Probabilistic Slow Features for Behavior Analysis

Lazaros Zafeiriou, *Student Member, IEEE*, Mihalis A. Nicolaou, *Member, IEEE*,
Stefanos Zafeiriou, *Member, IEEE*, Symeon Nikitidis,
and Maja Pantic, *Fellow, IEEE*

***Abstract*— A recently introduced latent feature learning technique for time-varying dynamic phenomena analysis is the so-called slow feature analysis (SFA). SFA is a deterministic component analysis technique for multidimensional sequences that, by minimizing the variance of the first-order time derivative approximation of the latent variables, finds uncorrelated projections that extract slowly varying features ordered by their temporal consistency and constancy. In this paper, we propose a number of extensions in both the deterministic and the probabilistic SFA optimization frameworks. In particular, we derive a novel deterministic SFA algorithm that is able to identify linear projections that extract the common slowest varying features of two or more sequences. In addition, we propose an expectation maximization (EM) algorithm to perform inference in a probabilistic formulation of SFA and similarly extend it in order to handle two and more time-varying data sequences. Moreover, we demonstrate that the probabilistic SFA (EM-SFA) algorithm that discovers the common slowest varying latent space of multiple sequences can be combined with dynamic time warping techniques for robust sequence time-alignment. The proposed SFA algorithms were applied for facial behavior analysis, demonstrating their usefulness and appropriateness for this task.**

***Index Terms*— Behavior analysis, linear dynamical system (LDS), slow feature analysis (SFA), temporal alignment.**

I. Introduction

SLOW feature analysis (SFA) was first proposed in [1] as an unsupervised methodology for finding slowly varying (invariant) features from rapidly temporal varying signals. The exploited slowness learning principle in [1] was motivated by the empirical observation that higher order meanings of sensory data, such as objects and their attributes, are often more persistent (i.e., change smoothly) than the independent activation of any single sensory receptor. For instance, the position and the identity of an object are visible for extended periods of time and change with time in a continuous fashion. Their change is slower than that of any primary sensory signal (like the responses of individual retinal receptors or the gray-scale values of a single pixel in a video camera), thus being more robust to subtle changes in the environment.

To identify the most slowly varying features, a trace optimization problem with generalized orthogonality constraints was formulated in [1] that assumes a discrete time input signal[1] and the low-dimensional output signal is obtained as a linear transformation of a nonlinear expansion of the input. The optimization problem proposed in [1] aims to minimize the magnitude of the approximated first-order time derivative of the extracted slowly varying features under the constraints that these are centered (i.e., have zero mean) and uncorrelated. Thus, the slowest varying features are identified by solving a generalized eigenvalue problem (GEP) for the joint diagonalization of the data covariance matrix and the covariance matrix of the first-order forward data differences.

Intuitively, SFA imitates the functionality of the receptive fields of the visual cortex [3], thus being appropriate for describing the evolution of time-varying visual phenomena. However, until today limited research has been conducted regarding its efficacy on computer vision problems [4]–[8]. Recently, SFA and its discriminant extensions have been successfully applied for human action recognition in [8], while hierarchical segmentation of video sequences using SFA was investigated in [7]. In [4], SFA was applied for object and object-pose recognition on a homogeneous background, while in [6] SFA for vector-valued functions was studied for blind source separation. Finally, an incremental SFA algorithm for change detection was proposed in [5].

Links between SFA and other component analysis techniques, such as independent component analysis (ICA) and Laplacian eigenmaps (LE) [9] were extensively studied in [10] and [11]. In [10], the equivalence between linear SFA and the second-order ICA algorithm, in the case of one time delay, is demonstrated. In [11], the relation between LE and SFA was studied. This paper demonstrated that SFA is a special case of kernel locality preserving projections [12] acquired by defining the data neighborhood structure using their temporal variations. In [13], it was shown that the projection bases provided by SFA are similar to those yielded by the maximum likelihood (ML) solution of a probabilistic generative model in the limit case where the noise variance tends to zero. The probabilistic generative model comprises a

Manuscript received December 11, 2013; revised January 21, 2015 and May 5, 2015; accepted May 16, 2015. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Project EP/J017787/1 (4DFAB). The work of S. Zafeiriou was supported by EPSRC under Project EP/L026813/1 and in part by the Adaptive Facial Deformable Models for Tracking. The work of M. Pantic was supported by the European Community Horizon 2020 [H2020/2014-2020] under Grant 645094.

L. Zafeiriou, M. A. Nicolaou, S. Zafeiriou, and S. Nikitidis are with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: l.zafeiriou12@imperial.ac.uk; mihalis@imperial.ac.uk; s.zafeiriou@imperial.ac.uk; s.nikitidis@imperial.ac.uk).

M. Pantic is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K., and also with the Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede 7522NB, The Netherlands (e-mail: m.pantic@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2435653

[1]Continuous time SFA has been proposed in [2], but because we assume discrete time signals in this paper, such works are out of our scope.

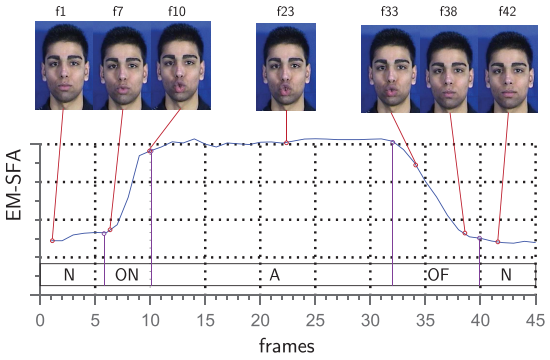

Fig. 1. Latent space obtained by EM-SFA, accurately capturing the transition between temporal phases of AUs. The ground truth is shown as N: neutral, ON: onset, A: apex, and OF: offset.

linear model for the generation of observations and imposes a Gaussian linear dynamical system (LDS) with diagonal covariances over the latent space.

In this paper, we study the application of SFA for unsupervised facial behavior analysis. Our motivation is based on the aforementioned theory on the close relationship between human perception and SFA. The application of SFA is further motivated in Fig. 1, which shows the resulting latent space obtained by EM-SFA, applied on a video sequence where the subject is activating action unit (AU) 22 (lip funneler). In general, when activating an AU, the following temporal phases are present. Neutral, when the face is expressionless. Onset, when the action initiates. Apex, when the action reach the peak intensity. Offset, when the muscles begin to relax. The action finally ends with neutral. It can be clearly observed in Fig. 1 that the latent space obtained by EM-SFA accurately captures the transitions between the temporal phases of the AU, providing an unsupervised method for detecting the temporal phases of AUs.

Summarizing the contributions of this paper, we propose the following theoretical novelties.

1) We propose the first expectation maximization (EM) algorithm for learning the model parameters of a probabilistic SFA (EM-SFA). In contrast to existing ML approaches [13], our approach allows for full probabilistic modeling of the latent distributions instead of mapping the variances to zero, as in ML.
2) We extend both deterministic and probabilistic SFA to enable us to find the common slowest varying features of two or more time-varying data sequences, thus allowing the simultaneous analysis of multiple data streams.
3) We study the relation of SFA to a very common component analysis technique applied to two sequences, canonical correlation analysis (CCA). Through this analysis, we highlight the basic variations of the two methods. In particular, we showed that CCA can be provided by a fully connected Markov random field (MRF) model that does not consider time dependence, while the proposed the probabilistic does that explicitly.

The novelties of this paper in terms of application can be summarized as follows.

1) We apply the proposed EM-SFA to facial behavior dynamics analysis and in particular for facial AUs analysis. More precisely, we demonstrate that it is possible to discover the dynamics of AUs in an unsupervised manner using EM-SFA. To the best of our knowledge, this is the first unsupervised approach that detects the temporal phases of AUs [other unsupervised approaches such as [14] focus on detecting facial expression methodology (i.e., constituting AUs) rather than their temporal phases].
2) We combine the common latent space derived by EM-SFA with dynamic time warping (DTW) [15] for the temporal alignment of dynamic facial behavior. We claim that using the slowest varying features for sequence alignment is well motivated by the principle of slowness, i.e., that slowly varying features correspond to target changes rather than rapidly varying ones, which most likely corresponds to noise [1]).

The rest of this paper is organized as follows. In Section II, we describe the deterministic SFA model, while in Section III, we introduce the probabilistic interpretation of SFA. Our proposed EM-SFA is presented in Section IV, both for one (Section IV-A) and multiple sequences (Section IV-E), while the latter method is incremented with warpings in Section VI-C. In Section V, we discuss the relationship between SFA and CCA. Finally, we evaluate the proposed models in Section VI, by a set of experiments with both synthetic (Section VI-A) and real (Sections VI-B and VI-C) data.

II. Deterministic Slow Feature Analysis

In order to identify the slowest varying features deterministic SFA considers the following optimization problem. Given an M-dimensional time-varying input sequence $\mathbf{X} = [\mathbf{x}_t, t \in [1, T]]$, where t denotes time and $\mathbf{x}_t \in \Re^M$ is the sample of observations at time t, SFA seeks to determine appropriate projection bases stored in the columns of matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N] \in \Re^{M\times N}$ ($N \ll M$), that in the low-dimensional space minimize the variance of the approximated first-order time derivative of the latent variables $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T] \in \Re^{N\times T}$ subject to zero mean, unit covariance, and decorrelation constraints

$$\begin{aligned} &\min_{\mathbf{V}} \quad \text{tr}[\dot{\mathbf{Y}}\dot{\mathbf{Y}}^T] \\ &\text{s.t.} \quad \mathbf{Y1} = \mathbf{0}, \quad \mathbf{YY}^T = \mathbf{I} \end{aligned} \tag{1}$$

where tr[.] is the matrix trace operator, $\mathbf{1}$ is a $T \times 1$ vector with all its elements equal to $(1/T)$, $\mathbf{I}$ is a $N \times N$ identity matrix and matrix $\dot{\mathbf{Y}}$ approximates the first-order time derivative of $\mathbf{Y}$, evaluated using the forward latent variable differences as follows:

$$\dot{\mathbf{Y}} = [\mathbf{y}_2 - \mathbf{y}_1, \mathbf{y}_3 - \mathbf{y}_2, \ldots, \mathbf{y}_T - \mathbf{y}_{T-1}]. \tag{2}$$

Considering the linear case where the latent space can be derived by projecting the input samples on a set of basis $\mathbf{V}$ where $\mathbf{Y} = \mathbf{V}^T\mathbf{X}$ and assuming that input data have been normalized such as to have zero mean, problem (1) can be reformulated to the following trace optimization problem:

$$\min_{\mathbf{V}} \ \text{tr}[\mathbf{V}^T\mathbf{AV}] \ \text{ s.t. } \mathbf{V}^T\mathbf{BV} = \mathbf{I} \tag{3}$$

where $\mathbf{B}$ is the input data covariance matrix and $\mathbf{A}$ is an $M \times M$ covariance matrix evaluated using the forward temporal differences of the input data, contained in matrix $\dot{\mathbf{X}}$

$$\mathbf{A} = \frac{1}{T-1}\dot{\mathbf{X}}\dot{\mathbf{X}}^T, \quad \mathbf{B} = \frac{1}{T}\mathbf{X}\mathbf{X}^T. \tag{4}$$

The solution of (3) can be found from the GEP [1]

$$\mathbf{A}\mathbf{V} = \mathbf{B}\mathbf{V}\mathbf{L} \tag{5}$$

where the columns of the projection matrix $\mathbf{V}$ are the generalized eigenvectors associated with the N-lower generalized eigenvalues contained sorted in the diagonal matrix $\mathbf{L}$.

