A general linear non-Gaussian state-space model: Identifiability, identification, and applications

Kun Zhang KZHANG@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems
Spemannstr. 38, 72076 Tübingen, Germany

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Dept of Computer Science, HIIT, and *Dept of Mathematics and Statistics*
University of Helsinki, Finland

Editor: Chun-Nan Hsu and Wee Sun Lee

Abstract

State-space modeling provides a powerful tool for system identification and prediction. In linear state-space models the data are usually assumed to be Gaussian and the models have certain structural constraints such that they are identifiable. In this paper we propose a non-Gaussian state-space model which does not have such constraints. We prove that this model is fully identifiable. We then propose an efficient two-step method for parameter estimation: one first extracts the subspace of the latent processes based on the temporal information of the data, and then performs multichannel blind deconvolution, making use of both the temporal information and non-Gaussianity. We conduct a series of simulations to illustrate the performance of the proposed method. Finally, we apply the proposed model and parameter estimation method on real data, including major world stock indices and magnetoencephalography (MEG) recordings. Experimental results are encouraging and show the practical usefulness of the proposed model and method.

Keywords: State-space model, Non-Gaussianity, Identifiability, Causality

1. Introduction

Suppose that we have multiple parallel times series which are observable. Usually, the source series of interest are not directly measurable, but hidden in them. In addition, the mixing system generating the observable series from the sources is unknown. For simplicity, we often assume that the mixing system is linear. The goal is to recover the latent interesting sources, as well as to model their dynamics. This problem was referred to as blind source separation (BSS, see e.g., books by Cichocki and Amari (2003) and by Hyvärinen et al. (2001)). In the literature, statistical independence has played a great role in BSS; in most BSS algorithms, the sources are assumed to be statistically independent. In the noiseless case, certain techniques have been proposed to solve this problem efficiently. For example, if the sources are non-Gaussian (or at most one of them is Gaussian), BSS can be solved by the independent component analysis (ICA) technique (Hyvärinen et al., 2001). If the sources are temporally correlated, simultaneous diagonalization of the cross-correlations makes source separation possible (Belouchrani et al., 1997).

©2011 Kun Zhang and Aapo Hyvärinen.

In practice the data are usually noisy, i.e., the observed data contain observation errors, and the latent source processes exhibit some temporal structures (which may include delayed influences between them). The state-space representation then offers a powerful modeling approach. Here we are particularly interested in the linear state-space model (SSM) or linear dynamic system (Kalman, 1960; van Overschee and de Moor, 1996). Denote by $\mathbf{x}_t = (x_{1t}, ..., x_{nt})^T$, $t = 1, ..., N$, the vector of the observed signals, and by $\mathbf{y}_t = (y_{1t}, ..., y_{mt})^T$ the vector of latent processes which are our main object of interest.[1] The observed data are assumed to be linear mixtures of the latent processes together with some noise effect, while the latent processes follow a vector autoregressive (VAR) model. Mathematically, we have

$$\mathbf{x}_t = \mathbf{A}\mathbf{y}_t + \mathbf{e}_t, \tag{1}$$

$$\mathbf{y}_t = \sum_{\tau=1}^{L} \mathbf{B}_\tau \mathbf{y}_{t-\tau} + \boldsymbol{\epsilon}_t, \tag{2}$$

where $\mathbf{e}_t = (e_{1t}, ..., e_{nt}^T)$ and $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, ..., \epsilon_{mt})^T$ denote the observation error and process noise, respectively. Moreover, $\mathbf{e}_t$ and $\boldsymbol{\epsilon}_t$ are both temporally white and independent of each other. One can see that because of the state transition matrices $\mathbf{B}_\tau$, y_{it} are generally dependent, even if ϵ_{it} are mutually independent.

In traditional SSMs, both $\boldsymbol{\epsilon}_t$ and $\mathbf{e}_t$ are assumed to be Gaussian; or equivalently, one makes use of their covariance structure, and the statistical properties beyond second-order are not considered. In Kalman filtering (Kalman, 1960), $\mathbf{A}$ and $\mathbf{B}_\tau$ are given, and the goal is to do inference, i.e., to estimate $\mathbf{y}_t$ based on $\{\mathbf{x}_t\}$. Learning of the parameters $\mathbf{A}$, $\mathbf{B}_\tau$, and the covariance matrices of $\mathbf{e}_t$ and $\boldsymbol{\epsilon}_t$ was also studied; see, e.g, van Overschee and de Moor (1991); Ghahramani and Hinton (1996). However, it is well-known that under the above assumptions, the SSM model is generally not identifiable; see e.g., Arun and Kung (1990), and consequently, one can not use this model to recover the latent processes y_{it}.

Under specific structural constraints on $\mathbf{B}_\tau$ or $\mathbf{A}$, the SSM model (1∼2) may become identifiable, so that it can be used to reveal the underlying structure of the data. Many existing models which are used for source separation or prediction of time series can be considered as special cases of this model. For instance, the temporal structure based source separation (Murata et al., 2001) assume that $\mathbf{B}_\tau$ are diagonal. The model also becomes identifiable with some other structural constraints on $\mathbf{A}$, as discussed in Xu (2002). However, one should note that in practice such constraints may not hold; for instance, for the electroencephalography (EEG) or magnetoencephalography (MEG) data, some underlying processes or sources may have delayed influences on others, and letting $\mathbf{B}_\tau$ be diagonal will destroy these types of connectivities.

On the other hand, distributional information also helps system identification. One can ignore the temporal information and perform system identification based on the non-Gaussianity of the data. For example, if the matrices $\mathbf{B}_\tau$ are zero and $e_i(t)$ are non-Gaussian, it is reduced to the noisy ICA problem or the independent factor analysis (IFA) model (Attias, 1999). In the noiseless case, ICA could recover the underlying linear mixing system up to trivial indeterminacies. But in the noisy case, the model is just partially

1. We use the terms latent processes, factors, and sources interchangeably in this paper, depending on application scenarios.

identifiable (Davies, 2004); in fact, the distributions of y_{it} could not be uniquely determined, even with a given mean and variance.

For generality, we would like to keep the flexibility of the SSM model (i.e., do not use specific structural constraints on $\mathbf{A}$ and $\mathbf{B}_\tau$), but still make it identifiable by incorporating some assumption which usually holds in practice. Non-Gaussianity of the distribution plays such a role: we assume that the process noise $\boldsymbol{\epsilon}_t$ has non-Gaussian and independent components. Hence our proposed model is called non-Gaussian SSM (NG-SSM). We will show that combining the temporal information and the distributional information makes NG-SSM fully identifiable. This enables finding the sources, or latent processes, successfully even when they are dependent. Note that a special case of the proposed model, which does not have the observation error $\mathbf{e}_t$, was recently exploited to analyze the connectivity between the brain sources for EEG or MEG (Haufe et al., 2010; Gómez-Herrero et al., 2008).

It is also interesting to note that NG-SSM could serve as another scheme to do Granger causality analysis (Granger, 1988; Hyvärinen et al., 2010). A time series z_{1t} is said to Granger cause another series z_{2t} if z_{1t} has incremental predictive power when forecasting z_{2t}. Granger causality can be readily extended to the case with more than two time series; see, e.g., Ding et al. (1996). VAR is a widely-used way to represent Granger causality. As (2) is a VAR model of y_{it} with contemporaneously independent residuals, from another point of view, one can see that NG-SSM finds the latent factors y_{it} which can be explained well by Granger causality. In this sense, NG-SSM provides a way to do "Granger causal factor" analysis.

