Temporal BYY Encoding, Markovian State Spaces, and Space Dimension Determination

Lei Xu, *Fellow, IEEE*

***Abstract*—As a complementary to those temporal coding approaches of the current major stream, this paper aims at the Markovian state space temporal models from the perspective of the temporal Bayesian Ying-Yang (BYY) learning with both new insights and new results on not only the discrete state featured Hidden Markov model and extensions but also the continuous state featured linear state spaces and extensions, especially with a new learning mechanism that makes selection of the state number or the dimension of state space either automatically during adaptive learning or subsequently after learning via model selection criteria obtained from this mechanism. Experiments are demonstrated to show how the proposed approach works.**

***Index Terms*—Gated multitemporal models, harmony learning, hidden Markov model, linear state spaces, space dimension, state selection, temporal Bayesian Ying-Yang (BYY) system, temporal factor analysis.**

I. Introduction

SIMULTANEOUSLY building up a bottom-up pathway for encoding an observed pattern into a representation space and a top-down pathway for reconstructing a pattern from an inner representation has been widely adopted as a fundamental idea in various studies of brain theory and neural networks. Typical examples include Carpenter and Grossberg's adaptive resonance theory (ART) [8], Kawato's theory on cerebellum and motor control [15], and Hinton and colleagues' Helmholtz machines and wake-sleep learning [9], [11]. Moreover, the LMSER self-organizing rule proposed in 1991 [42] is also an effort that uses a bidirectional architecture for statistical unsupervised learning.

The basic sprit of the least mean-square-error reconstruction (LMSER) self-organizing has been further developed into the Bayesian Ying Yang (BYY) harmony learning [40], which is firstly proposed in 1995 and then systematically developed in past years. Readers are referred to [30] and [31] for a recent systematical introduction. The BYY harmony learning formulates the two pathway sprit in a general statistical framework. First, a so-called BYY system is proposed for coordinately modeling the two pathways via two complement Bayesian representations of the joint distribution on an observation space and representation space such that a number of existing major learning problems and learning methods are revisited as special cases from a unified perspective. Second, a harmony learning theory is developed with a new learning mechanism that makes model selection implemented either *automatically* during parameter learning or *subsequently after* parameter learning via a new class of model selection criteria obtained from this mechanism. Third, this BYY harmony learning has motivated three types of regularization, namely a data smoothing technique that provides a new solution on the hyper-parameter in a Tikinov-like regularization [23], a normalization with a new conscience de-learning mechanism that has a nature similar to the rival penalized competitive learning (RPCL) [30], [33], [41], and a structural regularization by imposing certain structural constraints via designing a specific forward structure in a BYY system. Specifically, the harmony learning on various specific BYY systems with typical structures lead to various specific learning algorithms as well as the detailed forms for implementing regularization and model selection. The details are referred to [30], [31], [33].

In recent years, a large volume of physiological and behavioral data has emerged in supporting a key role for temporal coding in the brain. Using time as an extra degree of freedom in neural representation, temporal coding takes a very important role in various information processing tasks, including scene segmentation, figure-ground separation, classification, learning, associative memory, inference, motor control, and communication. Though ever increasing emphases in the literature of neural network have been put on the topics of nonlinear dynamics, oscillatory and chaotic networks, spiking neurons, and pulse-coupled networks, the studies on temporal coding via Markovian state spaces can be traced back to 60s in the literature of control theory and 70s in the literature of signal processing and speech recognition. One typical example is the well known Kalman filter [14] that bases on a linear state space, which has been extensively studied and applied in the fields of control systems and signal processing over decades. Another typical example is the well known hidden Markov model (HMM) that bases a discrete state space, which has been also widely studied and used not only in the field of speech processing and recognition but also recently in many tasks of learning, data mining, and bio-informatics [5]. As a complementary to those temporal coding approaches of the current major stream, this paper aims at providing not only a unified framework for these Markovian state space based temporal models in help of temporal BYY learning [32], [34], but also several new advances on both the HMM and the linear state spaces.

The task of learning on the Markovian state spaces is usually made via maximum likelihood (ML) learning. The ML learning works well when the number of states or the dimension of

Manuscript received May 30, 2003; revised January 1, 2004. This work was supported by a Grant from the Research Grant Council of the Hong Kong SAR under Project CUHK 4184/03E).

The author is with the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong, P.R. China (e-mail: lxu@cse.cuhk.edu.hk).

Digital Object Identifier 10.1109/TNN.2004.833302

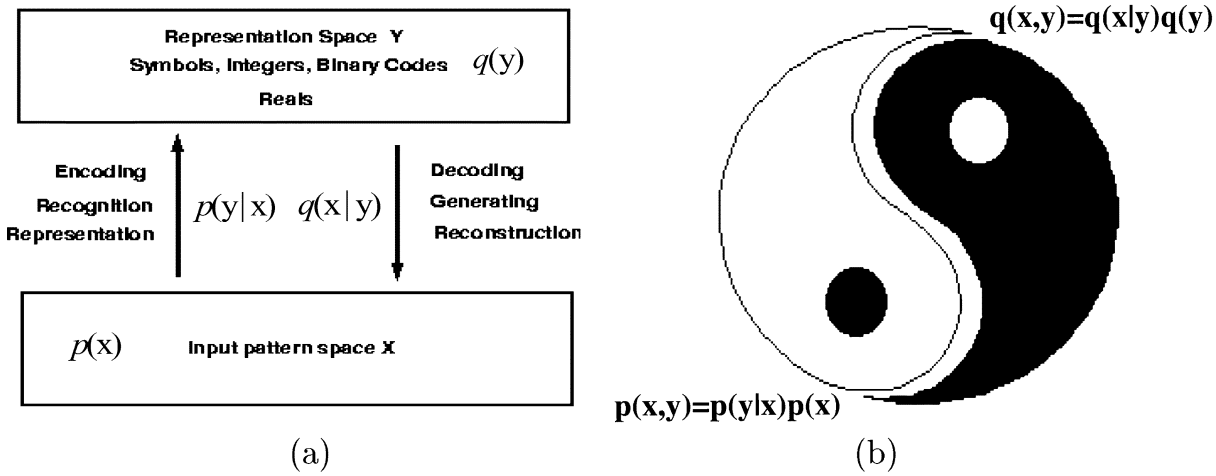


Fig. 1. BYY harmony learning.

the state space is known, which is however unknown in many practical problems. Efforts have been made toward this critical challenge, with a number of model selection criteria developed to evaluate a family of structures with different scales such that a best one is selected. Typical examples include the AIC criterion [3] as well as its extensions [4], the VC dimension based learning theory [24], Cross validation [22], Bayesian theory [16], [21], the minimum message length (MML) theory [25], and the minimum description length (MDL) theory [11], [18], [19]. However, only a rough solution is available from these existing model selection criteria in the cases of a small size of samples. Moreover, the criteria can only be used in a computational very expensive two-stage procedure that selects models after making ML learning on all the candidate models, but not applicable to making model selection during parameter learning. In contrast, the temporal BYY harmony learning provides a new learning mechanism that can make model selection and parameter learning under a same best harmony principle, and model selection is implemented either automatically during parameter learning under this harmony principle or via a new class of model selection criteria obtained from this mechanism after making parameter learning under the maximum likelihood principle.

In Section II, after introducing the fundamentals of BYY system and harmony learning, two types of temporal BYY systems are presented under a general framework for Markovian state space based temporal models. One is called the temporal BYY process system [32], [34], with its key points re-elaborated here. The other is called temporal BYY instantaneous system that is newly presented in this paper. Detailed algorithms and criteria have been provided not only on the discrete state featured HMM and extensions in Section III but also on the continuous state featured TFA, TNFA, and extensions in Section IV. Several experimental results are demonstrated in Section V before giving the concluding remarks in Section VI.

II. Temporal Bayesian Ying-Yang Systems and Harmony Learning

A. BYY System and Harmony Learning

As shown in Fig. 1, we consider the joint distribution of an observation $\mathbf{x}$ and its inner representation $\mathbf{y}$ in a learning system from two complement aspects. On one hand, each $\mathbf{x}$ is interpreted as generated via a backward path $q(\mathbf{x}|\mathbf{y})$ from an inner distribution $q(\mathbf{y})$ in a structure subject to certain learning tasks, i.e.,

$$q(\mathbf{x}) = \int q(\mathbf{x}|\mathbf{y})q(\mathbf{y})\mu(d\mathbf{y}) \tag{1}$$

where $\mu(\cdot)$ is a given measure. On the other hand, each $\mathbf{x}$ is interpreted as being mapped into an inner representation $\mathbf{y}$ via a forward path $p(\mathbf{y}|\mathbf{x})$

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\mu(d\mathbf{x}) \tag{2}$$

to match the target density $q(\mathbf{y})$. The two aspects reflect the two types of Bayesian decomposition of the joint density $q(\mathbf{x}|\mathbf{y})q(\mathbf{y}) = q(\mathbf{x},\mathbf{y}) = p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x})$. Without any constraint, the two should be conceptually identical. However, in a practical consideration, $p(\mathbf{x})$ is obtained from a set $\mathcal{X} = \{x_t\}_{t=1}^T$ of observed samples and other three components $p(\mathbf{y}|\mathbf{x})$, $q(\mathbf{x}|\mathbf{y})$, and $q(\mathbf{y})$ are also subject to certain structural constraints. Thus, we usually have two different but complementary Bayesian representations

$$p(\mathbf{x},\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \quad q(\mathbf{x},\mathbf{y}) = q(\mathbf{x}|\mathbf{y})q(\mathbf{y}) \tag{3}$$

which compliments to the famous Chinese ancient Ying-Yang philosophy with $p(\mathbf{x},\mathbf{y})$ called Yang machine that consists of the observation space (or called Yang space) by $p(\mathbf{x})$ and the forward pathway (or called Yang pathway) by $p(\mathbf{y}|\mathbf{x})$, and with $q(\mathbf{x},\mathbf{y})$ called Ying machine that consists of the invisible domain (or Ying space) by $q(\mathbf{y})$ and the Ying (or backward) pathway by $q(\mathbf{x}|\mathbf{y})$. Such a pair of Ying-Yang models is called BYY system.

Typically, $p(\mathbf{x})$ is obtained either from $\mathcal{X}$ directly or after a smoothing preprocessing featured by a smoothing parameter h. Thus, it can be denoted as $p(x|\mathcal{X})$ or $p(x|\mathcal{X},h)$. Our learning task is to specify all the aspects of $p(\mathbf{y}|\mathbf{x})$, $q(\mathbf{x}|\mathbf{y})$, and $q(\mathbf{y})$ as well as h (if any). Specifically, the task further consists of design of structures and learning of unknowns.

The structure of $q(\mathbf{y})$ describes the nature of the inner representation. A specific structure of $q(\mathbf{y})$ is specified in

three aspects. One is the representation format or the function form of $q(\mathbf{y})$, which features the key nature of the learning task that a specific BYY system performs. The other is a set $\mathbf{k}$ of integers called scale parameters. A collection of specific BYY systems with different specific values of $\mathbf{k}$ corresponds to a family of specific BYY systems (thus, specific learning models they perform) that share a same system configuration but in different scales. Another is a set θ_y of all unknown parameters in real numbers.

The implementing architecture of a BYY system is featured by a combination of the specific structures of $p(\mathbf{y}|\mathbf{x})$, $q(\mathbf{x}|\mathbf{y})$. There are three typical combinations as follows:

- A B-architecture that directly implements (1), with a structure-free $p(\mathbf{y}|\mathbf{x})$,
- A F-architecture that directly implements (2), with a structure-free $q(\mathbf{x}|\mathbf{y})$,
- A BI-architecture that bidirectionally implements both (1) and (2).

Each of three architectures can implement a learning task of the same nature, but from a different perspective and with a different performance. We say $p(u|v)$ is structural free if $p(u|v) \in \mathcal{P}^0_{u|v}$ with $\mathcal{P}^0_{u|v}$ consisting of all of functions in the format of $p(u|v)$ that satisfies $\int p(u|v)du = 1$, $p(u|v) \geq 0$. One of $p(\mathbf{y}|\mathbf{x})$, $q(\mathbf{x}|\mathbf{y})$ is designed as being structure-free means that there is no any priori constraint to impose on it and it will be specified during learning via other components in a BYY system. In contrast, a parametric $p(u|v) \in \mathcal{P}^S_{u|v}$ means that it comes from a family $\mathcal{P}^S_{u|v}$ with a prespecified structure based on certain priori requirements or knowledge and then a particular density is specified by a set of unknown parameters. Without losing generality, we use $\theta_{u|v}$ to denote not only this set of unknown parameters but also even a free $p(u|v)$ that can be regraded as a real set of infinite many of unknowns.

Specifically, the function forms and scale parameters of $p(\mathbf{y}|\mathbf{x})$, $q(\mathbf{x}|\mathbf{y})$ are selected according to the structure of $q(\mathbf{y})$. In other words, $\mathbf{k}$ of $q(\mathbf{y})$ actually represents the scales of a whole BYY system with the entire parameter set $\theta = \{\theta_y, \theta_{x|y}, \theta_{y|x}\}$. Therefore, the learning task consists of determining both θ and $\mathbf{k}$. The former is usually called parameter learning that searches a specific value of θ within a real domain Θ. The latter is usually called model selection that selects a specific value of $\mathbf{k}$ among a discrete domain $\mathbf{K}$ that corresponds a collection of all the candidate models.

The basic principle to implement both the two tasks is making the Ying machine and Yang machine be best harmony in a twofold sense as follows:

- difference between the two Bayesian representations in (3) should be minimized;
- resulting BYY system should be of the least complexity.

Mathematically, this principle can be implemented by [30]–[33], [40]

$$\max_{\theta,\mathbf{k}} H(\theta,\mathbf{k}), \quad H(\theta,\mathbf{k}) = H(p||q)$$
$$H(p||q) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \ln[q(\mathbf{x}|\mathbf{y})q(\mathbf{y})] \times \mu(d\mathbf{x})\mu(d\mathbf{y}) - Z_q,$$
$$Z_q = -\ln[\mu(\delta(\mathbf{x}))\mu(\delta(\mathbf{y}))]. \tag{4}$$

As discussed in [30] and [33], taking a regularization role for learning on a small size of samples, Z_q comes from the fact that the above harmony measure is derived from its original definition $H(p||q) = \sum_t p_t \ln q_t$ on discrete probability distributions. Specifically, $\mu(\delta(\mathbf{x}))$ and $\mu(\delta(\mathbf{y}))$ can be estimated in several specific choices. For examples, from $1 = \int q(\mathbf{x}|\mathbf{y})q(\mathbf{y})\mu(\delta(\mathbf{x}))\mu(\delta(\mathbf{y})) \approx \sum_{\mathbf{x}\in\mathcal{X}}\sum_{\mathbf{y}\in\mathcal{Y}} q(\mathbf{x}|\mathbf{y})q(\mathbf{y})\mu(\delta(\mathbf{x}))\mu(\delta(\mathbf{y}))$ we may get

$$\mu(\delta(\mathbf{y}))\mu(\delta(\mathbf{y})) = \frac{1}{\sum_{\mathbf{x}\in\mathcal{X}}\sum_{\mathbf{y}\in\mathcal{Y}} q(\mathbf{x}|\mathbf{y})q(\mathbf{y})} \tag{5}$$

where $\mathcal{Y}$ consists of a set of samples of $\mathbf{y}$ that is usually obtained in a correspondence to $\mathcal{X}$. Alternatively, we can even directly let $q(\mathbf{x}|\mathbf{y})$ to be replaced by $p(\mathbf{x})$ that is either given by (1) or obtained from the sample $\mathcal{X}$ [26].

One salient feature of $H(\theta,\mathbf{k})$ is that both parameter learning and model selection are implemented via maximizing this same criterion.

This feature is not shared by the conventional two-stage implementation of typical approaches in the existing literature of statistical learning. That is, parameter learning is made under maximum likelihood principle at Stage I on getting a best parameter set $\theta^*_{\mathbf{k}}$ for each candidate model that corresponds $\mathbf{k} \in \mathbf{K}$, where $\mathbf{K}$ consists of a given set of possible candidates. At Stage II, a best $\mathbf{k}^*$ is selected by $\min_{\mathbf{k}} J(\mathbf{k})$ and the value $J(\mathbf{k})$ at every candidate $\mathbf{k} \in \mathbf{K}$ has to be calculated basing on $\theta^*_{\mathbf{k}}$ that is resulted from parameter learning. Typical model selection criteria include MML/MDL [18], [19], [25], AIC [3] as well as extensions [4]. One serious problem of this two-stage implementation is that the computing cost is very expensive since parameter learning has to be made on every candidate $\mathbf{k}$ that should be enumerated among $\mathbf{K}$.

Instead of implementing the two stages with different criterion for each stage, model selection is not only made via the same criterion $H(\theta,\mathbf{k})$ that parameter learning is based on, but also implemented automatically during parameter learning. On the other hand, $H(\theta,\mathbf{k})$ in (4) can also be used in the conventional two stage way. At Stage I, we can make parameter learning by

$$\max_{\theta} H(\theta), \quad H(\theta) = H(\theta, \mathbf{k}) \tag{6}$$

with each $\mathbf{k} \in \mathbf{K}$ enumerated from small scales incrementally to large scales. With the obtained $\theta^*_{\mathbf{k}}$, model selection is implemented at Stage II via

$$\min_{\mathbf{k}} J(\mathbf{k}), \quad J(\mathbf{k}) = -H(\theta^*_{\mathbf{k}}, \mathbf{k}). \tag{7}$$

Moreover, when N is very small, this new model selection criterion can be further improved by the following generalized version

$$J_G(\mathbf{k}) = J(\mathbf{k}) + \frac{m_{eff} + m_{E_X}}{N} \tag{8}$$

which will be further discussed at the end of Section VI.

