

# Temporal Principal Component Analysis

## — Advances in Dual Auto-regressive Modeling for Blind Gaussian Process Identification

Yiu-ming Cheung

Department of Computer Science  
 Hong Kong Baptist University  
 Hong Kong, SAR, China  
 E-mail: ymc@comp.hkbu.edu.hk

*Abstract*— The recent paper (Cheung 2001) has studied the blind identification of Gaussian source process through a general temporal independent component analysis (ICA) approach named *dual auto-regressive modelling*. It is actually a temporal extension of the classical principal component analysis without considering the principal order of the components. In this paper, we will further show the identifiable condition of the general temporal PCA (TPCA), and analyze the solution property of a specific TPCA algorithm presented in (Cheung 2001). Also, a new component ordering method is suggested, which includes the classical PCA ordering as a special case.

*Keywords*— Temporal Independent Component Analysis, Dual Auto-Regressive Modelling, Temporal Principal Component Analysis, PCA Ordering.

### I. INTRODUCTION

As a typical statistical analysis tool, principal component analysis (PCA) has been widely used in a variety of application areas such as image processing, pattern recognition, data mining and time series analysis. Given a series of multivariate Gaussian-distributed observations  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$  with  $\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(k)}]^T$ , PCA considers the following model:

$$\mathbf{x}_t = \mathbf{A}\mathbf{y}_t, \quad 1 \leq t \leq N, \quad (1)$$

with  $\mathbf{y}_t = [y_t^{(1)}, y_t^{(2)}, \dots, y_t^{(j)}, \dots, y_t^{(k)}]^T$ , where the  $k$  components  $y_t^{(j)}$ 's (also called *sources* hereafter) are statistically independent each other, and are unknown as well as the true mixing matrix  $\mathbf{A}$ . PCA is just to use second-order statistics information of the observations to find out an appropriate de-mixing matrix  $\mathbf{W}$  such that the  $\mathbf{y}_t$ 's estimator:

$$\hat{\mathbf{y}}_t = \mathbf{W}\mathbf{x}_t, \quad 1 \leq t \leq N \quad (2)$$

with

$$\hat{\mathbf{y}}_t = [\hat{y}_t^{(1)}, \hat{y}_t^{(2)}, \dots, \hat{y}_t^{(k)}]^T$$

is also component-wise independent, where the principal order of  $\hat{y}_t^{(j)}$ 's is explicitly given out based on the eigenvalues of the observation covariance matrix. However, PCA implies a strong assumption that the observations are independently and identically distributed (iid); otherwise, this technique will lead  $\hat{y}_t^{(1)}, \hat{y}_t^{(2)}, \dots, \hat{y}_t^{(k)}$  not to be independent each other any more.

In the literature, temporal factor analysis (TFA) [9] as a generalized model of Eq.(1) has considered the time correlations among observations in extracting the Gaussian components, but without assigning the components' principal order. The paper [9] has presented an algorithm by minimizing a Kullback-Leibler divergence function with a first-order Taylor expansion approximation for temporal decoupling. Although that algorithm has been further improved in [4], [5] and its performance well demonstrated by the experiments, their solution properties thus far have not been investigated yet.

Alternatively, our recent paper [3] has studied the blind identification of Gaussian source process through a general temporal independent component analysis (ICA) approach named *dual auto-regressive (AR) modelling*. Without considering the principal order of the components, it actually performs principal component analysis on time-correlated observations, rather than iid ones. We therefore call it temporal PCA (TPCA) hereafter to further distinguish it and the classical PCA. In this paper, we will give out the identifiable condition of a general TPCA model. Moreover, we further analyze the solution property of a specific TPCA algorithm in [3]. The theoretical results have shown that the wave form of each component  $\mathbf{y}_t^{(j)}$  can always be identified (i.e., the  $k$  components  $\mathbf{y}_t^{(1)}, \mathbf{y}_t^{(2)}, \dots, \mathbf{y}_t^{(k)}$  can be recovered up to a constant scale and any permutation of component superscript indices) as long as the underlying model is identifiable. In addition, we propose a new component ordering method upon the fact that ICA does not give out the principal order of those independent components unlike PCA. In the literature, some methods have been suggested to determine the component order. For example, the components are sorted according to their non-Gaussianity [7]. Back and Trappenberg [1] suggested to select a subset of the components based on the mutual information between the observations and the individual components, which also provides another way to order the components. Furthermore, the paper [2] decides the component order according to the  $L_\infty$  norm of each individual component, and the paper [6] orders the components according to their joint contributions in data reconstruction. All of these existing methods perform ordering assignment after all the components have been extracted via an ICA approach. In contrast, this

new ordering method gives out the principal order of the components in advance as given a set of observations. Consequently, it can save considerable computing costs if only first several principal components need to be extracted in data analysis. Besides that, the proposed method can bring at least two extra advantages:

1. It is a natural extension of PCA ordering with including the latter as a special case.
2. If the time series of each source is stationary and all eigenvalues of the sample observation covariance matrix are distinct, the intrinsic indeterminacy of ICA on component scale and subscript indices can both be fixed in advance up to a constant sign.

The organization of this paper is as follows. Section II conducts the identifiable analysis of TPCA, and Section III analyzes the solution properties of the TPCA algorithm in [3]. In Section IV, the new component ordering for TPCA is presented and demonstrated by the experiment. Finally, we draw a conclusion in Section V.

## II. IDENTIFIABLE ANALYSIS OF TEMPORAL PCA COMPONENTS

Suppose each component in the model of Eq.(1) is generally non-Gaussian iid distributed with at most one Gaussian. It has been shown [8] that the wave form of those components in Eq.(1) are identifiable (i.e., the components are identified up to a constant scale and any permutation of indices) under the condition that the components are independent each other. Unfortunately, when some components are iid Gaussian distributed, it no longer holds upon the fact that the independence property among a set of iid Gaussian variables is invariant in a rotation transformation (i.e., multiply an orthogonal matrix). However, this is not true in general when each component is a Gaussian process with  $y_t^{(j)}$  and  $y_{t-\tau}^{(j)}$  time-correlated for  $\tau = 1, 2, \dots$ . In the paper [3], while observations are modelled by Eq.(1), each of  $k$  independent components  $y_t^{(1)}, y_t^{(2)}, \dots, y_t^{(k)}$  in Eq.(1) has been further described as a general AR process:

$$y_t^{(j)} = f_j(Y_{t-1}^{(j)}|\theta_j) + \varepsilon_t^{(j)}, \quad 1 \leq j \leq k, \quad (3)$$

where  $f_j(Y_{t-1}^{(j)}|\theta_j)$  with  $Y_{t-1}^{(j)} = \{y_{t-1}^{(j)}, y_{t-2}^{(j)}, \dots, y_0^{(j)}\}$  is a deterministic function of  $Y_{t-1}^{(j)}$ ,  $\theta_j$  denotes the unknown true parameter set in  $f_j$ , and  $\varepsilon_t^{(j)}$  is a Gaussian white noise. For simplicity, these  $k$  components can be further expressed in the matrix form:

$$\mathbf{y}_t = \mathbf{f}(\mathbf{Y}_{t-1}|\Theta) + \boldsymbol{\varepsilon}_t, \quad (4)$$

where  $\mathbf{f} = [f_1, f_2, \dots, f_k]^T$ ,  $\mathbf{Y}_{t-1} = [\mathbf{y}_{t-1}^T, \mathbf{y}_{t-2}^T, \dots, \mathbf{y}_0^T]^T$ ,  $\Theta = \{\theta_1, \theta_2, \dots, \theta_k\}$ , and  $\boldsymbol{\varepsilon}_t = [\varepsilon_t^{(1)}, \varepsilon_t^{(2)}, \dots, \varepsilon_t^{(k)}]^T$ . We name the model described by Eq.(1) and Eq.(4) as temporal PCA. In the remaining part of this section and Section III, we let the covariance  $\Sigma$  of  $\boldsymbol{\varepsilon}_t$  be the identity matrix  $\mathbf{I}$  without loss of generality due to the fact that the scale of each component in this model is unidentifiable.

