Learning Slow Features for Behaviour Analysis

Lazaros Zafeiriou[1], Mihalis A. Nicolaou[1], Stefanos Zafeiriou[1], Symeon Nikitidis[1] and Maja Pantic[1,2]
[1]Department of Computing, Imperial College London, UK
[2]EEMCS, University of Twente, NL
{l.zafeiriou12, mihalis, s.zafeiriou, s.nikitidis, m.pantic}@imperial.ac.uk

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 multi-dimensional sequences that by minimizing the variance of the first order time derivative approximation of the input signal 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.

1. Introduction

Slow Feature Analysis (SFA) was first proposed in [25] as an unsupervised methodology for finding slowly varying (invariant) features from rapidly temporal varying signals. The exploited slowness learning principle in [25] 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 [25] that assumes a discrete time input signal [1] and the low dimensional output signal is obtained as a linear transformation of a non-linear expansion of the input. The proposed in [25] optimization problem 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 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 [2], 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 [8, 13, 14, 15, 26]. Recently, SFA and its discriminant extensions have been successfully applied for human action recognition in [26], while hierarchical segmentation of video sequences using SFA was investigated in [15]. In [8] SFA was applied for object and object-pose recognition on a homogeneous background, while in [14] SFA for vector-valued functions was studied for blind source separation. Finally, an incremental SFA algorithm for change detection was proposed in [13].

Links between SFA and other other component analysis techniques, such as Independent Component Analysis (ICA) and Laplacian Eigenmaps (LE) [1] were extensively studied in [4, 20]. In [4], the equivalence between linear SFA and the second-order ICA algorithm, in the case of one time delay, is demonstrated. In [20], the relation between

[1]Continuous time SFA has been proposed in [24] but since in this paper we assume discrete time signals, such works are out of our scope.

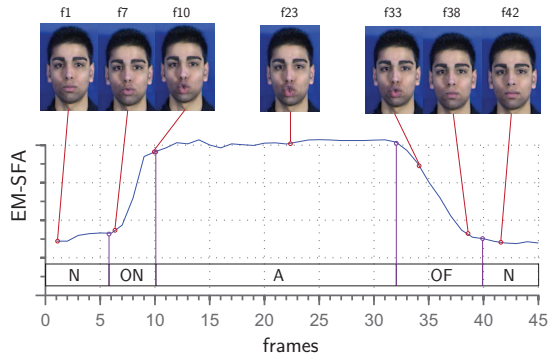

Figure 1: The latent space obtained by EM-SFA, accurately capturing the transition between temporal phases of action units. The ground truth is shown as N: Neutral, ON: Onset, A: Apex, OF: Offset.

LE and SFA was studied and exhibited that SFA is a special case of kernel Locality Preserving Projections (LPP) [9] acquired by defining the data neighborhood structure using their temporal variations. In [21], 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 that the noise variance tends to zero. The probabilistic generative model comprises a linear model for the generation of observations and imposes a Gaussian linear dynamical system with diagonal covariances over the latent space.

In this paper, we study the application of SFA for unsupervised facial behaviour analysis. Our motivation is based on the aforementioned theory on the close relationship between human perception and SFA. Our application is further motivated by Fig. 1. In more detail, in Fig. 1, we can see 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 recorded: Neutral, when the face is relaxed, Onset, when the action initiates, Apex, when the muscles reach the peak intensity and Offset when the muscles begin to relax. The action finally ends with Neutral. It can be clearly observed in the figure, 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.

Summarising the contributions of our paper, we propose the following theoretical novelties:

- 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 ([21]), our approach allows for full probabilistic modelling of the latent distributions instead of mapping the variances to zero, as in ML.
- 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.

The novelties of our paper in terms of application can be summarized as follows:

- We apply the proposed EM-SFA to facial behaviour dynamics analysis and in particular for facial Action Units (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 which detects the temporal phases of AUs (other unsupervised approaches such as [29] focus on detecting global structures (i.e. AUs or expressions) rather than their temporal phases).
- We combine the common latent space derived by EM-SFA with Dynamic Time Warping techniques [18] for the temporal alignment of dynamic facial behaviour. We claim that by using the slowest varying features for sequence alignment is well motivated by the principle of slowness as described above (i.e., slowly varying features correspond to meaningful changes rather than rapidly varying ones, which most likely correspond to noise [25]).

