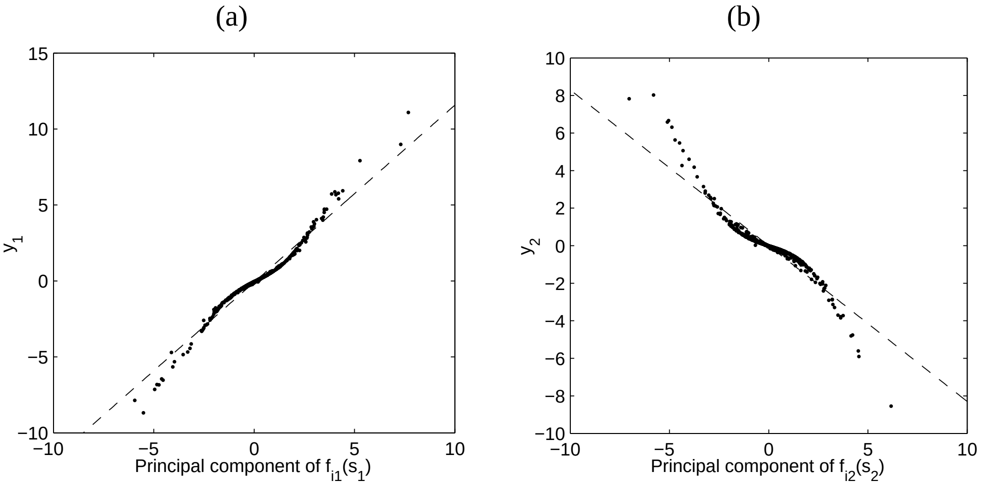

Figure 11: (a) y_1 recovered by MISEP with MND versus the PC of the contributions of s_1 to the DS mixtures used in Section 5.1. The SNR of y_1 w.r.t. the PC of the contributions of s_1 is 13.48dB. The dashed line is the linear function fitting the points best. (b) y_2 versus the PC of the contributions of s_2 to the DS mixtures. The SNR is 9.12dB.

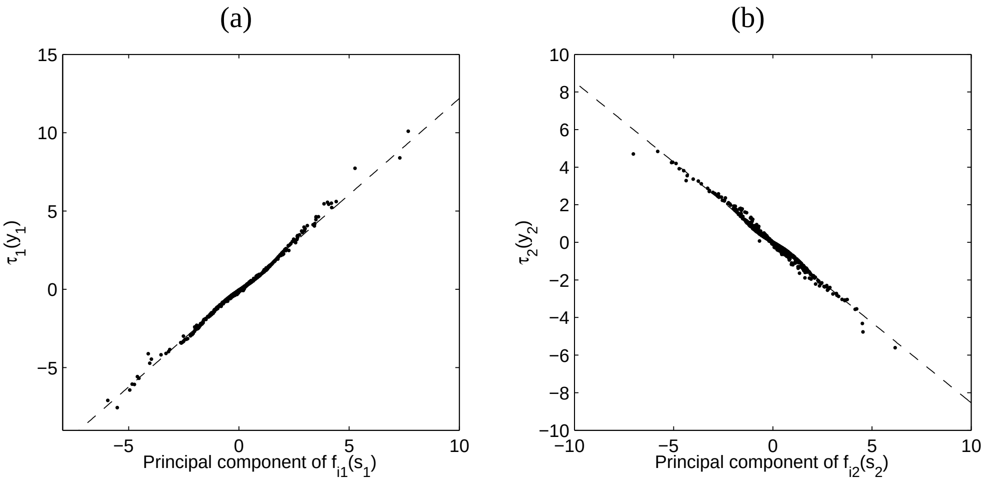

Figure 12: (a) $\tau_1(y_1)$ versus the PC of the contributions of s_1 to x_i. τ_1 is modelled by a 1-8-1 MLP and is learned by minimizing R_{MSE} (Eq. 6). The SNR is 20.99dB. (b) $\tau_2(y_2)$ versus the PC of the contributions of s_2 to x_i. The SNR is 18.64dB.

6.2 Causality Discovery by ICA: Basic Idea

The LiNGAM model assumes that the generation procedure of the observed data follows the following properties (Shimizu et al., 2006). *1.* It is recursive. This is, the observed variables x_i, $i = 1,\dots,n$, can be arranged in a causal order, such that no later variable causes any earlier variable. This causal order is denoted by $k(i)$. *2.* The value of x_i is a linear function of the values assigned to the earlier variables, plus a disturbance term e_i and an optional constant c_i: $x_i = \sum_{k(j)<k(i)} b_{ij}x_j + e_i + c_i$. *3.*

e_i are independent continuous-valued variables with non-Gaussian distributions (or at most one is Gaussian).

After centering of the variables, the causal relations among x_i can be written in the matrix form: $\mathbf{x} = \mathbf{B}\mathbf{x} + \mathbf{e}$, where $\mathbf{x} = (x_1, ..., x_n)^T$, $\mathbf{e} = (e_1, ..., e_n)^T$, and the matrix $\mathbf{B}$ can be permuted (by simultaneous equal row and column permutations) to strict lower triangularity if one knows the causal order $k(i)$ of x_i. We then have $\mathbf{e} = \mathbf{W}\mathbf{x}$, where $\mathbf{W} = \mathbf{I} - \mathbf{B}$. This is exactly the ICA separation procedure (Hyvärinen et al., 2001). Therefore, the LiNGAM model can be estimated by ICA. We can permute the rows of the ICA de-mixing matrix $\mathbf{W}$ such that it produces a matrix $\widetilde{\mathbf{W}}$ without any zero on its diagonal (or in practice, $\sum_i |\widetilde{\mathbf{W}}_{ii}|$ is maximized). Dividing each row of $\widetilde{\mathbf{W}}$ by the corresponding diagonal entry gives a new matrix $\widetilde{\mathbf{W}}'$ with all entries on its diagonal equal to 1. Finally, by applying equal row and column permutations on $\mathbf{B} = \mathbf{I} - \widetilde{\mathbf{W}}'$, we can find the matrix $\widetilde{\mathbf{B}}$ which is as close as possible to strictly lower triangularity. $\widetilde{\mathbf{B}}$ contains the causal relations of x_i. For details, see Shimizu et al. (2006).

