ISSN 1673-3460

Volume 6 · Number 2 · June 2011

Frontiers of Electrical and Electronic Engineering in China

高等教育出版社
HIGHER EDUCATION PRESS

Springer

Front. Electr. Electron. Eng. China 2011, 6(2): 300–317
DOI 10.1007/s11460-011-0149-8

RESEARCH ARTICLE

Penghui WANG, Lei SHI, Lan DU, Hongwei LIU, Lei XU, Zheng BAO

Radar HRRP statistical recognition with temporal factor analysis by automatic Bayesian Ying-Yang harmony learning

Abstract Radar high-resolution range profiles (HRRPs) are typical high-dimensional and inter-dimension dependently distributed data, the statistical modeling of which is a challenging task for HRRP-based target recognition. Supposing that HRRP samples are independent and jointly Gaussian distributed, a recent work [Du L, Liu H W, Bao Z. IEEE Transactions on Signal Processing, 2008, 56(5): 1931–1944] applied factor analysis (FA) to model HRRP data with a two-phase approach for model selection, which achieved satisfactory recognition performance. The theoretical analysis and experimental results reveal that there exists high temporal correlation among adjacent HRRPs. This paper is thus motivated to model the spatial and temporal structure of HRRP data simultaneously by employing temporal factor analysis (TFA) model. For a limited size of high-dimensional HRRP data, the two-phase approach for parameter learning and model selection suffers from intensive computation burden and deteriorated evaluation. To tackle these problems, this work adopts the Bayesian Ying-Yang (BYY) harmony learning that has automatic model selection ability during parameter learning. Experimental results show stepwise improved recognition and rejection performances from the two-phase learning based FA, to the two-phase learning based TFA and to the BYY harmony learning based TFA with automatic model selection. In addition, adding many extra free parameters to the classic FA model and thus becoming even worse in identifiability, the model of a general linear dynamical system is even inferior to the classic FA model.

Keywords radar automatic target recognition (RATR), high-resolution range profile (HRRP), temporal factor analysis (TFA), Bayesian Ying-Yang (BYY) harmony learning, automatic model selection

Received March 28, 2011; accepted April 15, 2011

Penghui WANG, Lan DU, Hongwei LIU, Zheng BAO
National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China

Lei SHI, Lei XU (✉)
Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
E-mail: lxu@cse.cuhk.edu.hk

1 Introduction

A high-resolution range profile (HRRP) is the amplitude of coherent summations of the complex time return from target scatterers in each range cell. It contains target structure information, such as target size, scatterer distribution, etc. Therefore, radar HRRP target recognition has received intensive attention from the radar automatic target recognition (RATR) community [1–10].

Statistical recognition methods have been extensively studied and successfully applied to HRRP-based RATR area [2,3,5–8]. By statistical recognition we mean the class membership of a measured HRRP is determined by its posterior probabilities of each class. Previous efforts [2,3,6–8] showed that statistical recognition is superior not only in its excellent performance but also in providing a measure for class prediction, and an opportunity to combine the probabilities with additional information, such as intelligence reports, and other sensor data. Thus, we will concentrate on HRRP-based statistical recognition in this paper.

For HRRP-based statistical recognition, one key problem is to choose an appropriate model that can describe HRRP's statistical property accurately. In the earlier literatures, some simple models such as independent Gaussian [5], independent Gamma [3] and Gaussian-Gamma compounded [6] model were proposed under the assumption that the range cells in an HRRP are mutually independent. Later, Du et al. [7,8] analyzed the statistical characteristics of HRRP from physical mechanism and found the independence assumption among different range cells in an HRRP is inappropriate. They suggested HRRPs follow a joint Gaussian distribution. For high-dimensional HRRP data whose sample size is relatively small, it would be inaccurate to estimate the Gaussian covariance matrix directly. Hence, the authors resorted to some joint Gaussian models with less free parameters, where, factor analysis (FA) obtained the best recognition performance [8].

All of the above models assume that HRRP samples are independent and identically distributed (i.i.d.). In this paper we will illustrate that the HRRP data are highly temporally correlated which motivates us to incorporate this temporal correlation into HRRP statistical modeling. To accomplish this task, we adopt a spatio-temporal model named as temporal factor analysis (TFA) [11–19], which has been widely used to model high-dimensional time series [13–17,20,21]. Briefly, TFA model is an extension of FA by considering the temporal relationship of HRRPs in a hidden state space.

Once a class of parametric models has been selected, we approach to the learning task consisting of parameter learning and model selection [22]. For a given model scale k, the parameters are usually learned by maximum likelihood (ML) method. Nevertheless, the model scale k is often unknown and can hardly be prespecified. A conventional model selection approach is implemented via a two-phase learning. In the first phase, parameter learning, which is usually performed under the ML principle, is repeated on a set of candidate model scales. In the second phase, one model scale k^* is selected among the candidates by a model selection criterion. Examples of such criteria include Akaike's information criterion (AIC) [23–25] and Schwarz's Bayesian inference criterion (BIC) [26].

However, for high-dimensional data like HRRPs, this two-phase learning inevitably suffers from two problems, i.e., huge computation and unreliable evaluation to the criterion, see Ref. [19, Sect. 2.1] for a detailed discussion. To tackle these problems, we adopt the automatic Bayesian Ying-Yang (BYY) harmony learning [13–18], which implements model selection automatically during parameter learning. Based on the TFA model, one specific adaptive BYY algorithm is developed in this paper.

In the recognition experiments based on measured HRRP data, we compare the performance of following models, i.e., the two-phase learning based linear dynamical systems (LDS-BIC) [27–29], the two-phase learning based FA (FA-BIC), the two-phase learning based TFA (TFA-BIC), and the BYY harmony learning based TFA with automatic model selection (auto-TFA-BYY), respectively. The recognition results show incrementally improved performances from LDS-BIC, to FA-BIC, to TFA-BIC, and to TFA-BYY. Moreover, TFA-BYY greatly reduces the computation compared with TFA-BIC. In the rejection experiments, the TFA model shows superior rejection ability over the FA model. These experimental observations verify the accuracy and efficiency of the model and algorithm proposed in this paper. In addition, adding many extra free parameters to the classic FA model and thus becoming even worse in identifiability, the LDS model is even inferior to the classic FA model.

The remaining of this paper is organized as follows. In Sect. 2, we give some background knowledge about HRRP-based target recognition, including an analysis of temporal correlation among HRRPs. Section 3 reviews previous works, and introduces the TFA model and its model selection. In Sect. 4, we further introduce BYY harmony learning for the TFA model with automatic model selection to overcome the drawbacks of two-phase procedure. Based on measured HRRP data, Sect. 5 experimentally shows improved performances of our model and learning algorithm. Finally, conclusions are drawn in Sect. 6.

2 Background for HRRP-based RATR

2.1 Background knowledge of HRRP

For high-resolution radar, the wavelength of radar signal is much smaller than the target size, and the electromagnetism characteristic of the target can be described by the scattering center model [30]. According to this model, a target consists of many scatterers distributed in several range cells along the radar line of sight (LOS). Intuitively, an HRRP can be viewed as a projection of radar returns from all these scatterers onto the radar LOS, as illustrated in Fig. 1. Formally, an HRRP $\boldsymbol{x}$ is the coherent summation amplitudes of the returns from target scatterers in each range cell, defined as follows:

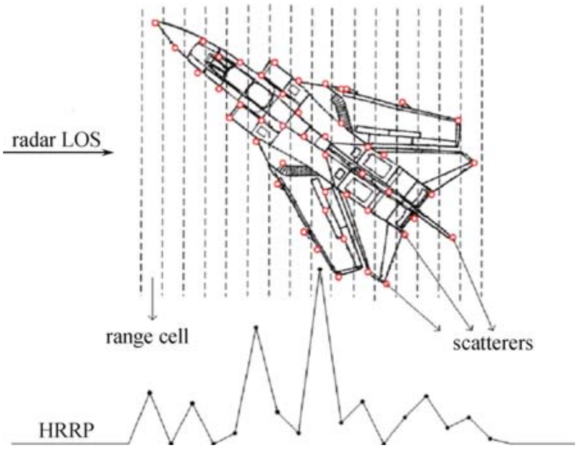


Fig. 1 Diagram of radar HRRP. The radar returns from all scatterers on the target are projected onto the LOS, resulting in an HRRP

$$\boldsymbol{x} = [|x_1|,\ |x_2|, \ldots,\ |x_d|]^{\mathrm{T}},$$
$$\text{with each}\ \ x_i = \sum_{n=1}^{L_i} \sigma_{ni} \mathrm{e}^{\mathrm{j}\phi_{ni}},\ i = 1, 2, \ldots, d, \quad (1)$$

where $|\cdot|$ and $[\cdot]^{\mathrm{T}}$ denote modulus operation and matrix transpose, d is the number of range cells, i.e., the dimensionality of HRRP data. Usually, the measured HRRPs are high-dimensional, e.g., $d = 256$ in Refs. [6–8] and $d = 128$ in this paper. In the ith ($i = 1, 2, \ldots, d$) range cell, there are L_i scatterers. Furthermore, for the nth ($n = 1, 2, \ldots, L_i$) scatter, σ_{ni} and ϕ_{ni} are the reflectivity strength and phase, respectively.

Before implementing the HRRP recognition task, we need to deal with three sensitivity problems of HRRP, namely, translation, amplitude-scale and target-aspect sensitivities [6]. The former two could be settled by translation alignment [1,6] and amplitude-scale normalization [6] preprocessing. As to the target-aspect sensitivity, a common strategy is to partition the consecutive HRRPs into small aspect-frames and build different models for each frame [6–8]. In the following, we assume the HRRPs in each frame have been translation aligned and amplitude-scale normalized.

2.2 Temporal correlation analysis of adjacent HRRPs

According to the scattering center model, when the target aspect changes in a short time, the reflectivities σ_{ni} ($n = 1, 2, \ldots, L_i$; $i = 1, 2, \ldots, d$) of all scatterers are approximately invariant, whereas the phases ϕ_{ni} ($n = 1, 2, \ldots, L_i$; $i = 1, 2, \ldots, d$) may change greatly. The authors in Ref. [6] classified the range cells into three types, where the first type consists of one predominant scatter and a multiple of small scatterers. The amplitudes of these range cells, which depend on the reflectivities of predominant scatterers, keep nearly invariant. Thus this type of range cells in contiguous HRRPs exhibits a high degree of correlation. The other two types contain none or more than one strong scatterer. The great change in scatterer phases leads to a dramatic fluctuation of their amplitudes which indicates poor correlation between them. Usually, the correlation of two adjacent HRRPs is mainly determined by the first type of range cell. To verify this statement, we calculate the average cross-correlation coefficients of measured HRRP data (the data are introduced in Sect. 5). Let $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ be two adjacent HRRPs of a target, and the cross-correlation coefficient ρ_t between them is given as follows:

$$\rho_t = \frac{\left|\boldsymbol{x}_t^{\mathrm{T}} \boldsymbol{x}_{t+1}\right|}{\|\boldsymbol{x}_t\|_2 \|\boldsymbol{x}_{t+1}\|_2}, \quad (2)$$

where t is the discrete time index, and $\|\cdot\|_2$ denotes L_2 norm. As shown in Fig. 2, the cross-correlation coefficients of each two adjacent HRRPs is close to one in most cases, which indicates the high dependence in successive HRRPs. Since the HRRPs are received sequentially, they can be viewed as a high-dimensional time series and we thus call the correlation above as temporal correlation.

Physically, the temporal correlation among HRRPs reflects the change of target spatial structure with time and HRRP sequences from different targets may have different temporal evolution characteristic, so it is important to use this knowledge as a discriminative feature for RATR [4]. In this paper we will make a detailed study on radar HRRP statistical recognition under the prerequisite that the consecutive HRRPs in a frame are temporally correlated.

2.3 Statistical recognition based on HRRP

Radar HRRP statistical recognition aims to classify the HRRPs of unknown targets into different classes using a Bayesian classifier. Considering a sequence of HRRPs $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ which is assumed to come from one of the C classes, a Bayesian classifier assigns $\boldsymbol{x}_t$ ($t = 1, 2, \ldots,\ T$) to the $\hat{c}(\boldsymbol{x}_t)$th class according to the following maximum a posterior (MAP) criterion [8]:

$$\begin{aligned}\hat{c}(\boldsymbol{x}_t) &= \arg\max_{c} p(c|\boldsymbol{x}_t)\\ &= \arg\max_{c}\ [p(\boldsymbol{x}_t|\,c)\,p(c)]\,, c = 1, 2, \ldots, C, \quad (3)\end{aligned}$$

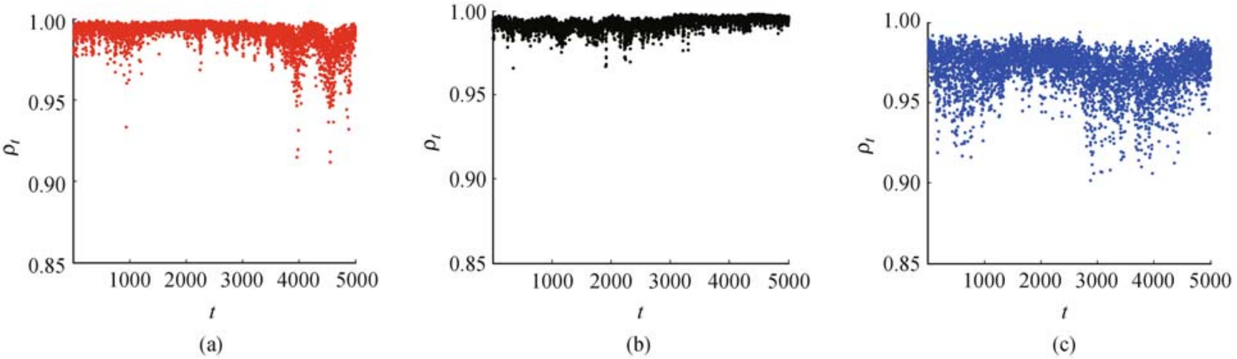


Fig. 2 Cross-correlation coefficients of adjacent HRRPs from three targets. 5001 consecutive HRRPs are randomly extracted from the measured HRRP data of each target. (a) An-26; (b) Cessna; (c) Yark-42

where $p\left(c\left|\boldsymbol{x}_t\right.\right)$ is the posterior probability of $\boldsymbol{x}_t$ being of class c, and $p\left(\boldsymbol{x}_t\right|c)$ and $p(c)$ are likelihood and prior probability for class c, respectively. Generally, equal prior probabilities are adopted, and Eq. (3) turns into

$$\hat{c}\left(\boldsymbol{x}_t\right)=\arg\max_c\, p\left(\boldsymbol{x}_t\right|c),c=1,2,\ldots,C. \tag{4}$$

Thus, the recognition task becomes estimating $p\left(\boldsymbol{x}|c\right)$.

In general, both nonparametric and parametric methods could be used to estimate $p\left(\boldsymbol{x}|c\right)$ [3]. However, nonparametric methods need impractical amounts of samples to guarantee the estimation accuracy [3,31] and are thus impractical for data lying in a high-dimensional space, e.g., HRRPs. This paper focuses on the parametric modeling instead, and $p\left(\boldsymbol{x}|c\right)$ is shortly denoted as $p(\boldsymbol{x})$ in the sequel without further specification.

3 Statistical models and learning methods

3.1 Previous studies on statistical models

Most early works [2,3,5,6] in the area of radar HRRP modeling assumed that the radar returns in each range cell are statistically independent, i.e., $p(\boldsymbol{x})=\prod_{i=1}^{d}p\left(x_i\right)$. For instance, in Refs. [2,3], an independent Gamma model was adopted for HRRP modeling; in Ref. [5], the author employed an independent Gaussian model to describe HRRP data.