Since the wave form of  $\mathbf{y}_t$ 's is recovered by Eq.(2), if the wave form is unidentifiable, there must exist a

matrix  $\mathbf{R} \neq \mathbf{P}\mathbf{D}$  with  $\mathbf{P}$  being a permutation matrix and  $\mathbf{D}$  a diagonal matrix such that  $\mathbf{W}\mathbf{A} = \mathbf{R}$ , but  $\mathbf{R}\mathbf{R}^T = \mathbf{I}$ . That is,  $\mathbf{R}$  must be an orthogonal matrix. We therefore just need to investigate the  $\mathbf{y}_t$ 's wave-form identification under the situation that  $\mathbf{R}$  is orthogonal. Consequently, we have the following results:

*Theorem 1:* In temporal PCA model, for any orthogonal matrix  $\mathbf{R} \neq \mathbf{P}\mathbf{D}$ , where  $\mathbf{P}$  and  $\mathbf{D}$  are a permutation matrix and a diagonal matrix respectively, if  $\mathbf{R}\mathbf{f}(\mathbf{Y}_{t-1}|\Theta) \neq \mathbf{f}(\mathbf{R}\mathbf{Y}_{t-1}|\tilde{\Theta})$ , where  $\Theta$  and  $\tilde{\Theta}$  are both the parameters of  $\mathbf{f}$ , but with the different values in general, then the wave form of  $\mathbf{y}_t$ 's is identifiable.

The proof is given in Appendix A. In particular, when a series of  $\mathbf{y}_t$  is an AR(1) process, described as:

$$\mathbf{y}_t = \boldsymbol{\Lambda}\mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad (5)$$

where  $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$  is the diagonal matrix with  $\lambda_j$ 's as the diagonal elements, the above theorem can be further refined as follows:

*Corollary 1:* If  $\lambda_j$ 's are distinct each other, the wave forms of  $\mathbf{y}_t$ 's are identifiable.

Since from Eq.(5), given any  $k \times k$  orthogonal matrix  $\mathbf{R}$ , we know that:

$$\mathbf{R}\mathbf{f}(\mathbf{Y}_{t-1}|\Theta) = \mathbf{R}\boldsymbol{\Lambda}\mathbf{y}_{t-1} \quad (6)$$

$$= \mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^T\mathbf{R}\mathbf{y}_{t-1}, \quad (7)$$

$$\mathbf{f}(\mathbf{R}\mathbf{Y}_{t-1}|\tilde{\Theta}) = \tilde{\boldsymbol{\Lambda}}\mathbf{R}\mathbf{y}_{t-1}, \quad (8)$$

where  $\tilde{\boldsymbol{\Lambda}}$  is a diagonal matrix. Hence, to prove that  $\mathbf{R}\mathbf{f}(\mathbf{Y}_{t-1}|\Theta) \neq \mathbf{f}(\mathbf{R}\mathbf{Y}_{t-1}|\tilde{\Theta})$ , we just need to prove the following theorem:

*Theorem 2:* For any  $k \times k$  orthogonal matrix  $\mathbf{R} \neq \mathbf{P}\mathbf{D}$ , where  $\mathbf{P}$  and  $\mathbf{D}$  are a permutation matrix, and a diagonal matrix respectively, if  $\lambda_j$ 's are distinct each other,  $\mathbf{R}\boldsymbol{\Lambda}\mathbf{R}^T$  must not be a diagonal matrix.

The proof of Theorem 2 is given out in Appendix B. Although we here just give out the result for AR(1) process, the AR(p) component process can actually be transformed as an AR(1) process to study. The results are similar to Corollary 1. We leave it elsewhere.

## III. ANALYSIS OF THE TEMPORAL PCA ALGORITHM

Given a series of observations  $\mathbf{x}_t$ 's, the paper [3] has presented a general maximum likelihood (ML) algorithm to estimate  $\hat{\mathbf{y}}_t$ 's as well as the parameter set  $\Theta$  by maximizing the log-likelihood function of the observations:

$$\begin{aligned} Q(\Theta_1) &= \ln p(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N; \Theta_1) \\ &= \sum_{t=1}^N \ln p(\mathbf{x}_t | \mathbf{X}_{t-1}; \Theta_1), \end{aligned} \quad (9)$$

where  $\mathbf{X}_0 = \mathbf{x}_0$ , and  $\Theta_1 = \{\Theta, \mathbf{A}\}$ . In the following, we will use the same notations for the true parameters and their estimates for simplicity, which can be distinct from the context without ambiguity. Particularly, when  $\mathbf{y}_t$ 's are described by Eq.(5), the paper [3] has shown that  $p(\mathbf{x}_t | \mathbf{X}_{t-1}; \Theta_1)$  is explicitly given out as

$$p(\mathbf{x}_t | \mathbf{X}_{t-1}; \Theta_1) = G(\mathbf{x}_t | \mathbf{A}\mathbf{A}\mathbf{A}^{-1}\mathbf{x}_{t-1}, \mathbf{A}\mathbf{A}^T), \quad (10)$$

where  $G(\cdot)$  denotes a Gaussian probability density function, and  $\mathbf{A}^{-1}$  should be replaced by its pseudo inverse if  $\mathbf{A}$  is either a non-square matrix or singular. The detailed adaptive implementation at time step  $t$  is therefore as follow:

1. Given  $\Theta_1^{\text{old}}$ , let  $\hat{\mathbf{y}}_t = \mathbf{W}\mathbf{x}_t$ , where  $\mathbf{W}$  is the inverse of  $\mathbf{A}$ ;
2. Update  $\Theta_1$  by

$$\Theta_1^{\text{new}} = \Theta_1^{\text{old}} + \eta \frac{\partial J_t(\Theta_1)}{\partial \Theta_1} \Big|_{\Theta_1^{\text{old}}} \quad (11)$$

with  $\frac{\partial J_t(\Theta_1)}{\partial \Theta_1}$  being

$$\begin{aligned} \frac{\partial J_t(\Theta_1)}{\partial \mathbf{A}} &= \mathbf{W}^T (\mathbf{W}\mathbf{z}_t \hat{\mathbf{y}}_{t-1}^T \mathbf{\Lambda} - \mathbf{\Lambda} \mathbf{W}\mathbf{z}_t \hat{\mathbf{y}}_{t-1}^T \\ &\quad + \mathbf{W}\mathbf{z}_t \mathbf{z}_t^T \mathbf{W}^T - \mathbf{I}), \\ \frac{\partial J_t(\Theta_1)}{\partial \mathbf{\Lambda}} &= \text{diag}(\mathbf{W}\mathbf{z}_t \hat{\mathbf{y}}_{t-1}^T), \end{aligned} \quad (12)$$

where  $\eta$  is a small positive learning rate,  $\mathbf{z}_t = \mathbf{x}_t - \mathbf{A}\mathbf{\Lambda}\hat{\mathbf{y}}_{t-1}$ , and  $\text{diag}(\mathbf{U})$  denotes a diagonal matrix with the major diagonal of  $\mathbf{U}$ .

It is clear that, as long as  $\eta$  is small enough, the parameters  $\mathbf{A}$  and  $\mathbf{\Lambda}$  learned by the above algorithm will converge to:

$$E\left[\frac{\partial J_t(\Theta_1)}{\partial \mathbf{A}}\right] = 0 \quad (13)$$

$$E\left[\frac{\partial J_t(\Theta_1)}{\partial \mathbf{\Lambda}}\right] = 0. \quad (14)$$

Then we have the following results:

*Theorem 3:* The parameters learned by Eq.(12) will guarantees to converge to a true solution, i.e., the wave form of  $\mathbf{y}_t$ 's can be identified, as long as the model of temporal PCA is identifiable.

The mathematical proof is be given in Appendix C.

#### IV. COMPONENT ORDERING IN TEMPORAL PCA

In Eq.(4), we know that the variance  $\Sigma$  of  $\boldsymbol{\varepsilon}_t$  is indeterminate in estimation, which therefore gives us a freedom to pre-assign it. We here let

$$\Sigma = \text{diag}(\lambda_{x,1}, \lambda_{x,2}, \dots, \lambda_{x,k}), \quad (15)$$

where  $\lambda_{x,j}$  is the  $j^{\text{th}}$  largest eigenvalue of the sample covariance matrix  $\Sigma_x$  of the observations. Suppose the covariance matrix of  $\mathbf{y}_0$  is  $\sigma^2 \mathbf{I}$ , where  $\sigma^2$  is a constant scalar. We then define the  $j^{\text{th}}$  component of  $\mathbf{y}_t$ 's is the  $j^{\text{th}}$  principal one. This ordering can be interpreted that the observations are decomposed into  $k$  independent components such that the first principal component is the one with the maximum variance, and the second one is with the second maximum variance, and so on. It can be seen that this ordering is actually a natural extension of the PCA ordering, and it degenerates to the latter when  $f_j$ 's are some constant functions.

