A Variational Bayesian Methodology for Hidden Markov Models utilizing Student's-t Mixtures

Sotirios P. Chatzis

Department of Electrical and Electronic Engineering
Imperial College London, UK
Exhibition Road, South Kensington Campus, SW7 2BT

Dimitrios I. Kosmopoulos

Institute of Informatics and Telecommunications
NSCR Dimokritos
P. Grigoriou & Neapoleos Str.
15310 Athens, Greece

Abstract

The Student's-t hidden Markov model (SHMM) has been recently proposed as a robust to outliers form of conventional continuous density hidden Markov models, trained by means of the expectation-maximization algorithm. In this paper, we derive a tractable variational Bayesian inference algorithm for this model. Our innovative approach provides an efficient and more robust alternative to EM-based methods, tackling their singularity and overfitting proneness, while allowing for the automatic determination of the optimal model size without cross-validation. We highlight the superiority of the proposed model over the competition using synthetic and real data. We also demonstrate the merits of our methodology in applications from diverse research fields, such as human computer interaction, robotics and semantic audio analysis.

Key words: Hidden Markov models; Student's-t distribution; variational Bayes; speaker identification; robotic task failure; violence detection

1 Introduction

The hidden Markov model (HMM) is increasingly being adopted in applications since it provides a convenient way of modeling observations appearing in a sequential manner and tending to cluster or to alternate between different possible components (subpopulations). Specifically, HMMs with continuous observation densities have been used in a wide spectrum of applications in ecology, encryption, image understanding, speech recognition, and machine vision applications [1]. The hidden observation densities associated with each state of a continuous HMM must be capable of approximating arbitrarily complex probability density functions. Finite Gaussian mixture models (GMMs) are the most common selection of emission distribution models in the continuous HMM literature [2]. Their popularity stems from the well-known capability of GMMs to successfully approximate unknown random distributions, including distributions with multiple modes, while also providing a simple and computationally efficient maximum-likelihood (ML) estimation framework using the expectation-maximization (EM) algorithm [3]. Nevertheless, GMMs do also suffer from a significant drawback concerning their parameters estimation procedure, which is well-known to be adversely affected by the presence of outliers in the data sets used for the model fitting.

To tackle these issues, we have proposed in [4] a novel form of continuous HMMs where the hidden state distributions are modeled using finite mixtures of multivariate Student's-t densities. The multivariate Student's-t distribu-

Email addresses: `soteri0s@me.com` (Sotirios P. Chatzis), `dkosmo@iit.demokritos.gr` (Dimitrios I. Kosmopoulos).

tion is a bell-shaped distribution with heavier tails compared to the Gaussian; as a consequence, Student's-t mixture models (SMMs) provide an alternative to GMMs means of probabilistic generative modeling with high robustness to training data outliers. The so-obtained Student's-t hidden Markov model (SHMM) has been considered in [4] under the ML paradigm using the EM algorithm; as it has been shown, the SHMM provides an effective, computationally efficient and application-independent means for outlier tolerant representation and classification of sequential data by means of continuous HMMs.

In this paper, we provide an alternative treatment of the SHMM under a *Bayesian* framework using a *variational approximation*, yielding the *variational Bayesian* SHMM (VB-SHMM). Variational Bayesian treatments of statistical models present significant advantages over ML-based alternatives: ML approaches have the undesirable property of being ill-posed since the likelihood function is unbounded from above [5,6,7]. This fact results in several very significant shortcomings. To begin with, a significant difficulty concerns the infinities which plague the likelihood function, associated with the collapsing of the bell-shaped component distributions onto individual data points and, hence, resulting in singular or near-singular covariance matrices [7]. Obviously, the adoption of a Bayesian model inference algorithm, providing posterior distributions over the model parameters instead of point-estimates, would allow for the natural resolution of these issues [5,6,7]. Another central issue ML treatments of generative models are confronted with concerns selection of the optimal model size. Maximum likelihood is unable to address this issue since it favors models of ever-increasing complexity, thus leading to over-fitting [17,10].

In our work, we conduct a *Bayesian treatment* of the SHMM, overcoming the problems of ML approaches elegantly, by marginalizing over the model parameters with respect to appropriate priors. The resulting model (marginal) likelihood can then be maximized with respect to the model size, in case one aims at optimal model selection, or combined with a prior over the model size if the goal is model averaging [17,16]. Our novel approach is based on *variational approximation* methods [8], which have recently emerged as a deterministic alternative to Markov chain Monte-Carlo (MCMC) algorithms for doing Bayesian inference for probabilistic generative models [9,10], with better scalability in terms of computational cost [11]. Variational Bayesian inference has previously been applied to relevance vector machines [12], GMMs [13], autoregressive models [14,15], SMMs [16,17], mixtures of factor analyzers [18,19,20], discrete HMMs [21], Gaussian HMMs [22], as well as HMMs with Poisson and autoregressive observation models [23], thereby ameliorating the singularity and overfitting problems of ML approaches.

The remainder of this paper is organized as follows: In Section 2, a brief review of the SHMM is provided. In Section 3, the proposed variational Bayesian treatment of the SHMM is carried out, yielding the variational Bayesian SHMM algorithm. In Section 4, the experimental evaluation of the proposed algorithm is conducted, considering a series of data modeling and classification applications and using real-world data sets. In the final section, our results are summarized and discussed.

2 The Student's-t HMM

Let us suppose an N-state HMM where the hidden emission density of each state is modeled by a K-component finite mixture model. Considering that the component distributions of the K-component finite mixture models modeling the HMM state densities are multivariate Student's-t distributions, the definition of the Student's-t HMM is obtained. The pdf of a d-dimensional Student's-t distribution with mean $\boldsymbol{\mu}$, precision $\boldsymbol{R}$, and ν degrees of freedom is given by

$$t(\boldsymbol{x}_t|\boldsymbol{\mu}, \boldsymbol{R}, \nu) = \frac{\mathit{\Gamma}\left(\frac{\nu+d}{2}\right)|\boldsymbol{R}|^{1/2}(\pi\nu)^{-d/2}}{\mathit{\Gamma}(\nu/2)\{1 + \mathrm{MD}(\boldsymbol{x}_t, \boldsymbol{\mu}|\boldsymbol{R}^{-1})/\nu\}^{(\nu+d)/2}} \tag{1}$$

where $\mathrm{MD}(\boldsymbol{x}_t, \boldsymbol{\mu}|\boldsymbol{R}^{-1})$ is the squared Mahalanobis distance between $\boldsymbol{x}_t, \boldsymbol{\mu}$ with covariance matrix (inverse precision) $\boldsymbol{R}^{-1}$ [24] and $\mathit{\Gamma}(.)$ is the Gamma function.

The SHMM can be modeled by the set of parameters $\Psi = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Theta}, \boldsymbol{\nu}\}$, where $\boldsymbol{\pi} = (\pi_i)_{i=1}^{N}$ is the initial-state probability vector, $\boldsymbol{A} = (a_{ij})_{i,j=1}^{N}$ is the $N \times N$ one-step transition matrix, $\boldsymbol{C} = (c_{ij})_{i,j=1}^{N,K}$ is the $N \times K$ mixture coefficient matrix, with c_{ij} denoting the mixing proportion of the jth component density of the hidden emission distribution of the ith SHMM state, $\boldsymbol{\Theta}$ is the $N \times K$ parameter matrix that comprises the means $\boldsymbol{\mu}_{ij}$ and the precisions $\boldsymbol{R}_{ij}$ of the constituent Student's-t densities of the model, that is $\boldsymbol{\Theta} = (\theta_{ij})_{i,j=1}^{N,K}$ where $\theta_{ij} = \{\boldsymbol{\mu}_{ij}, \boldsymbol{R}_{ij}\}$, and $\boldsymbol{\nu} = (\nu_{ij})_{i,j=1}^{N,K}$ is the NK vector of the degrees of freedom of the model component densities.

Let $X = \{\boldsymbol{x}_t\}_{t=1}^{T}$ be an observed data sequence, with $\boldsymbol{x}_t \in \mathcal{X} \subseteq \mathbb{R}^d$, modeled by an SHMM. The latent (unobserved) data associated with this sequence comprise the corresponding state sequence $S = \{s_t\}_{t=1}^{T}$, where $s_t = 1, \ldots, N$ is the indicator of the state the tth observation is emitted from, and the sequence of the corresponding mixture component indicators $L = \{l_t\}_{t=1}^{T}$, where $l_t = 1, \ldots, K$ indicates the mixture component density that generated the tth observation. The likelihood of the parameters set Ψ of the SHMM given the observable data X is, then, given by

$$p(X|\Psi) = \sum_{S,L} \pi_{s_1} \left[\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right] \left[\prod_{t=1}^{T} c_{s_t l_t} p(\boldsymbol{x}_t|\theta_{s_t l_t}, \nu_{s_t l_t})\right] \tag{2}$$

As it has been discussed in [24], there is no closed-form solution for likelihood maximization of a Student's-t distribution. However, a computationally elegant solution can be obtained [16,17] by exploiting the property of the Student's-t distribution [24]

$$t(\boldsymbol{x}_t|\boldsymbol{\mu}, \boldsymbol{R}, \nu) = \int_0^\infty \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}, u_t\boldsymbol{R})\mathcal{G}(u_t|\nu/2, \nu/2)\mathrm{d}u_t \tag{3}$$

which implies that a Student's-t density can be viewed as an infinite sum of Gaussians with the same mean and scaled precisions, where the precision scalars are Gamma-distributed latent variables depending on the degrees of freedom of the Student's-t density. Let us denote as $U = \{u_{s_t l_t}\}$ the sequence of the (latent) precision scalars associated with the observed data, depending on the corresponding unobserved state sequence and mixture component indicator sequence. Then, we have that

$$\boldsymbol{x}_t \sim t(\boldsymbol{\mu}_{s_t l_t}, \boldsymbol{R}_{s_t l_t}, \nu_{s_t l_t}) \tag{4}$$

is equivalent to

$$\boldsymbol{x}_t|u_{s_t l_t} \sim \mathcal{N}(\boldsymbol{\mu}_{s_t l_t}, u_{s_t l_t}\boldsymbol{R}_{s_t l_t}) \tag{5}$$

where

$$p(u_{s_t l_t}|\nu_{s_t l_t}) = \mathcal{G}(u_{s_t l_t}|\nu_{s_t l_t}/2, \nu_{s_t l_t}/2) \tag{6}$$

