Variational Inference and Learning of Piecewise-linear Dynamical Systems

Xavier Alameda-Pineda, Vincent Drouard, Radu Horaud

Variational Inference and Learning of Piecewise-linear Dynamical Systems

Xavier Alameda-Pineda, *Senior Member, IEEE*, Vincent Drouard and Radu Horaud

***Abstract*—Modeling the temporal behavior of data is of primordial importance in many scientific and engineering fields. Baseline methods assume that both the dynamic and observation equations follow linear-Gaussian models. However, there are many real-world processes that cannot be characterized by a single linear behavior. Alternatively, it is possible to consider a piecewise-linear model which, combined with a switching mechanism, is well suited when several modes of behavior are needed. Nevertheless, switching dynamical systems are intractable because their computational complexity increases exponentially with time. In this paper, we propose a variational approximation of piecewise linear dynamical systems. We provide full details of the derivation of two variational expectation-maximization algorithms, a filter and a smoother. We show that the model parameters can be split into two sets, static and dynamic parameters, and that the former parameters can be estimated off-line together with the number of linear modes, or the number of states of the switching variable. We apply the proposed method to the head-pose tracking, and we thoroughly compare our algorithms with several state of the art trackers.**

***Index Terms*—Switching state space model, linear dynamical system, inverse regression, Bayesian inference, variational approximation, expectation-maximization, Kalman filter.**

I. Introduction

Modeling the temporal behavior of data is of primordial importance in many scientific fields, such as signal processing [1], computer vision [2], [3], robotics [4], autonomous navigation [5], to cite just a few. The baseline model for addressing this problem is the linear dynamical system (LDS). The basic idea of LDS is to assume that both the dynamic and the observation equations follow linear-Gaussian models. This yields tractable learning and inference procedures, namely the Kalman filter (KF) [6] and the Kalman smoother (KS) [7].

In many situations, the latent (state) space, whose dynamics must be modeled, is embedded in a high-dimensional observed space. In general, the direct mapping, from the low-dimensional latent space to the high-dimensional feature (observed) space, as well as the inverse of this mapping, are non-linear. Moreover, the dynamics of the latent space may be non-linear as well. Several methods were proposed to deal with non-linear dynamical systems, e.g. Bayesian tracking with particle filters [8], [9], the extended Kalman filter (EKF) [10], and the unscented Kalman filter (UKF) [11]. Alternatively, it is possible to consider several linear dynamic models and to combine them with a switching mechanism that selects over time one among several linear regimes: this is referred to as the switching Kalman filter (SKF), the switching LDS, the jump-Markov linear system, or the switching state space model [12].

The mixture Kalman filter (MKF) [8] uses a sequential Monte Carlo method based on a random mixture of Gaussians to approximate the target distribution. It formulates non-linear systems into conditional or partial conditional LDSs. Outcomes of non-linear/non-Gaussian Bayesian trackers based on sequential importance sampling are reviewed and discussed in [9], most notably the problems of degeneracy, choice of importance density, and resampling. The basic idea of EKF is to linearize the equations using a first-order Taylor expansion and to apply the standard KF to the linearized model. The additional error due to linearization is not taken into account which may lead to sub-optimal performance. Rather than approximating a non-linear dynamical system with a linear one, UKF specifies the state distribution using a minimal set of deterministically selected sample points. The sample points, when propagated through the true non-linear system, capture the posterior state distribution accurately to the third order Taylor expansion. The stability of UKF was thoroughly investigated in the control literature, e.g. [13]. It was shown that the design of the noise covariance, of both state and observation equations, critically affects the performance of the filter.

The methods just outlined generally deal with a single non-linear or linearized, state equation. There are many real-world processes that cannot be characterized with a single state equation, but with multiple discrete modes of behavior, both in terms of their dynamics and of their observation model, in particular when the latter must predict (generate) high-dimensional observations from a low-dimensional state space.

We consider the problem of tracking the orientation of a person head/face (three rotation angles) from a sequence of images, referred to as *head-pose tracking* (HPT). A face detector provides input to a face descriptor immune to illumination changes, background conditions, as well as inter- and intra-person variabilities (shape and aspect). Face descriptors of choice are histograms of oriented gradients (HOGs) [14] and embeddings based on convolutional neural networks (CNNs) [15]–[18]. These high-dimensional feature vectors contain head pose information implicitly and a number of non-linear or piecewise-linear regression methods have been proposed to extract head pose, namely Gaussian process regression [19], support vector regression [20], kernel partial least squares [21],

X. Alameda-Pineda and R. Horaud are with INRIA Grenoble Rhône-Alpes, Montbonnot Saint-Martin, France.

V. Drouard is with Image Metrics, Manchester, UK.

This work is supported by the Multidisciplinary Institute in Artificial Intelligence (MIAI), Grenoble, and funded by the ANR under grand agreement ANR-19-P3IA-0003; by the ANR Young Researchers Programme ML3RI project (GA ANR-19-CE33-0008-01) ; and by the European Commission under the Horizon 2020 SPRING project (GA #871245).

deep inverse regression [18], or Gaussian mixture of linear regressions [22], [23].

Recently it was proposed to approximate non-linear high-dimensional to low-dimensional (high-to-low) mappings with mixtures of linear-Gaussian [22], [24] and linear-Student [25] regressions. These models adopt an *inverse regression* strategy, namely they learn a low-to-high mapping followed by the evaluation of a high-to-low mapping. The rationale of this way of doing is manyfold: (i) low-to-high regression learning avoids the estimation of a large number of parameters, hence it requires a relatively small amount of training data, (ii) the parameters of the high-to-low regression are analytically evaluated from the low-to-high parameters, (iii) the mixture model setting has the advantage of providing inference procedures using closed-form EM algorithms. It is interesting to note that these Gaussian/Student mixtures group data with similar regression associations together. Within the same cluster, the association can be considered as locally linear, which can then be resolved under the classical linear regression setting. This *piecewise linear* models are well suited to capture potentially non-linear relations. This was extensively discussed in [22] and in [24], and was successfully applied to both head-pose estimation [23] and audio-source localization [26], [27].

In this paper we propose a variational expectation-maximization algorithm to learn *piecewise-linear dynamic systems* (P-LDSs). A P-LDS can be viewed either as a piecewise-linear approximation of a non-linear dynamic system [9], or as a dynamic generalization of mixtures of linear regressions [22], [23], [26]. A P-LDS may also be viewed as soft version of switching LDS [12], [28]. The assignment variable of the piecewise-linear mixture model plays the role of the switching variable of both the dynamic and observation models and it is governed by the transition matrix of a corresponding hidden Markov model (HMM). It is well known that the complexity of these hybrid dynamical systems, i.e. systems that combine discrete- and continuous-valued latent variables, increases exponentially with time [12]. Indeed, for K linear models and after T time steps, the exact marginalized posterior distribution of the state is a Gaussian mixture with K^T components. Therefore, the problem of learning the parameters of such hybrid systems must be carried out via approximate solutions. Traditionally, inference of hybrid dynamical systems is based on approximating the posterior distribution with a simpler one, e.g. the generalized pseudo Bayes filter. In this paper we propose an alternative based on replacing the difficult-to-compute posterior with an approximate tractable posterior.

The remainder of this paper is organized as follows. Section II describes related work. Section III formulates P-LDS and analyses its intractability. Section IV describes in detail the proposed variational approximation model as well as the as two EM algorithms, a variational filter and a variational smoother. For the sake of completeness, Section V describes a GPB2 approximation of P-LDS. Section VI describes experimental results obtained with head-pose tracking. [1]

[1]Supplemental materials can be found at https://team.inria.fr/perception/research/learning-plds/.

II. Related work

The intractability of switching LDS (and of P-LDS) can be addressed using sampling methods: sequential Monte Carlo methods (particle filtering) have been used for this purpose. The main drawback is that sampling can be inefficient, which leads to slow convergence. To reduce the size of the state space, Rao-Blackwellisation may be employed. Instead of drawing the samples from the joint posterior of the discrete and continuous states, tractable sub-structures of the model can be utilized. Non parametric Bayesian inference of switching LDS was proposed in [29], where a switching mechanism is used for the dynamic model, while the observation model uses a standard LDS. This is problematic when the observation model is not linear. Moreover, a Gibbs sampler is used for inference, providing asimptotic properties tied to a high computational cost.

A general theory of Rao-Blackwellised particle filters (RBPFs) applied to dynamic Bayesian networks (DBNs) was proposed in [30]. In the case of switching LDS, marginalization of the joint posterior, namely analytic integration over the continuous variables, considerably reduces the sampling space. RBPF using Gibbs sampling was used for speech recognition [31], where the discrete state corresponds to phonemes and the continuous state corresponds to a time evolving representation of the observations. RBPF using Metropolis-Hastings sampling was used in [32] to analyse motion patterns of bees.

In [32] it was noticed that a naive exploration of the space of discrete variables is prohibitive and that data-driven (DD) MCMC sampling improves convergence. DD-MCMC requires supervised learning from a labeled training dataset, which may be cumbersome, if not prohibitive, because it is not practical to manually associate discrete-variable values with the observed vectors. More recently, [33] proposed a mixture of switching LDSs to analyse the dynamic behavior of pedestrians: an MCMC inference scheme uses both Gibbs and Metropolis-Hastings samplers. While potentially powerful, MCMC methods and their variants are non-analytic methods and typically suffer from slow convergence rates (they are only exact in the case of infinite size samples), especially in high-dimensional spaces, which is impractical in the case of tracking. A recent comparison between MCMC and variational inference emphasizes that the latter easily takes advantage of stochastic and distributed optimization [34].

The *generalized pseudo Bayes* (GPB) [35] and the GPB of order two (GPB2) [36], [37] algorithms belong to the *assumed density filtering* (ADF) [38] class of models which is widely used to approximate an intractable distribution with a simpler one. GPB collapses the mixture of K Gaussian components, resulting from considering all the possible linear models at each time step, into one Gaussian component. GPB2 is more sophisticated and more time consuming than GPB, as it collapses the K^2 Gaussian components, resulting from considering all the possible linear models when going from one time step to the next, into K Gaussians components. The GPB2 algorithm was applied to the analysis of motor cortical activity of hand movements in macaque monkeys

[1], and more recently it was used for tracking eye gaze [39] and for path prediction of pedestrians in the context of intelligent vehicles [5]. An offline extension of GPB2, called expectation correction was proposed in [40] and applied to speech recognition robust to noise [41].