6.3 With Nonlinear ICA with Minimal Nonlinear Distortion

We now consider a general case of the nonlinear distortion often encountered in the data generation procedure, provided that the nonlinear distortion is smooth and mild. We use the MLP structure described in Section 3.1.1, which is a linear transformation coupled with an ordinary MLP, as shown in Figure 13, to model the nonlinear transformation from the the observed variables x_i to the disturbance variables e_i.

According to Figure 13, we have $\mathbf{e} = \mathbf{W}\mathbf{x} + \mathbf{h}(\mathbf{x})$, and consequently $\mathbf{x} = (\mathbf{I} - \mathbf{W})\mathbf{x} - \mathbf{h}(\mathbf{x}) + \mathbf{e}$, where $\mathbf{h}(\mathbf{x})$ denotes the output of the MLP. As it is difficult to analyze the relations among x_i implied by the nonlinear transformation $\mathbf{h}(\mathbf{x})$, we expect that $\mathbf{h}(\mathbf{x})$ is weak such that its effect can be neglected. The *linear* causal relations among x_i can then be discovered by analyzing $\mathbf{W}$.

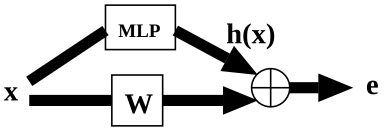

Figure 13: Structure used to model the transformation from the observed data x_i to independent disturbances e_i. $\mathbf{h}(\mathbf{x})$ accounts for nonlinear distortion if necessary.

In order to do causality discovery, the separation system in Figure 13 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*. The matrix $\mathbf{W}$ is sparse enough such that it can be permuted to lower triangularity. This can be enforced by incorporating the L_1 (Hyvärinen and Karthikesh, 2000) or smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001) on the entries of $\mathbf{W}$. *3*. The nonlinear mapping modeled by the MLP is weak enough such that we just care about the linear causal relations indicated by $\mathbf{W}$. To achieve that, we use MISEP with MND given in Section 3.1. In addition, we initialize the system with linear ICA results. That is, $\mathbf{W}$ is initialized by the linear ICA de-mixing matrix, and the initial values for weights in the MLP $\mathbf{h}(\mathbf{x})$ are very close to 0. The training process is terminated once the LiNGAM property holds for $\mathbf{W}$. After the algorithm terminates, $\frac{\mathrm{var}(h_i(\mathbf{x}))}{\mathrm{var}(e_i)}$ can be used to measure the level of nonlinear distortion in each channel, if needed.

6.4 Simulation Study

We examined the performance of the scheme discussed in Section 6.3 for identifying linear causal relations using simulated data. To make the nonlinear distortion in the data generation procedure weak, we used the structure in Figure 13 to generate the 8-dimensional observed data x_i from some independent and non-Gaussian variables e_i, that is, x_i are generated by a linear transformation coupled with a MLP.

The linear transformation in the data generation procedure was generated by $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$. It satisfies the LiNGAM property since $\mathbf{B}$ was made strict lower triangular. The magnitude of non-zero entries of $\mathbf{B}$ is uniformly distributed between 0.05 and 0.5, and the sign is random. To examine if spurious causal relations would be caused, we also randomly selected 9 entries in the strict lower triangular part of $\mathbf{B}$ and set them to zero. The disturbance variables were obtained by passing independent Gaussian variables through power non-linearities with the exponent between 1.5 and 2. The variances of e_i were randomly chosen between 0.2 and 1. These settings are similar to those in the simulation studies by Shimizu et al. (2006). The sample size is 1000. The nonlinear part is a 8-10-8 MLP with the arctan activation function in the hidden layer. The weights from the inputs to the hidden layer are between -3 and 3, that is, they are comparatively large, while those from the hidden layer to the outputs are small, such that the nonlinear distortion is weak. The nonlinear distortion level in the generation procedure is measured by the ratio of the variance of the MLP output to that of the linear output. We considered two cases where the nonlinear distortion level is 0.01 and 0.03, respectively.

We used the scheme detailed in Section 6.3 to identify the linear causal relations among x_i. The SCAD penalty was used, and there are 10 arctan hidden units connected to each output of the MLP. We repeated the simulation for 100 trials. In each trial the maximum iteration number was set to 800. The results are given in Table 1 (numbers in parentheses are corresponding standard errors). The failure rate (the chance that LiNGAM does not hold for $\mathbf{W}$ within 800 iterations), the percentages of correctly identified non-zero edges, correctly identified large edges (with the magnitude larger than 0.2), and spurious edges in the successful cases, and the resulting nonlinear distortion level $\frac{\mathrm{var}(h_i(\mathbf{x}))}{\mathrm{var}(e_i)}$ in the separation system are reported. We can see that $\mathbf{W}$ almost always satisfies the LiNGAM property, and that most causal relations (especially large ones) are successfully identified. The setting $\lambda = 0.12$, meaning that MND is explicitly incorporated, gives better results than $\lambda = 0$ does, although the difference is not large. This is not surprising because even with $\lambda = 0$, nonlinear ICA with the separation structure of Figure 13 and with $\mathbf{W}$ initialized by linear ICA could achieve MND to some extent. However, when $\lambda = 0.12$, the nonlinear distortion in the separation system is much weaker, and we found that estimated values of the entries of $\mathbf{B}$ are closer to the true ones. The penalization parameter for SCAD, λ_{SCAD}, plays an important role. A larger λ_{SCAD} would make $\mathbf{W}$ satisfy the LiNGAM property more easily, but as a price, in the result more causal relations tend to disappear or be weaker.