Later, Du et al. in Refs. [7,8] theoretically analyzed the spatial correlation characteristic of range cells in HRRP and suggested the usage of joint Gaussian distribution to describe HRRPs. Because the modeling of HRRPs in a frame is a typical small sample problem, the Gaussian covariance matrix estimated directly from HRRPs may be inaccurate even singular. This difficulty was circumvented in Refs. [7,8] by resorting to some parametric models with less free parameters and FA model provided the best recognition performance. For a d-dimensional observable vector $\boldsymbol{x}_t$, FA postulates that $\boldsymbol{x}_t$ is generated via a linear mapping from a low-dimensional hidden variable plus observation noise, i.e.,

$$\begin{aligned}&\boldsymbol{x}_t=\boldsymbol{A}\boldsymbol{y}_t+\boldsymbol{\mu}+\boldsymbol{e}_t,\\&\boldsymbol{y}_t\sim G\left(\boldsymbol{y}_t\right|\boldsymbol{0},\ \boldsymbol{I}_m),\\&\boldsymbol{e}_t\sim G(\boldsymbol{e}_t|\,\boldsymbol{0},\ \boldsymbol{\Psi}),\ \ t=1,2,\ldots,T,\end{aligned} \tag{5}$$

where $\boldsymbol{\mu}$ is the d-dimensional mean vector of $\boldsymbol{x}_t$, $\boldsymbol{y}_t$ represents the m-dimensional hidden variable with $m<d$, $G\left(\boldsymbol{y}_t\right|\boldsymbol{0},\ \boldsymbol{I}_m)$ denotes the Gaussian density of $\boldsymbol{y}_t$ with mean $\boldsymbol{0}$ and identity covariance matrix $\boldsymbol{I}_m$, $\boldsymbol{e}_t\sim G(\boldsymbol{e}_t|\,\boldsymbol{0},\ \boldsymbol{\Psi})$ is observation noise with $\boldsymbol{\Psi}$ being diagonal, $\boldsymbol{A}$ is a $d\times m$ projection matrix that maps the hidden variables to the observation vectors. According to the property of Gaussian variable, we have $p\left(\boldsymbol{x}_t\right)=G\left(\boldsymbol{x}_t\right|\boldsymbol{\mu},\ \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{\Psi}\big)$.

The task of estimating the set $\boldsymbol{\Theta}$ of all the unknown parameters in Eq. (5) is called *parameter learning*, which is usually implemented by the expectation-maximization (EM) algorithm under the ML principle [32]. The other task is selecting an appropriate hidden dimension m, which is usually called *model selection* since enumerating different values of m actually considers a set of candidate FA models of different scales. Model selection is conventionally tackled by a two-phase procedure, a traditional approach is a two-stage implementation, i.e., parameter learning is repeated on a set of candidate hidden dimensionalities among which one is selected by a model selection criterion, e.g., AIC [23–25] and BIC are used by Ref. [33] in a two-phase implementation.

Besides considering the inter-dimensional dependence by an FA model as above, recently paper [34] further considers both non-Gaussian and inter-dimensional dependence simultaneously by a mixture of factor analyzers [35,36] or local factor analysis (LFA), that is, a mixture of a number k of FA models with each FA model having its own parameter set $\boldsymbol{\Theta}_j$ and hidden dimension m_j.. Although learning can be still implemented by the EM algorithm with model selection made in a two-stage implementation, now there is a set $\boldsymbol{k}=\{k,\,\{m_j\}_{j=1}^{k}\}$ of $k+1$ integers to enumerate, which makes a two-stage implementation suffer more serious problems of extensive computation and unreliable estimation of the criterion (see Ref. [19, Sect. 2.1] for a detailed discussion). Instead, BYY harmony learning is used with a favorable nature of automatic model selection, which determines the component number and the hidden dimensionalities of LFA automatically during parameter learning (see the interpretations in Ref. [19], especially its Eqs. (3) and (4)). Experimental results show incremental improvements on recognition accuracy by three implementations, progressively from a two-phase learning based FA, to a two-phase learning based LFA, and then to BYY harmony learning based LFA with automatic model selection.

3.2 Linear dynamical system (LDS) versus TFA

The above efforts all treat the HRRPs in each frame as i.i.d. data. However, as we have analyzed in Sect. 2.2, there is a temporal dependence among HRRPs. Naturally, it arises the question whether Eq. (5) may be extended to take this temporal relation into HRRP modeling, which leads us to the following formulation:

$$\begin{aligned}&\boldsymbol{x}_t=\boldsymbol{A}\boldsymbol{y}_t+\boldsymbol{\mu}+\boldsymbol{e}_t,\\&\boldsymbol{y}_t=\tilde{\boldsymbol{B}}\boldsymbol{y}_{t-1}+\boldsymbol{\omega}_t,\ t=1,2,\ldots,T,\\&\boldsymbol{e}_t\sim G(\boldsymbol{e}_t|\,\boldsymbol{0},\ \boldsymbol{\Psi}),\ \boldsymbol{\omega}_t\sim G\left(\boldsymbol{\omega}_t\right|\boldsymbol{0},\ \boldsymbol{\Omega}),\end{aligned} \tag{6}$$

where time is indexed by discrete t, parameters $\boldsymbol{A}$, $\boldsymbol{\mu}$, $\boldsymbol{\Psi}$ have the same meaning as in Eq. (5), $\boldsymbol{e}_t$ and $\boldsymbol{\omega}_t$ are assumed to be uncorrelated random noises, $\boldsymbol{e}_t$ is assumed to be uncorrelated with $\boldsymbol{y}_t$ and $\boldsymbol{\omega}_t$ is uncorrelated with $\boldsymbol{y}_{t-1}$. Moreover, $\tilde{\boldsymbol{B}}$ is an $m \times m$ transition matrix. The temporal relation is embodied in the first-order vector autoregressive process in the hidden state sequence $\{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$. It should be noted that the model parameters $\boldsymbol{A}$, $\boldsymbol{\mu}$, $\tilde{\boldsymbol{B}}$, $\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}$ are time invariant.

The above formulation has been extensively studied in the literatures of neural networks, machine learning, and signal processing for recent decades. These efforts may be roughly classified by two different types of motivations. One type is featured by regarding Eq. (6) as a general state space model or linear dynamical system and then introducing the EM algorithm for its parameter estimation. It was originally derived by Shumway and Stoffer in Ref. [37]. Around the middle of 1990's, a number of efforts have been made on re-introducing the EM algorithm or making some extensions such as variational approach and Ying-Yang matching [11,27–29,37–43]. Though some of these efforts were made under the name of system identification [28,29], Eq. (6) is generally not identifiable while how special structures or constraints make the model become identifiable and stable is out of consideration in these studies, which is obviously different from those studies of control system theory.

Instead of jumping from FA to a general linear dynamical system, the other type of studies [12–19] originated from Ref. [11] and considered how the FA model in Eq. (5) is extended to take temporal dependence $\boldsymbol{y}_t = \tilde{\boldsymbol{B}}\boldsymbol{y}_{t-1} + \boldsymbol{\omega}_t$ into consideration, while still keeping the original motivation that the cross-dimensional independence of $\boldsymbol{y}_t$ in order to improve the model identifiability. These studies were under the name of TFA or *independent state space* with the following additional requirements:

$$\tilde{\boldsymbol{B}}, \boldsymbol{\Lambda}, \boldsymbol{\Psi} \text{ are all diagonal matrices.} \tag{7}$$

Also, the name of *temporal dependence reduction* (TDR) is used to cover temporal independent factor analysis (TIFA) and temporal independent component analysis (TICA) for a general temporal model beyond the linear relation $\boldsymbol{y}_t = \tilde{\boldsymbol{B}}\boldsymbol{y}_{t-1} + \boldsymbol{\omega}_t$. It should be noticed that letting $\tilde{\boldsymbol{B}} = \boldsymbol{0}$ makes Eqs. (6) and (7) degenerate to be equivalent to the FA model by Eq. (5), while letting $\tilde{\boldsymbol{B}}=\boldsymbol{0}$ in Eq. (6) without Eq. (7) does not necessarily so.

In addition to taking temporal dependence into consideration, Refs. [12–19] also aim at a model with a guaranteed stability and a further improvement on identifiability. Favorably, it has been shown in Sects. III and IV in Ref. [13] that Eq. (6) together with Eq. (7) indeed improves the identifiability of the FA model by Eq. (5) because the notorious rotation indeterminacy of Gaussian FA has been further removed due to $\boldsymbol{y}_t = \tilde{\boldsymbol{B}}\boldsymbol{y}_{t-1} + \boldsymbol{\omega}_t$ with $\tilde{\boldsymbol{B}} \neq \boldsymbol{0}$ known. In Ref. [15], not only the TFA model stability is ensured with each diagonal element $\tilde{b}_i$ of $\tilde{\boldsymbol{B}}$ satisfying $\left|\tilde{b}_i\right| < 1$, but also an identifiable family of TFA structures has been investigated.

The studies in Refs. [27–29,37,41,42] use the EM algorithm to implement the maximum likelihood learning, while the task of selecting an appropriate hidden dimension m of $\boldsymbol{y}_t$ is tackled by a two-stage implementation with help of a model selection criterion such as AIC or BIC. In contrast, the studies in Refs. [12–19] perform BYY harmony learning, by which model selection is made either automatically during learning or still in a two-stage implementation but with an improved selection criterion. The subsequent subsections further introduce details about each of the two types.

3.3 EM algorithm based two-phase learning: LDS vs. TFA

Given a family of parametric models, the task of modeling $p\left(\boldsymbol{x}_t\right)$ turns into estimating model parameters when a model scale m is given, where m is the component number of Gamma mixture model [2,3], and the hidden dimensionality for either FA [8] or TFA model, etc. The parameter set $\boldsymbol{\Theta}_m$ is estimated based on a given set of samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$ under the ML criterion, i.e., $\hat{\boldsymbol{\Theta}}_m = \arg\ \max_{\boldsymbol{\Theta}_m} L\left(\boldsymbol{X}\left|\boldsymbol{\Theta}_m\right.\right)$, which is often implemented by the EM algorithm. The task of selecting an appropriate dimensionality m is tackled by the following two-phase procedure:

Phase 1 Compute $\hat{\boldsymbol{\Theta}}_m = \arg\ \max_{\boldsymbol{\Theta}_m} L\left(\boldsymbol{X}\left|\boldsymbol{\Theta}_m\right.\right)$ via EM algorithm given in Ref. [27] for each $m \in \mathcal{M}$. Here $\mathcal{M}$ is a candidate set for m, a typical choice of which is $[1,\ d-1]$.

Phase 2 Find the optimal

$$m^* = \arg\ \min_m\ J(m),$$
$$J(m) = \begin{cases} -2L\big(\boldsymbol{X}\big|\hat{\boldsymbol{\Theta}}_m\big) + 2D(m), & \text{for AIC,} \\ -2L\big(\boldsymbol{X}\big|\hat{\boldsymbol{\Theta}}_m\big) + (\ln T)D(m), & \text{for BIC,} \end{cases}$$

where $L\big(\boldsymbol{X}\big|\hat{\boldsymbol{\Theta}}_m\big) = \ln q\big(\boldsymbol{X}\big|\hat{\boldsymbol{\Theta}}_m\big)$ is the log-likelihood of $\boldsymbol{X}$ based on ML estimator $\hat{\boldsymbol{\Theta}}_k$ under a given m, and $D(m)$ is the number of free parameters in the model.

In this paper, we investigate the performances of both the LDS by Eq. (6) and TFA jointly by Eqs. (6) and (7), in the following three typical scenarios:

- **LDS-general** Phase 1 uses the EM algorithm given in Ref. [27], with $\boldsymbol{A}, \tilde{\boldsymbol{B}}$ being two general matrices and $\boldsymbol{\Psi}, \boldsymbol{\Omega}$ being two general covariance matrices, while Phase 2 considers $D(m) = dm + mm$ (for $\boldsymbol{A}, \tilde{\boldsymbol{B}}$) $+0.5m(m+1) + 0.5d(d+1)$ (for $\boldsymbol{\Psi}, \boldsymbol{\Omega}$) $+\, d$ (for $\boldsymbol{\mu}$).

For any invertible matrix $\boldsymbol{\Phi}$, we have

$$\begin{aligned} &\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{\Phi}^{-1}\left(\boldsymbol{\Phi}\boldsymbol{y}_t\right) + \boldsymbol{\mu} + \boldsymbol{e}_t, \\ &\left(\boldsymbol{\Phi}\boldsymbol{y}_t\right) = \left(\boldsymbol{\Phi}\tilde{\boldsymbol{B}}\boldsymbol{\Phi}^{-1}\right)\left(\boldsymbol{\Phi}\boldsymbol{y}_{t-1}\right) + \boldsymbol{\Phi}\boldsymbol{\omega}_t. \end{aligned} \tag{8}$$

Let $\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{\Phi}^{-1}$, $\boldsymbol{y}'_t = \boldsymbol{\Phi}\boldsymbol{y}_t$, $\boldsymbol{y}'_{t-1} = \boldsymbol{\Phi}\boldsymbol{y}_{t-1}$, $\tilde{\boldsymbol{B}}' = \boldsymbol{\Phi}\tilde{\boldsymbol{B}}\boldsymbol{\Phi}^{-1}$, $\boldsymbol{\omega}'_t = \boldsymbol{\Phi}\boldsymbol{\omega}_t$, we again get the format $\boldsymbol{x}'_t =$

$\boldsymbol{A}'\boldsymbol{y}'_t + \boldsymbol{\mu} + \boldsymbol{e}_t,\ \boldsymbol{y}'_t = \tilde{\boldsymbol{B}}'\boldsymbol{y}'_{t-1} + \boldsymbol{\omega}'_t$. That is, we have an indeterminacy of any invertible matrix $\boldsymbol{\Phi}$.

- **LDS-constrained** In order to reduce the above indeterminacy and especially to reduce the notorious additive indeterminacy caused by $\boldsymbol{\Psi}$, Phase 1 is added with the following constraint:

$$\boldsymbol{\Psi} \text{ is diagonal, } \boldsymbol{\Omega} = \boldsymbol{I}, \tag{9}$$

for which a slight modification $\boldsymbol{\Psi}^{\text{new}} = \text{diag}\,[\boldsymbol{\Psi}'^{\text{new}}]$ is added after getting a new updating $\boldsymbol{\Psi}'^{\text{new}}$ by the EM algorithm given in Ref. [27]. Moreover, in Phase 2 we consider $D(m) = dm + mm$ (for $\boldsymbol{A},\ \tilde{\boldsymbol{B}}) + d + 1$ (for $\boldsymbol{\Psi},\ \boldsymbol{\Omega}) +$ d (for $\boldsymbol{\mu}$).

- **TFA** We further ensure the cross-dimensional independence of $\boldsymbol{y}_t$ by modifying Eq. (9) into

$$\tilde{\boldsymbol{B}} \text{ is diagonal, } \boldsymbol{\Psi} \text{ is diagonal, and } \boldsymbol{\Omega} = \boldsymbol{I}. \tag{10}$$

In the implementation by the EM algorithm given in Ref. [27], in addition to the above $\boldsymbol{\Psi}^{\text{new}} = \text{diag}\,[\boldsymbol{\Psi}'^{\text{new}}]$, another modification $\boldsymbol{B}^{\text{new}} = \text{diag}\big[\tilde{\boldsymbol{B}}^{\text{new}}\big]$ and $\tilde{\boldsymbol{B}}^{\text{new}} \leftarrow \boldsymbol{B}^{\text{new}}$ is also added after getting the new updating $\tilde{\boldsymbol{B}}^{\text{new}}$. In Phase 2, we let $D(m) = dm$ (for $\boldsymbol{A}) + m$ (for $\tilde{\boldsymbol{B}}) +$ $d + 1$ (for $\boldsymbol{\Psi},\ \boldsymbol{\Omega}) + d$ (for $\boldsymbol{\mu}$).

4 Automatic BYY harmony learning for TFA

4.1 Temporal BYY harmony learning

Firstly proposed in 1995 [44] and systematically developed in the past decade and half [19,45], not only BYY harmony learning theory provides a general statistical learning framework for parameter learning and model selection under a best harmony principle; but also BYY harmony learning on typical structures leads to new model selection criteria, new techniques for implementing regularization and a class of algorithms implement automatic model selection during parameter learning.