The rest of the paper is organised as follows. In Sec.2, we describe the deterministic SFA model, while in Sec. 3, we introduce the probabilistic interpretation of SFA. Our proposed EM-SFA is presented in Sec. 4, both for one (Sec. 4.1) and multiple sequences (Sec. 4.2), while the latter method is incremented with warpings in Sec. 5.3. Finally, we evaluate the proposed models in Sec. 5, by a set of experiments with both synthetic (Sec. 5.1) and real (5.2, 5.3) data.

2. 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 $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N] \in \Re^{M \times N}$ ($N << M$), 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:

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

where $tr[.]$ is the matrix trace operator, $\mathbf{1}$ is a $T \times 1$ vector with all its elements equal to $\frac{1}{T}$, $\mathbf{I}$ 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:

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

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:

$$\min_{\mathbf{V}} \quad tr[\mathbf{V}^T\mathbf{A}\mathbf{V}], \text{ s.t. } \mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{I}. \quad (3)$$

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 $\dot{\mathbf{X}}$

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

The solution of (3) can be found from the Generalized Eigenvalue Problem (GEP) [25]:

$$\mathbf{A}\mathbf{V} = \mathbf{B}\mathbf{V}\mathbf{L} \quad (5)$$

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 $\mathbf{L}$.

3. A 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:

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

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 the linear Gaussian dynamical system priors over the latent space $\mathbf{Y}$ are:

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

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 $\mathbf{\Sigma}_1 = [\delta_{i,j}\sigma^2_{n,1}]$ the prior over the latent space can be evaluated as:

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

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

In [21], it was shown that the ML solution of the above model in the deterministic case (i.e., $\sigma_x^2 \to 0$) with $T \to \infty$, where the conditional probability in (8) is simplified to $P(\mathbf{Y}|\theta_y) \approx \frac{1}{Z}\exp\Big[-\text{tr}\Big[\mathbf{Y}\mathbf{Y}^T\mathbf{\Lambda}^{(2)} + \dot{\mathbf{Y}}\dot{\mathbf{Y}}^T\mathbf{\Lambda}^{(1)}\Big]\Big]$, is evaluated as:

$$\begin{aligned} \mathbf{V} &= \arg\max_{\mathbf{V}, \sigma_x^2 \to 0} \log P(\mathbf{X}|\theta) \\ &= \arg\max \log \int_{\mathbf{Y}} P(\mathbf{X}|\mathbf{Y}, \theta_x)P(\mathbf{Y}|\theta_y)d\mathbf{Y} \end{aligned} \quad (9)$$

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 corresponds directly to inducing an order 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 $\sigma_n^2 = 1 - \lambda_n^2$ [21], which constraints the prior covariance of the latent variables to be one.

3.1. Extension to two sequences

The probabilistic interpretation of SFA discussed above can be extended to more than one sequences. Under this scenario, the method essentially uncovers the *common* slowly varying features extracted from the sequences at-hand. We define the following generative model,

$$\begin{aligned} \mathbf{x}_t^k &= \hat{\mathbf{V}}_k^{-1}\mathbf{y}_t + \boldsymbol{\epsilon}_t^k, \quad \boldsymbol{\epsilon}_t^k \sim \mathcal{N}(0, \sigma_k^2 I), k = 1, 2 \\ \mathbf{X}_t^{tot} &= \begin{bmatrix} \mathbf{x}_t^1 \\ \mathbf{x}_t^2 \end{bmatrix} = \mathbf{V}^{-1}\mathbf{y}_t + \begin{bmatrix} \boldsymbol{\epsilon}_t^1 \\ \boldsymbol{\epsilon}_t^2 \end{bmatrix} \end{aligned} \quad (10)$$

