Signal Modeling and Classification Using a Robust Latent Space Model based on t Distributions

Sotirios Chatzis, *Member, IEEE*, Dimitrios Kosmopoulos, *Member, IEEE*, Theodora Varvarigou, *Member, IEEE*

Abstract

Factor analysis is a statistical covariance modeling technique based on the assumption of normally distributed data. A mixture of factor analyzers can be hence viewed as a special case of Gaussian (normal) mixture models providing a mathematically sound framework for attribute space dimensionality reduction. A significant shortcoming of mixtures of factor analyzers is the vulnerability of normal distributions to outliers. Recently, the replacement of normal distributions with the heavier-tailed Student's-t distributions has been proposed as a way to mitigate these shortcomings and the treatment of the resulting model under an expectation-maximization (EM) algorithm framework has been conducted. In this paper we develop a Bayesian approach to factor analysis modelling based on Student's-t distributions. We derive a tractable variational inference algorithm for this model by expressing the Student's-t distributed factor analyzers as a marginalization over additional latent variables. Our innovative approach provides an efficient and more robust alternative to EM-based methods, resolving their singularity and overfitting proneness problems, while allowing for the automatic determination of the optimal model size. We demonstrate the superiority of the proposed model over well-known covariance modeling techniques in a wide range of signal processing applications.

I. Introduction

Factor analysis (FA) is a well-established linear latent variable scheme modeling the covariances between the elements of multivariate observations, by dividing them into two parts, an unobserved systematic part, taken as a linear combination of a relatively small number of unobserved latent variables called *factors*, and an unobserved *error* part, whose elements are considered as uncorrelated. Factor analysis is closely related to principal components analysis (PCA) [1], and might be considered as a generalization of both PCA and its probabilistic version, PPCA [1], overcoming their drawbacks which namely are: (a) PCA does not correspond to an underlying density function

for the data, and (b) both PCA and PPCA assume a uniform variation for the components of the feature vectors outside the principal subspace, which in general is a strong and restrictive assumption.

As a single FA model provides only a global linear model for the representation of the data in a lower-dimensional subspace, its applicability is limited. A global nonlinear approach can be obtained by postulating a finite mixture of linear submodels for the distribution of the full observation vector given the unobservable (latent) factor vectors, yielding the so-called mixture of factor anayzers (MFA) model [2]. Conventionally, the *factor* and the *error* vectors in the MFA model are considered to be normally distributed. This way, the MFA model can be viewed a special case of finite Gaussian (normal) mixture models (GMMs), where the covariance matrices of its component densities have a special form [3]. GMMs have been widely used in statistical signal modeling and classification applications. Their popularity stems from their provision of a sound statistical framework for the approximation of unknown, non-Gaussian multimodal distributions.

The MFA model has been considered as a more efficient alternative to conventional GMMs allowing for the reduction of the degree of freedom of the covariance matrices while maintaining the recognition performance [4]. Many treatments of it under a maximum-likelihood (ML) framework using variants of the expectation-maximization (EM) algorithm have been proposed. (e.g. [2], [5]). The MFA model has been applied to a large variety of signal modeling and classification problems. Among its most typical applications, we mention speech recognition [6], text-independent speaker recognition [7], [8], face detection [9]. Nevertheless, MFA model suffers from a significant shortcoming, common to every GMM-based or GMM-related model: the model parameter estimation procedure can be adversely affected by outliers in the training data [3].

The problem of providing protection against outliers in multivariate data is a very difficult problem and increases in difficulty with the dimension of the data [10]. Mixtures of Student's-t distributions (SMMs) have been proposed recently as an alternative to GMMs providing high robustness against outliers [11]. The Student's-t distribution is a bell-shaped distribution with heavier tails and one more parameter (degrees of freedom - DOF) comparing to the normal distribution and tends to a normal distribution for big DOF values. Hence, SMMs provide a much more robust approach to the fitting of GMMs, as observations that are atypical of a component are given reduced weight in the calculation of its parameters [3], [11]–[14].

In this paper, we exploit the outlier tolerance advantages of mixtures of Student's-t distributions in the context of factor analysis by proposing a novel MFA model, where the *factor* and the *error* distributions of each component factor analyzer are considered to be multivariate Student's-t distributions; the proposed model is treated under a *Bayesian* framework using a *variational approximation*, yielding the *Variational Bayes Mixture of Student's-t Factor Analyzers* (VB-MSFA) model. Concurrently with this work, another research group proposed a similar model for robust latent subspace modeling using mixtures of Student's-t factor analyzers, providing an *ML treatment* of it using an *EM variant* [15]. However, the undesirable property of ML of being ill-posed since the likelihood function is unbounded [3], [16], [17] results in several very significant shortcomings. To begin with, the EM algorithm can easily get caught in local maxima, and often many restarts are required before a good maximum is reached. Another 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 covariance matrices [3]. A further central issue is the choice of the number of components of a mixture model: ML methods fail to address this issue since the unbounded nature of the log likelihood function makes them prone to favouring models of ever increasing complexity and, hence, leading to overfitting.

In our work we conduct a *Bayesian treatment* of mixtures of Student's-t factor analyzers, overcoming the problems of ML approaches elegantly, by marginalizing over the model parameters with respect to appropriate priors, and maximizing the resulting marginal likelihood of the model with respect to the number of mixture components to obtain the optimal model size. Our approach is based on *variational approximation* methods [18], which have recently emerged as a deterministic alternative to Markov chain Monte-Carlo (MCMC) algorithms for doing Bayesian inference for finite mixture models [19], [20], with better scalability in terms of computational cost [21]. Variational Bayesian inference has previously been applied to GMMs (e.g. [22]), autoregressive models [23], [24], SMMs [13], [14] and conventional (Gaussian) MFA [25], thereby avoiding the singularity and overfitting problems of ML approaches.

The remainder of this paper is organized as follows: In Section II we begin with a brief presentation of the Student's-t factor analysis model (SFA) and of the mixture of Student's-t factor analyzers (MSFA) model. Subsequently, the Bayesian formulation of the MSFA model is introduced. In Section III, we conduct the Bayesian

treatment of the MSFA model using a variational inference approximation, yielding the *Variational Bayes*-MSFA (VB-MSFA) model. In Section IV, we experimentally demonstrate the advantages of the proposed method and its superiority over competing statistical signal analysis and classification models using a synthetic example and two applications. The last section concludes this paper.