Furthermore, as an alternative of (6), Stage I can also be made with parameter learning via

$$\min_{\theta} KL(\theta) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \ln \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{q(\mathbf{x}|\mathbf{y})q(\mathbf{y})} \mu(d\mathbf{x})\mu(d\mathbf{y}) \tag{9}$$

which has been systematically investigated in the early study of the BYY learning and actually shown being equivalent to the ML learning on $q(\mathbf{x})$ by (1) when the BYY system has a B-architecture [40].

B. Inner Representation Structures and Model Selection Mechanism

The structure of $q(\mathbf{y})$ for inner representation takes a crucial role in BYY harmony learning. Not only it should match the nature of a learning task to be perform (e.g., classification, clustering, feature extraction, temporal encoding, etc), but also $q(\mathbf{y})$, designed in an appropriate format with a scale large enough, is essential to enable the new model selection mechanism of $H(\theta, \mathbf{k})$ in (4). This mechanism will make the implementation of (6) result in a specific value of θ by which $\mathbf{k}$ is effectively reduced to an appropriate one. In other words, model selection is made automatically during making parameter learning. This is a unique advantage of $H(\theta, \mathbf{k})$ that is not shared by typical model selection approaches in the existing literature.

Before describing the inner representations for temporal encoding in Section II-C, it is helpful to make a systematic view on inner representations for typical learning tasks without considering temporal relation. For this purpose, we consider the special cases of $q(\mathbf{y})$ on a three-part representation $\mathbf{y} = \{y, z, \ell\}$ in which ℓ is a random variable that takes one of discrete values $1, \ldots, k$, z consists of $z^{(1)}, \ldots, z^{(m)}$ with each $z^{(j)}$ being a random variable that takes one of discrete values $1, \ldots, k_{\ell j}$, and y consists of $y^{(1)}, \ldots, y^{(m)}$ with each $y^{(j)}$ being a random variable that takes real numbers from some distribution. The three parts y, z, ℓ jointly have the following structure:

$$\begin{aligned}
q(\mathbf{y}) &= q(y, z, \ell) = q(y, z|\ell) q(\ell) \\
q(\ell) &= \sum_{j=1}^{k} \alpha_j \delta(\ell - j), \quad \alpha_\ell \geq 0, \ \sum_{j=1}^{k} \alpha_j = 1 \\
q(y, z|\ell) &= \prod_{j=1}^{m_\ell} q(y^{(j)}, z^{(j)}|\ell) \\
q(y^{(j)}, z^{(j)}|\ell) &= q(y^{(j)}|z^{(j)}, \ell) q(z^{(j)}|\ell) \\
q(z^{(j)}|\ell) &= \sum_{i=1}^{k_{\ell j}} \beta_{\ell j i} \delta(z^{(j)} - i), \quad \beta_{\ell j i} \geq 0, \\
\sum_{i=1}^{k_{\ell j}} \beta_{\ell j i} &= 1, \quad q(y^{(j)}|z^{(j)} = i, \ell) \\
&= G\left(y^{(j)}|\mu_{\ell j i}, \sigma^2_{\ell j i}\right). \qquad (10)
\end{aligned}$$

In this paper, we use $G(\xi|\mu, \Sigma)$ to denote that ξ comes from a Gaussian distribution with the mean μ and the variance matrix Σ. Also, $\delta(u)$ in the above and $\bar{\delta}(u)$ to be encountered later are defined as follows:

$$\delta(u) = \begin{cases} \delta(0), & \text{if } u = 0 \\ 0, & \text{otherwise} \end{cases}$$

with

$$\begin{aligned}
\delta(0) &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \\
\bar{\delta}(u) &= \begin{cases} 1, & \text{if } u = 0 \\ 0, & \text{otherwise.} \end{cases} \qquad (11)
\end{aligned}$$

First, we consider the case that $k = 1$, which is encountered in those learning tasks where the inner representation is $y = [y^{(1)}, \ldots, y^{(m)}]$ that is usually interpreted as hidden factors, feature vectors, etc. In this special case, it follows from (10) that

$$\begin{aligned}
q(y, z) &= \prod_{j=1}^{m} q(y^{(j)}, z^{(j)}) \\
q(y^{(j)}, z^{(j)}) &= q(y^{(j)}|z^{(j)}) q(z^{(j)}) \\
q(z^{(j)}) &= \sum_{i=1}^{k_j} \beta_{ji} \delta(z^{(j)} - i) \\
\beta_{ji} &\geq 0, \quad \sum_{i=1}^{k_j} \beta_{ji} = 1. \qquad (12)
\end{aligned}$$

As systematically reviewed in [27], a BYY system with

$$x = Ay + e, \ i.e., \ q(x|y) = G(x|Ay, \Sigma) \qquad (13)$$

and $q(y)$ by (12) leads to the classical factor analysis [2] when $y^{(j)}$ comes from a standard Gaussian with every $k_j = 1$. Moreover, if $k_j > 1$ and $y^{(j)}$ is real but non-Gaussian, we have

$$q(y) = \int q(y, z) dz = \prod_{j=1}^{m} \sum_{i=1}^{k_j} \beta_{ji} q\left(y^{(j)}|\theta_{ji}\right) \qquad (14)$$

which leads to not only the so-called learned mixture based ICA [27], [38] when $\Sigma \to 0$, but also the non-Gaussian factor analysis (NFA) that extends the classical factor analysis to the cases with non-Gaussian factors [27], [33]. Moreover, based on $H(\theta, \mathbf{k})$ in (4), the number m of factors can be decided during the implementation of learning. Furthermore, it has been shown recently in [26] that a BYY system with $q(\mathbf{y}) = q(y, z)$ by (12) in a general case leads to a new NFA version with the scales $\{k_j\}$ becoming selectable.

Second, we consider the case that $k = 1$, every $k_j = 2$ and every $y^{(j)}$ does not exist (i.e., $q(\mathbf{y}) = q(z^{(j)}) = \int q(y^{(j)}, z^{(j)}) dy^{(j)}$). In this case, the inner representation is a binary vector $z = [z^{(1)}, \ldots, z^{(m)}]$. Together with (13), we are lead to a binary factor analysis (BFA) [33], [36] and the LMSER learning [30], [42]. Again, the number m of factors can be decided during the maximization of $H(\theta, \mathbf{k})$ in (4).

Third, we consider the case of $k \neq 1$ and $q(\ell) = \int q(y, z, \ell) dy dz$, which is encountered in a learning task for classification or cluster analysis that classifies every observation to one of $\ell = 1, \ldots, k$ labels. In this case, it follows from (1) that:

$$q(x) = \sum_{j=1}^{k} \alpha_j q(x|\theta_j). \qquad (15)$$

When $q(x|\theta_j) = G(x|m_j, \Sigma)$, BYY learning by $H(\theta, \mathbf{k})$ in (4) can perform Gaussian mixture estimation, elliptic clustering, MSE clustering with the number k selected either automatically during learning [30], [33] or via criteria obtained from $H(\theta, \mathbf{k})$ [30], [33], [40]. With $q(x|\theta_j)$ extended beyond Gaussian in help of the so-called multiset mixture, BYY learning by $H(\theta, \mathbf{k})$ can detect multiple objects with various shapes [29].

With $q(\mathbf{y}) = q(y, z, \ell)$ in a general case by (10), the above tasks can be combined. For example, we have a local FA when every $y^{(j)}$ is Gaussian [30] or a local NFA [30] when $q(y|\ell)$ is given by (14), which performs a combined task with ℓ for classification and with $y = [y^{(1)}, \ldots, y^{(m_\ell)}]$ as extracted feature vector of the class ℓ. We can also get a local BFA and local LMSER [28], [30] with a binary vector $z = [z^{(1)}, \ldots, z^{(m)}]$ in place of $y = [y^{(1)}, \ldots, y^{(m_\ell)}]$.

Beyond unifying the results, $H(\theta, \mathbf{k})$ in (4) with $q(\mathbf{y}) = q(y, z, \ell)$ in a general case by (10) will also bring us new results. One example is that a local NFA can be performed with not only k and $\{m_\ell\}$ but also $\{k_{\ell j}\}$ all selectable. One other example is that a BFA and local BFA can be extended from $z^{(j)}$ taking binary values to k_j discrete values.

Why the BYY harmony learning by $H(\theta, \mathbf{k})$ has a new model selection mechanism has been explained from the perspectives of a least complexity in [30], [33], a best information transfer in [26], [31], and a generalized projection geometry in [26], respectively. In the following, we provide further intuitive insights on this mechanism, especially on which situations are suitable for deriving model selection criteria from $J(\mathbf{k}) = -H(\theta_{\mathbf{k}}, \mathbf{k})$ and on which situations are suitable for making learning by (6) with automatic model selection.

Observing $H(\theta, \mathbf{k})$ in (4), the mechanism that selects $\mathbf{k}$ can be intuitively observed from maximizing

$$\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \ln q(\mathbf{y})\mu(d\mathbf{x})\mu(d\mathbf{y}) = \int p(\mathbf{y}) \ln q(\mathbf{y})\mu(d\mathbf{y}),$$
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})\mu(d\mathbf{x})$$

which results in $q(\mathbf{y}) = p(\mathbf{y})$ if $p(\mathbf{y})$ falls in the parametric family of $q(\mathbf{y})$. That is, the BYY harmony learning by (4) contains a force that minimizes the entropy of the inner representation structure, i.e.,

$$H_y = -\int q(\mathbf{y}) \ln q(\mathbf{y})\mu(d\mathbf{y}) + c_\infty^y \tag{16}$$

where c_∞^y is infinite large constant that cancels out another infinite large constant $-c_\infty^y$ that comes from those $\delta(z^{(j)} - i)$'s and $\delta(\ell - j)$'s in $q(\mathbf{y})$ given by (10). Specifically, we have

$$c_\infty^y = \left(1 + \sum_{\ell=1}^{k} \alpha_\ell m_\ell\right) \ln \delta(0). \tag{17}$$

This c_∞^y comes from the term $-Z_q$. Specifically, it follows from (5) that Z_q consists of two parts

$$Z_q = Z_q^\infty + Z_q^b \tag{18}$$

where $Z_q^\infty = M \ln \delta(0)$ collects all the $\ln \delta(0)$-terms that comes from each δ function $\delta(u - c)$ in a density of a discrete variable u, including c_∞^y by (17), while Z_q^b is a bounded term contributed from not only those densities of continuous variables and but also the densities of discrete variables after turning each δ function $\delta(u - c)$ into its counter part $\bar{\delta}(u - c)$. Corresponding to $q(\mathbf{y})$ by (10), we have (19), shown at the bottom of the page. It further follows from (16) that we have

$$H_y = -\sum_{\ell=1}^{k} \alpha_\ell \ln \alpha_\ell + \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell} H_{\ell,j}$$
$$H_{\ell,j} = \begin{cases} 0.5[1 + \ln(2\pi)] + \sum_{i=1}^{k_{\ell j}} \beta_{\ell ji} \ln \frac{\sigma_{\ell ji}}{\beta_{\ell ji}} & \text{(a) real } y^{(j)} \\ -\sum_{i=1}^{k_{\ell j}} \beta_{\ell ji} \ln \beta_{\ell ji} & \text{(b) no } y^{(j)}. \end{cases} \tag{20}$$

Thus, it can be observed that the minimization of H_y leads to

$$\alpha_\ell = \delta(\ell - \ell^*), \quad \beta_{\ell ji} = \delta(i - i^*), \quad \sigma_{\ell ji}^2 = 0,$$

where ℓ^*, i^* can be selected arbitrarily. It is equivalent to that k, $k_{\ell j}$ are actually all pushed to the minimum 1. Moreover, $\sigma_{\ell ji}^2 = 0$ for all i means that $y^{(j)}$ is deterministic at the component ℓ and, thus, can be discarded. Furthermore, $\sigma_{\ell ji}^2 = 0$ for all i, j, ℓ means that m_ℓ is pushed to 0.

Of course, the above extreme situation will not really happen since $H(\theta, \mathbf{k})$ in (4) also contains

$$\int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \ln[q(\mathbf{x}|\mathbf{y})]\mu(d\mathbf{x})\mu(d\mathbf{y}),$$

which will increase as the scales k, m_ℓ, $k_{\ell j}$ increase. After the scales go beyond certain values, this increasing will stop and then remain unchanged when $p(\mathbf{x})$ is the true distribution. However, $H(\theta, \mathbf{k})$ will still increase slowly even after the scales go beyond these threshold values when $p(\mathbf{x})$ is actually obtained from a set $\mathcal{X} = \{x_t\}_{t=1}^T$ of a finite size. Finally, $H(\theta, \mathbf{k})$ will suddenly tend to infinite when the scales go too large.

The coordination of the above two tendencies makes $J(\mathbf{k}) = -H(\theta_{\mathbf{k}}, \mathbf{k})$ vary as illustrated in Fig. 2 where only one scale k is demonstrated. As shown in Fig. 2(a), $J(k)$ will decrease as k increases until a k^*. When $k > k^*$, making learning by (6) includes minimizing H_y, during which those extra α_ℓ, $\ell = k^* + 1, \ldots, k$ are pushed toward 0 and, thus, have no contribution to H_y. So, $J(k)$ will roughly remain flat for a certain period until reaching a breaking limit k^u. In such a learning process, we do not need to check $J(k)$ explicitly. Instead, given a $k > k^*$, model selection on k^* is automatically implied in the parameter learning by (6) via driving those extra α_ℓ, $\ell = k^* + 1, \ldots, k$ to 0. Similarly, we can understand the situations with not just k but also the other scales m_ℓ, $k_{\ell j}$, as well as their nontrivial degenerated cases.

$$Z_q^b = -\begin{cases} \ln\left\{\sum_{y \in Y} \sum_{\ell=1}^{k} \sum_{z} \alpha_\ell q(x|y, z, \ell) \prod_{j=1}^{m_\ell} \left[\beta_{\ell j z^{(j)}} G\left(y^{(j)}|\mu_{\ell j z^{(j)}}, \sigma_{\ell j z^{(j)}}^2\right)\right]\right\} & \text{(a) real } y^{(j)} \\ \ln\left\{\sum_{\ell=1}^{k} \sum_{z} \alpha_\ell q(x|z, \ell) \prod_{j=1}^{m_\ell} \sum_{i=1}^{k_{\ell j}} \beta_{\ell ji} \bar{\delta}(z^{(j)} - i)\right\} & \text{(b) no } y^{(j)}. \end{cases} \tag{19}$$

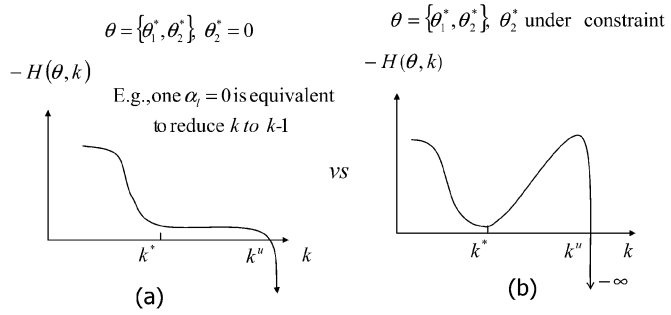


Fig. 2. (a) Automatic model selection during parameter learning with an initial k in a large enough value. (b) Model selection made after parameter learning on every k in a given interval $[k_d, k_u]$.

Carefully examining (20), we can observe that selectable scales via minimizing H_y are respectively the scales of random variables ℓ, z, y that are directly considered via $q(\mathbf{y}) = q(y, z, \ell)$. A similar observation may be expected even either $q(\mathbf{y}) = q(y, z, \ell)$ is in a structure different from (10) or $\mathbf{y}$ consists of other types of random variables jointly. However, what happens on those scales that their corresponding random variables are not directly considered in $q(\mathbf{y})$? For an example, we only make a direct consideration on a random label ℓ and random variables of y via $q(\mathbf{y}) = \int q(y, z, \ell)dz = q(y, \ell)$ with $q(y|\ell) = \prod_{j=1}^{m_\ell} \sum_{i=1}^{k_{\ell j}} \beta_{ji} G\left(y^{(j)}|\mu_{\ell ji}, \sigma^2_{\ell ji}\right)$. For (16), we have

$$H_y = -\sum_{\ell=1}^{k} \alpha_\ell \ln \alpha_\ell + \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell} H_{\ell,j},$$
$$H_{\ell,j} = -\sum_{i=1}^{k_{\ell j}} \beta_{ji} \int G\left(y^{(j)}|\mu_{\ell ji}, \sigma^2_{\ell ji}\right) \times \ln\left[\sum_{i=1}^{k_{\ell j}} \beta_{ji} G\left(y^{(j)}|\mu_{\ell ji}, \sigma^2_{\ell ji}\right)\right] \mu(dy^{(j)}).$$

Now, there is no a clear indication that minimizing $H_{\ell,j}$ will driving extra $\beta_{\ell ji}$ toward 0. Thus, we cannot guarantee that $\{k_{\ell j}\}$ can be selected automatically during learning.

Instead of automatical model selection, we may also simply fix

$$\alpha_\ell = \frac{1}{k} \quad \beta_{\ell ji} = \frac{1}{k_{\ell j}} \tag{21}$$

and then H_y in (20) becomes (22), shown at the bottom of the page, which will increase as the scales $k, m_\ell, k_{\ell j}$ increase. This can be clearly observed from its simplified case

$$H_y = \ln k + m \ln \kappa + H_y^r$$
$$H_y^r = \begin{cases} 0.5\ \mathrm{m}[1 + \ln(2\pi\sigma^2)] & \text{(a) real } y^{(j)} \\ 0 & \text{(b) no } y^{(j)} \end{cases}$$

by letting $m_\ell = m$, $k_{\ell j} = \kappa$ and $\sigma^2_{\ell ji} = \sigma^2$. So, $J(\mathbf{k}) = -H(\theta_{\mathbf{k}}, \mathbf{k})$ with H_y by (22) will have a shape as illustrated in Fig. 2(b) where only one scale k is demonstrated. That is, the model selection should be made by (7) via a two-stage implementation. A similar situation will happen when the parameters $\alpha_\ell, \beta_{\ell ji}, \sigma^2_{\ell ji}$ are learned by (9) or the ML learning on $q(\mathbf{x})$ by (1), during which extra parameters will no longer driven to 0 since there is no a clear force to minimize H_y.