Since the principal order of the components is completely determined by  $\Sigma_x$ , we need to estimate  $\Sigma_x$  adaptively when the observation  $\mathbf{x}_t$  is not available until the time step  $t$ . Consequently, we learn it together with the

other model parameters by the implementation procedures as described in Section III. That is, we replace Eq.(11) by

$$\begin{aligned} \bar{\mathbf{x}}^{\text{new}} &= (1 - \eta)\bar{\mathbf{x}}^{\text{old}} + \eta(\mathbf{x}_t - \bar{\mathbf{x}}^{\text{old}}), \\ \Sigma_x^{\text{new}} &= (1 - \eta\Sigma)\Sigma_x^{\text{old}} + \eta\Sigma(\mathbf{x}_t - \bar{\mathbf{x}}^{\text{new}})(\mathbf{x}_t - \bar{\mathbf{x}}^{\text{new}})^T, \\ \Theta_1^{\text{new}} &= \Theta_1^{\text{old}} + \eta \frac{\partial J_t(\Theta_1)}{\partial \Theta_1} \Big|_{\Theta_1^{\text{old}}} \end{aligned} \quad (16)$$

with Eq.(12) becoming

$$\begin{aligned} \frac{\partial J_t(\Theta_1)}{\partial \mathbf{A}} &= \mathbf{W}^T (\Sigma^{-1} \mathbf{W}\mathbf{z}_t \hat{\mathbf{y}}_{t-1}^T \mathbf{\Lambda} - \mathbf{\Lambda} \Sigma^{-1} \mathbf{W}\mathbf{z}_t \hat{\mathbf{y}}_{t-1}^T \\ &\quad + \Sigma^{-1} \mathbf{W}\mathbf{z}_t \mathbf{z}_t^T \mathbf{W}^T - \mathbf{I}), \\ \frac{\partial J_t(\Theta_1)}{\partial \mathbf{\Lambda}} &= \text{diag}[\Sigma^{-1} \mathbf{W}\mathbf{z}_t \hat{\mathbf{y}}_{t-1}^T], \end{aligned} \quad (17)$$

where  $\Sigma$  is given by Eq.(15),  $\bar{\mathbf{x}}$  is the sample mean of  $\mathbf{x}_t$ 's, and  $\eta_\Sigma$  is a small positive learning rate like  $\eta$ . In general, we can firstly initialize  $\bar{\mathbf{x}}$  at zero, and  $\Sigma$  at a random value with its intrinsic constraints satisfied. We then adaptively adjust  $\bar{\mathbf{x}}$  and  $\Sigma$  by Eq.(16). To make the covariance learned smoothly, by rule of thumb,  $\eta_\Sigma$  should be chosen much smaller than  $\eta$ , e.g.,  $\eta_\Sigma = 0.1\eta$ . As a result, it needs more data points for their learning. Alternatively, after scanning a small portion of observations, we can re-initialize them at the sample mean and covariance matrix of those past observations, respectively.

#### A. Simulation Results

As an example, we let the observations be generated by Eq.(1) with the true matrix

$$\mathbf{A} = \begin{pmatrix} 1.5 & 0.5 \\ 0.7 & 2.0 \end{pmatrix}, \quad (18)$$

and the source process described by Eq.(5) be

$$\mathbf{y}_t = \begin{pmatrix} 0.4 \\ -0.5 \end{pmatrix} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t, \quad (19)$$

where  $\boldsymbol{\varepsilon}_t = [\varepsilon_t^{(1)}, \varepsilon_t^{(2)}]^T$  is from zero-mean Gaussian distribution with the variance of  $\varepsilon_t^{(1)}$  being 0.1, and  $\varepsilon_t^{(2)}$  being 0.2. We initialized  $\mathbf{y}_0$  at zero, and set  $\eta = 0.001$  and  $\eta_\Sigma = 0.0001$ , respectively. After scanning 500,000 sample points, the learned parameter  $\mathbf{A}$  has been converged to

$$\mathbf{A} = \begin{pmatrix} 0.4239 & 0.5368 \\ 0.2064 & 2.1812 \end{pmatrix}, \quad (20)$$

and  $\mathbf{\Lambda}$  converged to

$$\mathbf{\Lambda} = \begin{pmatrix} 0.4006 & 0.0000 \\ 0.0000 & -0.4427 \end{pmatrix}. \quad (21)$$

Figure 1 presents a slide window to show the identification results, where it can be seen that the original independent components have been well identified. In this case, the algorithm gave out the variance of the first component 1.5955 and the second one 0.2262. That means the first one is the most principal as expected.