III. Probabilistic Interpretation of SFA

In this section, we discuss a probabilistic approach to SFA latent variable extraction. Let us assume the following linear generative model that relates the latent variable $\mathbf{y}_t$ with the observed samples $\mathbf{x}_t$ as:

$$\mathbf{x}_t = \mathbf{V}^{-T}\mathbf{y}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N\big(0, \sigma_x^2\mathbf{I}\big) \tag{6}$$

where $\mathbf{e}_i$ is the noise which is assumed to be an isotropic Gaussian model. Hence, the conditional probability is $P(\mathbf{x}_t|\mathbf{V}, \mathbf{y}_t, \sigma_x^2) = \mathcal{N}(\mathbf{V}^{-T}\mathbf{y}_t, \sigma_x^2\mathbf{I})$. Let us also assume that the linear Gaussian dynamical system priors over the latent space $\mathbf{Y}$ are

$$\begin{aligned} P\big(\mathbf{y}_t|\mathbf{y}_{t-1}, \lambda_{1:N}, \sigma_{1:N}^2\big) &= \prod_{n=1}^{N} P\big(y_{n,t}|y_{n,t-1}, \lambda_n, \sigma_n^2\big) \\ P\big(y_{n,t}|y_{n,t-1}, \lambda_n, \sigma_n^2\big) &= \mathcal{N}\big(\lambda_n y_{n,t-1}, \sigma_n^2\big) \\ P\big(y_{n,1}|\sigma_{n,1}^2\big) &= \mathcal{N}(0, \sigma_{n,1}^2). \end{aligned} \tag{7}$$

Defining the model parameters $\theta = \{\theta_x, \theta_y\}$, where $\theta_x = \{\mathbf{V}, \sigma_x^2\}$, $\theta_y = \{\mathbf{\Lambda}, \mathbf{\Sigma}, \mathbf{\Sigma}_1\}$ with matrices $\mathbf{\Lambda} = [\delta_{i,j}\lambda_n]$, $\mathbf{\Sigma} = [\delta_{i,j}\sigma_n^2]$, and $\mathbf{\Sigma}_1 = [\delta_{i,j}\sigma_{n,1}^2]$ the prior over the latent space can be evaluated as

$$\begin{aligned} &P(\mathbf{Y}|\theta_y) \\ &= \frac{1}{Z}\exp\left[-\sum_{n=1}^{N}\left(\frac{1}{2\sigma_{n,1}^2}y_{n,1} + \frac{1}{2\sigma_n^2}\sum_{t=2}^{T}[y_{n,t} - \lambda_n y_{n,t-1}]^2\right)\right] \\ &= \frac{1}{Z}\exp\big[-\mathrm{tr}\big[\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{(2)} + \dot{\mathbf{Y}}\dot{\mathbf{Y}}^T\mathbf{\Lambda}^{(1)} + (\mathbf{y}_1\mathbf{y}_1 + \mathbf{y}_T\mathbf{y}_T)\mathbf{\Lambda}^{(3)}\big]\big] \end{aligned} \tag{8}$$

where $Z = \int_{\mathbf{Y}} P(\mathbf{Y})d\mathbf{Y}$, $\mathbf{\Lambda}^{(1)} = [\delta_{i,j}(\lambda_n/\sigma_n^2)]$, $\mathbf{\Lambda}^{(2)} = [\delta_{i,j}((1-\lambda_n)^2/\sigma_n^2)]$, $\mathbf{\Lambda}^{(3)} = [\delta_{i,j}\lambda_n(1-\lambda_n)]$, and $\delta_{i,j} = 1$ for $i = j$ and 0 for $i \neq j$.

In [13], it was shown that the conditional probability in (8) for the deterministic case (i.e., taking the limit $\sigma_x^2 \to 0$) is simplified to

$$P(\mathbf{Y}|\theta_y) \approx \frac{1}{Z}\exp[-\mathrm{tr}[\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{(2)} + \dot{\mathbf{Y}}\dot{\mathbf{Y}}^T\mathbf{\Lambda}^{(1)}]]. \tag{9}$$

Thus, the ML solution for the basis matrix $\mathbf{V}$ of the above model is evaluated as

$$\begin{aligned} \mathbf{V} &= \arg\max_{\mathbf{V}, \sigma_x^2 \to 0} \log P(\mathbf{X}|\theta) \\ &= \arg\max_{\mathbf{V}, \sigma_x^2 \to 0} \log \int_{\mathbf{Y}} P(\mathbf{X}|\mathbf{Y}, \theta_x)P(\mathbf{Y}|\theta_y)d\mathbf{Y}. \end{aligned} \tag{10}$$

Completing the integrals and assuming a sufficient large number of data samples, the optimization problem in (10) results to the following optimization problem:

$$\mathbf{V} = \arg\max_{\mathbf{V}} \; T\log|\mathbf{V}| - \frac{T}{2}\mathrm{tr}[\mathbf{V}\mathbf{B}\mathbf{V}^T\mathbf{\Lambda}^{(2)} + \mathbf{V}\mathbf{A}\mathbf{V}^T\mathbf{\Lambda}^{(1)}] + c \tag{11}$$

where all terms independent of $\mathbf{V}$ are summarized by the constant c. Differentiating (11) with respect to $\mathbf{V}$ yields the same solution as (3) up to a scale factor.

In the ML solution the direction of $\mathbf{V}$ does not depend on σ_n^2 and λ_n. If $0 < \lambda_n < 1, \forall n$, then larger values of λ_n correspond to slower latent variables. This directly induces an ordering to the derived SFA slowly varying features. In order to recover the exact equivalent of the deterministic SFA algorithm, another limit is required to correct the scales. A natural approach is to set $\sigma_n^2 = 1 - \lambda_n^2$ [13], which constraints the prior covariance of the latent variables to be one.

IV. EM Approach for Probabilistic SFA

The ML approach for probabilistic SFA bears many disadvantages. First, the mapping of $\sigma_x^2 \to 0$ essentially reduces the model to a deterministic one, and serves mostly as a theoretical proof of the connection of the probabilistic interpretation and the deterministic model. Furthermore, the ML method approximates the latent Markov chain by employing first-order derivatives. In this section, we present a fully probabilistic treatment to SFA, which includes modeling full distributions along with both observation and latent variance (EM-SFA, Section IV-A). Furthermore, we extend EM-SFA to handle two distinct sequences (Section IV-E), while the extension for handling any number of multiple sequences is straight-forward.

A. EM-SFA for Single Sequence

In this section, we propose a complete probabilistic SFA algorithm using EM, while following the constraints discussed in Section III ($0 < \lambda_n < 1, \forall n$ and $\sigma_n^2 = 1 - \lambda_n^2$).[2] First, let us slightly modify the considered linear generative model such as $\mathbf{x}_t = \mathbf{V}\mathbf{y}_t + \mathbf{e}_t, \mathbf{e}_t \sim N(0, \sigma_x^2\mathbf{I})$.[3] Let us also define the new model parameters $\theta = \{\theta_x, \mathbf{\Sigma}_1, \mathbf{\Lambda}\}$ (since $\mathbf{\Sigma}$ is a function of $\mathbf{\Lambda}$).

To derive the probabilistic EM-SFA algorithm, we consider a linear generative model according to which each latent variable $\mathbf{y}_t$ is associated with an observation sample $\mathbf{x}_t$ and connected with a first-order Markov chain. Thus, the probability distribution $P(\mathbf{y}_t|\mathbf{y}_{t-1})$ for each latent variable $\mathbf{y}_t$ is only conditioned by the previous variable $\mathbf{y}_{t-1}$. Fig. 2(a) presents graphically the considered linear generative model for a single sequence. In addition, we considered that both the observed and the latent variables are continuous and Gaussian described

[2]The EM algorithm presented shares some similarities with the EM for LDS (see [16, Ch. 13], [17, Ch. 13], [18, Ch. 13], [19, Ch. 13]).

[3]In the ML problem $\mathbf{V}^{-1}$ was used instead in order to facilitate the computations in the case of $\sigma_x^2 \to 0$.

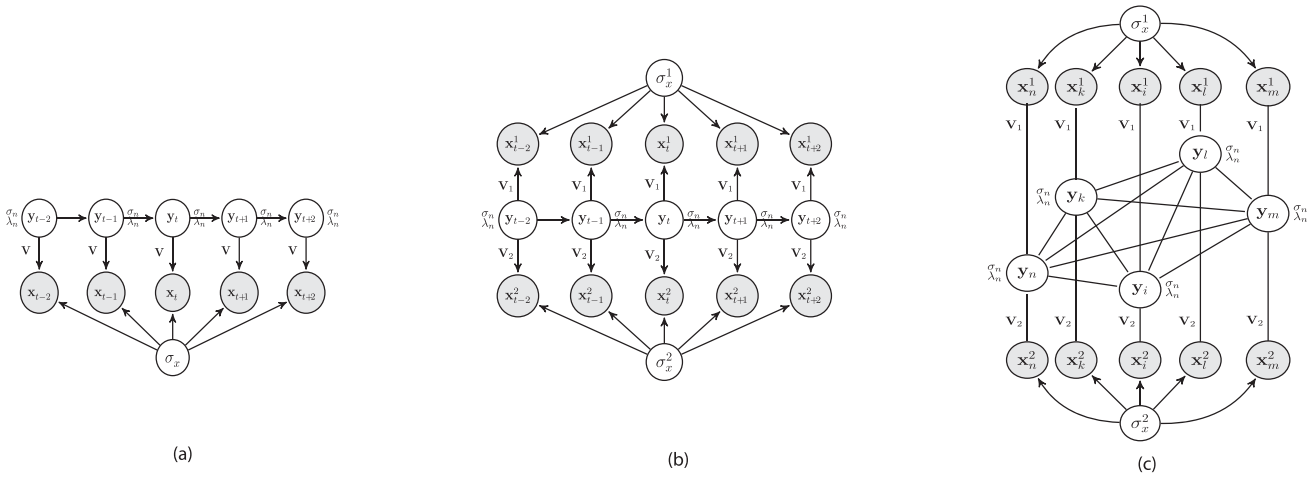

Fig. 2. Linear generative models considered by EM-SFA where latent variables $\mathbf{y}_t$ are connected by a first-order Markov chain. (a) For a single sequence each variable $\mathbf{y}_t$ is associated with a single observation sample. (b) For two sequences, each common latent variable is associated with two observation samples. (c) Undirected graphical model that considers fully connectivity among latent variables for deriving the probabilistic approach for CCA.

by the following generative models, respectively:

$$\mathbf{x}_t = \mathbf{V}\mathbf{y}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \sigma_x^2\mathbf{I}) \tag{12}$$

$$\mathbf{y}_t = \mathbf{\Lambda}\mathbf{y}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{13}$$

$$\mathbf{y}_1 = \mathbf{u}, \qquad \mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1) \tag{14}$$

with the special property that the covariance and the transition matrices are related as $\mathbf{\Lambda}^2 = \mathbf{I} - \mathbf{\Sigma}$. The special properties of the above LDS are (1) in the limit case where the variance is sent to zero $\sigma_x^2 \to 0$, the ML solution is the same as the one produced by SFA, and (2) most importantly there is a ranking of the latent features according to the value of the elements in the main diagonal of $\mathbf{\Lambda}$ or $\mathbf{\Sigma}$ (i.e., the larger the value of λ_n the slower the feature). This is a major advantage, since standard LDS does not have this property.

The joint distribution of such a model is given by

$$P(\mathbf{X}, \mathbf{Y}) = P(\mathbf{y}_1)\prod_{t=2}^{T} P(\mathbf{y}_t|\mathbf{y}_{t-1})\prod_{t=1}^{T} P(\mathbf{x}_t|\mathbf{y}_t) \tag{15}$$

where the initial latent variable, the transition, and emission distributions are given by

$$P(\mathbf{y}_1) = \mathcal{N}(\mathbf{y}_1|\mathbf{0}, \mathbf{\Sigma}_1) \tag{16}$$

$$P(\mathbf{y}_t|\mathbf{y}_{t-1}) = \mathcal{N}(\mathbf{y}_t|\mathbf{\Lambda}\mathbf{y}_{t-1}, \mathbf{\Sigma}) \tag{17}$$

$$P(\mathbf{x}_t|\mathbf{y}_t) = \mathcal{N}(\mathbf{x}_t|\mathbf{V}\mathbf{y}_t, \sigma_x^2\mathbf{I}). \tag{18}$$

Therefore, our objective is to determine the model parameters $\theta = \{\theta_x, \mathbf{\Sigma}_1, \mathbf{\Lambda}\}$ (since $\mathbf{\Sigma}$ is a function of $\mathbf{\Lambda}$) that maximize the joint likelihood, which is equivalent to maximizing the complete log likelihood defined by

$$\log P(\mathbf{X}, \mathbf{Y}|\theta) = \sum_{t=1}^{T} \log P(\mathbf{x}_t|\mathbf{y}_t, \theta_x) + \log P(\mathbf{y}_1|\mathbf{\Sigma}_1) + \sum_{t=2}^{T} \log P(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{\Lambda}). \tag{19}$$

The EM procedure requires the sufficient statistics of the posterior distributions $\mathbb{E}[\mathbf{y}_t|\mathbf{X}]$, $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}]$, and $\mathbb{E}[\mathbf{y}_t\mathbf{y}_{t-1}^T|\mathbf{X}]$ to be determined in order to become available for the maximization step.

B. Learning the Sufficient Statistics $\mathbb{E}[\mathbf{z}_t]$, $\mathbb{E}[\mathbf{z}_t\mathbf{z}_t^T]$, and $\mathbb{E}[\mathbf{z}_t\mathbf{z}_{t-1}^T]$

The sufficient statistics in EM-SFA (known as the E-step) are obtained similarly as for LDSs. In more detail, the estimation consists of the forward and backward steps, known as the Kalman filter [20] and Rauch–Tung–Streibel (RTS) [21], [22] equations, respectively. The algorithm of the inference step is described in Algorithm 1, where we consider the set of parameters ($\theta = \mathbf{V}, \sigma_x, \Sigma_1, \Lambda$) as known. The forward step, which is the first part of the algorithm, recovers the latent marginal $P(\mathbf{y}_t|\mathbf{x}_1, \ldots, \mathbf{x}_t) \sim \mathcal{N}(\mathbf{y}_t|\boldsymbol{\mu}_t, \mathbf{U}_t)$. Notice how the Kalman gain matrix $\mathbf{K}_t$ update differs from the traditional LDS via $\mathbf{P}_{t-1}$, where $\mathbf{P}_{t-1}$ is the variance of the resulting distribution obtained from the following integral:

$$\int \mathcal{N}(\mathbf{y}_{t-1}|\boldsymbol{\mu}_{t-1}, \mathbf{U}_{t-1})\mathcal{N}(\mathbf{y}_t|\mathbf{\Lambda}\mathbf{y}_{t-1}, \mathbf{\Sigma})d\mathbf{y}_{t-1} = \mathcal{N}(\mathbf{y}_t|\mathbf{\Lambda}\boldsymbol{\mu}_{t-1}, \mathbf{P}_{t-1}) \tag{20}$$

which needs to be evaluated in order to obtain the posterior marginal [16]. Since we have forced $\mathbf{\Sigma} = \mathbf{I} - \mathbf{\Lambda}^2$, in our case $\mathbf{P}_t$ is given by

$$\mathbf{P}_{t-1} = \mathbf{\Lambda}(\mathbf{U}_{t-1} - \mathbf{I})\mathbf{\Lambda}^T + \mathbf{I}. \tag{21}$$

Furthermore, we note that the initial parameters of EM-SFA are similar to the traditional LDS

$$\boldsymbol{\mu}_1 = \mathbf{K}_1\mathbf{x}_1 \tag{22}$$

$$\mathbf{U}_1 = (\mathbf{I} - \mathbf{K}_1\mathbf{V})\mathbf{\Sigma}_1 \tag{23}$$

$$c_1 = \mathcal{N}(\mathbf{x}_1|\mathbf{0}, \mathbf{V}\mathbf{\Sigma}_1\mathbf{V}^T + \sigma_x^2\mathbf{I}) \tag{24}$$

where

$$\mathbf{K}_1 = \mathbf{\Sigma}_1\mathbf{V}^T(\mathbf{V}\mathbf{\Sigma}_1\mathbf{V}^T + \sigma_x^2\mathbf{I})^{-1} \tag{25}$$

and c_1 is the normalization coefficient.