C. Temporal Inner Representations and Two Temporal BYY Systems

Conceptually, the BYY system in (3) can be used to learn temporal dependence via (4) or (9) by inserting a pair of a temporal process $\mathbf{x} = x_1 x_2 \cdots x_T$ and its inner temporal process $\mathbf{y} = \mathbf{y}_1 \mathbf{y}_2 \cdots \mathbf{y}_T$, which, thus, can be called a *temporal Bayesian Ying-Yang (TBYY)* process system (shortly TBYY p-system). In implementation, this situation is usually too complicated to be handled directly. Further simplification can be made by considering types of inner representation for temporal encoding.

We adopt the well known Markovian assumption that the current inner coding $\mathbf{y}_t$ depends on only a finite number of past codings $\boldsymbol{\omega}_t = \{\mathbf{y}_{t-\tau}, \tau = 1, \ldots, p\}$ and the current observation x_t. On the other hand, the current x_t can also be regarded as generated from the current coding $\mathbf{y}_t$ that covers the past information already. Moreover, "knowing that the event $x_t = \bar{x}_t$ happens already" means that the event is irrelevant to any environment. As a result, we have the following temporal relations:

$$q(\mathbf{y}) = \prod_{t=1}^{T} q(\mathbf{y}_t|\boldsymbol{\omega}_t)$$
$$q(\mathbf{x}|\mathbf{y}) = \prod_{t=1}^{T} q(x_t|\mathbf{y}_t)$$
$$p(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} p(\mathbf{y}_t|x_t, \boldsymbol{\omega}_t)$$
$$p(\mathbf{x}) = \prod_{t=1}^{T} G\left(x_t|\bar{x}_t, h_x^2 I\right). \tag{23}$$

$$H_y = \ln k + \frac{1}{k} \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} \ln k_{\ell j} + H_y^r$$
$$H_y^r = \begin{cases} 0.5[1 + \ln(2\pi)] \sum_{\ell=1}^{k} \alpha_\ell m_\ell + \frac{1}{k} \sum_{\ell=1}^{k} \sum_{j=1}^{m_\ell} \frac{\sum_{i=1}^{k_{\ell j}} \ln \sigma_{\ell ji}}{k_{\ell j}} & \text{(a) real } y^{(j)} \\ 0 & \text{(b) no } y^{(j)} \end{cases} \tag{22}$$

Moreover, the inner representation $\mathbf{y}_t$ can still own a structure similar to (10). We can let all the appearances of $\mathbf{y}$ in (10) to be replaced with $\mathbf{y}_t|\boldsymbol{\omega}_t$ for encoding temporal relations. However, it will be too tedious and also unnecessary to consider all the temporal relations between $\mathbf{y}_t$ and $\boldsymbol{\omega}_t$. Usually, we only need to consider several typical relations. In this paper, we focus on the following relations:

$$\begin{aligned}
q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q(y_t, z_t, \ell_t|\boldsymbol{\omega}_t) \\
&= q(y_t, z_t|\ell_t, \boldsymbol{\omega}_t) q(\ell_t|\boldsymbol{\omega}_t) \\
q(y_t, z_t|\ell_t, \boldsymbol{\omega}_t) &= \prod_{j=1}^{m_\ell} q\left(y_t^{(j)}, z_t^{(j)}|\ell_t, \boldsymbol{\omega}_t\right) \\
q\left(y_t^{(j)}, z_t^{(j)}|\ell_t, \boldsymbol{\omega}_t\right) &= q\left(y_t^{(j)}|z_t^{(j)}, \ell_t, \boldsymbol{\omega}_t\right) q\left(z_t^{(j)}|\ell_t, \boldsymbol{\omega}_t\right) \\
q(\ell_t|\boldsymbol{\omega}_t) &= q\left(\ell_t|\ell_t^p\right), \\
q\left(z_t^{(j)}|\ell_t, \boldsymbol{\omega}_t\right) &= q\left(z_t^{(j)}|\ell_t, \mathbf{z}_{j,t}^p\right) \\
q\left(y_t^{(j)}|z_t^{(j)}, \ell_t, \boldsymbol{\omega}_t\right) &= q\left(y_t^{(j)}|z_t^{(j)}, \ell_t, \mathbf{y}_{j,t}^p\right) \\
\ell_t^p &= \{\ell_{t-\tau}, \tau = 1, \ldots, p\}, \\
\mathbf{z}_{j,t}^p &= \left\{z_{t-\tau}^{(j)}, \tau = 1, \ldots, p\right\} \\
\mathbf{y}_{j,t}^p &= \left\{y_{t-\tau}^{(j)}, \tau = 1, \ldots, p\right\}.
\end{aligned} \tag{24}$$

The detailed and simplified forms of the above will be further discussed in the rest sections. In the sequel we first derive the implementable forms of the BYY harmony learning from (4).

Putting (23) into (4), we have

$$\begin{aligned}
H(p||q) &= \sum_{t=1}^{T} H_t \\
H_t &= \int p(\omega_t) H_t(p||q, \omega_t) \mu(d\omega_t) \\
H_t(p||q, \omega_t) &= \int p(\mathbf{y}_t|x_t, \omega_t) G\left(x_t|\bar{x}_t, h_x^2 I\right) \\
&\quad \times \ln[q(x_t|\mathbf{y}_t, \omega_t) q(\mathbf{y}_t|\omega_t)] \\
&\quad \times \mu(dx_t)\mu(d\mathbf{y}_t) - Z_q^{(t)}.
\end{aligned} \tag{25}$$

We can further simplify the above integral in a format $\int p(u)T(u)\mu(du)$ by a Taylor expansion of $T(u)$ around the mean $\bar{u} = \int up(u)\mu(du)$, resulting in [32], [34]

$$\begin{aligned}
&\int p(u)T(u)\mu(du) \\
&\quad \approx T(\bar{u}) + c_T Tr[\Sigma H(\bar{u})] \\
&c_T = \begin{cases} 0, & \text{the first order expansion only} \\ 0.5, & \text{up to the second order expansion} \end{cases}
\end{aligned} \tag{26}$$

where Σ is the covariance matrix of $p(u)$, $H(u)$ is the Hessian of $T(u)$, and $Tr[C]$ is the trace of matrix C. Thus, from (25) in the first-order expansion, we approximately have

$$\begin{aligned}
H_t &= \int p(\mathbf{y}_t|x_t, \bar{\omega}_t) G\left(x_t|\bar{x}_t, h_x^2 I\right) \\
&\quad \times \ln[q(x_t|\mathbf{y}_t) q(\mathbf{y}_t|\bar{\omega}_t)] \mu(dx_t)\mu(d\mathbf{y}_t) - Z_q^{(t)} \\
\bar{\omega}_t &= \{\bar{\mathbf{y}}_{t-\tau}, \tau = 1, \ldots, p\}, \\
\bar{\mathbf{y}}_{t-\tau} &= \int \mathbf{y}_{t-\tau} p(\mathbf{y}_{t-\tau}|\bar{\omega}_{t-1}) \mu(d\omega_t).
\end{aligned} \tag{27}$$

Following the derivation made in [30, 32] it further follows that maximizing $H(p||q)$ results in

$$\begin{aligned}
p(\mathbf{y}_t|x_t, \bar{\omega}_t) &= \delta(\mathbf{y}_t - \hat{\mathbf{y}}(x_t)) \quad \text{a B-architecture,} \\
\hat{\mathbf{y}}(x_t) &= \arg\max_{\mathbf{y}_t}[q(x_t|\mathbf{y}_t) q(\mathbf{y}_t|\bar{\omega}_t)], \\
&\quad f(x_t|\bar{\omega}_t, \theta_f), \qquad \text{a BI-architecture.}
\end{aligned} \tag{28}$$

Putting it into (27), we further have

$$\bar{\mathbf{y}}_{t-\tau} = \hat{\mathbf{y}}_{t-\tau}. \tag{29}$$

Then considering (26) again in its second-order expansion case with $G\left(x_t|\bar{x}_t, h_x^2 I\right)$ taking the position of $p(u)$, we can further get

$$\begin{aligned}
H_t &= \ln[q(\bar{x}_t|\hat{\mathbf{y}}_t) q(\hat{\mathbf{y}}_t|\bar{\omega}_t)] - Z_q(t) \\
&\quad + 0.5 h_x^2 Tr[\pi_q(\bar{x}_t)] \\
\bar{\omega}_t &= \{\hat{\mathbf{y}}_{t-\tau}, \tau = 1, \ldots, p\} \\
\pi_q(x_t) &= \begin{cases} \frac{\partial^2 \ln q(x_t|\bar{\mathbf{y}}_t)}{\partial x_t \partial x_t^T}, & y_t \text{ is discrete and regarded as} \\ & \text{irrelevant to a real } x_t, \\ \frac{\partial^2 \ln q(x_t|\hat{\mathbf{y}}(x_t))}{\partial x_t \partial x_t^T}, & \hat{\mathbf{y}}(x_t) \text{ with respect to a} \\ & \text{real } x_t \text{ is in consideration,} \\ 0, & x_t \text{ is discrete.} \end{cases}
\end{aligned} \tag{30}$$

Another type of temporal BYY system can be obtained by considering an instantaneous Ying-Yang pair

$$p(x_t, \mathbf{y}_t) = p(\mathbf{y}_t|x_t) p(x_t), \quad q(x_t, \mathbf{y}_t) = q(x_t|\mathbf{y}_t) q(\mathbf{y}_t) \tag{31}$$

subject to the satisfaction of the following temporal dependence:

$$q(\mathbf{y}_t) = \int q(\mathbf{y}_t|\omega_t) q(\omega_t) d\omega_t. \tag{32}$$

We call such a case the temporal BYY instantaneous system (shortly TBYY i-system).

Similar to (25) and (27), we have $H(p||q) = \sum_{t=1}^{T} H_t$ with

$$\begin{aligned}
H_t &= \int p(\mathbf{y}_t|x_t) G\left(x_t|\bar{x}_t, h_x^2 I\right) \\
&\quad \times \ln[q(x_t|\mathbf{y}_t) q(\mathbf{y}_t)] \mu(dx_t)\mu(d\mathbf{y}_t) \\
&\quad - Z_q^{(t)}, \text{ subject to (32).}
\end{aligned} \tag{33}$$

Also, similar to (28) and (30) we further have

$$\begin{aligned}
p(\mathbf{y}_t|x_t) &= \delta(\mathbf{y}_t - \hat{\mathbf{y}}_t) \\
\hat{\mathbf{y}}_t &= \max_{\mathbf{y}_t}[q(x_t|\mathbf{y}_t) q(\mathbf{y}_t)] \\
H_t &= \ln[q(\bar{x}_t|\hat{\mathbf{y}}_t) q(\hat{\mathbf{y}}_t)] - Z_q(t) \\
&\quad + 0.5 h_x^2 Tr[\pi_q(\bar{x}_t|\hat{\mathbf{y}}_t)] \\
&\quad \text{subject to (32).}
\end{aligned} \tag{34}$$

The temporal BYY harmony learning by (33) is conceptually different from that by (25). Strictly speaking, the temporal BYY harmony learning by (25) considers a best harmony of two representations of the joint distribution of two entire temporal processes, including all the temporal dependence in consideration. Thus, a TBYY p-system should be conceptually better than a TBYY i-system by (33) that bases only on an instantaneous Ying-Yang pair by (31). However, after the first-order approximation by (26), a TBYY p-system by (27) actually already degraded into an instantaneous implementation that is actually inferior to that by (33) since the information carried from $t-1$ to t is merely via a point estimate ω_t in (27) but via an integral

of (32) in (33). This can also be observed by approximating the integral of (32) using the first-order approximation by (26), resulting in

$$q(y_t) = \int q(y_t|\bar{\omega}_t)q(\omega_t)d\omega_t = q(y_t|\bar{\omega}_t). \tag{35}$$

In implementation, the maximization of $H(p||q)$ can be made recursively at each time t to increase $\sum_{t'=1}^{t} H_{t'}$ in a certain extent by updating all the parameters to learn, which is implementable since the past information carried by either $\bar{\omega}_t$ or the integral of (32) is already available. Instead of directly maximizing $H(p||q)$, alternatively we can conduct learning in two stages similar to what discussed around (7). That is, at Stage I, parameter learning is made by the ML learning or equivalently by (9) when the BYY system has a B-architecture. Then, at Stage II model selection is implemented by (7) with the criteria derived from $J(\mathbf{k}) = -H(\theta_{\mathbf{k}}, \mathbf{k})$.

Specifically, by considering special cases of the inner representation in (24), the maximization of $H(p||q)$ will lead us to not only the well known Hidden Markov model (HMM) and various extensions in Section III but also variants and extensions of the well known Kalman filter related linear state space models in Section IV, with adaptive algorithms developed in help of typical updating rules in Table I and with model selection made either during the implementation of the algorithms or via specific criteria derived from $J(\mathbf{k}) = -H(\theta_{\mathbf{k}}, \mathbf{k})$.

As discussed in [30] and [33], two specific settings of Z_q will result in two types of regularization. One is called normalization that causes a new conscience de-learning mechanism similar to that of the rival penalized competitive learning (RPCL) [41]. The other is called data smoothing that causes a Tikinov-like regularization [23] with h_x^2 acting a role similar to the hyperparameter but being estimated in an easy implementing way. For simplicity and clarify, the roles of Z_q and h_x are ignored in the subsequent two sections by simply letting $Z_q^b = 0$ in (18) while keeping Z_q^∞ for cancelling out its counterpart infinite term that comes from the 1st part of $H(p||q)$. Without the role of Z_q^b, the maximization of $H(p||q)$ will lead $h_x = 0$ automatically. Thus, we can also simply setting $h_x = 0$ in Sections III and IV.

III. HMM and State Selection

A. Hidden Markov Models and State Selection Criteria

We start at the simplest special case of (24) as follows:

$$\begin{aligned} q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q\left(\ell_t|\ell_t^p\right) = q(\ell_t|\ell_{t-1}) \\ &= \sum_{j=1}^{k} \alpha_{\ell_{t-1}j}\delta(\ell_t - j) \\ \alpha_{\ell_{t-1}j} \geq 0, &\quad \sum_{j=1}^{k} \alpha_{\ell_{t-1}j} = 1 \end{aligned} \tag{36}$$

that is, the inner representation is simply discrete value $\ell = 1, \ldots k$ that indicates k states and $\alpha_{\ell_{t-1}j}$ represents the transfer probability from the state ℓ_{t-1} to the state j. The process $\{\ell_t\}_{t=1}^T$ is simply a 1st Markov chain that is hidded behind the process $\{x_t\}_{t=1}^T$ and, thus, is called hidden Markov model (HMM).

TABLE I
Typical Adaptive Updating Rules

(a) Adaptively increasing $\eta \ln G(x_t|A_I y_t + a_I, \Sigma_I) - h^2 Tr[\Sigma_I^{-1}]$

$$\begin{aligned} &e_{t,I} = x_t - A_I^{(t)} y_t - a_I^{(t)},\ a_I^{(t+1)} = a_I^{(t)} + \eta e_{t,I},\ A_I^{(t+1)} = A_I^{(t)} + \eta e_{t,I} y_t^T, \\ &\Sigma_I^{(t+1)} = \begin{cases} (1-\eta)\Sigma_I^{(t)} + \eta(h^2 I + e_{t,I}e_{t,I}^T), & \text{if } \eta \geq 0, \\ S_I^{(t+1)} S_I^{(t+1)\,T},\ S_I^{(t+1)} = S_I^{(t)} + \eta G_{\Sigma_I} S_I^{(t)}, & \text{if } \eta < 0. \end{cases} \\ &G_{\Sigma_I} = \Sigma_I^{(t)\,-1}(h^2 I + e_{t,I}e_{t,I}^T)\Sigma_I^{(t)\,-1} - \Sigma_I^{(t)\,-1}. \end{aligned}$$

(b) Adaptively increasing $\eta[\ln G(y_t^{(j)}|b_{Ij}y_{t-1}^{(j)} + \mu_{Ij}, \lambda_{Ij}^2) - h_{yj}^2 \lambda_{Ij}^{-2}]$

$$\begin{aligned} &\varepsilon_{t,I}^{(j)} = y_t^{(j)} - b_{Ij}^{(t)} y_{t-1}^{(j)} - \mu_{Ij}^{(t)},\ \bar{b}_{Ij}^{(t+1)} = \bar{b}_{Ij}^{(t)} + \eta\varepsilon_{t,I}^{(j)} y_{t-1}^{(j)}(1 - \bar{b}_{Ij}^{(t)\,2}),\ b_{Ij}^{(t+1)} = \frac{e^{\bar{b}_{Ij}^{(t+1)}} - e^{-\bar{b}_{Ij}^{(t+1)}}}{e^{\bar{b}_{Ij}^{(t+1)}} + e^{-\bar{b}_{Ij}^{(t+1)}}}, \\ &\mu_{Ij}^{(t+1)} = \mu_{Ij}^{(t)} + \eta\varepsilon_{t,I}^{(j)},\ \begin{cases} \lambda_{Ij}^{(t+1)\,2} = (1-\eta)\lambda_{Ij}^{(t)\,2} + \eta[\varepsilon_{t,I}^{(j)\,2} + h_{yj}^2], & \text{if } \eta \geq 0, \\ \lambda_{Ij}^{(t+1)} = \lambda_{Ij}^{(t)} + \eta[\varepsilon_{t,I}^{(j)\,2} + h_{yj}^2 - \lambda_{Ij}^{(t)\,2}]\lambda_{Ij}^{(t)}, & \text{if } \eta < 0. \end{cases} \end{aligned}$$