Considering that the observation $\boldsymbol{X}$ is generated from its inner representation $\boldsymbol{R} = \{\boldsymbol{Y}, \boldsymbol{\Theta}\}$, where a parameter set $\boldsymbol{\Theta}$ represents the underlying structure of $\boldsymbol{X}$, and $\boldsymbol{Y}$ is the inner representation of $\boldsymbol{X}$ accordingly. Two types of decomposition $p\,(\boldsymbol{X}, \boldsymbol{R}) = p\,(\boldsymbol{R}|\boldsymbol{X})\,p(\boldsymbol{X})$ and $q\,(\boldsymbol{X}, \boldsymbol{R}) = q\,(\,\boldsymbol{X}|\,\boldsymbol{R})\,q(\boldsymbol{R})$ are called Yang machine and Ying machine, respectively. Such a Ying-Yang pair is called a BYY system, as depicted in the left of Fig. 3. The harmony measure is featured by following functional:

$$\begin{aligned} H\,(p||q) &= \int p\,(\,\boldsymbol{R}|\,\boldsymbol{X})\,p(\boldsymbol{X}) \ln\,[q\,(\,\boldsymbol{X}|\,\boldsymbol{R})\,q(\boldsymbol{R})]\,\mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{R} \\ &= \int p(\boldsymbol{\Theta}|\boldsymbol{X}) H\,(p||q, \boldsymbol{\Theta})\,\mathrm{d}\boldsymbol{\Theta}, \\ H\,(p||q, \boldsymbol{\Theta}) &= \int p\,(\,\boldsymbol{Y}|\,\boldsymbol{X})\,p(\boldsymbol{X}) \ln\,[q\,(\,\boldsymbol{X}|\,\boldsymbol{Y})\,q(\boldsymbol{Y})] \\ &\quad \cdot \mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{Y} + \ln\ q(\boldsymbol{\Theta}). \end{aligned} \tag{11}$$

Different from maximizing the likelihood function, an important nature of maximizing $H\,(p||q)$ is that it leads to not only a best matching between the Ying-Yang pair, but also a compact model with a least complexity. Such an ability can be observed from several perspectives, see Sect. 4.1 in Ref. [19]. Here we only introduce one of them due to space limit. On one hand, maximizing $H\,(p||q)$ in Eq. (11) forces Ying machine $q(\boldsymbol{X}, \boldsymbol{R})$ to match Yang machine $p(\boldsymbol{X}, \boldsymbol{R})$. Due to a finite sample size and practical constraints imposed on the Ying-Yang structures, a perfect equality $q(\boldsymbol{X}, \boldsymbol{R}) = p(\boldsymbol{X}, \boldsymbol{R})$ may not be really reached but still be approached as possible as it can. At this equality, $H\,(p||q)$ becomes the negative entropy that describes the complexity of the system. Further maximizing it will decrease the system complexity, which leads to a model selection. Here, it is only a brief introduction. Readers are referred to Ref. [19] for a recent systematic description on the BYY harmony learning.

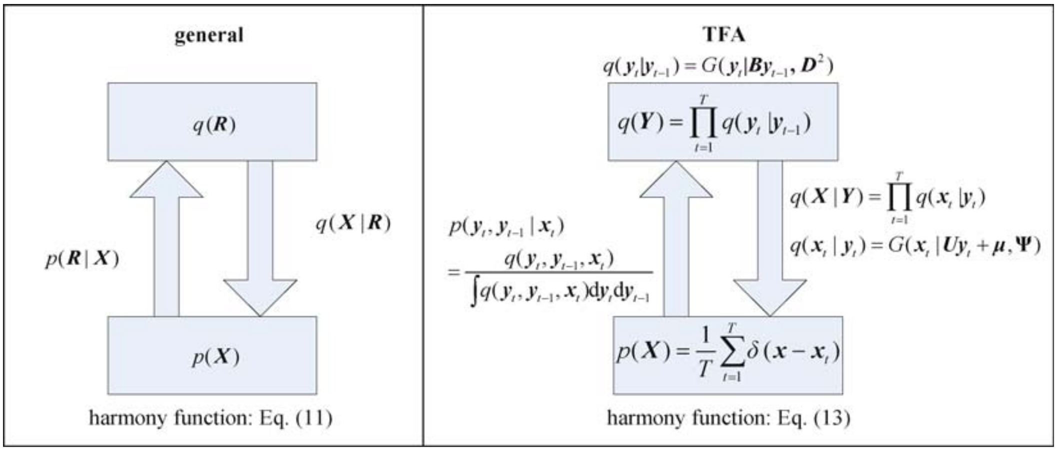

general
$q(\boldsymbol{R})$
$p(\boldsymbol{R}|\boldsymbol{X})$
$q(\boldsymbol{X}|\boldsymbol{R})$
$p(\boldsymbol{X})$
harmony function: Eq. (11)
TFA
$q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) = G(\boldsymbol{y}_t|\boldsymbol{B}\boldsymbol{y}_{t-1}, \boldsymbol{D}^2)$
$q(\boldsymbol{Y}) = \prod_{t=1}^{T} q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})$
$q(\boldsymbol{X}|\boldsymbol{Y}) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{y}_t)$
$q(\boldsymbol{x}_t|\boldsymbol{y}_t) = G(\boldsymbol{x}_t|\boldsymbol{U}\boldsymbol{y}_t + \boldsymbol{\mu}, \boldsymbol{\Psi})$
$p(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}|\boldsymbol{x}_t) = \frac{q(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t)}{\int q(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{x}_t)\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1}}$
$p(\boldsymbol{X}) = \frac{1}{T}\sum_{t=1}^{T}\delta(\boldsymbol{x} - \boldsymbol{x}_t)$
harmony function: Eq. (13)

Fig. 3 BYY system in the general form and specific structures for TFA model

In the sequel, we consider a noninformative prior on $\mathbf{\Theta}$, i.e., $\ln q(\mathbf{\Theta})$ is ignored, and also consider $p\left(\mathbf{\Theta}|\boldsymbol{X}\right)$ in a free structure. Maximizing $H\left(p||q\right)$ with respect to such a $p\left(\mathbf{\Theta}|\boldsymbol{X}\right)$ leads to $p\left(\mathbf{\Theta}|\boldsymbol{X}\right)=\delta\big(\mathbf{\Theta}-\hat{\mathbf{\Theta}}\big)$ with

$$
\begin{aligned}
&\hat{\mathbf{\Theta}}=\arg\ \max\nolimits_{\mathbf{\Theta}} H\left(p||q,\mathbf{\Theta}\right),\\
&H\left(p||q,\mathbf{\Theta}\right)\\
&\quad=\int p\left(\boldsymbol{Y}|\,\boldsymbol{X}\right)p(\boldsymbol{X})\ln\left[q\left(\boldsymbol{X}|\,\boldsymbol{Y}\right)q(\boldsymbol{Y})\right]\mathrm{d}\boldsymbol{X}\mathrm{d}\boldsymbol{Y}.
\end{aligned}
\tag{12}
$$

Given $\boldsymbol{X}=\{\boldsymbol{x}_t\}_{t=1}^{T}$, we get an empirical density $p(\boldsymbol{X})=\dfrac{1}{T}\sum_{t=1}^{T}\delta\left(\boldsymbol{x}-\boldsymbol{x}_t\right)$. This sequence $\boldsymbol{X}=\{\boldsymbol{x}_t\}_{t=1}^{T}$ is considered as generated from the hidden states $\boldsymbol{Y}=\{\boldsymbol{y}_t\}_{t=1}^{T}$, the Ying machine is designed to describe the first order Markovian dependence by $q(\boldsymbol{Y})=\prod_{t=1}^{T}q\left(\boldsymbol{y}_t\,|\boldsymbol{y}_{t-1},\theta_y\right)$ together with the Ying-pathway $q\left(\boldsymbol{X}|\,\boldsymbol{Y}\right)=\prod_{t=1}^{T}q\left(\boldsymbol{x}_t|\,\boldsymbol{y}_t,\theta_{x|y}\right)$. Putting them into Eq. (12), we are lead to a simplified version of Eq. (59) in Ref. [29] at the special case $h=0$, that is, we have

$$
\begin{aligned}
&H\left(p||q,\theta,m\right)=\sum_{t=1}^{T}H_t\left(p||q,\theta,m\right),\\
&H_t\left(p||q,\theta,m\right)=\int p\left(\boldsymbol{y}_t,\,\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t,\theta\right)\\
&\quad\cdot\ \ln[q\left(\boldsymbol{x}_t|\,\boldsymbol{y}_t,\theta_{x|y}\right)q\left(\boldsymbol{y}_t\,|\boldsymbol{y}_{t-1},\theta_y\right)]\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1}.
\end{aligned}
\tag{13}
$$

Moreover, $p\left(\boldsymbol{y}_t,\,\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t,\theta\right)$ is adopted from Eq. (60) in Ref. [19], that is, we have the following Bayesian inverse:

$$
\begin{aligned}
p\left(\boldsymbol{y}_t,\,\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t,\theta\right)&=\frac{q\left(\boldsymbol{y}_t,\boldsymbol{y}_{t-1},\boldsymbol{x}_t\right)}{\displaystyle\int q\left(\boldsymbol{y}_t,\boldsymbol{y}_{t-1},\boldsymbol{x}_t\right)\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1}}\\
&=\frac{q\left(\boldsymbol{x}_t|\,\boldsymbol{y}_t,\theta_{x|y}\right)q\left(\boldsymbol{y}_t\,|\boldsymbol{y}_{t-1},\theta_y\right)}{\displaystyle\int q\left(\boldsymbol{x}_t|\,\boldsymbol{y}_t,\theta_{x|y}\right)q\left(\boldsymbol{y}_t\,|\boldsymbol{y}_{t-1},\theta_y\right)\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1}}.
\end{aligned}
\tag{14}
$$

4.2 BYY harmony learning based TFA with automatic model selection

Traditionally, the studies on the classic FA are all made on the parameterization by Eq. (5). In Ref. [46, Item 9.4], an alternative FA parameterization has been proposed and implemented by the BYY harmony learning, which is featured by that the matrix $\boldsymbol{A}$ is restricted to be rectangular orthogonal matrix and $G\left(\boldsymbol{y}_t|\,\boldsymbol{0},\boldsymbol{I}_m\right)$ is relaxed to be $G\left(\boldsymbol{y}_t|\,\boldsymbol{0},\ \mathbf{\Lambda}\right)$ with a nonnegative diagonal matrix $\mathbf{\Lambda}$. For convenience, we refer the classic one simply by FA-A and the alternative one by FA-B. The two FA parameterizations make no difference on $q\left(\boldsymbol{X}|\mathbf{\Theta}\right)$ and thus are equivalent in term of the ML learning. In contrast, two FA parameterizations become different in term of the BYY harmony learning, as listed in Table 2 of Ref. [16].

Recently, it has been experimentally found in Ref. [47] that the FA-B outperforms the FA-A not only by the variational Bayes learning but also much considerably by the BYY harmony learning based criterion. It can be further understood analytically from the statements around Eqs. (28) and (29) in Ref. [45]. Even importantly, relaxing $\mathbf{\Lambda}=\mathrm{diag}\left[\lambda_1,\lambda_2,\ldots,\lambda_m\right]$ from being forced at $\boldsymbol{I}_m$ in $G\left(\boldsymbol{y}_t|\,\boldsymbol{0},\boldsymbol{I}_m\right)$ to the one in $G\left(\boldsymbol{y}_t|\,\boldsymbol{0},\ \mathbf{\Lambda}\right)$ provides a chance for automatic model selection. The BYY harmony learning drives some $\lambda_j\to 0$ when the jth dimension of $\boldsymbol{y}_t$ is extra. That is, automatic model selection can be made via discarding the jth dimension via checking $\lambda_j\to 0$. Further details are referred to Sect. 2.2 in Ref. [45] for an outline and to Ref. [47] for an extensive empirical comparison.

Actually, TFA jointly by Eqs. (6) and (10) corresponds to the temporal extension of FA-A, thus denoted by TFA-A. First addressed in Ref. [15], we consider a singular value decomposition (SVD) $\boldsymbol{A}=\boldsymbol{U}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{V}$, where $\boldsymbol{D}$ is diagonal, $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}_m$ and $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}=\boldsymbol{I}_m$. Let $\mathbf{\Phi}=\boldsymbol{V}$ or $\mathbf{\Phi}=\boldsymbol{D}\boldsymbol{V}$ in Eq. (8), we are lead to a stable-identifiable family with each one being equivalent to TFA-A. Moreover, the gradient flow for updating a general matrix $\boldsymbol{A}$ is replaced by the orthogonal flows of $\boldsymbol{U}$ and $\boldsymbol{V}$ on the Stiefel manifold, with a good numerical property in computation. Within this family, one instance (i.e., the case (d) given at the bottom on page 474 of Ref. [15]) is particularly recommended in the subsequent studies, (see Eq. (175) in Ref. [16], Eq. (69) and its extension Eq. (66) in Ref. [17]). This instance actually corresponds to the temporal extension of FA-B, thus denoted by TFA-B, with its details rewritten as follows:

$$
\begin{aligned}
&\boldsymbol{x}_t=\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu}+\boldsymbol{e}_t,\ \boldsymbol{y}_t=\boldsymbol{B}\boldsymbol{y}_{t-1}+\boldsymbol{\varepsilon}_t,\ t=1,2,\ldots,T,\\
&\boldsymbol{e}_t\sim G\left(\boldsymbol{e}_t|\,\boldsymbol{0},\ \mathbf{\Psi}\right),\ \ \boldsymbol{\varepsilon}_t\sim G\left(\boldsymbol{\varepsilon}_t|\,\boldsymbol{0},\boldsymbol{D}^2\right),\ \ \boldsymbol{y}_0=\boldsymbol{0}_m,\\
&\boldsymbol{B}=\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{D}^{-1},\ \ \boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}=\boldsymbol{I}_m,\\
&\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}=\boldsymbol{I}_m,\ \ \boldsymbol{D}=\mathrm{diag}\left[d_1,d_2,\ldots,d_m\right],\\
&\tilde{\boldsymbol{B}}=\mathrm{diag}\left[b_1,b_2,\ldots,b_m\right],\ b_j=\frac{\mathrm{e}^{s_j}-\mathrm{e}^{-s_j}}{\mathrm{e}^{s_j}+\mathrm{e}^{-s_j}},
\end{aligned}
\tag{15}
$$

for which we accordingly have

$$
\begin{aligned}
&q\left(\boldsymbol{y}_t\,|\boldsymbol{y}_{t-1},\theta_y\right)=G\left(\boldsymbol{y}_t|\,\boldsymbol{B}\boldsymbol{y}_{t-1},\boldsymbol{D}^2\right),\\
&q\left(\boldsymbol{x}_t|\,\boldsymbol{y}_t,\theta_{x|y}\right)=G\left(\boldsymbol{x}_t|\,\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu},\mathbf{\Psi}\right),\\
&p\left(\boldsymbol{y}_t,\,\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t,\theta\right)=\\
&\quad\frac{G\left(\boldsymbol{x}_t|\,\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu},\mathbf{\Psi}\right)G\left(\boldsymbol{y}_t|\,\boldsymbol{B}\boldsymbol{y}_{t-1},\boldsymbol{D}^2\right)}{\displaystyle\int G\left(\boldsymbol{x}_t|\,\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu},\mathbf{\Psi}\right)G\left(\boldsymbol{y}_t|\,\boldsymbol{B}\boldsymbol{y}_{t-1},\boldsymbol{D}^2\right)\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1}}.
\end{aligned}
\tag{16}
$$

The above Ying-Yang components for the TFA model are illustrated in the right of Fig. 3. Substituting them into Eq. (13) and maximizing $H\left(p||q,\theta,m\right)$ with respect to θ, the nature of least redundancy by the BYY harmony learning (see Sect. 2.2 in Ref. [19]) will provide an intrinsic force to push $d_i\to 0$ if the corresponding hidden dimension $\boldsymbol{y}_t$ is extra and thus discarded. Initialized large enough, m will automatically reduce to an appropriate dimension as learning proceeds.