By computing the marginal $\log P(\mathbf{X}_1, \mathbf{X}_2|\theta)$ (i.e. marginalising out the latent space) and taking the lim-

its $\lim\{\sigma_{x,1}^2, \sigma_{x,2}^2\} \to 0, T \to \infty$, we obtain

$$\begin{aligned}
&\log P(\mathbf{X}_1, \mathbf{X}_2|\theta) \qquad (11)\\
&= \log \int_{\mathbf{Y}} \prod_{t=1}^{T} P\left(\mathbf{X}_t^{tot}|\mathbf{y}_t, \theta_{x_1}, \theta_{x_2}\right) P(\mathbf{Y}|\theta_y) d\mathbf{Y}\\
&= \log \int_{\lim\{\sigma_{x,1}^2, \sigma_{x,2}^2\}\to 0} \delta\left(\mathbf{X}_t^{tot} - \mathbf{V}^{-1}\mathbf{y}_t\right) P(\mathbf{Y}|\theta_y) d\mathbf{Y}\\
&= c + T\left(log|\mathbf{V}_1| + log|\mathbf{V}_2|\right)\\
&- \frac{T}{2}\text{tr}\left[\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{B} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{\Lambda}^{(2)} + \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{A} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{\Lambda}^{(1)}\right]
\end{aligned}$$

where

$$\mathbf{B} = \begin{bmatrix} \mathbf{X}_1\mathbf{X}_1^T & \mathbf{X}_1\mathbf{X}_2^T \\ \mathbf{X}_2\mathbf{X}_1^T & \mathbf{X}_2\mathbf{X}_2^T \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} \dot{\mathbf{X}}_1\dot{\mathbf{X}}_1^T & \dot{\mathbf{X}}_1\dot{\mathbf{X}}_2^T \\ \dot{\mathbf{X}}_2\dot{\mathbf{X}}_1^T & \dot{\mathbf{X}}_2\dot{\mathbf{X}}_2^T \end{bmatrix} \qquad (12)$$

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

$$\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{B} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} \mathbf{\Lambda}^{(2)} + \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{A} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} \mathbf{\Lambda}^{(1)} = \mathbf{I} \qquad (13)$$

since $\mathbf{\Lambda}^{(2)}$ and $\mathbf{\Lambda}^{(1)}$ 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.

$$\begin{aligned}
\min_{\mathbf{V}} \quad & tr\left[\begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{A} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}\right]\\
\text{s.t.} \quad & \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}^T \mathbf{B} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} = \mathbf{I}. \qquad (14)
\end{aligned}$$

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

$$\mathbf{A} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} = \mathbf{B} \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix} \begin{bmatrix}\mathbf{L}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{L}_2\end{bmatrix}. \qquad (15)$$

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

4. An EM approach for probabilistic SFA

The ML approach for probabilistic SFA bears many disadvantages. Firstly, 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 modelling full distributions along with both observation and latent variance (EM-SFA, Sec. 4.1). Furthermore, we extend EM-SFA to handle two distinct sequences (Sec. 4.2), while the extension for handling any number of multiple sequences is straight-forward.

4.1. EM-SFA for Single Sequence

In this Section we propose a complete probabilistic SFA algorithm using EM, while following the constraints discussed in Sec. 3 ($0 < \lambda_n < 1, \forall\ n$ 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 $\mathbf{\Lambda}$).

In order to perform EM we need to define the complete log likelihood of the model as:

$$\begin{aligned}
\log P(\mathbf{X}, \mathbf{Y}|\theta) = &\sum_{t=1}^{T} \log P(\mathbf{x}_t|\mathbf{y}_t, \theta_x) + \log P(\mathbf{y}_1|\mathbf{\Sigma}_1)\\
&+ \sum_{t=2}^{T} \log P(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{\Lambda}) \qquad (16)
\end{aligned}$$

In the Expectation step we need to find the sufficient statistics given the observed data and the model parameters θ. The sufficient statistics $\mathbb{E}[\mathbf{y}_t|\mathbf{X}]$, $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}]$ and $\mathbb{E}[\mathbf{y}_t\mathbf{y}_{t-1}^T|\mathbf{X}]$ can be computed using forward and backward recursions,known as the Kalman or Rauch-Tung-Streibel (RTS) smoother [17]. In the Maximization step given the sufficient statistics obtained, we need to find the model parameters θ by optimising:

$$\theta_o = \arg\max_{\theta} \mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}[\log P(\mathbf{X}, \mathbf{Y}|\theta)] \qquad (17)$$

which can be split to three parts $\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[\sum_{t=1}^{T} \log P(\mathbf{x}_t|\mathbf{y}_t, \theta_x)]$, $\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[P(\mathbf{y}_1|\mathbf{\Sigma}_1)]$ and $\mathbb{E}_{P(\mathbf{Y}|\mathbf{X})}[\log \sum_{t=2}^{T} p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{\Lambda})]$. By expanding the first part, which provides updates for $\mathbf{V}^{new}$ and $(\sigma_x^{new})^2$, we obtain