(c) Adaptively increasing $\eta[\ln G(y_t^{(j)}|\mu_{Ij}, b_{Ij}^2\pi_j^2 + \lambda_{Ij}^2) - h_{yj}^2(b_{Ij}^2\pi_j^2 + \lambda_{Ij}^2)^{-1}]$

$$\begin{aligned} &\varepsilon_{t,I}^{(j)} = y_t^{(j)} - \mu_{Ij}^{(t)},\ \bar{b}_{Ij}^{(t+1)} = \bar{b}_{Ij}^{(t)} + \eta\pi_j^{(t)\,2} b_{Ij}^{(t)}(1 - \bar{b}_{Ij}^{(t)\,2})[\varepsilon_{t,I}^{(j)\,2} + h_{yj}^2 - (b_{Ij}^{(t)}\pi_j^{(t)})^2 - \lambda_{Ij}^{(t)\,2}], \\ &b_{Ij}^{(t+1)} = \frac{e^{\bar{b}_{Ij}^{(t+1)}} - e^{-\bar{b}_{Ij}^{(t+1)}}}{e^{\bar{b}_{Ij}^{(t+1)}} + e^{-\bar{b}_{Ij}^{(t+1)}}},\ \mu_{Ij}^{(t+1)} = \mu_{Ij}^{(t)} + \eta\varepsilon_{t,I}^{(j)},\ \pi_j^{(t+1)\,2} = (b_{Ij}^{(t+1)}\pi_j^{(t)})^2 + \lambda_{Ij}^{(t)\,2}, \\ &\begin{cases} \lambda_{Ij}^{(t+1)\,2} = (1-\eta)\lambda_{Ij}^{(t)\,2} + \eta[\varepsilon_{t,I}^{(j)\,2} + h_{yj}^2 - (b_{Ij}^{(t+1)}\pi_j^{(t)})^2], & \text{if } \eta \geq 0, \\ \lambda_{Ij}^{(t+1)} = \lambda_{Ij}^{(t)} + \eta[\varepsilon_{t,I}^{(j)\,2} + h_{yj}^2 - (b_{Ij}^{(t+1)}\pi_j^{(t)})^2 - \lambda_{Ij}^{(t)\,2}]\lambda_{Ij}^{(t)}, & \text{if } \eta < 0. \end{cases} \end{aligned}$$

In the classic form of HMM, x_t is a discrete variable from

$$q(x_t|\hat{\mathbf{y}}_t) = q(x_t|\ell_t) = \sum_{j=1}^{n} b_{\ell_t j}\delta(x_t - j) \tag{37}$$

which is called the emitting matrix that describes the probability of emitting the label j at the state ℓ_t. This classic HMM model has been widely studied and extensively applied in the literature of speech processing and recognition for over 30 years [17] and also in the literature of bio-informatics in the recent years [5]. The learning is usually made on $q\left(\{x_t\}_{t=1}^T\right)$ under the maximum likelihood principle, and implemented by the well known Baum-Welch or EM algorithm [17].

The classic HMM model has also been extended into the following variants:

a) **Gaussian HMM** Instead of a label, x_t emitted from the state j is a real variable or a real vector from a Gaussian density

$$q(x_t|\hat{\mathbf{y}}_t) = q(x_t|\ell_t) = G(x_t|\mu_{\ell_t}, \Sigma_{\ell_t}). \tag{38}$$

At time t, we have $q(x_t|\ell_{t-1}) = \sum_{j=1}^{k} G(x_t|\mu_j, \Sigma_j)$ $\alpha_{\ell_{t-1}j}$ is a temporal Gaussian mixture with its mixing proportions $\alpha_{\ell_{t-1}j}$ varying as t.

b) **Gaussian Mixture HMM** x_t emitted from the state j is a non-Gaussian real random variable via an inner representation with not only a temporal label ℓ_t but also a non-temporal label z_t as follows:

$$\begin{aligned} q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q(z_t|\ell_t)q(\ell_t|\ell_{t-1}) \\ &\quad q(\ell_t|\ell_{t-1}) \text{is given by (36)} \\ q(z_t|\ell_t) &= \sum_{i=1}^{m_{\ell_t}} \beta_{\ell_t i}\delta(z_t - i) \\ q(x_t|\hat{\mathbf{y}}_t) &= q(x_t|\ell_t, z_t) = G(x_t|\mu_{\ell_t z_t}, \Sigma_{\ell_t z_t}). \end{aligned} \tag{39}$$

In this case, we have that $q(x_t|\ell_t) = \sum_{i=1}^{m_{\ell_t}} \beta_{\ell_t i} G(x_t|\mu_{\ell_t i}, \Sigma_{\ell_t i})$ is a Gaussian mixture for x_t emitted from the state j. Thus, $q(x_t|\ell_{t-1}) = \sum_{j=1}^{k} q(x_t|\ell_t = j)\alpha_{\ell_{t-1} j}$ is a temporal non-Gaussian mixture.

c) **Independent HMM** The inner representation consists of m independent channels with each $z_t^{(j)}$ taking discrete values $i = 1, \ldots, k_j$ at each t, and x_t is emitted from z_t in a linear model as follows:

$$q(\mathbf{y}_t|\boldsymbol{\omega}_t) = \prod_{j=1}^{m} q\left(z_t^{(j)}|z_{t-1}^{(j)}\right)$$

$$q(z_0) = \prod_{j=1}^{m} q\left(z_0^{(j)}\right)$$

$$q\left(z_t^{(j)}|z_{t-1}^{(j)} = r\right) = \sum_{i=1}^{k_j} \beta_{ri}^{(j)} \delta\left(z_t^{(j)} - i\right)$$

$$q(x_t|\hat{\mathbf{y}}_t) = q(x_t|z_t) = G(x_t|Az_t + \mu, \Sigma) \quad (40)$$

where $q\left(z_0^{(j)}\right)$ can be initialized equally taking one of $1 \ldots, k_j$. This is an extension of the so-called binary independent factor analysis [28], [32], [36] with not only the first-order Markovian temporal relation among $z_1, \ldots, z_t$ taken in consideration but also every $k_j = 2$ extended to any other values.

d) **Multi-independent HMM** The above independent HMM can be further extended to multiple ones of a same parameterization structure with each located at a different μ_ℓ and in a different specification of parameters:

$$q(\mathbf{y}_t|\boldsymbol{\omega}_t) = q(\ell_t|\ell_{t-1}) \prod_{j=1}^{m_\ell} q\left(z_t^{(j)}|z_{t-1}^{(j)}, \ell_t\right)$$

$$q(z_0|\ell_0) = \prod_{j=1}^{m_{\ell_0}} q\left(z_0^{(j)}|\ell_0\right)$$

$$q(\ell_t|\ell_{t-1}) \text{ is given by (36)}$$

$$q\left(z_t^{(j)}|z_{t-1}^{(j)} = r, \ell_t\right) = \sum_{i=1}^{k_j} \beta_{\ell_t ri}^{(j)} \delta\left(z_t^{(j)} - i\right)$$

$$q(x_t|\hat{\mathbf{y}}_t) = q(x_t|z_t, \ell_t) = G(x_t|A_{\ell_t} z_t + \mu_{\ell_t}, \Sigma_{\ell_t}) \quad (41)$$

where $q\left(z_0^{(j)}|\ell_0\right)$ is initialized equally taking one of $1 \ldots, k_j$.

Similar to the classic HMM, Gaussian HMM and Gaussian mixture HMM have also been widely studied with parameters estimated under the maximum likelihood principle by the well known Baum-Welch algorithm or the temporal EM algorithm [17]. Moreover, similar implementation algorithms can also be developed on the independent HMM and multi-independent HMM.

One key problem is how to decide the scales of the state spaces (i.e., the numbers k, k_j, m). Different values of these numbers correspond to a family of different specific HMM models that share a same system configuration but in different scales of representation ability. The task of deciding them is, thus, called model selection. It is well known that maximum likelihood principle is usually weak on making model selection, especially on a small size of training samples. As introduced in Section II, the BYY harmony learning provides a new learning mechanism that makes model selection implemented either automatically during an adaptive learning or subsequently after learning via a new class of model selection criteria derived from this mechanism.

In sequel, we further derive the model selection criteria on the above discussed HMM and variants from (7) with $J(\mathbf{k}) = -H(\theta_{\mathbf{k}}, \mathbf{k})$. Similar to the derivation process in [30], [31], we can get the following criteria for the classic HMM and the Gaussian HMM:

$$J(k) = J_x(k) + J_\ell(k)$$

$$J_x(k) = \begin{cases} -\sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{n} b_{\ell j} \ln b_{\ell j} & \text{(a) } x_t = 1, \ldots, n \\ 0.5 \sum_{\ell=1}^{k} \alpha_\ell \ln |\Sigma_\ell| & \text{(b) } x_t \text{ from Gaussian at each } \ell \end{cases}$$

$$J_\ell(k) = -\begin{cases} \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{k} \alpha_{\ell j} \ln \alpha_{\ell j} & \text{(a) TBYY } p\text{-system} \\ \sum_{\ell=1}^{k} \alpha_\ell \ln \alpha_\ell & \text{(b) TBYY } i\text{-systemm} \end{cases} \quad (42)$$

where α_ℓ is the probability that the state ℓ has been visited, and $\alpha_{\ell j}$ is the transfer probability from the state ℓ to the state j. Both types of the parameters as well as $b_{\ell j}$ or Σ_ℓ are obtained via a parameter learning process. The part $J_\ell(k)$ has two choices that correspond to a TBYY p-system and a TBYY i-system, respectively.

Being further extended to a Gaussian mixture HMM, we can select both k and m_ℓ by

$$J(k, \{m_\ell\}) = J_x(k, \{m_\ell\}) + J_z(k, \{m_\ell\}) + J_\ell(k), \quad J_\ell(k) \text{as in (42)}$$

$$J_x(k, \{m_\ell\}) = 0.5 \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell} \beta_{\ell j} \ln |\Sigma_{\ell j}|$$

$$J_z(k) = -\sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell} \beta_{\ell j} \ln \beta_{\ell j}. \quad (43)$$

Moreover, for an independent HMM we can decide the number m of independent channels in help of

$$J(m) = J_x(m) + J_z(m), \quad J_x(m) = 0.5 \ln |\Sigma|$$

$$J_z(m) = -\begin{cases} \sum_{j=1}^{m} \sum_{r=1}^{k_j} \pi_r^{(j)} \sum_{i=1}^{k_j} \beta_{ri}^{(j)} \ln \beta_{ri}^{(j)} & \text{(a) TBYY } p\text{-system} \\ \sum_{j=1}^{m} \sum_{r=1}^{k_j} \pi_r^{(j)} \ln \pi_r^{(j)} & \text{(b) TBYY } i\text{-system.} \end{cases} \quad (44)$$

In each channel j, $\pi_r^{(j)}$ is the probability that the state r has been visited, and $\beta_{ri}^{(j)}$ is the transfer probability from the state r to the state i. Being further extended to a multi-independent HMM, we have (45), shown at the bottom of the next page.

Historically, the use of (7) for selecting m on the classic HMM was firstly suggested in [37] and then in [34] but in a cumbersome form. Also, selecting m on the independent HMM was firstly suggested in [35] via a direct use of (7) without any simplification. Then, a simplified form $J(m) = 0.5 \ln |\Sigma| + J_z(m)$ was obtained in [32] but still with a tedious second term $J_z(m)$. In contrast, the criteria proposed above are not only compact in their representation but also much simple to be computed accurately.

B. Recursive Harmony Learning and Automatic State Selection

Instead of the above two stage implementation, state selection can also be made automatically during implementing the parameter learning by (6) with $H(p\|q) = \sum_{t=1}^{T} H_t$ given either by (34) on a TBYY i-system or by (30) and (28) on a TBYY p-system.

Recursively from $t-1$ to t, parameters are updated to increase H_t with extra states discarded due to the mechanism discussed in Section II-B. In the following, the detailed algorithms are developed for the HMM models discussed in Section III-A. By these algorithms, parameters are updated in an iterative format $\theta^{(t+1)} = \theta^{(t)} + \eta_0 \delta\theta$, where $\delta\theta$ is either a gradient direction $\nabla_\theta H_t$ or a direction that has a positive projection on $\nabla_\theta H_t$, and $\eta_0 > 0$ is a learning step size. For convenience, the same notation η_0 is used everywhere. However, it should be understood that it may take different values for updating different parameters.

a) **Classic HMM and Gaussian HMM**: shown in (46), at the bottom of the page.

b) **Gaussian Mixture HMM**: shown in (47), at the bottom of the page.

$$
\begin{aligned}
J(k,\{m_\ell\}) &= J_x(k,\{m_\ell\}) + J_z(k,\{m_\ell\}) \\
&\quad + J_\ell(k), \quad J_\ell(k) \text{ as in (42)} \\
J_x(k,\{m_\ell\}) &= 0.5 \sum_{\ell=1}^{k} \alpha_\ell \ln |\Sigma_\ell| \\
J_z(k,\{m_\ell\}) &= -\begin{cases} \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell} \sum_{r=1}^{k_{j\ell}} \pi_{\ell r}^{(j)} \sum_{i=1}^{k_{j\ell}} \beta_{\ell r i}^{(j)} \ln \beta_{\ell r i}^{(j)} & \text{(a) TBYY } p\text{-system} \\ \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell} \sum_{r=1}^{k_{j\ell}} \pi_{\ell r}^{(j)} \ln \pi_{\ell r}^{(j)} & \text{(b) TBYY } i\text{-system.} \end{cases}
\end{aligned} \tag{45}
$$

(a) $\bar{\ell}_t = \max_\ell [q(x_t|\ell)\pi_\ell]$ with $\pi_\ell = \begin{cases} \alpha_{\bar{\ell}_{t-1}\ell}, & \text{(a) TBYY } p\text{-system} \\ \alpha_\ell^{t-1}, & \text{(b) TBYY } i\text{-system} \end{cases}$

(b) $\begin{cases} c_{j\bar{\ell}_t}^{(t)} = c_{j\bar{\ell}_t}^{(t-1)} + \eta_0 \left(\bar{\delta}_{\bar{\ell}_t j} - b_{j\bar{\ell}_t}^{(t-1)} \right), \quad b_{j\bar{\ell}_t}^{(t)} = \dfrac{e^{c_{j\bar{\ell}_t}^{(t)}}}{\sum_{j=1}^{m} e^{c_{j\bar{\ell}_t}^{(t)}}}, & \text{if } x_t = j \\ \text{update } \mu_\ell,\ \Sigma_\ell \text{ by Tab.1(a) with } I = \ell, A_I = 0 \text{ and } \eta = \eta_0 \bar{\delta}_{\bar{\ell}_t \ell}, & \text{if } x_t \text{ from } G(x_t|\mu_{\ell_t}, \Sigma_{\ell_t}). \end{cases}$

(c) $f_{\bar{\ell}_{t-1}\ell}^{(t)} = f_{\bar{\ell}_{t-1}\ell}^{(t-1)} + \eta_0 \hat{\eta}_t \left(\bar{\delta}_{\bar{\ell}_t \ell} - \alpha_{\bar{\ell}_{t-1}\ell}^{(t-1)} \right) \quad \hat{\eta}_t = \begin{cases} 1, & \text{(a) TBYY } p\text{-system} \\ \dfrac{\alpha_{\bar{\ell}_{t-1}}^{t-1}}{\sum_{j=1}^{k} \alpha_{\bar{\ell}_t j}^{(t-1)} \alpha_j^{(t-1)}} & \text{(b) TBYY } i\text{-system} \end{cases}$

$$\alpha_{\bar{\ell}_{t-1}\ell}^{(t)} = \frac{e^{f_{\bar{\ell}_{t-1}\ell}^{(t)}}}{\sum_{j=1}^{m} e^{f_{\bar{\ell}_{t-1}\ell}^{(t)}}}, \quad \alpha_\ell^{(t)} = \sum_{j=1}^{k} \alpha_{\ell j}^{(t)} \alpha_j^{(t-1)}.$$

(d) if π_ℓ keeps to be small such that it can be regarded as zero, we discard the state ℓ as well as all the relevant parameters, and let $k = k - 1$. (46)

(a) $(\bar{\ell}_t, \bar{j}_t) = \max_{\ell,j} [G(x_t|\mu_{\ell j}, \Sigma_{\ell j})\beta_{\ell j}\pi_\ell]$, π_ℓ as in (a) of (46).

(b) update $\mu_{\ell j}, \Sigma_{\ell j}$ by Tab.1(a) with $I = (\ell, j), \quad A_I = 0$ and $\eta = \eta_0 \bar{\delta}_{\bar{\ell}_t \ell}$

(c) update $\alpha_{\ell_{t-1}\ell}^{(t)}$ as in (c) of (46) $\quad b_{\bar{\ell}_t j}^{(t)} = b_{\bar{\ell}_{t-1} j}^{(t-1)} + \eta_0 \left(\bar{\delta}_{\bar{j}_t j} - \beta_{\bar{\ell}_{t-1} j}^{(t-1)} \right), \quad \beta_{\bar{\ell}_t j}^{(t)} = \dfrac{e^{b_{\bar{\ell}_t j}^{(t)}}}{\sum_{j=1}^{m_\ell} e^{b_{\bar{\ell}_t j}^{(t)}}}.$

(d) same as (d) in (46)

(e) if $\beta_{\ell j}$ keeps to be small such that it can be regarded as zero, we discard the sub-state $z_t = j$ under the state ℓ as well as all the relevant parameters, let $m_\ell = m_\ell - 1$. (47)

c) **Independent HMM and Multi-Independent HMM:** shown in (48), at the bottom of the page. In the particular case that $k = 1$ and, thus, $\pi_\ell = 1$, we can drop $\alpha_{\ell_{t-1}\ell}$ and all the subscripts of ℓ. In this case, the above algorithm with every $k_j = 2$ becomes an extension of the adaptive harmony learning algorithm for the independent HMM given by Table III and [32, p. 839].