4.3 A gradient based adaptive learning algorithm

Putting Eq. (16) into Eq. (13), learning algorithm is developed via maximizing $H(p||q,\theta,m)$ with respect to θ, for which past efforts may be roughly outlined as follows:

1) $q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\ \theta_y) \approx q(\boldsymbol{y}_t|\bar{\boldsymbol{y}}_{t-1},\ \theta_y)$ is used and temporal information is carried through $\bar{\boldsymbol{y}}_{t-1}$ recursively by the past value of

$$\bar{\boldsymbol{y}}_t = \arg\max_{\boldsymbol{y}_t}\left[G(\boldsymbol{x}_t|\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu},\boldsymbol{\Psi})\,G\left(\boldsymbol{y}_t|\boldsymbol{B}\bar{\boldsymbol{y}}_{t-1},\boldsymbol{D}^2\right)\right], \tag{17}$$

such that the problem by Eq. (13) reduces into learning FA-B of $G\left(\boldsymbol{y}_t|\boldsymbol{\nu}_t,\boldsymbol{D}^2\right)$ with its mean in a regression structure $\boldsymbol{\nu}_t = \boldsymbol{B}\bar{\boldsymbol{y}}_{t-1}$. Via $\bar{\boldsymbol{y}}_t$, the task is decoupled into adaptively learning a typical FA-B by $G(\boldsymbol{x}_t|\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu},\boldsymbol{\Psi})\,G\left(\boldsymbol{y}_t|\boldsymbol{0},\boldsymbol{D}^2\right)$ and learning a linear regression by $G\left(\boldsymbol{y}_t|\boldsymbol{B}\bar{\boldsymbol{y}}_{t-1},\boldsymbol{D}^2\right)$, e.g., with help of the algorithm given by Eqs. (78)–(80) in Ref. [14]. Further improved algorithms are given in Ref. [18] (e.g., Algorithm III with the option of regression parameterization in Fig. 8) and in Ref. [19] (e.g., the Ying-Yang alternation procedure given in Fig. 8), by updating $p(\boldsymbol{y}_t,\boldsymbol{y}_{t-1}|\theta)$ via some learning regularization for alleviating being stuck at local optimal solutions.

2) $q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\theta_y) \approx q(\boldsymbol{y}_t|\theta_y)$ is used and temporal information is carried recursively by

$$q(\boldsymbol{y}_t|\theta_y) = \int q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\theta_y)\,q(\boldsymbol{y}_{t-1}|\theta_y)\,\mathrm{d}\boldsymbol{y}_{t-1}, \tag{18}$$

by which the problem by Eq. (13) reduces into learning an FA model featured by $G(\boldsymbol{y}_t|\boldsymbol{0},\boldsymbol{\Lambda}_t)$ with $\boldsymbol{\Lambda}_t = \boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}}+\boldsymbol{D}^2$ that is usually no longer diagonal, e.g., with help of the algorithm given by Eqs. (171)–(173) in Ref. [16]. Moreover, a further improved algorithm is given in Ref. [18] (e.g., Algorithm III with the option of marginalization in Fig. 8), with help of updating $p(\boldsymbol{y}_t,\boldsymbol{y}_{t-1}|\theta)$ via some learning regularization.

3) $q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\theta_y)$ is considered in the integral over $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ without approximation. As introduced from Eq. (59) to Eq. (60) in Ref. [19] and also sketched in its Fig. 13, one way is featured by jointly getting

$$\begin{aligned}\left\{\boldsymbol{y}_t^*,\boldsymbol{y}_{t-1}^*\right\} &= \arg\ \max_{\{\boldsymbol{y}_t,\ \boldsymbol{y}_{t-1}\}}\left[G(\boldsymbol{x}_t|\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu},\boldsymbol{\Psi})\right.\\ &\qquad\left.\cdot\,G\left(\boldsymbol{y}_t|\boldsymbol{B}\boldsymbol{y}_{t-1},\boldsymbol{D}^2\right)\right],\\ \left\{\bar{\boldsymbol{y}}_t,\bar{\boldsymbol{y}}_{t-1}\right\} &= \arg\max_{\{\boldsymbol{y}_t,\boldsymbol{y}_{t-1}\}}\left[G(\boldsymbol{x}_t|\boldsymbol{U}\boldsymbol{y}_t+\boldsymbol{\mu},\boldsymbol{\Psi})\right.\\ &\qquad\left.\cdot\,G\left(\boldsymbol{y}_t|\boldsymbol{B}\boldsymbol{y}_{t-1},\boldsymbol{D}^2\right)G(\boldsymbol{y}_{t-1}|\boldsymbol{0},\boldsymbol{\Lambda}_{t-1})\right],\end{aligned} \tag{19}$$

such that the integral over $\boldsymbol{y}_t$ and $\boldsymbol{y}_{t-1}$ is removed for implementing learning. Another algorithm is given by Eqs. (92) and (93) in Ref. [45], featured by a double loop learning procedure that iteratively solves two FA-B problems.

In this paper, we propose another adaptive learning algorithm via analytically computing the integral over $\boldsymbol{y}_t,\ \boldsymbol{y}_{t-1}$. Putting Eq. (16) into Eq. (13) and making the mathematical derivation (see Appendix A), we obtain the following expression of harmony functional:

$$\begin{aligned}&H(p||q,\theta,m) = \sum_{t=1}^{T} H_t(p||q,\theta,m),\\ &H_t(p||q,\theta,m)\\ &= \begin{cases} -\dfrac{1}{2}\Big\{\mathrm{Tr}\left[\left(\boldsymbol{I}_d-\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\Psi}^{-1}\boldsymbol{O}\right] \\ \quad +\ln\left|\boldsymbol{D}^2\right|+\ln|\boldsymbol{\Psi}|\Big\}+\mathrm{const}, & t=1,\\ -\dfrac{1}{2}\Big\{\mathrm{Tr}\left[\boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1}\boldsymbol{O}\right]+\ln\left|\boldsymbol{D}^2\right| \\ \quad +\mathrm{Tr}\left[(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\right]+\ln|\boldsymbol{\Psi}|\Big\} \\ \quad +\mathrm{const}, & t\neq 1,\end{cases}\\ &\boldsymbol{K} = \boldsymbol{C}+\boldsymbol{U}\boldsymbol{B}(\boldsymbol{V}\boldsymbol{D})^{\mathrm{T}} f_t(\tilde{\boldsymbol{B}})\boldsymbol{V}\boldsymbol{D}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}},\\ &f_t(\tilde{\boldsymbol{B}}) = \left(\boldsymbol{I}_m-\tilde{\boldsymbol{B}}^{2(t-1)}\right)\left(\boldsymbol{I}_m-\tilde{\boldsymbol{B}}^2\right)^{-1},\\ &\boldsymbol{C} = \boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{D}(\boldsymbol{U}\boldsymbol{D})^{\mathrm{T}},\ \boldsymbol{O} = (\boldsymbol{x}_t-\boldsymbol{\mu})(\boldsymbol{x}_t-\boldsymbol{\mu})^{\mathrm{T}},\end{aligned} \tag{20}$$

where Tr[·] is the matrix trace, const is a constant term.

To maximize the above $H(p||q,\theta,m)$ with respect θ, a gradient-based adaptive algorithm is sketched in Table 1. The extra structure is removed via checking $d_i \to 0$ and discarding the corresponding dimension y_i (see the line labeled by *** in Table 1).

Before closing, we further consider a variant algorithm. Equation (13) comes from Eq. (12) that considers the whole sequence $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$. Alternatively, with temporal relation via Eq. (18), we may also make the BYY harmony learning on an FA model instantaneously at time t. The former is a typical example of *temporal Bayesian Ying-Yang* process system (TBYY p-system), while the latter is a typical example of *temporal Bayesian Ying-Yang* instantaneous system (TBYY i-system). The two systems are different. Conceptually, TBYY p-system is preferred for a stationary sequence $\boldsymbol{X}$, otherwise TBYY i-system is preferred, e.g., for an HRRP sequence that is not long enough to be stationary well. Readers are referred to Sect. 6.2 of Ref. [16] for more details about TBYY p-system and TBYY i-system.

In this paper, we also make a further investigation on this issue. We consider BYY harmony learning instantaneously at time t, t–1 still with temporal relation via Eq. (18). That is, with $q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\theta_y)$ replaced by $q(\boldsymbol{y}_t,\boldsymbol{y}_{t-1}|\theta_y) = q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\theta_y)\,q(\boldsymbol{y}_{t-1}|\theta_y)$, we have

$$\begin{aligned}H_t(p||q,\theta,m) = \textstyle\int p(\boldsymbol{y}_t,\boldsymbol{y}_{t-1}|\boldsymbol{x}_t,\theta)\ln\left[q\left(\boldsymbol{x}_t|\boldsymbol{y}_t,\theta_{x|y}\right)\right.\\ \left.\cdot\,q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1},\theta_y)\,q(\boldsymbol{y}_{t-1}|\theta_y)\right]\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1},\end{aligned} \tag{21}$$

with $q(\boldsymbol{y}_{t-1}|\theta_y)$ still given by Eq. (18).

Accordingly, Eq. (20) is modified by one additional term that comes from $\int p(\boldsymbol{y}_{t-1}|\theta_y)\ln q(\boldsymbol{y}_{t-1}|\theta_y)\,\mathrm{d}\boldsymbol{y}_{t-1}$, that is, for $t\neq 1$ we have

Table 1 Automatic BYY harmony learning algorithm for TFA model

```
Input: An observed data sequence $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$.
Output: TFA model parameters $\boldsymbol{\Theta} = \left\{\boldsymbol{\mu},\ \boldsymbol{\Psi},\ \boldsymbol{U},\ \boldsymbol{V},\ \boldsymbol{D},\ \tilde{\boldsymbol{B}}\right\}$ together with model scale $m$.
Initialization: Randomly initialize $m$ with large enough value; set iteration index $\tau = 0$ and $H(\tau) = -\infty$.
repeat
  for $t = 1 : T$
  First compute the following temporary variables at time $t$:
  $\boldsymbol{T}_{\boldsymbol{VD}} = \boldsymbol{VD},\ \boldsymbol{T}_{\boldsymbol{UD}} = \boldsymbol{UD},\ \boldsymbol{B} = (\boldsymbol{T}_{\boldsymbol{VD}})^{\mathrm{T}}\, \tilde{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{D}^{-1},\ \boldsymbol{T}_{\boldsymbol{UB}} = \boldsymbol{UB},\ \boldsymbol{O} = (\boldsymbol{x}_t - \boldsymbol{\mu})(\boldsymbol{x}_t - \boldsymbol{\mu})^{\mathrm{T}},\ \boldsymbol{C} = \boldsymbol{\Psi} + \boldsymbol{T}_{\boldsymbol{UD}}(\boldsymbol{T}_{\boldsymbol{UD}})^{\mathrm{T}}.$
  if $t = 1,\ \boldsymbol{\Lambda}_{t-1} = \boldsymbol{0}$;
  else
    $f_t(\tilde{\boldsymbol{B}}) = (\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^{2(t-1)})(\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2)^{-1},\ \boldsymbol{\Lambda}_{t-1} = (\boldsymbol{T}_{\boldsymbol{VD}})^{\mathrm{T}} f_t(\tilde{\boldsymbol{B}})\boldsymbol{T}_{\boldsymbol{VD}},$
    $\boldsymbol{K} = \boldsymbol{C} + \boldsymbol{T}_{\boldsymbol{UB}}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{T}_{\boldsymbol{UB}})^{\mathrm{T}},\ \boldsymbol{\Gamma} = \boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1},\ \boldsymbol{\Delta} = \boldsymbol{K}^{-1}\boldsymbol{O}\boldsymbol{K}^{-1},\ \boldsymbol{P} = \boldsymbol{T}_{\boldsymbol{UB}}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}},\ \boldsymbol{T}_{\boldsymbol{U}} = \boldsymbol{T}_{V} - \boldsymbol{K}^{-1},\ \boldsymbol{\Sigma} = \boldsymbol{D}^2,$
    $\boldsymbol{T}_{\boldsymbol{V}} = \boldsymbol{\Gamma} - \left[\boldsymbol{\Gamma\Sigma K}^{-1} + \left(\boldsymbol{\Gamma\Sigma K}^{-1}\right)^{\mathrm{T}}\right],\ \boldsymbol{T} = \boldsymbol{T}_{\boldsymbol{U}} + \boldsymbol{\Delta},\ \boldsymbol{T}_{\tilde{\boldsymbol{B}}1} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{T}_{\boldsymbol{UB}},\ \boldsymbol{T}_{\tilde{\boldsymbol{B}}2} = (\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2)^{-1}\boldsymbol{T}_{\boldsymbol{VD}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{T}_{\tilde{\boldsymbol{B}}1}(\boldsymbol{T}_{\boldsymbol{VD}})^{\mathrm{T}},$
    $\boldsymbol{T}_{\boldsymbol{V}1} = \tilde{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{T}_{\boldsymbol{UB}})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{T}_{\boldsymbol{UD}},\ \boldsymbol{T}_{\boldsymbol{V}2} = f_t(\tilde{\boldsymbol{B}})\boldsymbol{T}_{\boldsymbol{VD}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{T}_{\tilde{\boldsymbol{B}}1}\boldsymbol{D},\ \boldsymbol{T}_{\boldsymbol{V}3} = \tilde{\boldsymbol{B}}\boldsymbol{T}_{\boldsymbol{VD}}\boldsymbol{T}_{\tilde{\boldsymbol{B}}1}\boldsymbol{\Lambda}_{t-1}\boldsymbol{D}^{-1};$
  end
  Then update TFA model parameters in a gradient manner:
  when $t = 1,$ only update $\boldsymbol{\mu},\ \boldsymbol{\Psi},\ \boldsymbol{U},\ \boldsymbol{D}$.
  $\boldsymbol{\mu}^{\mathrm{new}} = \begin{cases} \boldsymbol{\mu} + \eta\left(\boldsymbol{I}_d - \boldsymbol{C}^{-1}\boldsymbol{UU}^{\mathrm{T}}\right)(\boldsymbol{x}_t - \boldsymbol{\mu}), & t = 1, \\ \boldsymbol{\mu} + \eta\boldsymbol{\Gamma}(\boldsymbol{x}_t - \boldsymbol{\mu}), & t \neq 1, \end{cases}$
  $\boldsymbol{\Psi}^{\mathrm{new}} = \mathrm{diag}\left[(1-\eta)\boldsymbol{\Psi} + \eta\boldsymbol{\Psi}\boldsymbol{G}_{\boldsymbol{\Psi}}\boldsymbol{\Psi}\right],\ \boldsymbol{G}_{\boldsymbol{\Psi}} = \begin{cases} -\boldsymbol{C}^{-1}\boldsymbol{UU}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{O}\boldsymbol{C}^{-1} + \boldsymbol{\Psi}^{-1}\boldsymbol{O}\left(\boldsymbol{I}_m - \boldsymbol{C}^{-1}\boldsymbol{UU}^{\mathrm{T}}\right)\boldsymbol{\Psi}^{-1}, & t = 1, \\ -\boldsymbol{T}, & t \neq 1, \end{cases}$
  diag[·] sets all off-diagonal elements of a matrix to zero.
  Update $\boldsymbol{U}$ and $\boldsymbol{V}$ by gradient on the Stiefel manifold (see Ref. [48] for details):
  $\boldsymbol{U}^{\mathrm{new}} = \boldsymbol{U} + \eta\left(\boldsymbol{G}_{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{G}_{\boldsymbol{U}}^{\mathrm{T}}\boldsymbol{U}\right),$ with $\boldsymbol{G}_{\boldsymbol{U}} = \begin{cases} -\boldsymbol{C}^{-1}\left(\boldsymbol{O}\boldsymbol{\Psi}^{-1}\boldsymbol{UU}^{\mathrm{T}} + \boldsymbol{UU}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{O}\right)\boldsymbol{C}^{-1}\boldsymbol{U\Sigma} \\ \quad -\boldsymbol{C}^{-1}\boldsymbol{O}\boldsymbol{\Psi}^{-1}\boldsymbol{U} - \boldsymbol{\Psi}^{-1}\boldsymbol{O}\boldsymbol{C}^{-1}\boldsymbol{U}, & t = 1, \\ -\left(\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{P} + \boldsymbol{T}\boldsymbol{T}_{\boldsymbol{UD}}\boldsymbol{D}\right), & t \neq 1, \end{cases}$
  $\boldsymbol{V}^{\mathrm{new}} = \boldsymbol{V} + \eta\left(\boldsymbol{G}_{\boldsymbol{V}} - \boldsymbol{V}\boldsymbol{G}_{\boldsymbol{V}}^{\mathrm{T}}\boldsymbol{V}\right),$ with $\boldsymbol{G}_{\boldsymbol{V}} = -\left(\boldsymbol{T}_{\boldsymbol{V}1} + \boldsymbol{T}_{\boldsymbol{V}2} + \boldsymbol{T}_{\boldsymbol{V}3}\right),$
  $\boldsymbol{D}^{\mathrm{new}} = (1-\eta)\boldsymbol{D} + \eta\,\mathrm{diag}\left(\boldsymbol{D}\boldsymbol{G}_{\boldsymbol{D}}\boldsymbol{D}\right),$ with $\boldsymbol{G}_{\boldsymbol{D}} = \begin{cases} -\boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{UU}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{O}\boldsymbol{C}^{-1}\boldsymbol{UD}, & t = 1, \\ -\left[(\boldsymbol{T}_{\boldsymbol{UD}})^{\mathrm{T}}\boldsymbol{T} + \boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\left(\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{T}_{\boldsymbol{UB}}\right)^{\mathrm{T}}\right]\boldsymbol{U}, & t \neq 1, \end{cases}$
  if $d_i \to 0$ then discard hidden dimension $y_i$ and let $m^{\mathrm{new}} = m - 1$;          ***
  Let $b_i = \dfrac{\mathrm{e}^{s_i} - \mathrm{e}^{-s_i}}{\mathrm{e}^{s_i} + \mathrm{e}^{-s_i}}$, where $b_i$ and $s_i$ are the $i$th diagonal elements of diagonal matrices $\tilde{\boldsymbol{B}}$ and $\boldsymbol{S}$, respectively.
  We update $b_i$ indirectly via $s_i$ to ensure the stability of the model.
  $\boldsymbol{S}^{\mathrm{new}} = \mathrm{diag}\left[\boldsymbol{S} + \eta\boldsymbol{G}_{\tilde{\boldsymbol{B}}}(\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2)\right],$ with $\boldsymbol{G}_{\tilde{\boldsymbol{B}}} = -\boldsymbol{T}_{\boldsymbol{VD}}\boldsymbol{T}_{\tilde{\boldsymbol{B}}1}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{T}_{\boldsymbol{VD}})^{-1} + \boldsymbol{T}_{\tilde{\boldsymbol{B}}2}\left[(t-1)\tilde{\boldsymbol{B}}^{(2t-3)} - f_t(\tilde{\boldsymbol{B}})\tilde{\boldsymbol{B}}\right];$
  end for
Calculate $H(\tau)$ by Eq. (20), if $|H(\tau) - H(\tau-1)| < \xi|H(\tau-1)|$, terminate the algorithm, else let $\tau^{\mathrm{new}} = \tau + 1$,
continue the algorithm. In our implementation, $\xi$ is set as $10^{-5}$.
```