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

Subsequently, by setting the derivatives for $\mathbf{V}^{new}$ and

[2] The EM algorithm presented shares some similarities with the EM for LDS c.f., Chap. 13 of [3], [19], [7], [5]

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

$(\sigma_x^{new})^2$ equal to zero we obtain the updates

$$\mathbf{V}^{new} = \left(\sum_{t=1}^{T}\mathbf{x}_t\,\mathbb{E}[\mathbf{y}_t^T|\mathbf{Y}]\right)\left(\sum_{t=1}^{T}\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{Y}]\right)^{-1} \quad (18)$$

$$(\sigma_x^{new})^2 = \frac{1}{MT}\sum_{t=1}^{T}\Big(tr(\mathbf{x}_t\mathbf{x}_t^T) - 2\mathbf{x}_t^T\mathbf{V}^{new}\,\mathbb{E}[\mathbf{y}_t|\mathbf{Y}] + tr(\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{Y}](\mathbf{V}^{new})^T\mathbf{V}^{new})\Big). \quad (19)$$

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:

$$\mathbf{\Sigma}_1^o = \arg\max_{\mathbf{\Sigma}_1}\mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}[P(\mathbf{y}_1|\mathbf{\Sigma}_1)] \quad (20)$$

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:

$$\begin{aligned}\mathbf{\Lambda} &= \arg\max_{\mathbf{\Lambda}}\mathop{\mathbb{E}}_{P(\mathbf{Y}|\mathbf{X})}[\log\sum_{t=2}^{T}p(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{\Lambda})]\\ &= \arg\max_{\mathbf{\Lambda}} -\frac{1}{2}\sum_{t=2}^{T}\Bigg[\sum_{n=1}^{N}\ln(1-\lambda_n^2)\\ &+ \frac{1}{1-\lambda_n^2}\sum_{n=1}^{N}((\mathbb{E}[y_{n,t}^2|\mathbf{X}] - 2\lambda_d\,\mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]\\ &+ \lambda_n^2\,\mathbb{E}[y_{n,t-1}^2|\mathbf{X}])\Bigg] + const \end{aligned} \quad (21)$$

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

$$\sum_{t=2}^{T}\Big((\lambda_n^{new})^3 - \mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}](\lambda_n^{new})^2 + (\mathbb{E}[y_{n,t}^2|\mathbf{X}] + \mathbb{E}[y_{n,t-1}^2|\mathbf{X}] - 1)\lambda_n^{new} - \mathbb{E}[y_{n,t}y_{n,t-1}|\mathbf{X}]\Big) = 0 \quad (22)$$

The above equation yields three solutions for λ_n^{new}. We retain the solution which satisfies $0 < \lambda_n^{new} < 1$. Due to space limitations the detailed solution of the cubic equation is provided in the supplementary materials.

4.2. 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. To do so let us assume the following generative model for the samples of the following time varying input sequences $\mathbf{X}_1 = \left[\mathbf{x}_t^1,\, t\in[1,T]\right] \in \Re^{M_1\times T}$ and $\mathbf{X}_2 = \left[\mathbf{x}_t^2,\, t\in[1,T]\right] \in \Re^{M_2\times T}$:

$$\mathbf{x}_t^k = \mathbf{V}_k\mathbf{y}_t + \mathbf{e}_t^k, \mathbf{e}_t^k \sim \mathcal{N}(0,\sigma_{x,k}^2\mathbf{I}), k = 1,2 \quad (23)$$

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

$$\begin{aligned}&\log P(\mathbf{X}_1,\mathbf{X}_2,\mathbf{Y}|\theta) =\\ &\log P(\mathbf{y}_1|0,\mathbf{\Sigma_1}) + \sum_{t=2}^{T}\log P(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{\Lambda}) +\\ &\sum_{t=1}^{T}\log P(\mathbf{x}_t^1|\mathbf{y}_t,\mathbf{V}_1,\sigma_{x,1}^2) + \sum_{t=1}^{T}\log P(\mathbf{x}_t^2|\mathbf{y}_t,\mathbf{V}_2,\sigma_{x,2}^2)\end{aligned} \quad (24)$$