II. Problem Formulation

A. Mixtures of Student's-t Factor Analyzers

Let us denote as $\boldsymbol{x}_1$,...., $\boldsymbol{x}_n$ a random sample of size n at a P-dimensional random vector $\boldsymbol{X}$. Factor analysis models the observed variables $\boldsymbol{x}_j, j = 1, ..., n$, as

$$\boldsymbol{x}_j = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}_j + \boldsymbol{e}_j \tag{1}$$

where $\boldsymbol{y}_j$ is a Q-dimensional ($Q < P$) vector of latent variables called factors, $\boldsymbol{\mu}$ is the mean of the observations $\boldsymbol{x}_j$, $\boldsymbol{\Lambda}$ is a $P \times Q$ matrix of factor loadings (parameters), and $\boldsymbol{e}_j$ is the model error. We assume that $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), ..., (\boldsymbol{x}_n, \boldsymbol{y}_n)$ are independent, identically distributed (i.i.d).

In Student's-t factor analysis [15], the factor vectors $\boldsymbol{y}_j$ are assumed to be i.i.d. following the t distribution as

$$\boldsymbol{y}_j \sim t(\boldsymbol{0}, \boldsymbol{I}_Q, \nu) \tag{2}$$

independently of the errors $\boldsymbol{e}_j$, which are also assumed to be i.i.d. following the t distribution as

$$\boldsymbol{e}_j \sim t(\boldsymbol{0}, \boldsymbol{\Psi}, \nu) \tag{3}$$

where $\boldsymbol{\Psi}$ is a diagonal matrix $\boldsymbol{\Psi} = \mathrm{diag}(\sigma_1^2, .., \sigma_P^2)$ and where $\boldsymbol{I}_Q$ denotes the $Q \times Q$ identity matrix and ν stands for the degrees of freedom of the t distribution. The σ_i^2 are called the uniquenesses. The probability density function (pdf) of a t distribution $t(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with mean $\boldsymbol{\mu}$, inner product matrix $\boldsymbol{\Sigma}$, and ν degrees of freedom, is given by

$$t(\boldsymbol{x}_j|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+P}{2}\right)|\boldsymbol{\Sigma}|^{-1/2}}{(\pi\nu)^{P/2}\Gamma(\nu/2)\{1 + d(\boldsymbol{x}_j, \boldsymbol{\mu}|\boldsymbol{\Sigma})/\nu\}^{(\nu+P)/2}} \tag{4}$$

where $\Gamma(s)$ is the Gamma function and $d(\boldsymbol{x}_j, \boldsymbol{\mu}|\boldsymbol{\Sigma})$ is the squared Mahalanobis distance between $\boldsymbol{x}_j, \boldsymbol{\mu}$ with

covariance matrix $\boldsymbol{\Sigma}$

$$d(\boldsymbol{x}_j, \boldsymbol{\mu}|\boldsymbol{\Sigma}) = (\boldsymbol{x}_j - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}) \tag{5}$$

An alternative definition of the factor vector distributions and of the error vector distributions can be derived considering the properties of the Student's-t distribution; following [26], the Student's-t distribution can be represented as an infinite mixture of scaled Gaussians with the same mean, yielding

$$t(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \int_0^\infty \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\mathcal{G}(u|\nu/2, \nu/2)\mathrm{d}u \tag{6}$$

where $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for a normal distribution and the random variable $u > 0$ follows a Gamma distribution which depends only on the DOF, ν, of the considered Student's-t distribution, i.e.

$$u \sim \mathcal{G}(\nu/2, \nu/2) \tag{7}$$

The pdf of the Gamma distribution, $\mathcal{G}(u|\alpha, \beta)$, is given by

$$\mathcal{G}(u|\alpha, \beta) = u^{\alpha-1}\frac{\beta^\alpha e^{-\beta u}}{\Gamma(a)} \tag{8}$$

Using (6) we yield

$$\boldsymbol{y}_j|u_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_Q/u_j) \tag{9}$$

independently of the errors, and

$$\boldsymbol{e}_j|u_j \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}/u_j) \tag{10}$$

where $u_j \sim \mathcal{G}(\nu/2, \nu/2)$.

From the definition of the factor analysis model (1) and of the distributions of the factor and the error vectors, we derive that conditional on the factors $\boldsymbol{y}_j$, the observations $\boldsymbol{x}_j$ are independently distributed as $t(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}_j, \boldsymbol{\Psi}, \nu)$:

$$\boldsymbol{x}_j|\boldsymbol{y}_j \sim t(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}_j, \boldsymbol{\Psi}, \nu) \tag{11}$$

or, alternatively,

$$\boldsymbol{x}_j|\boldsymbol{y}_j, u_j \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}_j, \boldsymbol{\Psi}/u_j) \tag{12}$$

Integrating out $\boldsymbol{y}_j$, by calculating the convolution of $t(\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{y}_j, \boldsymbol{\Psi}, \nu)$, which is the conditional distribution $\boldsymbol{x}_j|\boldsymbol{y}_j$, and $t(0, \boldsymbol{I}_Q, \nu)$, which is the marginal distribution of $\boldsymbol{y}_j$, we derive that, unconditionally, the observations $\boldsymbol{x}_j$ are i.i.d. according to a t distribution with mean $\boldsymbol{\mu}$ and positive definite inner product matrix equal to $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}$, i.e.,

$$\boldsymbol{x}_j \sim t(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) \tag{13}$$

or, alternatively,

$$\boldsymbol{x}_j|u_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}/u_j) \tag{14}$$

where $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi} \tag{15}$$

Now, let us consider a g-component mixture of Student's-t factor analyzers, with means $\boldsymbol{\mu}_i$, factor loadings matrices $\boldsymbol{\Lambda}_i$, and diagonal error inner product matrices $\boldsymbol{\Psi}_i$, with mixing proportions π_i, $i = 1, ..., g$; that is,

$$\boldsymbol{x}_j = \boldsymbol{\mu}_i + \boldsymbol{\Lambda}_i\boldsymbol{y}_{ij} + \boldsymbol{e}_{ij} \text{ with probability } \pi_i \tag{16}$$

where $\boldsymbol{y}_{ij}$ is the factor vector of the j-th observation given that it comes from the i-th component analyzer

$$\boldsymbol{y}_{ij} \sim t(\boldsymbol{0}, \boldsymbol{I}_Q, \nu_i) \tag{17}$$

and $\boldsymbol{e}_{ij}$ is the error of the model for the j-th observation given that it comes from the i-th component analyzer

$$\boldsymbol{e}_{ij} \sim t(\boldsymbol{0}, \boldsymbol{\Psi}_i, \nu_i) \tag{18}$$