Our contribution is two-fold. First, we prove the identifiability of NG-SSM, and more specifically, we show how the non-Gaussianity of the process noise makes the model fully identifiable. Second, we present an identification method which, as illustrated by numerical studies, clearly outperforms the previous identification approach. The rest of this paper is organized as follows. In Section 2 we give a rigorous proof of the identifiability of the proposed model. A simple and efficient method for system identification is presented in Section 3, followed by some simulation results in Section 4. Section 5 reports the results of analyzing financial and MEG data with the proposed model. Finally we conclude the paper and give further discussions in Section 6.

2. Identifiability

We consider the SSM (1∼2) under the non-Gaussianity assumption on $\boldsymbol{\epsilon}_t$. Denote by $\boldsymbol{\Sigma}_\mathbf{e}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ the covariance matrices of $\mathbf{e}_t$ and $\boldsymbol{\epsilon}_t$, respectively. Without loss of generality, we assume that the data are zero-mean. Further, we make the following assumptions on the proposed NG-SSM.

A1. $n \geq m$, both the observation error $\mathbf{e}_t$ and process noise $\boldsymbol{\epsilon}_t$ are temporally white, and $\boldsymbol{\epsilon}_t$ has i.i.d. mutually independent components.

A2. $\mathbf{A}$ is of full column rank, and $\mathbf{B}_L$ is of full rank.

A3. The VAR process (2) is stable. That is, the roots of $\det(\mathbf{I} - \sum_{\tau=1}^{L} \mathbf{B}_\tau z^{-\tau}) = 0$, where z is the delay operator and $\mathbf{I}$ the identity matrix, lie inside the complex unit circle, i.e., with modulus smaller than one.

A4. The process noise has a unit variance, i.e., $\mathbf{\Sigma}_{\boldsymbol{\epsilon}} = \mathbf{I}$. This assumption is to avoid the scaling indeterminacy of ϵ_{it} or y_{it}.

A5. At most one component of the process noise $\boldsymbol{\epsilon}_t$ is Gaussian, and the observation error $\mathbf{e}_t$ is either Gaussian or non-Gaussian.

Differently from the Kalman filtering task which aims to do inference with given parameters, the goal in this contribution is to learn $(\mathbf{A}, \{\mathbf{B}_\tau\}_{\tau=1}^L, \mathbf{\Sigma}_{\mathbf{e}})$, as well as to recover the latent process $\mathbf{y}_t$.

Now we show the identifiability of the proposed NG-SSM model (1∼2). First, one can see that without the non-Gaussianity assumption of the data, the temporal information helps to partially identify the model. Similar results have been mentioned in the literature (Arun and Kung, 1990; van Overschee and de Moor, 1996); however, for completeness of the theory and consistency of the presentation, we give a rigorous formulation and proof; see Lemma 1.

We then prove that the non-Gaussianity of the process noise further allows NG-SSM to be fully identifiable. For simplicity of the presentation, in Theorem 2 we consider the case with $n = m$. The case with $n > m$ is considered later in Theorem 3. We start by proving that using only the second-order temporal information in the data, the observation error covariance matrix $\mathbf{\Sigma}_{\mathbf{e}}$, as well as other quantities such as $\mathbf{A}\mathbf{B}^k\mathbf{A}^T$, $k = 0, 1, 2, ...$, are identifiable, as stated in the following lemma.

Lemma 1 *Consider the model given by (1∼2) with $n = m$ and given L. If the assumptions A1∼A4 hold, by making use of the autocovariances of the observed data $\mathbf{x}_t$, the noise covariance $\mathbf{\Sigma}_{\mathbf{e}}$ and $\mathbf{A}\mathbf{B}^k\mathbf{A}^T$ can be uniquely determined; furthermore, $\mathbf{A}$ and $\mathbf{B}_\tau$ can be identified up to some rotation transformations. That is, suppose that the NG-SSM model with parameters $(\mathbf{A}, \{\mathbf{B}_\tau\}_{\tau=1}^L, \mathbf{\Sigma}_{\mathbf{e}})$ and that with $(\tilde{\mathbf{A}}, \{\tilde{\mathbf{B}}_\tau\}_{\tau=1}^L, \tilde{\mathbf{\Sigma}}_{\tilde{\mathbf{e}}})$ are observationally equivalent; we then have $\tilde{\mathbf{\Sigma}}_{\tilde{\mathbf{e}}} = \mathbf{\Sigma}_{\mathbf{e}}$, $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{U}$, $\tilde{\mathbf{B}}_\tau = \mathbf{U}^T\mathbf{B}_\tau\mathbf{U}$, where $\mathbf{U}$ is a m-dimensional orthogonal matrix.*[2]

Proof Define the autocovariance function of $\mathbf{y}$ at lag k as $\mathbf{R}_{\mathbf{y}}(k) = \mathbb{E}(\mathbf{y}_t\mathbf{y}_{t+k}^T)$, and similarly for $\mathbf{R}_{\mathbf{x}}(k)$. Clearly $\mathbf{R}_{\mathbf{y}}(-k) = \mathbf{R}_{\mathbf{y}}(k)^T$ and $\mathbf{R}_{\mathbf{x}}(-k) = \mathbf{R}_{\mathbf{x}}(k)^T$. Due to (2), we have

$$\mathbf{R}_{\mathbf{y}}(k) = \mathbb{E}\Big[\mathbf{y}_t\Big(\sum_{\tau=1}^L \mathbf{B}_\tau\mathbf{y}_{t+k-\tau} + \boldsymbol{\epsilon}_{t+k}\Big)^T\Big] = \begin{cases}\sum_{\tau=1}^L \mathbf{R}_{\mathbf{y}}(k-\tau)\mathbf{B}_\tau^T, & \text{for } k \neq 0;\\ \sum_{\tau=1}^L \mathbf{R}_{\mathbf{y}}(\tau)^T\mathbf{B}_\tau^T + \mathbf{I}, & \text{for } k = 0.\end{cases} \tag{3}$$

Let $\tilde{\mathbf{x}}_t \triangleq \mathbf{A}\mathbf{y}_t$, so $\mathbf{x}_t = \tilde{\mathbf{x}}_t + \mathbf{e}_t$. From (1) one can see that

$$\mathbf{R}_{\tilde{\mathbf{x}}}(k) = \mathbf{A}\mathbf{R}_{\mathbf{y}}(k)\mathbf{A}^T. \tag{4}$$

Further considering the fact that $\mathbf{A}$ is invertible, combining (3) and (4) gives that $\mathbf{R}_{\tilde{\mathbf{x}}}(k)$ satisfies the recursive relationship:

$$\mathbf{R}_{\tilde{\mathbf{x}}}(k) = \begin{cases}\sum_{\tau=1}^L \mathbf{R}_{\tilde{\mathbf{x}}}(k-\tau)\mathbf{C}_\tau^T, & \text{for } k \neq 0,\\ \sum_{\tau=1}^L \mathbf{R}_{\tilde{\mathbf{x}}}(k-\tau)\mathbf{C}_\tau^T + \mathbf{A}\mathbf{A}^T, & \text{for } k = 0,\end{cases} \tag{5}$$

2. Note that here we already assumed that $\mathbb{V}\mathrm{ar}(\epsilon_{it}) = 1$. Otherwise, given $\mathbf{\Sigma}_{\mathbf{e}}$ and $\tilde{\mathbf{\Sigma}}_{\tilde{\mathbf{e}}}$, both of which are diagonal with positive diagonal entries, $\tilde{\mathbf{A}}$ can be represented as $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{\Sigma}_{\mathbf{e}}^{1/2}\mathbf{U}\tilde{\mathbf{\Sigma}}_{\tilde{\mathbf{e}}}^{-1/2}$ and the expression for $\tilde{\mathbf{B}}_\tau$ is more complex.