The backward step aims to recover the posterior marginal for $\mathbf{y}_t$ given all observations. The updates can be found in the second part of Algorithm 1 and are derived similarly to the updates for the traditional LDS [16]. Having recovered

Algorithm 1: Inference In SFA

Data: $\mathbf{\Lambda}, \mathbf{V}, \sigma_x^2, \mathbf{\Sigma}_1$
Result: $\mathbb{E}[\mathbf{y}_t]$, $\hat{\mathbf{U}}_t$, $\hat{\mathbf{U}}_{t,t-1}$

%(forward step)
for $\mathbf{t = 1}$ *to* $\mathbf{T}$ **do**
  $\mathbf{P}_{t-1} \leftarrow \mathbf{\Lambda}(\mathbf{U}_{t-1} - \mathbf{I})\mathbf{\Lambda}^T + \mathbf{I}$ if $\mathbf{t > 1}$
  $\mathbf{K}_t \leftarrow \mathbf{P}_{t-1}\mathbf{V}^T(\mathbf{V}\mathbf{P}_{t-1}\mathbf{V}^T + \sigma_x^2\mathbf{I})^{-1}$
  $\boldsymbol{\mu}_t \leftarrow \mathbf{\Lambda}\boldsymbol{\mu}_{t-1} + \mathbf{K}_t(\mathbf{x}_t - \mathbf{V}\mathbf{\Lambda}\boldsymbol{\mu}_{t-1})$ if $\mathbf{t > 1}$
  $\mathbf{U}_t \leftarrow (\mathbf{I} - \mathbf{K}_t\mathbf{V})\mathbf{P}_{t-1}$
initialize $\hat{\mathbf{U}}_{T,T-1} \leftarrow (\mathbf{I} - \mathbf{K}_T\mathbf{V})\mathbf{\Lambda}\mathbf{P}_{t-1}$
%(backward step)
for $\mathbf{t = T}$ *to* $\mathbf{2}$ **do**
  $\mathbf{J}_{t-1} \leftarrow \mathbf{U}_{t-1}\mathbf{\Lambda}^T(\mathbf{P}_{t-1})^{-1}$
  $\mathbb{E}[\mathbf{y}_{t-1}] \leftarrow \boldsymbol{\mu}_{t-1} + \mathbf{J}_{t-1}(\mathbb{E}[\mathbf{y}_t] - \mathbf{\Lambda}\boldsymbol{\mu}_{t-1})$
  $\hat{\mathbf{U}}_{t-1} \leftarrow \mathbf{U}_{t-1} + \mathbf{J}_{t-1}(\hat{\mathbf{U}}_t - \mathbf{P}_{t-1})\mathbf{J}_{t-1}^T$
  $\hat{\mathbf{U}}_{t,t-1} \leftarrow \mathbf{U}_t\mathbf{J}_{t-1}^T + \mathbf{J}_t(\hat{\mathbf{U}}_{t+1,t} - \mathbf{\Lambda}\mathbf{U}_t)\mathbf{J}_{t-1}^T$ if $\mathbf{t < T}$

the statistics from Algorithm 1, the sufficient statistics of our model are given by

$$\mathbb{E}[\mathbf{y}_t] = \hat{\boldsymbol{\mu}}_t \tag{26}$$

$$\mathbb{E}\left[\mathbf{y}_t\mathbf{y}_t^T\right] = \hat{\mathbf{U}}_t + \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_t^T \tag{27}$$

$$\mathbb{E}\left[\mathbf{y}_t\mathbf{y}_{t-1}^T\right] = \mathbf{J}_{t-1}\hat{\mathbf{U}}_t + \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_{t-1}^T \tag{28}$$

where $\mathbf{J}_t = \mathbf{U}_t\mathbf{\Lambda}(\mathbf{P}_t)^{-1}$.

C. Learning the Parameters

So far, we have considered that the model parameters $\theta = \{\mathbf{V}, \sigma_x, \mathbf{\Sigma}_1, \mathbf{\Lambda}\}$ are known in order to evaluate the sufficient statistics. In this section, we provide the detailed derivation of the model parameters by assuming that the sufficient statistics are fixed. Therefore, considering the complete likelihood given by (19) we wish to determine those parameters by optimizing

$$\begin{aligned}\theta_o &= \arg\max_{\theta} \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}\left[\log P(\mathbf{X}, \mathbf{Y}|\theta)\right] \\ &= \arg\max_{\theta} \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}\left[\sum_{t=1}^{T}\log P(\mathbf{x}_t|\mathbf{y}_t, \theta_x)\right] \\ &\quad + \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}\left[P(\mathbf{y}_1|\mathbf{\Sigma}_1)\right] + \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}\left[\sum_{t=2}^{T}\log P(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{\Lambda})\right]\end{aligned} \tag{29}$$

which can be split into three parts. Expanding the first part $\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[\sum_{t=1}^{T}\log P(\mathbf{x}_t|\mathbf{y}_t, \theta_x)]$, which involves parameters $\mathbf{V}$ and σ_x^2, we derive

$$\begin{aligned}&\{\mathbf{V}^{\text{new}}, (\sigma_x^{\text{new}})^2\} \\ &= \arg\max_{\mathbf{V}, \sigma_x^2} \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}\left[\sum_{t=1}^{T}\log P(\mathbf{x}_t|\mathbf{y}_t, \theta_x)\right] \\ &= \arg\max_{\mathbf{V}, \sigma_x^2} -\frac{NT}{2}\ln\left(2\pi\sigma_x^2\right) \\ &\quad - \frac{1}{2\sigma_x^2}\sum_{t=1}^{T}\left(\text{tr}(\mathbf{x}_t\mathbf{x}_t^T) - 2\mathbf{x}_t^T\mathbf{V}\,\mathbb{E}[\mathbf{y}_t|\mathbf{X}] + \text{tr}(\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}]\mathbf{V}^T\mathbf{V})\right).\end{aligned}$$

Setting the derivatives with respect to $\mathbf{V}^{\text{new}}$ and $(\sigma_x^{\text{new}})^2$ equal to zero, we obtain the updates

$$\mathbf{V}^{\text{new}} = \left(\sum_{t=1}^{T}\mathbf{x}_t\,\mathbb{E}\left[\mathbf{y}_t^T|\mathbf{Y}\right]\right)\left(\sum_{t=1}^{T}\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{Y}]\right)^{-1} \tag{30}$$

$$\begin{aligned}\left(\sigma_x^{\text{new}}\right)^2 = \frac{1}{NT}\sum_{t=1}^{T}&\left(\text{tr}(\mathbf{x}_t\mathbf{x}_t^T) - 2\mathbf{x}_t^T\mathbf{V}^{\text{new}}\,\mathbb{E}\left[\mathbf{y}_t|\mathbf{Y}\right]\right. \\ &\left. + \text{tr}\left(\mathbb{E}\left[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{Y}\right](\mathbf{V}^{\text{new}})^T\mathbf{V}^{\text{new}}\right)\right).\end{aligned} \tag{31}$$

By maximizing the second part $\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[P(\mathbf{y}_1|\mathbf{\Sigma}_1)]$, we find the updates for the observed variance, $\mathbf{\Sigma}_1$ as

$$\mathbf{\Sigma}_1^o = \arg\max_{\mathbf{\Sigma}_1} \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}[P(\mathbf{y}_1|\mathbf{\Sigma}_1)] \tag{32}$$

from which we derive $\mathbf{\Sigma}_1^o = \mathbb{E}[\mathbf{y}_1\mathbf{y}_1^T|\mathbf{X}]$.

Finally, for parameters $\mathbf{\Lambda}$, by applying the constraint $\sigma_n^2 = 1 - \lambda_n^2$ we maximize the third part $\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[\sum_{t=2}^{T}\log P(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{\Lambda})]$

$$\begin{aligned}\mathbf{\Lambda} &= \arg\max_{\mathbf{\Lambda}} \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}\left[\log\sum_{t=2}^{T}p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{\Lambda})\right] \\ &= \arg\max_{\mathbf{\Lambda}} -\frac{1}{2}\sum_{t=2}^{T}\left[\sum_{n=1}^{N}\ln\left(1-\lambda_n^2\right) + \frac{1}{1-\lambda_n^2}\sum_{n=1}^{N}\right. \\ &\quad \times\left(\mathbb{E}\left[y_{n,t}^2|\mathbf{X}\right] - 2\lambda_n\,\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]\right. \\ &\quad \left.\left. + \lambda_n^2\,\mathbb{E}\left[y_{n,t-1}^2|\mathbf{X}\right]\right)\right] + \text{const}\end{aligned} \tag{33}$$

where by computing the first-order derivative with respect to λ_n, we derive the following cubic equation:

$$\begin{aligned}\sum_{t=2}^{T}&\left(\left(\lambda_n^{\text{new}}\right)^3 - \mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]\left(\lambda_n^{\text{new}}\right)^2\right. \\ &+ \left(\mathbb{E}\left[y_{n,t}^2|\mathbf{X}\right] + \mathbb{E}\left[y_{n,t-1}^2|\mathbf{X}\right] - 1\right)\lambda_n^{\text{new}} \\ &\left. - \mathbb{E}\left[y_{n,t}y_{n,t-1}|\mathbf{X}\right]\right) = 0.\end{aligned} \tag{34}$$

The discriminant of the above cubic equation is given as follows:

$$\begin{aligned}\Delta = \sum_{t=2}^{T}&\left(18\left(\mathbb{E}\left[y_{n,t}^2|\mathbf{X}\right] + \mathbb{E}\left[y_{n,t-1}^2|\mathbf{X}\right] - 1\right)\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]^2\right. \\ &- 4\,\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]^4 + \left(\mathbb{E}\left[y_{n,t}^2|\mathbf{X}\right] + \mathbb{E}\left[y_{n,t-1}^2|\mathbf{X}\right] - 1\right)^2 \\ &\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]^2 - 4\left(\mathbb{E}\left[y_{n,t}^2|\mathbf{X}\right] + \mathbb{E}\left[y_{n,t-1}^2|\mathbf{X}\right] - 1\right)^3 \\ &\left. - 27\,\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]^2\right).\end{aligned} \tag{35}$$

According to the discriminant value, we can consider the following cases.

1) If $\Delta > 0$, then the equation has three distinct real roots.
2) If $\Delta = 0$, then the equation has a multiple root and all its roots are real.
3) If $\Delta < 0$, then the equation has one real root and two nonreal complex conjugate roots.

The three solutions of (34)

$$\lambda_{n_1} = \sum_{t=2}^{T}\left(\frac{\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]}{3} - \frac{1}{3}\sqrt[3]{\frac{1}{2}\left[a_t + \sqrt{a_t^2+b_t}\right]} - \frac{1}{3}\sqrt[3]{\frac{1}{2}\left[a_t - \sqrt{a_t^2+b_t}\right]}\right) \quad (36)$$

$$\lambda_{n_2} = \sum_{t=2}^{T}\left(\frac{\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]}{3} + \frac{1+i\sqrt{3}}{6}\sqrt[3]{\frac{1}{2}\left[a_t + \sqrt{a_t^2+b_t}\right]} + \frac{1-i\sqrt{3}}{6}\sqrt[3]{\frac{1}{2}\left[a_t - \sqrt{a_t^2+b_t}\right]}\right) \quad (37)$$

$$\lambda_{n_3} = \sum_{t=2}^{T}\left(\frac{\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]}{3} + \frac{1-i\sqrt{3}}{6}\sqrt[3]{\frac{1}{2}\left[a_t + \sqrt{a_t^2+b_t}\right]} + \frac{1+i\sqrt{3}}{6}\sqrt[3]{\frac{1}{2}\left[a_t - \sqrt{a_t^2+b_t}\right]}\right) \quad (38)$$

where the constants a_t and b_t are given by

$$a_t = -2\,\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]^3 + 9\,\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]\left(\mathbb{E}\left[y_{n,t}^2|\mathbf{X}\right] + \mathbb{E}\left[y_{n,t-1}^2|\mathbf{X}\right] - 1\right) - 27\,\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}] \quad (39)$$

$$b_t = -4(\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]^2 - 3(\mathbb{E}\left[y_{n,t}^2|\mathbf{X}\right] + \mathbb{E}\left[y_{n,t-1}^2|\mathbf{X}\right] - 1))^3. \quad (40)$$

From the above solutions, we retain for each feature the one that satisfies the condition $0 < \lambda_n < 1$.