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 $\mathbb{E}[\mathbf{y}_t|\mathbf{X}_1,\mathbf{X}_2]$, $\mathbb{E}[\mathbf{y}_t|\mathbf{X}_1,\mathbf{X}_2]$, $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}_1,\mathbf{X}_2]$ 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 Sec. 4.1. Applying the maximization step on the joint log likelihood (24) we obtain the updates for $\mathbf{V}_1,\mathbf{V}_2$ and $\sigma_{x,1}^2,\sigma_{x,2}^2$ as:

$$\mathbf{V}_k^{new} = \left(\sum_{t=1}^{T}\mathbf{x}_t^k\,\mathbb{E}[\mathbf{y}_t^T|\mathbf{X}^{tot}]\right)\left(\sum_{t=1}^{T}\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}^{tot}]\right)^{-1} \quad (25)$$

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

Regarding $\mathbf{\Lambda}$ and $\mathbf{\Sigma_1}$ the updates are given by (21) and (20), applied using the derived $\mathbb{E}[\mathbf{y}_t|\mathbf{X}_1,\mathbf{X}_2]$, $\mathbb{E}[\mathbf{y}_t\mathbf{y}_t^T|\mathbf{X}_1,\mathbf{X}_2]$ 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.

4.3. 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. This entails that we disregard high frequency features which are likely to be noisy. We note that recently, time series alignment was performed on a space recovered by the application of Canonical Correlation Analysis (CCA) ([27]). A simple, commonly used [27] and optimal method for finding the warpings is Dynamic Time Warping (DTW)[4], which we employ in our case. Given two sequences

[4]Other methods that can be used include e.g., [28], while for related work from functional data analysis, please c.f., [10, 11, 12].

$\mathbf{X}_1 \in \Re^{M_1 \times T_1}$ and $\mathbf{X}_2 \in \Re^{M_2 \times T_2}$ of different lengths $T_1 \neq T_2$, DTW aims is to find the warpings $\mathbf{\Delta}_1 \in \Re^{T_1 \times T}$ and $\mathbf{\Delta}_2 \in \Re^{T_2 \times T}$ such that the observation sequences will have common length of size T. The augmentation of EM-SFA with DTW is presented in Algorithm 1.

Algorithm 1: EMSFA with DTW

Data: $\mathbf{X}_1, \mathbf{X}_2, iter, q$
Result: $\mathbf{\Delta}_1, \mathbf{\Delta}_2, \mathbb{E}[\mathbf{Y}|\mathbf{X}_1^{\Delta}, \mathbf{X}_2^{\Delta}]$

while *not converged* **do**
  if $iter = 1$ **then**
    $(\mathbf{\Delta}_1, \mathbf{\Delta}_2) \leftarrow \text{DTW}(\mathbf{X}_1, \mathbf{X}_2)$
  else
    $(\mathbf{\Delta}_1, \mathbf{\Delta}_2) \leftarrow \text{DTW}(\mathbb{E}[\mathbf{Y}|\mathbf{X}_1], \mathbb{E}[\mathbf{Y}|\mathbf{X}_2])$
  $\mathbf{X}_1^{\Delta} \leftarrow \mathbf{X}_1\mathbf{\Delta}_1, \mathbf{X}_2^{\Delta} \leftarrow \mathbf{X}_2\mathbf{\Delta}_2$
  while *not converged* **do**
    Update θ (Eq. (25,26, 21,20)
    Update $\mathbf{\Sigma}$ acc. to $\sigma_n^2 = 1 - \lambda_n^2$
    $\mathbb{E}[\mathbf{Y}|\mathbf{X}_1^{\Delta}, \mathbf{X}_2^{\Delta}] \leftarrow \text{RTS}(\mathbf{X}_{tot}^{\Delta}, \mathbf{\Lambda}, \mathbf{\Sigma}, \mathbf{V}, \sigma_{x,tot}^2, \mathbf{\Sigma_1})$
  $\sigma_{x,1}^2, \sigma_{x,2}^2 \leftarrow \sigma_x^{tot}\mathbf{I}_M = \begin{pmatrix} \sigma_{x,1}^2\mathbf{I}_{M_1} & \mathbf{0} \\ \mathbf{0} & \sigma_{x,2}^2\mathbf{I}_{M_2} \end{pmatrix}$
  $\mathbf{V}_1, \mathbf{V}_2 \leftarrow \mathbf{V} = \begin{pmatrix} \mathbf{V}_1 \\ \mathbf{V}_2 \end{pmatrix}$
  $\mathbb{E}[\mathbf{Y}|\mathbf{X}_k] \leftarrow \text{RTS}(\mathbf{X}_k, \mathbf{\Lambda}, \mathbf{\Sigma}, \mathbf{V}_k, \sigma_{x,k}^2, \mathbf{\Sigma_k})$, k=1,2