where $\mathbf{C}_\tau \triangleq \mathbf{A}\mathbf{B}_\tau\mathbf{A}^{-1}$. Also bearing in mind that $\mathbf{R}_\mathbf{x}(k) = \mathbf{R}_{\tilde{\mathbf{x}}}(k)$ for $k \neq 0$ and $\mathbf{R}_\mathbf{x}(0) = \mathbf{R}_{\tilde{\mathbf{x}}}(0) + \mathbf{\Sigma_e}$, (5) can be written in the matrix form:

$$\begin{bmatrix} \mathbf{R}_\mathbf{x}(0) - \mathbf{A}\mathbf{A}^T \\ \mathbf{R}_\mathbf{x}(1) \\ \vdots \\ \mathbf{R}_\mathbf{x}(L) \\ \hdashline \mathbf{R}_\mathbf{x}(L+1) \\ \vdots \\ \mathbf{R}_\mathbf{x}(2L) \end{bmatrix} = \mathbf{H} \cdot \underbrace{\begin{bmatrix} \mathbf{C}_1^T \\ \mathbf{C}_2^T \\ \vdots \\ \mathbf{C}_L^T \end{bmatrix}}_{\triangleq \mathbf{C}} - \mathbf{\Sigma_e} \cdot \begin{bmatrix} -\mathbf{I} \\ \mathbf{C}_1^T \\ \vdots \\ \mathbf{C}_L^T \\ \hdashline \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}, \tag{6}$$

where

$$\mathbf{H} \triangleq \begin{bmatrix} \mathbf{R}_\mathbf{x}(1)^T & \mathbf{R}_\mathbf{x}(2)^T & \dots & \mathbf{R}_\mathbf{x}(L)^T \\ \mathbf{R}_\mathbf{x}(0) & \mathbf{R}_\mathbf{x}(1)^T & \dots & \mathbf{R}_\mathbf{x}(L-1)^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_\mathbf{x}(L-1) & \mathbf{R}_\mathbf{x}(L-2) & \dots & \mathbf{R}_\mathbf{x}(0) \\ \hdashline \mathbf{R}_\mathbf{x}(L) & \mathbf{R}_\mathbf{x}(L-1) & \dots & \mathbf{R}_\mathbf{x}(1) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_\mathbf{x}(2L-1) & \mathbf{R}_\mathbf{x}(2L-2) & \dots & \mathbf{R}_\mathbf{x}(L) \end{bmatrix}.$$

The lower block of $\mathbf{H}$ is actually $\mathbb{E}(\overrightarrow{\mathbf{x}}_t \overrightarrow{\mathbf{x}}_{t+L}^T)$, where $\overrightarrow{\mathbf{x}}_t = (\mathbf{x}_t^T, \mathbf{x}_{t-1}^T, ..., \mathbf{x}_{t-L+1}^T)^T$. Let $\overrightarrow{\mathbf{C}} \triangleq \left[\begin{array}{c:c} \mathbf{C} & \begin{matrix} \mathbf{I}_{m(L-1)} \\ \mathbf{0}_{m \times m(L-1)} \end{matrix} \end{array}\right]$, where $\mathbf{I}_{m(L-1)}$ denotes the $m(L-1)$-dimensional identity matrix, and $\mathbf{0}_{m \times m(L-1)}$ the $m \times m(L-1)$ zero matrix. As $\det(\overrightarrow{\mathbf{C}}) = \det(\mathbf{C}_L)$, which is not zero due to Assumption A2, $\overrightarrow{\mathbf{C}}$ is nonsingular. One can easily see that $\mathbb{E}(\overrightarrow{\mathbf{x}}_t \overrightarrow{\mathbf{x}}_{t+1}^T) = \mathbb{E}(\overrightarrow{\mathbf{x}}_t \overrightarrow{\mathbf{x}}_t^T)\overrightarrow{\mathbf{C}}$. Consequently $\mathbb{E}(\overrightarrow{\mathbf{x}}_t \overrightarrow{\mathbf{x}}_{t+L}^T) = \mathbb{E}(\overrightarrow{\mathbf{x}}_t \overrightarrow{\mathbf{x}}_t^T)\overrightarrow{\mathbf{C}}^L$, and the lower block of $\mathbf{H}$ is thus non-singular. Hence, we can find the unique solution for $\mathbf{C}_\tau$, $\tau = 1, 2, ..., L$ from the bottom mL equations of (6). Substituting the solutions for $\mathbf{C}_\tau$ into the top $2m$ equations, we can get the unique solution for $\mathbf{\Sigma_e}$ and $\mathbf{A}\mathbf{A}^T$.

Since $(\mathbf{A}, \{\mathbf{B}_\tau\}_{\tau=1}^L, \mathbf{\Sigma_e})$ and $(\tilde{\mathbf{A}}, \{\tilde{\mathbf{B}}_\tau\}_{\tau=1}^L, \tilde{\mathbf{\Sigma}}_{\tilde{\mathbf{e}}})$ produce the same $\mathbf{R}_\mathbf{x}(k)$, we have $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{A}\mathbf{A}^T$ and $\tilde{\mathbf{A}}\tilde{\mathbf{B}}_\tau\tilde{\mathbf{A}}^{-1} = \mathbf{A}\mathbf{B}\mathbf{A}^{-1}$. As $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{A}\mathbf{A}^T$ and $\tilde{\mathbf{A}}$ is nonsingular, we have $\mathbf{A}^{-1}\tilde{\mathbf{A}} \cdot (\mathbf{A}^{-1}\tilde{\mathbf{A}})^T = \mathbf{I}$, i.e., $\mathbf{A}^{-1}\tilde{\mathbf{A}} = \mathbf{U}$, or $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{U}$, where $\mathbf{U}$ is an orthogonal matrix. Furthermore, multiplying both sides of $\tilde{\mathbf{A}}\tilde{\mathbf{B}}_\tau\tilde{\mathbf{A}}^{-1} = \mathbf{A}\mathbf{B}\mathbf{A}^{-1}$ and $\tilde{\mathbf{A}}\tilde{\mathbf{A}}^T = \mathbf{A}\mathbf{A}^T$ gives $\mathbf{A}\mathbf{B}\mathbf{A}^T = \tilde{\mathbf{A}}\tilde{\mathbf{B}}_\tau\tilde{\mathbf{A}}^T = \mathbf{A}\mathbf{U}\tilde{\mathbf{B}}_\tau\mathbf{U}^T\mathbf{A}^T$, i.e., $\tilde{\mathbf{B}}_\tau = \mathbf{U}^T\mathbf{B}_\tau\mathbf{U}$. ■

Next, we show how the assumption of non-Gaussian distributions leads to identifiability in the basic case of $n = m$. A typical example in which the distributional information guarantees the model identifiability is ICA. In ICA, one has a set of observable signals, which are assumed to be linear mixtures of some hidden independent sources, but the mixing procedure is unknown. In the square case (with an equal number of sources and

observable signals), if at most one of the sources is non-Gaussian, then ICA is able to estimate the mixing procedure and recover the sources up to some trivial indeterminacies. Similarly, here considering the non-Gaussianity of the process noise allows the NG-SSM model with $n = m$ to be fully identifiable, as given by the following theorem.