Hence, from (13), (14) we obtain that, unconditionally, the pdf of the observations $\boldsymbol{x}_j$ is given by

$$p(\boldsymbol{x}_j|\{\pi_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i\}_{i=1}^g) = \sum_{i=1}^{g} \pi_i t(\boldsymbol{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \nu_i) \tag{19}$$

or alternatively,

$$p(\boldsymbol{x}_j|\boldsymbol{u}_j;\{\pi_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i\}_{i=1}^g)=\sum_{i=1}^g \pi_i\mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i/u_{ij}) \tag{20}$$

where

$$\boldsymbol{\Sigma}_i=\boldsymbol{\Lambda}_i\boldsymbol{\Lambda}_i^T+\boldsymbol{\Psi}_i \tag{21}$$

$$u_{ij}\sim\mathcal{G}(\nu_i/2,\nu_i/2) \tag{22}$$

and $\boldsymbol{u}_j=(u_{ij})$. Here, u_{ij} is the Gamma distributed scaling variable of the j-th observation given than it comes from the i-th component factor analyzer. Finally, from (11), (12) we yield that conditional on the factor vectors $\boldsymbol{y}_{ij}$ it holds

$$p(\boldsymbol{x}_j|\boldsymbol{Y}_j;\{\pi_i,\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i,\boldsymbol{\Psi}_i,\nu_i\}_{i=1}^g)=\sum_{i=1}^g \pi_i t(\boldsymbol{x}_j|\boldsymbol{\mu}_i+\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij},\boldsymbol{\Psi}_i,\nu_i) \tag{23}$$

$$p(\boldsymbol{x}_j|\boldsymbol{Y}_j,\boldsymbol{u}_j;\{\pi_i,\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i,\boldsymbol{\Psi}_i\}_{i=1}^g)=\sum_{i=1}^g \pi_i\mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_i+\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij},\boldsymbol{\Psi}_i/u_{ij}) \tag{24}$$

where we denote as $\boldsymbol{Y}_j$ the factors matrix $\boldsymbol{Y}_j=(\boldsymbol{y}_{1j}...\boldsymbol{y}_{gj})$.

B. The Variational Bayes Mixture of Student's-t Factor Analyzers

Let us consider a set of observations $X=\{\boldsymbol{x}_j\}_{j=1}^n$ drawn independently from a mixture of Student's-t factor analyzers (MSFA model) with density given by (19). From eq. (19) or (23) it can be observed that an MSFA model has in essence the form of a finite mixture of Student's-t distributions (SMM model). As it has been discussed in [3], there is no closed-form solution for likelihood maximization of an SMM model. However, a tractable solution for a Student's-t mixture, and hence for the MSFA model, can be obtained [14] by exploiting eq. (6), i.e., by considering the alternative expressions (20) and (24) of the MSFA model density, and, hence, by viewing the scaling variables u_{ij} as implicit latent (hidden) variables where a Gamma prior is imposed.

In order to further obtain a tractable variational treatment of the MSFA model, we consider the conditional on the, latent, factor vectors, $\boldsymbol{y}_{ij}$, and scale vectors, $\boldsymbol{u}_j$, expression of the MSFA model density, given by (24), and we re-express it in terms of a marginalization over a set of binary latent variables denoting the label of the component factor analyzer that each one of the observable data $\boldsymbol{x}_j,\ j=1,...,n$ derive from. Let us denote as $Z=\{\boldsymbol{z}_j\}_{j=1}^n$ the

set of label indicator vectors, $\boldsymbol{z}_j = (z_{ij})$, with $z_{ij} \in \{0, 1\}$ and such that $z_{ij} = 1$ if $\boldsymbol{x}_j$ is viewed as generated by the i-th mixture component analyzer, $z_{ij} = 0$ otherwise. We also denote as $U = \{\boldsymbol{u}_j\}_{j=1}^n$ the set of scale vectors, and $Y = \{\boldsymbol{Y}_j\}_{j=1}^n$ the set of factor matrices. Then, for each observation, $\boldsymbol{x}_j,\ j = 1, ..., n$, the corresponding latent variables are the corresponding factors matrix $\boldsymbol{Y}_j$, the scale vector $\boldsymbol{u}_j$, and the label indicator vector $\boldsymbol{z}_j$. Therefore, from (24) and using (17), (22) it follows that, for a fixed number of component analyzers, g, the latent variable model of the MSFA model is specified as follows:

$$p(\boldsymbol{z}_j|\boldsymbol{\pi}) = \prod_{i=1}^{g} \pi_i^{z_{ij}} \tag{25}$$

$$p(\boldsymbol{u}_j|\boldsymbol{z}_j, \boldsymbol{\nu}) = \prod_{i=1}^{g} \mathcal{G}(u_{ij}|\nu_i/2, \nu_i/2)^{z_{ij}} \tag{26}$$

$$p(\boldsymbol{Y}_j|\boldsymbol{u}_j, \boldsymbol{z}_j) = \prod_{i=1}^{g} \mathcal{N}(\boldsymbol{y}_{ij}|\boldsymbol{0}, \boldsymbol{I}_Q/u_{ij})^{z_{ij}} \tag{27}$$

$$p(\boldsymbol{x}_j|\boldsymbol{Y}_j, \boldsymbol{u}_j, \boldsymbol{z}_j; \{\boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i, \boldsymbol{\Psi}_i\}_{i=1}^g) = \prod_{i=1}^{g} \mathcal{N}(\boldsymbol{x}_j|\boldsymbol{\mu}_i + \boldsymbol{\Lambda}_i \boldsymbol{y}_{ij}, \boldsymbol{\Psi}_i/u_{ij})^{z_{ij}} \tag{28}$$

where $\boldsymbol{\pi} = (\pi_i)$, $\boldsymbol{\nu} = (\nu_i)$. To complete the Bayesian formulation of the MSFA model we need to impose appropriate prior distributions over the model parameters. For convenience, we choose priors conjugate to the likelihood terms (25)-(28), as this selection greatly simplifies inference and interpretability [18]. This way, the prior for the mixing proportions vector is chosen to follow a Dirichlet distribution (i.e. conjugate to the multinomial distribution $p(\boldsymbol{z}_j|\boldsymbol{\pi})$, given by (25))

$$p(\boldsymbol{\pi}|\boldsymbol{a}) = \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{a}) = \frac{\Gamma(a_0)}{\prod_{i=1}^{g} \Gamma(a_i)} \prod_{i=1}^{g} \pi_i^{a_i - 1} \tag{29}$$