Alternatively, structural variational inference and learning techniques consider a parameterized distribution which is in some sense close to the desired posterior distribution, as well as easier to compute. Variational models modify the structure of the posterior by removing dependencies between variables, i.e. the joint posterior distribution P is approximated by a tractable *variational* distribution Q with variational parameters $\boldsymbol{\theta}$. The Kullback-Leibler divergence between Q and P is minimized with respect to $\boldsymbol{\theta}$. In the case of switching LDSs, [42], [28] and [43] propose to remove some of the dependencies between the continuous and discrete latent variables, thus yielding tractable solvers. The mixed-state DBN proposed in [42] is an HMM driving the LDS bias. In [28] an HMM switches between several LDSs, each LDS having a different latent variable with its own dynamic regime. Both [42] and [28] lead to an EM algorithm whose maximization step (learning) satisfies a set of fixed-point equations in the variational parameters. The variational model proposed in [43] is more general than the one proposed in [42] and in [28]. Nevertheless, their approximation breaks the dependencies between the HMM and the LDS as well as the temporal dependencies.

The variational model of [43] was applied to speech recognition [44] and to speech production [45], while the model of [42] was applied to human motion capture [46]. It is interesting to note that in spite of the recent popularity gained by variational models, e.g. [47], as they provide tractable solutions to various intractable inference problems, e.g. [2], [3], [48]–[56], at the best of our knowledge, variational inference of switching LDS has not been addressed for the last decade.

More recently, learning and inference of dynamical systems have been addressed in the framework of deep generative models (DGMs), where the linear-Gaussian transition and emission distributions of LDS are replaced with non-linear Gaussian models. In detail, the mean and covariance of a Gaussian distribution are modeled with neural networks. Because of this non-linear dependencies, direct optimization of the corresponding data log-likelihood function is intractable. This issue is solved by maximization of a variational lower bound of the log-likelihood. For example, [57] uses a recurrent neural network (RNN) to model the mean and diagonal covariance matrix. The proposed structured inference network corresponds to a deep Kalman smoother, that needs both past and future observations. This formulation belongs to a wider class of non-linear Gaussian state-space models that were recently reviewed in [58]. Deep neural networks can also be used within structured variational inference for pixel-level prediction tasks [59], [60], but we are not aware of any works addressing switching LDS with this methodology.

With respect to the related work just outlined, this paper has the following contributions. We propose a variational approximation of P-LDS. Unlike the variational model of [28], which switches between several linear regimes, we propose a piecewise-linear model that approximates non-linear dynamic models. Unlike [43], the proposed variational approximation doesn't break the temporal dependencies. Unlike the RNN-based non-linear state-space model of [57], the proposed variational piecewise-linear model yields a closed-form solver and it has fewer parameters.

We provide full details of an EM algorithm that can be indifferently used either as a *variational filter* or as a *variational smoother*. Unlike the models of [42] and of [28] that lead to solving a set of fixed-point equations, we develop a closed-form solution for learning the model parameters. Moreover, the proposed method benefits from a closed-form EM algorithm for the off-line estimation of the static parameters, namely, those associated with the observation model. This has two practical outcomes: (i) it reduces the task of learning to the estimation of the parameters of the dynamic model and (ii) it allows to learn the number of states of the switching variable (or, equivalently, the number of linear models) based on a model selection principle, i.e. the Bayes information criterion, [22], [25]. For the sake of completeness, we describe a GPB2 algorithm for P-LDS that slightly differs from the standard GPB2 for switching LDS in that it only needs to estimate the parameters of the dynamic model.

III. Problem Formulation

Let $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^L$ be a latent (or state) random variable and $\boldsymbol{Y} \in \mathcal{Y} \subset \mathbb{R}^D$ be an observation variable. Without loss of generality it will be assumed that the dimensionality of the observation space is much higher than the dimensionality of the latent space, $D \gg L$. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ denote realizations of $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Let $t \in \mathbb{N}$ be the discrete time index: $\boldsymbol{X}_t$ denotes the latent variable at t and the notation $\boldsymbol{X}_{1:t}$ is a shorthand for the temporal sequence $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t$. In an LDS, the observed vectors are connected to the latent vectors through an observation equation. We will consider the following piecewise-linear observation model. It is assumed that at any time t a realization $(\boldsymbol{y}_t, \boldsymbol{x}_t)$ of $(\boldsymbol{Y}_t, \boldsymbol{X}_t)$ is such that $\boldsymbol{y}_t$ is generated from $\boldsymbol{x}_t$ by a linear function $\boldsymbol{y} = l_k(\boldsymbol{x})$ plus an error term. At each time step t, a discrete latent variable Z_t is introduced, such that $Z_t = k$ if and only if $\boldsymbol{y}_t$ is the image of $\boldsymbol{x}_t$ by l_k, with $k \in \{1, \ldots, K\} \subset \mathbb{N}$. The *piecewise* linear function that maps the state $\boldsymbol{X}_t$ onto the observation $\boldsymbol{Y}_t$ is:

$$\boldsymbol{y}_t = \sum_{k=1}^{K} \mathbb{I}(Z_t = k)(\mathbf{A}_{Z_t}\boldsymbol{x}_t + \boldsymbol{b}_{Z_t} + \boldsymbol{e}_{Z_t}), \tag{1}$$

where matrix $\mathbf{A}_{Z_t=k} \in \mathbb{R}^{D\times L}$ and vector $\boldsymbol{b}_{Z_t=k} \in \mathbb{R}^D$ define l_k, $\mathbb{I}(\cdot)$ is the indicator function and $\boldsymbol{e} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ is a zero-mean Gaussian noise vector with covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$. The description of the model is completed by a similar piecewise linear dynamic equation:

$$\boldsymbol{x}_t = \sum_{k=1}^{K} \mathbb{I}(Z_t = k)(\mathbf{C}_{Z_t}\boldsymbol{x}_{t-1} + \boldsymbol{w}_{Z_t}), \tag{2}$$

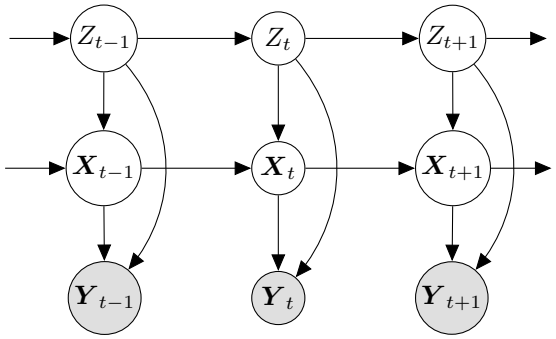

Fig. 1. Graphical representation of the switching linear dynamic model that show the dependencies between the latent variables (white circles) and the observed variables (gray circles).

where $\mathbf{C}_{Z_t} \in \mathbb{R}^{L\times L}$ is the state transition matrix and $\boldsymbol{w} \sim \mathcal{N}(0, \mathbf{Q})$ is a zero-mean Gaussian noise vector with covariance matrix $\mathbf{Q} \in \mathbb{R}^{L\times L}$. To summarize, the k-th LDS is defined by the following probabilistic model, see Fig. 1:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, Z_t = k) = \mathcal{N}(\boldsymbol{x}_t; \mathbf{C}_{Z_t}\boldsymbol{x}_{t-1}, \mathbf{Q}_{Z_t}), \quad (3)$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, Z_t = k) = \mathcal{N}(\boldsymbol{y}_t; \mathbf{A}_{Z_t}\boldsymbol{x}_t + \boldsymbol{b}_{Z_t}, \boldsymbol{\Sigma}_{Z_t}), \quad (4)$$

$$p(\boldsymbol{x}_1|Z_1 = k) = \mathcal{N}(\boldsymbol{x}_1; \boldsymbol{\gamma}_{Z_1}, \boldsymbol{\Gamma}_{Z_1}), \quad (5)$$

$$p(Z_1 = k) = \pi_{Z_1}, \quad (6)$$

where $\{\boldsymbol{\gamma}_k, \boldsymbol{\Gamma}_k, \pi_k\}_{k=1}^K$, $\boldsymbol{\gamma}_k \in \mathbb{R}^L$, $\boldsymbol{\Gamma}_k \in \mathbb{R}^{L\times L}$ and π_k, define the Gaussian mixture parameters of the initial state.

So far we did not specify the temporal behavior of the discrete hidden variables $Z_{1:t}$ that allow to select both the observation model (1) and the dynamic model (2). We assume that $Z_{1:t}$ obey a first-order Markov chain:

$$p(Z_t = i) = \sum_{j=1}^{K} p(Z_t = i|Z_{t-1} = j)p(Z_{t-1} = j)$$

$$\tau_{ij} = p(Z_t = i|Z_{t-1} = j), \quad 1 \leq i, j \leq K \quad (7)$$

where τ_{ij} is an entry of a state transition matrix $\mathbf{T} \in \mathbb{R}^{K\times K}$ which defines the switching behavior: at any time, the system evaluates a convex combination of K Gaussian-linear observation models and dynamic regimes. The model above is described by the following parameters that we group into two sets:

$$\boldsymbol{\theta} = \{\mathbf{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k, \pi_k, \boldsymbol{\gamma}_k, \boldsymbol{\Gamma}_k\}_{k=1}^{k=K}, \quad (8)$$

$$\boldsymbol{\phi} = \{\mathbf{C}_j, \mathbf{Q}_j, \tau_{ij}\}_{i,j=1}^{i,j=K}. \quad (9)$$

The *static* parameters $\boldsymbol{\theta}$ in (8) characterize the observation model (1) and (4), the initial distribution of $\boldsymbol{x}$, (5), and the prior (6): they do not depend on the dynamics of the sequence. Hence, $\boldsymbol{\theta}$ can be learned from a training set of input-output pairs $\{\boldsymbol{y}_n, \boldsymbol{x}_n\}_{n=1}^N$ in the following way. Let $\mathbb{E}_{r_Z}[\log p(\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N}, Z_{1:N}; \boldsymbol{\theta})]$ be the expected complete-data log-likelihood, where the expectation is taken over the responsibilities $r_Z = p(Z_{1:N}|\boldsymbol{y}_{1:N}, \boldsymbol{x}_{1:N}\boldsymbol{\theta})$. The parameter set $\boldsymbol{\theta}$ can be estimated via a closed-form EM procedure, i.e. [22], that alternates between the evaluation of the responsibilities (posteriors) and the maximization of the expectation that was just defined. The number of linear models K, can be estimated via model selection, using the Bayes information criterion (BIC), or empirically, based on the mean absolute error between the predicted outputs and the ground-truth values, i.e [23].