$$
\begin{aligned}
&H_t\left(p||q, \theta, m\right) \\
&= -\frac{1}{2}\Big\{\ln\left|\boldsymbol{D}^2\right| + \ln\left|\boldsymbol{\Lambda}_{t-1}\right| + \mathrm{Tr}\left[\boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1}\boldsymbol{O}\right] \\
&\quad + \mathrm{Tr}\left[(\boldsymbol{UB})^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{UB}\boldsymbol{\Lambda}_{t-1}\right] + \ln\left|\boldsymbol{\Psi}\right|\Big\} + \mathrm{const}.
\end{aligned} \tag{22}
$$

In Table 1, the gradients of $\boldsymbol{V},\ \boldsymbol{D},\ \tilde{\boldsymbol{B}}$ (when $t \neq 1$) are thus changed into

$$
\begin{aligned}
\boldsymbol{G}_{\boldsymbol{V}} &= -\left(\boldsymbol{T}_{\boldsymbol{V}1} + \boldsymbol{T}_{\boldsymbol{V}2} + \boldsymbol{T}_{\boldsymbol{V}3} + f_t\big(\tilde{\boldsymbol{B}}\big)\boldsymbol{T}_{\boldsymbol{VD}}\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{D}\right), \\
\boldsymbol{G}_{\boldsymbol{D}} &= -[(\boldsymbol{T}_{\boldsymbol{UD}})^{\mathrm{T}}\boldsymbol{T} + \boldsymbol{B}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{T}_{\boldsymbol{UB}})^{\mathrm{T}}]\boldsymbol{U} \\
&\quad -\boldsymbol{V}^{\mathrm{T}}f_t(\tilde{\boldsymbol{B}})\boldsymbol{VD}\boldsymbol{\Lambda}_{t-1}^{-1}, \\
\boldsymbol{G}_{\tilde{\boldsymbol{B}}} &= \left[\boldsymbol{T}_{\tilde{\boldsymbol{B}}\boldsymbol{2}} - \left(\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2\right)^{-1}\boldsymbol{T}_{\boldsymbol{VD}}\boldsymbol{\Lambda}_{t-1}^{-1}\left(\boldsymbol{T}_{\boldsymbol{VD}}\right)^{\mathrm{T}}\right] \\
&\quad \cdot\left[(t-1)\tilde{\boldsymbol{B}}^{(2t-3)} - f_t\big(\tilde{\boldsymbol{B}}\big)\tilde{\boldsymbol{B}}\right] - \boldsymbol{T}_{\boldsymbol{VD}}\boldsymbol{T}_{\tilde{\boldsymbol{B}}1} \\
&\quad \cdot\boldsymbol{\Lambda}_{t-1}\left(\boldsymbol{T}_{\boldsymbol{VD}}\right)^{-1}.
\end{aligned} \tag{23}
$$

5 Experimental results

5.1 Data description

The experiments presented in this paper are based on the same measured data of three planes as in Refs. [8,34]. The parameters of radar and planes are given in Tables 2 and 3 and the projections of plane trajectories onto ground plane are segmented as displayed in Fig. 4. We take the 5th and 6th segments of An-26, the 6th and 7th segments of Cessna and the 2nd and 5th segments of Yark-42 as training samples, while the remaining data are left for testing. These training data almost cover all of the target-aspect angles. There are 25600 HRRPs in each segment except the 5th segment of Yark-42 which has 10240 HRRPs. The training data are divided into frames by the equal interval partition method with each frame containing 1024 HRRPs. Thus, there are 50/50/35 frames for An-26/Cessna/Yark-42, respectively.

The HRRPs of three targets are measured at different time and their signal-to-noise ratios (SNRs) are slightly different. To avoid the use of SNR as discriminative information, we discard part of range cells containing only noise and the dimensionality of the truncated HRRP is 128.

Table 2 Parameters of radar

parameters	values
center frequency/MHz	5520
bandwidth/MHz	400
pulse repetition frequency/Hz	400

Table 3 Parameters of planes

plane type	length/m	width/m	height/m
Yark-42	36.38	34.88	9.83
An-26	23.80	29.20	9.83
Cessna	14.40	15.90	4.57

5.2 Recognition performance

In this section, we implement two types of recognition experiments. The first type is based on the LDS, FA and TFA models with the two-phase model selection, whereas the second type is LFA and TFA models learned automatically by BYY learning. Since in Ref. [8] BIC shows better performance than AIC, only BIC is considered here for the two-phase model selection. To make the final decision by Eq. (4), we compute the likelihood of each testing sequence $\boldsymbol{X}=\{\boldsymbol{x}_1,\ \boldsymbol{x}_2,...,\boldsymbol{x}_N\}$ by following equation (see Appendix B for mathematical derivation):

$$
\begin{aligned}
&p(\boldsymbol{X}|c)=\prod_{t=2}^{N}p\left(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}\right)p(\boldsymbol{x}_1),\\
&p\left(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}\right)=\begin{cases}
\dfrac{p\left(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}\right)}{p\left(\boldsymbol{x}_{t-1}\right)}=\dfrac{G\left(\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{x}_{t-1}\end{bmatrix}\middle|\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix},\begin{bmatrix}\boldsymbol{U}\boldsymbol{\Lambda}_t\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{\Psi}, & \boldsymbol{\Sigma}_{t,t-1}\\ \boldsymbol{\Sigma}_{t,t-1}^{\mathrm{T}}, & \boldsymbol{U}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{\Psi}\end{bmatrix}\right)}{G\left(\boldsymbol{x}_{t-1}|\,\boldsymbol{\mu},\boldsymbol{U}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{\Psi}\right)}\\
\qquad=G\left(\boldsymbol{x}_t|\,\boldsymbol{F}\boldsymbol{x}'_{t-1}+\boldsymbol{\mu},\ \boldsymbol{W}\right),\ \text{with } \boldsymbol{x}'_{t-1}=\boldsymbol{x}_{t-1}-\boldsymbol{\mu}, & \text{for TFA model,}\\
\dfrac{p(\boldsymbol{x}_t,\boldsymbol{x}_{t-1})}{p(\boldsymbol{x}_{t-1})}=\dfrac{G\left(\begin{bmatrix}\boldsymbol{x}_t\\ \boldsymbol{x}_{t-1}\end{bmatrix}\middle|\begin{bmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{bmatrix},\begin{bmatrix}\boldsymbol{A}\boldsymbol{\Lambda}_t\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{\Psi}, & \boldsymbol{\Sigma}_{t,t-1}\\ \boldsymbol{\Sigma}_{t,t-1}^{\mathrm{T}}, & \boldsymbol{A}\boldsymbol{\Lambda}_{t-1}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{\Psi}\end{bmatrix}\right)}{G\left(\boldsymbol{x}_{t-1}|\,\boldsymbol{\mu},\boldsymbol{A}\boldsymbol{\Lambda}_{t-1}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{\Psi}\right)}\\
\qquad=G\left(\boldsymbol{x}_t|\,\boldsymbol{F}\boldsymbol{x}'_{t-1}+\boldsymbol{\mu},\ \boldsymbol{W}\right), & \text{for LDS model,}\\
p(\boldsymbol{x}_t)\qquad=G\left(\boldsymbol{x}_t|\,\boldsymbol{\mu},\ \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{\Psi}\right), & \text{for FA model,}
\end{cases}\\
&\boldsymbol{\Lambda}_t=\begin{cases}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}}+\boldsymbol{D}^2, & \text{for TFA model,}\\ \tilde{\boldsymbol{B}}\boldsymbol{\Lambda}_{t-1}\tilde{\boldsymbol{B}}^{\mathrm{T}}+\boldsymbol{\Omega}, & \text{for LDS model,}\end{cases}\\
&\boldsymbol{\Sigma}_{t,t-1}=\mathrm{E}(\boldsymbol{x}_t-\boldsymbol{\mu})(\boldsymbol{x}_{t-1}-\boldsymbol{\mu})^{\mathrm{T}}=\begin{cases}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^{\mathrm{T}}, & \text{for TFA model,}\\ \boldsymbol{A}\tilde{\boldsymbol{B}}\boldsymbol{\Lambda}_{t-1}\boldsymbol{A}^{\mathrm{T}}, & \text{for LDS model,}\end{cases}\\
&\boldsymbol{F}=\begin{cases}\boldsymbol{\Sigma}_{t,t-1}(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^{\mathrm{T}})^{-1}, & \text{for TFA model,}\\ \boldsymbol{\Sigma}_{t,t-1}(\boldsymbol{\Psi}+\boldsymbol{A}\boldsymbol{\Lambda}_{t-1}\boldsymbol{A}^{\mathrm{T}})^{-1}, & \text{for LDS model,}\end{cases}\\
&\boldsymbol{W}=\begin{cases}\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\mathrm{T}}+(\boldsymbol{U}\boldsymbol{B}-\boldsymbol{F}\boldsymbol{U})\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}, & \text{for TFA model,}\\ \boldsymbol{\Psi}+\boldsymbol{A}\boldsymbol{\Omega}\boldsymbol{A}^{\mathrm{T}}+(\boldsymbol{A}\tilde{\boldsymbol{B}}-\boldsymbol{F}\boldsymbol{A})\boldsymbol{\Lambda}_{t-1}(\boldsymbol{A}\tilde{\boldsymbol{B}})^{\mathrm{T}}, & \text{for LDS model.}\end{cases}
\end{aligned}
\tag{24}
$$

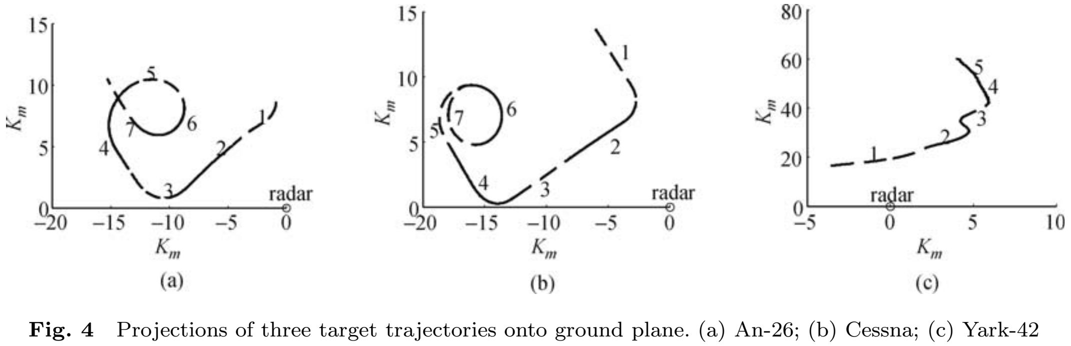


Fig. 4 Projections of three target trajectories onto ground plane. (a) An-26; (b) Cessna; (c) Yark-42

where N is the length of each testing sequence, and here we select $N = 1, 2, 3, 4$. For a more reliable evaluation, each experiment is repeated 50 times with randomly initialized parameters.

The mean and standard deviation values of average correct recognition rate (ACRR) and the average time cost per frame by different models are listed in Table 4. The results in Table 4 verify our analysis in Sects. 3 and 4 which are summarized as follows:

1) Compared with the FA model, the LDS model [27–29] has too many extra free parameters and thus becoming even worse in identifiability, thus the recognition performance of the LDS model is actually inferior to that of the classic FA model, and its performance will be improved via adding constraints on the free parameters, which is confirmed by the first three rows of Table 4.

2) The TFA model [11–19] extends the FA model to the temporal case, and improves the identifiability of the FA model (see Sects. III and IV in Ref. [13]), by removing the notorious rotation indeterminacy, thus it outperforms the FA model, which is confirmed by the two rows located at the center of Table 4.

3) In the testing phase, the TFA model not only combines the information carried by each individual HRRP just as the FA model dose, but also further exploits the temporal dependence among HRRPs which can be viewed as an advanced form of information combination. This is another reason why the performance of the TFA model is superior to that of the FA model.

4) Owing to its unique model selection ability, similar to Ref. [34], the BYY harmony learning based TFA further outperforms the two-phase learning based TFA in both estimation accuracy and computational efficiency.

5) When there is no or only a little temporal dependence ($N < 3$) existing in the testing sequence, the LFA-BYY model, capturing the non-Gaussian property of HRRPs, shows better performance than others; however, with the increase in the length of testing sequence, more temporal dependence is used as discriminating information, which makes the TFA-BYY model prevail over the LFA-BYY model.

Also, we observe that TFA-BYY-i slightly outperforms TFA-BYY-p, which indicates that HRRP sequences are not long enough to be well stationary and thus TBYY i-system is more preferred. In addition, the performance of the FA model in the above table appears not as good as in Refs. [8,34]. The reason is previously given at the end of Sect. 5.1, that is, this paper uses the truncated HRRPs to remove those superficial discriminative contributions due to that the HRRPs of three targets are measured at different time and SNR.