where $a_0 = \sum_i a_i$. Concerning the entries of the factor loading matrices, $\boldsymbol{\Lambda}_i$, we choose an *hierarchical* conjugate prior distribution, in order to allow for the conduction of *automatic relevance determination (ARD)* [27]. Each column of each factor loading matrix is imposed a Gaussian prior with mean zero and a different precision hyperparameter

$$p(\boldsymbol{\Lambda}_i|\boldsymbol{\phi}_i) = \prod_{k=1}^{Q} \mathcal{N}(\boldsymbol{\lambda}_{ik}|\boldsymbol{0}, \boldsymbol{I}_P/\phi_{ik}) \tag{30}$$

where $\boldsymbol{\lambda}_{ik}$ denotes the k-th column of $\boldsymbol{\Lambda}_i$ and ϕ_{ik} is the same precision parameter for each entry in the corresponding column. Since the number of hyperparameters in the factor loading matrix precision vector, $\boldsymbol{\phi}_i = (\phi_{ik})$, increases with the dimensionality of the i-th analyzer, we choose to impose a Gamma distributed hyper-prior on every element of each precision vector $\boldsymbol{\phi}_i$; that is

$$p(\boldsymbol{\phi}_i|\gamma_0, \omega_0) = \prod_{k=1}^{Q} \mathcal{G}(\phi_{ik}|\gamma_0, \omega_0) \tag{31}$$

where γ_0 and ω_0 are the shape and inverse-scale hyper-hyperparameters of the Gamma prior imposed on each precision vector $\boldsymbol{\phi}_i$. We note that, since the spherical Gaussian prior (30) is separable into each of its P dimensions, it can equivalently be expressed as a Gaussian with axis-aligned elliptical covariance on each row of each analyzer, i.e.

$$p(\boldsymbol{\Lambda}_i|\boldsymbol{\phi}_i) = \prod_{l=1}^{P} \mathcal{N}(\boldsymbol{\lambda}_{il}|\mathbf{0}, \mathrm{diag}(\boldsymbol{\phi}_i)^{-1}) \tag{32}$$

where $\boldsymbol{\lambda}_{il}$ denotes the l-th row of $\boldsymbol{\Lambda}_i$. We might also notice that marginalizing the priors of $\boldsymbol{\Lambda}_i$ over $\boldsymbol{\phi}_i$, using (30) and (31), the obtained marginal priors of the factor loading matrices, $\boldsymbol{\Lambda}_i$, are also Student's-t distributed. In Section III.D we will show how the ARD method [27] provides an exponentially tractable solution to the factor analyzer dimensionality inference problem. Finally, the conjugate priors for the means of the factor analyzers are choosen to be

$$p(\boldsymbol{\mu}_i|\boldsymbol{m}_0, \boldsymbol{s}_0) = \mathcal{N}(\boldsymbol{\mu}_i|\boldsymbol{m}_0, \mathrm{diag}(\boldsymbol{s}_0)^{-1}) \tag{33}$$

We note that, as an aside, we do not impose a prior over the number of component analyzers, g. This will be induced instead by maximization of the variational bound of the MSFA model marginal likelihood, as will be discussed in Section III.D. We have also not placed priors on the factor space dimensionality, Q, of each factor analyzer, since this will be controlled by means of application of the ARD method, as we have already mentioned. For computational convenience no prior distribution is imposed over the error precision matrices, $\boldsymbol{\Psi}_i$, of the component analyzers. Finally, no conjugate prior exists for the component analyzer degrees of freedom vector, $\boldsymbol{\nu}$. The latter two MSFA model parameters will be estimated as model hyperparameters, by optimization as a part of the variational inference procedure discussed next.

III. VARIATIONAL BAYESIAN INFERENCE

Having introduced prior distributions over the MSFA model parameters, the formulation of the *Variational Bayes*-MSFA (VB-MSFA) model is complete. We can, therefore, 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 us denote as $\boldsymbol{\theta}_S = (\boldsymbol{\pi}, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_1, ..., \boldsymbol{\Lambda}_g, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_g)$ the set of the (stochastic) model parameters, and as $\boldsymbol{\theta}$ the set of all the stochastic variables associated with the VB-MSFA model, that is, the considered model parameters, $\boldsymbol{\theta}_S$, and the assumed latent variables, Y, U, Z, i.e., $\boldsymbol{\theta} = (\boldsymbol{\theta}_s, Y, U, Z)$. The variational Bayesian treatment of the VB-MSFA model is conducted by introducing an arbitrary distribution $q(\boldsymbol{\theta}) = q(Y, U, Z, \boldsymbol{\theta}_S)$ and considering the well-known equality for the log marginal likelihood (log evidence), $\log p(X)$ [21]

$$\log p(X) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{34}$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \log \frac{p(X, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \tag{35}$$

In eq. (34), $\mathrm{KL}(q||p)$ stands for the Kullback-Leibler (KL) divergence between the arbitrary distribution $q(\boldsymbol{\theta})$, which is considered as the (approximate) variational posterior over the model variables, and $p(\boldsymbol{\theta}|X)$ which is the true posterior over the model parameters; it is given by

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|X)}{q(\boldsymbol{\theta})} \mathrm{d}\boldsymbol{\theta} \tag{36}$$

Since the KL divergence is a non-negative quantity, it follows from (36) that $\mathcal{L}(q)$ is a lower bound of the log evidence, i.e.

$$\log p(X) \geq \mathcal{L}(q) \tag{37}$$

and would become exact if $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|X)$. Hence, maximizing the lower bound of the log evidence, $\mathcal{L}(q)$, so that it becomes as tight as possible, i.e. minimizing the KL divergence between the true and the variational posterior, a good variational approximation of the VB-MSFA model can be obtained.