The parameters $\boldsymbol{\phi}$ characterize the *dynamic* behavior of both the continuous (3) and discrete (7) state variables and hence they must be estimated from a temporal sequence of observations. It is interesting to note that when the observation space is of high dimension, the strategy consisting of independently estimating the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ simplifies the tasks of P-LDS inference and learning by drastically reducing the number of parameters. Let, for example, $D = 1000$ (the dimension of the observation space), $L = 10$ (the dimension of the latent space), and $K = 10$ (the number of linear-Gaussian models). If we assume diagonal covariance matrices $\boldsymbol{\Sigma}_k$, $\dim(\boldsymbol{\theta}) \approx 10^5$ and $\dim(\boldsymbol{\phi}) \approx 10^3$.

A. Computational Intractability

Exact model inference, i.e. estimation of the dynamic parameters $\boldsymbol{\phi}$, is faced with an intractability problem. Indeed, let's analyze the complexity of computing the posterior distribution, namely the conditional probability of the state variable $\boldsymbol{x}_t$ given the present and past observations $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$. This distribution can be obtained by marginalization over the continuous and discrete variables given the observations:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \sum_{z_{1:t}=1}^{K} \int_{\boldsymbol{x}_{1:t-1}} p(\boldsymbol{x}_t, \boldsymbol{x}_{1:t-1}, z_{1:t}|\boldsymbol{y}_{1:t}) d\boldsymbol{x}_{1:t-1}, \quad (10)$$

where $\sum_{z_{1:t}=1}^K$, $\int_{\boldsymbol{x}_{1:t-1}}$ and $d\boldsymbol{x}_{1:t-1}$ are shorthands for $\sum_{z_1=1}^K \cdots \sum_{z_t=1}^K$, for $\int_{\boldsymbol{x}_1} \cdots \int_{\boldsymbol{x}_{t-1}}$ and for $d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_{t-1}$, respectively. Using the first-order Markov dependencies shown in the graphical model of Fig. 1, the joint probability (right hand side of (10)) can be factorized as:

$$p(\boldsymbol{x}_{1:t}, z_{1:t}|\boldsymbol{y}_{1:t}) \propto p(\boldsymbol{x}_1, z_1|\boldsymbol{y}_1)$$

$$\times \prod_{t'=2}^{t} p\left(\boldsymbol{y}_{t'}|\boldsymbol{x}_{t'}, z_{t'}\right) p\left(\boldsymbol{x}_{t'}|\boldsymbol{x}_{t'-1}, z_{t'}\right) p\left(z_{t'}|z_{t'-1}\right). \quad (11)$$

Substituting the factors of (11) with their expressions (3)-(7) and using standard properties of Gaussian distributions and using marginalization, the joint distribution (11) writes:

$$p(\boldsymbol{x}_t, \boldsymbol{x}_{1:t-1}, z_{1:t}|\boldsymbol{y}_{1:t}) = \beta_t \, \mathcal{N}\left([\boldsymbol{x}_1 : \boldsymbol{x}_t]; \boldsymbol{\kappa}_t, \mathbf{K}_t\right), \quad (12)$$

where the notation $[\boldsymbol{x}_1 : \boldsymbol{x}_t]$ denotes vertical concatenation of vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$, and where the weight β_t, mean $\boldsymbol{\kappa}_t$, and covariance $\mathbf{K}_t$ depend on the model parameters (8) and (9) and on $Z_{1:t}$. Therefore, the predictive distribution (10) is a Gaussian mixture with a number of components that increases exponentially with time, i.e. there are K^t components after t time steps, which is intractable.

This phenomenon appears not only when attempting to evaluate $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ (filtering), but as well as when $p(\boldsymbol{x}_{1:T}|\boldsymbol{y}_{1:T})$ (smoothing) is evaluated. While the former is used for on-line

prediction, i.e. when the model is already trained, the latter is part of the E-step of any algorithm used for learning the parameters, and therefore equally essential. In the following, we present our variational approximation to perform inference as well as the complete VEM algorithm for learning. We also discuss the derivation of the GPB2 algorithm [36] in the context of the proposed formulation.

IV. Variational Inference and Learning

In this section we present a variational approximation of P-LDS and we derive a variational EM algorithm with tractable inference (expectation) and closed-form parameter learning (maximization). We assume that the continuous and discrete variables are independent, a posteriori: consequently, the joint distribution over $\boldsymbol{x}_{1:T}$ and $z_{1:T}$ is approximated with the following factorization:

$$p(\boldsymbol{x}_{1:T}, z_{1:T}|\boldsymbol{y}_{1:T}) \approx q(\boldsymbol{x}_{1:T})q(z_{1:T}). \tag{13}$$

This follows the same philosophy as the factorial hidden Markov models [61]. However, here we deal with hybrid states, namely discrete and continuous, therefore the derivation [61] does not apply and we need to derive a new EM algorithm. Notice that the proposed model is different than the model of [28]. Indeed, the latter switches between several continuous states, with their own linear dynamic regimes, while the proposed model approximates a possibly non-linear dynamic regime with a piecewise-linear model, similar to GPB2, i.e. (44).

As with HMM and LDS learning, we assume that the entire sequence of observations is available for training and the challenge consists of inferring the entire chain of state variables and of estimating the model parameters, namely the parameter vectors $\boldsymbol{\theta}$, i.e. (8) (observation model) and $\boldsymbol{\phi}$, i.e. (9) (dynamic model). As already outlined in Section III, the parameters $\boldsymbol{\theta}$ don't depend on time and they can be estimated using the algorithm of [22]. Therefore, we only need to estimate the dynamic parameters $\boldsymbol{\phi}$. Formally, we maximize the expected complete-data log-likelihood:

$$\mathcal{Q}(\boldsymbol{\phi}) = \mathbb{E}_{p(\boldsymbol{x}_{1:T}, z_{1:T}|\boldsymbol{y}_{1:T})}\big[\log p(\boldsymbol{x}_{1:T}, z_{1:T}, \boldsymbol{y}_{1:T}|\boldsymbol{\phi})\big], \tag{14}$$

where the posterior distribution $p(\boldsymbol{x}_{1:T}, z_{1:T}|\boldsymbol{y}_{1:T})$ is evaluated with the model parameters at the previous iteration $\boldsymbol{\phi}^{\text{old}}$, implicit in the previous equation to simplify the reading.

A. Inference: The Expectation Steps

The two posterior distributions (13) write:

$$\log q(z_{1:T}) = \mathbb{E}_{q(\boldsymbol{x}_{1:T})}\big[\log p(\boldsymbol{x}_{1:T}, z_{1:T}|\boldsymbol{y}_{1:T})\big] + \text{const}, \tag{15}$$

$$\log q(\boldsymbol{x}_{1:T}) = \mathbb{E}_{q(z_{1:T})}\big[\log p(\boldsymbol{x}_{1:T}, z_{1:T}|\boldsymbol{y}_{1:T})\big] + \text{const}, \tag{16}$$

These distributions are alternatively evaluated, as explained in detail below.

1) E-Z step: By developing (15), ignoring the constant terms and using (3)-(7) we obtain:

$$\begin{aligned}
q(z_{1:T}) \propto\; & \mathcal{N}(\boldsymbol{y}_1; \mathbf{A}_{z_1}\boldsymbol{\eta}_1 + \boldsymbol{b}_{z_1}, \boldsymbol{\Sigma}_{z_1}) \\
& \times \exp\Big(-\tfrac{1}{2}\text{tr}\Big(\mathbf{A}_{z_1}^\top \boldsymbol{\Sigma}_{z_1}^{-1}\mathbf{A}_{z_1}\mathbf{V}_1\Big)\Big) \\
& \times \mathcal{N}(\boldsymbol{\eta}_1; \boldsymbol{\gamma}_{z_1}, \boldsymbol{\Gamma}_{z_1}) \exp\big(-\tfrac{1}{2}\text{tr}\big(\boldsymbol{\Gamma}_{z_1}^{-1}\mathbf{V}_1\big)\big) \\
& \times \prod_{t\geq 2}\Bigg(\mathcal{N}(\boldsymbol{y}_t; \mathbf{A}_{z_t}\boldsymbol{\eta}_t + \boldsymbol{b}_{z_t}, \boldsymbol{\Sigma}_{z_t}) \\
& \quad \times \exp\Big(-\tfrac{1}{2}\text{tr}\Big(\mathbf{A}_{z_t}^\top \boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{A}_{z_t}\mathbf{V}_t\Big)\Big) \\
& \quad \times \mathcal{N}(\boldsymbol{\eta}_t; \mathbf{C}_{z_t}\boldsymbol{\eta}_{t-1}, \mathbf{Q}_{z_t}) \\
& \quad \times \exp\Big(-\tfrac{1}{2}\text{tr}\Big(\mathbf{C}_{z_t}^\top \mathbf{Q}_{z_t}^{-1}\mathbf{C}_{z_t}\mathbf{V}_{t-1} + \mathbf{Q}_{z_t}^{-1}\mathbf{V}_t\Big)\Big) \\
& \quad \times \exp\big(\text{tr}\big(\mathbf{Q}_{z_t}^{-1}\mathbf{C}_{z_t}\mathbf{W}_t\big)\big)\ \tau_{z_{t-1}z_t}\Bigg),
\end{aligned} \tag{17}$$

where $\boldsymbol{\eta}_t = \mathbb{E}_{q(\boldsymbol{x}_{1:T})}[\boldsymbol{x}_t]$ is the posterior mean of $\boldsymbol{x}_t$, $\mathbf{V}_t = \mathbb{E}_{q(\boldsymbol{x}_{1:T})}[\boldsymbol{x}_t\boldsymbol{x}_t^\top] - \boldsymbol{\eta}_t\boldsymbol{\eta}_t^\top$ is the posterior covariance of $\boldsymbol{x}_t$ and $\mathbf{W}_t = \mathbb{E}_{q(\boldsymbol{x}_{1:T})}[\boldsymbol{x}_{t-1}\boldsymbol{x}_t^\top] - \boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_t^\top$ is the posterior cross-covariance of $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$.