D. Inference and Learning in SFA

Next, we briefly describe the algorithm's implementation for computing the posterior means and covariances. In particular, this algorithm has been divided into two parts: 1) it uses the observations from $\mathbf{y}_1$ to $\mathbf{y}_t$ (forward recursion) known as the Kalman filter [23] and 2) it exploits the observations from $\mathbf{y}_{t+1}$ to $\mathbf{y}_T$ (backward recursion) [24].

The EM step for learning the SFA algorithm is given in Algorithm 2 and concerning a single sequence of observations.

Algorithm 2: LEARNING IN SFA

Data: $\mathbf{X}, iter, q$
Result: $\mathbf{\Lambda}, \mathbf{V}, \sigma_x^2, \mathbf{\Sigma}, \mathbf{\Sigma}_1$
initialize $\mathbf{\Lambda}, \mathbf{V}, \sigma_x^2, \mathbf{\Sigma}, \mathbf{\Sigma}_1$
set $\alpha \leftarrow \sum_t \mathbf{x}_t\mathbf{x}_t^T$
SFAInference($\mathbf{\Lambda}, \mathbf{V}, \sigma_x^2, \mathbf{\Sigma}_1$) **% E step**
while *log likelihood* $> q$ *or maxiter* $< iter$ **do**
  initialize $\delta \leftarrow \mathbf{0}$, $\delta_1 \leftarrow \mathbf{0}$, $\gamma \leftarrow \mathbf{0}$
  for $\mathbf{t = 1}$ *to* $\mathbf{T}$ **do**
    $\delta \leftarrow \delta + \mathbf{x}_t\,\mathbb{E}[\mathbf{y}_t]^T$
    $\delta_1 \leftarrow \delta_1 + \mathbf{x}_t^T\mathbf{V}\,\mathbb{E}[\mathbf{y}_t]$
    $\gamma \leftarrow \gamma + \mathbb{E}[\mathbf{y}_t]\,\mathbb{E}[\mathbf{y}_t]^T + \hat{\mathbf{U}}_t$
    if $t > 1$ **then**
      for $\mathbf{n = 1}$ *to* $\mathbf{N}$ **do**
        update λ_{n1}, λ_{n2} and λ_{n3} (Eqs. (36), (37) and (38))
  % M step
  $\mathbf{V} \leftarrow \delta\gamma^{-1}$
  $\sigma_x^2 \leftarrow \frac{1}{TM}(\mathrm{tr}(\alpha) - 2\delta_1 + \mathrm{tr}(\gamma\,\mathbf{V}^T\mathbf{V}))$
  for $\mathbf{i = 1}$ *to* ***3*** **do**
    if $0 > \lambda_{n,i} > 1$ **then**
      $\lambda_n \leftarrow \lambda_{n,i}$
  $\mathbf{\Lambda} \leftarrow [\delta_{i,j}\lambda_n]$
  $\mathbf{\Sigma} \leftarrow \mathbf{I} - \mathbf{\Lambda}^2$

E. EM-SFA for Two Sequences

In the following, we propose a generative probabilistic model for finding the common higher order, slowest varying feature between the two sequences. The corresponding graphical model is shown in Fig. 2(b). To do so, let us assume the following generative model for the samples of the following time-varying input sequences $\mathbf{X}_1 = [\mathbf{x}_t^1, t \in [1, T]] \in \Re^{M_1\times T}$ and $\mathbf{X}_2 = [\mathbf{x}_t^2, t \in [1, T]] \in \Re^{M_2\times T}$:

$$\mathbf{x}_t^k = \mathbf{V}_k\mathbf{y}_t + \mathbf{e}_t^k, \quad \mathbf{e}_t^k \sim \mathcal{N}(0, \sigma_{x,k}^2\mathbf{I}), \quad k = 1, 2 \quad (41)$$

where each sequence has different loads $\mathbf{V}_1$ and $\mathbf{V}_2$ and noise, while both sequences share a common latent space $\mathbf{Y}$ with $P(\mathbf{Y}|\theta_y)$ given by (8). The complete joint likelihood distribution $P(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y})$ is of the form

$$\log P(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}|\theta) = \log P(\mathbf{y}_1|0, \mathbf{\Sigma_1}) + \sum_{t=2}^{T}\log P(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{\Lambda}) + \sum_{t=1}^{T}\log P\left(\mathbf{x}_t^1|\mathbf{y}_t, \mathbf{V}_1, \sigma_{x,1}^2\right) + \sum_{t=1}^{T}\log P\left(\mathbf{x}_t^2|\mathbf{y}_t, \mathbf{V}_2, \sigma_{x,2}^2\right) \quad (42)$$

where now $\theta = \{\theta_x^1, \theta_x^2, \mathbf{\Sigma}_1, \mathbf{\Lambda}\}$ with $\theta_x^1 = \{\mathbf{V}_1, \sigma_{x,1}^2\}$ and $\theta_x^2 = \{\mathbf{V}_2, \sigma_{x,2}^2\}$.

For the two-sequence SFA, in the expectation step we need to compute $\mathbb{E}[\mathbf{y}_t|\mathbf{X}_1, \mathbf{X}_2]$, $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}_1, \mathbf{X}_2]$, and $\mathbb{E}[\mathbf{y}_t\mathbf{y}_{t-1}^T|\mathbf{X}_1, \mathbf{X}_2]$ which can be also performed using RTS smoothing, as in Section IV-A. Applying the maximization step on the joint log likelihood (42), we obtain the updates for $\mathbf{V}_1, \mathbf{V}_2$, and $\sigma_{x,1}^2, \sigma_{x,2}^2$ as

$$\mathbf{V}_k^{\text{new}} = \left(\sum_{t=1}^{T}\mathbf{x}_t^k\,\mathbb{E}\left[\mathbf{y}_t^T|\mathbf{X}^{\text{tot}}\right]\right)\left(\sum_{t=1}^{T}\mathbb{E}\left[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}^{\text{tot}}\right]\right)^{-1} \quad (43)$$

$$(\sigma_{x,k}^{\text{new}})^2 = \frac{1}{M_kT}\sum_{t=1}^{T}\Big(\mathrm{tr}\big(\mathbf{x}_t^k(\mathbf{x}_t^k)^T\big) - 2(\mathbf{x}_t^k)^T\mathbf{V}_k^{\text{new}}\,\mathbb{E}\left[\mathbf{y}_t|\mathbf{X}^{\text{tot}}\right] + \mathrm{tr}\big(\mathbb{E}\left[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}^{\text{tot}}\right](\mathbf{V}_k^{\text{new}})^T\mathbf{V}_k^{\text{new}}\big)\Big), \quad k = 1, 2. \quad (44)$$

Regarding $\mathbf{\Lambda}$ and $\mathbf{\Sigma_1}$ the updates are given by (32) and (33), applied using the derived $\mathbb{E}[\mathbf{y}_t|\mathbf{X}_1, \mathbf{X}_2]$, $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}_1, \mathbf{X}_2]$,

Algorithm 3: EMSFA With DTW

Data: $\mathbf{X}_1, \ldots, \mathbf{X}_K, iter, q$
Result: $\Delta_1, \ldots, \Delta_K, \mathbb{E}[\mathbf{Y}|\mathbf{X}_1^{\Delta}, \ldots, \mathbf{X}_K^{\Delta}]$
while *not converged* **do**
  if $iter = 1$ **then**
    $(\Delta_1, \ldots, \Delta_K) \leftarrow \text{DTW}(\mathbf{X}_1, \ldots, \mathbf{X}_K)$
  else
    $(\boldsymbol{\Delta}_1, \ldots, \Delta_K) \leftarrow \text{DTW}(\mathbb{E}[\mathbf{Y}|\mathbf{X}_1], \ldots, \mathbb{E}[\mathbf{Y}|\mathbf{X}_K])$
  $\mathbf{X}_1^{\Delta} \leftarrow \mathbf{X}_1\Delta_1, \ldots, \mathbf{X}_K^{\Delta} \leftarrow \mathbf{X}_K\Delta_K$
  while *not converged* **do**
    Update θ (Eqs. (43), (44), (33) and (32))
    Update $\boldsymbol{\Sigma}$ acc. to $\sigma_n^2 = 1 - \lambda_n^2$
    $\mathbb{E}[\mathbf{Y}|\mathbf{X}_1^{\Delta}, \ldots, \mathbf{X}_K^{\Delta}] \leftarrow$ $\text{RTS}(\mathbf{X}_{\text{tot}}^{\Delta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{V}, \sigma_{x,tot}^2, \boldsymbol{\Sigma}_1)$
  $\sigma_{x,1}^2, \ldots, \sigma_{x,K}^2 \leftarrow \sigma_x^{\text{tot}}\mathbf{I}_M =$ $\begin{pmatrix} \sigma_{x,1}^2\mathbf{I}_{M_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \ldots & \ldots\ldots & \ldots & \\ \mathbf{0} & \mathbf{0} & \ldots & \sigma_{x,K}^2\mathbf{I}_{M_K} \end{pmatrix}$
  $\mathbf{V}_1, \ldots, \mathbf{V}_K \leftarrow \mathbf{V} = \begin{pmatrix} \mathbf{V}_1 \\ \ldots \\ \mathbf{V}_K \end{pmatrix}$
  $\mathbb{E}[\mathbf{Y}|\mathbf{X}_k] \leftarrow \text{RTS}(\mathbf{X}_k, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{V}_k, \sigma_{x,k}^2, \boldsymbol{\Sigma}_1)$, $k = 1, 2, \ldots, K$

and $\mathbb{E}[\mathbf{y}_t\mathbf{y}_{t-1}^T|\mathbf{X}_1, \mathbf{X}_2]$. Using the above expositions the K-sequence case can be trivially derived.

F. Aligning Observed Sequences

In this section, we propose an algorithm that uses the latent spaces provided by the two-sequence EM-SFA for time series alignment. We claim that since the two-sequence EM-SFA provides the slowest varying common features, these features would be well-suited for time series alignment. In essence, this translates to aligning the slowest varying features from two sequences which means that we disregard high-frequency features that are likely to be noisy. We note that recently, time series alignment was performed on a space recovered by the application of CCA [25]. A simple, commonly used [25], and optimal method for finding the warpings is DTW,[4] which we employ in our case. Given K sequences $\mathbf{X}_1 \in \Re^{M_1 \times T_1}, \ldots, \mathbf{X}_K \in \Re^{M_K \times T_K}$ of different lengths $T_1 \neq \ldots \neq T_K$, our aim is to find the warpings $\Delta_1 \in \Re^{T_1 \times T}, \ldots, \Delta_K \in \Re^{T_K \times T}$ such that the common latent space will have common length of size T. The alignment algorithm is presented in Algorithm 3.

V. CCA and SFA: What Is the Difference?

In this section, we will shed further light on the relation of the proposed two-sequence EM-SFA and CCA, and show that CCA can be derived from a generative probabilistic model having as a fully connected MRF prior over the latent space. Fig. 2(c) shows graphically such model. First, it is important to note that both techniques follow the same generative model

$$\mathbf{x}_t^k = \hat{\mathbf{V}}_k^{-1}\mathbf{y}_t + \boldsymbol{\epsilon}_t^k, \quad \boldsymbol{\epsilon}_t^k \sim \mathcal{N}(0, \sigma_k^2 I), \quad k = 1, 2$$
$$\mathbf{X}_t^{\text{tot}} = \begin{bmatrix} \mathbf{x}_t^1 \\ \mathbf{x}_t^2 \end{bmatrix} = \mathbf{V}^{-1}\mathbf{y}_t + \begin{bmatrix} \boldsymbol{\epsilon}_t^1 \\ \boldsymbol{\epsilon}_t^2 \end{bmatrix}. \tag{45}$$

Based on the above generative model, we derive a ML solution for SFA (Section V-A) as well as for CCA (Section V-B). Finally, we attempt to determine the relationship of SFA and CCA by studying the resulting optimization problem.