Under this regard, and using (3), the likelihood of the SHMM (2) eventually becomes

$$p(X|\Psi) = \sum_{S,L} \pi_{s_1} \int \mathrm{d}u_{s_t l_t} \left[\prod_{t=1}^{T-1} a_{s_t s_{t+1}}\right] \left[\prod_{t=1}^{T} c_{s_t l_t} p(\boldsymbol{x}_t|\theta_{s_t l_t}, u_{s_t l_t}) p(u_{s_t l_t}|\nu_{s_t l_t})\right] \tag{7}$$

3 Variational Bayesian Inference for the SHMM

Variational Bayesian inference for the SHMM comprises introduction of a set of prior distributions over the model parameters and further maximization of the log marginal likelihood (log evidence) of the resulting model. For convenience, we choose priors conjugate to the considered observable and latent data, as this selection greatly simplifies inference and interpretability [8]. This way, the prior for the initial-state probabilities vector is chosen to follow a Dirichlet distribution

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\phi}^{\pi}) = \mathcal{D}(\pi_1, \dots \pi_N|\phi_1^{\pi}, \dots, \phi_N^{\pi}) \tag{8}$$

In the same fashion, we choose

$$p(\boldsymbol{A}) = \prod_{i=1}^{N} \mathcal{D}(a_{i1}, \dots, a_{iN}|\phi_{i1}^{A}, \dots, \phi_{iN}^{A}) \tag{9}$$

$$p(\boldsymbol{C}) = \prod_{i=1}^{N} \mathcal{D}(c_{i1}, \dots, c_{iK}|\phi_{i1}^{C}, \dots, \phi_{iK}^{C}) \tag{10}$$

Under the equivalent expression (5) of the Student's-t distribution, we let the joint (conjugate exponential) prior on the means and the precisions of the mixture component densities of the SHMM hidden states be

$$p(\{\boldsymbol{\mu}_{ij}, \boldsymbol{R}_{ij}\}_{i,j=1}^{N,K}) = \prod_{i=1}^{N}\prod_{j=1}^{K} \mathcal{NW}(\boldsymbol{\mu}_{ij}, \boldsymbol{R}_{ij}|\lambda_{ij}, \boldsymbol{m}_{ij}, \eta_{ij}, \boldsymbol{S}_{ij}) \tag{11}$$

where $\mathcal{NW}(\boldsymbol{\mu}_{ij}, \boldsymbol{R}_{ij}|\lambda_{ij}, \boldsymbol{m}_{ij}, \eta_{ij}, \boldsymbol{S}_{ij})$ is a Normal-Wishart distribution with hyperparameters $\lambda_{ij}, \boldsymbol{m}_{ij}, \eta_{ij}$, and $\boldsymbol{S}_{ij}$. Finally, no conjugate prior exists for the degrees of freedom ν_{ij} of the model component densities. Instead, these parameters will be estimated as model hyperparameters, by optimization as a part of the variational inference procedure discussed next.

Having introduced prior distributions over the SHMM parameters, the formulation of the *variational Bayesian* SHMM is complete. The graph of the

VB-SHMM can be found in Fig. 1. Therefore, we can proceed to the estimation of the marginal likelihood of the data. Exact inference in our Bayesian model is intractable. Nevertheless, the choice of conjugate exponential prior distributions for the model parameters allows for the derivation of an elegant variational framework. Let $\tilde{\Psi}$ be the set of the stochastic model variables of the VB-SHMM, that is, the latent model variables and the model parameters on which a conjugate exponential prior has been imposed, $\tilde{\Psi} = \{S, L, U, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Theta}\}$. The variational Bayesian treatment of the VB-SHMM is conducted by introducing an arbitrary distribution $q(\tilde{\Psi}) = q(S, L, U, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Theta})$ and considering the well-known equality for the log evidence, $\log p(X)$ [11]

$$\log p(X) = F(q) + \mathrm{KL}(q||p) \tag{12}$$

where

$$F(q) = \int \mathrm{d}\tilde{\Psi} q(\tilde{\Psi}) \log \frac{p(X, \tilde{\Psi})}{q(\tilde{\Psi})} \tag{13}$$

In (12), $\mathrm{KL}(q||p)$ stands for the Kullback-Leibler (KL) divergence between the arbitrary distribution $q(\tilde{\Psi})$, which is considered as the approximate (variational) posterior over the model variables, and $p(\tilde{\Psi}|X)$ which is the true posterior over the model variables; it is given by

$$\mathrm{KL}(q||p) = -\int \mathrm{d}\tilde{\Psi} q(\tilde{\Psi}) \log \frac{p(\tilde{\Psi}|X)}{q(\tilde{\Psi})} \tag{14}$$

Since the KL divergence is a non-negative quantity, it follows from (12) that $F(q)$ is a strict lower bound of the log evidence, i.e.

$$\log p(X) \geq F(q) \tag{15}$$

and would become exact if $q(\tilde{\Psi}) = p(\tilde{\Psi}|X)$. Hence, maximizing the lower bound of the log evidence (variational free energy), $F(q)$, so that it becomes as tight as possible, i.e. minimizing the KL divergence between the true and the variational posterior, a good variational inference scheme for the VB-SHMM is obtained.

In order to yield a tractable expression for the variational free energy of the VB-SHMM, we assume that the joint variational posterior over the stochastic variables associated with the VB-SHMM, $q(\tilde{\Psi}) = q(S, L, U, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Theta})$, factorizes over the latent variables and the model parameters as

$$q(\tilde{\Psi}) = q(S, L, U, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Theta}) \approx q(S, L)q(U)q(\boldsymbol{\pi})q(\boldsymbol{A})q(\boldsymbol{C})q(\boldsymbol{\Theta}) \tag{16}$$

where $q(S, L) = q(L|S)q(S)$. Factorization of $q(\tilde{\Psi})$ on the form (16) is a common approach in variational Bayesian inference (see e.g. [16,25,21,26]). Then, having chosen a family of approximating (variational) posterior distributions, we can now search for the optimal member of this family by maximization of the variational free energy, thus increasing $F(q)$ on $\log p(X)$, the exact log marginal likelihood.

From (13) and (16), the variational free energy, $F(q)$, reads

$$\begin{aligned} F(q) =& \int \mathrm{d}S\mathrm{d}L\mathrm{d}U\mathrm{d}\boldsymbol{\pi}\mathrm{d}\boldsymbol{A}\mathrm{d}\boldsymbol{C}\mathrm{d}\boldsymbol{\Theta} q(S, L)q(U)q(\boldsymbol{\pi})q(\boldsymbol{A})q(\boldsymbol{C})q(\boldsymbol{\Theta}) \Bigg[\log \pi_{s_1} \\ &+ \sum_{t=1}^{T-1} \log a_{s_t s_{t+1}} + \sum_{t=1}^{T} \log c_{s_t l_t} + \sum_{t=1}^{T} \log p(\boldsymbol{x}_t | \theta_{s_t l_t}, u_{s_t l_t}) \\ &+ \log p(U|\boldsymbol{\nu}) + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{A}) + \log p(\boldsymbol{C}) + \log p(\boldsymbol{\Theta}) \\ &- \log q(S, L) - \log q(U) - \log q(\boldsymbol{\pi}) - \log q(\boldsymbol{A}) - \log q(\boldsymbol{C}) - \log q(\boldsymbol{\Theta}) \Bigg] \\ =& F(q(\boldsymbol{\pi})) + F(q(\boldsymbol{A})) + F(q(\boldsymbol{C})) + F(q(\boldsymbol{\Theta})) + F(q(S, L)) + F(q(U)) \end{aligned} \tag{17}$$

where, the analytical expressions of the terms constituting $F(q)$ are provided in the Appendix. From (17), it follows that $F(q)$ is a non-convex function of the variational posterior distribution [27]. As a consequence, there will in general exist multiple maxima of $F(q)$, and, hence, the solution obtained from the variational inference procedure will depend, indeed, on the initialization. This issue can be easily addressed by performing multiple optimizations from

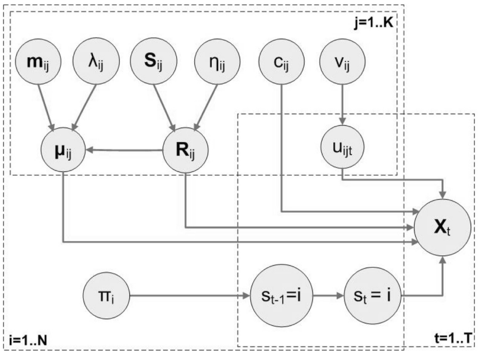


Figure 1. Graphical representation of the VB-SHMM. $\{\boldsymbol{X}_t\}_{t=1}^{T}$ is the observed sequence. The arrows represent conditional dependencies. The plates indicate independent copies for states i, component mixtures j, and data samples t. Some variables are related to more than one plates. The dependency of the observed variables results from the Normal-Wishart distribution, and the variables s_t, s_{t-1} represent the current and the previous state at time t.

different random starts, and retaining the solution yielding the largest value of the variational free energy, $F(q)$. We note that a benefit of the upper bounded nature of $F(q)$, as a result of the adoption of the proposed variational Bayesian approach, is that this optimization procedure allows us to use the entire training set in a single pass of training and does not require cross-validation, as is the case with ML approaches (such as the EM algorithm and its derivatives) [17], where the optimized objective function is unbounded.

3.1 Variational Posteriors

The expressions of the variational posteriors over the VB-SHMM variables can be derived by maximizing $F(q)$ with respect to each one of the factors of $q(\tilde{\Psi})$ in turn, holding the others fixed, in an iterative manner where each iteration resembles an EM algorithm iteration [28]; on the E-step of these iterations, the variational posteriors over the VB-SHMM latent variables, $q(S, L)$ and $q(U)$, are updated, while, on the M-step, updating of the variational posteriors over the model parameters is conducted. At the end of each iteration, the value of the variational free energy, $F(q)$, is estimated and used to apply a variational inference convergence criterion. We note that, as a consequence of the conjugate exponential structure of our model, the resulting optimal factors of the variational posterior distribution, $q(\tilde{\Psi})$, are expected to take the same functional form as the corresponding conditional (prior) distributions [27]. We also mention that, by construction, the iterative, consecutive updating of the interdependent distributions of the considered factors of $q(\tilde{\Psi})$ is guaranteed to monotonically and maximally increase the variational free energy $F(q)$ [18].