For comparison, we also used linear ICA with the de-mixing matrix penalized by SCAD[11] for causality discovery. The result is reported in Table 2. Even when λ_{SCAD} is very large, which causes many causal relations to disappear, as seen from the table, there is still a high probability that the resulting de-mixing matrix fails to satisfy the LiNGAM property. These results show that for

[11] The algorithm can be derived by maximizing the ICA likelihood penalized by the SCAD penalty on each entry of the de-mixing matrix. We used the natural gradient learning rule, with the score function adaptively estimated from the data.

Nonl. level in $\mathcal{F}$	Settings $(\lambda, \lambda_{SCAD})$	Fail. in 800 iter.	Edges identified	Large edges identified	Spurious edges	Nonl. level $\frac{\text{var}(h_i(\mathbf{x}))}{\text{var}(e_i)}$
0.01	$(0.12, 0.06)$	3%	88% (11%)	99% (3%)	7% (10%)	~ 0.03
	$(0, 0.06)$	3%	87% (12%)	97% (4%)	7% (9%)	~ 0.06
0.03	$(0.12, 0.10)$	1%	79% (14%)	92% (7%)	9% (11%)	~ 0.08
	$(0, 0.10)$	1%	76% (12%)	89% (8%)	8% (10%)	~ 0.13

Table 1: Simulation results of identifying linear causal relations among x_i with the nonlinear ICA structure Figure 13 and the SCAD penalty (100 trials). Numbers in parentheses are corresponding standard errors.

Nonl. level in $\mathcal{F}$	Settings	Fail. rate	Edges identified	Large edges identified	Spurious edges
0.01	$\lambda_{SCAD} = 0.2$	41%	67% (13%)	79% (15%)	4% (5%)
0.03	$\lambda_{SCAD} = 0.25$	54%	52% (12%)	58% (17%)	4% (7%)

Table 2: Simulation results of identifying linear causal relations among x_i by linear ICA with SCAD penalized de-mixing matrix (100 trials).

the data whose generation procedure has weak nonlinear distortion and approximately satisfies the LiNGAM property, nonlinear ICA with MND, together with the SCAD penalty, is useful to identify their linear causal relations.

6.5 Empirical Results

The Hong Kong stock market has some structural features different from the US and UK markets (Ho et al., 2004). One typical feature is 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.

6.5.1 DATA

Here we aim at discovering the causality network among 14 stocks selected from the Hong Kong stock market.[12] The selected 14 stocks are constituents of Hang Seng Index (HSI).[13] They are almost the largest companies of the Hong Kong stock market. We used 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, we used simple linear interpolation 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 $\mathbf{x}_t = (x_{1t}, ..., x_{14,t})^T$. Each return series contains 1331 samples.

Recently ICA has been exploited as a possible way to explain the driving forces for financial returns (Back, 1997; Kiviluoto and Oja, 1998; Chan and Cha, 2001). We conjecture that nonlinear ICA would be more suitable than linear ICA to serve this task, since it seems reasonable that the

[12] For saving space, they are not listed here; see the legend in Figure 15.

[13] The only exception is Hang Lung Development Co. Ltd (0010.hk), which was removed from HSI on Dec. 2, 2002.

ICA mixing model varies slightly for returns at different levels. So we use nonlinear ICA with MND to analyze the stock returns and to do causality discovery. However, we should be aware that it is probably very hard to discover causal relations among the selected stocks, since the financial data are somewhat non-stationary, the data generation mechanism is not clear, and there may be many confounder variables.

6.5.2 RESULTS

We first applied a standard ICA algorithm to perform ICA on the data $\mathbf{x}_t$. The natural gradient algorithm (Amari et al., 1996) with the score function adaptively estimated from data was adopted. We used the LiNGAM software[14] to permute $\mathbf{W}$ and obtain the matrix $\mathbf{B} = \mathbf{I} - \widetilde{\mathbf{W}}'$. $\mathbf{B}$ 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 0.24, which is very large. We also exploited linear ICA with the de-mixing matrix penalized by SCAD to do causality discovery. It was found that the learned de-mixing matrix $\mathbf{W}$ does not follow LiNGAM for $\lambda_{SCAD} \leq 0.25$. The value 0.25 for λ_{SCAD} is so large that statistical independence between outputs is affected. (In fact, most correlations between outputs have a magnitude larger than 0.1 when $\lambda_{SCAD} = 0.25$.) We may conclude that the data do not satisfy the LiNGAM model.

We then adopted the method proposed in Section 6.3. The SCAD penalty was applied to entries of $\mathbf{W}$ with $\lambda_{SCAD} = 0.04$. The regularization parameter for nonlinear ICA with MND (Eqs. 11 and 16–19) was $\lambda = 0.14$. After 195 epochs, $\mathbf{W}$ satisfies the LiNGAM assumption and the training process is terminated. Figure 14 shows the scatter plot of each output e_i and its linear part, from which we can see that the nonlinear distortion is weak. Based on the learned $\mathbf{W}$, we found the linear causal relations among these stocks, as shown in Figure 15. This figure was plotted using the LiNGAM software.