A. Deterministic SFA for Two Sequences

We consider the generative model formulated in (45). By computing the marginal $\log P(\mathbf{X}_1, \mathbf{X}_2|\theta)$ (i.e., marginalizing out the latent space) and taking the limits $\lim\{\sigma_{x,1}^2, \sigma_{x,2}^2\} \to 0$, $T \to \infty$, we obtain

$$\begin{aligned} &\log P(\mathbf{X}_1, \mathbf{X}_2|\theta) \qquad (46)\\ &= \log \int_{\mathbf{Y}} \prod_{t=1}^{T} P\big(\mathbf{X}_t^{\text{tot}}|\mathbf{y}_t, \theta_{x_1}, \theta_{x_2}\big) P(\mathbf{Y}|\theta_y) d\mathbf{Y} \\ &= \log \int_{\lim\{\sigma_{x,1}^2,\ \sigma_{x,2}^2\} \to 0} \delta\big(\mathbf{X}_t^{\text{tot}} - \mathbf{V}^{-1}\mathbf{y}_t\big) P(\mathbf{Y}|\theta_y) d\mathbf{Y} \\ &= c + T\big(\log|\mathbf{V}_1| + \log|\mathbf{V}_2|\big) \\ &\quad -\frac{T}{2}\text{tr}\left[\begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \mathbf{B} \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \boldsymbol{\Lambda}^{(2)} + \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \mathbf{A} \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \boldsymbol{\Lambda}^{(1)}\right] \end{aligned}$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{X}_1\mathbf{X}_1^T & \mathbf{X}_1\mathbf{X}_2^T \\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{X}_2\mathbf{X}_2^T \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} \dot{\mathbf{X}}_1\dot{\mathbf{X}}_1^T & \dot{\mathbf{X}}_1\dot{\mathbf{X}}_2^T \\ \dot{\mathbf{X}}_2\dot{\mathbf{X}}_1^T & \dot{\mathbf{X}}_2\dot{\mathbf{X}}_2^T \end{bmatrix}. \tag{47}$$

By taking the derivatives and solving for the loadings $\mathbf{V}_1$ and $\mathbf{V}_2$, we arrive at the condition

$$\begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \mathbf{B} \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix} \boldsymbol{\Lambda}^{(2)} + \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \mathbf{A} \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix} \boldsymbol{\Lambda}^{(1)} = \mathbf{I} \tag{48}$$

since $\boldsymbol{\Lambda}^{(2)}$ and $\boldsymbol{\Lambda}^{(1)}$ are diagonal, then the projection bases $\mathbf{V}_1, \mathbf{V}_2$ are given by joint diagonalization of $\mathbf{B}$ and $\mathbf{A}$. Hence, the ML solution of the above probabilistic model gives the same (up to a scale) projection bases as the following trace optimization problem:

$$\begin{aligned} \min_{\mathbf{V}} \quad & \text{tr}\left[\begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \mathbf{A} \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}\right] \\ \text{s.t.} \quad & \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}^T \mathbf{B} \begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix} = \mathbf{I} \end{aligned} \tag{49}$$

which can be solved by keeping the smallest eigenvalues of the following GEP:

$$\mathbf{A}\begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix} = \mathbf{B}\begin{bmatrix}\mathbf{V}_1 \\ \mathbf{V}_2\end{bmatrix}\begin{bmatrix}\mathbf{L}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_2\end{bmatrix}. \tag{50}$$

It is straightforward to extend the above methodology such as to identify the common slowest varying features of multiple sequences.

[4] Other methods that can be used include, for example, [26], while for related work from functional data analysis, please see [27]–[29].

B. CCA for Finding the Common Latent Space

Next, we show that CCA can be given as a limit case of a probabilistic generative model with a fully connected MRF prior. To the best of our knowledge, this is the first time that CCA is described by such a model.[5] Let us consider again the generative model in (45) and assume a fully connected MRF prior $P(\mathbf{Y})$ over our samples, i.e., each latent node $\mathbf{y}_i$ is connected to all other latent nodes $\mathbf{y}_j$, $j \neq i$

$$\begin{aligned} &P(\mathbf{Y}|\theta_y) \\ &= \frac{1}{Z}\exp\left\{-\frac{1}{2}\sum_{n=1}^{N}\frac{1}{T-1}\sum_{i=1,j=1}^{T}\frac{1}{\sigma_n^2}(y_{n,i}-\lambda_n y_{n,j})^2\right\} \\ &= \frac{1}{Z}\exp\left\{-\frac{1}{2}\left(\mathrm{tr}\left[\mathbf{\Lambda}_{\mathrm{CCA}}^{(1)}\mathbf{Y}\mathbf{Y}^T\right]+\mathrm{tr}\left[\mathbf{\Lambda}_{\mathrm{CCA}}^{(2)}\mathbf{Y}\mathbf{M}\mathbf{Y}^T\right]\right)\right\} \end{aligned} \quad (51)$$

where $\mathbf{M} \triangleq -(1/T)\mathbf{1}\mathbf{1}^T$, $\mathbf{\Lambda}_{\mathrm{CCA}}^{(1)} \triangleq [\delta_{mn}(\lambda_n^2+1/\sigma_n^2)]$, $\mathbf{\Lambda}_{\mathrm{CCA}}^{(2)} \triangleq [\delta_{mn}(2\lambda_n/\sigma_n^2)]$.

Following similar steps as in Section III by taking the $\lim_{\sigma_{x,1}^2,\sigma_{x,2}^2\to 0}\int_{\mathbf{Y}}\log P(\mathbf{X}|\mathbf{Y})P(\mathbf{Y}|\theta_y)d\mathbf{Y}$ we arrive at:

$$\begin{aligned} L(\mathbf{V}_1,\mathbf{V}_2) = c + T\left(\log|\mathbf{V}_1|+\log|\mathbf{V}_2|\right) \\ -\frac{T}{2}\mathrm{tr}\left[\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T\mathbf{B}\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T\mathbf{\Lambda}_{\mathrm{CCA}}^{(1)}\right]. \end{aligned} \quad (52)$$

Forcing $(\partial\mathcal{L}/\partial\mathbf{V}_1)=0$ and $(\partial\mathcal{L}/\partial\mathbf{V}_2)=0$, we obtain

$$\begin{aligned} &\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T\begin{bmatrix}\mathbf{X}_1\mathbf{X}_1^T & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\mathbf{X}_2^T\end{bmatrix}\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}\mathbf{\Lambda}_{\mathrm{CCA}}^{(1)} \\ &+\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T\begin{bmatrix}\mathbf{0} & \mathbf{X}_1\mathbf{X}_2^T\\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}\mathbf{\Lambda}_{\mathrm{CCA}}^{(1)} = \mathbf{I} \end{aligned} \quad (53)$$

hence, the solution is given by joint diagonalization of

$$\begin{bmatrix}\mathbf{X}_1\mathbf{X}_1^T & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\mathbf{X}_2^T\end{bmatrix} \text{ and } \begin{bmatrix}\mathbf{0} & \mathbf{X}_1\mathbf{X}_2^T\\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{0}\end{bmatrix}$$

which is equivalent to the CCA problem [30]. As we mentioned above, deterministic CCA is a methodology that is often used for feature extraction combined with alignment of time series (such as with DTW [25]). We claim that the proposed EM-SFA is more suitable for aligning time series, since it incorporates temporal constraints (via the first-order Markov prior), while CCA incorporates a fully connected MRF prior over the latent space, as seen in (51).

VI. Experimental Results

For demonstrating the effectiveness of our proposed methods, experiments were conducted both on synthetic (Section VI-A) and real (Sections VI-B and VI-C) data.

A. Synthetic Data

In this section, we demonstrate the experimental results of our proposed algorithms on synthetic data. We use the dimensionality reduction toolbox [31] in order to generate randomly scaled synthetic examples of 1000 data points each. In Fig. 3, we visualize a comparison between the resulting latent space of the two-sequence EM-SFA and deterministic two-sequence SFA, when applying the algorithms on the two sequences presented in Fig. 3(a) and (b). It is easy to observe that the latent spaces derived by both two-sequence EM-SFA [Fig. 3(d)] and deterministic two-sequence SFA [Fig. 3(c)] follow the same shape, but because the EM-SFA employs an iterative smoothing procedure, the result is much smoother than the deterministic projections.

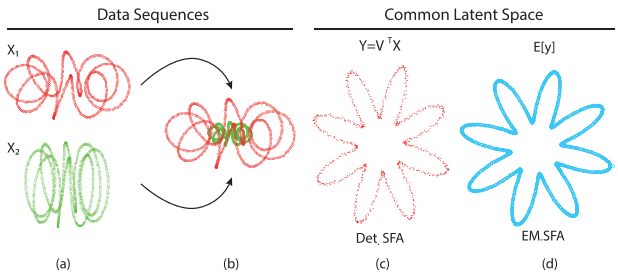

Fig. 3. (a) and (b) Application of deterministic SFA and EM-SFA on two synthetic data sequences $\mathbf{X}_1, \mathbf{X}_2$. (c) and (d) Resulting common latent space.

Next, we examine the ability of the deterministic two-sequence SFA, two-sequence EM-SFA, and CCA algorithms to identify the common latent space of two synthetic sequences contaminated by noise. To derive our synthetic example, we added Gaussian noise on the two randomly scaled helix sequences shown in Fig. 4(a) and (b), while subsequently we applied the examined algorithms on the noisy sequences shown in Fig. 4(c). As it can be observed, two-sequence EM-SFA was able to accurately extract the common latent space of the two sequences shown in Fig. 4(e), while the common latent spaces identified by both the deterministic two-sequence SFA and the CCA algorithms are noisy as can be observed in Fig. 4(d) and (f). This can be attributed to the fact that the extracted common latent space by two-sequence EM-SFA is smoothed and filtered by the RTS algorithm.

B. Real Data 1: Unsupervised AU Temporal Phase Segmentation

Regarding real data, we employ the publicly available MMI database [32] and the UvA-Nemo Smile (UNS) [33] that display both posed and spontaneous expressions. The MMI consists of around 400 videos of 19 subjects annotated in terms of FAUs and their temporal phases, i.e., neutral, onset, apex, and offset. The UNS is a large-scale database having $>$1000 smile videos (597 spontaneous and 643 posed) from 400 subjects. Throughout this section, we use trackings of facial expressions for each subject, resulting in 68 landmark points. The employed tracker is a person-independent implementation of active appearance models, using the normalized gradient features proposed in [34] and is presented in [35].

1) Experiments in MMI Database: For the first experiment, our goal is to measure how effectively EM-SFA can detect the temporal phases of AUs in comparison with a deterministic SFA, the traditional LDSs, and more complex non-LDSs such as dynamic Gaussian process latent variable models (GPLVMs) [36]. In this experiment, for each

[5]The probabilistic approach for CCA in [30] is radically different than ours, since it models only individual random variables; in our case we consider the whole set of variables at once.

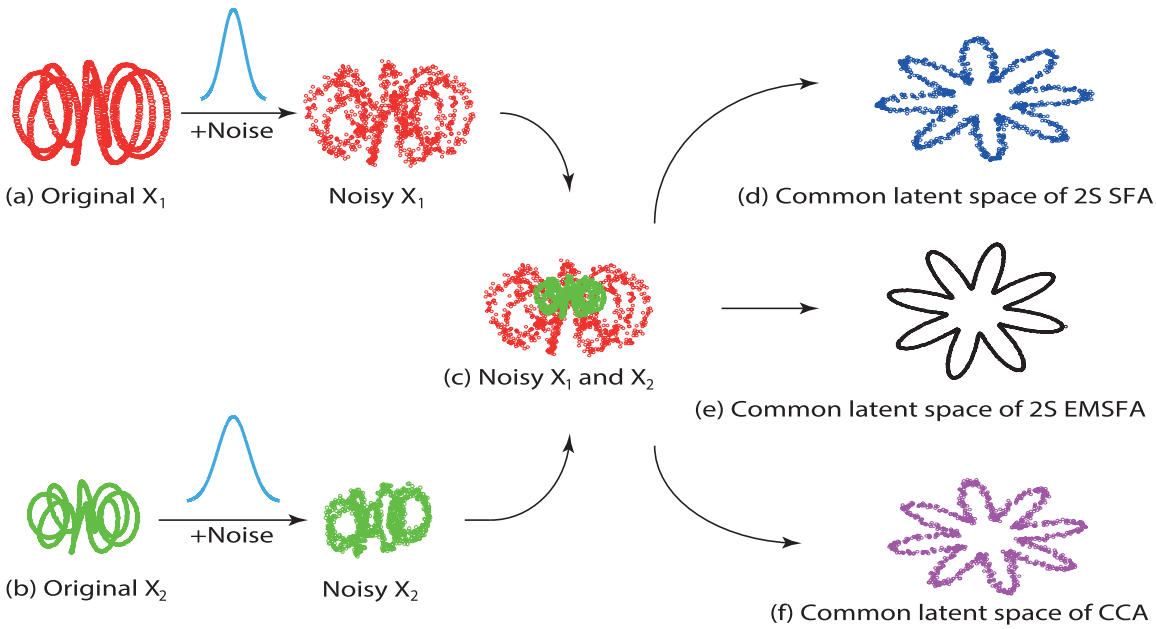

Fig. 4. Synthetic data sequences contaminated by Gaussian noise. (a) and (b) Initial randomly scaled helix sequences. (c) Two noisy sequences. The common latent space derived by (d) deterministic two-sequence SFA, (e) two-sequence EM-SFA, and (f) CCA algorithms.

AU present in the data, we apply the compared algorithms based on the corresponding region of the face (mouth, eyes, brows). We subsequently evaluate the latent space obtained by all methods, and compare with the annotated ground truth.

The temporal dynamics of posed expressions are typically described by the following temporal segments.

1) *Neutral:* Where there is no facial motion.
2) *Onset:* The facial motion starts until it reaches an apex.
3) *Apex:* The point of the strongest possible facial deformation. The person who displays the expression usually stands still for some moments.
4) *Offset:* The reverse path from the apex to the relaxed neutral position.