5. Experimental Results

For demonstrating the effectiveness of our proposed methods, experiments were conducted both on synthetic (Sec. 5.1) and real (Sec. 5.2, 5.3) data.

5.1. Synthetic Data

In this section we demonstrate the experimental results of our proposed algorithms on synthetic data. We use the Dimensionality Reduction Toolbox to generate randomly scaled synthetic examples of 1000 data points each. In Fig. 2, we can see a comparison between the resulting latent space of EM-SFA and deterministic SFA, when applying the algorithms on the two sequences presented in Fig. 2(a,b). It is easy to observe that the latent spaces derived by both EM-SFA (d) and deterministic SFA Fig. 2(c) are essentially equivalent. Due to lack of space, further synthetic examples are shown in the supplementary materials.

5.2. Real Data 1: Unsupervised AU Temporal Phase Segmentation

Regarding real data, we employ the publicly available MMI database [16], which consists of more than 400 videos annotated in terms of facial Action Units (AUs) and their temporal phases, i.e. Neutral, Onset, Apex and Offset. Throughout this section, we use trackings of facial expressions for each subject. The employed tracker is a person-independent implementation of Active Appearance Models (AAMs), using the normalised gradient features proposed in [6]. The implementation used, presented in [22], firstly detects the faces of the subjects by applying the Viola-Jones face detector [23] and subsequently tracks 68 2-dimensional facial points.

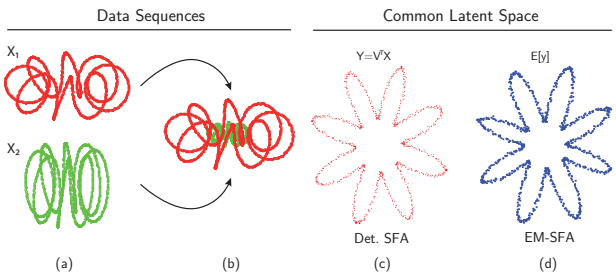

Figure 2: Application of deterministic SFA and EM-SFA on two synthetic data sequences $\mathbf{X}_1, \mathbf{X}_2$ (a,b). The resulting common latent space is shown in (c),(d).

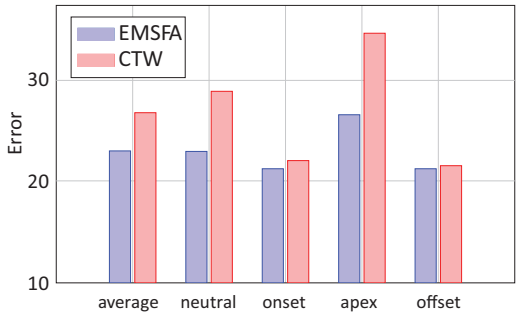

Figure 4: Results obtained when comparing EM-SFA with DTW to CTW, for all temporal phases of AUs.

For the first experiment, our goal is to measure how effectively EM-SFA can detect the temporal phases of AUs, in comparison to deterministic SFA and traditional Linear Dynamic Systems (LDS). In this experiment, for each 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 to the annotated ground truth.

To facilitate the comparison with the ground truth, we map the recovered latent space to the temporal phases of AUs. This is done by finding the points in which the first order derivative of the obtained latent space (most slowly varying feature) crosses zero and switches to positive or negative. In more detail, when the derivative switches to positive and then back to zero we obtain points x_1 and x_2 and when the derivative switches to a negative value and back we obtain the points x_3 and x_4. These points are clearly illustrated in Fig. 3(a) in green bullets. Subse-

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

Table 1: Performance of SFA, EMSFA and LDS in terms of extracting the ground truth from Actions Units related to mouth, eyes and brows, evaluated on all AU temporal phases and the expression peak.