Let us denote as $< \chi >_\xi$ the mean of the expression χ with respect to the probability density function ξ. We begin with the updates of the variational posteriors over the VB-SHMM parameters (M-step of the algorithm). From (17), and by collecting all the quantities related to $q(\boldsymbol{A})$ together, we have

$$F(q(\boldsymbol{A})) = \int \mathrm{d}\boldsymbol{A} q(\boldsymbol{A}) \log \left[\frac{\prod_{i=1}^{N} \prod_{j=1}^{N} a_{ij}^{\omega_{ij}^{A}-1}}{q(\boldsymbol{A})} \right] \tag{18}$$

where

$$\omega_{ij}^{A} = \sum_{t=1}^{T-1} \gamma_{ijt}^{A} + \phi_{ij}^{A} \tag{19}$$

and

$$\gamma_{ijt}^{A} \triangleq q(s_t = i, s_{t+1} = j) \tag{20}$$

Then, from (18) and using Gibbs inequality, maximization of $F(q(\boldsymbol{A}))$ yields the variational posterior

$$q(\boldsymbol{A}) = \prod_{i=1}^{N} \mathcal{D}(a_{i1}, \ldots, a_{iN} | \omega_{i1}^{A}, \ldots, \omega_{iN}^{A}) \tag{21}$$

In the same fashion, we can optimize $F(q)$ with respect to $q(\boldsymbol{\pi})$ and $q(\boldsymbol{C})$, to obtain their expressions; this yields

$$q(\boldsymbol{\pi}) = \mathcal{D}(\pi_1, \ldots, \pi_N | \omega_1^{\pi}, \ldots, \omega_N^{\pi}) \tag{22}$$

where

$$\omega_i^{\pi} = \gamma_i^{\pi} + \phi_i^{\pi} \tag{23}$$

$$\gamma_i^{\pi} \triangleq q(s_1 = i) \tag{24}$$

and

$$q(\boldsymbol{C}) = \prod_{i=1}^{N} \mathcal{D}(c_{i1}, \ldots, c_{iK} | \omega_{i1}^{C}, \ldots, \omega_{iK}^{C}) \tag{25}$$

where

$$\omega_{ij}^{C} = \sum_{t=1}^{T} \gamma_{ijt}^{C} + \phi_{ij}^{C} \tag{26}$$

$$\gamma_{ijt}^{C} \triangleq q(s_t = i, l_t = j) \tag{27}$$

Finally, from (17), and by collecting all the quantities related to $q(\boldsymbol{\Theta})$ together, we have

$$F(q(\boldsymbol{\Theta})) = \int \mathrm{d}\boldsymbol{\Theta} q(\boldsymbol{\Theta}) \int \mathrm{d}U q(U) \log \left\{ \frac{\prod_{i=1}^{N} \prod_{j=1}^{K} p(\theta_{ij}) \prod_{t=1}^{T} \left[p(\boldsymbol{x}_t | \theta_{ij}, u_{ijt})\right]^{\gamma_{ijt}^{C}}}{q(\boldsymbol{\Theta})} \right\} \tag{28}$$

where

$$u_{ijt} \triangleq u_{s_t l_t} | s_t = i, l_t = j \tag{29}$$

Using the expression of $F(q(\boldsymbol{\Theta}))$ given by (28), log evidence maximization w.r.t. $q(\boldsymbol{\Theta})$ yields

$$q(\boldsymbol{\Theta}) = \prod_{i=1}^{N} \prod_{j=1}^{K} q(\theta_{ij}) \tag{30}$$

with

$$q(\theta_{ij}) = q(\boldsymbol{\mu}_{ij}, \boldsymbol{R}_{ij}) = \mathcal{NW}(\boldsymbol{\mu}_{ij}, \boldsymbol{R}_{ij} | \tilde{\lambda}_{ij}, \tilde{\boldsymbol{m}}_{ij}, \tilde{\eta}_{ij}, \tilde{\boldsymbol{S}}_{ij}) \tag{31}$$

where we introduce the notation

$$\tilde{\gamma}_{ij} \triangleq \sum_{t=1}^{T} \gamma_{ijt}^{C} \left\langle u_{ijt} \right\rangle_{q(u_{ijt})} \tag{32}$$

$$\bar{\boldsymbol{x}}_{ij} \triangleq \frac{\sum_{t=1}^{T} \gamma_{ijt}^{C} \left\langle u_{ijt} \right\rangle_{q(u_{ijt})} \boldsymbol{x}_t}{\tilde{\gamma}_{ij}} \tag{33}$$

$$\boldsymbol{\Delta}_{ij} \triangleq \sum_{t=1}^{T} \gamma_{ijt}^{C} \left\langle u_{ijt} \right\rangle_{q(u_{ijt})} \left(\boldsymbol{x}_t - \bar{\boldsymbol{x}}_{ij}\right) \left(\boldsymbol{x}_t - \bar{\boldsymbol{x}}_{ij}\right)^{\mathrm{T}} \tag{34}$$

and, it holds

$$\tilde{\eta}_{ij} = \eta_{ij} + \sum_{t=1}^{T} \gamma_{ijt}^{C} \tag{35}$$

$$\tilde{\boldsymbol{S}}_{ij} = \boldsymbol{S}_{ij} + \boldsymbol{\Delta}_{ij} + \frac{\lambda_{ij}\tilde{\gamma}_{ij}}{\lambda_{ij} + \tilde{\gamma}_{ij}} \left(\boldsymbol{m}_{ij} - \bar{\boldsymbol{x}}_{ij}\right) \left(\boldsymbol{m}_{ij} - \bar{\boldsymbol{x}}_{ij}\right)^{\mathrm{T}} \tag{36}$$

$$\tilde{\lambda}_{ij} = \lambda_{ij} + \tilde{\gamma}_{ij} \tag{37}$$

$$\tilde{\boldsymbol{m}}_{ij} = \frac{\lambda_{ij}\boldsymbol{m}_{ij} + \tilde{\gamma}_{ij}\bar{\boldsymbol{x}}_{ij}}{\lambda_{ij} + \tilde{\gamma}_{ij}} \tag{38}$$

Let us, now, consider the expressions of the variational posteriors over the latent model variables (E-step of the algorithm). We begin with the variational posteriors over the precision scalars $u_{s_t l_t}$. From (17), we have

$$\begin{aligned} F(q(U)) = & \sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{t=1}^{T} \int \mathrm{d}u_{ijt} q(u_{ijt}) \Bigg[\log \frac{p(u_{ijt}|\nu_{ij})}{q(u_{ijt})} \\ & + \gamma_{ijt}^{C} \int \mathrm{d}\theta_{ij} q(\theta_{ij}) \log p(\boldsymbol{x}_t|\theta_{ij}, u_{ijt})\Bigg] \end{aligned} \tag{39}$$

Then, maximization of $F(q)$ with respect to $q(u_{ijt})$ yields

$$\begin{aligned} \log q(u_{ijt}) \propto & \left(\frac{\nu_{ij}}{2} - 1\right) \log u_{ijt} - \frac{\nu_{ij}}{2} u_{ijt} + \gamma_{ijt}^{C} \frac{d}{2} \log u_{ijt} \\ & - \gamma_{ijt}^{C} \frac{1}{2} \left[\frac{d}{\tilde{\lambda}_{ij}} + \tilde{\eta}_{ij} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{ij}\right)^{\mathrm{T}} \tilde{\boldsymbol{S}}_{ij}^{-1} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{ij}\right)\right] u_{ijt} \end{aligned} \tag{40}$$

and, hence,

$$q(U) = \prod_{i=1}^{N}\prod_{j=1}^{K}\prod_{t=1}^{T} q(u_{ijt}) \tag{41}$$

where

$$q(u_{ijt}) = \mathcal{G}(\alpha_{ijt}, \beta_{ijt}) \tag{42}$$

and

$$\alpha_{ijt} = \frac{\nu_{ij} + \gamma_{ijt}^{C} d}{2} \tag{43}$$

$$\beta_{ijt} = \frac{1}{2}\left\{\nu_{ij} + \gamma_{ijt}^{C} \frac{d}{\tilde{\lambda}_{ij}} + \gamma_{ijt}^{C} \tilde{\eta}_{ij} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{ij}\right)^{\mathrm{T}} \tilde{\boldsymbol{S}}_{ij}^{-1} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{ij}\right)\right\} \tag{44}$$

Finally, concerning the joint variational posterior over the state indicator sequence and the mixture component indicator sequence $q(S, L)$, from (17) we have

$$F(q(S,L)) = \sum_{S,L} q(S,L)\mathrm{log}\frac{\pi^*_{s_1}\prod_{t=1}^{T-1} a^*_{s_t s_{t+1}} \prod_{t=1}^{T} c^*_{s_t l_t} p^*(\boldsymbol{x}_t|\theta_{s_t l_t}, u_{s_t l_t})}{q(S,L)} \tag{45}$$

which yields the optimizer

$$q(S,L) = \frac{1}{W}\pi^*_{s_1}\prod_{t=1}^{T-1} a^*_{s_t s_{t+1}} \prod_{t=1}^{T} c^*_{s_t l_t} p^*(\boldsymbol{x}_t|\theta_{s_t l_t}, u_{s_t l_t}) \tag{46}$$

where

$$\pi^*_i \triangleq \exp\left[\langle \mathrm{log}\pi_i\rangle_{q(\boldsymbol{\pi})}\right] \tag{47}$$

$$a^*_{ij} \triangleq \exp\left[\langle \mathrm{log}a_{ij}\rangle_{q(\boldsymbol{A})}\right] \tag{48}$$

$$c^*_{ij} \triangleq \exp\left[\langle \mathrm{log}c_{ij}\rangle_{q(\boldsymbol{C})}\right] \tag{49}$$

$$p^*(\boldsymbol{x}_t|\theta_{s_t l_t}, u_{s_t l_t}) \triangleq \exp\left[\langle \mathrm{log}p(\boldsymbol{x}_t|\theta_{s_t l_t}, u_{s_t l_t})\rangle_{q(U),q(\boldsymbol{\Theta})}\right] \tag{50}$$

and the normalizing constant W is given by

$$W = \sum_{S,L}\pi^*_{s_1}\prod_{t=1}^{T-1} a^*_{s_t s_{t+1}} \prod_{t=1}^{T} c^*_{s_t l_t} p^*(\boldsymbol{x}_t|\theta_{s_t l_t}, u_{s_t l_t}) \tag{51}$$