For more details regarding the temporal segments of posed behavior, please refer to [37] and [38]. The MMI database contains videos that have been annotated with regard to the above-mentioned categories of temporal segments. In more detail, the annotations are as follows: 0 for neutral, 1 for onset, 2 for apex, and 3 for onset. In order to facilitate the description of dynamics, we consider the fact that there is a monotonic increase of facial deformation in the onset region and monotonic decrease of facial deformation in the offset region. Put simply, the intensity of the activated facial muscles monotonically increases until it reaches the apex, and subsequently monotonically decreases. As understandable, the apex of the expression can be discovered by locating zero-crossings of the first derivative. In conclusion, the zero-crossings of the first derivative represent changing points during the activation and relaxation of human facial muscles during posed expressions. An example of the annotations is shown in Fig. 7 along with the gradient of the annotation. As can be seen, the zero-crossings of the derivative indeed represent the changing points between the different temporal phases.

To facilitate the comparison with the ground truth, which is annotated in terms of the temporal phases of facial AUs, we map the recovered latent space to the temporal phases of AUs. This is accomplished using a subset of the critical points of the obtained latent space (the most slowly varying feature). In particular, we are interested in a specific set of zero-crossings of the first-derivative. In more detail, we obtain four points, $x_1, \ldots, x_4$, which correspond to a particular frame in the video [the points are clearly indicated in Fig. 5(a) with circles]. x_1 is the first point where there is a zero-crossing in the first-order derivative, transitioning to positive (signifying an increase in the expression intensity). This point corresponds to the beginning of the onset phase, thus ending the neutral phase of the expression. The x_2 point is obtained by taking the next zero-crossing, which indicates the beginning of the apex phase, i.e., the intensity of the expression has stopped increasing. The zero-crossing of the first-derivative where the value switches to negative is x_3 (signifying the intensity decrease) marks the beginning of the offset phase, while x_4, which is the next zero-crossing of the first-derivative indicates the end of the offset phase and the beginning of the neutral phase. Summarizing, the neutral phase spans from the first frame to x_1 and from x_4 to the end of the video, the onset phase from x_1 to x_2, the apex phase from x_2 to x_3, and finally, the offset phase from x_3 to x_4.

The overall results for the applied methods are summarized in Table I. The presented results show that EM-SFA outperforms deterministic SFA, LDS,[6] and GPLVM on the unsupervised detection of the temporal phases of AUs, for all temporal phases and for all relevant regions of the face. The relevant AUs used for each region of the face are as follows.

1) *Mouth:* Upper lip raiser, nasolabial deepener, lip corner puller, cheek puffer, dimpler, lip corner depressor, lower

[6]In the reported results for the general LDS, we used a diagonal transition matrix. Results with full transition matrix were even worse.

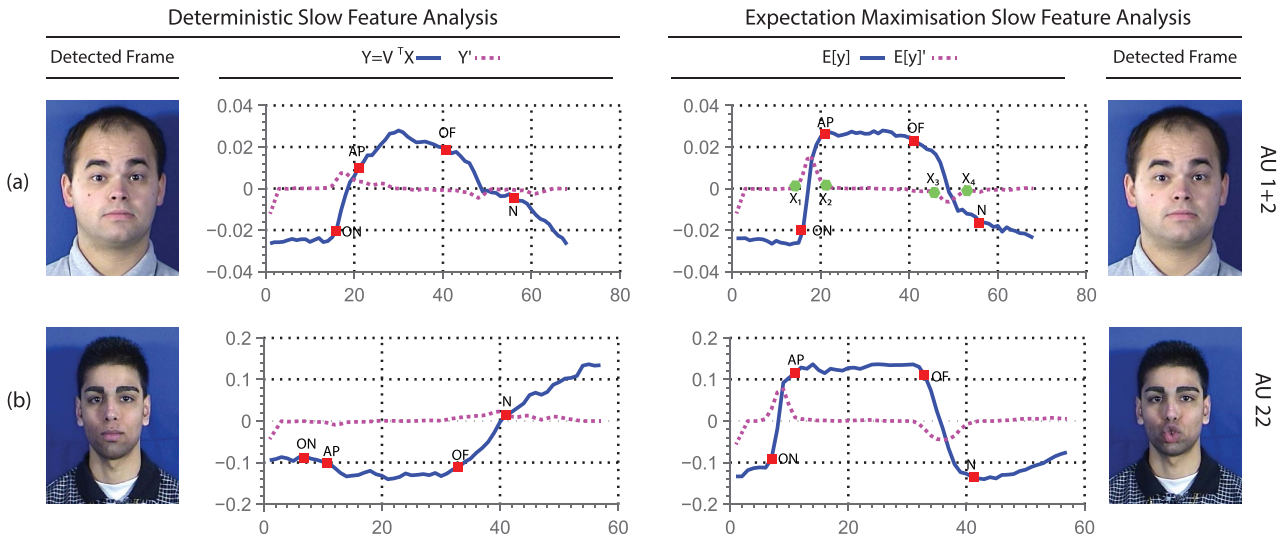

Fig. 5. Comparing the derived latent space (i.e., slowest varying feature) for SFA and EM-SFA, obtained when applying the algorithms on two different videos depicting a subject performing (a) AU 1-2 (Outer Brow Raiser) and (b) AU 22 (Lip Funneler).

TABLE I

PERFORMANCE OF SFA, EMSFA, LDS, AND GPLVM IN TERMS OF EXTRACTING THE GROUND TRUTH FROM AUS RELATED TO MOUTH, EYES, AND BROWS, EVALUATED ON ALL AU TEMPORAL PHASES AND THE EXPRESSION PEAK IN MMI DATABASE

Accuracy (%)															
	Neutral			Onset			Apex			Offset			Expr. Peak		
Method	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows	Mouth	Eyes	Brows
EMSFA	**88.15**	**83.59**	**78.68**	**93.78**	**85**	**100**	**67.76**	**26.67**	**54.59**	**90.05**	**31.48**	**95.52**	**87.5**	**50**	**100**
SFA	69.48	58.77	69.97	90.67	60	87.5	51.97	2	42.35	87.06	22.22	83.58	41.67	7.14	36.36
LDS	67.37	53.16	67.57	91.19	50	81.25	47.86	6.67	45.41	87.56	18.52	77.61	79.17	2	63.64
GPLVM	73.79	68.92	68.49	87.5	65.57	87.68	59.23	13.67	48.82	88.10	18.04	81.07	77.76	25.71	59.08

lip depressor, chin raiser, lip puckerer, lip stretcher, lip funneler, lip tightener, lip pressor, lips part, jaw drop, mouth stretch, and lip suck.

2) *Eyes:* Upper lid raiser, cheek raiser, lid tightener, nose wrinkler, eyes closed, blink, wink, eyes turn left, and eyes turn right.
3) *Brows:* Inner brow raiser, outer brow raiser, and brow lowerer.

Furthermore, in Table I, we show the results for detecting the peak of the expression, i.e., when the intensity of the expression is maximal. This corresponds to the maximum of the derived latent space, and in theory correspond to a frame which is annotated as an apex frame. In this scenario, EM-SFA outperforms all compared methods. We note that the low performance in terms of apex and expression peak for eyes, is due to the fact that most eye-related AUs in the data were blinks, which have a very short apex (most of the times just 1 frame). Therefore, it is very challenging to capture it with slow varying features. Nevertheless, EM-SFA appears to capture the blink apex much better than the compared methods. In Fig. 5, we can visually evaluate the performance of EM-SFA and deterministic SFA on the given problem. Two examples are shown, in Fig. 5(a) and (b), both methods manage to capture the apex of the expression as well as to segment the temporal phases according to the ground truth, with EM-SFA performing better. In Fig. 5(b), deterministic SFA fails to capture the dynamics of the AU, while EM-SFA accurately captures the transition.

A second set of experiments were conducted in order to show how accurately the latent space can capture the dynamics of the behavior. In this set of experiment, we measure the similarity between the ground truth and the extracted latent space by monitoring the alignment cost using the DTW algorithm. Fig. 8(a)–(c) plots the percentage of videos versus the DTW-error for mouth, eyes, and eye-brows related AUs. It is obvious that EM-SFA latent space largely outperforms the other latent space in terms of characterization of the dynamics.

2) Experiments in UNS Database: Next, we test the performance of the examined methods in UNS database. Specifically, we applied all the tested methods for the recognition of temporal segments of 100 videos with posed smiles. An example of the best feature that describes the dynamics of a posed smile of the UNS database can be seen in Fig. 6. The performance was measured similar to the experiments performed on the MMI database. Table II summarizes the performance of the compared methods in terms of the extraction of the temporal segments, while Fig. 10 plots the DTW-error versus the percentage of videos. As can be seen the latent space, produced by the proposed EM-SFA largely outperforms

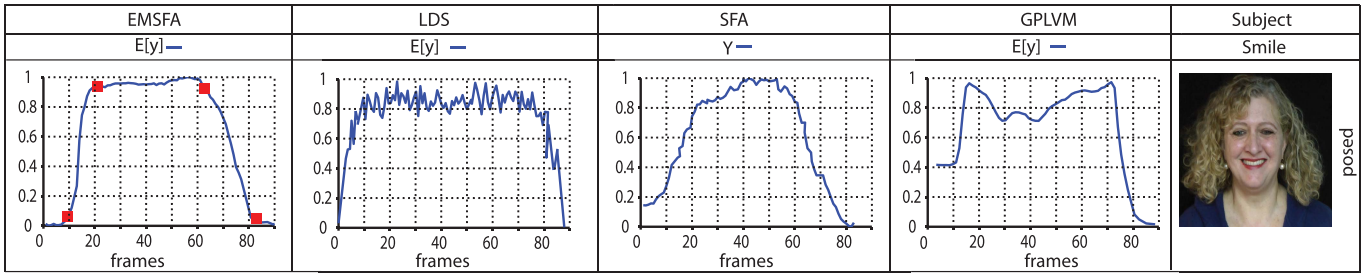

Fig. 6. Extracted features by applying the EMSFA, LDS, SFA, and GPLVM on a video sequence from the UNS database on a subject performing posed smile. Rectangle markers: annotated ground truth where the AU temporal phase changes.

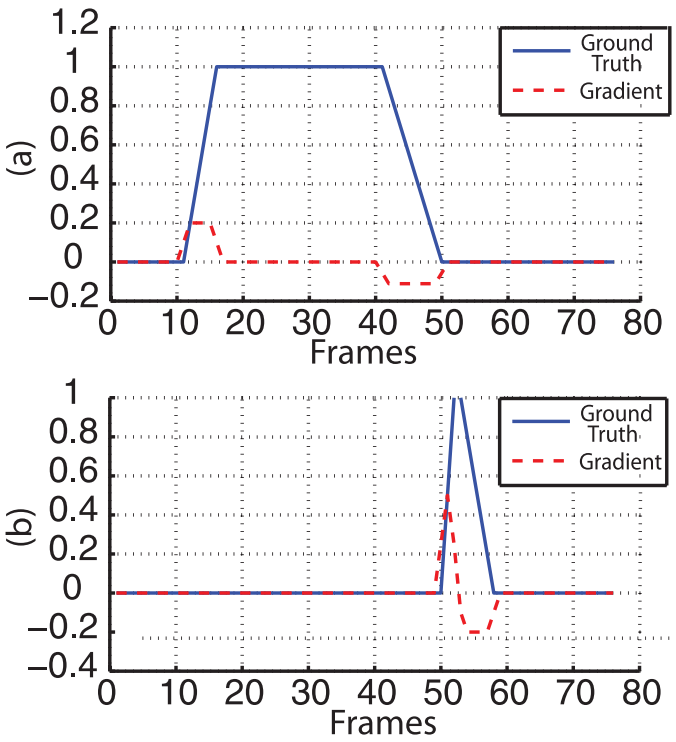

Fig. 7. Ground truth of the various temporal segments of the facial behaviour: (a) of posed AU 12 and (b) of a spontaneous blink (AU 45).

TABLE II
PERFORMANCE OF SFA, EMSFA, LDS, AND GPLVM IN TERMS OF EXTRACTING THE GROUND TRUTH FROM DELIBERATE AND SPONTANEOUS SMILES, EVALUATED ON ALL AU TEMPORAL PHASES IN UNS DATABASE

	Accuracy (%)				
Method	Neutral	Onset	Apex	Offset	Overall
EMSFA	**72.91**	**56.78**	57.82	**39.23**	**56.68**
SFA	61.04	42.99	**60.06**	35.84	47.98
LDS	39.52	19.33	16.11	23.32	24.57
GPLVM	63.87	42.20	43.65	36.88	49.98

all the other methods in the characterization of the dynamics of posed smiles.

Characterization and analysis of human behavior in terms of temporal dynamics is a crucial and understudied problem which is particularly important for the analysis of multiple behavioral events, e.g., it has been argued that temporal dynamics of facial behavior represent a critical factor for distinction between spontaneous and posed facial behavior [39]. In particular, posed behavior, as described above, typically has an onset-apex-offset curve as can be seen in Fig. 9, while the spontaneous behavior may have multiple apexes. This is evidently shown in Fig. 11. As can be seen by the visualizations, the slowest varying feature from EMSFA can accurately describes the dynamics of the complex behavioral phenomenon.