5.3 Rejection of unknown target

Assessing the performance of statistical models should not simply be relied on ACRRs. An important issue in radar target recognition is how to distinguish in-class targets and out-of-class targets, i.e., the so-called confusers. When we cannot guarantee that all testing HRRP sequences belong to the training set classes, rejecting those HRRPs with a low degree of membership to these classes becomes important. In this experiment, 18000 HRRPs generated by simulation software — XPATCH, are adopted as the confusers. To graphically compare the rejection ability of the TFA and FA model, we adopt the receiver operating characteristics (ROC) curve in the detection theory [49]. For a given a detection threshold γ, we consider two evaluation indexes as follows:

1) Detection probability P_{d} is the percentage of in-class targets correctly classified;

2) False alarm probability P_{f} is the percentage of confusers wrongly classified as in-class targets. Both P_{d} and P_{f} can be computed via the following equations:

$$P_{\mathrm{f}} = \frac{M_1}{N_1},\ P_{\mathrm{d}} = \frac{M_2}{N_2}, \tag{25}$$

where$N_1$is the total number of in-class testing sequences, N_2 is the total number of confuser sequences, M_1 and M_2 are the numbers of in-class testing sequences and confuser sequences whose likelihoods are larger than γ.

Table 4 Mean and standard deviation values of ACRRs (in percentage), training time cost per frame (in minute) and testing time cost per sequence (in second, $N = 3$)

model	model description	ACRR/%				training	testing time/s
		N=1	N=2	N=3	N=4	time/min	(N=3)
LDS-G-BIC Eq. (6)	$\boldsymbol{\Psi}$, $\boldsymbol{\Omega}$ are general covariance matrices	91.4±3.6	91.8±3.4	92.5±3.1	92.9±2.9	72.5	3.3
LDS-C-BIC	$\boldsymbol{\Omega} = \boldsymbol{I}, \boldsymbol{\Psi}$ is diagonal	91.6±3.5	92.1±3.3	93.0±3.0	93.3±2.6	69.7	3.3
FA-BIC	$\boldsymbol{\Omega} = \boldsymbol{I}, \tilde{\boldsymbol{B}} = \boldsymbol{0}, \boldsymbol{\Psi}$ is diagonal	92.3±3.3	93.1±3.0	94.1±2.6	94.5±2.5	4.9	0.4
TFA-BIC	$\boldsymbol{\Omega} = \boldsymbol{I}, \tilde{\boldsymbol{B}}, \boldsymbol{\Psi}$ are diagonal	92.2±3.0	93.5±2.7	94.6±2.6	95.2±2.2	62.0	3.3
LFA-BYY	Eq. (3) in Ref. [34]	93.7±3.1	94.6±2.8	95.0±2.6	95.4±2.5	22.4	2.1
TFA-BYY-p Eq. (16)	$\boldsymbol{\Omega} = \boldsymbol{I}, \tilde{\boldsymbol{B}}, \boldsymbol{\Psi}$ are diagonal, with Eq. (20)	92.9±2.8	94.4±2.4	95.3±2.1	96.1±1.9	15.2	3.3
TFA-BYY-i Eq. (18)	$\boldsymbol{\Omega} = \boldsymbol{I}, \tilde{\boldsymbol{B}}, \boldsymbol{\Psi}$ are diagonal, with Eqs. (18) and (22)	93.0±2.8	94.5±2.3	95.5±2.0	96.2±1.8	15.2	3.3

By changing γ, we will get an ROC curve by plotting P_{d} against P_{f} as shown in Fig. 5. The TFA models always achieve a higher P_{d} than the FA-BIC model for a same P_{f}. The good rejection performance of TFA model can be attributed to the fact that the temporal correlation utilized acts as a discriminant to pull the in-class targets out of the confusers. Moreover, the TFA-BYY is further superior to TFA-BIC which is mainly due to the model selection ability of BYY learning.

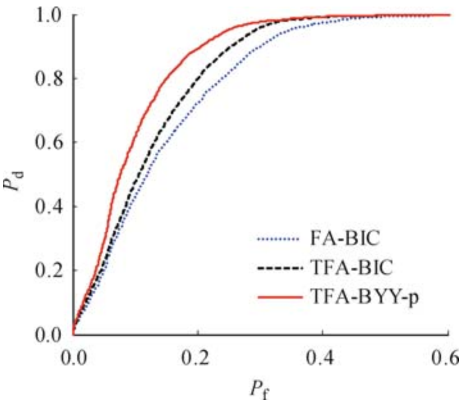


Fig. 5 ROC curves of FA-BIC, TFA-BIC and TFA-BYY-p models (N=3). Since the TFA-BYY-p and TFA-BYY-i obtain the similar ROC curves, only the ROC curve of TFA-BYY-p is given here

6 Conclusions

Existing statistical models for HRRP-based radar target recognition assume that HRRPs are temporally independent, whereas theoretical analysis and experimental results based on measured data show the independence assumption regarding the HRRPs is inappropriate. To incorporate the temporal correlation between adjacent HRRPs, this paper adopts the TFA for the modeling task. Moreover, to tackle the two problems of the conventional two-phase approach for model selection, i.e., huge computation and unreliable evaluation, the BYY harmony learning is employed with model selection implemented automatically during parameter learning. Experimental results show incrementally improved performances from the two-phase learning based LDS, to the two-phase learning based FA, further to the two-phase learning based TFA and to the BYY harmony learning based TFA with automatic model selection. First, TFA obtains better recognition and rejection performances than FA due to its integration of temporal correlation among HRRPs. Second, the BYY harmony learning based TFA with automatic model selection outperforms models obtained by a two-phase learning, evaluated by both recognition accuracy and time cost. Moreover, TFA-BYY-p slightly outperforms TFA-BYY-i. In addition, adding many extra free parameters to the classic FA model and thus becoming even worse in identifiability, the LDS model is actually inferior to the classic FA model.

Appendix A Derivation of Eq. (20) from Eq. (13)

We compute $p(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}|\boldsymbol{x}_t) = p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_t)\, p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{x}_t)$ from

$$p(\boldsymbol{y}_{t-1}|\boldsymbol{x}_t) = \frac{q(\boldsymbol{x}_t, \boldsymbol{y}_{t-1})}{q(\boldsymbol{x}_t)} \text{ and } p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{x}_t) = \frac{q(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{y}_{t-1})}{q(\boldsymbol{x}_t|\boldsymbol{y}_{t-1})}.$$

We start from

$$\begin{aligned} q(\boldsymbol{x}_t|\boldsymbol{y}_{t-1}) &= \int q(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{y}_{t-1})\,\mathrm{d}\boldsymbol{y}_t = \int q(\boldsymbol{x}_t|\boldsymbol{y}_t)\, q(\boldsymbol{y}_t|\boldsymbol{y}_{t-1})\,\mathrm{d}\boldsymbol{y}_t \\ &= \int \frac{\exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{x}_t - \boldsymbol{\mu} - \boldsymbol{U}\boldsymbol{y}_t\right)^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\left(\boldsymbol{x}_t - \boldsymbol{\mu} - \boldsymbol{U}\boldsymbol{y}_t\right) + \left(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{y}_{t-1}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{y}_{t-1}\right)\right]\right\}}{(2\pi)^{(d+m)/2}(|\boldsymbol{\Psi}||\boldsymbol{\Sigma}|)^{1/2}}\mathrm{d}\boldsymbol{y}_t \\ &= \frac{\exp\left[-\frac{1}{2}\left(\boldsymbol{x}'_t - \boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1}\right)^{\mathrm{T}}\left(\boldsymbol{\Psi} + \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right)^{-1}\left(\boldsymbol{x}'_t - \boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1}\right)\right]}{(2\pi)^{d/2}|\boldsymbol{\Psi} + \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}|^{1/2}} \\ &= G\left(\boldsymbol{x}_t|\boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right), \end{aligned}$$

where $\boldsymbol{x}'_t = \boldsymbol{x}_t - \boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \boldsymbol{D}^2$.

We further have $p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}, \boldsymbol{x}_t) = \dfrac{q(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{y}_{t-1})}{q(\boldsymbol{x}_t|\boldsymbol{y}_{t-1})} = G(\boldsymbol{y}_t|\boldsymbol{\mu}_1, \boldsymbol{Z}_1)$, with $\boldsymbol{\mu}_1 = \boldsymbol{M}^{-1}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}'_t\right)$, $\boldsymbol{Z}_1 = \boldsymbol{M}^{-1}$, $\boldsymbol{M} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}$ for which it suffices to observe its exponential term

$$\begin{aligned} \exp\Bigg\{-\frac{1}{2}\Big[&(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{y}_{t-1})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{B}\boldsymbol{y}_{t-1}) + \left(\boldsymbol{x}'_t - \boldsymbol{U}\boldsymbol{y}_t\right)^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\left(\boldsymbol{x}'_t - \boldsymbol{U}\boldsymbol{y}_t\right) \\ &-\boldsymbol{x}'^{\mathrm{T}}_t\boldsymbol{C}^{-1}\boldsymbol{x}'_t - 2\boldsymbol{x}'^{\mathrm{T}}_t\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1} - \boldsymbol{y}^{\mathrm{T}}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1}\Big]\Bigg\} \end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{ -\frac{1}{2}\left[\boldsymbol{y}_t^{\mathrm{T}}\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\right)\boldsymbol{y}_t - 2\boldsymbol{y}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1} - 2\boldsymbol{y}_t^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t' + \boldsymbol{x}_t'^{\mathrm{T}}\left(\boldsymbol{\Psi}^{-1}-\boldsymbol{C}^{-1}\right)\boldsymbol{x}_t'\right.\right. \\
&\quad \left.\left. + \boldsymbol{y}_{t-1}^{\mathrm{T}}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B} - (\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\right)\boldsymbol{y}_{t-1} - 2\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_t-\boldsymbol{\mu}_1\right)^{\mathrm{T}}\boldsymbol{Z}_1^{-1}\left(\boldsymbol{y}_t-\boldsymbol{\mu}_1\right)\right\},
\end{aligned}
$$

where $\boldsymbol{C}=\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}$.

Next, from Eq. (18) and $q\left(\boldsymbol{x}_t\right)=\int q\left(\boldsymbol{x}_t|\,\boldsymbol{y}_{t-1}\right)q\left(\boldsymbol{y}_{t-1}\right)\mathrm{d}\boldsymbol{y}_{t-1}=G\left(\boldsymbol{x}_t|\,\boldsymbol{\mu},\boldsymbol{\Psi}+\boldsymbol{U}\left(\boldsymbol{\Sigma}+\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}}\right)\boldsymbol{U}^{\mathrm{T}}\right)$, we get

$$
p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)=\frac{q\left(\boldsymbol{x}_t,\boldsymbol{y}_{t-1}\right)}{q\left(\boldsymbol{x}_t\right)}=\frac{q\left(\boldsymbol{x}_t|\,\boldsymbol{y}_{t-1}\right)q\left(\boldsymbol{y}_{t-1}\right)}{q\left(\boldsymbol{x}_t\right)}=G\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{\mu}_2,\boldsymbol{Z}_{2(t-1)}\right),
$$

with $\boldsymbol{\mu}_2=\boldsymbol{P}_{t-1}\boldsymbol{x}_t'$, $\boldsymbol{Z}_{2(t-1)}=\left[\boldsymbol{\Lambda}_{t-1}^{-1}+(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\right]^{-1}$, $\boldsymbol{P}_{t-1}=\left[\boldsymbol{\Lambda}_{t-1}^{-1}+(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\right]^{-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{C}^{-1}$, $\boldsymbol{\Lambda}_{t-1}=(\boldsymbol{V}\boldsymbol{D})^{\mathrm{T}}f_t\left(\tilde{\boldsymbol{B}}\right)\boldsymbol{V}\boldsymbol{D}$, and $f_t(\tilde{\boldsymbol{B}})=\left(\boldsymbol{I}_m-\tilde{\boldsymbol{B}}^{2(t-1)}\right)\left(\boldsymbol{I}_m-\tilde{\boldsymbol{B}}^2\right)^{-1}$, which can be observed from its exponential term

$$
\begin{aligned}
&\exp\left\{-\frac{1}{2}\left[\boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{y}_{t-1}+\boldsymbol{x}_t'^{\mathrm{T}}\left(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{x}_t' - 2\boldsymbol{x}_t'^{\mathrm{T}}\left(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1}\right.\right. \\
&\quad\left.\left. + \boldsymbol{y}_{t-1}^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}})^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{y}_{t-1} - \boldsymbol{x}_t'^{\mathrm{T}}\left(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}+\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\right)^{-1}\boldsymbol{x}_t'\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{y}_{t-1}^{\mathrm{T}}\left[\boldsymbol{\Lambda}_{t-1}^{-1}+(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\left(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{U}\boldsymbol{B}\right]\boldsymbol{y}_{t-1}+\boldsymbol{y}_{t-1}^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\left(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{x}_t'\right. \\
&\quad\left.+\frac{1}{2}\boldsymbol{x}_t'^{\mathrm{T}}\left[\left(\boldsymbol{\Psi}+\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\right)^{-1}-\left(\boldsymbol{\Psi}+\boldsymbol{U}\left(\boldsymbol{\Sigma}+\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}}\right)\boldsymbol{U}^{\mathrm{T}}\right)^{-1}\right]\boldsymbol{x}_t'\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_{t-1}-\boldsymbol{P}_{t-1}\boldsymbol{x}_t'\right)^{\mathrm{T}}\left(\boldsymbol{\Lambda}_{t-1}^{-1}+(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\right)\left(\boldsymbol{y}_{t-1}-\boldsymbol{P}_{t-1}\boldsymbol{x}_t'\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_{t-1}-\boldsymbol{\mu}_2\right)^{\mathrm{T}}\boldsymbol{Z}_{2(t-1)}^{-1}\left(\boldsymbol{y}_{t-1}-\boldsymbol{\mu}_2\right)\right\}.
\end{aligned}
$$

From the above ones, we get $p(\boldsymbol{y}_t,\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t)$ by Eq. (16), which is put into Eq. (13) jointly with $q(\boldsymbol{y}_t|\,\boldsymbol{y}_{t-1})q(\boldsymbol{x}_t|\,\boldsymbol{y}_t)$, resulting in

$$
H_t=-\frac{1}{2}\text{'integration'}+\text{'term'},
$$

where

$$
\begin{aligned}
\text{'term'} &= -\frac{d+m}{2}\ln(2\pi)-\frac{1}{2}\ln|\boldsymbol{\Psi}|-\frac{1}{2}\ln|\boldsymbol{\Sigma}|, \\
\text{'integration'} &= \int p\left(\boldsymbol{y}_t,\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\left[\left(\boldsymbol{y}_t-\boldsymbol{B}\boldsymbol{y}_{t-1}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t-\boldsymbol{B}\boldsymbol{y}_{t-1}\right)+\left(\boldsymbol{x}_t'-\boldsymbol{U}\boldsymbol{y}_t\right)^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\left(\boldsymbol{x}_t'-\boldsymbol{U}\boldsymbol{y}_t\right)\right]\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1} \\
&= \int p\left(\boldsymbol{y}_t|\,\boldsymbol{y}_{t-1},\boldsymbol{x}_t\right)p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\left[\boldsymbol{y}_t^{\mathrm{T}}\boldsymbol{M}\boldsymbol{y}_t-2\boldsymbol{y}_t^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}-2\boldsymbol{y}_t^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right. \\
&\quad\left.+\boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right]\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1} \\
&= \int p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\left\{\mathrm{Tr}\left[\left(\boldsymbol{Z}_1+\boldsymbol{\mu}_1\boldsymbol{\mu}_1^{\mathrm{T}}\right)\boldsymbol{M}\right]-2\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}-2\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right. \\
&\quad\left.+\mathrm{Tr}\left[\boldsymbol{y}_{t-1}\boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right]+\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right\}\mathrm{d}\boldsymbol{y}_{t-1}.
\end{aligned}
$$