Comparing the expression of the variational posterior distribution $q(S,L)$, given by (46), with the expression of the conditional probability $p(S,L|X,\Psi)$ computed in the context of the EM algorithm for the SHMM [4], we notice that, in essence

$$q(S,L) = p(S,L|X,\Psi^*) \tag{52}$$

where $\Psi^* = \{\boldsymbol{\pi}^*, \boldsymbol{A}^*, \boldsymbol{C}^*, \boldsymbol{\Theta}^*, \boldsymbol{\nu}\}$ and $\boldsymbol{\Theta}^* = \{\theta^*_{ij}\}_{i,j=1}^{N,K}$ is such that

$$p^*(\boldsymbol{x}_t|\theta_{s_t l_t}, u_{s_t l_t}) = p(\boldsymbol{x}_t|\theta^*_{s_t l_t}) \tag{53}$$

Therefore, the probabilities γ^A_{ijt}, γ^π_i, and γ^C_{ijt}, defined in (20), (24), and (27), respectively, that constitute the variational posterior $q(S,L)$, can be easily computed by means of the forward-backward algorithm, as described in [2], by using the set of the posterior expected values Ψ^* as the optimized values (“point”-estimates) of the SHMM parameters.

3.2 Hyperparameter Selection

Let us, now, deal with the proper selection of the values of the hyperparameters of the model priors, i.e. $\{\boldsymbol{\nu}, \phi_i^{\pi}, \phi_{ij}^{A}, \phi_{ij}^{C}, \lambda_{ij}, \boldsymbol{m}_{ij}, \eta_{ij}, \boldsymbol{S}_{ij}\}_{i,j=1}^{N,K}$. We begin with the degrees of freedom of the model component densities. Taking derivatives of $F(q)$ with respect to ν_{ij} we obtain that ν_{ij} is given by the solution of the equation

$$\log\frac{\nu_{ij}}{2} + 1 - \psi\left(\frac{\nu_{ij}}{2}\right) + \frac{1}{\sum_{t=1}^{T}\gamma_{ijt}^{C}}\sum_{t=1}^{T}\gamma_{ijt}^{C}\left(\langle \log u_{ijt}\rangle_{q(u_{ijt})} - \langle u_{ijt}\rangle_{q(u_{ijt})}\right) = 0 \quad (54)$$

For the rest of the prior hyperparameters of the VB-SHMM, instead of determining their optimal expression with respect to the model's variational free energy, we select instead a set of proper *ad hoc* values. This is preferable due to the fact that the benefit from optimizing their values is not significant, when a good *ad hoc* value selection can be conducted [18,17]; on the contrary, the computational burden imposed by this procedure is significant, mainly due to the open-form formulas required to be computed (see e.g. [18,17]). For instance, a good selection might be obtained by setting $\lambda_{ij} = 1$, $\boldsymbol{m}_{ij} = \boldsymbol{0}$, so as to obtain broad components over $\boldsymbol{\mu}_{ij}$, and $\eta_{ij} = 20$, $\boldsymbol{S}_{ij} = 200\boldsymbol{I}$ to allow for more moderate components over $\boldsymbol{R}_{ij}$. On the other hand, concerning the hyperparameters of the Dirichlet priors over the Markov chain probabilities, i.e. ϕ_i^{π}, ϕ_{ij}^{A}, and ϕ_{ij}^{C}, interpreting them as effective numbers of prior observations, one may set their values to $\phi = 10^{-3}$.

3.3 Model Size Selection

Proper selection of the model size is a significant and generally difficult problem in the field of probabilistic generative models. In our variational Bayesian setting, we do not impose a prior over the model size parameters, that is the number of states, N, and the number of mixture component densities per model state, K. Instead, the adoption of the proposed Bayesian approach allows the optimal values of the number of states and of the mixture components to be obtained by merely running the variational inference procedure for different values of N and K and selecting the one that yields the biggest value of the variational free energy, $F(q)$, since this approximates the log marginal likelihood for the model. Note that, on the contrary, in ML approaches, using some variant of the EM algorithm, usually cross-validation techniques are employed against an independent data set to select an appropriate model complexity, a method which imposes a heavy computational burden and is also prone to well-known over-fitting problems [7].

3.4 Approximation of the Predictive Density

Let us consider an already estimated VB-SHMM, that is an SHMM for which the proposed variational Bayesian treatment has already been conducted. The ultimate goal of Bayesian learning is, given a set of test data, to perform density estimation with respect to the trained model. Let us suppose a test sequence $X' = \{\boldsymbol{x}_t'\}_{t=1}^{T'}$ and a VB-SHMM trained using the training sequence $X = \{\boldsymbol{x}_t\}_{t=1}^{T}$. To conduct density estimation, the predictive density of the test set with respect to the trained model

$$p(X'|X) = \int \mathrm{d}\tilde{\Psi} p(\tilde{\Psi}|X) p(X'|\tilde{\Psi}) \tag{55}$$

has to be estimated. Replacing the unknown actual posterior $p(\tilde{\Psi}|X)$ with the variational Bayesian one, and using Jensen's inequality, (55) reads

$$\begin{aligned}
\log q(X'|X) =& \log \sum_{S,L} q(S,L) \Bigg[\int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \pi_{s_1} \int \mathrm{d}\boldsymbol{A} q(\boldsymbol{A}) \prod_{t=1}^{T'-1} a_{s_t s_{t+1}} \int \mathrm{d}\boldsymbol{C} q(\boldsymbol{C}) \\
& \times \prod_{t=1}^{T'} c_{s_t l_t} \int \mathrm{d}\boldsymbol{\Theta} q(\boldsymbol{\Theta}) \int \mathrm{d}U q(U) \prod_{t=1}^{T'} p(\boldsymbol{x}_t^{'}|\theta_{s_t l_t}, u_{s_t l_t}) \Bigg] \\
& \approx \sum_{S,L} q(S,L) \Bigg[\int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \log \pi_{s_1} + \int \mathrm{d}\boldsymbol{A} q(\boldsymbol{A}) \sum_{t=1}^{T'-1} \log a_{s_t s_{t+1}} \\
& + \int \mathrm{d}\boldsymbol{C} q(\boldsymbol{C}) \sum_{t=1}^{T'} \log c_{s_t l_t} + \int \mathrm{d}\boldsymbol{\Theta} q(\boldsymbol{\Theta}) \int \mathrm{d}U q(U) \sum_{t=1}^{T'} \log p(\boldsymbol{x}_t^{'}|\theta_{s_t l_t}, u_{s_t l_t}) \Bigg]
\end{aligned} \tag{56}$$

Therefore, the log predictive density for the test sequence X' can be approximated by

$$\log q(X'|X) \approx \mathrm{Pred}(X') \tag{57}$$

where

$$\mathrm{Pred}(X') = \sum_{S,L} q(S,L) \log \frac{\pi_{s_1}^* \prod_{t=1}^{T'-1} a_{s_t s_{t+1}}^* \prod_{t=1}^{T'} c_{s_t l_t}^* p^*(\boldsymbol{x}_t^{'}|\theta_{s_t l_t}, u_{s_t l_t})}{q(S,L)} \tag{58}$$

It is apparent that, similar to estimation of the $q(S,L)$, computation $\mathrm{Pred}(X')$ consists in merely employing the forward-backward algorithm, as described in [2], using the set of the posterior expected values Ψ^* as the optimized values ("point"-estimates) of the model parameters. Note that exactly the same procedure would be employed to obtain the likelihood of a sequence with respect to an SHMM trained using the EM algorithm [4].

4 Experimental Evaluation

In this section, we provide a thorough experimental evaluation of the VB-SHMM algorithm, in a series of sequential data modeling applications from diverse domains. Our experiments have been developed in Matlab R2008a, and were executed on a Macintosh platform with an Intel Core 2 Duo 2 GHz CPU, and 2 GB RAM, running Mac OS X 10.5.

4.1 Synthetic data

We begin with a toy example on simulated data, to demonstrate the notion behind the use of the Student's-t observations distribution in HMM-based sequential data modeling, and the advantages of variational-Bayes over expectation-maximization. The considered synthetic data was obtained by generating five realizations of 400 samples each. In each realization, we drew 100 data points from each one of the bivariate Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, and $\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$, where the parameters $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^3$ are given by

$$\boldsymbol{\mu}_1 = [-6\ 1.5],\ \boldsymbol{\mu}_2 = [0\ 0],\ \boldsymbol{\mu}_3 = [6\ 1.5]$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 5\ 4 \\ \\ 4\ 5 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 5 & -4 \\ \\ -4 & 5 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 1.56 & 0 \\ \\ 0 & 1.56 \end{bmatrix}$$

The resulting 300 simulated points were further augmented by 100 samples drawn from a uniform distribution on the interval $[-10, 10]$ (outliers). These data are used to formulate a time series of multivariate observations, where the observations at time instances $t = 1 : 100$ come from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, the observations at time instances $t = 101 : 200$ come from $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, the observations at time instances $t = 201 : 300$ come from $\mathcal{N}(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$, and the last 100 observations are the outliers.

Using the obtained data, we train one GHMM and one SHMM under both the EM and the VB paradigm. The trained models comprise $N = 3$ states with $K = 1$ mixture components per state. We emphasize that, in our ex-

Table 1
Simulated Experiment: Obtained Gaussian and Student's-t models using the EM algorithm.

Model	Gaussian	Student's-t
State 1 output mean	$\boldsymbol{\mu} = [-6.07\ \ -2.91]$	$\boldsymbol{\mu} = [-5.91\ 1.40]$
State 2 output mean	$\boldsymbol{\mu} = [-3.21\ 1.78]$	$\boldsymbol{\mu} = [0.23\ \ -0.18]$
State 3 output mean	$\boldsymbol{\mu} = [5.11\ 0.74]$	$\boldsymbol{\mu} = [6.02\ 1.69]$

Table 2
Simulated Experiment: Obtained Gaussian and Student's-t models using the VB algorithm.