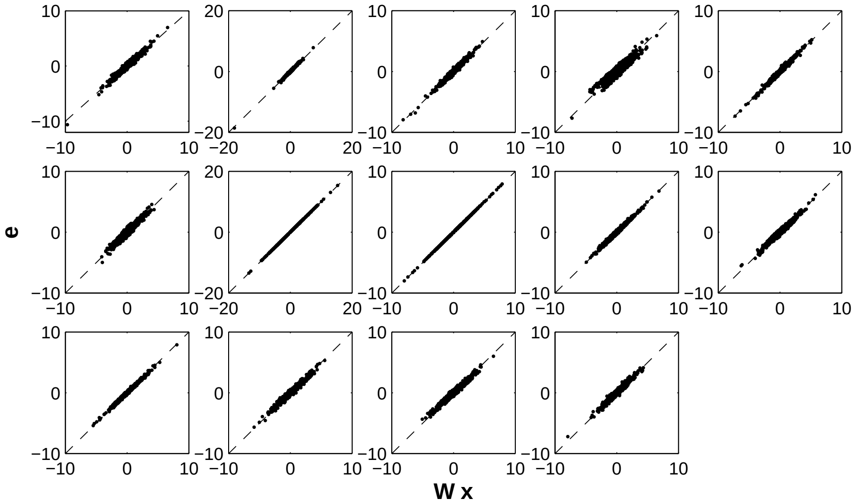

Figure 14: Scatter plot of each output of the system in Figure 13 and its linear part. The nonlinear distortion level $\frac{\text{var}(h_i(\mathbf{x}))}{\text{var}(e_i)}$ is 0.0485, 0.0145, 0.0287, 0.2075, 0.0180, 0.0753, 0, 0.0001, 0.0193, 0.0652, 0.0146, 0.0419, 0.0544, and 0.0492, respectively, for the 14 outputs e_i.

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

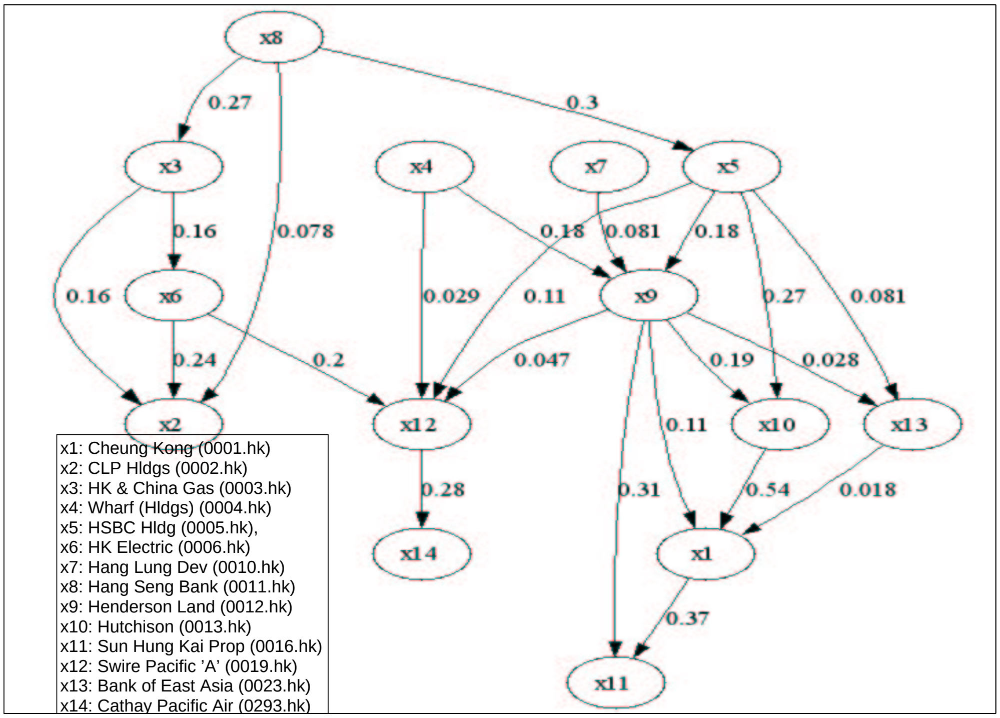
x8
0.27
0.3
x3
x4
x7
x5
0.16
0.078
0.18
0.081
0.18
0.16
x6
0.029
0.11
x9
0.27
0.081
0.24
0.2
0.047
0.19
0.028
x2
x12
0.11
x10
x13
0.28
0.31
0.54
0.018
x14
x1
0.37
x11
x1: Cheung Kong (0001.hk)
x2: CLP Hldgs (0002.hk)
x3: HK & China Gas (0003.hk)
x4: Wharf (Hldgs) (0004.hk)
x5: HSBC Hldg (0005.hk),
x6: HK Electric (0006.hk)
x7: Hang Lung Dev (0010.hk)
x8: Hang Seng Bank (0011.hk)
x9: Henderson Land (0012.hk)
x10: Hutchison (0013.hk)
x11: Sun Hung Kai Prop (0016.hk)
x12: Swire Pacific 'A' (0019.hk)
x13: Bank of East Asia (0023.hk)
x14: Cathay Pacific Air (0293.hk)

Figure 15: Causal diagram of the 14 stocks.

Figure 15 gives some interesting findings. *1.* Ownership relations tend to cause causal relations. If A is a holding company of B, there tends to be a causal relation from B to A. There are two significant relations $x_8 \rightarrow x_5$ and $x_{10} \rightarrow x_1$. 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 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. Here x_5 and x_8 are the two largest banks in Hong Kong. 4. Returns of stocks in Hang Seng Property Index tend to depend on many other stocks, while they hardly influence other stocks. Note that Here x_1, x_9, and x_{11} are in Hang Seng Property Index.