For the reason discussed in Section II-B, those extra states will be driven out of the TBYY system in a sense that they act their roles with probability zero. In other words, appropriate states are selected automatically with the extra states becoming out of functions. In a view of computational efficiency, keeping the extra states in the system will waste many computing costs. We can further remove the extra states by detecting and, thus, discarding them during learning by step (d) and step (e) in these algorithms.

IV. TFA, Identifiable State Spaces, and Space Dimension

A. From Kalman Filter to TFA

We further consider the cases of real variables included in the inner representation. One special case is

$$\begin{aligned} q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q(y_t|y_{t-1}) = G(y_t|By_{t-1}, \Lambda), \text{ or} \\ y_t &= By_{t-1} + \varepsilon_t \text{ with } G(\varepsilon_t|0, \Lambda) \text{ and} \\ Ey_{t-1}\varepsilon_t^T &= 0. \end{aligned} \tag{49}$$

That is, the inner representation is a multi-channel autoregressive (AR) model that is guaranteed to be stable when

$$\|B\| < 1, \quad \|B\| = \lambda_{\max}^B \tag{50}$$

where $\lambda_{\max}^B$ is the positive root of the largest eigen-value of BB^T.

From this inner representation, an observation x_t is generated via the following linear state space:

$$\begin{aligned} q(x_t|y_t) &= G(x_t|Ay_t, \Sigma), \text{ or} \\ x_t &= Ay_t + e_t, \text{ with } G(e_t|0, \Sigma) \text{ and} \\ Ey_t e_t^T &= 0, \quad Ee_t\varepsilon_t^T = 0. \end{aligned} \tag{51}$$

In the literature of control theory, the model by (49) and (51) has actually been widely studied for decades since the early sixties [14], by which the parametric matrices A, B as well as the variances of Gaussian e_t, ε_t have to be known, the task is to estimate $\hat{y}_t$ upon observing x_t and then get $\hat{x}_t = A\hat{y}_t$ as the result after filtering out the noise in x_t. The optimal solution is given by the well know Kalman filter. As shown in [6], the estimate y_t given by the Kalman filter is equivalent to

$$\begin{aligned} \hat{y}_t &= E[y_t|x_t] = \int y_t p(y_t|x_t) dy_t \\ p(y_t|x_t) &= \frac{q(x_t|y_t)q(y_t)}{\int q(x_t|y_t)q(y_t)dy_t}, \\ \text{s.t. } q(y_t) &= \int q(y_t|\bar{y}_{t-1})q(\bar{y}_{t-1})d\bar{y}_{t-1}. \end{aligned} \tag{52}$$

Alternatively, it can be observed in [32] that $\min_{p(y_t|x_t)} KL$ by (9) also results in the above $p(y_t|x_t)$. Moreover, we have the following equivalence:

$$E[y_t|x_t] = \hat{y}_t = \max_{y_t}[q(x_t|y_t)q(y_t)]. \tag{53}$$

That is, in this case both the learning by maximizing $H(p\|q)$ and the learning by minimizing $KL(p\|q)$ is equivalent to the well known Kalman filter on estimating $\hat{y}_t$. Moreover, a criterion has been derived from $H(p\|q)$ for determining an appropriate m as the dimension of the state space [27].

(a) $(\bar{z}_t, \bar{\ell}_t) = \arg\max_{z_t,\ell} \left[G(x_t|A_\ell z_t + \mu_\ell, \Sigma_\ell)\pi_\ell \prod_{j=1}^{m_\ell} \pi_{\ell z_t^{(j)}}^{(j)} \right]$ with π_ℓ as in (a) of (46)

$$\pi_{\ell i}^{(j)} = \begin{cases} \beta_{\ell r i}^{(j)}, r = \bar{z}_{t-1}^{(j)}, & \text{(a) TBYY } p\text{-system} \\ \sum_{r=1}^{k_{j\ell}} \beta_{\ell r i}^{(j)} \pi_{\ell r}^{(j)}, & \text{(b) TBYY } i\text{-system} \end{cases}$$

(b) update $\mu_\ell, A_\ell, \Sigma_\ell$ by Tab.1(a) with $I = \ell$, $\eta = \eta_0 \bar{\delta}_{\bar{\ell}_t \ell}$ and z_t as y_t

(c) update $\alpha_{\ell_{t-1}\ell}^{(t)}$ as in (c) of (46) $\hat{\eta}_t = \begin{cases} 1, & \text{(a)TBYY } p\text{-system} \\ \frac{\beta_{\bar{\ell}_t r i}^{(j)(t-1)}}{\sum_{r=1}^{k_{j\ell}} \beta_{\bar{\ell}_t r i}^{(j)(t-1)} \pi_{\ell r}^{(j)(t-1)}}, & \text{(b) TBYY } i\text{-system} \end{cases}$

$$\bar{i}_{\bar{\ell}_t} = \bar{z}_t^{(j)} \text{at} \quad \ell = \bar{\ell}_t,\ b_{\bar{\ell}_t r i}^{(j)(t)} = b_{\bar{\ell}_t r i}^{(j)(t-1)} + \eta_0 \hat{\eta}_t \left(\bar{\delta}_{\bar{i}_t i}^{(j)} - \beta_{\bar{\ell}_t r i}^{(j)(t-1)} \right)$$

$$\beta_{\bar{\ell}_t r i}^{(j)(t)} = \frac{e^{b_{\bar{\ell}_t r i}^{(j)(t)}}}{\sum_{j=1}^{m} e^{b_{\bar{\ell}_t r i}^{(j)(t)}}}, \quad \pi_{\ell i}^{(j)} = \sum_{r=1}^{k_{j\ell}} \beta_{\bar{\ell}_t r i}^{(j)(t)} \pi_{\ell r}^{(j)(t-1)}.$$

(d) same as (d) in (46)

(e) if $\pi_{\ell i}^{(j)}$ keeps to to be small such that it can be regarded as zero, we discard the state $z_t^{(j)} = i$ at ℓ as well as all the related parameters, and let $k_{j\ell} = k_{j\ell} - 1$. Moreover, if both $k_{j\ell} = 1$ and $\pi_{\ell i}^{(j)} = 1$, we delete $z_t^{(j)}$ as well as all the related parameters, let $m_\ell = m_\ell - 1$. (48)

The above equivalence occurs under the assumption that the matrices A, B, Λ, Σ are known, as requested in the literature of the Kalman filter studies. However, this is too restrictive in many temporal modeling problems where it is difficult to know A, B in advance. In the rest of this section, we will show how this restrictive assumption can be considerably relaxed. Maximizing $H(p\|q)$ not only goes beyond the Kalman filter but also provides a new model selection mechanism that is not possessed by minimizing $KL(p\|q)$.

Knowing only $\{x_t\}$, the problem of estimating all the unknowns y_t, A, B, as well as Σ and Λ suffer intrinsic indeterminacy. The indeterminacy can be observed from the perspective of making a linear mapping $y'_t = \phi y_t$, with a general matrix ϕ. It follows from (49) and (51) that we get the same form $x_t = A'y'_t + e_t$, $y'_t = B'y'_{t-1} + \varepsilon'_t$ with

$$A' = A\phi^{-1}, \quad B' = \phi B\phi^{-1}, \quad \varepsilon'_t = \phi\varepsilon_t \tag{54}$$

and ε'_t remains to be a Gaussian in the form of ε_t as in (49). That is, y_t in model is not identifiable. Of course, we expect that y_t is identifiable.

For this purpose, we need to specify the meaning that y_t is identifiable. Using the notation $\mathcal{D}$ for the family of all the diagonal matrices and $\mathcal{P}$ for the family of all the permutation matrices, we observe that

$$y'_t = \phi y_t, \quad \phi = \Pi D, \quad D \in \mathcal{D}, \ \Pi \in \mathcal{P} \tag{55}$$

will keep an one-to-one correspondence between the m series $\left\{y_t^{(j')}\right\}_{t=1}^T$, $j = 1, \ldots, m$ and the m series $\left\{y_t^{(i)}\right\}_{t=1}^T$, $i = 1, \ldots, m$. That is, for each series, $\left\{y_t^{(i)}\right\}_{t=1}^T$ there must be one and only one series $\left\{y_t^{(j')}\right\}_{t=1}^T$ such that both the series have a same waveform with differences only in a constant scale and a permutation (i.e., an order change $i \neq j$). The differences are ignorable in many applications and, thus, y_t is usually said to be identifiable if we are able to specify y_t via y'_t in a sense of (55). Also, we say that y_t is identifiable in a strong sense if $D = I$ (i.e., there is no scale indeterminacy) or in a weak sense if $D \in \mathcal{D}$ but $D \neq I$.

Restricting A or B to be known and unchanged (i.e., $A' = A$ or $B' = B$), the only choice is $\phi = I$ and, thus, y_t is identifiable in a strong sense. That is, there is no identifiable problem in the Kalman filter studies. Releasing all the restrictions, it follows from (54) that y_t is not identifiable. Between the two extremes, we can require that y_t becomes identifiable as one or more of A', B' and ε'_t retain certain invariant nature of the counterpart from A, B and ε_t.

Getting two more notations

$$\begin{aligned} \mathcal{O}_{nm} &= \{U \in \mathcal{O}_{nm} : U \text{ is a } n \times m \\ &\qquad \text{matrix and } UU^T = I\} \\ \mathcal{D}^* &= \{D \in \mathcal{D} : \text{all diagonal elements of} \\ &\qquad D \text{ are different}\} \end{aligned} \tag{56}$$

we make further discussions as follows:

- Requiring $A' \in \mathcal{O}_{nm}$, $A \in \mathcal{O}_{nm}$, we are lead to $\phi \in \mathcal{O}$, i.e., y_t are not identifiable due to an arbitrary unknown rotation ϕ. Requiring $A' \in \mathcal{D}$, $A \in \mathcal{D}$, we are lead to $\phi \in \mathcal{D}$, i.e., y_t becomes identifiable in a weak sense. However, restricted in $A \in \mathcal{D}$, $x_t = Ay_t + e_t$ has a very limited regression ability.
- Similarly requiring $B' \in \mathcal{O}_{mm}$, $B \in \mathcal{O}_{mm}$ leads to $\phi \in \mathcal{O}$ while requiring $B' \in \mathcal{D}^*$, $B \in \mathcal{D}^*$, we are lead to ϕ that satisfies (55).
- Requiring both ε'_t and ε_t remain to be independent among their components, i.e., $G(\varepsilon'_t|0, \Lambda')$ and $G(\varepsilon_t|0, \Lambda)$ with both $\Lambda' \in \mathcal{D}$ and $\Lambda \in \mathcal{D}$, we are lead to $\phi = DU\Lambda^{-0.5}$, $U \in \mathcal{O}_{nm}$ and $D \in \mathcal{D}$.

We can observe that B takes an important role. For the degenerated case $B = 0$, $x_t = Ay_t + e_t$, $y_t = \varepsilon_t$ is actually the classical factor analysis that has been widely studied in the literature of statistics for several decades [2], for which we have to suffer the indeterminacy of an arbitrary unknown rotation. The rotation can be removed by requiring $B \in \mathcal{D}^*$. It further follows from $y_t = By_{t-1} + \varepsilon_t$ that only the diagonal part of Λ is actually useful. Thus, we see the state space model by (49) and (51) has been restricted into

$$\begin{aligned} q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q(y_t|y_{t-1}) \\ &= \prod_{j=1}^m G\left(y_t^{(j)}|b_j y_{t-1}^{(j)}, \lambda_j^2\right) \\ \text{together with } q(x_t|y_t) &= G(x_t|Ay_t, \Sigma) \end{aligned} \tag{57}$$

where it follows from (50) that we have $|b_j| < 1$, which is always satisfied by letting

$$\begin{aligned} B &= \text{diag}[b_1, \ldots, b_m], \quad b_j = \frac{e^{c_j} - e^{-c_j}}{e^{c_j} + e^{-c_j}} \\ C &= \text{diag}[c_1, \ldots, c_m]. \end{aligned} \tag{58}$$

If y_0 is initialized to be independent among its components, e.g., $G(y_0|0, I)$, we have m mutually independent AR series $\{y_t^{(j)}\}_{t=1}^T$, $j = 1, \ldots, m$. As firstly proposed in [32], [34] we, thus, call (57) temporal factor analysis (TFA) since it extends the classic factor analysis with temporal dependence encoded by $B \neq 0$. Being different from those Kalman filter studies, A, B as well as Σ in (57) are unknowns to be learned.

However, the indeterminacy of an arbitrary unknown rotation cannot be completely removed for a $B \in \mathcal{D}$ but $B \notin \mathcal{D}^*$. E.g., when $B = bI$, we have $B' = \phi B\phi^{-1} = bI$ for any $\phi \in \mathcal{O}_{mm}$. When $B \neq I$ but having m' same diagonal elements, ϕ also does not satisfy (55) but contain an arbitrary unknown rotation in the corresponding m' dimensional subspace. This case is not difficult to understand, the duplications of m' diagonal elements in B means that this m' dimensional temporal link actually can be replaced by a one-dimensional link and, thus, this part degenerated back to a $m' - 1$ dimensional classic factor analysis.

Though y_t becomes identifiable in a weak sense is already enough in many practical applications, the scale indeterminacy $y'_t = \phi y_t$, $\phi \in \mathcal{D}$ leads to the indeterminacy of the parameters A and Λ, which makes parameter learning on A and Λ difficult to

TABLE II
A GENERAL ADAPTIVE LEARNING ALGORITHM

(a) $(\bar{y}_t, \bar{z}_t, \bar{\ell}_t) = arg \max_{y_t, z_t, \ell} L(y_t, z_t, \ell)$ *with*

$$L(y_t, z_t, \ell) = G(x_t|A_\ell y_t + \mu_\ell, \Sigma_\ell)\pi_\ell \prod_{j=1}^{m_\ell} q(y_t^{(j)}|y_{t-1}^{(j)}, \ell, z_t^{(j)})\beta_{\ell j z_t^{(j)}},$$

$$\pi_\ell = \begin{cases} \alpha_{\bar{\ell}_{t-1}\ell}, & \text{(a) TBYY p-system,} \\ \alpha_\ell^{t-1}, & \text{(b) TBYY i-system;} \end{cases}$$

$$q(y_t^{(j)}|y_{t-1}^{(j)}, \ell, z_t^{(j)}) = \begin{cases} G(y_t^{(j)}|b_{\ell j z_t^{(j)}} y_{t-1}^{(j)} + \mu_{\ell j z_t^{(j)}}, \lambda^2_{\ell j z_t^{(j)}}), & \text{(a) TBYY p-system,} \\ G(y_t^{(j)}|\mu_{\ell j z_t^{(j)}}, b^2_{\ell j z_t^{(j)}} \pi_{\ell j z_t^{(j)}} + \lambda^2_{\ell j z_t^{(j)}}), & \text{(b) TBYY i-system.} \end{cases}$$

(b) *update* $A_\ell, \mu_\ell, \Sigma_\ell$, by Tab.1(a) with $I = \ell, \eta = \eta_0 \bar{\delta}_{\bar{\ell}_t \ell}$,

$(c1)$ $f^{(t)}_{\bar{\ell}_{t-1}\ell} = f^{(t-1)}_{\bar{\ell}_{t-1}\ell} + \eta_0 \hat{\eta}_t (\bar{\delta}_{\bar{\ell}_t \ell} - \alpha^{(t-1)}_{\bar{\ell}_{t-1}\ell})$, $\hat{\eta}_t = \begin{cases} 1, & \text{(a) TBYY p-system,} \\ \frac{\alpha^{t-1}_{\bar{\ell}_{t-1}}}{\sum_{j=1}^k \alpha^{(t-1)}_{\bar{\ell}_t j} \alpha_j^{(t-1)}}, & \text{(b) TBYY i-system;} \end{cases}$

$$\alpha^{(t)}_{\bar{\ell}_{t-1}\ell} = e^{f^{(t)}_{\bar{\ell}_{t-1}\ell}} / \sum_{j=1}^m e^{f^{(t)}_{\bar{\ell}_{t-1}\ell}}, \ \alpha_\ell^{(t)} = \textstyle\sum_{j=1}^k \alpha^{(t)}_{\ell j} \alpha_j^{(t-1)}.$$

$(c2)$ $c^{(t)}_{\bar{\ell}_t j \bar{z}_t^{(j)}} = c^{(t-1)}_{\bar{\ell}_t j \bar{z}_t^{(j)}} + \eta_0 \hat{\eta}_t (\bar{\delta}^{(j)}_{\bar{z}_t^{(j)} i} - \beta^{(t-1)}_{\bar{\ell}_t j \bar{z}_t^{(j)}})$, $\beta^{(t)}_{\bar{\ell}_t j \bar{z}_t^{(j)}} = e^{c^{(t)}_{\bar{\ell}_t j \bar{z}_t^{(j)}}} / \sum_{j=1}^m e^{c^{(t)}_{\bar{\ell}_t j \bar{z}_t^{(j)}}}$.

$(c3)$ with $h_{yj} = 0, \lambda^2_{\ell j 1} = 1, I = (\ell j i), \eta = \eta_0 \bar{\delta}_{\bar{\ell}_t \ell} \prod_{j=1}^{m_\ell} \bar{\delta}^{(j)}_{\bar{z}_t^{(j)} i}$,

update $b_{\ell j i}, \mu_{\ell j i}, \lambda^2_{\ell j i}$ for all ℓ, j and $i \geq 2$ by $\begin{cases} \text{Tab.1(b),} & \text{(a) TBYY p-system,} \\ \text{Tab.1(c),} & \text{(b) TBYY i-system.} \end{cases}$

(d) *if* π_ℓ *keeps to be small such that it can be regarded as zero, we discard the state* ℓ *as well as all the relevant parameters, and let* $k = k - 1$.