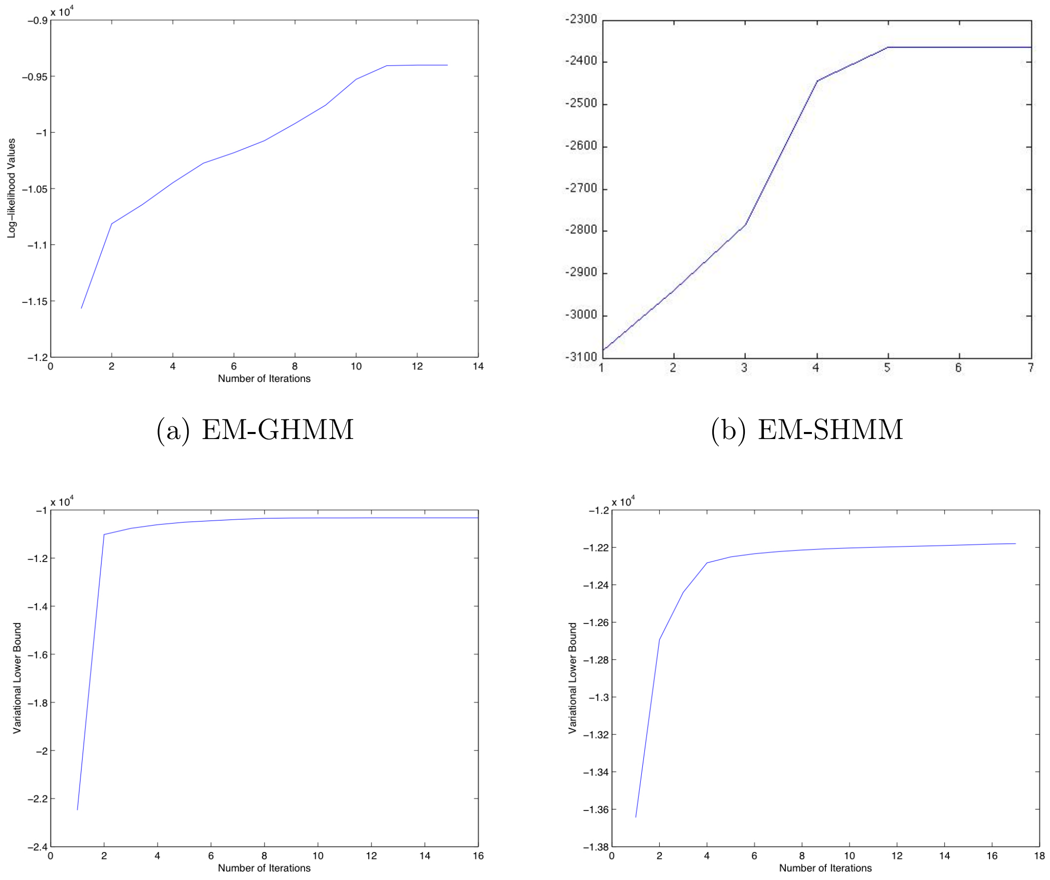
Model	Gaussian	Student's-t
State 1 output mean	$\tilde{\boldsymbol{m}} = [-7.94\ 3.33]$	$\tilde{\boldsymbol{m}} = [-5.96\ 1.49]$
State 2 output mean	$\tilde{\boldsymbol{m}} = [-2.45\ 0.83]$	$\tilde{\boldsymbol{m}} = [0.07\ \ -0.04]$
State 3 output mean	$\tilde{\boldsymbol{m}} = [6.32\ 1.75]$	$\tilde{\boldsymbol{m}} = [6.01\ 1.53]$

periment, we employ a common initialization scheme for the parameters of the EM-trained models (obtained by random selection of the Markov chain probabilities and a run of the k-means algorithm), as well as the posterior hyperparameters of the VB-trained models. This way, the comparison of the evaluated algorithms is just.

In Tables 1 and 2, we provide the point estimates of the GHMM and SHMM means obtained using the EM algorithm, and the joint mean-precision posterior hyperparameter means of the GHMM and SHMM models trained using the VB approach. As we observe, the VB algorithm obtains a much better estimate of the actual distributions of the modeled set of observations. Indeed, when comparing the EM-trained models with the corresponding VB-trained models, we notice that the VB algorithm completely outperforms EM, which yielded a rather poor result. Furthermore, we emphasize that the proposed VB inference algorithm for the SHMM yielded a nearly perfect result, with the estimated mean-precision posterior hyperparameter means being almost identical to the actual means of the modeled data distributions.

Table 3
Simulated Experiment: Execution times of the evaluated algorithms.

Model	Time (in Sec.)
VB-SHMM	26.08
VB-GHMM	25.39
EM-SHMM	8.22
EM-GHMM	8.08



(a) EM-GHMM

(b) EM-SHMM

(c) VB-GHMM

(d) VB-SHMM

Figure 2. Simulated Experiment: Convergence of the evaluated algorithms

In Fig. 2, we illustrate the convergence rates of the evaluated algorithms. It is evident that all the considered algorithms converge comparably fast, in a monotonic fashion, as theoretically expected. Finally, in Table 3, we provide a comparison of the computational requirements of the considered algorithms. It is apparent that training the considered models under the VB framework is relatively more expensive in terms of computational requirements compared to the EM algorithm. We note, however, that in both cases, the computation time needed is of the same order of magnitude.

4.2 Text-dependent speaker identification under the presence of noise

Here, we evaluate our method in classification of noisy sequential data. We consider a text-dependent speaker identification task, using the Japanese Vowels Data Set [29] from the UCI machine learning repository [30]. In this data set, the pass-phrase used for speaker identification purposes comprises two Japanese vowels, /ae/, successively uttered by nine male speakers. For each utterance, a 12-degree linear prediction analysis is applied to obtain a discrete-time series with 12 LPC cepstrum coefficients. Further, given the considerably low noise level in this data set, to make the task of the evaluated models more challenging, Gaussian noise is added to the available data. The added noise has zero mean and diagonal covariance matrix, with the ith element defined as $\sigma_{ii} = \alpha\cdot[\max(x_i^n)\text{-}\min(x_i^n)]$, where x_i^n is the ith dimension of the nth training vector and α is a noise variance factor.

There are 640 time series in total in our data set. We use a set of 270 time series for training (30 for each speaker) and the rest 370 time series for testing. For each one of the speakers, we train one Student's-t HMM and one Gaussian HMM as the classifier, using either of the considered approaches (variational Bayes and expectation-maximization). In detail, for each speaker, first, we run the VB-SHMM and the VB-GHMM algorithms to conveniently and dependably obtain the optimal model size (number of states, N, and number of component distributions, K) of the Student's-t and Gaussian hidden Markov models. The criterion for this selection is the maximization of the value of the lower bound of the model log evidence; this is a dependable criterion for model order selection under a Bayesian setting, as described in Section 3.3. Further,

Table 4
Text-dependent speaker identification: Recognition error rates for optimal number of states and mixture components (N, K are average values for all classes and α is the noise variance factor). Maximum allowed model size (i.e., number of states × number of mixture components) was set equal to 200

Model	VB-SHMM			VB-GHMM			EM-SHMM	EM-GHMM
α	Error rate	N	K	Error rate	N	K	Error rate	Error rate
0	0.03	13.66	4.33	0.03	14	4.77	0.02	0.02
0.05	0.12	17.33	4.77	0.14	19.33	4.77	0.13	0.13
0.10	0.24	17.33	4.44	0.48	16.44	4.44	0.5	0.64
0.15	0.36	16.22	4.22	0.43	8.33	3.22	0.48	0.52
0.20	0.37	14.77	4	0.45	8.33	3	0.52	0.68

Table 5
Text-dependent speaker identification: Recognition error rates for maximum number of states $N = 5$ (N, K are average values for all classes and α is the noise factor)

MAX N=5	VB-SHMM			VB-GHMM			EM-SHMM	EM-GHMM
α	Error rate	N	K	Error rate	N	K	Error rate	Error rate
0	0.04	4.77	5	0.04	4.66	4.77	0.02	0.02
0.05	0.18	4.77	4.88	0.20	3.22	2.22	0.2	0.23
0.10	0.21	4.88	4.55	0.29	3	2.11	0.31	0.41
0.15	0.32	4.77	4.88	0.51	3.77	3.55	0.53	0.71
0.20	0.46	4.55	4.77	0.57	3.66	4	0.61	0.65

we use the trained VB-SHMM and VB-GHMM models to classify our test set, thus obtaining the error rates of the models. Finally, to examine the advantages of the variational Bayesian approach over maximum-likelihood, we also train GHMM and SHMM models using the EM algorithm, with the model size selected equal to the “optimal” model size as determined by the VB algorithm for the corresponding models. The obtained results are illustrated in Table 4.

Further, we repeat the above experiment by limiting the maximum number of model states N. This experimental setting allows for the comparison of Gaussian and Student’s-t models with approximately equal demands in com-

Table 6
Text-dependent speaker identification: Recognition error rates for maximum number of states $N = 10$ (N, K are average values for all classes and α is the noise factor)

MAX N=10	VB-SHMM			VB-GHMM			EM-SHMM	EM-GHMM
α	Error rate	N	K	Error rate	N	K	Error rate	Error rate
0	0.04	8.88	4.77	0.03	8.44	5	0.02	0.02
0.05	0.16	9.88	4.88	0.21	8.22	4.11	0.21	0.32
0.10	0.28	9.66	5	0.29	2.55	2.77	0.3	0.41
0.15	0.35	9.66	5	0.45	5.22	3.77	0.49	0.57
0.20	0.45	9.22	4.77	0.54	7.22	3.55	0.61	0.65

Table 7
Text-dependent speaker identification: Recognition error rates for maximum number of states $N = 15$ (N, K are average values for all classes and α is the noise factor)

MAX N=15	VB-SHMM			VB-GHMM			EM-SHMM	EM-GHMM
α	Error rate	N	K	Error rate	N	K	Error rate	Error rate
0	0.03	11.2	4.77	0.03	12.88	4.88	0.02	0.02
0.05	0.12	14	4.77	0.13	14.77	5	0.11	0.11
0.10	0.29	14.22	4.55	0.36	10.44	3.88	0.38	0.50
0.15	0.34	12.55	4.77	0.37	6.88	3.11	0.41	0.59
0.20	0.35	12.22	4.66	0.51	7.66	3.77	0.55	0.56

Table 8
Text-dependent speaker identification: Mean execution times of the evaluated algorithms for optimal model sizes

Model	Time (in Sec.)
EM-SHMM	21.13
EM-GHMM	20.07
VB-SHMM	307.60
VB-GHMM	302.01

putational resources for their application, while also providing a relative flexibility in choosing a proper model size, within the set limit, which describes the training data in the best way. The results for MAX $N = 5, 10, 15$ are given in Tables 5, 6, and 7, correspondingly. In the same tables, apart from the VB-trained models, we also present for reference purposes the results obtained by the EM-trained models.