Moreover, we integrate out $\boldsymbol{y}_{t-1}$ in parts as follows:

$$
\begin{aligned}
\text{'part1'}:&\int p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\mathrm{Tr}\left[\left(\boldsymbol{Z}_1+\boldsymbol{\mu}_1\boldsymbol{\mu}_1^{\mathrm{T}}\right)\left(\boldsymbol{\Sigma}^{-1}+\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\right)\right]\mathrm{d}\boldsymbol{y}_{t-1} \\
&= \int p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\left\{\mathrm{Tr}\left[\boldsymbol{M}\boldsymbol{Z}_1+\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}+\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}+\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\right.\right.\right. \\
&\quad\left.\left.\left.+\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}\boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{M}^{-1}\right]\right\}\mathrm{d}\boldsymbol{y}_{t-1} \\
&= m+\mathrm{Tr}\left\{\left[\left(\boldsymbol{\Psi}^{-1}\boldsymbol{U}+\boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}+\left(\boldsymbol{\Psi}^{-1}\boldsymbol{U}+\boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1}\right]\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\right\} \\
&\quad+\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{Z}_{2(t-1)}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}\right],
\end{aligned}
$$

$$\begin{aligned}
&\text{'part2'}: \int p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\left(-2\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}\right)\mathrm{d}\boldsymbol{y}_{t-1}\\
&\qquad = \int p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\left(-2\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1} - 2\boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}\right)\mathrm{d}\boldsymbol{y}_{t-1}\\
&\qquad = -\,2\mathrm{Tr}\left[\left(\boldsymbol{\Psi}^{-1}\boldsymbol{U} + \boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1}\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{Z}_{2(t-1)}\right],\\
&\text{'part3'}: \int p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\left(-2\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right)\mathrm{d}\boldsymbol{y}_{t-1} = -2(\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{M}^{-1} + \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1})\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\\
&\qquad = -\,2\mathrm{Tr}\left[\left(\boldsymbol{\Psi}^{-1}\boldsymbol{U} + \boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\right],\\
&\text{'part4'}: \int p\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{x}_t\right)\mathrm{Tr}\left(\boldsymbol{y}_{t-1}\boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\right)\mathrm{d}\boldsymbol{y}_{t-1} = \mathrm{Tr}\left[\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\left(\boldsymbol{\mu}_2\boldsymbol{\mu}_2^{\mathrm{T}} + \boldsymbol{Z}_{2(t-1)}\right)\right]\\
&\qquad = \mathrm{Tr}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{Z}_{2(t-1)}\right) + \mathrm{Tr}\left(\boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1}\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\right),\\
&\text{'part5'}: \mathrm{Tr}\left(\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\right).
\end{aligned}$$

To sum up, we get

$$\begin{aligned}
&\text{'integration'} = \text{'part1'} + \text{'part2'} + \text{'part3'} + \text{'part4'} + \text{'part5'} = \text{'terms with } \boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\text{'} + \text{'terms without } \boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\text{'},\\
&\text{'terms with } \boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\text{'} = -\,\mathrm{Tr}\left\{\left[\boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1} + \boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1} - \boldsymbol{B}\boldsymbol{P}_{t-1}\right) - \boldsymbol{\Psi}^{-1}\right.\right.\\
&\qquad\qquad \left.\left. + \boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{M}^{-1}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1} + \boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1}\right)\right]\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\right\},\\
&\text{'terms without } \boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\text{'} = m + \mathrm{Tr}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{Z}_{2(t-1)}\right) - \mathrm{Tr}\left(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{Z}_{2(t-1)}\right) = 2m - \mathrm{Tr}\left(\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{Z}_{2(t-1)}\right),\\
&\text{'integration'} = 2m - \mathrm{Tr}\left(\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{Z}_{2(t-1)}\right) - \mathrm{Tr}\left[\left(\boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1} + \boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\right.\right.\\
&\qquad\left.\left. + \boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1} + \boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1} - \boldsymbol{P}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{P}_{t-1} - \boldsymbol{\Psi}^{-1}\right)\boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}\right].
\end{aligned}$$

Noticing $\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}^{-1}$, we further have

$$\mathrm{Tr}(\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{Z}_{2(t-1)}) = \mathrm{Tr}\left\{\boldsymbol{\Lambda}_{t-1}^{-1}\left[\boldsymbol{\Lambda}_{t-1} - \boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}(\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}} + \boldsymbol{C})^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\right]\right\},$$

and thus get

$$\text{'integration'} = m + \mathrm{Tr}\left[(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\right] + \mathrm{Tr}\left(\boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1}\boldsymbol{O}\right),$$

where

$$\boldsymbol{O} = \boldsymbol{x}_t'\boldsymbol{x}_t'^{\mathrm{T}}, \quad \boldsymbol{K} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}} + \boldsymbol{C}.$$

Finally, we have

$$H_t = \text{'term'} - \frac{1}{2}\text{'integration'} = -\frac{1}{2}\left\{\ln|\boldsymbol{\Psi}| + \ln|\boldsymbol{D}^2| + \mathrm{Tr}(\boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1}\boldsymbol{O}) + \mathrm{Tr}\left[(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\right]\right\} + \mathrm{const},$$

from which we get Eq. (20).

Appendix B Derivation of gradients for all TFA model parameters

First, we compute the partial derivative of H_t with respect to $\boldsymbol{\Psi}$:

$$\begin{aligned}
\frac{\partial H_t}{\partial \boldsymbol{\Psi}} &= -\frac{1}{2}\boldsymbol{\Psi}^{-1} - \frac{1}{2}\mathrm{diag}\left[-\boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1}\boldsymbol{O}\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1}\boldsymbol{O}\boldsymbol{K}^{-1}\boldsymbol{C}\mathbf{K}^{-1} + \boldsymbol{K}^{-1}\boldsymbol{O}\boldsymbol{K}^{-1} - \boldsymbol{K}^{-1} + \boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1}\right]\\
&= -\frac{1}{2}\boldsymbol{\Psi}^{-1} - \frac{1}{2}\mathrm{diag}\left[\boldsymbol{T}_{\boldsymbol{V}} - \boldsymbol{K}^{-1} + \boldsymbol{\Delta}\right],
\end{aligned}$$

with $\boldsymbol{T}_{\boldsymbol{V}} = \boldsymbol{\Gamma} - \left[\boldsymbol{\Gamma}\boldsymbol{O}\boldsymbol{K}^{-1} + \left(\boldsymbol{\Gamma}\boldsymbol{O}\boldsymbol{K}^{-1}\right)^{\mathrm{T}}\right]$, $\boldsymbol{\Gamma} = \boldsymbol{K}^{-1}\boldsymbol{C}\boldsymbol{K}^{-1}$ and $\boldsymbol{\Delta} = \boldsymbol{K}^{-1}\boldsymbol{O}\boldsymbol{K}^{-1}$.

Second, we compute the partial derivatives of H_t with respect to $\boldsymbol{V}$, $\boldsymbol{U}$ and $\boldsymbol{D}$, respectively:

$$\begin{aligned}
\frac{\partial H_t}{\partial \boldsymbol{V}} &= -\left[\tilde{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{D} + \tilde{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{D}^{-1} + f_t(\tilde{\boldsymbol{B}})\boldsymbol{V}\boldsymbol{D}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{D}\right],\\
\frac{\partial H_t}{\partial \boldsymbol{U}} &= -\left(\boldsymbol{T}_{\boldsymbol{V}} - \boldsymbol{K}^{-1} + \boldsymbol{\Delta}\right)\boldsymbol{U}\boldsymbol{\Sigma} - \boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}},\\
\frac{\partial H_t}{\partial \boldsymbol{D}} &= -\boldsymbol{D}^{-1} - \mathrm{diag}\left[\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{T}_{\boldsymbol{V}} - \boldsymbol{K}^{-1})\boldsymbol{U}\boldsymbol{D} + \boldsymbol{D}^{-1}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}(\boldsymbol{T}_{\boldsymbol{V}} - \boldsymbol{K}^{-1})\boldsymbol{U} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Delta}\boldsymbol{U}\boldsymbol{D} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{K}^{-1}\boldsymbol{U}\ \boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\right]\\
&= -\boldsymbol{D}^{-1} - \mathrm{diag}\left[\boldsymbol{U}^{\mathrm{T}}\left(\boldsymbol{T}_{\boldsymbol{V}} - \boldsymbol{K}^{-1} + \boldsymbol{\Delta}\right)\boldsymbol{U}\boldsymbol{D} + \boldsymbol{U}^{\mathrm{T}}\,\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1}\right].
\end{aligned}$$

Next, we get the partial derivatives of H_t with respect to $\tilde{\boldsymbol{B}}$ and $\boldsymbol{S}$:

$$\begin{aligned}\frac{\partial H_t}{\partial \tilde{\boldsymbol{B}}} &= \mathrm{diag}\Big\{-\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}} + (\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2)^{-1}\boldsymbol{V}\boldsymbol{D}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\Big[(t-1)\tilde{\boldsymbol{B}}^{2t-3} - f_t(\tilde{\boldsymbol{B}})\tilde{\boldsymbol{B}}\Big]\Big\},\\ \frac{\partial H_t}{\partial \boldsymbol{S}} &= \mathrm{diag}\left[\left(\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2\right)\frac{\partial H_t}{\partial \tilde{\boldsymbol{B}}}\right],\end{aligned}$$

where

$$\tilde{\boldsymbol{B}} = \mathrm{diag}\,(b_1, b_2, \ldots, b_m),\ \ \boldsymbol{S} = \mathrm{diag}\,(s_1, s_2, \ldots, s_m)\,,\ \text{and}\ b_i = \frac{\exp(s_i) - \exp(-s_i)}{\exp(s_i) + \exp(-s_i)}.$$

Particularly, at $t = 1$, the TFA-B model degenerates to FA-B:

$$\boldsymbol{x}_1 = \boldsymbol{U}\boldsymbol{y}_1 + \boldsymbol{\mu} + \boldsymbol{e}_1,\ \ \boldsymbol{y}_1 = \boldsymbol{\varepsilon}_1,$$

for which we get

$$H_1 = \int p\left(\boldsymbol{y}_1|\,\boldsymbol{x}_1\right)\ln\left[q\left(\boldsymbol{x}_1|\,\boldsymbol{y}_1\right)q\left(\boldsymbol{y}_1|\,\boldsymbol{y}_0\right)\right]\mathrm{d}\boldsymbol{y}_1 = -\frac{1}{2}\ln|\boldsymbol{\Psi}| - \frac{1}{2}\ln|\boldsymbol{D}^2| - \frac{1}{2}\mathrm{Tr}\left[\left(\boldsymbol{I}_d - \boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\Psi}^{-1}\boldsymbol{O}\right],$$

from $p\left(\boldsymbol{y}_1|\,\boldsymbol{x}_1\right) = G\left(\boldsymbol{y}_1|\,\boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}), (\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U} + \boldsymbol{\Sigma}^{-1})^{-1}\right)$ and $q\left(\boldsymbol{y}_1|\,\boldsymbol{y}_0\right) = G(\boldsymbol{y}_1|\,\boldsymbol{0}, \boldsymbol{\Sigma})$, and accordingly we have

$$\begin{aligned}\frac{\partial H_1}{\partial \boldsymbol{\Psi}} &= -\frac{1}{2}\boldsymbol{\Psi}^{-1} - \frac{1}{2}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{O}\boldsymbol{C}^{-1} + \frac{1}{2}\boldsymbol{\Psi}^{-1}\boldsymbol{O}\left(\boldsymbol{I}_d - \boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\Psi}^{-1},\\ \frac{\partial H_1}{\partial \boldsymbol{D}} &= -\boldsymbol{D}^{-1} - \mathrm{diag}\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{O}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{D}\right),\\ \frac{\partial H_1}{\partial \boldsymbol{U}} &= -\frac{1}{2}\left[\boldsymbol{C}^{-1}\left(\boldsymbol{O}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}} + \boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{O}\right)\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{\Sigma} - \boldsymbol{C}^{-1}\boldsymbol{O}\boldsymbol{\Psi}^{-1}\boldsymbol{U} - \boldsymbol{\Psi}^{-1}\boldsymbol{O}\boldsymbol{C}^{-1}\boldsymbol{U}\right].\end{aligned}$$

For the instantaneous TFA by Eq. (21), we get Eq. (22) with the following newly added term put in Eq. (20):

$$\int q\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{\theta}_y\right)\ln q\left(\boldsymbol{y}_{t-1}|\,\boldsymbol{\theta}_y\right)\mathrm{d}\boldsymbol{y}_{t-1} = -\frac{m}{2} - \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Lambda}_{t-1}|\,.$$

Correspondingly, the partial derivatives of $\boldsymbol{V}$, $\boldsymbol{D}$ and $\tilde{\boldsymbol{B}}$ for the instantaneous TFA model are changed to

$$\begin{aligned}\frac{\partial H_t}{\partial \boldsymbol{V}} &= -\Big[\tilde{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{D} + \tilde{\boldsymbol{B}}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{D}^{-1} + f_t(\tilde{\boldsymbol{B}})\boldsymbol{V}\boldsymbol{D}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{D} + f_t(\tilde{\boldsymbol{B}})\boldsymbol{V}\boldsymbol{D}\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{D}\Big],\\ \frac{\partial H_t}{\partial \boldsymbol{D}} &= -\boldsymbol{D}^{-1} - \mathrm{diag}\left[\boldsymbol{U}^{\mathrm{T}}\left(\boldsymbol{T}_{\boldsymbol{V}} - \boldsymbol{K}^{-1} + \boldsymbol{\Delta}\right)\boldsymbol{U}\boldsymbol{D} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{D}^{-1} - \boldsymbol{V}^{\mathrm{T}}f_t\left(\tilde{\boldsymbol{B}}\right)\boldsymbol{V}\boldsymbol{D}\boldsymbol{\Lambda}_{t-1}^{-1}\right],\\ \frac{\partial H_t}{\partial \tilde{\boldsymbol{B}}} &= \mathrm{diag}\Big\{-\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\ \boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}} + \left(\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2\right)^{-1}\boldsymbol{V}\boldsymbol{D}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\left[(t-1)\tilde{\boldsymbol{B}}^{2t-3} - f_t\left(\tilde{\boldsymbol{B}}\right)\tilde{\boldsymbol{B}}\right]\\ &\quad - \left(\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2\right)^{-1}\boldsymbol{V}\boldsymbol{D}\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\left[(t-1)\tilde{\boldsymbol{B}}^{2t-3} - f_t\left(\tilde{\boldsymbol{B}}\right)\tilde{\boldsymbol{B}}\right]\Big\}\\ &= \mathrm{diag}\Big\{-\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}\ \boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{D}^{-1}\boldsymbol{V}^{\mathrm{T}} + \left(\boldsymbol{I}_m - \tilde{\boldsymbol{B}}^2\right)^{-1}\boldsymbol{V}\boldsymbol{D}\left[(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}\boldsymbol{T}_{\boldsymbol{V}}\boldsymbol{U}\boldsymbol{B} - \boldsymbol{\Lambda}_{t-1}^{-1}\right]\\ &\quad\cdot\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}\left[(t-1)\tilde{\boldsymbol{B}}^{2t-3} - f_t\left(\tilde{\boldsymbol{B}}\right)\tilde{\boldsymbol{B}}\right]\Big\}.\end{aligned}$$

Appendix C Derivation of Eq. (24)

For an observation sequence $\boldsymbol{X}$, we have

$$p(\boldsymbol{X}) = p\left(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\right) = \prod_{t=2}^{N} p\left(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}\right)p\left(\boldsymbol{x}_1\right),\quad p\left(\boldsymbol{x}_1\right) = G(\boldsymbol{x}_1|\,\boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^{\mathrm{T}}).$$

We can get

$$p\left(\boldsymbol{x}_t|\,\boldsymbol{x}_{t-1}\right) = \frac{p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}\right)}{p\left(\boldsymbol{x}_{t-1}\right)},$$

from $p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}\right) = \int p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{y}_t, \boldsymbol{y}_{t-1}\right)\mathrm{d}\boldsymbol{y}_t\mathrm{d}\boldsymbol{y}_{t-1}$ and $p(\boldsymbol{x}_{t-1}) = G(\boldsymbol{x}_{t-1}|\,\boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{U}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^{\mathrm{T}})$.