One may notice that it is possible to group the terms that depend on z in (17), thus yielding:

$$q(z_{1:T}) \propto \rho^{\text{I}}_{1,z_1}\prod_{t\geq 2}\Big(\rho^{\text{I}}_{t,z_t}\tau_{z_{t-1}z_t}\Big), \tag{18}$$

which is equivalent to an HMM with observation probabilities ρ^{I}_{t,z_t}. Therefore, by computing the standard forward-backward algorithm for HMMs, one can easily obtain the forward ρ^{F}_{t,z_t} and backward ρ^{B}_{t,z_t} probabilities to eventually obtain the posteriori probability $p(z_t|\boldsymbol{y}_{1:T}) \approx q(z_t)$:

$$q(z_t) = \rho_{t,z_t} \propto \frac{\rho^{\text{F}}_{t,z_t}\rho^{\text{B}}_{t,z_t}}{\sum_z \rho^{\text{F}}_{t,z}\rho^{\text{B}}_{t,z}}. \tag{19}$$

The estimation of the transition parameters $\tau_{zz'}$ requires the joint posteriori probability distribution of z_{t-1}, z_t which can be easily computed as:

$$q(z_{t-1}, z_t) = \rho^{\text{J}}_{t,z_{t-1}z_t} \propto \rho^{\text{F}}_{t-1,z_{t-1}}\tau_{z_{t-1}z_t}\rho^{\text{I}}_{t,z_t}\rho^{\text{B}}_{t,z_t}. \tag{20}$$

2) E-X step: By developing (16), ignoring the constant terms and using (3)-(7) we obtain:

$$\begin{aligned}
& \log q(\boldsymbol{x}_{1:T}) \propto \\
& -\tfrac{1}{2}\Big(\boldsymbol{x}_1^\top(\mathbf{V}_1^{\text{I}})^{-1}\boldsymbol{x}_1 - 2\boldsymbol{x}_1^\top(\mathbf{V}_1^{\text{I}})^{-1}\boldsymbol{\eta}_1^{\text{I}} \\
& + \boldsymbol{x}_1^\top(\overline{\boldsymbol{\Gamma}})^{-1}\boldsymbol{x}_1 - 2\boldsymbol{x}_1^\top(\overline{\boldsymbol{\Gamma}})^{-1}\overline{\boldsymbol{\gamma}} \\
& + \sum_{t\geq 2}\big(\boldsymbol{x}_t^\top(\mathbf{V}_t^{\text{I}})^{-1}\boldsymbol{x}_t - 2\boldsymbol{x}_t^\top(\mathbf{V}_t^{\text{I}})^{-1}\boldsymbol{\eta}_t^{\text{I}} + \boldsymbol{x}_t^\top(\overline{\mathbf{Q}}_t)^{-1}\boldsymbol{x}_t \\
& - 2\boldsymbol{x}_t^\top\overline{\mathbf{R}}_t\boldsymbol{x}_{t-1} + \boldsymbol{x}_{t-1}^\top(\overline{\mathbf{S}}_t)^{-1}\boldsymbol{x}_{t-1}\big)\Big),
\end{aligned} \tag{21}$$

where the following definitions hold:

$$\boldsymbol{\eta}_t^{\mathrm{I}} = \mathbf{V}_t^{\mathrm{I}}\Big(\sum_z \rho_{t,z}\mathbf{A}_z^\top \boldsymbol{\Sigma}_z^{-1}(\boldsymbol{y}_t - \boldsymbol{b}_z)\Big), \tag{22}$$

$$(\mathbf{V}_t^{\mathrm{I}})^{-1} = \sum_z \rho_{t,z}\mathbf{A}_z^\top \boldsymbol{\Sigma}_z^{-1}\mathbf{A}_z, \tag{23}$$

$$\overline{\boldsymbol{\gamma}} = \overline{\boldsymbol{\Gamma}}\sum_z \rho_{1,z}\boldsymbol{\Gamma}_z^{-1}\boldsymbol{\gamma}_z, \tag{24}$$

$$(\overline{\boldsymbol{\Gamma}})^{-1} = \sum_z \rho_{1,z}\boldsymbol{\Gamma}_z^{-1}, \tag{25}$$

$$(\overline{\mathbf{Q}}_t)^{-1} = \sum_z \rho_{t,z}\mathbf{Q}_z^{-1}, \tag{26}$$

$$\overline{\mathbf{R}}_t = \sum_z \rho_{t,z}\mathbf{Q}_z^{-1}\mathbf{C}_z, \tag{27}$$

$$(\overline{\mathbf{S}}_t)^{-1} = \sum_z \rho_{t,z}\mathbf{C}_z^\top\mathbf{Q}_z^{-1}\mathbf{C}_z. \tag{28}$$

To be valid, the last equation requires that $\rho_{t,z}\mathbf{C}_z^\top\mathbf{Q}_z^{-1}\mathbf{C}_z$ is invertible for all values of z, this implies that $\mathbf{C}_z$ is a full rank matrix for all values of z, which is a very mild assumption. This is very close to a standard LDS (Kalman filter), but different enough in that standard forward-backward recursions cannot be applied. Indeed, in a standard LDS the following relationship holds: $\overline{\mathbf{R}}_t^\top \overline{\mathbf{Q}}_t^{\ -1}\overline{\mathbf{R}}_t = \overline{\mathbf{S}}_t^{\ -1}$, which is not the case in general. This condition is equivalent to impose the same Gaussian dynamic model to all the realizations of Z_t (which clearly corresponds to a Kalman filter). However, from (21), one may easily see that the joint posterior distribution is a high-dimensional Gaussian, and therefore the marginals will also be Gaussian. Even if the relationship does not correspond to a standard LDS, it is still possible to write the forward-backward equations that efficiently solve in an exact manner the inference of $q(\boldsymbol{x}_t)$.

With the above notations the forward and backward recursions write, respectively:

$$\begin{aligned}\boldsymbol{\eta}_t^{\mathrm{F}} &= \mathbf{V}_t^{\mathrm{F}}\left((\mathbf{V}_t^{\mathrm{I}})^{-1}\boldsymbol{\eta}_t^{\mathrm{I}} + \overline{\mathbf{R}}_t\overline{\mathbf{S}}_t(\overline{\mathbf{S}}_t + \mathbf{V}_{t-1}^{\mathrm{F}})^{-1}\boldsymbol{\eta}_{t-1}^{\mathrm{F}}\right),\\ (\mathbf{V}_t^{\mathrm{F}})^{-1} &= (\mathbf{V}_t^{\mathrm{I}})^{-1} + (\overline{\mathbf{Q}}_t)^{-1} - \overline{\mathbf{R}}_t((\overline{\mathbf{S}}_t)^{-1} + (\mathbf{V}_{t-1}^{\mathrm{F}})^{-1})^{-1}\overline{\mathbf{R}}_t^\top,\end{aligned}$$

and:

$$\begin{aligned}\boldsymbol{\eta}_t^{\mathrm{B}} &= \mathbf{V}_t^{\mathrm{B}}\overline{\mathbf{R}}_t\boldsymbol{\Omega}_{t+1}^{-1}\left((\mathbf{V}_{t+1}^{\mathrm{I}})^{-1}\boldsymbol{\eta}_{t+1}^{\iota} + (\mathbf{V}_{t+1}^{\mathrm{B}})^{-1}\boldsymbol{\eta}_{t+1}^{\leftarrow}\right),\\ (\mathbf{V}_t^{\mathrm{B}})^{-1} &= \overline{\mathbf{S}}_{t+1}^{-1} - \overline{\mathbf{R}}_{t+1}^\top\boldsymbol{\Omega}_{t+1}^{-1}\overline{\mathbf{R}}_{t+1},\\ \boldsymbol{\Omega}_{t+1}^{-1} &= (\mathbf{V}_{t+1}^{\mathrm{I}})^{-1} + \overline{\mathbf{Q}}_{t+1}^{-1} + (\mathbf{V}_{t+1}^{\mathrm{B}})^{-1}.\end{aligned}$$

The forward is initialized with $(\mathbf{V}_1^{\mathrm{F}})^{-1} = \overline{\boldsymbol{\Gamma}}^{-1} + (\mathbf{V}_1^{\mathrm{I}})^{-1}$ and $\boldsymbol{\eta}_1^{\mathrm{F}} = \mathbf{V}_1^{\mathrm{F}}\left(\overline{\boldsymbol{\Gamma}}^{-1}\overline{\boldsymbol{\gamma}} + (\mathbf{V}_1^{\mathrm{I}})^{-1}\boldsymbol{\eta}_1^{\mathrm{I}}\right)$. The backward recursion starts at $t = T-1$ with $(\mathbf{V}_T^{\mathrm{B}})^{-1} = \mathbf{0}$ (and therefore the value of $\boldsymbol{\eta}_T^{\mathrm{B}}$ has no effect). Together, they allow us to write the posterior probability of $\boldsymbol{x}_t$:

$$\begin{aligned}q(\boldsymbol{x}_t|\boldsymbol{y}_{1:T}) &= \mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\eta}_t,\mathbf{V}_t),\\ \boldsymbol{\eta}_t &= \mathbf{V}_t\left((\mathbf{V}_t^{\mathrm{B}})^{-1}\boldsymbol{\eta}_t^{\mathrm{B}} + (\mathbf{V}_t^{\mathrm{F}})^{-1}\boldsymbol{\eta}_t^{\mathrm{F}}\right),\\ (\mathbf{V}_t)^{-1} &= (\mathbf{V}_t^{\mathrm{B}})^{-1} + (\mathbf{V}_t^{\mathrm{F}})^{-1}.\end{aligned} \tag{29}$$

In order to estimate the parameters of the dynamics, $\mathbf{C}_z$ and $\mathbf{Q}_z$, one needs the joint posterior probability of $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$:

$$\begin{aligned}q(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}) &= \mathcal{N}\left(\boldsymbol{x}_t,\boldsymbol{x}_{t-1};\boldsymbol{\eta}_t^{\mathrm{J}},\mathbf{V}_t^{\mathrm{J}}\right),\\ \boldsymbol{\eta}_t^{\mathrm{J}} &= \mathbf{V}_t^{\mathrm{J}}\begin{pmatrix}(\mathbf{V}_t^{\mathrm{I}})^{-1}\boldsymbol{\eta}_t^{\mathrm{I}} + (\mathbf{V}_t^{\mathrm{B}})^{-1}\boldsymbol{\eta}_t^{\mathrm{B}}\\ (\mathbf{V}_{t-1}^{\mathrm{F}})^{-1}\boldsymbol{\eta}_{t-1}^{\mathrm{F}}\end{pmatrix},\\ (\mathbf{V}_t^{\mathrm{J}})^{-1} &= \begin{pmatrix}\boldsymbol{\Omega}_t^{-1} & -\overline{\mathbf{R}}_t\\ -\overline{\mathbf{R}}_t^\top & \overline{\mathbf{S}}_t^{-1} + (\mathbf{V}_{t-1}^{\mathrm{F}})^{-1}\end{pmatrix},\end{aligned} \tag{30}$$

from which we compute the matrix $\mathbf{W}_t$, required in (17), by taking the upper-right block of $\mathbf{V}_t^{\mathrm{J}}$.