From the provided results, it is obvious that the VB-SHMM constantly outperforms the VB-GHMM, while both models perform much better than the corresponding EM-trained models, especially for higher noise variances α. For all methods, as the number N increases from $N = 5$ to $N = 15$, the performance is also enhanced, while letting free the values of N, the effect of overfitting becomes evident. This is especially true for the EM-trained models, as theoretically expected; notice, however, that a similar phenomenon was also witnessed for the VB algorithms, to a much lesser extent whatsoever. It is also obvious that, as the noise increases, the performance of the VB-GHMM is significantly affected. This is not the case for the VB-SHMM, which maintains its representation capabilities and appears to have higher tolerance to noise. Finally, in Table 8 we provide the mean algorithm execution times for optimal model sizes.

4.3 Semantic Characterization of Audio Scenes Based on Content

The significance of audio content in the semantic characterization of multimedia has recently motivated the development of various techniques for content-based scene classification in audio signals. Audio streams, in general, contain a lot of artifacts and outliers, that cannot be easily eliminated by a potential model training sample. Furthermore, to allow for the effective semantic classification of audio data, usually a large number of audio features has to be extracted, thus increasing significantly the dimension of the formulated feature space over which classification or categorization algorithms are carried out. These open issues motivate the application of the VB-SHMM in audio scene classification based on content.

The data set used to carry out our tests consists 487, 20 min. audio samples extracted by several movie genres. Each sample has been divided into semantically coherent audio segments (scenes), and a groundtruth semantic classification has been assigned to each one of these scenes by human experts; we consider 5 semantic classes of audio scene content: music, speech, gunshots, fights, and screams. Each audio scene is represented by a 8-dimensional feature vector comprising each segment's

- spectral rolloff median (SRM);
- zero crossing rate (ZCR), measuring the number of time-domain zero crossings, divided by the frame's length;
- two spectrogram features, the standard deviation and the maximum value of the means obtained over the spectrogram windows;
- a chroma feature [33], expressing the deviation between the obtained chroma coefficients of each segment;
- spectral rolloff [32];
- energy entropy; and
- pitch.

Our experimental setup is the following: We divide our data set into two subsets, one comprising 50 sequences which we use as our training set, and one comprising 437 sequences, used as our test set. Similar to the previous experiment, we contaminate the original data sets with additive Gaussian noise of various variances, in order to make the sequence segmentation task even more demanding for the evaluated models.

We train one EM-GHMM, one VB-GHMM, one EM-SHMM, and one VB-SHMM comprising 5 states, to model the 5 considered semantic classes of audio content. Afterwards, each one of the available test sequences is presented to the trained CHMMs, and the Viterbi algorithm is employed to attribute each sample segment to the "most probable" model state. This way, the semantic

Table 9
Semantic Characterization of Audio Scenes Based on Content: Sequence segmentation error rates obtained by the evaluated algorithms for $\alpha = 0$.

Model	EM-GHMM	EM-SHMM	VB-GHMM	VB-SHMM
$K = 5$	0.5249	0.5172	0.5295	0.4025
$K = 6$	0.5497	0.5144	0.5236	0.409
$K = 7$	0.5249	0.5190	0.5263	0.4152
$K = 8$	0.5428	0.5213	0.5195	0.3856
$K = 9$	0.5299	0.5195	0.5274	0.4251
$K = 10$	0.5684	0.5199	0.5249	0.3992

classifications of the available test segments, as determined by each evaluated model, are obtained. Finally, these results are compared to the groundtruth sequence segmentations, and the error rates of the considered methods are calculated. To ensure the objective comparison of the evaluated methods, our experiment is repeated 50 times, each time using different random starts for the evaluated model training algorithms.

The obtained results for a noise variance factor α equal to 0, 0.05 and 0.10 are given in Tables 9, 10, and 11, respectively. As we observe, the variational methods clearly outperform the point estimators especially when the noise levels become higher. Additionally, the proposed variational method based on the Student's-t distribution is the ultimate winner in all the considered cases. This verifies again the merits of our approach. Finally, in Table 12 we provide the mean algorithm execution times for optimal model sizes.

4.4 Detection of robotic task execution failures

In this experiment, we verify the applicability of our methodology using data from the field of robotics. More specifically, we use five datasets to classify types of failures during the execution of robotic tasks using sequential data from a force/torque sensor. We consider benchmark datasets from the UCI

Table 10
Semantic Characterization of Audio Scenes Based on Content: Sequence segmentation error rates obtained by the evaluated algorithms for $\alpha = 0.05$.

Model	EM-GHMM	EM-SHMM	VB-GHMM	VB-SHMM
$K = 5$	0.7565	0.7217	0.6110	0.4987
$K = 6$	0.7098	0.6966	0.5936	0.4810
$K = 7$	0.7172	0.7135	0.6169	0.5098
$K = 8$	0.7140	0.6902	0.6110	0.4983
$K = 9$	0.6531	0.6700	0.6064	0.4980
$K = 10$	0.7254	0.7080	0.6041	0.4992

Table 11
Semantic Characterization of Audio Scenes Based on Content: Sequence segmentation error rates obtained by the evaluated algorithms for $\alpha = 0.1$.

Model	EM-GHMM	EM-SHMM	VB-GHMM	VB-SHMM
$K = 5$	0.7597	0.7355	0.6604	0.5328
$K = 6$	0.7455	0.7368	0.6499	0.5233
$K = 7$	0.7597	0.7469	0.6581	0.5273
$K = 8$	0.7245	0.7240	0.6577	0.5289
$K = 9$	0.6929	0.7053	0.6458	0.5186
$K = 10$	0.7629	0.7299	0.6371	0.5100

Table 12
Semantic Characterization of Audio Scenes Based on Content: Mean execution times of the evaluated algorithms for optimal model sizes

Model	Time (in Sec.)
VB-SHMM	21.13
VB-GHMM	20.07
EM-SHMM	8.69
EM-GHMM	3.25

repository [30], previously used in [34] in a similar experimental context. Each of the five datasets define a different learning problem; the considered problems comprise:

(1) failures in approach to grasp position (normal, collision, front collision, obstruction);

(2) failures in transfer of a part (normal, front collision, back collision, collision to the right, collision to the left);

(3) position of part after a transfer failure (normal, slightly moved, moved, lost);

(4) failures in approach to ungrasp position (normal, collision, obstruction);

(5) failures in motion with part (normal, bottom collision, bottom obstruction, collision in part, collision in tool).

Each feature in the considered datasets represents a force or a torque measured after failure detection; each failure instance is characterized in terms of 15 force/torque samples collected at regular time intervals starting immediately after failure detection. The total observation window for each failure instance was of 315 ms, and each sequence contains six feature vectors.

The five datasets contain 88, 47, 47, 117 and 164 samples correspondingly [30]. From each dataset we have used the first five samples for training and the rest for testing, except for dataset 2, where only the first three samples were used, due to the limited availability of samples. To model each one of the described classes, HMMs with two states and a single-component distribution were fitted, using both the commonplace EM approaches, as well as the variational Bayesian approach for Gaussian and Student's-t models.

The overall accuracy results are displayed in Table 13, where the best obtainable performance for each of the aforementioned methods is reported. It is apparent that the VB-SHMM outperforms its rivals. As this experiment makes more than evident, the derivation of posterior distributions over the model parameters instead of point estimators offers clear advantages when only few training samples are available. Additionally, the employment of Student's-t observation models offers a lot in alleviating the effect of outliers on the

Table 13
Detection of robotic task execution failures: total error rates for each of the five classification tasks

	EM-GHMM	EM-SHMM	VB-GHMM	VB-SHMM
dataset 1	43.29	43.29	43.29	8.96
dataset 2	56.25	56.25	50.00	25.00
dataset 3	44.83	27.59	24.14	10.34
dataset 4	48.04	48.04	42.16	36.27
dataset 5	50.36	50.36	51.08	41.73

Table 14
Detection of robotic task execution failures: Mean execution times of the evaluated algorithms for optimal model sizes

Model	Time (in Sec.)
EM-SHMM	0.76
EM-GHMM	0.76
VB-SHMM	1.47
VB-GHMM	1.62

trained models. Finally, in Table 14 we provide the mean algorithm execution times for optimal model sizes. As we observe, in cases of clearly limited availability in training samples, the extra computational burden imposed by the variational Bayesian methodologies is clearly insignificant compared to the enhanced pattern recognition effectiveness they allow for.

5 Discussion

Hidden Markov models are a well-established technique for sequential data modeling and classification. Typically, HMMs with continuous observation distributions employ Gaussian mixture models as their hidden state densities. Nevertheless, this selection might considerably undermine the HMM performance when noise contaminates the training data, due to the well-known intolerance of GMMs to outliers. To mitigate this shortcoming, the replacement of Gaussian mixture models with finite mixture models of the heavy-tailed Student's-t distribution has been recently proposed as a promising solution; the resulting Student's-t HMM has been treated under a maximum-likelihood framework using the EM algorithm [4].

In this paper, we proposed a Bayesian treatment of the SHMM using a variational approximation. The so-obtained, VB-SHMM, provides significant advantages over possible alternative maximum-likelihood-based regards of the SHMM model using the EM algorithm and its variations, as the objective function optimized by the VB-SHMM inference algorithm is bounded from above, contrary to the ill-posed construction of the EM algorithm. This interesting property of the proposed VB-SHMM model makes it a favorable selection in many practical applications, where only limited training data sets, contaminated with outliers and noise, are obtainable.