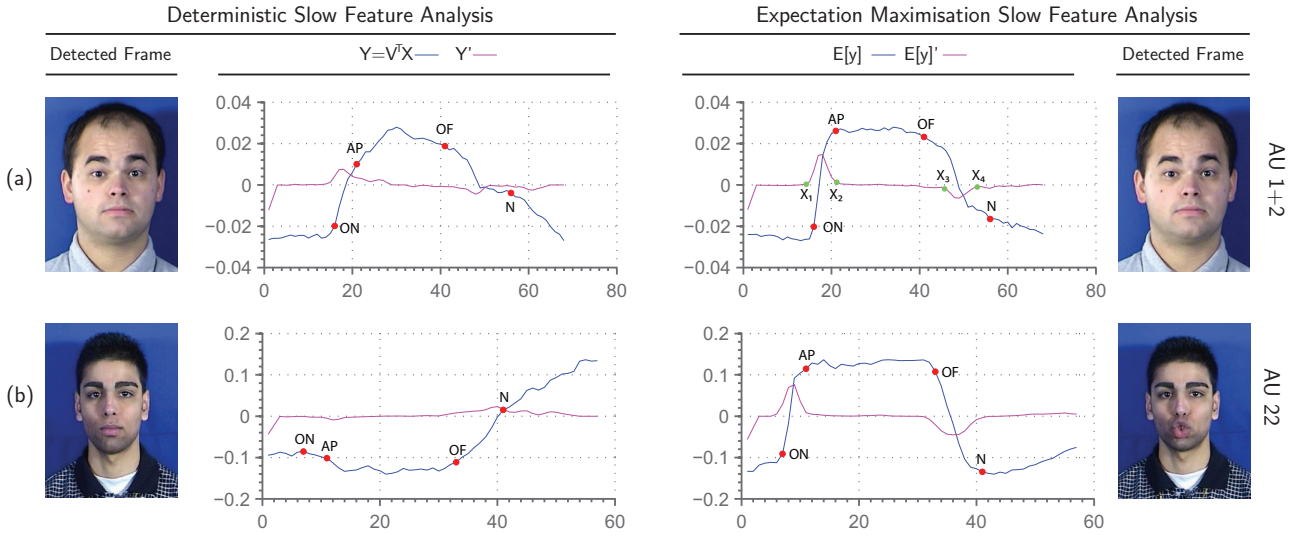

Figure 3: Comparing the derived latent space (i.e. slowest varying feature) for SFA and EM-SFA, obtained when applying the algorithms on two different videos. The space ($E[y]$) along with the gradient ($E[y]'$) is shown. The true points were the AU temporal phase changes are shown with red markers. (ON - change from neutral to onset, AP - change from onset to apex, OF - change from apex to offset, N - change from offset to neutral).

quently, we corresponded the points from 0 to x_1 to Neutral, from x_1 to x_2 to Onset, from x_2 to x_3 to Apex and from x_3 to x_4 to Offset, while the rest of the frames are considered as Neutral. The overall results for the applied methods are summarized in Table 1. The presented results show that EM-SFA overperforms deterministic SFA and LDS on the unsupervised detection of the temporal phases of AUs, for all temporal phases and for all relevant regions of the face. Furthermore, in Table 1 ta 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 should 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 small apex (most of the times just 1 frame). Therefore, it is very challenging to capture it. Nevertheless, EM-SFA appears to capture the blink Apex much better than compared methods. In Fig. 3, we can visually evaluate the performance of EM-SFA and deterministic SFA on the given problem. Two examples are shown, in (a), both methods manage to capture the apex of the expression as well as segment the temporal phases according to the ground truth, with EM-SFA performing better. In example (b), deterministic SFA fails to capture the dynamics of the AU, while EM-SFA accurately captures the transition.