Theorem 2 *Consider the model given by (1∼2) with $n = m$ and given L. Suppose that the assumptions A1∼A5 hold. Then the model is identifiable. In particular, suppose that the model with parameters $(\mathbf{A}, \{\mathbf{B}_\tau\}_{\tau=1}^L, \mathbf{\Sigma_e})$ and that with $(\tilde{\mathbf{A}}, \{\tilde{\mathbf{B}}_\tau\}_{\tau=1}^L, \tilde{\mathbf{\Sigma}}_{\tilde{\mathbf{e}}})$ are observationally equivalent; we then have $\tilde{\mathbf{\Sigma}}_{\tilde{\mathbf{e}}} = \mathbf{\Sigma_e}$, $\tilde{\mathbf{A}} = \mathbf{AP}$, $\tilde{\mathbf{B}}_\tau = \mathbf{P}^T\mathbf{B}_\tau\mathbf{P}$, and $\tilde{\mathbf{y}}_t = \mathbf{P}^T\mathbf{y}_t$, where $\mathbf{P}$ is a signed permutation matrix (a permutation matrix with non-zero entries being 1 or -1). The distribution of ϵ_{it} can also be uniquely determined up to the sign indeterminacy.*[3]

Proof Lemma 1 gives the relationships between the two parameter sets $(\mathbf{A}, \{\mathbf{B}_\tau\}_{\tau=1}^L)$ and $(\tilde{\mathbf{A}}, \{\tilde{\mathbf{B}}_\tau\}_{\tau=1}^L)$ which produce the same $\mathbf{x}_t$, under the assmptions A1∼A4. Based on Lemma 1, we have

$$
\begin{aligned}
&\mathbf{x}_t = \mathbf{A}\Big[\big(\mathbf{I} - \sum_{\tau=1}^L \mathbf{B}_\tau z^{-\tau}\big)^{-1}\Big]\boldsymbol{\epsilon}_t + \mathbf{e}_t = \tilde{\mathbf{A}}\Big[\big(\mathbf{I} - \sum_{\tau=1}^L \tilde{\mathbf{B}}_\tau z^{-\tau}\big)^{-1}\Big]\tilde{\boldsymbol{\epsilon}}_t + \tilde{\mathbf{e}}_t \\
\Longrightarrow &\mathbf{A}\big(\mathbf{I} - \sum_{\tau=1}^L \mathbf{B}_\tau z^{-\tau}\big)^{-1}\boldsymbol{\epsilon}_t + \mathbf{e}_t = \mathbf{AU}\Big[\big(\mathbf{I} - \sum_{\tau=1}^L \mathbf{U}^T\mathbf{B}_\tau\mathbf{U} z^{-\tau}\big)^{-1}\Big]\tilde{\boldsymbol{\epsilon}}_t + \tilde{\mathbf{e}}_t \\
\Longrightarrow &\Big[\big(\mathbf{I} - \sum_{\tau=1}^L \mathbf{B}_\tau z^{-\tau}\big)^{-1}\Big]\boldsymbol{\epsilon}_t + \mathbf{A}^{-1}\mathbf{e}_t = \Big[\big(\mathbf{I} - \sum_{\tau=1}^L \mathbf{B}_\tau z^{-\tau}\big)^{-1}\Big]\mathbf{U}\tilde{\boldsymbol{\epsilon}}_t + \mathbf{A}^{-1}\tilde{\mathbf{e}}_t \\
\Longrightarrow &\Big[\big(\mathbf{I} - \sum_{\tau=1}^L \mathbf{B}_\tau z^{-\tau}\big)^{-1}\Big](\boldsymbol{\epsilon}_t - \mathbf{U}\tilde{\boldsymbol{\epsilon}}_t) = \mathbf{A}^{-1}(\tilde{\mathbf{e}}_t - \mathbf{e}_t) \\
\Longrightarrow &(\boldsymbol{\epsilon}_t - \mathbf{U}\tilde{\boldsymbol{\epsilon}}_t) = \Big[\mathbf{I} - \sum_{\tau=1}^L \mathbf{B}_\tau z^{-\tau}\Big]\big(\mathbf{A}^{-1}(\tilde{\mathbf{e}}_t - \mathbf{e}_t)\big). \qquad (7)
\end{aligned}
$$

Let $\mathbf{d}_t \triangleq \mathbf{A}^{-1}(\tilde{\mathbf{e}}_t - \mathbf{e}_t)$. The right-hand side (RHS) of (7) is a moving average (MA) process of $\mathbf{d}_t$, and its autocovariance function at lag L is $\mathbf{R}_{RHS}(L) = \mathbb{E}\big[-\sum_{j=0}^L \mathbf{B}_j\mathbf{d}_{t-j} \cdot (-\sum_{i=0}^L \mathbf{B}_i\mathbf{d}_{t+L-j})^T\big] = \mathbb{E}(\mathbf{d}_t\mathbf{d}_t^T)\mathbf{B}_L^T$, where we have defined $\mathbf{B}_0 \triangleq -\mathbf{I}$. On the other hand, the left-hand side of (7) is i.i.d., and hence $\mathbb{E}(\mathbf{d}_t\mathbf{d}_t^T)\mathbf{B}_L^T = \mathbf{0}$. As $\mathbf{B}_L$ is nonsingular, we must have $\mathbb{E}(\mathbf{d}_t\mathbf{d}_t^T) = \mathbf{0}$, i.e., $\tilde{\mathbf{e}}_t = \mathbf{e}_t$ and $\boldsymbol{\epsilon}_t = \mathbf{U}\tilde{\boldsymbol{\epsilon}}_t$.

We then consider the condition $\boldsymbol{\epsilon}_t = \mathbf{U}\tilde{\boldsymbol{\epsilon}}_t$, or $\tilde{\boldsymbol{\epsilon}}_t = \mathbf{U}^T\boldsymbol{\epsilon}_t$. Assumption A5 states that at most one of ϵ_{it} is Gaussian. From the identifiability of the noiseless ICA model with a square mixing matrix (see Theorem 11 of Comon (1994) or Theorem 10.3.1 of Kagan et al. (1973)), $\mathbf{U}$ must be a signed permutation matrix. Furthermore, the distributions of $\tilde{\boldsymbol{\epsilon}}_{it}$ are the same as those of $\boldsymbol{\epsilon}_{it}$ (up to the permutation and sign indeterminacies). ■

3. If $\mathbb{V}\mathrm{ar}(\epsilon_{it})$ can be arbitrary, we have $\tilde{\mathbf{A}} = \mathbf{AP\Lambda}$ and $\tilde{\mathbf{B}}_\tau = \mathbf{\Lambda}^{-1}\mathbf{P}^T\mathbf{BP\Lambda}$, where $\mathbf{\Lambda}$ is a diagonal matrix with positive diagonal entries.

Finally, we consider the general case and show the identifiability of the NG-SSM model. The identifiability in the case with $n \geq m$ follows as a consequence of the above theorem and linear algebra.

Theorem 3 *Under the assumptions A1∼A5, the model given by (1∼2) with $n \geq m$ and given L is identifiable up to arbitrary permutations and signs of y_{it}.*

Proof Let $\mathbf{A}_i^T$ be the ith row of $\mathbf{A}$. Recall that $\mathbf{A}$ is of full column rank. Then for any i, by making use of linear algebra, one can show that there always exist $(m-1)$ rows of $\mathbf{A}$ such that they, together with $\mathbf{A}_i^T$, form a full-rank matrix, denoted by $\bar{\mathbf{A}}_i$. According to Theorem 2, from the observed data corresponding to $\bar{\mathbf{A}}_i$, $\{\mathbf{B}_\tau\}_{\tau=1}^L$ and $\bar{\mathbf{A}}_i$ can be uniquely determined up to the permutation and sign indeterminacies. That is, all rows of $\mathbf{A}$ are identifiable, and thus $\mathbf{A}$ is identifiable. Furthermore, $\mathbb{C}\mathrm{ov}(\mathbf{A}\mathbf{y}_t)$ is determined by $\mathbf{A}$ and $\{\mathbf{B}_\tau\}_{\tau=1}^L$. As $\mathbb{C}\mathrm{ov}(\mathbf{x}_t) = \mathbb{C}\mathrm{ov}(\mathbf{A}\mathbf{y}_t) + \mathbf{\Sigma_e}$, $\mathbf{\Sigma_e}$ is also identifiable. ■

3. Identification

In the Gaussian case, one can efficiently find an estimate of all parameters in the SSM model from an infinite number of possible solutions. By considering $\mathbf{y}_t$ as missing values, one can find the complete data likelihood $\log p(\{\mathbf{y}_t\}_{t=1}^n, \{\mathbf{x}_t\}_{t=1}^n)$ of the model (1∼2) in closed-form. The EM algorithm to maximize the data likelihood can then be derived, allowing partial identification of the model; see, for instance, Ghahramani and Hinton (1996). Alternatively, by rewriting the SSM model as some large block matrix formula, one can adopt the subspace state space system identification (4SID) methods (van Overschee and de Moor, 1991). For a comparison between these two types of methods, see Smith et al. (2000).