(e) *if* $\beta^{(t)}_{\ell j i}$ *keeps to to be small such that it can be regarded as zero, we discard the state* $z_t^{(j)} = i$ *at* ℓ *as well as all the related parameters, and let* $k_{j\ell} = k_{j\ell} - 1$. *Moreover, if both* $k_{j\ell} = 1$ *and* $\beta^{(t)}_{\ell j i} = 1$, *we delete* $z_t^{(j)}$ *as well as all the related parameters, let* $m_\ell = m_\ell - 1$.

Remarks

(1) When $k = 1, k_{j\ell} = 1$, ℓ and $z_t^{(j)}$ take only one value 1, and all the appearances of π, α, β and their updating equations can be dropped. It is simplified into an adaptive TFA algorithm.

(2) When $k_{j\ell} = 1$, all $z_t^{(j)}$ take only one value 1, and all the appearances of β and their updating equations can be dropped. It is simplified into an adaptive HMM-TFA algorithm.

(3) When $k = 1$, ℓ takes only one value 1, and all the appearances of π, α and their updating equations can be dropped. It is simplified into an adaptive TNFA algorithm.

(4) In general, it is an adaptive HMM-TNFA algorithm.

implement. This scale indeterminacy can be removed via fixing the scale of Λ. That is, in (57) we simply set

$$\lambda_j^2 = 1, \quad j = 1, \ldots, m \text{ or } \Lambda = I. \tag{59}$$

A direct consequence of (59) is that the space dimension m is, thus, fixed and the automatic selection ability as discussed in Section II-B is lost. To select an appropriate m, we need to enumerate a number values of m and make parameter learning by the TFA algorithm given in Table II(a) at every value of m. Then, we select m by (60), shown at the bottom of the page, where $1 - b_j^2$ comes from $var\left(y_t^{(j)}\right) = b_j^2 var\left(y_{t-1}^{(j)}\right) + 1$ and $var\left(y_t^{(j)}\right) = var\left(y_{t-1}^{(j)}\right)$ as $t \to \infty$, with $var[u]$ denoting the variance of u.

B. HMM Gated TFA, HMM Gated TNFA, and Coupled State Spaces

The above TFA can be further extended along three directions. One is considering typical special cases of (10) and (24), which leads us to the following models:

a) **HMM gated TFA** A number of different TFA models work collaboratively with the engagement of each TFA gated by a hidden Markov chain. That is, we have

$$q(\mathbf{y}_t|\boldsymbol{\omega}_t) = q(y_t|\boldsymbol{\omega}_t, \ell_t) q(\ell_t|\ell_{t-1})$$
$$q(y_t|\boldsymbol{\omega}_t, \ell_t) = \prod_{j=1}^{m_{\ell_t}} G\left(y_t^{(j)}|b_{\ell_t j} y_{t-1}^{(j)} + \mu_{\ell_t j}, \lambda^2_{\ell_t j}\right)$$
$$q(\ell_t|\ell_{t-1}) \text{ is given by (36)}$$
$$q(x_t|\hat{\mathbf{y}}_t) = q(x_t|y_t, \ell_t) = G(x_t|A_{\ell_t} y_t + \mu_{\ell_t}, \Sigma_{\ell_t}). \tag{61}$$

$$\min_m J(m), J(m) = J_x(m) + J_y(m), \quad J_x(m) = 0.5 \ln |\Sigma|$$
$$J_y(m) = 0.5m[1 + \ln(2\pi)] + \begin{cases} 0, & \text{(a) TBYY } p\text{-system} \\ -0.5 \sum_{j=1}^m \ln\left(1 - b_j^2\right) & \text{(b) TBYY } i\text{-system.} \end{cases} \tag{60}$$

We have $q(x_t|\ell_{t-1}) = \sum_{\ell_t=1}^{k} q(x_t|\ell_t,\ell_{t-1})q(\ell_t|\ell_{t-1})$ which can be regarded as an extension of the Gaussian HMM by (38). The role of each TFA is gated by $q(\ell_t|\ell_{t-1})$ that varies as t. Particularly, when $q(\ell_t|\ell_{t-1})$ degenerates to $q(\ell_t)$ without temporal relation, we are lead to a mixture of k different TFA models at different locations, which is, thus, called local TFA [27], [32].

b) **Temporal Non-Gaussian Factor Analysis (TNFA)** For a TFA, we have m mutually independent channels $\{y_t^{(j)}\}_{t=1}^{T}$, $j = 1,\ldots,m$ with each being an AR series. The limitation of an AR series can be further extended by the following inner representation space:

$$\begin{aligned} q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= \prod_{j=1}^{m} q\left(y_t^{(j)}|y_{t-1}^{(j)},z_t^{(j)}\right) q\left(z_t^{(j)}\right) \\ q\left(z_t^{(j)}\right) &= \sum_{r=1}^{k_j} \beta_{jr}\delta\left(z_t^{(j)}-r\right) \\ q\left(y_t^{(j)}|y_{t-1}^{(j)},z_t^{(j)}\right) &= G\left(y_t^{(j)}|b_{jz_t^{(j)}}y_{t-1}^{(j)}+\mu_{jz_t^{(j)}},\lambda^2_{jz_t^{(j)}}\right) \\ q(x_t|\hat{\mathbf{y}}_t) &= q(x_t|y_t) = G(x_t|Ay_t,\Sigma). \end{aligned} \tag{62}$$

We have

$$q\left(y_t^{(j)}|y_{t-1}^{(j)}\right) = \sum_{r=1}^{k_j} \beta_{jr} G\left(y_t^{(j)}|b_{jr}y_{t-1}^{(j)}+\mu_{jr},\lambda^2_{jr}\right)$$

that is a mixture of AR series. In the special case that every k_j is 1, (62) returns to a TFA. Also, ignoring temporal relation by simply setting every $b_{jr} = 0$, (62) returned to the so-called non-Gaussian factor analysis (NFA) that extends the classic factor analysis with decorrelated Gaussian factors replaced by independent non-Gaussian real factors [27], [30], [32].

c) **HMM gated TNFA** Similar to (61) we can let a number of different TNFA models work collaboratively as follows:

$$\begin{aligned} q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q(\ell_t|\ell_{t-1})q(y_t,z_t|\boldsymbol{\omega}_t,\ell_t), \\ q(y_t,z_t|\boldsymbol{\omega}_t,\ell_t) &= \prod_{j=1}^{m_\ell} q\left(y_t^{(j)}|y_{t-1}^{(j)},\ell_t,z_t^{(j)}\right) \\ &\quad \times q\left(z_t^{(j)}|\ell_t\right) \\ q(\ell_t|\ell_{t-1}) &\text{is given by (36)} \\ q\left(y_t^{(j)}|y_{t-1}^{(j)},\ell_t,z_t^{(j)}\right) &= G\left(y_t^{(j)}|b_{\ell_t j z_t^{(j)}}y_{t-1}^{(j)}\right. \\ &\quad \left.+\mu_{\ell_t j z_t^{(j)}},\lambda^2_{\ell_t j z_t^{(j)}}\right) \\ q\left(z_t^{(j)}|\ell_t\right) &= \sum_{r=1}^{k_{j\ell}} \beta_{\ell_t jr}\delta\left(z_t^{(j)}-r\right) \\ q(x_t|\hat{\mathbf{y}}_t) &= q(x_t|y_t,\ell_t) \\ &= G(x_t|A_{\ell_t}y_t+\mu_{\ell_t},\Sigma_{\ell_t}) \end{aligned} \tag{63}$$

which can be regarded as an extension of the Gaussian mixture HMM by (47) with a Gaussian mixture for x_t replaced by a TNFA process.

d) **Other Extensions** There are also several other useful extensions. For an example, both the above TNFA and HMM gated TNFA can be further extended by considering

$$\begin{aligned} q(x_t|\hat{\mathbf{y}}_t) &= q(x_t|y_t,\ell_t,z_t) \\ &= G(x_t|A_{\ell_t z_t}y_t+\mu_{\ell_t z_t},\Sigma_{\ell_t z_t}). \end{aligned} \tag{64}$$

Shown in Table II is a general adaptive learning algorithm that implements parameter learning for a HMM gated TNFA directly and for TFA, HMM gated TFA, and TNFA as special cases (see Table II(a)–(c)), respectively, derived from maximizing $H(p\|q)$ with H_t either by (34) on a TBYY i-system or by (30) and (28) on a TBYY p-system. Similar to (59) for removing the scale indeterminacy in a TFA, we set $\lambda^2_{\ell j} = 1, \forall\, \ell, j$ for a HMM gated TFA and set $\lambda^2_{\ell j1} = 1$ in Table II(c) for a HMM gated TNFA (including TNFA as a special case). The reason of setting $\lambda^2_{\ell j1} = 1$ only at $i = 1$ for a HMM gated TNFA is that the scale indeterminacy of $y_t^{(j)}$ will affect the variances of $q\left(y_t^{(j)}|y_{t-1}^{(j)},z_t^{(j)}=r\right)$, $r = 1,\ldots,k_j$ by only a same unknown scale and, thus, is able to normalize only one of these variances to 1.

Similar to step (d) an step (e) of the algorithms in Section III-B, it follows from the model selection mechanism explained in Section II-B that all the extra states indexed by ℓ and $z_t^{(j)}$, $\forall\, j$ will be discarded by Table II(d) and (e) during learning. In an extreme case, it leads to $k = 1$, $k_j = 1$, $\forall\, j$. However, there is at least one dimension that is not able to be removed due to the setting of $\lambda^2_{\ell j1} = 1$. Similar to (60), we need to make parameter learning via enumerating a number values of m or/and m_ℓ, and select them by $\max_{m,\{m_\ell\}} J$ via (65), shown at the bottom of the next page.

The criteria can also be used on selecting k, k_j, $k_{j\ell}$ together with m, m_ℓ after making parameter learning either under the maximum likelihood principle or still via the algorithm of Table II but simply setting $\alpha_\ell = 1/k$, $\beta_{jr} = 1/k_j$, and $\beta_{\ell jr} = 1/k_{j\ell}$.

Another direction to extend TFA is considering an alternative parameterization for $x_t = Ay_t + e_t$. Instead of fixing (59), the scale indeterminacy can also be removed via fixing the scale of A via imposing $A \in \mathcal{O}_{nm}$. However, doing so directly on the model by (57) will narrow down the representation ability of $x_t = Ay_t + e_t$. Instead of (57) that imposes strong constraints $B \in \mathcal{D}^*$ and $\Lambda = I$ but no constraint on A, we can impose the constraint $A \in \mathcal{O}_{nm}$ while relax the constraints on B and Λ as follows:

$$\begin{aligned} y_t &= By_{t-1}+\varepsilon_t, \quad B = DWD^{-1} \\ W &= W^T, \quad D \in \mathcal{D}^* \\ q(x_t|y_t) &= G(x_t|Ay_t,\Sigma), \quad A \in \mathcal{O}_{nm} \\ q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q(y_t|y_{t-1}) \\ &= G(\varepsilon_t|0,\Lambda), \quad \Lambda \text{ is positive definite} \end{aligned} \tag{66}$$

where B is no longer diagonal and, thus, m channels $\left\{y_t^{(j)}\right\}_{t=1}^{T}$, $j = 1,\ldots,m$ are no longer mutually independent. So, we call this model a coupled state space.

For a mapping $y'_t = \phi y_t$, requiring $A \in \mathcal{O}_{nm}$ means that it is only possible for $\phi \in \mathcal{O}_{mm}$, and then requiring $\phi B\phi^T = \phi DWD^{-1}\phi^T$ to keep the format of B in (66) further makes ϕ become a permutation matrix. In other words, y_t becomes identifiable in a weak sense. However, the indeterminacy of an

arbitrary unknown rotation cannot be completely removed for a $D \in \mathcal{D}$ but $D \notin \mathcal{D}^*$, e.g., when $B = bI$, we have $\phi B \phi^T = D'W'D'^{-1}$, $D' = I$, $W' = \phi W \phi^T$, $W' = W'^T$ for any $\phi \in \mathcal{O}_{mm}$. Also, ϕ may contain an arbitrary unknown rotation in the corresponding m' dimensional subspace when $D \neq I$ but having m' same diagonal elements.

In fact, the model by (57) is equivalent to a special case of (66). We consider a general matrix A in the following singular value decomposition:

$$\begin{aligned} &A = UDV^T, \quad U \in \mathcal{O}_{nm}, \quad V \in \mathcal{O}_{mm} \\ &D = \text{diag}[d_1, \ldots, d_m]. \end{aligned} \tag{67}$$

Considering a linear mapping

$$y_{t\prime} = \phi y_t, \quad \phi = DV^T \tag{68}$$

it follows from (54) that $A' = U$, $\varepsilon'_t = DV^T\varepsilon_t$ and, thus, $E\left[\varepsilon'_t \varepsilon'^T_{t'}\right] = DV^T E\left[\varepsilon_t \varepsilon_t^T\right] VD = \Lambda'$. That is, (57) becomes equivalent to

$$\begin{aligned} &y'_t = B'y'_{t-1} + \varepsilon'_t, \quad \Lambda' = E\left(\varepsilon'_t \varepsilon'^T_t\right) = DV^T \Lambda V D \\ &B' = DW'D^{-1}, \quad W' = V^T B V, \quad W' = W'^T \\ &x_t = A'y'_t + e_t, \quad A' \in \mathcal{O}_{nm} \end{aligned} \tag{69}$$

which is actually a special case of (66) in which Λ and B are bundled via D and V.

Being different from the case of (57), maximizing $\ln G(\varepsilon_t|0, \Lambda)$ in (66) will push the variance of $\varepsilon_t^{(j)}$ toward zero if the dimension $y_t^{(j)}$ is extra. This can be observed more clearly by considering $\Lambda = \text{diag}\left[\lambda_1^2, \ldots, \lambda_m^2\right]$ where we have $\sum_{j=1}^m \ln G\left(\varepsilon_t^{(j)}|0, \lambda_j^2\right)$. If λ_j^2 is pushed to zero due to $\max \ln G\left(\varepsilon_t^{(j)}|0, \lambda_j^2\right)$, the autoregressive process of the dimension $y_t^{(j)}$ will decade quickly and soon go out of function to the whole system. In other words, model selection on m will happen automatically during parameter learning. Moreover, we can save computing resources by simply discarding $y_t^{(j)}$, $y_{t-1}^{(j)}$, $\varepsilon_t^{(j)}$, eliminating the jth column of A, and also eliminating the corresponding column and row of A and B.

Similar to Table II, adaptive learning algorithm can also be directly derived from maximizing $H(p\|q)$ with H_t by (34) on a TBYY i-system or by (30) and (28) on a TBYY p-system. The key difference is that we should now consider the constraints on A and B, which can be made by considering (68) and

$$\begin{aligned} \ln q\left(y'_t|y'_{t-1}, \ell, z_t\right) &= \ln q(y_t|y_{t-1}, \ell, z_t) - \ln|D_\ell| \\ q(y_t|y_{t-1}, \ell, z_t) &= \prod_{j=1}^{m_\ell} q\left(y_t^{(j)}|y_{t-1}^{(j)}, \ell, z_t^{(j)}\right). \end{aligned}$$

Then, learning can be implemented in help of Table II via with the following modifications and supplements: shown in (70) at

---

$$\begin{aligned} &J = J_\ell + J_x + J_y + J_z, \; J_\ell \text{ is same as in (42)} \\ &J_z = -\begin{cases} 0, & \text{HMM gated TFA} \\ \sum_{j=1}^{m}\sum_{r=1}^{k_j} \beta_{jr} \ln \beta_{jr}, & \text{TNFA} \\ \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell}\sum_{r=1}^{k_{j\ell}} \beta_{\ell jr} \ln \beta_{\ell jr} & \text{HMM gated TNFA.} \end{cases} \\ &J_x = 0.5\begin{cases} \ln|\Sigma| & \text{TFA and TNFA} \\ \sum_{\ell=1}^{k} \alpha_\ell \ln|\Sigma_\ell| & \text{HMM gated TFA and HMM gated TNFA.} \end{cases} \\ &J_y = 0.5[1 + \ln(2\pi)]\bar{m} + \begin{cases} J_y^p & \text{(a) TBYY } p\text{-system} \\ J_y^I & \text{(b) TBYY } i\text{-system} \end{cases} \\ &\text{where } \bar{m} = \begin{cases} m & \text{TNFA} \\ \sum_{\ell=1}^{k} \alpha_\ell m_\ell & \text{HMM gated TFA and HMM gated TNFA.} \end{cases} \\ &J_y^p = -0.5\begin{cases} 0 & \text{HMM gated TFA} \\ \sum_{j=1}^{m}\sum_{r=2}^{k_j} \beta_{jr} \ln \lambda_{jr}^2 & \text{TNFA} \\ \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell}\sum_{r=2}^{k_{j\ell}} \beta_{\ell jr} \ln \lambda_{\ell jr}^2 & \text{HMM gatedTNFA.} \end{cases} \\ &J_y^I = 0.5\begin{cases} \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell} \ln\left(1 - b_{\ell j}^2\right) & \text{HMM gated TFA} \\ \sum_{j=1}^{m}\sum_{r=1}^{k_j} \beta_{jr} \ln\left(1 - b_{jr}^2\right) & \text{TNFA} \\ \sum_{\ell=1}^{k} \alpha_\ell \sum_{j=1}^{m_\ell}\sum_{r=1}^{k_{j\ell}} \beta_{\ell jr} \ln\left(1 - b_{\ell jr}^2\right) & \text{HMM gated TNFA.} \end{cases} \end{aligned} \tag{65}$$

the bottom of the page. We can get O_{A_ℓ} and O_{V_ℓ} in help of two approaches. One is presented by [30, Eq. (80), (81)] and the other is given in [10]. The details are similar to that made in the nontemporal cases in [26].