Our experimental results provide strong evidence towards the efficacy of the VB-SHMM approach, and its increased effectiveness in sequential data modeling and classification applications. As we have shown, the VB-SHMM outperforms EM-based approaches, as well as the VB-GHMM method, in a number of demanding applications entailing modeling of noisy sequential data. We have seen that these strengths of the proposed model become even more apparent when reducing the availability of training data in the same application (see, e.g., sections 4.2 and 4.4); this is totally expectable, since the provision of posterior densities over the model parameters allows for a much better modeling of the hidden dynamics of the modeled datasets when limited training samples are available, compared to the point-estimates the EM algorithm yields [9,10].

An additional merit of our method concerns optimal model size determination for an HMM fitted to a given dataset. ML has a clear difficulty in model selection, due to the unbounded nature of the objective function it employs. On the contrary, our variational Bayesian approach does not suffer from such problems, thus allowing for the effective and efficient conduction of the model selection procedure by merely maximizing the model marginal likelihood with respect to the model size. Model selection under our approach was extensively studied in the experimental section of our paper. Indeed, we employed our approach to obtain the sizes of all trained models, while we also experimented with setting a low upper limit on the desired model sizes (see, e.g., section 4.2), to study how the algorithm performs under such constrained environments. As we noticed, constraining the desired model size below what our algorithm obtains when no constraint is imposed yields clearly underperforming models, thus providing conspicuous indication that our model selection approach does not exhibit any proneness to favoring too big models, as we theoretically expected.

Concerning the computational complexities of the competing algorithms, as we have observed, VB inference requires in general more computational time than the EM algorithm. However, as already mentioned, the VB algorithm is most useful for problems with limited labeled data availability, since, in this case, the point model estimates provided by ML are inadequate for representing the uncertainty associated with the true posterior. Furthermore, when only small training data sets are available, while ML is faster than VB, both are very fast. Hence, for the case of limited labeled data and/or unknown model size, for which VB is most important, the proposed VB-SHMM is the method of choice. If large volumes of labeled data are available, and the desired model size is already known, an ML solution is adequate, and VB may be avoided if desired.

Appendix

We have

$$F(q(\boldsymbol{A})) = \sum_{S} q(S) \sum_{t=1}^{T-1} \left\langle \log a_{s_t s_{t+1}} \right\rangle_{q(\boldsymbol{A})} + \left\langle \log p(\boldsymbol{A}) \right\rangle_{q(\boldsymbol{A})} - \left\langle \log q(\boldsymbol{A}) \right\rangle_{q(\boldsymbol{A})} \quad (59)$$

where

$$\left\langle \log p(\boldsymbol{A}) \right\rangle_{q(\boldsymbol{A})} = \sum_{i=1}^{N} \left\{ \log \mathit{\Gamma} \left(\sum_{j=1}^{N} \phi_{ij}^{A} \right) + \sum_{j=1}^{N} \left[(\phi_{ij}^{A} - 1) \left\langle \log a_{ij} \right\rangle_{q(\boldsymbol{A})} - \log \mathit{\Gamma}(\phi_{ij}^{A}) \right] \right\} \quad (60)$$

$$\left\langle \log q(\boldsymbol{A}) \right\rangle_{q(\boldsymbol{A})} = \sum_{i=1}^{N} \left\{ \log \mathit{\Gamma} \left(\sum_{j=1}^{N} \omega_{ij}^{A} \right) + \sum_{j=1}^{N} \left[(\omega_{ij}^{A} - 1) \left\langle \log a_{ij} \right\rangle_{q(\boldsymbol{A})} - \log \mathit{\Gamma}(\omega_{ij}^{A}) \right] \right\} \quad (61)$$

$$\left\langle \log a_{ij} \right\rangle_{q(\boldsymbol{A})} = \psi(\omega_{ij}^{A}) - \psi \left(\sum_{j=1}^{N} \omega_{ij}^{A} \right) \quad (62)$$

and $\psi()$ is the digamma function. Similar, it holds

$$F(q(\boldsymbol{\pi})) = \sum_{S} q(S) \left\langle \log \pi_{s_1} \right\rangle_{q(\boldsymbol{\pi})} + \left\langle \log p(\boldsymbol{\pi}) \right\rangle_{q(\boldsymbol{\pi})} - \left\langle \log q(\boldsymbol{\pi}) \right\rangle_{q(\boldsymbol{\pi})} \quad (63)$$

where

$$\left\langle \log p(\boldsymbol{\pi}) \right\rangle_{q(\boldsymbol{\pi})} = \log \mathit{\Gamma} \left(\sum_{i=1}^{N} \phi_{i}^{\pi} \right) + \sum_{i=1}^{N} \left[(\phi_{i}^{\pi} - 1) \left\langle \log \pi_{i} \right\rangle_{q(\boldsymbol{\pi})} - \log \mathit{\Gamma}(\phi_{i}^{\pi}) \right] \quad (64)$$

$$\left\langle \log q(\boldsymbol{\pi}) \right\rangle_{q(\boldsymbol{\pi})} = \log \mathit{\Gamma} \left(\sum_{i=1}^{N} \omega_{i}^{\pi} \right) + \sum_{i=1}^{N} \left[(\omega_{i}^{\pi} - 1) \left\langle \log \pi_{i} \right\rangle_{q(\boldsymbol{\pi})} - \log \mathit{\Gamma}(\omega_{i}^{\pi}) \right] \quad (65)$$

$$\left\langle \log \pi_{i} \right\rangle_{q(\boldsymbol{\pi})} = \psi(\omega_{i}^{\pi}) - \psi \left(\sum_{i=1}^{N} \omega_{i}^{\pi} \right) \quad (66)$$

and

$$F(q(\boldsymbol{C})) = \sum_{S,L} q(S,L) \sum_{t=1}^{T} \left\langle \log c_{s_t l_t} \right\rangle_{q(\boldsymbol{C})} + \left\langle \log p(\boldsymbol{C}) \right\rangle_{q(\boldsymbol{C})} - \left\langle \log q(\boldsymbol{C}) \right\rangle_{q(\boldsymbol{C})} \quad (67)$$

where

$$\langle \log p(\boldsymbol{C}) \rangle_{q(\boldsymbol{C})} = \sum_{i=1}^{N} \left\{ \log \mathit{\Gamma} \left(\sum_{j=1}^{K} \phi_{ij}^{C} \right) + \sum_{j=1}^{K} \left[(\phi_{ij}^{C} - 1) \langle \log c_{ij} \rangle_{q(\boldsymbol{C})} - \log \mathit{\Gamma}(\phi_{ij}^{C}) \right] \right\} \tag{68}$$

$$\langle \log q(\boldsymbol{C}) \rangle_{q(\boldsymbol{C})} = \sum_{i=1}^{N} \left\{ \log \mathit{\Gamma} \left(\sum_{j=1}^{K} \omega_{ij}^{C} \right) + \sum_{j=1}^{K} \left[(\omega_{ij}^{C} - 1) \langle \log c_{ij} \rangle_{q(\boldsymbol{C})} - \log \mathit{\Gamma}(\omega_{ij}^{C}) \right] \right\} \tag{69}$$

$$\langle \log c_{ij} \rangle_{q(\boldsymbol{C})} = \psi(\omega_{ij}^{C}) - \psi \left(\sum_{j=1}^{K} \omega_{ij}^{C} \right) \tag{70}$$

Concerning the term $F(q(\boldsymbol{\Theta}))$, we have

$$F(q(\boldsymbol{\Theta})) = \sum_{S,L} q(S,L) \sum_{t=1}^{T} \langle \log p(\boldsymbol{x}_t | \theta_{s_t l_t}, u_{s_t l_t}) \rangle_{q(U), q(\boldsymbol{\Theta})} + \langle \log p(\boldsymbol{\Theta}) \rangle_{q(\boldsymbol{\Theta})} - \langle \log q(\boldsymbol{\Theta}) \rangle_{q(\boldsymbol{\Theta})} \tag{71}$$

where

$$\begin{aligned} \langle \log p(\boldsymbol{x}_t | \theta_{ij}, u_{ijt}) \rangle_{q(U), q(\boldsymbol{\Theta})} = & -\frac{d}{2} \log 2\pi + \frac{1}{2} \langle \log |\boldsymbol{R}_{ij}| \rangle_{q(\boldsymbol{\Theta})} + \frac{d}{2} \langle \log u_{ijt} \rangle_{q(U)} \\ & - \frac{\langle u_{ijt} \rangle_{q(U)}}{2} \left[\left\langle \left(\boldsymbol{x}_t - \boldsymbol{\mu}_{ij} \right)^{\mathrm{T}} \boldsymbol{R}_{ij} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_{ij} \right) \right\rangle_{q(\boldsymbol{\Theta})} \right] \end{aligned} \tag{72}$$

$$\left\langle \left(\boldsymbol{x}_t - \boldsymbol{\mu}_{ij} \right)^{\mathrm{T}} \boldsymbol{R}_{ij} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_{ij} \right) \right\rangle_{q(\boldsymbol{\Theta})} = \frac{d}{\tilde{\lambda_{ij}}} + \tilde{\eta_{ij}} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}_{ij}} \right)^{\mathrm{T}} \tilde{\boldsymbol{S}_{ij}}^{-1} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}_{ij}} \right) \tag{73}$$

$$\langle \log |\boldsymbol{R}_{ij}| \rangle_{q(\boldsymbol{\Theta})} = -\log \left| \frac{\tilde{\boldsymbol{S}_{ij}}}{2} \right| + \sum_{k=1}^{d} \psi \left(\frac{\tilde{\eta_{ij}} + 1 - k}{2} \right) \tag{74}$$

$$\langle u_{ijt} \rangle_{q(U)} = \frac{\alpha_{ij}}{\beta_{ijt}} \tag{75}$$

$$\langle \log u_{ijt} \rangle_{q(U)} = \psi(\alpha_{ij}) - \log \beta_{ijt} \tag{76}$$