5.3. Real Data 2: Temporal Alignment

In this section, we present results on aligning pairs of videos from the MMI database, where the same AU is activated. The goal of this experiment is to evaluate the derived space of EM-SFA to the obtained space when using 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 (CTW [27]). Of major importance to this claim is the modelling of dynamics in EM-SFA, on the contrary to traditional CCA, which is

does not account for temporal dependencies. Results are presented in Fig. 4. The error we used is the percentage of misaligned frames for each pair of videos, normalised per frame (i.e. divided by the aligned video length). We present results on average (for the entire video) and results regarding the apex, as well as neutral, onset and offset. It is clear from our results that the derived space of SFA is better suited for the alignment of temporal sequences than the space obtained by applying CCA.

6. Conclusions

In this paper, we presented a novel, probabilistic approach to Slow Feature Analysis. Specifically, we extended SFA to a fully probabilistic EM model (EM-SFA), while we augmented both deterministic and EM-SFA to handle multiple sequences. With a 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, which in an unsupervised manner can accurately capture the dynamics of sequences, such as facial expressions.

7. Acknowledegments

Mihalis A. Nicolaou and Lazaros Zafeiriou were funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 288235 (FROG). Maja Pantic by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB). The work of Stefanos Zafeiriou and Symeon Nikitidis was partially funded by the EPSRC project EP/J017787/1 (4D-FAB).

References

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. *NIPS*, 2001.

[2] P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex cell properties. *Vision*, 5(6), 2005.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] T. Blaschke, P. Berkes, and L. Wiskott. What is the relation between slow feature analysis and independent component analysis? *Neural computation*, 18(10):2495–2508, 2006.

[5] A. B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE TPAMI*, 30(5):909–926, 2008.

[6] T. F. Cootes and C. J. Taylor. On representing edge structure for model matching. In *CVPR*. IEEE, 2001.

[7] 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. *Speech and Audio Processing, IEEE Transactions on*, 1(4):431–442, 1993.

[8] M. Franzius, N. Wilbert, and L. Wiskott. Invariant object recognition and pose estimation with slow feature analysis. *Neural computation*, 23(9):2289–2323, 2011.

[9] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, 2003.

[10] A. Kneip and J. O. Ramsay. Combining registration and fitting for functional models. *Journal of the American Statistical Association*, 103(483):1155–1165, 2008.

[11] S. A. Kurtek, A. Srivastava, and W. Wu. Signal estimation under random time-warpings and nonlinear signal alignment. In *Advances in Neural Information Processing Systems*, pages 675–683, 2011.

[12] X. Liu and H.-G. Müller. Functional convex averaging and synchronization for time-warped random curves. *JASA*, 99(467):687–699, 2004.

[13] S. Liwicki, S. Zafeiriou, and M. Pantic. Incremental slow feature analysis with indefinite kernel for online temporal video segmentation. In *ACCV*, 2012.

[14] H. Q. Minh and L. Wiskott. Slow feature analysis and decorrelation filtering for separating correlated sources. In *ICCV*, pages 866–873. IEEE, 2011.

[15] F. Nater, H. Grabner, and L. Van Gool. Temporal relations in videos for unsupervised activity analysis. In *BMVC*, 2011.

[16] M. Pantic et al. Web-based database for facial expression analysis. In *EEE ICME*, pages 317–321, 2005.

[17] S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neur. comput.*, 11(2):305–345, 1999.

[18] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech and Sig. Processing*, 26(1):43–49, 1978.

[19] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the em algorithm. *Journal of time series analysis*, 3(4):253–264, 1982.

[20] H. Sprekeler. On the relation of slow feature analysis and laplacian eigenmaps. *Neural computation*, 23(12):3287–3302, 2011.

[21] R. Turner and M. Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007.

[22] G. Tzimiropoulos, J. Alabort-i Medina, S. Zafeiriou, and M. Pantic. Generic active appearance models revisited. In *Computer Vision–ACCV 2012*, pages 650–663. Springer, 2013.

[23] P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[24] L. Wiskott. Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, 15(9):2147–2177, 2003.

[25] L. Wiskott and T. J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

[26] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *TPAMI*, 34(3):436–450, 2012.

[27] F. Zhou and F. De la Torre. Canonical time warping for alignment of human behavior. In *NIPS*, December 2009.

[28] F. Zhou and F. De la Torre. Generalized time warping for multi-modal alignment of human motion. In *IEEE CVPR 2012*, pages 1282–1289. IEEE, 2012.

[29] F. Zhou, F. De la Torre, and J. F. Cohn. Unsupervised discovery of facial events. In *IEEE CVPR*, 2010.