In order to yield a tractable expression for the lower bound of the VB-MSFA model log evidence, we assume that the joint variational posterior of the stochastic variables associated with the VB-MSFA model, $q(\boldsymbol{\theta}) = q(Y, U, Z, \boldsymbol{\theta}_S)$, factorizes over the latent variables and the model parameters, i.e.

$$q(\boldsymbol{\theta}) = q(Y, U, Z, \boldsymbol{\theta}_S) \approx q(Y, U, Z)q(\boldsymbol{\theta}_S) \tag{38}$$

We further assume for convenience that the variational posterior over the model parameters factorizes as

$$q(\boldsymbol{\theta}_S) \approx q(\boldsymbol{\pi}) \prod_{i=1}^{g} q(\boldsymbol{\mu}_i)q(\boldsymbol{\phi}_i)q(\boldsymbol{\Lambda}_i) \tag{39}$$

Therefore, the introduced variational posterior distribution $q(\boldsymbol{\theta})$ is assumed to factorize on the form

$$q(\boldsymbol{\theta}) = q(Y, U, Z, \boldsymbol{\theta}_S) \approx q(Y, U, Z)q(\boldsymbol{\pi}) \prod_{i=1}^{g} q(\boldsymbol{\mu}_i)q(\boldsymbol{\phi}_i)q(\boldsymbol{\Lambda}_i) \tag{40}$$

The factorization of $q(\boldsymbol{\theta})$ on the form (40) is a common approach in variational Bayesian inference and is often referred to as the variational Bayes-expectation maximization (VB-EM) approximation [13].

Having chosen a family of approximating (variational) posterior distributions, we can now search for the optimal member of this family by maximization of the lower bound of the log marginal likelihood (variational lower bound), $\mathcal{L}(q)$, thus increasing the variational lower bound on $\log p(q)$, the exact log marginal likelihood. The organization of the remainder of this section is as follows: In the following subsection, we shall derive the expression of the variational lower bound of the VB-MSFA model, $\mathcal{L}(q)$, defined by (35), under the assumption that the variational posterior distribution $q(\boldsymbol{\theta})$ factorizes on the form (40). In subsection III.B, the expressions of the variational posteriors over the model stochastic variables (parameters and latent variables) shall be extracted by optimization of the variational lower bound. In subsection III.C, the problem of hyperparameter value selection shall be tackled. In subsection III.D, we describe the model size selection procedure, as well as the factor subspace dimension selection procedure in terms of the ARD mechanism. Finally, in subsection III.E, the expression of the VB-MSFA model predictive density shall be derived.

A. Variational Lower Bound

From eq. (35) and under the assumption (40) we obtain the following expression for the variational lower bound, $\mathcal{L}(q)$

$$\begin{aligned}\mathcal{L}(q) = &\int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \log \frac{p(\boldsymbol{\pi}|\boldsymbol{a})}{q(\boldsymbol{\pi})} + \sum_{i=1}^{g} \left\{ \int \mathrm{d}\boldsymbol{\phi}_i q(\boldsymbol{\phi}_i) \left[\log \frac{p(\boldsymbol{\phi}_i|\gamma_0, \omega_0)}{q(\boldsymbol{\phi}_i)} + \int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i) \log \frac{p(\boldsymbol{\Lambda}_i|\boldsymbol{\phi}_i)}{q(\boldsymbol{\Lambda}_i)} \right] \right. \\ &\left. + \int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i) \log \frac{p(\boldsymbol{\mu}_i|\boldsymbol{m}_0, \boldsymbol{s}_0)}{q(\boldsymbol{\mu}_i)} \right\} + \sum_{i=1}^{g} \sum_{j=1}^{n} q(z_{ij}=1) \left\{ \int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \log \frac{p(z_{ij}=1|\boldsymbol{\pi})}{q(z_{ij}=1)} \right. \\ &+ \int \mathrm{d}u_{ij} q(u_{ij}|z_{ij}=1) \left[\log \frac{p(u_{ij})}{q(u_{ij}|z_{ij}=1)} + \int \mathrm{d}\boldsymbol{y}_{ij} q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \log \frac{p(\boldsymbol{y}_{ij}|u_{ij})}{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)} \right. \\ &\left. \left. \left. + \int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i) \int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i) \int \mathrm{d}\boldsymbol{y}_{ij} q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, \boldsymbol{y}_{ij}, u_{ij}, z_{ij}=1) \right] \right\}\end{aligned} \tag{41}$$

The analytical expression of $\mathcal{L}(q)$ is derived in Appendix.

From (35) it can be shown that the lower bound of the log evidence, $\mathcal{L}(q)$, is a non-convex function of the variational posterior distribution [28]. As a consequence, there will in general exist multiple maxima of $\mathcal{L}(q)$, and, hence, the solution obtained from the variational inference procedure will depend on the initialization. This issue can be easily addressed by performing multiple optimizations from different random starts, and retaining the solution yielding the largest value of the variational bound, $\mathcal{L}(q)$. We note that a benefit 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 maximum likelihood approaches [14].

B. Variational Posteriors

The expressions of the variational posteriors over the VB-MSFA model variables are derived by maximizing $\mathcal{L}(q)$ with respect to each one of the factors of $q(\boldsymbol{\theta})$ in turn, holding the others fixed, in an iterative manner [29]. At the end of each iteration, the value of the variational lower bound, $\mathcal{L}(q)$, 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, $q(\boldsymbol{\theta})$, are expected to take the same functional form as the corresponding conditional (prior) distributions comprising $p(X, \boldsymbol{\theta})$ [28]. We also mention that, by construction, the lower bound cannot decrease after the update of a factor of $q(\boldsymbol{\theta})$. Moreover, the iterative, consecutive updating of the interdependent distributions of the considered factors of $q(\boldsymbol{\theta})$ is guaranteed

to monotonically and maximally increase the lower bound $\mathcal{L}(q)$ [25].

Let us denote as $< \chi >_{\xi}$ the mean of the expression χ with respect to the probability density function ξ. The complete derivations of the expressions of the variational posteriors over the VB-MSFA model variables, as well as the expressions of some auxiliary quantities (means) used here can be found in Appendix. We begin with considering the update of the variational posterior over the factor vectors $\boldsymbol{y}_{ij}$, $q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij} = 1)$. From the expression of the lower bound of the log evidence, given by (41), we yield

$$q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij} = 1) = \mathcal{N}(\boldsymbol{y}_{ij}|\bar{\boldsymbol{y}}_{ij}, \boldsymbol{\Sigma}_i^y/u_{ij}) \tag{42}$$

where

$$\boldsymbol{\Sigma}_i^y = \left(\boldsymbol{I}_Q + \left\langle \boldsymbol{\Lambda}_i^T \boldsymbol{\Psi}_i^{-1} \boldsymbol{\Lambda}_i \right\rangle_{q(\boldsymbol{\Lambda}_i)}\right)^{-1} \tag{43}$$

$$\bar{\boldsymbol{y}}_{ij} = \boldsymbol{\Sigma}_i^y \left\langle \boldsymbol{\Lambda}_i \right\rangle_{q(\boldsymbol{\Lambda}_i)}^T \boldsymbol{\Psi}_i^{-1} \left(\boldsymbol{x}_j - \left\langle \boldsymbol{\mu}_i \right\rangle_{q(\boldsymbol{\mu}_i)}\right) \tag{44}$$