However, when the process noise ϵ_{it} is non-Gaussian, one usually has to resort to simulation-based methods, such as particle filtering, to do the inference, and parameter estimation also becomes computationally difficult. For reasons of computational efficiency of the algorithm, especially for large-scale problems (e.g., for MEG data analysis), we propose an approximate but very efficient method which consists of two steps.

3.1 Step 1: Dimension reduction and noise suppression by recovering the subspace of the latent processes

In the case where $n = m$, this step is skipped and one directly performs the second step given in Subsection 3.2. Otherwise in the first step we reduce the noise effect and estimate the subspace of the latent processes $\mathbf{y}_t$.

In the proposed NG-SSM model, one can see that the subspace spanned by the latent processes y_{it} is temporally colored, while the observation error $\mathbf{e}_t$ is assumed to be temporally white. Based on these properties, in this step we extract the subspace of $\mathbf{y}_t$ from $\mathbf{x}_t$ by finding a m-dimensional *colored subspace*:

$$\check{\mathbf{x}}_t = \mathbf{W}_c \mathbf{x}_t \tag{8}$$

and making the remaining subspace temporally as white as possible. In this way the effect of the observation error $\mathbf{e}_t$ in the extracted subspace $\check{\mathbf{x}}_t$ is greatly reduced. Consequently

$\check{\mathbf{x}}_t$ and $\mathbf{y}_t$ approximately lie in the same subspace. In other words, $\check{\mathbf{x}}_t$ is expected to be a linear square transformation of $\mathbf{y}_t$, i.e.,

$$\check{\mathbf{x}}_t = \check{\mathbf{A}}\mathbf{y}_t. \tag{9}$$

How to find the transformation in (8) is discussed below, and how to find $\check{\mathbf{A}}$ and estimate $\mathbf{y}_t$ from $\check{\mathbf{x}}_t$ will be explained in Subsection 3.2.

At first glance, a natural way to find the transformation in (8) is to resort to so-called colored subspace analysis (CSA, Theis (2010)). We adopt the algorithm based on joint low-rank approximation (JLA) for CSA, due to its attractive theoretical properties (Theis, 2010). However, in practice we found that when the data are very noisy, its solution is highly sensitive to initialization conditions, and hence we need develop an approach which avoids local optima.

A closer inspection of (2), which specifies the generating process of the latent precesses $\mathbf{y}_t$, together with Assumption A2, shows that $\mathbf{y}_t$ corresponds to the subspace in $\mathbf{x}_t$ which can be predicted best from historical values, while the subspace which is mainly caused by the observation error can not. In fact, by exploiting the VAR model, the subspace of $\mathbf{y}_t$, or (8), can be estimated efficiently, and the solution is guaranteed to be global optimal. First, let us whiten the data. Assume the eigenvalue decomposition (EVD) of the covariance matrix of $\mathbf{x}_t$ is $\mathbb{C}\mathrm{ov}(\mathbf{x}_t) = \mathbf{E}_1\mathbf{D}_1\mathbf{E}_1^T$, where $\mathbf{D}_1$ is a diagonal matrix consisting of the eigenvalues and columns of $\mathbf{E}_1$ are corresponding eigenvectors. The whitened data are

$$\ddot{\mathbf{x}}_t = \mathbf{D}_1^{-1/2}\mathbf{E}_1^T\mathbf{x}_t.$$

Note that components of $\ddot{\mathbf{x}}_t$ are uncorrelated and have unit variance.

Second, one fits the VAR model with L lags on $\ddot{\mathbf{x}}_t$:

$$\ddot{\mathbf{x}}_t = \sum_{\tau=1}^{L} \mathbf{M}_\tau \ddot{\mathbf{x}}_{t-\tau} + \tilde{\boldsymbol{\varepsilon}}_t,$$

where $\mathbf{M}_\tau$ denote the coefficient matrices, and $\tilde{\boldsymbol{\varepsilon}}_t$ is the residual series. Parameters involved in the above equation can be estimated efficiently, with the solution given in closed-form. Note that the subspace in $\ddot{\mathbf{x}}_t$ which can be predicted well coincides with the subspace of $\check{\mathbf{x}}_t$ given in (8), and that it corresponds to that of $\tilde{\boldsymbol{\varepsilon}}_t$ which has small variance.

Let the EVD decomposition of the covariance matrix of $\tilde{\boldsymbol{\varepsilon}}_t$ be $\mathbb{C}\mathrm{ov}(\boldsymbol{\varepsilon}_t) = \mathbf{E}_2\mathbf{D}_2\mathbf{E}_2^T$. Let $\mathbf{P}$ be the matrix consisting of the eigenvectors corresponding to the m smallest eigenvalues. Third, one can see that $\mathbf{P}^T\tilde{\boldsymbol{\varepsilon}}_t$ gives the m-dimensional minor subspace of the error $\tilde{\varepsilon}_t$, and consequently, $\mathbf{P}^T\ddot{\mathbf{x}}_t$ corresponds to the subspace of $\mathbf{x}_t$ that can be predicted best. That is, the transformation in (8) is determined by

$$\mathbf{W}_c = \mathbf{P}^T\mathbf{D}_1^{-1/2}\mathbf{E}_1^T.$$

In our experiments, we use this solution for CSA. In our simulation studies we found that this scheme always leads to a good performance, i.e., the learned $\check{\mathbf{x}}_t$ provides a good estimate of the subspace of the latent processes.

3.2 Step 2: Multichannel blind deconvolution and post-processing

In the noiseless case, Haufe et al. (2010) discussed the relationship between the NG-SSM model and the multichannel blind deconvolution (MBD, Cichocki and Amari (2003)), or convolutive ICA problem. Here after dimension reduction in the first step, we also use MBD for parameter estimation. From (2) and (9) one can see that

$$\boldsymbol{\epsilon}_t = \Big[\mathbf{I} - \sum_{\tau=1}^{L} \mathbf{B}_\tau z^{-\tau}\Big]\mathbf{y}_t = \check{\mathbf{A}}^{-1}\check{\mathbf{x}}_t - \sum_{\tau=1}^{L} \mathbf{B}_\tau \check{\mathbf{A}}^{-1}\check{\mathbf{x}}_{t-\tau} = \sum_{k=0}^{L} \mathbf{W}_k \check{\mathbf{x}}_{t-k}, \tag{10}$$

where $\mathbf{W}_0 \triangleq \check{\mathbf{A}}^{-1}$, and $\mathbf{W}_\tau \triangleq -\mathbf{B}_\tau \check{\mathbf{A}}^{-1}$ for $\tau > 0$. Recall that $\boldsymbol{\epsilon}_t$ is assumed to be both spatially and temporally independent. Consequently, $\mathbf{W}_k$ in (10) can be estimated by using the MBD technique, which aims to make the output sequences spatially and temporally independent.