The third direction to extend TFA is considering another type of coupled state space as follows:

$$\begin{aligned} q(\mathbf{y}_t|\boldsymbol{\omega}_t) &= q(y_t|\boldsymbol{\omega}_t,\ell_t)q(\ell_t|\ell_{t-1}) \\ q(y_t|\boldsymbol{\omega}_t,\ell_t) &= G\left(y_t|B_{\ell_t}\boldsymbol{\omega}_t+\nu_{\ell_t},\lambda_{\ell_t}^2\right) \\ & \quad q(\ell_t|\ell_{t-1}) \text{ is given by (36)} \\ q(x_t|\hat{\mathbf{y}}_t) &= q(x_t|y_t,\ell_t) \\ &= G\left(x_t|A_{\ell_t}y_t+\mu_{\ell_t},\sigma_{\ell_t}^2\right) \end{aligned} \tag{71}$$

with both A_ℓ and B_ℓ having no constraint but requiring samples being known in each pair $\bar{x}_t$, $\bar{y}_t$. From knowing $\bar{y}_{t-1}$ we can predict $\bar{y}_t$ and $\bar{x}_t$ by

$$\begin{aligned} \hat{y}_t &= E(y_t|y_{t-1}) \\ &= \sum_{\ell_t=1}^{k} p(\ell_t|\bar{y}_{t-1})(B_{\ell_t}\boldsymbol{\omega}_t+\nu_{\ell_t}) \\ p(\ell_t|\bar{y}_{t-1}) &= \frac{\pi_{\ell_t}G\left(y_t|B_{\ell_t}\boldsymbol{\omega}_t+\nu_{\ell_t},\lambda_{\ell_t}^2\right)}{\sum_{\ell_t=1}^{k}\pi_{\ell_t}G\left(y_t|B_{\ell_t}\boldsymbol{\omega}_t+\nu_{\ell_t},\lambda_{\ell_t}^2\right)} \\ \hat{x}_t &= E(x_t|y_{t-1}) = \sum_{\ell_t=1}^{k} p(\ell_t|\bar{y}_{t-1})(C_{\ell_t}\boldsymbol{\omega}_t+c_{\ell_t}) \\ C_{\ell_t} &= A_{\ell_t}B_{\ell_t} \quad c_{\ell_t} = A_{\ell_t}\nu_{\ell_t}+\mu_{\ell_t} \end{aligned} \tag{72}$$

which can be regarded as an extension of the so-called normalized extended RBF net [30], [33], [39], with temporal relation considered via a number of multi-channel AR processes gated by a HMM process.

Learning can be implemented by a simplified version of Table II as shown in (73), at the bottom of the page.

V. Experimental Demonstrations

A. On Classical HMM and Independent HMM

We start at considering a classical HMM. Shown in Fig. 3(a) is a snapshot of the observation series with nine observed labels, which are generated from a HMM of four hidden states as shown in Fig. 3(b) in a comparison with the estimated states. There are 39 error bits out of the 800 estimated bits, i.e., the error rate is 0.049. Shown in Fig. 3(c) is a $J(k)$ curve by (42) with the correct state number four found as its minimum.

To illustrate how automatic model selection works, learning is also implemented by (46). Initially, the number m of states is set at 5 and $\pi^0 = [0.2, 0.2, 0.2, 0.2, 0.2]$ is set equally. After a learning process with 27 iterative steps, we get $\pi = [0.11, 0.12, 0.58, 0.15, 0.04]$. The one of 0.04 can be regarded as redundant and, thus, is discarded. Hereafter, the parameters converge to the desired values after 59 iterative steps.

Furthermore, an independent HMM is considered as shown in Fig. 4, with a ten-dimensional observation generated from five independent channels via $x_t = Az_t + e_t$ under a Gaussian noise from $G(e_t|0, 0.05I_{10})$. Fig. 4(b) shows a comparison of the original states with the estimated states. There are 33 error bits out of the 800 estimated bits, i.e., the error rate is 0.041. Shown in Fig. 4(c) is a $J(k)$ curve by (44), with the correct number 5 detected as its minimum.

B. On TFA p-System and TFA i-System

We further consider the observations $\{x_t\}_{t=1}^T$ generated from $x_t = Ay_t + e_t$, $y_t = By_{t-1} + \varepsilon_t$ as given in Table III(a) with $G(\varepsilon|0, I_3)$ and $G(e_t|0, 0.05I_3)$. What is used as known

---

(a) Modifying Tab.2(a), we get $(\bar{y}_t, \bar{z}_t, \bar{\ell}_t) = \arg\max_{y_t,z_t,\ell}[L(y_t, z_t, \ell) - \ln|D_\ell|]$

(b) update μ_ℓ, Σ_ℓ as in Tab.2(b), then map $\nabla_{A_\ell} \ln G(\bar{x}_t|A_\ell\bar{y}_t + \mu_\ell, \Sigma_\ell) = \bar{y}_t(\bar{x}_t - A_\ell\bar{y}_t - \mu_\ell)^T\Sigma_\ell^{-1}$ into its projection O_{A_ℓ} on the manifold $A_\ell^TA_\ell = I$, update $A_\ell^{\text{new}} = A_\ell^{\text{old}} + \eta O_{A_\ell}$.

(c) After implementing Tab.2(c3), map $G_{V_\ell} = \nabla_{V_\ell} \ln q\left(y_t'|y_{t-1}', \ell, z_t\right) = V_\ell\left[y_t\phi^T(y_t) + y_{t-1}\psi^T(y_{t-1})\right]$ into its projection O_{V_ℓ} on the manifold $V_\ell^TV_\ell = I$, update $V_\ell^{\text{new}} = V_\ell^{\text{old}} + \eta O_{V_\ell}$
also get $G_{D_\ell} = \nabla_{V_\ell} \ln q\left(y_t'|y_{t-1}', \ell, z_t\right) = \left\{V_\ell\left[y_t\phi^T(y_t) + y_{t-1}\psi^T(y_{t-1})\right]V_\ell D_\ell - I\right\}D_\ell^{-2}$
update $D_\ell^{\text{new}} = D_\ell^{\text{old}} + \eta G_{D_\ell}$ where $\phi(y_t) = \nabla_{y_t} \ln q(y_t|y_{t-1}), \psi(y_{t-1}) = \nabla_{y_{t-1}} \ln q(y_t|y_{t-1})$.

(d) After Tab.2(d)& 2(e), if $d_\ell^{(j)}$ is pushed to zero ,discarding $y_t^{(j)}, y_{t-1}^{(j)}, \varepsilon_t^{(j)}$
eliminating the j-th column of A, and the corresponding column and row of A and B. (70)

(a) $\bar{\ell}_t = \arg\max_\ell\left[G\left(\bar{x}_t|A_\ell\bar{y}_t + \mu_\ell, \sigma_\ell^2\right)G\left(y_t|B_{\ell_t}\boldsymbol{\omega}_t + \nu_\ell, \lambda_\ell^2\right)\pi_\ell\right] \quad \pi_\ell = \begin{cases} \alpha_{\bar{\ell}_{t-1}\ell}, & \text{(a) TBYY } p\text{-system} \\ \alpha_\ell^{t-1}, & \text{(b) TBYY } i\text{-system} \end{cases}$

(b) implement Tab.2(b) and Tab.2(c1) and discard Tab.2(c2)

(c) update $B_\ell, \nu_\ell, \lambda_\ell^2$ by Tab.1(a) with $I = \ell, \eta = \eta_0\bar{\delta}_{\bar{\ell}_t\ell}$, also with $B_\ell, \nu_\ell, \lambda_\ell^2$ in places of $A_\ell, \mu_\ell, \Sigma_\ell$.

(d) implement Tab.3(d). (73)

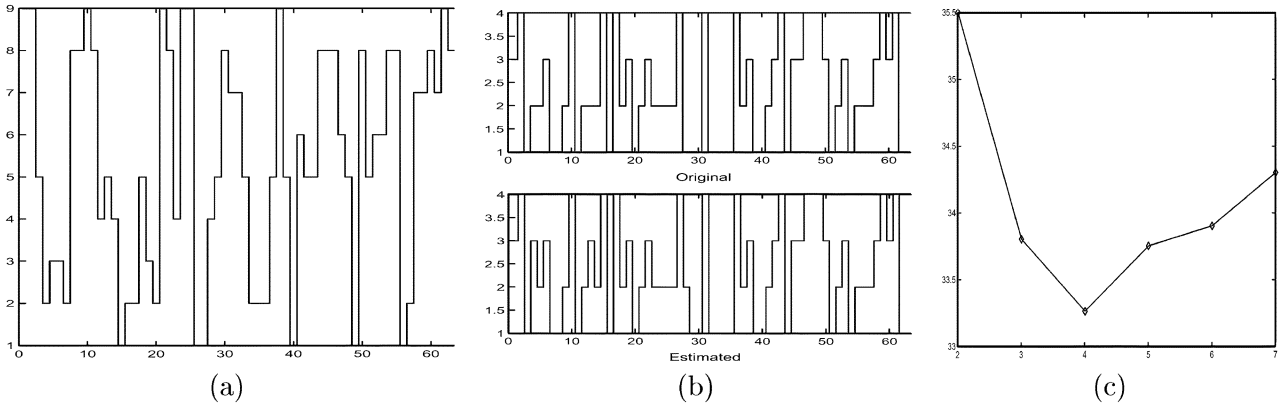


Fig. 3. An illustration on classical HMM. (a) Snapshot of the observation x_t. (b) Snapshot of the original and estimated series of hidden states. (c) $J(k)$ curve for selecting k.

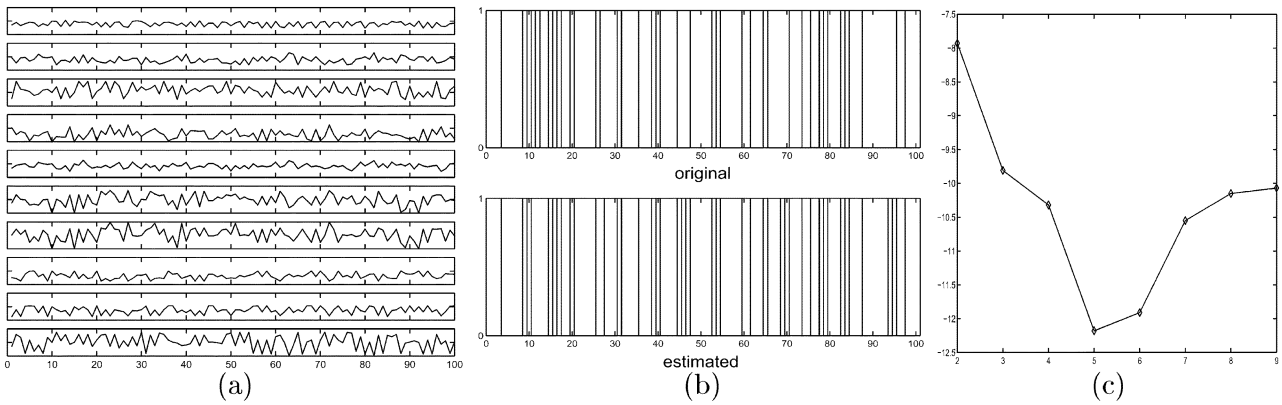


Fig. 4. Illustration on independent HMM. (a) Snapshot of the observation x_t. (b) Snapshot of the original and estimated series of one hidden channel. (c) $J(m)$ curve for selecting m.

TABLE III
EXPERIMENTS ON TWO TYPES OF TFA

$$A = \begin{pmatrix} 1.2 & 0.8 & 0.4 \\ 0.4 & -1.1 & 0.6 \\ 1.5 & 0.5 & 1.0 \end{pmatrix}$$
$$B = diag[0.5, -0.2, 0.6]$$
$$\hat{A} = \begin{pmatrix} 0.79 & 0.40 & 1.21 \\ -1.11 & 0.62 & 0.43 \\ 0.54 & 1.01 & 1.51 \end{pmatrix}$$
$$\hat{B} = diag[0.24, 0.65, 0.54]$$

(a)

State	p-TFA	i-TFA
1	0.1341	0.0898
2	0.1125	0.0647
3	0.1521	0.1158
Mean	0.1329	0.0901

(b)

$$R_p = \begin{pmatrix} 1.0 & & \\ 0.0810 & 1.0 & \\ 0.1874 & 0.1235 & 1.0 \end{pmatrix}$$

$$R_i = \begin{pmatrix} 1.0 & & \\ -0.0689 & 1.0 & \\ 0.0340 & -0.0558 & 1.0 \end{pmatrix}$$

(c)

in learning is only $\{x_t\}_{t=1}^T$ under the independence assumption among components of y_t. The estimated $\hat{A}$, $\hat{B}$ on a TFA p-system are listed in Table III(a) in comparison with the original A, B, it can be observed that a quite good identification is obtained though there is a unsolvable indeterminacy of a permutation $123 \rightarrow 231$.

Shown in Table III(b) and (c) are the performances obtained on a TFA p-system (shortly p-TFA) and TFA i-system (shortly i-TFA), respectively. Shown in Table III(b) are the mean square error between the estimated state $\hat{y}_t^{(j)}$ and the original state $y_t^{(j)}$, with a snapshot of $\hat{y}_t$ and y_t shown in Fig. 5. A comparison of correlation coefficients between $\hat{y}_t^{(i)}$ and $\hat{y}_t^{(j)}$ is also shown in Table III(c). We can see that I-TFA outperforms P-TFA in a sense that both the dependence between source components is reduced and the smaller mean square errors are achieved.

C. On Determining Dimension and Financial Application

We let the observations $\{x_t\}_{t=1}^N$, $N = 1000$ be generated from $x_t = Ay_t + e_t$, $y_t = By_{t-1} + \varepsilon_t$ with $B = \text{diag}[0.12, 0.21, -0.15, 0.24, -0.13]$ and A is a full rank 30×5 matrix, where ε_t and e_t are distributed with $G(\varepsilon|0, I)$ and $G(e_t|0, \Sigma)$ respectively, with the diagonal and off-diagonal elements of the symmetrical Σ being random variables from the uniform distributions $U(0.1, 0.25)$ and $U(0, 0.01)$, receptively.

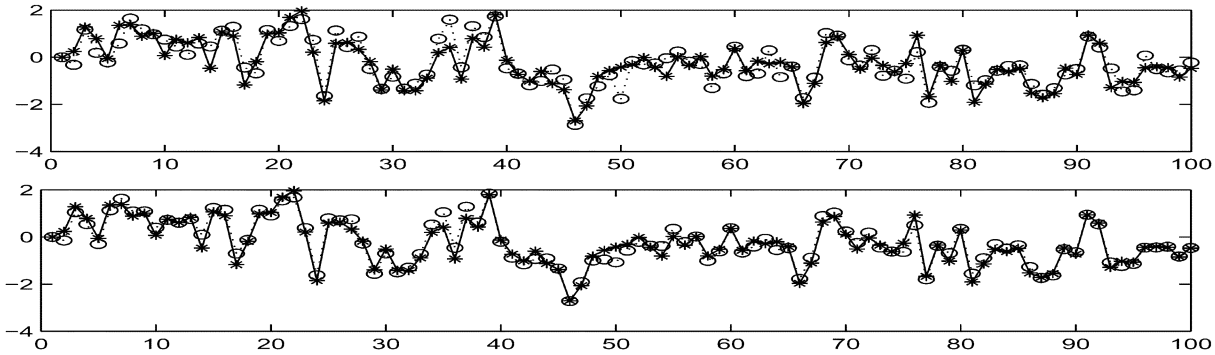


Fig. 5. Snapshots of one factor with the top for p-system and the bottom for i-system. "$*$"-original and "o"-estimated.

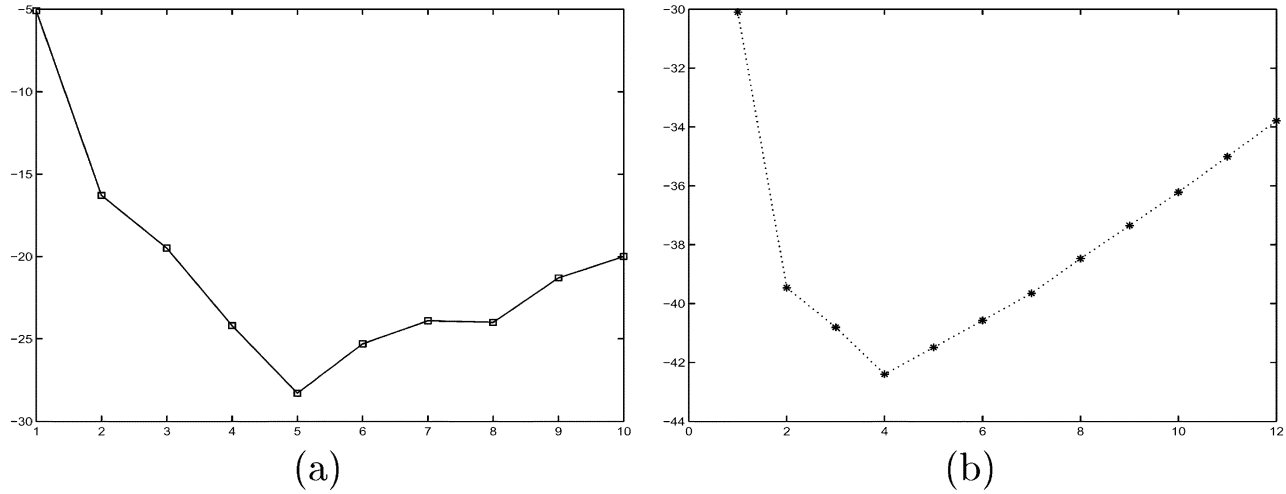


Fig. 6. $J(m)$ curve for selection m with TFA. (a) Simulated data. (b) Financial data.