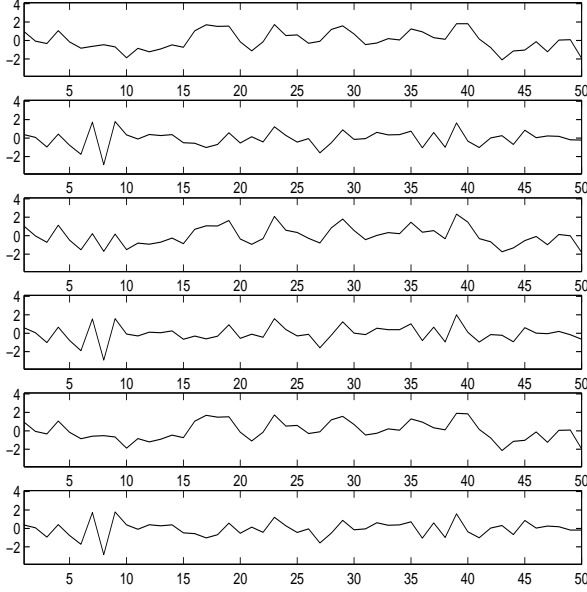


Fig. 1. The identification results. The first two rows are the slide windows of two sources, and the middle two rows are the corresponding observations. The last two are the identified results.

## V. CONCLUSION

This paper has not only given out the identifiable condition of temporal PCA, but also shown that the TPCA algorithm in [3] guarantees to converge a correct solution as long as the model is identifiable. Further, a new component ordering method is suggested, and demonstrated by the experiment.

### Acknowledgments

The work described in this paper was supported by a Faculty Research Grant of Hong Kong Baptist University (Project Code: FRG/01-02/II-24).

## REFERENCES

- [1] A.D. Back and T.P. Trappenberg, "Input Variable Selection Using Independent Component Analysis", *Proceedings of International Joint Conference on Neural Networks*, Vol. 2, pp. 989–992, 1999.
- [2] A.D. Back and A.S. Weigend, "A First Application of Independent Component Analysis to Extracting Structure From Stock Returns", *International Journal of Neural System*, Vol. 8, No. 4, pp. 473–484, 1997.
- [3] Y.M. Cheung, "Dual Auto-Regressive Modelling Approach To Gaussian Process Identification", *IEEE International Conference on Multimedia and Expo*, pp. 1256–1259, August 22–25, Tokyo, Japan, 2001.
- [4] Y.M. Cheung and L. Xu, "A Precise Approach of Temporal Factor Analysis", *Proceedings of International Conference on Neural Information Processing (ICONIP'2000)*, Vol. 2, pp. 1371–1376, Korea, 2000.
- [5] Y.M. Cheung and L. Xu, "Further Studies on Temporal Factor Analysis", *Proceedings of International Conference on Neural Information Processing (ICONIP'2000)*, Vol. 1, pp. 465–469, Korea, 2000.
- [6] Y.M. Cheung and L. Xu, "Independent Component Or-

dering in ICA Time Series Analysis", *Neurocomputing*, Vol. 41, pp. 145–152, 2001.

- [7] A. Hyvärinen, "Survey on Independent Component Analysis", *Neural Computing*, Surveys 2, pp. 94–128, 1999.
- [8] L. Tong, Y. Inouye and R.W. Liu, "Waveform-preserving Blind Estimation of Multiple Independent Sources", *IEEE Transactions on Signal Processing*, Vol. 41, No. 7, pp. 2461–2470, 1993.
- [9] L. Xu, "Temporal BYY Learning for State Space Approach, Hidden Markov Model and Blind Source Separation", *IEEE Transactions on Signal Processing*, Vol. 48, No. 7, pp. 2132–2144, July, 2000.

## APPENDIX

### I. PROOF OF THEOREM 1

We let

$$\tilde{\mathbf{y}}_t = \mathbf{R}\mathbf{y}_t, \quad (22)$$

with  $\mathbf{R} \neq \mathbf{PD}$ , where  $\mathbf{P}$  is a permutation matrix, and  $\mathbf{D}$  is a diagonal matrix. Suppose that  $\tilde{\mathbf{y}}_t$ 's also come from the same model as  $\mathbf{y}_t$ 's. Then, to guarantee the statistical independence among the components of  $\tilde{\mathbf{y}}_t$ , we must have:

$$\mathbf{x}_t = \tilde{\mathbf{A}}\tilde{\mathbf{y}}_t \quad (23)$$

$$\tilde{\mathbf{y}}_t = \mathbf{f}(\tilde{\mathbf{Y}}_{t-1}|\tilde{\Theta}) + \tilde{\mathbf{e}}_t \quad (24)$$

with the covariance matrix of  $\tilde{\mathbf{e}}_t$  is equal to  $\mathbf{I}$ . However, from Eq.(4), we have:

$$\begin{aligned} \tilde{\mathbf{y}}_t &= \mathbf{R}\mathbf{f}(\mathbf{Y}_{t-1}|\Theta) + \mathbf{R}\mathbf{e}_t \\ &= \mathbf{R}\mathbf{f}(\mathbf{Y}_{t-1}|\Theta) + \tilde{\mathbf{e}}_t. \end{aligned} \quad (25)$$