Under the assumption that at most one of ϵ_{it} is Gaussian, MBD can estimate $\mathbf{W}_k$ uniquely up to only the scaling, permutation, and time shift indeterminacies. Here, the permutation indeterminacy is trivial, and to tackle the scaling indeterminacy, we fix the variance of $\hat{\epsilon}_{it}$ to one. With proper initializations (say, with large values for $\mathbf{W}_0$), we can avoid the time shift indeterminacy. In our experiments we used the natural gradient-based algorithm (Cichocki and Amari, 2003) for MBD to determine $\mathbf{W}_k$.

Now suppose that we have obtained the estimate $\hat{\mathbf{W}}_k$. According to (10), one can see that the estimate of $\check{\mathbf{A}}$ and that of $\mathbf{B}_\tau$ can be constructed as

$$\hat{\check{\mathbf{A}}} = \hat{\mathbf{W}}_0^{-1}, \text{ and } \hat{\mathbf{B}}_\tau = -\hat{\mathbf{W}}_\tau \hat{\check{\mathbf{A}}} = -\hat{\mathbf{W}}_\tau \hat{\mathbf{W}}_0^{-1}.$$

Recalling that in (8) we used $\mathbf{W}_c$ to extract the colored subspace, i.e., $\check{\mathbf{x}}_t = \mathbf{W}_c \mathbf{x}_t = \check{\mathbf{A}}\mathbf{y}_t$, one can then contruct the estimate of $\mathbf{A}$ as

$$\hat{\mathbf{A}} = \mathbf{W}_c^\dagger \hat{\check{\mathbf{A}}} = \mathbf{W}_c^\dagger \hat{\mathbf{W}}_0^{-1},$$

where † denotes the pseudo-inverse. The proposed two-step method is denoted by CSA+MBD.

3.3 Significance assessment

We use the bootstapping-based method to assess the significance of each influence from $y_{i,t-\tau}$ ($\tau > 0$) to y_{jt} or the total effect from $\{y_{i,t-\tau}\}_{\tau=1}^{L}$ to y_{jt}, denoted by $S_{j\leftarrow i}$, following the pathway proposed in Sec. 6 of Hyvärinen et al. (2010).

4. Simulations

We used simulations to illustrate the performance of the proposed method CSA+MBD for parameter estimation of the NG-SSM model. The observed data $\mathbf{x}_t$ were generated according to (1∼2) with $n = 10$, $m = 4$ and the sample size $N = 1000$. ϵ_{1t} and ϵ_{2t} are i.i.d. super-Gaussian: they were generated by passing Gaussian samples through the power nonlinearity with the exponent 1.5 and keeping the original signs. ϵ_{3t} and ϵ_{4t} are uniformly distributed (and thus sub-Gaussian). The lag number of the VAR process (2) was set to $L = 2$. We used a heuristic way to enforce the stability of the process, by using

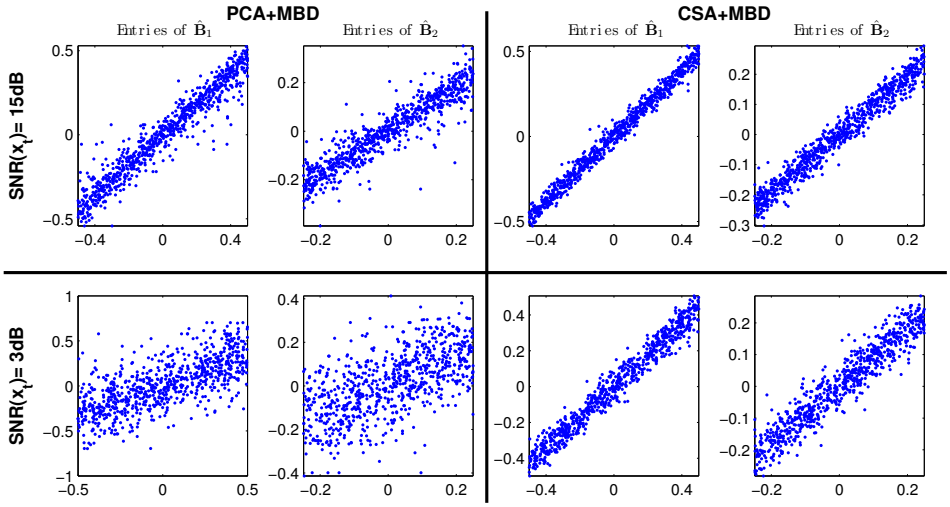

Figure 1: Scatterplots of the entries of $\hat{\mathbf{B}}_\tau$, $\tau = 1, 2$, vs. the true ones with different levels of $\mathbf{e}_t$ and different simulation methods. The average SNRs in $\mathbf{x}_t$ are 15dB (top) and 3dB (bottom). The two methods are PCA+MBD (left) and CSA+MBD (right).

small values for the entries of $\mathbf{B}_\tau$: the entries of $\mathbf{B}_1$ were uniformly distributed between -0.5 and 0.5, and those of $\mathbf{B}_2$ were drawn from the uniform distribution between -0.25 and 0.25. Entries of the mixing matrix $\mathbf{A}$ were drawn uniformly from $[-1.5, 1.5]$. The covariance matrix of $\mathbf{e}_t$ was constructed as the product of a $n \times n$ random matrix and its transpose. We use the average signal-to-noise ratio (SNR) in $\mathbf{x}_t$, which is defined as $10 \log_{10} \big(\sum_{i=1}^n \mathbb{V}\mathrm{ar}(x_{it} - e_{it}) / \sum_{i=1}^n \mathbb{V}\mathrm{ar}(e_{it}) \big)$, to measure the noise level in the observed data.

Principal component analysis (PCA) is a widely used approach for dimension reduction by finding the components with maximal variations. One can use PCA instead of CSA in Step 1 of the proposed method, leading to the method PCA+MBD, which was used in Haufe et al. (2010). We compare our method CSA+MBD with PCA+MBD. We considered two noise levels with the average SNRs in $\mathbf{x}_t$ being 15dB and 3dB, respectively. Each case was repeated for 50 random replications. Since good estimates of $\mathbf{A}$ often result in good estimates of the influence matrices $\mathbf{B}_\tau$, for conciseness of the presentation, here we only show how well $\mathbf{B}_\tau$ were recovered in Figure 1.

Figure 1 gives the scatterplots of the estimated entries of $\mathbf{B}_\tau$ and the true ones. Note that we already permuted $\hat{y}_{it}$ and adjusted their signs such that they and the true latent processes y_{it} have the same order and positive correlations (the scaling indeterminacy is avoided by enforcing Assumption A4). In addition, we report the SNRs of $\hat{\mathbf{B}}_\tau$ and $\hat{\mathbf{y}}_{it}$ w.r.t. the true ones in Table 1. For completeness of the comparison, we also show the results by ICA: note that here although the latent factors y_{it} are not mutually independent, one can still apply ICA to find the the components which are mutually *as independent as possible*; one can then fit the VAR model (2) on the estimated "independent" component to estimate

SNR($\mathbf{x}_t$)	Method	SNR($\hat{\mathbf{B}}_1$)	SNR($\hat{\mathbf{B}}_2$)	SNR($\hat{y}_{1t}$)	SNR($\hat{y}_{2t}$)	SNR($\hat{y}_{3t}$)	SNR($\hat{y}_{4t}$)
15dB	PCA+MBD	25.3	18.6	27.6 (11.6)	26.1 (10.6)	25.3 (9.4)	27.1 (10.5)
	FastICA	6.1	4.3	12.6 (4.4)	12.6 (4.5)	8.8 (4.3)	10.2 (5.5)
	CSA+MBD	**41.1**	**31.5**	**34.4** (7.9)	**33.1** (9.1)	**34.9** (10.4)	**32.7** (7.6)
3dB	PCA+MBD	5.4	2.2	13.2 (10.7)	12.6 (5.8)	10.6 (7.4)	8.2 (8.3)
	FastICA	5.4	2.8	10.4 (4.7)	10.2 (4.0)	7.6 (3.9)	7.2 (4.1)
	CSA+MBD	**25.5**	**17.7**	**20.4** (5.9)	**20.3** (4.5)	**20.4** (4.9)	**20.0** (4.4)

Table 1: SNR of the estimated $\mathbf{B}_\tau$ and the recovered latent processes y_{it} at different observation error levels and with different methods. Numbers in parentheses indicate standard deviations.