We perform TFA on both p-system and i-system with m increased from 1 to 10. The obtained $J(m)$ curve by (60) is shown in Fig. 6(a), with the correct factor number 5 detected. Also, learning was made by (70). Started from $D = \text{diag}[7.44, 2.02, 1.71, 1.70, 1.49, 1.45, 1.29]$, we get $D = \text{diag}[9.23, 3.19, 1.96, 1.54, 1.23, 0.03, 0.00]$ as learning tends to converged. That is, the corrected number of 5 factors has been automatically determined during learning.

As discussed in [32], the TFA has also been applied to financial arbitrage pricing theory (APT) on stock data that consists of daily closing prices of 30 Hang Seng Index (HSI) constituents covering the period from January 1, 1998 to December 31, 1999. According to two popular approaches, namely MLFA based LR statistic and eigenvalues analysis, either 11 factors or only one factor is detected, both of which are far from three or four that have been normally agreed by the finance community. In contrast, the minimum of $J(m)$ as shown in Fig. 6(b) provides a reasonable number 4.

VI. Concluding Remarks

The temporal BYY process system and temporal BYY instantaneous system have been rather systematically investigated for the Markovian state space based temporal coding, with a new mechanism that acts via inner representations such that model selection is made either automatically during adaptive learning or subsequently after learning via criteria obtained from this mechanism. Detailed algorithms and criteria have been provided not only on the discrete state featured HMM and extensions but also on the continuous state featured TFA, TNFA, and extensions, with via experimental demonstrations.

Referring the last paragraph in Section II-C, two regularization techniques by Z_q has been ignored in Sections III and IV. However, it is not difficult to add on the role of Z_q. The so-called data smoothing regularization can be added by considering appropriate values of $h = h_x$ in Table I for the cases with $p(\mathbf{y}_t|x_t, \bar{\omega}_t) = \delta(\mathbf{y}_t - \hat{\mathbf{y}}(x_t))$ or both h, h_{yj} in Table I for the cases with $p(\mathbf{y}_t|x_t, \bar{\omega}_t) = G(\mathbf{y}_t|\hat{\mathbf{y}}(x_t), h_{yj}^2 I)$, which acts via being added to the diagonal elements of covariance matrix. Also, the values of h, h_{yj} can be determined as what made in [30]. A normalization regularization can also be imposed by considering Z_q given by (19). In Sections II–V, the consequence of ignoring Z_q in all the algorithms and especially in Tables I and II is that a winner-take-all (WTA) competition is conducted in step (a), which affects the subsequent steps via $\overline{\delta}$ functions such as $\overline{\delta}_{\overline{\ell}_t \ell}$, $\overline{\delta}_{\overline{\ell}_t j}$, $\overline{\delta}_{\overline{i}i}$, and $\overline{\delta}^{(j)}_{\overline{z}_t^{(j)} i}$. The consequence of reconsidering Z_q given by (19) is that these δ functions will be modified such that the WTA competition is replaced by either a soft-competition via the posteriori probabilities $p_t(\ell)$ or a rival penalized competition [41].

Specially, taking the classic HMM, Gaussian HMM, and Gaussian Mixture HMM, HMM gated TFA as examples, we consider to replace $\overline{\delta}_{\overline{\ell}_t \ell}$ by the following soft-competition, shown in (74), at the bottom of the next page. With such replacements, a WTA competition is replaced with a soft-competition via the posteriori probabilities $p_t(\ell)$. As

Hidden Markov Model (HMM) and Extensions				
Classic **HMM**	Gaussian **HMM**	Gaussian Mixture **HMM**	Independent **HMM**	Multi-Independent **HMM**
$(n+k-2)k$	$(n+k-1)k$ $+k\dim(\Sigma_l)$	$k(k-2)+$ $[1+n+\dim(\Sigma_{lz})]\sum_{l=1}^{k} m_l$	$nm+n+\dim(\Sigma)$ $+\sum_{j=1}^{m} k_j(k_j-1)$	$k[k-1+nm+n+\dim(\Sigma)]$ $+\sum_{l=1}^{k}\sum_{j=1}^{m_l} k_{lj}(k_{lj}-1)$
Temporal Factor Analysis (TFA) and Extensions				
TFA-p	TFA-i	HMM gated TFA	TNFA	HMM gated TFA
$nm+m$ $+\dim(\Sigma)$	$nm+m$ $+\dim(\Sigma)$	$k[k-1+n+\dim(\Sigma_l)]$ $+(n+3)\sum_{l=1}^{k} m_l$	$nm+\dim(\Sigma)$ $+3\sum_{j=1}^{m} k_j^2$	$k[k-1+n+\dim(\Sigma_l)]+$ $\sum_{l=1}^{k}(nm_l+m_l+4\sum_{j=1}^{m_l} k_{jl})$

Fig. 7. Typical examples of m_θ, where n, m are dimensions of x, y, and $\dim(\Sigma)$ denotes the number of free parameters in Σ.

explained in [30], [33], this soft-competition can remedy the problem of a WTA competition that makes learning trapped into a local optimal solution. However, it also loses the automatic selection ability as discussed in Section II-B. Instead, model selection should be made via the criteria given in Section III-A.

Substitutions can also be made with a rival penalized competition such that the RPCL learning [41] and BYY harmony learning with normalization [30], [33] become in action. E.g., we can replace $\bar{\delta}_{\bar{\ell}_t\ell}$ with the following general format, shown in (75), at the bottom of the page, where L_κ consists of the first κ labels of $\{1,\ldots,k\}$ that correspond the first $\kappa \leq k$ largest values of Q_ℓ. It becomes again equivalent to the RPCL learning when $\kappa = 2$ [30], [33].

Briefly in [31] and then systematically in [26], BYY harmony learning has been further justified from an information theoretic perspective and a generalized projection geometry, with comparative discussions on its relations to and differences from the studies of not only the minimum message length (MML) [25], Bayesian approach, AIC [3], the minimum description length (MDL) [18], [19], and the bit-back based MDL [11], but also the maximum likelihood, information geometry [1], Helmholtz machines [9] and variational approximation [20]. Moreover, made in [12], [13] on both the Gaussian mixture problem and the factor analysis problem, experiments have shown that the criteria derived from this BYY harmony mechanism outperforms typical existing criteria such as AIC [3], CAIC [4], MDL [18], [19] or equivalently BIC [21]. Though made on data without considering temporal relation, the key issues of these studies apply to temporal BYY harmony learning.

When the size N of samples is quite small, model selection by (7) can be further improved with $J(\mathbf{k})$ replaced by $J_G(\mathbf{k})$ in (8), in which m_{eff} is an effective number of free parameters in the BYY system. In a crude approximation, we can simply let $m_{eff} = m_\theta$ with m_θ being the number of free parameters in θ. Moreover, m_{E_X} is also a number that relates to m_θ and the size N of samples. The detailed discussions are refereed to Section II-B of [26].

Specifically, we can apply the above $J_G(\mathbf{k})$ with $J(\mathbf{k}) = J(k)$ given by (42) for the classic HMM and the Gaussian HMM, $J(\mathbf{k}) = J(k,\{m_\ell\})$ by (43) for the Gaussian mixture HMM, $J(\mathbf{k}) = J(m)$ by (44) for the independent HMM, and $J(\mathbf{k}) = J(k,\{m_\ell\})$ by (45) for the independent channels. We can also apply the above $J_G(\mathbf{k})$ with $J(\mathbf{k}) = J(m)$ given by (60) for TBYY p-system and TBYY i-system, and $J(\mathbf{k}) = J(m,\{m_\ell\})$ given by (65) for the HMM gated TFA, temporal non-Gaussian factor analysis (TNFA), and HMM gated TNFA. Correspondingly, the detailed m_θ for each case is given in Fig. 7.

$$p_t(\ell) = \frac{Q_\ell}{\sum_{\ell=1}^{k} Q_\ell}$$
$$Q_\ell = \begin{cases} q(x_t|\ell)\pi_\ell, & \text{for (46)} \\ G(x_t|\mu_{\ell\bar{j}_t}, \Sigma_{\ell\bar{j}_t})\beta_{\ell\bar{j}_t}\pi_\ell & \text{for (47)} \\ \pi_\ell G(x_t|A_{\ell_t}y_t+\mu_{\ell_t}, \Sigma_{\ell_t})\prod_{j=1}^{m_{\ell_t}} G\left(y_t^{(j)}|b_{\ell_t j}y_{t-1}^{(j)}+\mu_{\ell_t j}, \lambda_{\ell_t j}^2\right), & \text{for (61) \& Tab.1(a).} \end{cases} \tag{74}$$

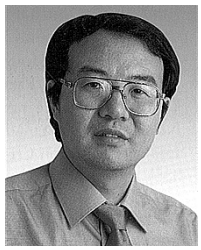
$$p_t(\ell) = \bar{\delta}_{\bar{\ell}_t\ell} - \lambda q_t(\ell), \quad \lambda \geq 0 \text{ is a small number}$$
$$q_t(\ell) = \begin{cases} 0, & \text{(a) same as the case by (74)} \\ \bar{\delta}_{\ell_c\ell}, \ell_c = \max_{\ell\neq\ell_t} Q_\ell, & \text{(b) RPCL learning} \\ \frac{Q_\ell}{\sum_{\ell\in L_\kappa} Q_\ell} & \text{(c) BYY harmony learning with normalization.} \end{cases} \tag{75}$$

ACKNOWLEDGMENT

The author would like to thank Z. Y. Liu and K. C. Chiu for experiments.

REFERENCES

[1] S. Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Netw.*, vol. 8, no. 9, pp. 1379–1408, 1995.
[2] T. W. Anderson and H. Rubin, "Statistical inference in factor analysis," in *Proc. 3rd Berkeley Symp. Mathemtaics Statistical Probabilities*, vol. 5, Berkeley, CA, 1956, pp. 111–150.
[3] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automat. Contr.*, vol. AC–19, pp. 714–723, June 1974.
[4] H. Bozdogan, "Model selection and Akaike's information criterion: The general theory and its analytical extension," *Psychometrika*, vol. 52, pp. 345–370, 1987.
[5] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*. Cambridge, MA: MIT Press, 2001.
[6] R. G. Brown and P. Y. C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*. New York: Wiley, 1997.
[7] K. C. Chiu and L. Xu *et al.*, "Stock price and index forecasting by arbitrage pricing theory-based gaussian TFA learning," in *Intelligent Data Engineering and Automated Learning—IDEAL 2002*, H. yin *et al.*, Eds. New York: Springer-Verlag, 2002, vol. LNCS 2412, pp. 366–371.
[8] G. A. Carrpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vis. Graph. Image Process.*, vol. 37, pp. 54–115, 1987.
[9] P. Dayan *et al.*, "The Helmholtz machine," *Neural Computat.*, vol. 7, no. 5, pp. 889–904, 1995.
[10] T. Edelman and S. Smith, "The geometry of algorithms with orthogonal constraints," *SIAM J. Matrix Anal. Appl.*, vol. 20, pp. 303–353, 1998.
[11] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and Helmholtz free energy," *Advances in NIPS*, vol. 6, pp. 3–10, 1994.
[12] X. L. Hu and L. Xu, "A comparative study of several cluster number selection criteria," in *Intelligent Data Engineering and Automated Learning—IDEAL*, Lecture Notes in Computer Science 2690, 2003, pp. 195–202.
[13] ——, "A comparative investigation on subspace dimension determination," Neural Netw., 2004, to be published.
[14] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. ASME, J Basic Eng.*, March 1960.
[15] M. Kawamoto, "Cerebellum and motor cobtrol," in *The Handbook of Brain Theory and Neural Networks*, Second ed, M. Arbib, Ed. Cambridge, MA: MIT Press, 2002, pp. 190–195.
[16] A. A. Neath and J. E. Cavanaugh, "Regression and time series model selection using variants of the Schwarz information criterion," *Communications Statist. A*, vol. 26, pp. 559–580, 1997.
[17] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[18] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.*, vol. 14, no. 3, pp. 1080–1100, 1986.
[19] ——, "Hypothesis selection and testing by the MDL principle," *Comput. J.*, vol. 42, no. 4, pp. 260–269, 1999.
[20] L. Saul and M. I. Jordan, "Exploiting tractable structures in intractable networks," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 1995, vol. 8, pp. 486–492.
[21] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, pp. 461–464, 1978.
[22] M. Stone, "Cross-validation: A review," *Math. Oper. Statist.*, vol. 9, pp. 127–140, 1978.
[23] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ILL-Posed Problems*. Washington, DC: Winston, 1977.
[24] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer-Verlag, 1995.
[25] C. S. Wallace and D. R. Dowe, "Minimum message length and Kolmogorov complexity," *Comput. J.*, vol. 42, no. 4, pp. 270–280, 1999.
[26] L. Xu, "Advances on BYY harmony learning: Information theoretic perspective, generalized projection geometry, and independent factor autodetermination," *IEEE Trans. Neural Networks.*, vol. 15, pp. 885–902, July 2004.
[27] ——, "Independent component analysis and extensions with noise and time: A Bayesian Ying-Yang learning perspective," *Neural Inform. Process. Lett. Rev.*, vol. 1, no. 1, pp. 1–52, 2003.
[28] ——, "BYY learning, regularized implementation, and model selection on modular networks with one hidden layer of binary units," *Neurocomput.*, vol. 51, pp. 227–301, 2003.
[29] ——, "Data smoothing regularization, multi-sets-learning, and problem solving strategies," *Neural Netw.*, vol. 15, no. 5–6, pp. 817–825, 2003.
[30] ——, "BYY harmony learning, structural RPCL, and topological self-organizing on mixture models," *Neural Netw.*, vol. 15, pp. 1125–1151, 2000a.
[31] ——, "Bayesian Ying Yang harmony Learning," in *The Handbook of Brain Theory and Neural Networks*, Second ed, M. A. Arbib, Ed. Cambridge, MA: MIT Press, 2002b, pp. 1231–1237.
[32] ——, "BYY harmony learning, independent state space and generalized apt financial analyzes," *IEEE Trans. Neural Networks*, vol. 12, pp. 822–849, July 2001.
[33] ——, "Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, three-layer nets and ME-RBF-SVM models," *Int. J. Neural Syst.*, vol. 11, no. 1, pp. 43–69, 2001.
[34] ——, "Temporal BYY learning for state space approach, hidden Markov model and blind source separation," *IEEE Trans. Signal Processing*, vol. 48, pp. 2132–2144, July 2000.
[35] ——, "Temporal Bayesian Ying-Yang dependence reduction, blind source separation and principal independent components," in *Proc. Int. Joint Conf. Neural Networks '99*, vol. 2, Washington, D.C., pp. 1071–1076.
[36] ——, "Bayesian kullback Ying-Yang dependence reduction theory," *Neurocomput.*, vol. 22, no. 1–3, pp. 81–112, 1998.
[37] ——, "Bayesian Ying-Yang system and theory as a unified statistical learning approach: (v) temporal modeling for temporal perception and control," in *Proc. Int. Conf. Neural Information Processing (ICONIP'98)*, vol. 2, Kitakyushu, Japan, Oct. 21–23, 1998, pp. 877–884.
[38] L. Xu, C. C. Cheung, and S.-I. Amari, "Learned parametric mixture based ICA algorithm," *Neurocomput.*, vol. 22, no. 1–3, pp. 69–80, 1998.
[39] L. Xu, "RBF nets, mixture experts, and Bayesian Ying-Yang learning," *Neurocomput.*, vol. 19, no. 1–3, pp. 223–257, 1998.
[40] ——, "A Unified Learning Scheme: Bayesian-Kullback Ying-Yang machine," in *Advances in Neural Information Processing Systems*, D. S. Touretzky *et al.*, Eds. Cambridge, MA: MIT Press, 1995–1996, vol. 8, pp. 444–450.
[41] L. Xu, A. Krzyzak, and E. Oja, "Rival penalized competitive learning for clustering analysis, RBF net and curve detection," *IEEE Trans. Neural Networks*, vol. 4, pp. 636–649, July 1993.
[42] L. Xu, "Least mean square error reconstruction for self-organizing neural-nets," *Neural Netw.*, vol. 6, pp. 627–648, 1991–93.



Lei Xu (SM'94–F'01) received the Ph.D. degree from Tsinghua University, Beijing, China, in 1987.

He is a Chair Professor with the Department of Computer Science and Engineering, the Chinese University of Hong Kong (CUHK), Hong Kong. He joined the National Key Lab on Machine Perception, Peking University, Beijing, China, in 1987, where he became one of ten university-level exceptionally promoted young Associate Professors in 1988 and was exceptionally promoted to a Full Professor in 1992. From 1989 to 1993, he worked at several universities in Finland, Canada, and the United States, including Harvard University and the Massachusetts Institute of Technology, both in Cambridge, MA. He joined CUHK in 1993 as a Senior Lecturer, became a Professor in 1996 and then took the current Chair Professor position in 2002. He has published over 100 academic journal papers, with several well cited contributions to pattern recognition and statistical learning for neural networks. He has given a number of keynote/plenary/invited/tutorial talks in international major neural networks (NN) conferences, such as WCNN, IEEE-ICNN, IJCNN, ICONIP, etc.

Prof. Xu is one of the past Governors of the International Neural Networks Society, a past President of Asia-Pacific NN Assembly, a past Chair of the Com-putational Finance Technical Committee of the IEEE NN Society, and an As-sociate Editor for six international journals on NN, including Neural Networks and the IEEE TRANSACTIONS ON NEURAL NETWORKS from 1994 to 1998. He was an ICONIP '96 Program Committee Chair, a Joint-ICANN-ICONIP '03 Program Committee Co-Chair and a General Chair of IDEAL '98, IDEAL '00, and IEEE CIFER '03. He has served as one of the program committee members in international major NN conferences over the past decade, including the International Joint Conference on Neural Networks, the World Conference on Neural Networks, and the IEEE International Conference on Neural Networks. He has received several Chinese national prestigious academic awards, including the National Nature Science Prize, as well as international awards, including the 1995 INNS Leadership Award. He is a Fellow of the International Association on Pattern Recognition and a Member of the European Academy of Sciences.