7. Conclusion

We have proposed the "minimal nonlinear distortion" principle to overcome the ill-posedness of the nonlinear ICA problem. With this principle, the nonlinear ICA solution whose estimated mixing system is close to linear would be preferred. This principle was implemented by a regularization technique that minimizes the mean square error of the best linear reconstruction of the observed mixtures. We explained how the proposed principle overcomes trivial and non-trivial indeterminacies in nonlinear ICA solutions. Experimental results on synthetic data in various situations showed that nonlinear ICA with minimal nonlinear distortion behaves very well and confirmed our theoretical claims. Since nonlinearity is usually encountered in practice and is not very strong in many cases, nonlinear ICA with minimal nonlinear distortion is expected to be capable of solving some real-life problems. Its successful application to causality discovery in the Hong Kong stock market illustrated the applicability of the method and the validity of the "minimal nonlinear distortion" principle for some real problems. The result also supports the independent factor model in finance to some extent. Finally, it should be noted that solutions to nonlinear ICA or nonlinear BSS rely heavily on the prior information on the sources or the mixing mappings. “Minimal nonlinear distortion” is one type of such information for some problems. If more precise prior information, such as the form of the mixing mapping, the temporal structure of the sources, etc., is available, the separation result may be more meaningful.

Acknowledgments

This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administration Region, China. We are very grateful to the action editor and the anonymous referees for their valuable comments and suggestions. The first author would like to thank Haixuan Yang, Gang Li, and Wan Zhang for helpful discussions.

Appendix A. Gradient of R_{MSE}

Let $\mathbf{H} = diag\{h'(a_1), h'(a_2), ..., h'(a_M)\}$, and $\mathbf{W}_j^{(2)}$ denote the j-th column of $\mathbf{W}^{(2)}$. We have

$$\begin{aligned}
\frac{\partial R_{MSE}(\theta)}{\partial \mathbf{W}^{(1)}} &= E\Big\{\sum_{i=1}^{n} K_i \cdot \frac{\partial y_i}{\partial \mathbf{W}^{(1)}}\Big\} = E\Big\{\sum_{i=1}^{n} K_i \cdot \Big[\sum_{j=1}^{M} \frac{\partial y_i}{\partial a_j} \cdot \frac{\partial a_j}{\partial \mathbf{W}^{(1)}}\Big]\Big\} \\
&= E\Big\{\sum_{j=1}^{M} \Big[\Big(\frac{\partial \mathbf{y}}{\partial a_j}\Big)^T \mathbf{K}\Big] \cdot \frac{\partial a_j}{\partial \mathbf{W}^{(1)}}\Big\} = E\Big\{\sum_{j=1}^{M} \big[h'(a_j) \cdot \mathbf{W}_j^{(2)T} \cdot \mathbf{K}\big] \cdot \frac{\partial a_j}{\partial \mathbf{W}^{(1)}}\Big\} \\
&= E\{\mathbf{H} \cdot \mathbf{W}^{(2)T} \cdot \mathbf{K} \cdot \mathbf{x}^T\}, \qquad (16) \\
\frac{\partial R_{MSE}(\theta)}{\partial \mathbf{W}^{(d)}} &= E\Big\{\sum_{i=1}^{n} K_i \cdot \frac{\partial y_i}{\partial \mathbf{W}^{(d)}}\Big\} \\
&= E\{\mathbf{K}\mathbf{x}^T\}, \qquad (17) \\
\frac{\partial R_{MSE}(\theta)}{\partial \mathbf{b}^{(2)}} &= E\{\mathbf{K}\}, \qquad (18) \\
\frac{\partial R_{MSE}(\theta)}{\partial \mathbf{b}^{(1)}} &= E\{\mathbf{H} \cdot \mathbf{W}^{(2)T} \cdot \mathbf{K}\}. \qquad (19)
\end{aligned}$$

Appendix B. Gradient of P_{ij} in Eq. 8

Noting that $\frac{\partial}{\partial\theta}\left(\frac{\partial^2 y_l}{\partial x_i \partial x_j}\right) = \frac{\partial^2}{\partial x_i \partial x_j}\left(\frac{\partial y_l}{\partial\theta}\right)$ since θ is independent from x_i, we can obtain the following rule after tedious derivation:

$$
\begin{aligned}
\frac{P_{ij}}{\partial w_{lm}^{(2)}} &= \frac{\partial^2 y_l}{\partial x_i \partial x_j} \cdot \frac{\partial^2 z_m}{\partial x_i \partial x_j}, \qquad (20)\\
\frac{\partial P_{ij}}{\partial w_{mk}^{(1)}} &= \Delta_{ijm} \cdot \left\{ h''(a_m)[w_{mi}^{(1)} \cdot \delta_{kj} + w_{mj}^{(1)} \cdot \delta_{ik}] + h'''(a_m) \cdot w_{mj}^{(1)} \cdot w_{mi}^{(1)} \cdot x_k \right\},\\
\frac{\partial P_{ij}}{\partial b_m^{(1)}} &= \Delta_{ijm} \cdot h'''(a_m) \cdot w_{mj}^{(1)} \cdot w_{mi}^{(1)},\\
\frac{\partial P_{ij}}{\partial \mathbf{W}^{(d)}} &= \mathbf{0},\\
\frac{\partial P_{ij}}{\mathbf{b}^{(2)}} &= \mathbf{0},
\end{aligned}
$$

where $\Delta_{ijm} = \sum_{l=1}^{n} w_{lm}^{(2)} \cdot \frac{\partial^2 y_l}{\partial x_i \partial x_j}$, $\frac{\partial^2 y_l}{\partial x_i \partial x_j} = \sum_{m=1}^{M} w_{lm}^{(2)} \cdot \frac{\partial^2 z_m}{\partial x_i \partial x_j}$, $\frac{\partial^2 z_m}{\partial x_i \partial x_j} = h''(a_m) \cdot w_{mi}^{(1)} \cdot w_{mj}^{(1)}$, and δ_{ik} is the Kronecker delta function.