Using eq. (41) and (42), we obtain that the variational posterior for the set of scaling variables, $U = \{\boldsymbol{u}_j\}_{j=1}^n$, is given by

$$q(u_{ij}|z_{ij} = 1) = \mathcal{G}(u_{ij}|\alpha_{ij}, \beta_{ij}) \tag{45}$$

where

$$\alpha_{ij} = \frac{\nu_i + P}{2} \tag{46}$$

$$\begin{aligned}\beta_{ij} =& \frac{1}{2}\left\{\nu_i + \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)}^T \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)} - 2\left(\boldsymbol{x}_j - \left\langle \boldsymbol{\mu}_i \right\rangle_{q(\boldsymbol{\mu}_i)}\right)^T \boldsymbol{\Psi}_i^{-1} \left\langle \boldsymbol{\Lambda}_i \right\rangle_{q(\boldsymbol{\Lambda}_i)} \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)} \right. \\ & \left. + \left\langle (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Psi}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \right\rangle_{q(\boldsymbol{\mu}_i)} + \mathrm{tr}\left[\boldsymbol{\Psi}_i^{-1} \mathrm{tr}\left(\left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)} \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)}^T \left\langle \boldsymbol{\Lambda}_i^T \boldsymbol{\Lambda}_i \right\rangle_{q(\boldsymbol{\Lambda}_i)}\right)\right]\right\}\end{aligned} \tag{47}$$

and tr stands for the trace of a matrix.

Concerning the set of the label indicator vectors, $Z = \{\boldsymbol{z}_j\}_{j=1}^n$, optimizing eq. (41) with respect to $q(z_{ij} = 1)$

and using (42) and (45), it follows that

$$q(z_{ij}=1) = \exp\Bigg\{ -\Big\langle \frac{u_{ij}}{2}\mathrm{tr}\left[\boldsymbol{\Psi}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij}\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij}\right)^T\right]\Big\rangle_{q(\boldsymbol{\mu}_i),q(\boldsymbol{\Lambda}_i),q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1),q(u_{ij}|z_{ij}=1)}$$
$$-\frac{1}{2}\log|\boldsymbol{\Psi}_i| + \log\Gamma(\alpha_{ij}) - (\alpha_{ij}-1)\left\langle \log u_{ij}\right\rangle_{q(u_{ij}|z_{ij}=1)} + \beta_{ij}\left\langle u_{ij}\right\rangle_{q(u_{ij}|z_{ij}=1)} - \alpha_{ij}\log(\beta_{ij})$$
$$+\frac{1}{2}\log|\boldsymbol{\Sigma}_i^y| + \frac{P-Q}{2}\left\langle \log u_{ij}\right\rangle_{q(u_{ij}|z_{ij}=1)} + \left\langle \log\boldsymbol{\pi}\right\rangle_{q(\boldsymbol{\pi})}\Bigg\}\exp(\text{const.}) \tag{48}$$

Furthermore, we have to take into account that the variational posterior distribution of the label indicator vectors, $q(\boldsymbol{z}_j)$, has to be normalized for each data point $\boldsymbol{x}_j$ so that it holds $\sum_{i=1}^{g} q(z_{ij}=1)=1$. Under this consideration, the expression of the variational posterior $q(z_{ij}=1)$ eventually becomes

$$q(z_{ij}=1) = \frac{r_{ij}}{\sum_{i=1}^{g} r_{ij}} \tag{49}$$

where

$$r_{ij} \triangleq \exp\Bigg\{ -\Big\langle \frac{u_{ij}}{2}\mathrm{tr}\left[\boldsymbol{\Psi}_i^{-1}\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij}\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij}\right)^T\right]\Big\rangle_{q(\boldsymbol{\mu}_i),q(\boldsymbol{\Lambda}_i),q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1),q(u_{ij}|z_{ij}=1)}$$
$$-\frac{1}{2}\log|\boldsymbol{\Psi}_i| + \log\Gamma(\alpha_{ij}) - (\alpha_{ij}-1)\left\langle \log u_{ij}\right\rangle_{q(u_{ij}|z_{ij}=1)} + \beta_{ij}\left\langle u_{ij}\right\rangle_{q(u_{ij}|z_{ij}=1)} - \alpha_{ij}\log(\beta_{ij}) \tag{50}$$
$$+\frac{1}{2}\log|\boldsymbol{\Sigma}_i^y| + \frac{P-Q}{2}\left\langle \log u_{ij}\right\rangle_{q(u_{ij}|z_{ij}=1)} + \left\langle \log\boldsymbol{\pi}\right\rangle_{q(\boldsymbol{\pi})}\Bigg\}$$

The variational posterior of the mixing proportions vector $\boldsymbol{\pi}$ is given by

$$q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\hat{\boldsymbol{a}}) \tag{51}$$

where,

$$\hat{a}_i = a_i + \sum_{j=1}^{n} q(z_{ij}=1) \tag{52}$$

The update of the variational posterior over the precision parameter for the *k*-th column of the *i*-th factor loading matrix, $q(\phi_{ik})$, can be proved to follow a Gamma distribution as

$$q(\phi_{ik}) = \mathcal{G}(\phi_{ik}|\gamma_{ik},\omega_{ik}) \tag{53}$$

where,

$$\gamma_{ik} = \gamma_0 + \frac{P}{2} \tag{54}$$

$$\omega_{ik} = \omega_0 + \frac{\left\langle \boldsymbol{\lambda}_{ik}^T \boldsymbol{\lambda}_{ik} \right\rangle_{q(\boldsymbol{\lambda}_{ik})}}{2} \tag{55}$$

We note that, this update of the variational posterior over the precision parameters, $q(\phi_{ik})$, comprises the key-step for the ARD mechanism placed over the columns of the factor loading matrices.