Since  $\mathbf{R}\mathbf{f}(\mathbf{Y}_{t-1}|\Theta) \neq \mathbf{f}(\mathbf{R}\mathbf{Y}_{t-1}|\tilde{\Theta})$ , we then have:

$$\tilde{\mathbf{y}}_t \neq \mathbf{f}(\tilde{\mathbf{Y}}_{t-1}|\tilde{\Theta}) + \tilde{\mathbf{e}}_t. \quad (26)$$

Hence, it is impossible that  $\tilde{\mathbf{y}}_t$ 's comes from the same model as  $\mathbf{y}_t$ 's with the different parameters. That is, the Gaussian rotation problem cannot occur. Hence, the wave form of  $\mathbf{y}_t$ 's is identifiable.

**Q.E.D.**

### II. PROOF OF THEOREM 2

Suppose  $\mathbf{R}\mathbf{A}\mathbf{R}^T = \tilde{\mathbf{D}}$ , where  $\tilde{\mathbf{D}}$  is a diagonal matrix. Since  $\mathbf{R}$  is an orthogonal matrix, then for each row and column, there must exist a non-zero element. Hence, we can let

$$\tilde{\mathbf{R}} = \tilde{\mathbf{P}}\mathbf{R} \quad (27)$$

where  $\tilde{\mathbf{P}}$  is a permutation matrix such that the main diagonal elements  $\tilde{r}_{ii}$ 's of  $\tilde{\mathbf{R}}$  are non-zero. Then we have:

$$\tilde{\mathbf{R}}\mathbf{A}\tilde{\mathbf{R}}^T = \tilde{\mathbf{P}}\mathbf{R}\mathbf{A}\mathbf{R}^T\tilde{\mathbf{P}}^T \quad (28)$$

$$= \tilde{\mathbf{P}}\tilde{\mathbf{D}}\tilde{\mathbf{P}}^T \quad (29)$$

$$= \tilde{\mathbf{A}}, \quad (30)$$

where  $\tilde{\mathbf{A}}$  is still a diagonal matrix. From Eq.(28), we have

$$\tilde{\mathbf{R}}\mathbf{A} = \tilde{\mathbf{A}}\tilde{\mathbf{R}}. \quad (31)$$

Since

$$\tilde{\mathbf{R}}\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 \tilde{r}_{11} & \lambda_2 \tilde{r}_{12} & \dots & \lambda_k \tilde{r}_{1k} \\ \lambda_1 \tilde{r}_{21} & \lambda_2 \tilde{r}_{22} & \dots & \lambda_k \tilde{r}_{2k} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_1 \tilde{r}_{k1} & \lambda_2 \tilde{r}_{k2} & \dots & \lambda_k \tilde{r}_{kk} \end{pmatrix} \quad (32)$$

and

$$\tilde{\mathbf{\Lambda}}\tilde{\mathbf{R}} = \begin{pmatrix} \tilde{\lambda}_1 \tilde{r}_{11} & \tilde{\lambda}_1 \tilde{r}_{12} & \dots & \tilde{\lambda}_1 \tilde{r}_{1k} \\ \tilde{\lambda}_2 \tilde{r}_{21} & \tilde{\lambda}_2 \tilde{r}_{22} & \dots & \tilde{\lambda}_2 \tilde{r}_{2k} \\ \vdots & \vdots & \dots & \vdots \\ \tilde{\lambda}_k \tilde{r}_{k1} & \tilde{\lambda}_k \tilde{r}_{k2} & \dots & \tilde{\lambda}_k \tilde{r}_{kk} \end{pmatrix}, \quad (33)$$