$\mathbf{B}_\tau$. We adopted the symmetric FastICA algorithm (Hyvärinen and Oja, 1997) with the tanh nonlinearity to perform ICA.

From Table 1 one can see that in both low-noise and high-noise situations, CSA+MBD successfully estimated $\mathbf{B}_\tau$ and the latent processes $\mathbf{y}_t$ with very high SNRs. Its performance is clearly better than PCA+MBD. This illustrates the validity of the proposed two-step method CSA+MBD, and also documents that properly initialized CSA performs well for dimension reduction and noise suppression (at least much better than PCA) in the estimation of the proposed NG-SSM. Not surprisingly, since the latent processes are not mutually independent, here ICA gives the poorest performance. Even in the lower noise situation (SNR = 15dB), the estimated $\mathbf{B}_\tau$ are very noisey (as seen from the low SNRs).

5. Real-world Applications

5.1 Financial data

We exploited the NG-SSM model to investigate some underlying structures of nine world major stock indices, including Dow Jones Industrial Average Index (DJI) and Composite Index (NAS) in USA, FTSE 100 Index (FTSE) in England, DAX Index in Germany, CAC 40 Index in France, Nikkei 225 (N225) in Japan, Hang Seng Index (HSI) in Hong Kong, Taiwan Weighted Index (TWI), and Shanghai Stock Exchange Composite Index (SSEC) in China. We used the daily dividend/split adjusted closing prices from Dec. 4, 2001 to Jul. 11, 2006. For the days when the price is not available, we used linear interpolation to estimate the price. Denoting the closing price of the ith index on day t by P_{it}, the corresponding return was calculated by $x_{it} = (P_{it} - P_{i,t-1})/P_{i,t-1}$. The data for analysis are the 9-dimensional parallel return series with 1200 samples.

We set the dimensionality of the latent process $\mathbf{y}_t$ to $m = 9$ and the number of lags to $L = 1$. After estimating all parameters, we further selected four estimated processes y_{it} of interest for further analysis. The corresponding columns of the estimated mixing matrix $\hat{\mathbf{A}}$ are given in Figure 2, with large numbers given in the figure. These latent processes are interesting to us for two reasons. First, they have large contributions to $\mathbf{x}_t$, as measured by the norm of the corresponding columns of $\hat{\mathbf{A}}$, and moreover, each of them contributes significantly to at least three indices, while others mainly contribute to one or two, so

they seem to be "common" factors of the indices. Second, compared to other estimated processes, they have stronger influences on each other, as measured by the corresponding entries of $\hat{\mathbf{B}}_1$. From Figure 2 one can see that $\hat{y}_{3t}$ and $\hat{y}_{4t}$ are closely related to the indices in America as well as those in Europe, while $\hat{y}_{1t}$ and $\hat{y}_{2t}$ contribute mainly to those in Europe and those in Asia, respectively. Figure 3 shows the entries of $\hat{\mathbf{B}}_t$ corresponding to these processes. One can see that the positive influences are mainly $\hat{y}_{4,t-1} \to \hat{y}_{1t}$, $\hat{y}_{3,t-1} \to \hat{y}_{2t}$, and $\hat{y}_{1,t-1} \to \hat{y}_{2t}$. These relationships seem meaningful and natural; in fact they are consistent with the direction of the information flow caused by the time difference between America, Europe, and Asia. It also suggests that the stock markets in Asia (except for that in China, as seen from the fact that SSEC is not strongly related to any of the four factors shown in Figure 2) are significantly influenced by those in America and Europe. However, this type of information will be ignored if one uses ICA to analyze their relationships, due to the assumption of independence between the latent factors.

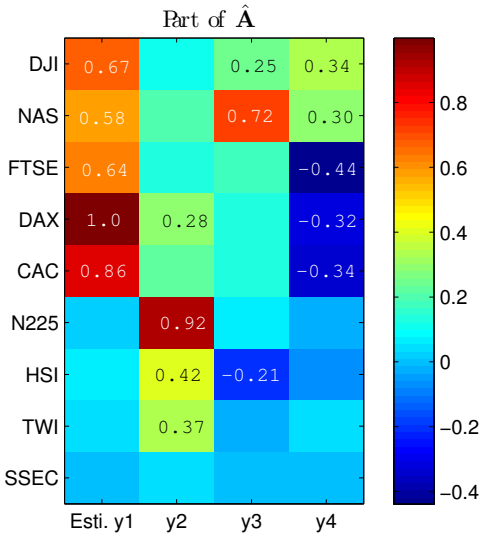

Figure 2: Part of $\hat{\mathbf{A}}$. $\hat{y}_{it}$ are sorted in descending order of their contributed variances to all x_{it}. The signs of $\hat{y}_{it}$ have been adjusted such that the sum of each column of $\hat{\mathbf{A}}$ is postive.

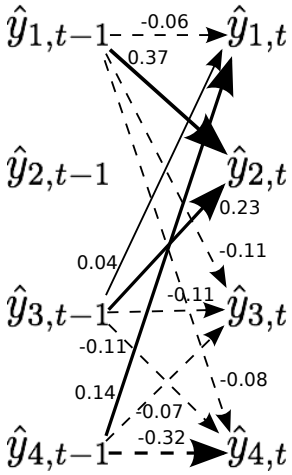

Figure 3: The entries of the influence matrix $\mathbf{B}_1$ corresponding to the four factors. Refer to Figure 2 for the relationships between $\hat{y}_{it}$ and the stock indices. Solid (dashed) lines show positive (negative) influences.

5.2 MEG data

Finally we applied NG-SSM on MEG data. The raw recordings consisted of the 204 gradiometer channels, and were obtained from a healthy volunteer, lasting about 8.2 minutes.

The data were down-sampled to 75 Hz from 600Hz.[4] As preprocessing, we used a bandpass filter to filter out the very low frequency part (lower than 4 Hz) and amplify the signals in the frequency bands 4-26Hz. These bands are believed to be informative, as Alpha, Beta, and Theta waves are mainly in this frequency range.

We used the proposed two-step method, CSA+MND, to extract the sources y_{it}. We empirically set the number of sources and number of lags to be $m = 25$ and $L = 15$ (corresponding to 0.2 second), respectively. Correspondingly in the first step the data dimensionality was reduced to 25. MBD was then applied on the extracted subspace, and finally 25 sources were obtained, together with $\{\mathbf{B}_\tau\}_{\tau=1}^{L}$ which imply the influences between them.

We found that the total effect $S_{j\leftarrow i}$ is significant at 1% level for at most 50% of the pairs $\{\mathbf{y}_{i,t-\tau}\} \to \mathbf{y}_{jt}$ with $i \neq j$. Therefore for better readability, we just give 11 sources which have the strongest effects (indicated by the smallest p-values) relative to each other. The topographic distributions of those sources, together with their influences, are given in Figure 4. Note the thicker the line, the stronger the effect. One can see that in many cases sources that have significant influences in between tend to be located near each other. Source #10 is of particular interest: it influences many other sources, but is hardly affected by others itself. For comparison, we also give the sources produced by FastICA (Hyvärinen and Oja, 1997) that have the strongest correlations to the sources given in Figure 5. One can see that sources found by NG-SSM usually have sharper or better-defined locations than those by FastICA, especially for sources #4, #5, #7, #11, and #10.