B. Learning: The Maximization Step

The estimation of the dynamic parameters is carried out by taking the derivative of the auxiliary function (14) with respect $\boldsymbol{\phi}$. Given the formulas derived in the previous section, the terms of the auxiliary function that depend on $\boldsymbol{\phi}$ are:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\phi}) = &\sum_{t\geq 2}\mathbb{E}_{q(\boldsymbol{x}_t,\boldsymbol{x}_{t-1})q(z_t)}\big[\log p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},z_t)\big]\\ &+\sum_{t\geq 2}\mathbb{E}_{q(z_t,z_{t-1})}\big[\log p(z_t|z_{t-1})\big].\end{aligned} \tag{31}$$

Taking the expectation with respect to all probabilities, including the discrete state variables Z_t, and using the dynamic models, we obtain:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\phi}) = &\frac{1}{2}\sum_{t\geq 2}\sum_z \rho_{tz}\int_{\boldsymbol{x}_t,\boldsymbol{x}_{t-1}}\mathcal{N}((\boldsymbol{x}_t;\boldsymbol{x}_{t-1});\boldsymbol{\eta}_t^{\mathrm{J}},\mathbf{V}_t^{\mathrm{J}})\\ &\times\Big(\log|\mathbf{Q}_z^{-1}| - (\boldsymbol{x}_t - \mathbf{C}_z\boldsymbol{x}_{t-1})^\top\mathbf{Q}_z^{-1}(\boldsymbol{x}_t - \mathbf{C}_z\boldsymbol{x}_{t-1})\Big)\mathrm{d}\boldsymbol{x}_t\boldsymbol{x}_{t-1}\\ &+\sum_{t\geq 2}\sum_{z,z'}\rho_{t,zz'}\log\tau_{zz'}.\end{aligned} \tag{32}$$

Taking the expectation with respect to the continuous variables $\boldsymbol{x}_t$, we obtain:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\phi}) = &\frac{1}{2}\sum_{t\geq 2}\sum_z \rho_{tz}\Big(\log|\mathbf{Q}_z^{-1}|\\ &-(\boldsymbol{\eta}_t - \mathbf{C}_z\boldsymbol{\eta}_{t-1})^\top\mathbf{Q}_z^{-1}(\boldsymbol{\eta}_t - \mathbf{C}_z\boldsymbol{\eta}_{t-1})\\ &-\mathrm{tr}(\mathbf{Q}_z^{-1}(\mathbf{C}_z\mathbf{V}_{t-1}\mathbf{C}_z^\top + \mathbf{V}_t - 2\mathbf{C}_z\mathbf{W}_t))\Big)\\ &+\sum_{t\geq 2}\sum_{z,z}\rho_{t,zz'}\log\tau_{zz'}.\end{aligned} \tag{33}$$

The optimal values of the transition parameters correspond to a standard HMM model and are given by:

$$\tau_{zz'} = \frac{1}{T-1}\sum_{t\geq 2}\rho_{t,zz'}. \tag{34}$$

The optimal value of $\mathbf{Q}_z$ is obtained by taking the derivative of $\mathcal{Q}$ with respect to $\mathbf{Q}_z^{-1}$ and setting this derivative equal to zero:

$$\begin{aligned}\mathbf{Q}_z = &\frac{1}{\sum_{t\geq 2}\rho_{tz}}\sum_{t\geq 2}\rho_{tz}\Big((\boldsymbol{\eta}_t - \mathbf{C}_z\boldsymbol{\eta}_{t-1})(\boldsymbol{\eta}_t - \mathbf{C}_z\boldsymbol{\eta}_{t-1})^\top\\ &+\mathbf{C}_z\mathbf{V}_{t-1}\mathbf{C}_z^\top + \mathbf{V}_t - 2\mathbf{C}_z\mathbf{W}_t\Big).\end{aligned} \tag{35}$$

Taking the derivative of $\mathcal{Q}$ with respect to $\mathbf{C}_z$ is more involved since one needs to take the derivative of the matrix-trace operator. After setting the derivatives equal to zero we obtain:

$$\mathbf{C}_z = \sum_{t\geq 2} \rho_{tz}(\mathbf{W}_t^\top + \boldsymbol{\eta}_t\boldsymbol{\eta}_{t-1}^\top)\Big(\sum_{t\geq 2}\rho_{tz}(\mathbf{V}_{t-1}+\boldsymbol{\eta}_{t-1}\boldsymbol{\eta}_{t-1}^\top)\Big)^{-1} \tag{36}$$

V. GPB2 Inference and Learning

GPB2 [36] is a commonly used algorithm to deal with the intractability of switching LDSs. For the sake of completeness, we describe the GPB2 algorithm for the particular case of P-LDS. As above, only the dynamic parameters $\boldsymbol{\phi}$, i.e. (9), need to be estimated. The central idea of GPB2 is to recursively collapse a K^2-component mixture into a K-component mixture, as explained below. This implies that at each time index t the conditional posterior $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ is successively approximated. To do that, the marginalization chain in (10) is truncated, yielding:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \sum_{z_{t-1}=1}^{K}\sum_{z_t=1}^{K}\int_{\boldsymbol{x}_{t-1}} p(\boldsymbol{x}_{t-1},\boldsymbol{x}_t,z_{t-1},z_t|\boldsymbol{y}_{1:t})d\boldsymbol{x}_{t-1}, \tag{37}$$

with:

$$\begin{aligned} &p(\boldsymbol{x}_{t-1},\boldsymbol{x}_t,z_{t-1},z_t|\boldsymbol{y}_{1:t}) \propto \\ &p(\boldsymbol{y}_t|\boldsymbol{x}_t,z_t)p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},z_t)p(z_t|z_{t-1})p(\boldsymbol{x}_{t-1},z_{t-1}|\boldsymbol{y}_{1:t-1}). \end{aligned} \tag{38}$$

GPB2 yields the following approximation:

$$\begin{aligned} &p(\boldsymbol{x}_{t-1},z_{t-1}|\boldsymbol{y}_{1:t-1}) \\ &\qquad\approx \tilde{\rho}^{\mathrm{F}}_{t-1,z_{t-1}}\mathcal{N}(\boldsymbol{x}_{t-1}|\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t-1,z_{t-1}},\tilde{\mathbf{V}}^{\mathrm{F}}_{t-1,z_{t-1}}), \end{aligned} \tag{39}$$

where $\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}}$, $\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}}$ and $\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}}$ play the same role as $\rho^{\mathrm{F}}_{t,z_{t-1}}$, $\boldsymbol{\eta}^{\mathrm{F}}_{t,z_{t-1}}$ and $\mathbf{V}^{\mathrm{F}}_{t,z_{t-1}}$ in the previous section, but they take different numerical values because they are computed using the GPB2 approximation instead of the variational one.

By using (3), (4), (7) and (38), the approximation in (39) boils down to the following relationship:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) \approx \sum_{z_{t-1}=1}^{K}\sum_{z_t=1}^{K}\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}z_t}\mathcal{N}(\boldsymbol{x}_{t-1}|\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}z_t},\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}z_t}), \tag{40}$$

where the priors $\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}z_t}$, the means $\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}z_t}$, and the covariances $\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}z_t}$ are given by:

$$\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}z_t} = \tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}}\tau_{z_{t-1}z_t}\mathcal{N}\left(\boldsymbol{d}_{t,z_{t-1}z_t};0,\mathbf{S}_{t,z_{t-1}z_t}\right), \tag{41}$$

$$\begin{aligned} \tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}z_t} &= \tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}z_t} \\ &\quad\times\left(\mathbf{A}_{z_t}^\top\boldsymbol{\Sigma}_{z_t}^{-1}(\boldsymbol{y}_t-\boldsymbol{b}_{z_t})+\mathbf{P}_{t,z_{t-1}z_t}\mathbf{C}_{z_t}\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}}\right) \end{aligned} \tag{42}$$

$$\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}z_t} = \left(\mathbf{A}_{z_t}^\top\boldsymbol{\Sigma}_{z_t}^{-1}\mathbf{A}_{z_t}+\mathbf{P}_{t,z_{t-1}z_t}\right)^{-1}, \tag{43}$$

with:

$$\begin{aligned} \boldsymbol{d}_{t,z_{t-1}z_t} &= \boldsymbol{y}_t - \mathbf{A}_{z_t}(\mathbf{C}_{z_t}\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}}) - \boldsymbol{b}_{z_t}, \\ \mathbf{S}_{t,z_{t-1}z_t} &= \boldsymbol{\Sigma}_{z_t} + \mathbf{A}_{z_t}(\mathbf{Q}_{z_t}+\mathbf{C}_{z_t}\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}}\mathbf{C}_{z_t}^\top)\mathbf{A}_{z_t}^\top, \\ \mathbf{P}_{t,z_{t-1}z_t} &= \left(\mathbf{Q}_{z_t}+\mathbf{C}_{z_t}\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}}\mathbf{C}_{z_t}^\top\right)^{-1}. \end{aligned}$$

Consequently, the dynamic model expands the K-component GMM hypothesized in (39) into a K^2-component GMM. GPB2 collapses this K^2 components into K components by moment matching, thus obtaining:

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) \approx \sum_{z_t=1}^{K}\tilde{\rho}^{\mathrm{F}}_{t,z_t}\mathcal{N}(\boldsymbol{x}_t|\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_t},\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_t}), \tag{44}$$

where $\tilde{\rho}^{\mathrm{F}}_{t,z_t},\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_t},\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_t}$ are given by:

$$\tilde{\rho}^{\mathrm{F}}_{t,z_t} = \sum_{z_{t-1}=1}^{K}\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}z_t}, \tag{45}$$

$$\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_t} = \sum_{z_{t-1}=1}^{K}\frac{\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}z_t}}{\tilde{\rho}^{\mathrm{F}}_{t,z_t}}\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}z_t}, \tag{46}$$

$$\begin{aligned} \tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_t} &= \sum_{z_{t-1}=1}^{K}\frac{\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}z_t}}{\tilde{\rho}^{\mathrm{F}}_{t,z_{t-1}}} \\ &\quad\times\left(\tilde{\mathbf{V}}^{\mathrm{F}}_{t,z_{t-1}z_t}+(\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}z_t}-\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_t})(\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_{t-1}z_t}-\tilde{\boldsymbol{\eta}}^{\mathrm{F}}_{t,z_t})^\top\right). \end{aligned} \tag{47}$$

Therefore, at each time index t, the filtering distribution is approximated with a Gaussian mixture with K components, i.e. (44). The same moment matching technique can be recursively applied to the backward (or smoothing) distribution, thus obtaining a K-component Gaussian mixture model for $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:T})$. Finally, it is straightforward to apply the moment matching technique to approximate the joint posterior distribution $p(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:T})$ with a Gaussian mixture with K components, so that the estimation of the transition parameters $\boldsymbol{\phi}$ is done with the same update formulas as in the variational case, see (34), (35) and (36), but using the posterior distribution provided by GPB2.