In order to show the usefulness of EMSFA for discrimination of posed versus spontaneous behavior, we have conducted an experiment using the UNS database. We have randomly selected 200 videos of posed and spontaneous smiles, and subsequently we utilized half for training and the remaining half for testing. The training videos were used to train an SVM using a DTW kernel [40] on the slowest varying feature from EMSFA, SFA, and from a LDS. The SVMs exploiting the EMSFA features achieve the highest recognition rate $\sim$73%, while the SVM with features derived via SFA and LDS achieves 64% and 59%, respectively.

C. Real Data 2: Temporal Alignment

In this section, we present results on aligning 50 pairs of videos from the MMI database, where the same AU is activated. The goal of this experiment is to evaluate the space obtained by EM-SFA to that obtained by CCA. Our claim is that the space derived by SFA (essentially recovering the slowest varying feature) will enable better alignment (when combined with DTW) than CCA (when combined with DTW, i.e., CTW [25]). Of major importance to this claim is the modeling of dynamics in EM-SFA, which contrary to the traditional CCA, accounts for temporal dependencies. Results are presented in Fig. 12. The error we used is the percentage of misaligned frames for each pair of videos, normalized per frame (i.e., divided by the aligned video length). We present results on average (for the entire video) and results per temporal phase: neutral, onset, and offset. As can be seen from our results that the derived space of SFA + DTW is better suited for the alignment of temporal sequences than the other compared space obtained by applying CTW.

D. Real Data 3: Conflict Detection

In this section, we aim to show that the latent space obtained by EM-SFA can detect when conflict arises in spontaneous, naturalistic scenarios, where two persons are debating. For

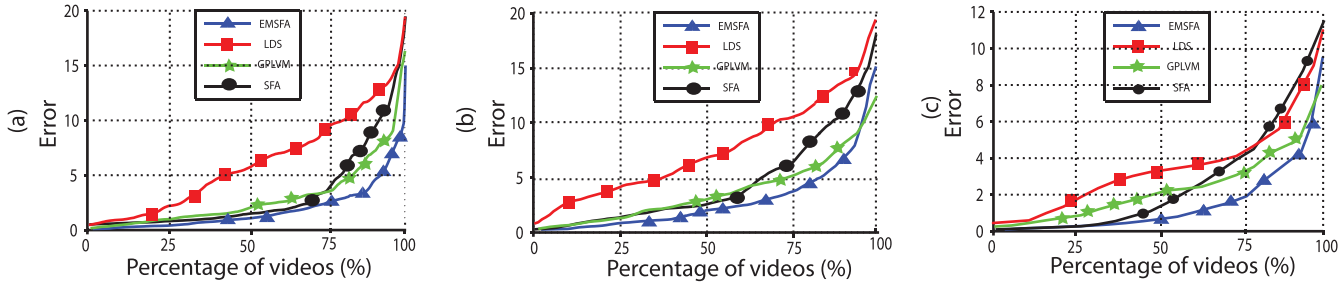

Fig. 8. Error versus the percentage of used videos in MMI database. The plots compare the performance of the tested methods in (a) mouth-related AUs, (b) eyes-related AUs, and (c) brows-related AUs).

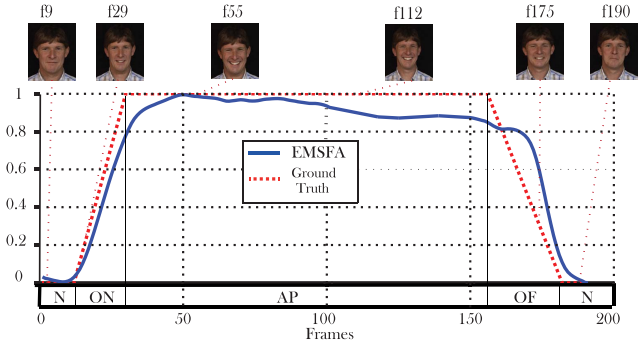

Fig. 9. Ground truth of a posed smile from the UNS database, along with the slowest varying feature from EMSFA.

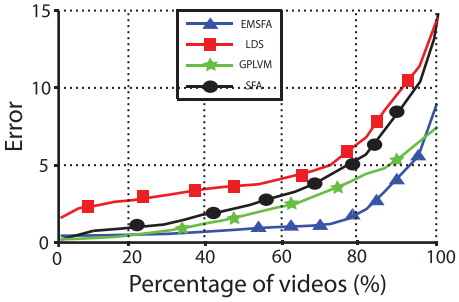

Fig. 10. Error versus the percentage of used videos in UNS posed smiles.

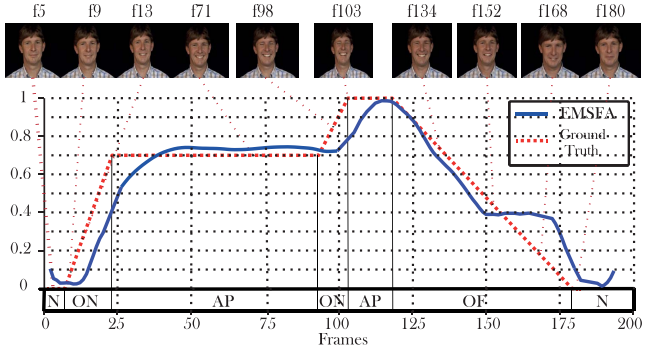

Fig. 11. Ground truth of a spontaneous smile from the UNS database, along with the slowest varying feature from EMSFA.

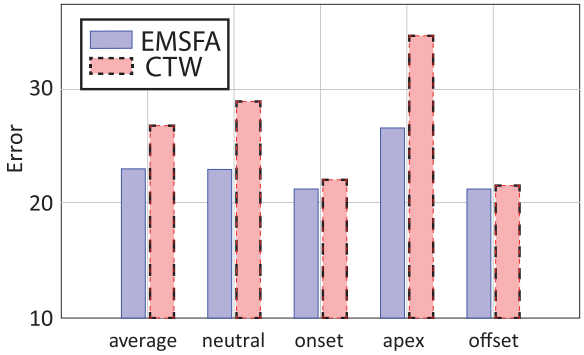

Fig. 12. Results obtained when comparing EM-SFA with DTW to CTW, for all temporal phases of AUs.

this experiment, we used segments, taken from a Greek TV political debate shown in 2011–2012, while the financial crisis was taking place (see Fig. 13 for relevant stills from the videos). Conflict episodes can be defined as situations where people interrupt each other or attempt to speak out of their turn; cases where people speak politely waiting for their turn are considered to be nonconflict episodes. In general, conflict has been associated to behavioral cues, such as nodes, frowns, and blinks [41]. Regarding the experimental setting, we first tracked 68 facial features of each of the persons participating in the recorded discussion (the tracking is the same as that described in Section VI-B). The frames of the utilized videos have been annotated in terms of conflict (one) and nonconflict (zero). In total, we have annotated 15 003 frames (5274 conflict frames and 9729 nonconflict frames).

In order to evaluate the performance of EM-SFA in detecting conflict and nonconflict episodes, we apply EM-SFA for two-sequences on the features extracted from both participants of each video. Subsequently, we extract the common, most slowly varying feature from both speakers, which we compare with the given ground-truth annotation. We expect this common, slowly varying feature to express the occurrence of conflict in the video. This is motivated by the fact that EM-SFA, as shown in Section VI-B, can detect the apex of the expression simply by examining the obtained latent space (slowest varying feature). Therefore, when applied on two subjects, EM-SFA can similarly detect the cooccurrence of the apex of both expressions, which in our scenario translates to a conflict episode.

In the example shown in Fig. 13, the left speaker (female) is speaking intensively throughout the video. The corresponding most slowly varying feature for the left speaker [Fig. 13(a)] is able to capture this. On the other hand, the right speaker (male) is trying to speak over the left speaker only for a particular set of frames, approximately from frame 60 to 200, while he remains a listener for the rest of the video. The corresponding most slowly varying feature for the speaker

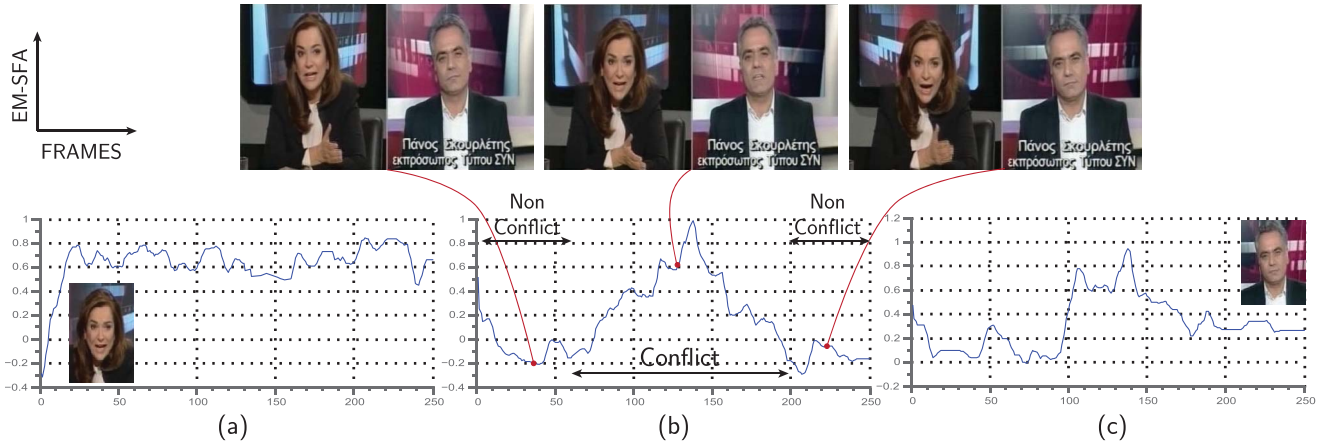

Fig. 13. Most slowly varying feature for (a) left speaker, (b) both speakers, and (c) right speaker. The slowest varying feature extracted via EM-SFA applied on both speakers can clearly detect the conflict occurrences in the video.

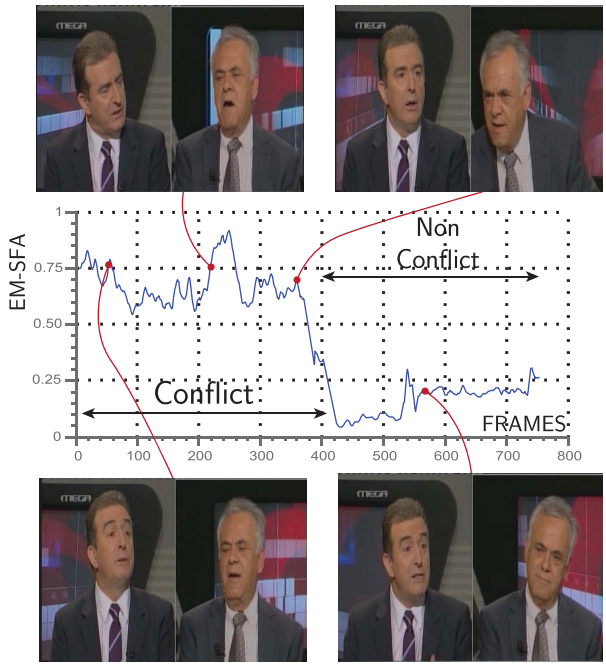

Fig. 14. Common most slowly varying feature applied on two speakers. The conflict episode is clearly detected by simply examining the latent space.

on the right is shown in Fig. 13(c). It is interesting to observe that in Fig. 13(b), where the common slowest varying feature is depicted, the conflict episode is clearly detected by EM-SFA, while also correctly identifying the nonconflict areas.

Another example of detecting conflict and nonconflict areas is shown in Fig. 14. In this video, the speakers are initially engaging in conflict trying to interrupt one another. This happens approximately up to frame 400. In what follows, the speaker on the right listens patiently while the speaker on the right speaks with less tension. In the same figure, we show the slowest varying feature extracted from both videos. It is clear from examining the common slowest varying feature obtained from EM-SFA that the conflict is clearly detected in the first part of the video, while the rest of the video contains only nonconflict episodes.

TABLE III
AVERAGE ACCURACY OF SFA, EMSFA, CCA FROM VIDEOS WHERE TWO PEOPLE ARE DEBATING

	Accuracy (%)		
Method	Conflict region	Non-conflict region	Overall
EMSFA	**69.95**	**56.42**	**63.19**
SFA	64.82	46.80	55.81
CCA	61.24	40.08	50.66

In order to provide quantitative results below, we provide two set of experiments. In the first one, we measure the reconstruction error and in the second one we measure the classification accuracy. The results clearly indicate that EMSFA outperforms compared methods in the most challenging scenarios, where human interactions are analyzed in an entirely uncontrolled scenario, without providing any supervision (i.e., label information) to the method.

1) Reconstruction Experiments: First, we examine the ability of the applied methods to reconstruct the input videos. To this end, we extracted three features for each of the applied methods and we measured to what extend can reconstruct the input signals. More precisely, we measured the mean reconstruction error (using the ℓ_2 norm) between the extracted original and the reconstructed sequence. Fig. 15(a) and (b) plots the mean reconstruction error versus the percentage of videos for each of the applied method on both conflict [Fig. 15(a)] and nonconflict [Fig. 15(b)] videos. As can be seen in both cases, the EM-SFA latent space not only captures the dynamics of the phenomenon but the extracted features are more expressive as well.