Appendix C. Proof of Lemma 1

Proof. The mean square error of reconstructing $\mathbf{d}$ from y with the linear transformation $\mathbf{a}$ is

$$
\begin{aligned}
& E\{||\mathbf{d} - \mathbf{a} \cdot y||^2\} = E\{(\mathbf{d} - \mathbf{a} \cdot y)^T (\mathbf{d} - \mathbf{a} \cdot y)\}\\
= \; & E\{\mathbf{d}^T \cdot \mathbf{d} - 2\mathbf{a}^T \mathbf{d} \cdot y + \mathbf{a}^T \mathbf{a} \cdot y^2\}\\
= \; & E\left\{\mathbf{a}^T \mathbf{a} \cdot \left(y - \frac{\mathbf{a}^T \mathbf{d}}{\mathbf{a}^T \mathbf{a}}\right)^2 - \frac{(\mathbf{a}^T \mathbf{d})^2}{\mathbf{a}^T \mathbf{a}} + \mathbf{d}^T \mathbf{d}\right\}\\
= \; & \mathbf{a}^T \mathbf{a} \cdot E\left\{\left(y - \frac{\mathbf{a}^T \mathbf{d}}{\mathbf{a}^T \mathbf{a}}\right)^2\right\} - E\left\{\frac{(\mathbf{a}^T \mathbf{d})^2}{\mathbf{a}^T \mathbf{a}}\right\} + E\{\mathbf{d}^T \mathbf{d}\}. \qquad (21)
\end{aligned}
$$

The first term of Eq. 21 is always non-negative. No matter what value $\mathbf{a}$ takes, in order to minimize Eq. 21, we should choose

$$y = \mathbf{a}^T \mathbf{d} \cdot (\mathbf{a}^T \mathbf{a})^{-1} \qquad (22)$$

to make this term vanish, meaning that y is the linear combination of d_i with the coefficients $\mathbf{a} \cdot (\mathbf{a}^T \mathbf{a})^{-1}$.

Next, when the first term of Eq. 21 vanishes, minimizing this function w.r.t. $\mathbf{a}$ is reduced to maximizing $E\{(\mathbf{a}^T \mathbf{d})^2 \cdot (\mathbf{a}^T \mathbf{a})^{-1}\} = E\{\mathbf{a}^T \mathbf{d}\mathbf{d}^T \mathbf{a} \cdot (\mathbf{a}^T \mathbf{a})^{-1}\}$. Letting $\mathbf{a}' = \mathbf{a}/\sqrt{\mathbf{a}^T \mathbf{a}}$, this is equivalent to the constrained optimization problem: $\max \mathbf{a}'^T \cdot E\{\mathbf{d}\mathbf{d}^T\} \cdot \mathbf{a}'$, s.t. $\mathbf{a}'^T \mathbf{a}' = 1$. Clearly this is the PCA problem. So $\mathbf{a}'$ is the eigenvector of $E\{\mathbf{d}\mathbf{d}^T\}$ associated with the largest eigenvalue, and according to Eq. 22, y is the principal component of $\mathbf{d}$ multiplied by a constant.

Now let us consider the case where y is constrained to be zero-mean. Let $\bar{\mathbf{d}} = E\{\mathbf{d}\}$, and $\tilde{\mathbf{d}} = \mathbf{d} - \bar{\mathbf{d}}$. We have $E\{||\mathbf{d} - \mathbf{a} \cdot y||^2\} = E\{(\tilde{\mathbf{d}} - \mathbf{a} \cdot y + \bar{\mathbf{d}})^T (\tilde{\mathbf{d}} - \mathbf{a} \cdot y + \bar{\mathbf{d}})\} = E\{(\tilde{\mathbf{d}} - \mathbf{a} \cdot y)^T (\tilde{\mathbf{d}} - \mathbf{a} \cdot y)\} + \bar{\mathbf{d}}^T \bar{\mathbf{d}}$. $\bar{\mathbf{d}}^T \bar{\mathbf{d}}$ can be considered as a constant. Using the result above, we can see that when R_y is minimized, y is the principal component of $\tilde{\mathbf{d}}$ multiplied by a constant. (Q.E.D)

Appendix D. Proof of Theorem 1

Proof: As it has been assumed here that each output of nonlinear ICA depends only on one of the sources, we can denote by $h_j(s_j)$ the estimate of s_j produced by nonlinear ICA. For the sake of simplicity, we make both x_i and $h_j(s_j)$ zero-mean, that is, $E\{x_i\} = E\{h_j(s_j)\} = 0$. So the matrix $\mathbf{A}^*$ in Eq. 2 is $n \times n$. Denote by a^*_{ij} the (i, j)th entry of $\mathbf{A}^*$. R_{MSE} defined by Eq. 2 is

$$\begin{aligned}
R_{MSE} &= \sum_i E\Big\{x_i - \sum_j a^*_{ij}h_j(s_j)\Big\}^2 = \sum_i E\Big\{\sum_j \Big[f_{ij}(s_j) - a^*_{ij}h_j(s_j)\Big]\Big\}^2 \\
&= \sum_i \Big\{\sum_j E\big(f_{ij}(s_j) - a^*_{ij}h_j(s_j)\big)^2 \\
&\qquad + \sum_{k\neq l} E\big[\big(f_{ik}(s_k) - a^*_{ik}h_k(s_k)\big)\cdot\big(f_{il}(s_l) - a^*_{il}h_l(s_l)\big)\big]\Big\}.
\end{aligned}$$

As $E\{h_k(s_k)h_l(s_l)\} = E\{h_k(s_k)f_{il}(s_l)\} = 0$ for $k \neq l$, the above equation becomes

$$\begin{aligned}
R_{MSE} &= \sum_i \Big\{\sum_j E\big(f_{ij}(s_j) - a^*_{ij}h_j(s_j)\big)^2 + \sum_{k\neq l} E\big(f_{ik}(s_k)f_{il}(s_l)\big)\Big\} \\
&= \sum_j \Big\{\sum_i E\big(f_{ij}(s_j) - a^*_{ij}h_j(s_j)\big)^2\Big\} + \text{const.}
\end{aligned}$$