6. Discussion

We have proposed a general linear state-space model with the non-Gaussianity assumption of the process noise. The model takes into account the time-delayed interactions of the latent processes and the observation error (or measurement noise) in the observed data. Thanks to the non-Gaussianity assumption, we proved that this model, although very flexible, is in general identifiable, which enables simultaneously recovering the latent processes and estimating their delayed interactions (or roughly speaking, delayed causal relations). A computationally efficient method which consists of two separate steps has been given for system identification. Simulation results suggest good performance of the proposed method, and experimental results on real data illustrate the usefulness of the model.

References

K. S. Arun and S. Y. Kung. Balanced approximation of stochastic systems. *SIAM Journal on Matrix Analysis and Applications*, 11:42–68, 1990.

H. Attias. Independent factor analysis. *Neural Computation*, 11(4):803–851, 1999.

A. Belouchrani, K. Abed-Meraim, J. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Trans. on Signal Processing*, 45(2):434–444, 1997.

4. We are very grateful to Pavan Ramkumar and Riitta Hari of Aalto University, Finland, for access to the MEG data.

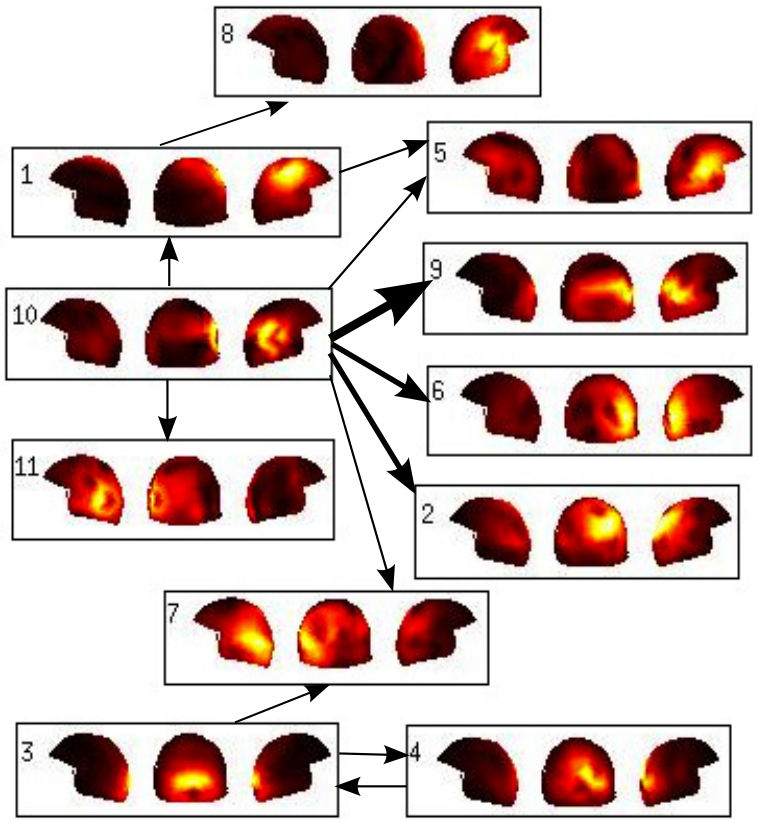

Figure 4: Some of the estimated sources for MEG signals and their relationships. The thicker the line, the stronger the influence.

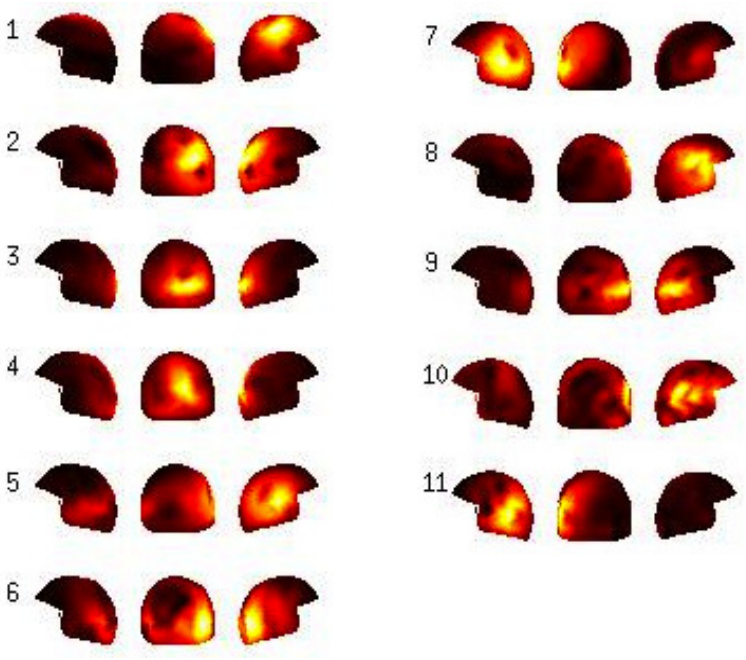

Figure 5: The sources estimated by FastICA which are most correlated with the sources in Figure 4.

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

P. Comon. Independent component analysis – a new concept? *Signal Processing*, 36: 287–314, 1994.

M. Davies. Identifiability issues in noisy ica. *IEEE Signal Processing Letters*, 11:470–473, 2004.

M. Ding, Y. Chen, and S. L. Bressler. Granger causality: Basic theory and application to neuroscience. In B. Schelter, M. Winterhalder, and J. Timmer, editors, *Handbook of Time Series Analysis*, pages 437–460, Wienheim, 1996. Wiley.

Z. Ghahramani and G. E. Hinton. Parameter estimation for LDS. Technical report, Department of Computer Science, University of Toronto, Toronto, Canada, 1996. U. Toronto Tech. Report.

G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J. L. Cantero. Measuring directional coupling between eeg sources. *NeuroImage*, 43:497–508, 2008.

C. Granger. Some recent developments in a concept of causality. *Journal of Economietrics*, 39:199–211, 1988.

S. Haufe, R. Tomioka, G. Nolte, K. R. Müller, and M. Kawanabe. Modeling sparse connectivity between underlying brain sources for EEG/MEG. *IEEE Trans Biomed Eng*, (8): 1954 – 1963, 2010.

A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

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

A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11: 1709–1731, 2010.

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

R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41:1–24, 2001.

G. A. Smith, A. J. Robinson, and M. Niranjan. A comparison between the em and subspace algorithms for the time-invariant linear dynamical system. Technical Report Tech. rep. CUED/F-INFENG/TR.366, Department of Computer Science, University of Toronto, Engineering Dept., Cambridge Univ., UK, 2000.

F. J. Theis. Colored subspace analysis: Dimension reduction based on a signals autocorrelation structure. *IEEE Transactions on Circuits and Systems – I: Regular Papers*, 57: 1463 – 1474, 2010.

P. van Overschee and B. de Moor. Subspace algorithms for the stochastic identification problem. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pages 1321–1326, Brighton, UK, 1991.

P. van Overschee and B. de Moor. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, Dordrecht, Netherlands, 1996.

L. Xu. Temporal factor analysis (TFA): stable-identifiable family, orthogonal flow learning, and automated model selection. In *Proceedings of the 2002 International Joint Conference on Neural Networks*, pages 472–476, Honolulu, HI , USA, 2002.