2) Detection Experiments: Finally, we evaluate the performance of the examined methods to identify conflict and nonconflict regions in 20 videos where two people are debating. The videos were manually annotated in terms of conflict and nonconflict regions. The normalized accuracy for each region (conflict, nonconflict) is measured as a percentage of the number of correctly identified frames for each region (e.g., correctly identified conflict frames), divided by the total number of frames corresponding to the region in the ground truth (e.g., the true number of conflict frames in video). This

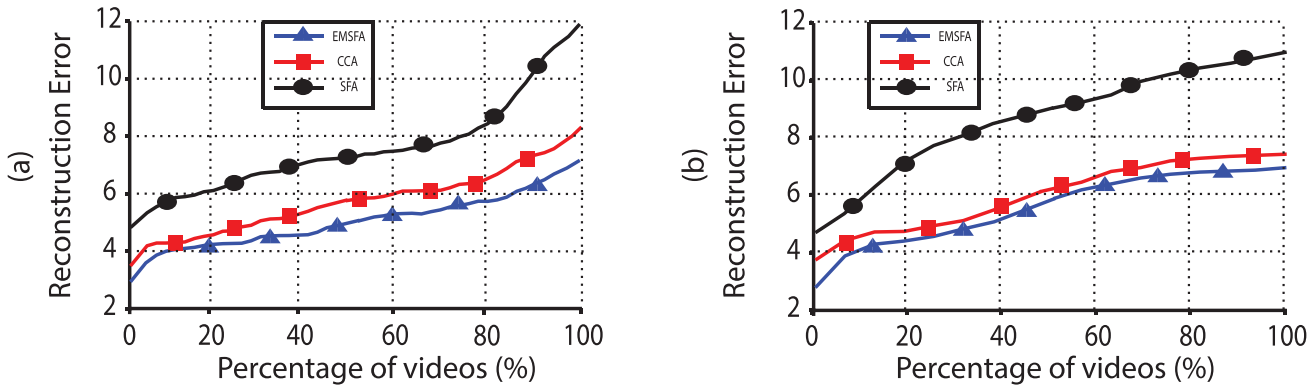

Fig. 15. Average reconstruction error of the input videos versus the percentage of videos on (a) conflict videos and (b) nonconflict videos.

can be formulated as

$$\frac{|\mathbf{n}^r \cap \mathbf{g}^r|}{|\mathbf{g}^r|} \tag{54}$$

where $\mathbf{n}^r$ is the quantized feature containing the frames which correspond to a specific region (r) and $\mathbf{g}^r$ is the ground truth for this region. To facilitate the comparison between the resulting latent features and the ground truth, we first normalized them to lie between 0 and 1 and then we quantized them by rounding. The results for each region and the overall results for the whole video are presented in Table III.

VII. Conclusion

In this paper, we presented a novel probabilistic approach to SFA. Specifically, we extended SFA to a fully probabilistic EM model (EM-SFA), while we augmented both deterministic and EM-SFA to handle multiple sequences. Furthermore, we combined EM-SFA algorithm with DTW techniques in order to align in time sequences of different lengths. We provide insights on the relationship between SFA and CCA. In particular we show that probabilistic CCA is a static model while the proposed one takes explicitly into account the time dependence. With an extended set of experiments we have shown the applicability of these novel models on both synthetic and real data. Our results show that the EM-SFA is a flexible component analysis model, in which an unsupervised manner can accurately capture the dynamics of sequences, such as facial expressions.

References

[1] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural Comput.*, vol. 14, no. 4, pp. 715–770, Apr. 2002.

[2] L. Wiskott, "Slow feature analysis: A theoretical analysis of optimal free responses," *Neural Comput.*, vol. 15, no. 9, pp. 2147–2177, Sep. 2003.

[3] P. Berkes and L. Wiskott, "Slow feature analysis yields a rich repertoire of complex cell properties," *J. Vis.*, vol. 5, no. 6, pp. 579–602, Jul. 2005.

[4] M. Franzius, N. Wilbert, and L. Wiskott, "Invariant object recognition and pose estimation with slow feature analysis," *Neural Comput.*, vol. 23, no. 9, pp. 2289–2323, Sep. 2011.

[5] S. Liwicki, S. Zafeiriou, and M. Pantic, "Incremental slow feature analysis with indefinite kernel for online temporal video segmentation," in *Proc. 11th Asian Conf. Comput. Vis.*, Daejeon, Korea, Nov. 2012, pp. 162–176.

[6] H. Q. Minh and L. Wiskott, "Slow feature analysis and decorrelation filtering for separating correlated sources," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 866–873.

[7] F. Nater, H. Grabner, and L. Van Gool, "Temporal relations in videos for unsupervised activity analysis," in *Proc. 22nd Brit. Mach. Vis. Conf.*, Dundee, Scotland, Nov. 2011, pp. 233–237.

[8] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 436–450, Mar. 2012.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 14. Whistler, BC, Canada, Dec. 2002, pp. 585–591.

[10] T. Blaschke, P. Berkes, and L. Wiskott, "What is the relation between slow feature analysis and independent component analysis?" *Neural Comput.*, vol. 18, no. 10, pp. 2495–2508, Oct. 2006.

[11] H. Sprekeler, "On the relation of slow feature analysis and Laplacian eigenmaps," *Neural Comput.*, vol. 23, no. 12, pp. 3287–3302, Dec. 2011.

[12] X. He and P. Niyogi, "Locality preserving projections," in *Proc. Adv. Neural Inf. Process. Syst.*, Whistler, BC, Canada, Dec. 2004, pp. 153–160.

[13] R. Turner and M. Sahani, "A maximum-likelihood interpretation for slow feature analysis," *Neural Comput.*, vol. 19, no. 4, pp. 1022–1038, Apr. 2007.

[14] F. Zhou, F. De la Torre, and J. F. Cohn, "Unsupervised discovery of facial events," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 2574–2581.

[15] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[16] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag, 2006.

[17] R. H. Shumway and D. S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Ser. Anal.*, vol. 3, no. 4, pp. 253–264, Jul. 1982.

[18] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, "ML estimation of a stochastic linear system with the EM algorithm and its application to speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 4, pp. 431–442, Oct. 1993.

[19] A. B. Chan and N. Vasconcelos, "Modeling, clustering, and segmenting video with mixtures of dynamic textures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 5, pp. 909–926, May 2008.

[20] G. Welch and G. Bishop, "An introduction to the Kalman filter," Dept. Comput. Sci., Univ. North Carolina Chapel Hill, Chapel Hill, NC, USA, Tech. Rep. TR 95-041, Jul. 2006.

[21] S. Haykin, Ed., *Kalman Filtering and Neural Networks*. New York, NY, USA: Wiley, 2001.

[22] S. Roweis and Z. Ghahramani, "A unifying review of linear Gaussian models," *Neural Comput.*, vol. 11, no. 2, pp. 305–345, Feb. 1999.

[23] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. Fluids Eng.*, vol. 82, no. 1, pp. 35–45, Mar. 1960.

[24] S. Sarkka, "Unscented Rauch–Tung–Striebel smoother," *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 845–849, Apr. 2008.

[25] F. Zhou and F. De la Torre, "Canonical time warping for alignment of human behavior," in *Proc. Adv. Neural Inf. Process. Syst.*, Whistler, BC, Canada, Dec. 2009, pp. 2286–2294.

[26] F. Zhou and F. De la Torre, "Generalized time warping for multi-modal alignment of human motion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1282–1289.

[27] A. Kneip and J. O. Ramsay, "Combining registration and fitting for functional models," *J. Amer. Statist. Assoc.*, vol. 103, no. 483, pp. 1155–1165, Sep. 2008.
[28] S. A. Kurtek, A. Srivastava, and W. Wu, "Signal estimation under random time-warpings and nonlinear signal alignment," in *Proc. Adv. Neural Inf. Process. Syst.*, Sierra Nevada, Spain, Dec. 2011, pp. 675–683.
[29] X. Liu and H.-G. Müller, "Functional convex averaging and synchronization for time-warped random curves," *J. Amer. Statist. Assoc.*, vol. 99, no. 467, pp. 687–699, Sep. 2004.
[30] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Dept. Statist., Univ. California, Berkeley, Berkeley, CA, USA, Tech. Rep. 688, Nov. 2005.
[31] L. J. P. van der Maaten, E. O. Postma, and H. J. van den Herik, "Dimensionality reduction: A comparative review," *J. Mach. Learn. Res.*, vol. 10, nos. 1–41, pp. 66–71, Oct. 2009.
[32] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, Jul. 2005, pp. 317–321.
[33] H. Dibeklioğlu, A. A. Salah, and T. Gevers, "Are you really smiling at me? Spontaneous versus posed enjoyment smiles," in *Proc. 12th Eur. Conf. Comput. Vis.*, Florence, Italy, Oct. 2012, pp. 525–538.
[34] T. F. Cootes and C. J. Taylor, "On representing edge structure for model matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Kauai, HI, USA, Jul. 2001, pp. I-1114–I-1119.
[35] G. Tzimiropoulos, J. Alabort-i-Medina, S. Zafeiriou, and M. Pantic, "Generic active appearance models revisited," in *Proc. 11th Asian Conf. Comput. Vis.*, Daejeon, Korea, Nov. 2012, pp. 650–663.
[36] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 283–298, Feb. 2008.
[37] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1940–1954, Nov. 2010.
[38] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert, "Recognition of 3D facial expression dynamics," *Image Vis. Comput.*, vol. 30, no. 10, pp. 762–773, Oct. 2012.
[39] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[40] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "Online handwriting recognition with support vector machines—A kernel approach," in *Proc. 8th Int. Workshop Frontiers Handwriting Recognit.*, Freiburg, Germany, Aug. 2002, pp. 49–54.
[41] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools," *Image Vis. Comput.*, vol. 31, no. 2, pp. 203–221, Feb. 2013.

Lazaros Zafeiriou (S'13) received the B.Sc. degree in automation from the Alexander Technical Educational Institute of Thessaloniki, Thessaloniki, Greece, in 2005, and the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, in 2010. He is currently pursuing the Ph.D. degree with the Department of Computing, Imperial College London, London, U.K., under the supervision of Prof. M. Pantic.

He has been a member of the IBUG Group with Imperial College London since 2012. He is currently involved in design of algorithms that consider both the spatial and the temporal (dynamic) nature of human behavior. His current research interests include statistical machine learning (with an emphasis on component analysis, time-series analysis, and automatic human behavior analysis).

Mihalis A. Nicolaou (S'10–M'15) received the B.Sc. (Hons.) (Ptychion) degree in informatics and telecommunications from the University of Athens, Athens, Greece, in 2008, and the M.Sc. (Hons.) degree in advanced computing and the Ph.D. (Hons.) degree from the Department of Computing, Imperial College London, London, U.K., in 2009 and 2014, respectively.

He is currently a Post-Doctoral Associate with the Department of Computing, Imperial College London. His current research interests include component analysis and time-series analysis, with a particular focus on the machine analysis of human behavior, in particular, with regards to continuous and dimensional emotion descriptions.

Dr. Nicolaou has received several awards and scholarships during his studies.

Stefanos Zafeiriou (M'09) is currently a Senior Lecturer (equivalent to Associate Professor) in Pattern Recognition/Statistical Machine Learning for Computer Vision with the Department of Computing, Imperial College London, London, U.K. He has co-authored over 40 journal papers in novel statistical machine learning methodologies applied to computer vision problems, such as 2-D/3-D face analysis, deformable object fitting and tracking, shape from shading, and human behavior analysis, published in the most prestigious journals in his field of research and many papers in top conferences, such as the Computer Vision and Pattern Recognition Conference (CVPR), the Conference on Computer Vision (ICCV), the European Conference on Computer Vision (ECCV), and the International Conference on Machine Learning.

Mr. Zafeiriou was a recipient of the Prestigious Junior Research Fellowships from Imperial College London to start his own independent research group in 2011. He has received various awards during his doctoral and post-doctoral studies. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and the *Image and Vision Computing Journal*. He has been a Guest Editor of over five journal special issues, and has co-organized over five workshops/special sessions in top venues, such as CVPR/Conference on Automatic Face and Gesture Recognition/ICCV/ECCV. His students are frequent recipients of very prestigious and highly competitive fellowships, such as the Google Fellowship, the Intel Fellowship, and the Qualcomm Fellowship.

Symeon Nikitidis received the B.Sc. degree in informatics from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2004, the M.Sc. degree in advanced computing from the University of Glasgow, Glasgow, U.K., in 2005, and the Ph.D. degree in informatics from the Aristotle University of Thessaloniki, in 2013.

He was a Research and Teaching Assistant with the Department of Informatics, Aristotle University of Thessaloniki, from 2006 to 2012. Since 2012, he has been a Research Associate with the Department of Computing, Imperial College London, London, U.K. His current research interests include statistical machine learning, digital signal and image processing, pattern recognition, and computer vision.

Maja Pantic (M'98–SM'06–F'12) is currently a Professor of Affective and Behavioral Computing with the Department of Computing, Imperial College London, London, U.K., and the Department of Computer Science, University of Twente, Enschede, The Netherlands.

Prof. Pantic received various awards for her work on automatic analysis of human behavior, including the European Research Council Starting Grant Fellowship in 2008 and the Roger Needham Award in 2011. She currently serves as the Editor-in-Chief of the *Image and Vision Computing Journal*, and an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON CYBERNETICS.