VI. Experimental Validation

In this section we present experimental evaluations of the proposed P-LDS variational EM filtering and smoothing algorithms, namely head pose tracking (HPT) from a video, e.g. Figure 2. The observed data consist of high-dimensional feature vectors, e.g. histogram of oriented gradients (HOG), but any other visual descriptors could be used in practice. As already explained in Section III, the static parameters can be estimated offline, from a training set of high-dimensional feature vectors (inputs) and the associated ground-truth head-pose angles (outputs). In order to choose the number K of linear-Gaussian components, which is equivalent to the number of states of the switch variable, we use the result of [23] which shows that the Bayesian information criterion (BIC) model [22] may be replaced with an empiric score based on the mean absolute error (MAE) between the predicted head pose and the ground truth. Based on this, we set $K = 25$ in

TABLE I
SUMMARY OF PRINCIPAL FEATURES OF THE DATASETS USED FOR EMPIRICAL EVALUATION.

Dataset	Biwi-Kinect [62]	EYEDIAP [63]	Vernissage [64]
#Recordings	24	94	10
#Participants	20	16	20
Sensor type	RGB-D	RGB-D	RGB
Pitch range	$[-60°, +60°]$	$[-40°, +40°]$	$[-90°, +90°]$
Yaw range	$[-75°, +75°]$	$[-50°, +50°]$	$[-90°, +90°]$
Roll range	$[-20°, +20°]$	N/A	N/A
Annotation method	Automatic fitting with a deformable 3D shape model		Optical motion capture device

all our experiments. It is worthwhile to notice that the off-line training procedure is shared by the variational filter, the variational smoother and GPB2.

A. Experimental Setup

We empirically evaluate the performance of the proposed methods with three publicly available datasets: Biwi-Kinect [62], EYEDIAP [63] and Vernissage [64] (see Table I):

- The Biwi-Kinect dataset contains 24 videos of 20 people (16 men and 4 women) recorded with a Kinect camera. During the recordings, people were asked to move their heads freely in front of the camera. 3D head pose (pitch, yaw, and roll angles) annotations were obtained automatically and accurately for each video frame using the face-shift software. The angle values range from $-60°$ to $60°$ for pitch, $-75°$ to $75°$ for yaw and $-20°$ to $20°$ for roll. The dataset provides RGB and depth images as well as the calibration matrices. The 3D nose positions are provided as well.
- The EYEDIAP dataset is intended for both eye-gaze and head-pose estimation. It contains 94 videos of 16 people recorded using different configurations, e.g. static and rotating heads. The dataset provides RGB videos, with both HD and VGA resolution, and depth videos with the associated calibration matrices. Annotations of both head-pose and eye-gaze are provided for each video frame. The angle values range from $-40°$ to $40°$ for pitch and $-50°$ to $50°$ for yaw.
- The Vernissage dataset contains 10 recordings of 20 people interacting with each and with a robot. Each recording comprises two people. The dataset was recorded with a camera embedded into the head of a robot head. A network of infrared cameras combined with optical markers placed on the participants' heads provide accurate ground-truth head positions and orientations.

B. Implementation details

Facial features are computed as follows. A face detector provides bounding boxes for each frame and for each video. Then a high-dimensional feature vector is extracted from each bounding box using histogram of oriented gradients (HOG) descriptors, obtained along the implementation described in [23] which yields vector-valued observations of dimension $D = 1888$. For all datasets and for each face identity, we split the corresponding videos into two disjoint sets: a training set and a test set. Since the datasets are annotated, i.e. there are ground-truth head-pose parameter values associated with each frame. As already mentioned, we use the method of [23] to estimate $\boldsymbol{\theta}$.

The dynamic parameters are initialized as follows. First we set $\mathbf{C}_k = \mathbf{I}$ since we noticed that the simultaneous estimation of $\mathbf{C}_k$ and $\mathbf{Q}_k$ is subject to instabilities in the estimation of the dynamic parameters. The covariance matrices $\left\{\mathbf{C}_j, \mathbf{Q}_j\right\}_{j=1}^{K}$ are initialized with the identity matrix. The entries of the transition matrix $\{\tau_{ij}\}_{i,j=1}^{K}$ are initialized in the following way. We compute the pairwise Bhattacharrya distances [65] between the K Gaussian components defined by (5). The variational EM algorithm alternates between inference of $q(Z_t)$ (E-Z step) and of $q(\boldsymbol{x}_t)$ (E-X step). The variational parameters $\boldsymbol{\eta}_t$ and $\mathbf{V}_t$ are initialized with their previously estimated values, namely: $\boldsymbol{\eta}_t = \boldsymbol{\eta}_{t-1}$ and $\mathbf{V}_t = \mathbf{V}_{t-1}$.

We implemented both the variational filter and the variational smoother described in detail in the previous section. The main difference between the filter and the smoother is the amount of information that is used at inference time. Indeed, while the filter uses only causal information (i.e. past observations) the smoother uses also non-causal information (i.e. past and future observations). This is why, a priori, one expects the smoother to have better performance, at the price of being a completely off-line algorithm that cannot be used for real-time applications.

C. Results and Discussion

The proposed methods were compared with the following state-of-the-art HPT methods:

- *Flandmarks* [66], [67] combines 2D face landmark detection with head pose estimation and with tracking. Head pose is estimated in the following way: the 3D landmarks of a mean face are projected onto the image plane and the error between the projected landmarks and the observed landmarks is minimized over the pose parameters. At each time step this non-linear minimizer is initialized with the pose parameters computed at the previous time step. The publicly available implementation of this method only computes the pitch and yaw angles.
- *OpenFace* [68] is an open source software package for facial behavior analysis, i.e. facial landmark detection and tracking, head pose and eye gaze estimation. It extracts 68 2D facial landmarks using conditional local neural fields and tracks them over time with a three-layer CNN trained to predict landmark detection errors [69]. Once 2D landmarks are detected and tracked, they are used in conjunction with a 3D facial model to compute head pose parameters.
- *ICP tracking* [70] uses both depth and color (RGB-D images) from which 16 3D facial landmarks are manually

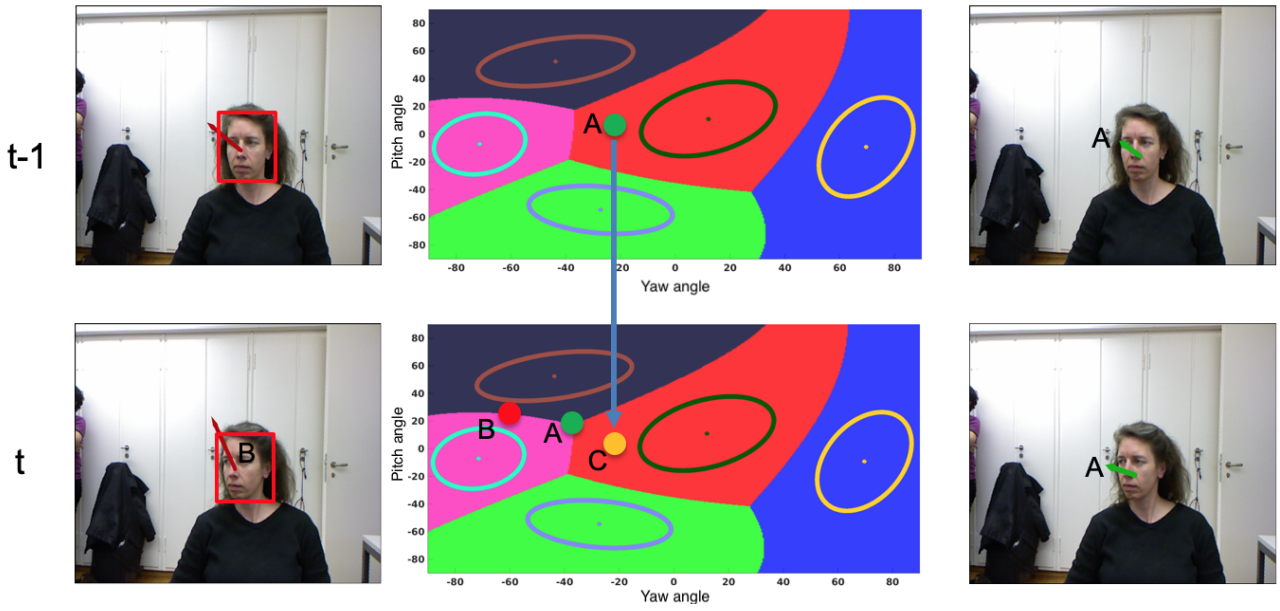

Fig. 2. The proposed variational P-LDS algorithm applied to the problem of head pose tracking (HPT). The central column shows the Gaussian mixture that models the latent space, i.e. eq. (5). The parameters of this mixture don't vary over time and they are learnt from a training set of input-output instances of the observed and latent variables. In this example we show the likelihood function associated with the latent variables of head pose, namely the yaw and pitch angles. The observed pose at t (red dot denoted B) is estimated from a high-dimensional feature vector that describes a face (left column). The variational means (green dots denoted A and shown with green arrows onto the right column) are inferred by the E-X step of the algorithm (22) based on the current dynamic prediction (orange dot denoted C) and the current observation (red dot denoted B).

extracted. The landmarks are tracked based on estimating the rigid motion between consecutive frames.

These HPT methods make use of 2D facial landmarks, whose detection is known to be sensitive to large head poses that induce partial occlusions of the face. This stays in contrast with the proposed method that directly exploits high-dimensional descriptors of faces. For completeness, we also compared our algorithms with an implementation of the Kalman filter (LDS), namely (1)-(6) with $K = 1$, with the HPE method of [23] and with GPB2, i.e. Section V and [71]. To quantitatively evaluate HPT we compute average and standard deviation of the absolute error between the estimated pose parameters and the ground-truth parameters provided with each annotated dataset.

The results obtained with the Biwi-Kinect, EYEDIAP and Vernissage datasets are summarized in tables II, III and IV. The HPT method of [66], [67] does not estimate the roll angle. Moreover, this method relies on 2D landmark detection. Therefore, it fails whenever all the landmarks are not properly detected. This is the reason for which this method did not provide publishable results, when applied to the EYEDIAP and Vernissage datasets. The method of [70] takes as input RGB-D images and manually annotated facial landmarks, hence it could only be applied to the EYEDIAP dataset. In general, the proposed descriptor-based trackers perform better than the landmark-based trackers. A lower standard deviation corresponds to a higher precision and a better method repeatability.