$$\langle \log p(\boldsymbol{\Theta}) \rangle_{q(\boldsymbol{\Theta})} = \sum_{i=1}^{N} \sum_{j=1}^{K} \langle \log p(\theta_{ij}) \rangle_{q(\boldsymbol{\Theta})} \tag{77}$$

$$\langle \log q(\boldsymbol{\Theta}) \rangle_{q(\boldsymbol{\Theta})} = \sum_{i=1}^{N} \sum_{j=1}^{K} \langle \log q(\theta_{ij}) \rangle_{q(\boldsymbol{\Theta})} \tag{78}$$

$$\begin{aligned}\left\langle \log p(\theta_{ij})\right\rangle_{q(\boldsymbol{\Theta})} =& -\log Z(\eta_{ij},\boldsymbol{S}_{ij}) - \frac{d}{2}\log 2\pi + \frac{d}{2}\log\lambda_{ij} \\ & - \frac{\tilde{\eta}_{ij}\lambda_{ij}}{2}\left(\tilde{\boldsymbol{m}}_{ij}-\boldsymbol{m}_{ij}\right)^{\mathrm{T}}\tilde{\boldsymbol{S}_{ij}}^{-1}\left(\tilde{\boldsymbol{m}}_{ij}-\boldsymbol{m}_{ij}\right) - \frac{\lambda_{ij}d}{2\tilde{\lambda_{ij}}} \\ & + \frac{\eta_{ij}-d}{2}\left[-\log\left|\frac{\tilde{\boldsymbol{S}_{ij}}}{2}\right| + \sum_{k=1}^{d}\psi\left(\frac{\tilde{\eta_{ij}}+1-k}{2}\right)\right] - \frac{\tilde{\eta}_{ij}}{2}\mathrm{tr}\left[\boldsymbol{S}_{ij}\left(\tilde{\boldsymbol{S}_{ij}}\right)^{-1}\right]\end{aligned} \tag{79}$$

$$\begin{aligned}\left\langle \log q(\theta_{ij})\right\rangle_{q(\boldsymbol{\Theta})} =& -\log Z(\tilde{\eta_{ij}},\tilde{\boldsymbol{S}_{ij}}) - \frac{d}{2}\log 2\pi + \frac{d}{2}\log\tilde{\lambda_{ij}} - \frac{\tilde{\eta}_{ij}d}{2} - \frac{d}{2} \\ & + \frac{\tilde{\eta_{ij}}-d}{2}\left[-\log\left|\frac{\tilde{\boldsymbol{S}_{ij}}}{2}\right| + \sum_{k=1}^{d}\psi\left(\frac{\tilde{\eta_{ij}}+1-k}{2}\right)\right]\end{aligned} \tag{80}$$

$$Z(\eta_{ij},\boldsymbol{S}_{ij}) = \pi^{d(d-1)/4}\left|\frac{\boldsymbol{S}_{ij}}{2}\right|^{-\eta_{ij}/2}\prod_{k=1}^{d}\mathit{\Gamma}\left(\frac{\eta_{ij}+1-k}{2}\right) \tag{81}$$

Finally, regarding the term $F(q(U))$, it holds

$$F(q(U)) = \sum_{i=1}^{N}\sum_{j=1}^{K}\sum_{t=1}^{T}\left[\left\langle \log p(u_{ijt}|\nu_{ij})\right\rangle_{q(u_{ijt})} - \left\langle \log q(u_{ijt})\right\rangle_{q(u_{ijt})} + \gamma_{ijt}^{C}\left\langle \log p(\boldsymbol{x}_t|\theta_{ij},u_{ijt})\right\rangle_{q(\theta_{ij})}\right] \tag{82}$$

where

$$\left\langle \log p(u_{ijt}|\nu_{ij})\right\rangle_{q(u_{ijt})} = \frac{\nu_{ij}}{2}\log\frac{\nu_{ij}}{2} - \log\mathit{\Gamma}\left(\frac{\nu_{ij}}{2}\right) + \left(\frac{\nu_{ij}}{2}-1\right)\left\langle \log u_{ijt}\right\rangle_{q(u_{ijt})} - \frac{\nu_{ij}}{2}\left\langle u_{ijt}\right\rangle_{q(u_{ijt})} \tag{83}$$

$$\left\langle \log q(u_{ijt})\right\rangle_{q(u_{ijt})} = (\alpha_{ij}-1)\left\langle \log u_{ijt}\right\rangle_{q(u_{ijt})} + \alpha_{ij}\log\beta_{ijt} - \beta_{ijt}\left\langle u_{ijt}\right\rangle_{q(u_{ijt})} - \log\mathit{\Gamma}(\alpha_{ij}) \tag{84}$$

$$\left\langle u_{ijt}\right\rangle_{q(u_{ijt})} = \frac{\alpha_{ij}}{\beta_{ijt}} \tag{85}$$

$$\left\langle \log u_{ijt}\right\rangle_{q(u_{ijt})} = \psi(\alpha_{ij}) - \log\beta_{ijt} \tag{86}$$

References

[1] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models.* Springer Series in Statistics, New York, 2005.

[2] L.R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, pp. 245–255, 1989.

[3] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society, B*, vol. 39, no. 1, pp. 1–38, 1977.

[4] S. Chatzis, D. Kosmopoulos, and T. Varvarigou, “Robust sequential data modeling using an outlier tolerant hidden Markov model,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1657-1669, September 2009.

[5] K. Yamazaki and S. Watanabe, “Singularities in mixture models and upper bounds of stochastic complexity,” *Neural Networks*, vol. 16, no. 7, pp. 1029–1038, 2003.

[6] C. Archambeau, J.A. Lee, and M. Verleysen, “On the convergence problems of the EM algorithm for finite Gaussian mixtures,” in *Eleventh European symposium on artificial neural networks*, 2003, pp. 99–106.

[7] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley Series in Probability and Statistics, New York, 2000.

[8] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[9] J. Diebolt and C.P. Robert, “Estimation of finite mixture distributions through Bayesian sampling,” *J. Roy. Statist. Soc. B*, vol. 56, pp. 363–375, 1994.

[10] S. Richardson and P.J. Green, “On Bayesian analysis of mixtures with unknown number of components,” *J. Roy. Statist. Soc. B*, vol. 59, pp. 731–792, 1997.

[11] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, “An introduction to variational methods for graphical models,” in *Learning in Graphical Models*, M.I. Jordan, Ed., pp. 105–162. Kluwer, Dordrecht, 1998.

[12] C.M. Bishop and M.E. Tipping, “Variational relevance vector machines,” in *Proc. 16th Conf. Uncertainty in Artificial Intelligence*, 2000, pp. 46–53.

[13] C. Constantinopoulos and A. Likas, "Unsupervised learning of Gaussian mixtures based on variational component splitting," *IEEE Trans. Neural Networks*, vol. 18, pp. 745–755, 2007.

[14] S.J. Roberts and W.D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Trans. Signal Processing*, vol. 50, pp. 2245–2257, 2002.

[15] V. Smidl and A. Quinn, "Mixture-based extension of the AR model and its recursive Bayesian identification," *IEEE Trans. Signal Processing*, vol. 53, pp. 3530–3542, 2005.

[16] C. Archambeau and M. Verleysen, "Robust Bayesian clustering," *Neural Networks*, vol. 20, pp. 129–138, 2007.

[17] Markus Svensén and Christopher M. Bishop, "Robust Bayesian mixture modelling," *Neurocomputing*, vol. 64, pp. 235–252, 2005.

[18] Z. Ghahramani and M. Beal, "Variational inference for Bayesian mixture of factor analysers," *Advances Neural Information Processing Systems*, vol. 12, pp. 449–455, 1999.

[19] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[20] S. Chatzis, D. Kosmopoulos, and T. Varvarigou, "Signal modeling and classification using a robust latent space model based on t distributions," *IEEE Trans. Signal Processing*, vol. 56, no. 3, pp. 949-963, March 2008.

[21] D. MacKay, "Ensemble learning for hidden Markov models," Tech. Rep., Dept. of Physics, Univ. of Cambridge, 1997.

[22] S. Ji, B. Krishnapuram, and L. Carin, "Variational Bayes for continuous hidden Markov models and its application to active learning," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 522–532, 2006.

[23] I. Rezek and S. J. Roberts, "Ensemble hidden Markov models with extended observation densities for biosignal analysis," in *Probabilistic Modeling*

in Biomedicine and Medical Bioinformatics, Eds Dirk Husmeier, Richard Dybowski, and Stephen Roberts, Eds. Springer Verlag, 2005.

[24] C. Liu and D. Rubin, “ML estimation of the t distribution using EM and its extensions, ECM and ECME,” *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.

[25] H. Attias, “A variational Bayesian framework for graphical models,” in *Proc. Ann. Conf. Neural Information Processing Systems*, 2000.

[26] T. Jaakkola and M.I. Jordan, “Bayesian parameter estimation via variational methods,” *Statistics and Computing*, vol. 10, pp. 25–37, 2000.

[27] J. Winn and C.M. Bishop, “Variational message passing,” *J. Machine Learning Research*, vol. 6, pp. 661–694, 2005.

[28] D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, New York, 1987.

[29] Mineichi Kudo, Jun Toyama, and Masaru Shimbo, “Multidimensional curve classification using passing-through regions,” *Pattern Recognition Letters*, vol. 20, no. 11-13, pp. 1103–1111, 1999.

[30] A. Asuncion and D.J. Newman, “UCI machine learning repository,” 2007. URL: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[31] Tom M. Mitchell, Rebecca Hutchinson, Radu S. Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman, “Learning to decode cognitive states from brain images,” *Machine Learning*, vol. 57, no. 1, pp. 145–175, 2004.

[32] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, “Violence Content Classification Using Audio Features,” *in Proc. Advances in Artificial Intelligence* , pp. 502-507, 2006.

[33] Mark A. Bartsch, and Gregory H. Wakefield, “Audio Thumbnailing of Popular Music Using Chroma-Based Representations,” *IEEE Trans Multimedia*, no. 1, vol. 7, pp. 96-104, 2005.

[34] Luis M. Camarinha-matos and Luis Seabra Lopes and Student Member and Josc Barata, "Integration and Learning in Supervision of Flexible Assembly

Systems," *IEEE Transactions on Robotics and Automation*, no. 12, pp. 202-219, 1996.