One can see that minimization of the above function can be achieved by minimizing $\sum_i E\big(f_{ij}(s_j) - a^*_{ij}h_j(s_j)\big)^2$ independently for each j. That is, $h_j(s_j)$ and a^*_{ij} are adjusted to minimize $\sum_i E\big(f_{ij}(s_j) - a^*_{ij}h_j(s_j)\big)^2$. According to Lemma 1, $h_j(s_j)$ produced by nonlinear ICA with MND is the first principal component of $\mathbf{f}_{*j}(s_j) = [f_{1j}(s_j), \cdots, f_{nj}(s_j)]^T$, multiplied by a constant. (Q.E.D)

Appendix E. Proof of Theorem 2

Proof. Denote by $h_j(s_j)$ the estimate of s_j produced by nonlinear ICA, and assume that both x_i and $h_j(s_j)$ zero-mean. Denote by a^*_{ij} the (i, j)th entry of $\mathbf{A}^*$. Note that $\sum_{j,k,l} \nabla_{i,jkl}\cdot s_js_ks_l = \sum_j \nabla_{i,jjj}\cdot s_js_js_j + 3\cdot\sum_j\sum_{k\neq j}\nabla_{i,jkk}\cdot s_js_k^2 + \sum_j\sum_{k\neq j}\sum_{\substack{l\neq j\\ l\neq k}}\nabla_{i,jkl}\cdot s_js_ks_l$, and that $E\{s_j\} = 0$ and $E\{s_j^2\} = 1$. R_{MSE} defined by Eq. 2 becomes

$$\begin{aligned}
&R_{MSE} = \sum_i E\Big\{x_i - \sum_j a^*_{ij}h_j(s_j)\Big\}^2 \\
= \;& \sum_i E\Big\{f_i(\mathbf{0}) + \sum_j \nabla_{i,j}\cdot s_j + \frac{1}{2}\sum_{j,k}\nabla_{i,jk}\cdot s_js_k + \frac{1}{6}\sum_{j,k,l}\nabla_{i,jkl}\cdot s_js_ks_l - \sum_j a^*_{ij}h_j(s_j)\Big\}^2 \\
= \;& \sum_{i=1}^n E\Big\{f_i(\mathbf{0}) + \sum_j\Big[\nabla_{i,j}\cdot s_j + \frac{1}{2}\nabla_{i,jj}\cdot s_j^2 + \frac{1}{6}\nabla_{i,jjj}\cdot s_j^3 + \frac{3}{6}\sum_{k\neq j}\nabla_{i,jkk}\cdot s_js_k^2 \\
& -a^*_{ij}h_j(s_j)\Big] + \frac{1}{2}\sum_j\sum_{k\neq j}\nabla_{i,jk}\cdot s_js_k + \frac{1}{6}\sum_j\sum_{k\neq j}\sum_{\substack{l\neq j\\ l\neq k}}\nabla_{i,jkl}\cdot s_js_ks_l\Big\}^2. \qquad (23)
\end{aligned}$$

Bearing in mind that s_j are mutually independent, and also taking all the terms independent of $h_j(s_j)$ and a^*_{ij} as constants, we can re-write Eq. 23 as

$$
\begin{aligned}
& R_{MSE} \\
= \; & \sum_i E\Big\{ \sum_j \Big[\nabla_{i,j}\cdot s_j + \frac{1}{2}\nabla_{i,jj}\cdot s_j^2 + \frac{1}{6}\nabla_{i,jjj}\cdot s_j^3 + \frac{3}{6}\sum_{k\neq j}\nabla_{i,jkk}\cdot s_j s_k^2 - a^*_{ij}h_j(s_j)\Big]\Big\}^2 + \text{const} \\
= \; & \sum_i E\Big\{ \sum_j \Big[\nabla_{i,j}\cdot s_j + \frac{1}{2}\nabla_{i,jj}\cdot s_j^2 + \frac{1}{6}\nabla_{i,jjj}\cdot s_j^3 - a^*_{ij}h_j(s_j)\Big] + \frac{1}{2}\sum_j\sum_{k\neq j}\nabla_{i,jkk}\cdot s_j s_k^2 \Big\}^2 + \text{const} \\
= \; & \sum_i E\Big\{ \Big[\sum_j \Big(\nabla_{i,j}\cdot s_j + \frac{1}{2}\nabla_{i,jj}\cdot s_j^2 + \frac{1}{6}\nabla_{i,jjj}\cdot s_j^3 - a^*_{ij}h_j(s_j)\Big)\Big]^2 \\
& - \sum_j \Big(a^*_{ij}h_j(s_j)\cdot \sum_{k\neq j}\nabla_{i,jkk}\cdot s_j s_k^2 \Big)\Big\} + \text{const} \\
= \; & \sum_i E\Big\{ \sum_j \Big(\nabla_{i,j}\cdot s_j + \frac{1}{2}\nabla_{i,jj}\cdot s_j^2 + \frac{1}{6}\nabla_{i,jjj}\cdot s_j^3 - a^*_{ij}h_j(s_j)\Big)^2 \\
& - \sum_j \Big(a^*_{ij}h_j(s_j)\cdot \sum_{k\neq j}\nabla_{i,jkk}\cdot s_j \Big\} + \text{const} \\
= \; & \sum_i\sum_j E\Big\{ \Big(\nabla_{i,j} + \frac{1}{2}\sum_{k\neq j}\nabla_{i,jkk}\Big)\cdot s_j + \frac{1}{2}\nabla_{i,jj}\cdot s_j^2 + \frac{1}{6}\nabla_{i,jjj}\cdot s_j^3 - a^*_{ij}h_j(s_j)\Big\}^2 + \text{const} \\
= \; & \sum_j \Big[\sum_i E\big(D_{ij}(s_j) - a^*_{ij}h_j(s_j)\big)^2 \Big] + \text{const.}
\end{aligned}
$$

Note that there is no dependence relationship between $h_j(\cdot)$, as well as a^*_{ij}, with different j. To minimize the above function, we just need to adjust $h_j(s_j)$ and a^*_{ij} to minimize $\sum_i E\big(D_i(s_j) - a^*_{ij}h_j(s_j)\big)^2$, independently for each j. According to Lemma 1, $h_j(s_j)$ is the first principal component of $\tilde{\mathbf{D}}_{*j}(s_j) = [\tilde{D}_{1j}(s_j), \cdots, \tilde{D}_{nj}(s_j),]^T$, multiplied by a constant. (Q.E.D)

References

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

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

S. Amari, A. Cichocki, and H.H. Yang. A new learning algorithm for blind signal separation. In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8, pages 757–763, Cambridge, MA, 1996. MIT Press.