we then have

$$\begin{aligned} & \begin{pmatrix} \lambda_1 \tilde{r}_{11} & \lambda_2 \tilde{r}_{12} & \dots & \lambda_k \tilde{r}_{1k} \\ \lambda_1 \tilde{r}_{21} & \lambda_2 \tilde{r}_{22} & \dots & \lambda_k \tilde{r}_{2k} \\ \vdots & \vdots & \dots & \vdots \\ \lambda_1 \tilde{r}_{k1} & \lambda_2 \tilde{r}_{k2} & \dots & \lambda_k \tilde{r}_{kk} \end{pmatrix} \\ = & \begin{pmatrix} \tilde{\lambda}_1 \tilde{r}_{11} & \tilde{\lambda}_1 \tilde{r}_{12} & \dots & \tilde{\lambda}_1 \tilde{r}_{1k} \\ \tilde{\lambda}_2 \tilde{r}_{21} & \tilde{\lambda}_2 \tilde{r}_{22} & \dots & \tilde{\lambda}_2 \tilde{r}_{2k} \\ \vdots & \vdots & \dots & \vdots \\ \tilde{\lambda}_k \tilde{r}_{k1} & \tilde{\lambda}_k \tilde{r}_{k2} & \dots & \tilde{\lambda}_k \tilde{r}_{kk} \end{pmatrix}. \end{aligned} \quad (34)$$

Since  $\tilde{r}_{ii}$ 's are non-zero, we have

$$\mathbf{\Lambda} = \tilde{\mathbf{\Lambda}}. \quad (35)$$

By putting Eq.(35) into Eq.(34) and comparing the elements of both sides one-by-one, we have

$$(\lambda_i - \lambda_j)\tilde{r}_{ij} = 0, \quad \text{for } \forall 1 \leq i, j \leq k \quad \text{and} \quad i \neq j. \quad (36)$$

Since  $\lambda_i - \lambda_j \neq 0$  for  $\forall i \neq j$ , we then have  $\tilde{r}_{ij} = 0$  for any  $1 \leq i, j \leq k$  and  $i \neq j$ . That is,  $\tilde{\mathbf{R}}$  is a diagonal matrix, denoted as  $\mathbf{D}$ . Hence, from Eq.(27), we know that  $\mathbf{R} = \tilde{\mathbf{P}}\mathbf{D}$ , which is contradictory to the condition that  $\mathbf{R}$  is not a product of a permutation matrix and a diagonal matrix. Hence, Theorem 2 holds.

**Q.E.D.**

### III. PROOF OF THEOREM 3

Given the log-likelihood function of Gaussian-distributed observations as described by Eq.(10), we know that, as  $N \rightarrow \infty$ , the ML estimates of the true parameters  $\mathbf{A}^*$  and  $\mathbf{\Lambda}^*$  that make Eq.(13) and Eq.(14) hold must also satisfy

$$\mathbf{A}\mathbf{\Lambda}\mathbf{A}^{-1} = \mathbf{A}^*\mathbf{\Lambda}^*\mathbf{A}^{*-1} \quad (37)$$

$$\mathbf{A}\mathbf{A}^T = \mathbf{A}^*\mathbf{A}^{*T} \quad (38)$$

upon the fact that the ML estimates of the Gaussian distribution unbiasedly tend to their true value as  $N \rightarrow \infty$ . We let:

$$\mathbf{A} = \mathbf{A}^*\mathbf{R}, \quad (39)$$

where  $\mathbf{R}$  is a  $k \times k$  matrix. By putting Eq.(39) into Eq.(38), we can get

$$\mathbf{R}\mathbf{R}^T = \mathbf{I}. \quad (40)$$

That is,  $\mathbf{R}$  must be an orthogonal matrix. Hence, from Eq.(38), we have

$$\mathbf{A}^*\mathbf{R}\mathbf{A}\mathbf{R}^T\mathbf{A}^{*-1} = \mathbf{A}^*\mathbf{\Lambda}^*\mathbf{A}^{*-1}. \quad (41)$$

Consequently, we have

$$\mathbf{R}\mathbf{A}\mathbf{R}^T = \mathbf{\Lambda}^*. \quad (42)$$

From Theorem 2, we know that  $\mathbf{R}$  satisfying Eq.(42) must be an orthogonal matrix that is a product of a permutation matrix  $\mathbf{P}$  and a diagonal matrix  $\mathbf{D}$ . That is,  $\mathbf{A}$  is the estimate of  $\mathbf{A}^*\mathbf{P}\mathbf{D}$ . Hence,  $\mathbf{A}$  is a true solution. Further, we know that  $\mathbf{\Lambda} = \mathbf{R}\mathbf{\Lambda}^*\mathbf{R}^T$  is also a true solution.

**Q.E.D.**