Now, let us consider the variational posterior of the factor loading matrices, $q(\boldsymbol{\Lambda}_i)$. Denoting as $(\boldsymbol{M})_{m,n}$ the (m,n) element of a matrix $\boldsymbol{M}$, and as $(\boldsymbol{v})_l$ the *l*-th element of a vector $\boldsymbol{v}$, eq. (41) yields that $q(\boldsymbol{\Lambda}_i)$ factorizes over the rows, $\boldsymbol{\lambda}_{il}$, of the factor loading matrix, $\boldsymbol{\Lambda}_i$, and it holds

$$q(\boldsymbol{\Lambda}_i) = \prod_{l=1}^{P} q(\boldsymbol{\lambda}_{il}) \tag{56}$$

$$q(\boldsymbol{\lambda}_{il}) = \mathcal{N}\left(\boldsymbol{\lambda}_{il} | \boldsymbol{m}_{il}^*, \boldsymbol{S}_{il}^*\right) \tag{57}$$

where

$$\boldsymbol{S}_{il}^* = \left[\left(\boldsymbol{\Psi}_i^{-1}\right)_{ll} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij})} \left\langle \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^T \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} + \left\langle \mathrm{diag}(\boldsymbol{\phi}_i) \right\rangle_{q(\boldsymbol{\phi}_i)} \right]^{-1} \tag{58}$$

$$\boldsymbol{m}_{il}^* = \boldsymbol{S}_{il}^* \left(\boldsymbol{\Psi}_i^{-1}\right)_{ll} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij})} \left[(\boldsymbol{x}_j)_l - (\left\langle \boldsymbol{\mu}_i \right\rangle_{q(\boldsymbol{\mu})})_l \right] \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} \tag{59}$$

and $\boldsymbol{\lambda}_{il}$ denotes the *l*-th row of the *i*-th factor loading matrix $\boldsymbol{\Lambda}_i$.

Finally, the expression of the variational posterior over the factor analyzer means, $q(\boldsymbol{\mu}_i)$, is given by

$$q(\boldsymbol{\mu}_i) = \mathcal{N}\left(\boldsymbol{\mu}_i | \boldsymbol{m}_i, \boldsymbol{S}_i\right) \tag{60}$$

where,

$$\boldsymbol{S}_i = \left[\mathrm{diag}(\boldsymbol{s}_0) + \boldsymbol{\Psi}_i^{-1} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij}|z_{ij}=1)} \right]^{-1} \tag{61}$$

$$\boldsymbol{m}_i = \boldsymbol{S}_i \left[\mathrm{diag}(\boldsymbol{s}_0)\boldsymbol{m}_0 + \boldsymbol{\Psi}_i^{-1} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij}|z_{ij}=1)} \left(\boldsymbol{x}_j - \left\langle \boldsymbol{\Lambda}_i \right\rangle_{q(\boldsymbol{\Lambda}_i)} \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} \right) \right] \tag{62}$$

C. Hyperparameter Selection

After having acquired the expressions of the variational posteriors over the model variables, i.e. the considered latent model variables and the model parameters on which we have imposed a conjugate prior, we also need to determine the values of the model hyperparameters set, i.e. $\{\boldsymbol{a}, \gamma_0, \omega_0, \boldsymbol{m}_0, \boldsymbol{s}_0, \{\boldsymbol{\Psi}_i\}_{i=1}^g, \boldsymbol{\nu}\}$.

We firstly consider the selection of the values of the hyperparameters that comprise model parameters with no conjugate prior imposed on them, i.e., $\boldsymbol{\Psi}_i$, ν_i. We shall estimate their optimal expression by means of variational lower bound optimization. Let us begin with considering the expression of $\boldsymbol{\Psi}_i$. Taking derivatives of $\mathcal{L}(q)$ with respect to $\boldsymbol{\Psi}_i^{-1}$ we yield

$$\boldsymbol{\Psi}_i^{-1} = \mathrm{diag}\left\{ \frac{1}{\sum_{j=1}^n q(z_{ij}=1)} \sum_{j=1}^n q(z_{ij}=1) \left\langle u_{ij} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i - \boldsymbol{\Lambda}_i \boldsymbol{y}_{ij}\right) \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i - \boldsymbol{\Lambda}_i \boldsymbol{y}_{ij}\right)^T \right\rangle_{q(\boldsymbol{\theta})} \right\} \tag{63}$$

where diag stands for the operator which sets off-diagonal terms to zero.

Concerning ν_i, from eq. (41) it is easily deduced that the maximization of the variational lower bound with respect to ν_i is equivalent to the maximization of the expected complete-data log likelihood with respect to ν_i. Then, from the relevant literature (e.g. [15]) we obtain that the update of ν_i is the solution of the equation

$$\log\frac{\nu_i}{2} + 1 - \psi\left(\frac{\nu_i}{2}\right) + \frac{1}{\sum_{j=1}^n q(z_{ij}=1)} \sum_{j=1}^n q(z_{ij}=1) \left(\langle \log u_{ij} \rangle_{q(u_{ij}|z_{ij}=1)} - \langle u_{ij} \rangle_{q(u_{ij}|z_{ij}=1)} \right) = 0 \tag{64}$$

For the rest of the hyperparameters of the VB-MSFA model, instead of determining their optimal expression with respect to the model's variational lower bound, we select instead a set of proper *ad hoc* values. This is preferable due to the fact that the benefit from determining their expressions by optimization of the model log evidence is not significant, when a good *ad hoc* value selection can be conducted; on the contrary, the computational burden required to carry out optimization of the variational lower bound with respect to these hyperparameters is significant, mainly due to the open-form formulas required to be computed (see e.g. [14], [25]).

Usually, the ad hoc values for the hyperparameters of a model treated under a variational Bayes framework are selected such that broad prior distributions are obtained [18]. Therefore, a good selection for the hyperparameters of the prior on $\boldsymbol{\mu}_i$ is $\boldsymbol{m}_0 = 0$, $\boldsymbol{s}_0 = 10^{-3}\boldsymbol{I}_P$, so as to obtain broad distributions. In the same fashion, we also choose broad priors for the shape hyperparameter, γ_0, and the inverse-scale hyperparameter, ω_0, of the factor loading matrix

precision parameters, ϕ_{ik}, which can be obtained by setting $\gamma_0 = \omega_0 = 10^{-3}$. Finally, concerning the prior over the mixing proportions vector, $\boldsymbol{\pi}$, we obtain broad priors for the hyperparameters, $\boldsymbol{a}$, by setting $a_i = 10^{-3}\ \forall i$.

D. Model Size Selection and Factor Subspace Dimensionality Inference Using ARD

Let us begin with the model size selection problem. As we have already explained, we do not impose a prior over the model size parameter, g, i.e. the number of component factor analyzers, but, instead, we estimate its optimal value by maximization of the variational lower bound of the VB-MSFA model's log marginal likelihood. Indeed, the adoption of the proposed Bayesian approach allows the optimal value of mixture components, g, to be obtained by merely running the variational inference procedure for different numbers of component analyzers, g, and selecting the one that yields the biggest value of the variational bound, $\mathcal{L}(q)$, since this approximates the log marginal likelihood for the model. On the contrary, in maximum likelihood approaches, 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 [3].