L.A. Baccala and K. Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biol. Cybern.*, 84:463–474, 2001.

A.D. Back. A first application of independent component analysis to extracting structure from stock returns. *International Journal of Neural Systems*, 8(4):473–484, August 1997.

A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

C.M. Bishop. Curvature-driven smoothing: a learning algorithm for feedforward networks. *IEEE Transactions on Neural Networks*, 4(5):882–884, 1993.

C.M. Bishop. Regularization and complexity control in feed-forward networks. In *Proc. International Conference on Artificial Neural Networks (ICANN'95)*, volume 1, pages 141–148, 1995.

G. Burel. Blind separation of sources: a nonlinear neural algorithm. *Neural Networks*, 5(6):937–947, 1992.

J.F. Cardoso. Blind signal separation: Statistical principles. *Proceeding Of The IEEE, special issue on blind identification and estimation*, 9(10):2009–2025, 1998.

L. Chan and S.M. Cha. Selection of independent factor model in finance. In *proceedings of 3rd International Conference on Independent Component Analysis and blind Signal Separation*, San Diego, California, USA, December 2001.

S.S. Chen and R.A. Gopinath. Gaussianization. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 423–429. MIT Press, 2001.

A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, UK, corrected and revisited edition, 2003.

Y. Dodge and V. Rousson. On asymmetric properties of the correlation coefficient in the regression setting. *The American Statistician*, 55(1):51–54, 2001.

J. Eriksson and V. Koivunen. Blind identification of class of nonlinear instantaneous ICA models. In *Proc. of the XI European SIgnal Proc. Conf. (EUSIPCO 2002)*, volume 2, pages 7–10, Toulouse, France, Sept. 2002.

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

C. Granger. Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329-352, 1980.

W.E.L. Grimson. A computational theory of visual surface interpolation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 298:395–427, 1982.

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

R.Y. Ho, R. Strange, and J. Piesse. The structural and institutional features of the Hong Kong stock market: Implications for asset pricing. Research Paper 027, The Management Centre Research Papers, King's College London, 2004.

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

A. Hyvärinen and R. Karthikesh. Sparse priors on the mixing matrix in independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 477–452, Helsinki, Finland, 2000.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley & Sons, Inc, 2001.

C. Jutten and A. Taleb. Source separation: From dusk till dawn. In *2nd International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, pages 15–26, Helsinki, Finland, 2000.

A.M. Kagan, Y.V. Linnik, and C.R. Rao. *Characterization Problems in Mathematical Statistics*. Wiley, New York, 1973.

T. Khanna and J.W. Rivkin. Interorganizational ties and business group boundaries: Evidence from an emerging economy. *Organization Science*, 17(3):333-352, 2006.

K. Kiviluoto and E. Oja. Independent component analysis for parallel financial time series. In *Proc. ICONIP'98*, volume 2, pages 895–898, Tokyo, Japan, 1998.

H. Lappalainen and A. Honkela. Bayesian nonlinear independent component analysis by multilayer perceptron. In M.Girolami, editor, *Advances in Independent Component Analysis*, pages 93–121. Spring-Verlag, 2000.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

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

M. Reale and G. Tunnicliffe Wilson. Identification of vector ar models with recursive structural errors using conditional independence graphs. *Statistical Methods and Applications*, 10(1-3): 49–65, 2001.

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

A. Taleb. A generic framework for blind source separation in structured nonlinear models. *IEEE Transactions on Signal Processing*, 50(8):1819–1830, 2002.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Trans. on Signal Processing*, 47(10):2807–2820, 1999.

Y. Tan, J. Wang, and J. M. Zurada. Nonlinear blind source separation using a radial basis function network. *IEEE Trans. on Neural Networks*, 12(1):124–134, 2001.

A.N. Tikhonov and V.A. Arsenin. *Solutions of Ill-posed Problems*. Winston & Sons, Washington, 1977.

H. Valpola. Nonlinear independent component analysis using ensemble learning: Theory. In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pages 251–256, Helsinki, Finland, 2000.

L. Xu. Least mean square error reconstruction principle for self-organizing neural-nets. *Neural Networks*, 6:627–648, 1993.

H.H. Yang, S. Amari, and A. Cichocki. Information-theoretic approach to blind separation of sources in nonlinear mixture. *Signal Processing*, 64(3):291–300, 1998.

K. Zhang and L. Chan. Kernel-based nonlinear independent component analysis. In *Proc. Int. Workshop on Independent Component Analysis and Signal Separation (ICA2007)*, pages 301–308, London, UK, Sept. 2007a.

K. Zhang and L. Chan. Nonlinear independent component analysis with minimum nonlinear distortion. In *the 24th Annual International Conference on Machine Learning (ICML 2007)*, pages 1127–1134, Corvallis, OR, US, Jun. 2007b.

K. Zhang and L. Chan. Extensions of ICA for causality discovery in the Hong Kong stock market. In *Proc. 13th International Conference on Neural Information Processing (ICONIP 2006)*, pages 400–409, Hong Kong, 2006.