In order to assess the statistical differences in performance between the various methods, we used the Wilcoxon signed-rank test [72], inspired from [73], which is a non-parametric

TABLE II
AVERAGE (AVG.) AND STANDARD DEVIATION (STD.) OF THE ABSOLUTE ERROR (IN DEGREES) FOR THE PITCH, YAW AND ROLL ANGLES (WHEN APPLICABLE) ON THE BIWI-KINECT DATASET. THE LANDMARK-BASED METHOD OF [66], [67] ONLY ESTIMATES THE PITCH AND THE ROLL ANGLES. THE BEST RESULTS ARE IN BOLD AND THE SECOND BEST ARE IN SLANTED BOLD. THE RESULTS THAT DID NOT PASS THE WILCOXON STATISTICAL TEST ARE MARKED WITH AN ASTERISK. THE SAME FACE DETECTOR WAS USED BY ALL METHODS.

	Pitch		Yaw		Roll	
Method	Avg.	Std.	Avg.	Std.	Avg.	Std.
HPE_GLLiM_HOG [23]	10.54	13.38	11.15	17.93	5.23	5.99
HPT_Flandmarks [66], [67]	13.12	10.79	21.1	14.16	–	–
HPT_OpenFace [68]	9.23	15.69	29.43*	25.74	10.72*	11.33
HPT_GLLiM_Kalman	10.35	13.19	***10.97***	17.75	5.12	5.93
HPT_GPB2_HOG	**9.03**	10.89	**8.77**	***13.42***	***4.75***	5.11
HPT_VarFilter_HOG	9.39	***8.95***	11.81	14.06	4.96	***5.01***
HPT_VarSmoother_HOG	***9.08***	**8.64**	11.06	***13.47***	***4.87***	**4.95**

statistical hypothesis test that is commonly used to compare two related samples in order to assess the null hypothesis that the median difference between pairs of observations is zero. It can be used as an alternative to the paired Student's t-test (t-test for matched pairs) when the population cannot be assumed to be normally distributed. For most of the comparisons, there were no statistical differences. The only notable exception is OpenFace [68], which performs considerably worse than the other methods on two of the three datasets: the yaw and roll estimation for the Biwi-Kinect dataset, and the pitch and yaw estimation for the EYEDIAP dataset. These estimation are marked with an asterisk in Table II and Table III, respectively.

As an example, Figure 3 shows a ground-truth yaw (left-

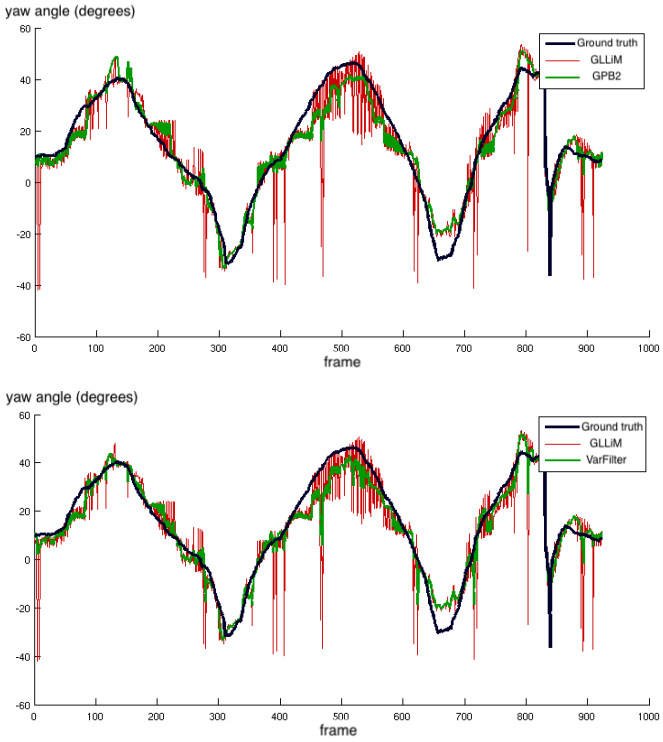

Fig. 3. An example from the Biwi-Kinect dataset of a person that rotates his head from left to right and then from right to left (yaw angle). The ground truth trajectory is shown with a blue curve, the result of head-pose estimation (HPE_GLLiM) is shown in red. The results obtained with the HPT_GPB2 algorithm (top) and with the proposed HPT_VarFilter algorithm (bottom), respectively. Notice that the large errors that are produced by HPE_GLLiM are eliminated by both trackers.

right rotation) trajectory as well as trajectories obtained with HPE_GLLiM [23], with HPT_GPB2 (top) and with the proposed variational filter (bottom). Overall, the performance of the proposed variational EM algorithms is comparable with the performance of GPB2. Notice that the large errors produced by HPE_GLLiM are filtered by both trackers.

We measured the CPU time needed to compute one time step. This comprises the following processes: (i) extraction of a HOG descriptor from a face, (ii) head pose estimation, and (iii) tracking, where the tracker can be either the GPB2 or the variational filter. Using an Intel-Xeon and Matlab, it takes 1.0 second to extract a HOG descriptor and to estimated head pose, 8.55 seconds to run GPB2 and 2.45 seconds to run the variational filter, respectively.

VII. Conclusions

In this paper we addressed the problem of learning and inference of piecewise LDS. The latter belongs to the switching LDS class of models, which is known to be intractable because of the combinatorial explosion, over time, of the modes of the latent space. The standard way of dealing with this problem is to use the GPB2 algorithm which collapses a K^2-component mixture into a K-component one based on moment matching – a computationally demanding process.

TABLE III
AVERAGE (AVG.) AND STANDARD DEVIATION (STD.) OF THE ABSOLUTE ERROR (IN DEGREES) FOR THE PITCH AND YAW ANGLES ON THE EYEDIAP DATASET. THE METHOD OF [70] USES BOTH COLOR AND DEPTH DATA AND IT REQUIRES MANUALLY ANNOTATED 2D LANDMARKS. THE BEST RESULTS ARE IN BOLD AND THE SECOND BEST ARE IN SLANTED BOLD. THE RESULTS THAT DID NOT PASS THE WILCOXON STATISTICAL TEST ARE MARKED WITH AN ASTERISK. THE SAME FACE DETECTOR WAS USED BY ALL METHODS.

	Pitch		Yaw	
Method	Avg.	Std.	Avg.	Std.
HPE_GLLiM_HOG [23]	6.29	7.80	***7.80***	10.39
ICP tracking [70]	**4.17**	**5.59**	**6.89**	14.42
OpenFace [68]	15.39*	12.85	22.21*	16.32
HPT_GLLiM_Kalman	***6.21***	***7.75***	10.62	**10.31**
HPT_GPB2_HOG	6.68	8.75	8.44	***10.91***
HPT_VarFilter_HOG	6.96	8.04	11.38	11.44
HPT_VarSmoother_HOG	6.78	7.88	10.66	10.99

TABLE IV
AVERAGE (AVG.) AND STANDARD DEVIATION (STD.) OF THE ABSOLUTE ERROR (IN DEGREES) FOR THE PITCH AND YAW ANGLES ON THE VERNISSAGE DATASET. THE BEST RESULTS ARE IN BOLD AND THE SECOND BEST ARE IN SLANTED BOLD. THE SAME FACE DETECTOR WAS USED BY ALL METHODS.

	Pitch		Yaw	
Method	Avg.	Std.	Avg.	Std.
HPE_GLLiM_HOG [23]	23.95	23.18	***11.03***	8.57
OpenFace [68]	21.30	***18.80***	13.18	10.67
HPT_GLLiM_Kalman	23.94	23.18	***11.03***	8.56
HPT_GPB2_HOG	**20.24**	20.62	**10.21**	**7.80**
HPT_VarFilter_HOG	21.06	19.96	13.76	8.25
HPT_VarSmoother_HOG	***20.37***	**18.58**	12.92	***7.89***

Alternatively, we propose a variational approximation of P-LDS and two associated EM algorithms: a variational filter and a variational smoother. Both the filter and the smoother are based on closed-form expressions, which guarantees efficient computation and fast convergence. The proposed variational filter is of the order of 3.5 times faster than the GPB2 method. Not surprisingly, the most time-consuming part of GPB2 is the evaluation of the parameters of the K^2-component Gaussian mixture, followed by the evaluation of the parameters of the approximating K-component mixture. This collapsing process resides at the heart of GPB2 and it cannot be avoided. In contrast, the M-step of the proposed VEM can be skipped, once the parameters are learnt, as done in [2], which can further accelerate the algorithm.

We applied the proposed algorithms to the problem of head-pose tracking. We presented a series of experiments using several datasets. We carried out a benchmark that included our algorithms and several state-of-the-art tracking algorithms. We note that the variational-based tracker compares well with GPB2 and with the other trackers. It should be noted, however,

that the landmark based methods, e.g. [66], [67], fail to track in many cases. The best performing method is [70]. Nevertheless, this method has limitations because it requires manual annotation of facial landmarks and it can only be applied to RGB-D images. In contrast, the proposed method is based on extracting high-dimensional descriptors from RGB images, and appears to be more robust against facial self occlusions than the landmark-based methods.

REFERENCES

[1] W. Wu, M. J. Black, D. Mumford, Y. Gao, E. Bienenstock, and J. P. Donoghue, "Modeling and decoding motor cortical activity using a switching Kalman filter," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 6, pp. 933–942, 2004.

[2] S. Ba, X. Alameda-Pineda, A. Xompero, and R. Horaud, "An on-line variational Bayesian model for multi-person tracking from cluttered scenes," *Computer Vision and Image Understanding*, vol. 153, pp. 64–76, 2016.

[3] M. Byeon, M. Lee, K. Kim, and J. Y. Choi, "Variational inference for 3-D localization and tracking of multiple targets using multiple cameras," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[4] S. Thrun, W. Burgard, and D. Fox, *Probalistic Robotics*. MIT Press, 2005.

[5] J. F. Kooij, F. Flohr, E. A. Pool, and D. M. Gavrila, "Context-based path prediction for targets with switching dynamics," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 239–262, 2019.

[6] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering*, vol. 82, no. Series D, pp. 35–45, March 1960.

[7] H. E. Rauch, C. Striebel, and F. Tung, "Maximum likelihood estimates of linear dynamic systems," *AIAA Journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[8] R. Chen and J. S. Liu, "Mixture Kalman filters," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 62, no. 3, pp. 493–508, 2000.

[9] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, February 2002.

[10] A. H. Jazwinski, *Stochastic processes and filtering theory*. Courier Corporation, 2007.

[11] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.

[12] K. P. Murphy, *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

[13] K. Xiong, H. Zhang, and C. Chan, "Performance evaluation of UKF-based nonlinear filtering," *Automatica*, vol. 42, no. 2, pp. 261–270, 2006.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, June 2005, pp. 886–893 vol. 1.