We start from

$$
\begin{aligned}
&p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{y}_t, \boldsymbol{y}_{t-1}\right)\\
&\quad = p\left(\boldsymbol{x}_t \middle| \boldsymbol{y}_t\right) p\left(\boldsymbol{y}_t \middle| \boldsymbol{y}_{t-1}\right) p\left(\boldsymbol{x}_{t-1} \middle| \boldsymbol{y}_{t-1}\right) p\left(\boldsymbol{y}_{t-1}\right)\\
&\quad \propto \exp\Bigg\{-\frac{1}{2}\Bigg\{\left[\boldsymbol{y}_t - \boldsymbol{M}^{-1}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right)\right]^{\mathrm{T}} \boldsymbol{M}\left[\boldsymbol{y}_t - \boldsymbol{M}^{-1}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right)\right]\\
&\qquad + \boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t' + \boldsymbol{x}_{t-1}'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_{t-1}' - 2\boldsymbol{x}_{t-1}'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{y}_{t-1} + \boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{y}_{t-1} + \boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1}\\
&\qquad + \boldsymbol{y}_{t-1}^{\mathrm{T}}\boldsymbol{\Lambda}_{t-1}^{-1}\boldsymbol{y}_{t-1} - \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right)^{\mathrm{T}}\boldsymbol{M}^{-1}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\boldsymbol{y}_{t-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'\right)\Bigg\}\Bigg\},
\end{aligned}
$$

where

$$
\begin{aligned}
&p\left(\boldsymbol{x}_t \middle| \boldsymbol{y}_t\right) = G\left(\boldsymbol{x}_t \middle| \boldsymbol{U}\boldsymbol{y}_t + \boldsymbol{\mu}, \boldsymbol{\Psi}\right),\ p\left(\boldsymbol{y}_t \middle| \boldsymbol{y}_{t-1}\right) = G\left(\boldsymbol{y}_t \middle| \boldsymbol{B}\boldsymbol{y}_{t-1}, \boldsymbol{\Sigma}\right),\ p(\boldsymbol{y}_{t-1}) = G\left(\boldsymbol{y}_{t-1} \middle| \boldsymbol{0}, \boldsymbol{\Lambda}_{t-1}\right),\\
&\boldsymbol{x}_t' = \boldsymbol{x}_t - \boldsymbol{\mu}, \boldsymbol{x}_{t-1}' = \boldsymbol{x}_{t-1} - \boldsymbol{\mu},\ \boldsymbol{\Sigma} = \boldsymbol{D}^2,\ \text{and}\ \ \boldsymbol{M} = \boldsymbol{\Sigma}^{-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U},
\end{aligned}
$$

from which we get

$$
\begin{aligned}
&p(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1})\\
&\quad = \int p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{y}_t, \boldsymbol{y}_{t-1}\right) \mathrm{d}\boldsymbol{y}_t\\
&\quad \propto \exp\left\{-\frac{1}{2}\left[\left(\boldsymbol{y}_{t-1} - \boldsymbol{S}\boldsymbol{u}_t\right)^{\mathrm{T}}\boldsymbol{S}^{-1}\left(\boldsymbol{y}_{t-1} - \boldsymbol{S}\boldsymbol{u}_t\right) - \boldsymbol{u}_t^{\mathrm{T}}\boldsymbol{S}\boldsymbol{u}_t + \boldsymbol{x}_t'^{\mathrm{T}}\left(\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\right)\boldsymbol{x}_t' + \boldsymbol{x}_{t-1}'^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_{t-1}'\right]\right\},
\end{aligned}
$$

with $\boldsymbol{u}_t = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_{t-1}' + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_t'$, $\boldsymbol{S} = \left[\boldsymbol{\Lambda}_{t-1}^{-1} + \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{U} + \boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\right]^{-1}$ and $\boldsymbol{C} = \boldsymbol{\Psi} + \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}$,

$$
\begin{aligned}
&p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}\right)\\
&\quad = \int p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}\right) \mathrm{d}\boldsymbol{y}_{t-1}\\
&\quad \propto \exp\Bigg\{-\frac{1}{2}\Big[\boldsymbol{x}_{t-1}'^{\mathrm{T}}\left(\boldsymbol{I}_d - \boldsymbol{\Psi}^{-1}\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{\Psi}^{-1}\boldsymbol{x}_{t-1}' - 2\boldsymbol{x}_t'^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{S}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{x}_{t-1}'\\
&\qquad + \boldsymbol{x}_t'^{\mathrm{T}}\left(\boldsymbol{I}_d - \boldsymbol{C}^{-1}\boldsymbol{U}\boldsymbol{B}\boldsymbol{S}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{U}^{\mathrm{T}}\right)\boldsymbol{C}^{-1}\boldsymbol{x}_t'\Big]\Bigg\}.
\end{aligned}
$$

Finally, we obtain

$$
p\left(\boldsymbol{x}_t \middle| \boldsymbol{x}_{t-1}\right) = \frac{p\left(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}\right)}{p\left(\boldsymbol{x}_{t-1}\right)} = G(\boldsymbol{x}_t | \boldsymbol{F}\boldsymbol{x}_{t-1}' + \boldsymbol{\mu}, \boldsymbol{W}),
$$

with $\boldsymbol{F} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^{\mathrm{T}}\left(\boldsymbol{\Psi} + \boldsymbol{U}\boldsymbol{\Lambda}_{t-1}\boldsymbol{U}^{\mathrm{T}}\right)^{-1}$ and $\boldsymbol{W} = \boldsymbol{C} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}} - \boldsymbol{F}\boldsymbol{U}\boldsymbol{\Lambda}_{t-1}(\boldsymbol{U}\boldsymbol{B})^{\mathrm{T}}$.

Acknowledgements The work described in this paper was supported by a grant of the General Research Fund (GRF) from the Research Grant Council of the Hong Kong SAR (No. CUHK4180/10E), the National Natural Science Foundation of China (Grant Nos. 60901067 and 61001212), Program for New Century Excellent Talents in University (No. NCET-09-0630), Program for Changjiang Scholars and Innovative Research Team in University (No. IRT0954), and the Fundamental Research Funds for the Central Universities.

References

1. Kosir P, DeWal R. Feature alignment techniques for pattern recognition. In: Proceedings of IEEE National Conference on Aerospace and Electronics. 1994, 1: 128–132
2. Webb A R. Gamma mixture models for target recognition. Pattern Recognition, 2000, 33(12): 2045–2054
3. Copsey K, Webb A R. Bayesian Gamma mixture model approach to radar target recognition. IEEE Transactions on Aerospace and Electronic Systems, 2003, 39(4): 1201–1217
4. Seibert M, Waxman A M. Adaptive 3-D object recognition from multiple views. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1992, 14(2): 107–124
5. Jacobs S P. Automatic target recognition using high-resolution radar range profiles. Dissertation for the Doctoral Degree. St. Louis: Washington University, 1999
6. Du L, Liu H W, Bao Z, Zhang J Y. A two-distribution compounded statistical model for radar HRRP target recognition. IEEE Transactions on Signal Processing, 2006, 54(6): 2226–2238
7. Du L, Liu H W, Bao Z. Radar HRRP statistical recognition based on hypersphere model. Signal Processing, 2008, 88(5): 1176–1190
8. Du L, Liu H W, Bao Z. Radar HRRP statistical recognition: parametric model and model selection. IEEE Transactions on Signal Processing, 2008, 56(5): 1931–1944
9. Zhu F, Zhang X D, Hu Y F. Gabor filter approach to joint feature extraction and target recognition. IEEE Transactions on Aerospace and Electronic Systems, 2009, 45(1): 17–30
10. Wong S K. High range resolution profiles as motion-invariant features for moving ground targets identification in SAR-

based automatic target recognition. IEEE Transactions on Aerospace and Electronic Systems, 2009, 45(3): 1017–1039

11. Xu L. Bayesian Ying-Yang system and theory as a unified statistical learning approach: (v) temporal modeling for temporal perception and control. In: Proceedings of the International Conference on Neural Information Processing. 1998, 2: 877–884

12. Xu L. Temporal Bayesian Ying-Yang dependence reduction, blind source separation and principal independent components. In: Proceedings of International Joint Conference on Neural Networks. 1999, 2: 1071–1076

13. Xu L. Temporal BYY learning for state space approach, hidden Markov model, and blind source separation. IEEE Transactions on Signal Processing, 2000, 48(7): 2132–2144

14. Xu L. BYY harmony learning, independent state space, and generalized APT financial analyses. IEEE Transactions on Neural Networks, 2001, 12(4): 822–849

15. Xu L. Temporal factor analysis: stable-identifiable family, orthogonal flow learning, and automated model selection. In: Proceedings of International Joint Conference on Neural Networks. 2002, 472–476

16. Xu L. Independent component analysis and extensions with noise and time: a Bayesian Ying-Yang learning perspective. Neural Information Processing — Letters and Reviews, 2003, 1(1): 1–52

17. Xu L. Temporal BYY encoding, Markovian state spaces, and space dimension determination. IEEE Transactions on Neural Networks, 2004, 15(5): 1276–1295

18. Xu L. Learning algorithms for RBF functions and subspace based functions. Handbook of Research on Machine Learning, Applications and Trends: Algorithms, Methods and Techniques. Hershey: IGI Global, 2009, 60–94

19. Xu L. Bayesian Ying-Yang system, best harmony learning and five action circling. Frontiers of Electrical and Electronic Engineering in China, 2010, 5(3): 281–328

20. Chiu K C, Xu L. Arbitrage pricing theory based Gaussian temporal factor analysis for adaptive portfolio management. Decision Support Systems, 2004, 37(4): 485–500

21. Chiu K C, Xu L. Optimizing financial portfolios from the perspective of mining temporal structures of stock returns. In: Proceedings of the 3rd International Conference on Machine Learning. 2003, 266–275

22. Burnham K P, Anderson D. Model Selection and Multi-Model Inference. New York: Springer, 2002

23. Akaike H. Factor analysis and AIC. Psychometrika, 1987, 52(3): 317–332

24. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 1974, 19(6): 714–723

25. Bozdogan H. Model selection and Akaike's information criterion (AIC): the general theory and its analytical extension. Psychometrika, 1987, 52(3): 345–370

26. Anderson T W, Rubin H. Statistical inference in factor analysis. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. 1956, 5: 111–150

27. Ghahramani Z, Hinton G. Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96–2, 1996

28. Roweis S, Ghahramani Z. A unifying review of linear Gaussian models. Neural Computation, 1999, 11(2): 305–345

29. Ghahramani Z, Hinton G E. Variational learning for switching state-space models. Neural Computation, 2000, 12(4): 831–864

30. Carrara W G, Goodman R S, Majewski R M. Spotlight Synthetic Aperture Radar — Signal Processing Algorithms. Boston: Arthech House, 1995

31. Parzen E. On the estimation of a probability density function and mode. Annals of Mathematical Statistics, 1962, 33(3): 1065–1076

32. Rubin D B, Thayer D T. EM algorithms for ML factor analysis. Psychometrika, 1982, 47(1): 69–76

33. Schwarz G. Estimating the dimension of a model. Annals of Statistics, 1978, 6(2): 461–464

34. Shi L, Wang P, Liu H, Xu L, Bao Z. Radar HRRP statistical recognition with local factor analysis by automatic Bayesian Ying-Yang harmony learning. IEEE Transactions on Signal Processing, 2011, 59(2): 610–617

35. Salah A A, Alpaydin E. Incremental mixtures of factor analyzers. In: Proceedings of the 17th International Conference on Pattern Recognition. 2004, 1: 276–279

36. Tipping M E, Bishop C M. Mixtures of probabilistic principal component analyzers. Neural Computation, 1999, 11(2): 443–482

37. Shumway R H, Stoffer D S. An approach to time series smoothing and forecasting using the EM algorithm. Journal of Time Series Analysis, 1982, 3(4): 253–264

38. Xu L. YING-YANG machine for temporal signals. In: Proceedings of 1995 IEEE International Conference on Neural Networks and Signal Processing. 1995, I: 644–651

39. Xu L. Bayesian Ying Yang system and theory as a unified statistical learning approach: (ii) from unsupervised learning to supervised learning and temporal modeling. In: Wong K M, King I, Yeung D Y, eds. Proceedings of Theoretical Aspects of Neural Computation: A Multidisciplinary Perspective. 1997, 25–42

40. Xu L. Temporal BYY learning and its applications to extended Kalman filtering, hidden Markov model, and sensor-motor integration. In: Proceedings of International Joint Conference on Neural Networks. 1999, 2: 949–954

41. Shumway R H, Stoffer D S. Dynamic linear models with switching. Journal of the American Statistical Association, 1991, 86(415): 763–769

42. Elliott R J, Aggoun L, Moore J B. Hidden Markov Models: Estimation and Control. New York: Springer-Verlag, 1995

43. Digalakis V, Rohlicek J R, Ostendorf M. ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition. IEEE Transactions on Speech and Audio Processing, 1993, 1(4): 431–442

44. Xu L. Bayesian-Kullback coupled YING-YANG machines: unified learning and new results on vector quantization. In: Proceedings of the International Conference on Neural Information Processing. 1995, 977–988 (A further version in NIPS8. In: Touretzky D S, et al. eds. Cambridge: MIT Press, 444–450)

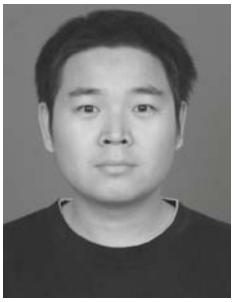
45. Xu L. Another perspective of BYY harmony learning: representation in multiple layers, co-decomposition of data covariance matrices, and applications to network biology. Frontiers of Electrical and Electronic Engineering in China, 2011, 6(1): 86–119

46. Xu L. Bayesian Ying Yang system and theory as a unified statistical learning approach: (i) unsupervised and semiunsupervised learning. In: Amari S, Kassabov N, eds. Brain-Like Computing and Intelligent Information Systems. New Zealand: Springer-Verlag, 1997, 241–274

47. Tu S, Xu L. Parameterizations make different model selections: empirical findings from factor analysis. Frontiers of Electrical and Electronic Engineering in China, 2011 (in Press)

48. Xu L. Data smoothing regularization, multi-sets-learning, and problem solving strategies. Neural Networks, 2003, 16(5–6): 817–825

49. Egan J P. Signal Detection Theory and ROC Analysis. San Diego: Academic Press, 1975



Penghui WANG received the B.Eng. degree in communication engineering from the National University of Defense Technology, Changsha, China, in 2005. He is currently working toward the Ph.D degree with the National Laboratory of Radar Signal Processing, Xidian University. His research interests include radar signal processing, radar automatic target recognition (RATR), and pattern recognition.

Lei SHI received the B.Eng. degree in computer science and technology from the University of Science and Technology of China, Hefei, in 2005. He is currently a Ph.D student with the Department of Computer Science and Engineering, the Chinese University of Hong Kong. His research interests include statistical learning and neural computing.

Lan DU received the B.S., M.S. and Ph.D degrees in electronic engineering from Xidian University, Xi'an, China, in Jul. 2001, Mar. 2004 and Jun. 2007 respectively. Her doctoral dissertation was granted Top 100 Doctoral Dissertation in China in 2009. She is currently an Associate Professor in Xidian University. Her main research interests are in the fields of statistical signal processing and machine learning with application to radar target recognition.

Hongwei LIU received the B.Eng. degree from Dalian University of Technology in electronic engineering in 1992, and the M.Eng. and Ph.D degrees in electronic engineering from Xidian University, Xi'an, China, in 1995 and 1999, respectively. He is currently the Director and a Professor with the National Laboratory of Radar Signal Processing, Xidian University. His research interests include radar automatic target recognition (RATR), radar signal processing, and adaptive signal processing.

Lei XU, IEEE Fellow (2001–) and Fellow of International Association for Pattern Recognition (2002–), and Academician of European Academy of Sciences (2002–); a Chair Professor with the Chinese University of Hong Kong, a Chang Jiang Chair Professor with Peking University, China, and an Honorary Professor with Xidian University (See Front. Electr. Electron. Eng. China, 2011, 6(1): 119 for a detailed introduction).

Zheng BAO received the B.Eng. degree from the Communication Engineering Institution of China in 1953. Currently, he is a Professor at Xidian University, Xi'an, China. He is the author or coauthor of six books and has published more than 300 papers. His current research work focuses on the areas of space-time adaptive processing (STAP), radar imaging (SAR/ISAR), radar automatic target recognition (RATR), over-the-horizon radar (OTHR) signal processing, and passive coherent location (PCL). Prof. Bao is a member of the Chinese Academy of Sciences.