[15] B. Ahn, J. Park, and I. S. Kweon, "Real-time head orientation from a monocular camera using deep neural network," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 82–96.

[16] S. S. Mukherjee and N. M. Robertson, "Deep head pose: Gaze-direction estimation in multimodal video," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2094–2107, 2015.

[17] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[18] S. Lathuilière, R. Juge, P. Mesejo, R. Munoz-Salinas, and R. Horaud, "Deep mixture of linear inverse regressions applied to head-pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] C. E. Rasmussen, *Gaussian processes for machine learning*. MIT Press, 2006.

[20] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[21] H. Abdi, "Partial least square regression (PLS regression)," *Encyclopedia for research methods for the social sciences*, pp. 792–795, 2003.

[22] A. Deleforge, F. Forbes, and R. Horaud, "High-dimensional regression with Gaussian mixtures and partially-latent response variables," *Statistics and Computing*, vol. 25, no. 5, pp. 893–911, 2015.

[23] V. Drouard, R. Horaud, A. Deleforge, S. Ba, and G. Evangelidis, "Robust head-pose estimation based on partially-latent mixture of linear regressions," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1428–1440, 2017.

[24] C.-C. Tu, F. Forbes, B. Lemasson, and N. Wang, "Prediction with high dimensional regression via hierarchically structured Gaussian mixtures and latent variables," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 68, no. 5, pp. 1485–1507, 2019.

[25] E. Perthame, F. Forbes, and A. Deleforge, "Inverse regression approach to robust non-linear high-to-low dimensional mapping," *Journal of Multivariate Analysis*, 2017.

[26] A. Deleforge, R. Horaud, Y. Y. Schechner, and L. Girin, "Co-localization of audio sources in images using binaural features and locally-linear regression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 718–731, 2015.

[27] X. Li, L. Girin, R. Horaud, and S. Gannot, "Estimation of the direct-path relative transfer function for supervised sound-source localization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2171–2186, 2016.

[28] Z. Ghahramani and G. E. Hinton, "Variational learning for switching state-space models," *Neural computation*, vol. 12, no. 4, pp. 831–864, 2000.

[29] E. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "Bayesian nonparametric inference of switching dynamic linear models," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1569–1585, 2011.

[30] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-Blackwellised particle filtering for dynamic Bayesian networks," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 176–183.

[31] A.-V. I. Rosti and M. J. Gales, "Rao-blackwellised Gibbs sampling for switching linear dynamical systems," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. 809–812.

[32] S. M. Oh, J. M. Rehg, T. Balch, and F. Dellaert, "Learning and inferring motion patterns using parametric segmental switching linear dynamic systems," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 103–124, 2008.

[33] J. F. Kooij, G. Englebienne, and D. M. Gavrila, "Mixture of switching linear dynamics to discover behavior patterns in object tracks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 322–334, 2016.

[34] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[35] Y. Bar-Shalom and X.-R. Li, *Estimation and tracking: Principles, techniques, and software*. Artech House, 1993.

[36] Y. Bar-Shalom and T. E. Fortmann, *Tracking and Data Association*. Academic Press, 1988.

[37] K. P. Murphy, "Dynamic Bayesian networks: representation, inference and learning," Ph.D. dissertation, University of California, Berkeley, 2002.

[38] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 1998, pp. 33–42.

[39] B. Massé, S. Ba, and R. Horaud, "Tracking gaze and visual focus of attention of people involved in social interaction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 11, pp. 2711–2724, 2018.

[40] D. Barber, "Expectation correction for smoothed inference in switching linear dynamical systems," *Journal of Machine Learning Research*, vol. 7, no. Nov, pp. 2515–2540, 2006.

[41] B. Mesot and D. Barber, "Switching linear dynamical systems for noise robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1850–1858, 2007.

[42] V. Pavlovic, B. J. Frey, and T. S. Huang, "Variational learning in mixed-state dynamic graphical models," in *Proceedings of Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 522–530.

[43] L. J. Lee, H. Attias, and L. Deng, "Variational inference and learning for segmental switching state space models of hidden speech dynamics," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2003.

[44] L. J. Lee, H. Attias, L. Deng, and P. Fieguth, "A multimodal variational approach to learning and inference in switching state space models," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. V, 2004, pp. 505–508.

[45] L. Deng, "Switching dynamic system models for speech articulation and acoustics," in *Mathematical Foundations of Speech and Language Processing*. Springer, 2004, pp. 115–133.
[46] V. Pavlovic, J. M. Rehg, and J. MacCormick, "Learning switching linear models of human motion," in *Proceedings of Neural Information Processing Systems*, 2000.
[47] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 2008–2026, 2019.
[48] Z. Ma, A. E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, and J. Guo, "Variational Bayesian matrix factorization for bounded support data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 4, pp. 876–889, 2014.
[49] Z. Ma, P. K. Rana, J. Taghia, M. Flierl, and A. Leijon, "Bayesian estimation of Dirichlet mixture model with variational inference," *Pattern Recognition*, vol. 47, no. 9, pp. 3143–3157, 2014.
[50] J. Taghia, Z. Ma, and A. Leijon, "Bayesian estimation of the von-Mises Fisher mixture model with variational inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1701–1715, 2014.
[51] J. Taghia and A. Leijon, "Variational inference for Watson mixture model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1886–1900, 2015.
[52] A. Deleforge, F. Forbes, S. Ba, and R. Horaud, "Hyper-spectral image analysis with partially latent regression and spatial Markov dependencies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 6, pp. 1037–1048, 2015.
[53] Z. Ma, J. Xie, Y. Lai, J. Taghia, J.-H. Xue, and J. Guo, "Insights into multiple/single lower bound approximation for extended variational inference in non-Gaussian structured data modeling," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.
[54] Y. Ban, X. Alameda-Pineda, F. Badeig, S. Ba, and R. Horaud, "Tracking a varying number of people with a visually-controlled robotic head," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 4144–4151.
[55] Y. Ban, X. Alameda-Pineda, C. Evers, and R. Horaud, "Tracking multiple audio sources with the von Mises distribution and variational EM," *IEEE Signal Processing Letters*, vol. 26, no. 6, pp. 798–802, 2019.
[56] Y. Ban, X. Alameda-Pineda, L. Girin, and R. Horaud, "Variational Bayesian inference for audio-visual tracking of multiple speakers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[57] R. G. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 2101–2109.
[58] L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda, "Dynamical variational autoencoders: A comprehensive review," 2020.
[59] D. Xu, W. Ouyang, X. Alameda-Pineda, E. Ricci, X. Wang, and N. Sebe, "Learning deep structured multi-scale features using attention-gated crfs for contour prediction," in *Advances in Neural Information Processing Systems*, 2017, pp. 3961–3970.
[60] D. Xu, X. Alameda-Pineda, W. Ouyang, E. Ricci, X. Wang, and N. Sebe, "Probabilistic graph attention network with conditional kernels for pixel-wise prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[61] Z. Ghahramani and M. I. Jordan, "Factorial hidden markov models," in *Advances in Neural Information Processing Systems*, 1996.
[62] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, February 2013.
[63] K. A. F. Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: a database for the development and evaluation of gaze estimation algorithms from RGB and RGB-D cameras," in *ACM Symposium on Eye Tracking Research and Applications*, March 2014, pp. 255–258.
[64] D. B. Jayagopi, S. Sheikhi, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede, V. Khalidov, L. Nguyen, B. Wrede, and D. Gatica-Perez, "The vernissage corpus: A multimodal human-robot-interaction dataset," Ecole Polytechnique Federale de Lausanne (EPFL, Tech. Rep., 2012.
[65] A. Bhattacharyya, "On a measure of divergence between two statistical population defined by their population distributions," *Bulletin Calcutta Mathematical Society*, vol. 35, pp. 99–109, July 1943.
[66] M. Uřičář, V. Franc, and V. Hlaváč, "Detector of facial landmarks learned by the structured output SVM," in *International Conference on Computer Vision Theory and Applications*, February 2012.
[67] J. Cech, V. Franc, and J. Matas, "A 3D approach to facial landmarks: detection, refinement, and tracking," in *International Conference on Pattern Recognition*. IEEE, 2014, pp. 2173–2178.
[68] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016.
[69] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 354–361.
[70] K. A. F. Mora and J.-M. Odobez, "Gaze estimation from multimodal kinect data," in *IEEE CVPRW*, June 2012.
[71] V. Drouard, S. Ba, and R. Horaud, "Switching linear inverse-regression model for tracking head pose," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2017, pp. 1232–1240.
[72] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics*. Springer, 1992, pp. 196–202.
[73] S. Lathuilière, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2065 – 2081, 2020.

Xavier Alameda-Pineda received M.Sc. degrees in mathematics (2008), in telecommunications (2009) and in computer science (2010) and a Ph.D. in mathematics and computer science (2013) from Université Joseph Fourier. Since 2016, he is a research scientist at Inria Grenoble Rhône-Alpes, with the Perception team. He served as Area Chair at ICCV'17, of ICIAP'19 and of ACM MM'19. He is the recipient of several paper awards and of the ACM SIGMM Rising Star Award in 2018. He is the coordinator of the H2020 ICT SPRING project. His scientific interests lie in computer vision, machine learning and signal processing for robotics and multimodal social behavior analysis.

Vincent Drouard received the M.Sc. in Applied Mathematics from the School of Engineering, Université Nice Sophia-Antipolis in 2013 and the Ph.D. in Computer Science from Université Grenoble Alpes in 2017. During the period 2014-2017 he was a member of the Perception team at Inria Grenoble Rhône-Alpes. Currently he is a research scientist with Image Metrics, Manchester, UK. He received the best student paper award at IEEE International Conference on Image Processing (ICIP'15).

Radu Horaud received the B.Sc. degree in Electrical Engineering (1977), the M.Sc. degree in Control Engineering (1979), and the Ph.D. degree in Computer Science (1981), all from Grenoble INP, France. In the past he held research positions with SRI International (1982-1984) and with CNRS (1984-1998). Since 1998 he has been director of research with Inria Grenoble Rhône-Alpes, where he leads the Perception team. His research interests include computer vision, machine learning, audio signal processing, audio-visual analysis, and robotics. Radu Horaud and his collaborators received numerous best paper awards. He was an area editor of *Computer Vision and Image Understanding*, an associate editor of the *International Journal of Computer Vision*. and an editorial board member of the *International Journal of Robotics Research*. Radu Horaud was awarded two ERC grants, an advanced grant for the project *Vision and Hearing in Action* (2014-2019) and a proof of concept grant (2018-2019).