Concerning the factor subspace dimensionality inference problem, as we have already mentioned, we solve it using automatic relevance determination (ARD) [27]. The notion of ARD is to continually create new components while detecting when a component model starts to overfit. The overfit manifests itself as a *precision hyperparameter posterior* tending to infinity, indicating that a single data value is being modeled by the component. Hence, in the case of the VB-MSFA model, the ARD mechanism can be implemented by imposing a hierarchical prior over the factor loading matrices, to discourage large factor loadings, with the width of each prior being controlled by a *Gamma distributed precision hyperparameter*, ϕ_{ik}, as illustrated in eq. (30), (31). If one of these precisions tends to infinity, $\phi_{ik} \to \infty$, then the outgoing weights ($\boldsymbol{\lambda}_{ik}$ column entries) for the k-th factor in the *i*-th analyzer will have to be very close to zero in order to maintain a high likelihood under this prior, which in turn leads the analyzer to ignore this factor, and, hence, the corresponding direction in latent subspace is effectively 'switched off'.

Under these considerations, the factor subspace dimensionality inference procedure for the VB-MSFA model, on the basis of the applied ARD technique, comprises the initial conduction of the learning procedure with the factor subspace dimension Q set to its maximum value, $Q = P - 1$, and, further, the removal of those of the factors that yield a very large mean (tending to infinity) for the variational posterior of their precision hyperparameter. We note

that, as the redundant factors do not actually model data, their removal must induce a log evidence increase.

E. Estimation of the Predictive Density

Let us consider an already estimated VB-MSFA model. In order to perform density estimation or classification of a test set $X' = \{\boldsymbol{x}_j^{'}\}_{j=1}^{n'}$ with respect to this model, we need to estimate the predictive density

$$p(X'|X) = \frac{p(X', X)}{p(X)} = \int \mathrm{d}\boldsymbol{\theta} p(\boldsymbol{\theta}|X) p(X'|\boldsymbol{\theta}) \tag{65}$$

where X is the training data. Using the variational posterior in place of the actual posterior distribution we yield

$$p(X'|X) \approx \int \mathrm{d}\boldsymbol{\theta} q(\boldsymbol{\theta}) p(X'|\boldsymbol{\theta}) \tag{66}$$

Then, from (40) and denoting $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}_i\}_{i=1}^{g}$ and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_i\}_{i=1}^{g}$, we obtain

$$\begin{aligned}
\log p(X'|X) &\approx \log \textstyle\int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \int \mathrm{d}\boldsymbol{\Lambda} q(\boldsymbol{\Lambda}) \int \mathrm{d}\boldsymbol{\mu} q(\boldsymbol{\mu}) \left[\prod_{j=1}^{n'} \sum_{i=1}^{g} p(z_{ij}=1|\boldsymbol{\pi}) p(\boldsymbol{x}_j^{'}|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, z_{ij}=1) \right] \\
&\geq \textstyle\sum_{j=1}^{n'} \int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \int \mathrm{d}\boldsymbol{\Lambda} q(\boldsymbol{\Lambda}) \int \mathrm{d}\boldsymbol{\mu} q(\boldsymbol{\mu}) \left[\log \sum_{i=1}^{g} p(z_{ij}=1|\boldsymbol{\pi}) p(\boldsymbol{x}_j^{'}|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, z_{ij}=1) \right] \\
&= \textstyle\sum_{j=1}^{n'} \int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \int \mathrm{d}\boldsymbol{\Lambda} q(\boldsymbol{\Lambda}) \int \mathrm{d}\boldsymbol{\mu} q(\boldsymbol{\mu}) \left[\log \sum_{i=1}^{g} q(z_{ij}=1) \frac{p(z_{ij}=1|\boldsymbol{\pi}) p(\boldsymbol{x}_j^{'}|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, z_{ij}=1)}{q(z_{ij}=1)} \right] \\
&\geq \textstyle\sum_{j=1}^{n'} \int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \int \mathrm{d}\boldsymbol{\Lambda} q(\boldsymbol{\Lambda}) \int \mathrm{d}\boldsymbol{\mu} q(\boldsymbol{\mu}) \sum_{i=1}^{g} q(z_{ij}=1) \log \frac{p(z_{ij}=1|\boldsymbol{\pi}) p(\boldsymbol{x}_j^{'}|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, z_{ij}=1)}{q(z_{ij}=1)} \\
\Rightarrow \log p(X'|X) &\geq \sum_{j=1}^{n'} \sum_{i=1}^{g} q(z_{ij}=1) \left\{ \int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi}) \log \frac{p(z_{ij}=1|\boldsymbol{\pi})}{q(z_{ij}=1)} + \int \mathrm{d}u_{ij} q(u_{ij}|z_{ij}=1) \left[\log \frac{p(u_{ij})}{q(u_{ij}|z_{ij}=1)} + \right. \right. \\
&\quad \int \mathrm{d}\boldsymbol{y}_{ij} q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \left(\int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i) \int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i) \log p(\boldsymbol{x}_j^{'}|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, \boldsymbol{y}_{ij}, u_{ij}, z_{ij}=1) \right. \\
&\quad \left. \left. \left. + \log \frac{p(\boldsymbol{y}_{ij}|u_{ij})}{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)} \right) \right] \right\}
\end{aligned} \tag{67}$$

Hence, the predictive density can be lower-bounded by a quantity identical to the lower bound of the log evidence, except for the KL penalty terms on the VB-MSFA model parameters which have been removed, and the use of the test data set instead of the training set. This lower bound can be used as a computationally tractable estimation of the predictive density in real applications. Its analytical expression is provided in Appendix.

IV. APPLICATIONS

In the following we present a synthetic and two real applications to evaluate our method. To demonstrate the superiority of the proposed, VB-MSFA model over competing approaches, we compare its modeling and classification performance with the performance obtained by the VB-MFA model [25], an ML approach to the MSFA model (ML-MSFA), as presented in [15], an ML treatment of the MFA model (ML-MFA), as presented in [2], and the methods MPPCA [1] and tPPCA [12]. We underline that, in all experiments, the time needed by each model for training and testing was of the same order if no cross validation is done. More detailed evaluation was not possible due to specific implementation choices, which could affect the results. However, since our method does not need cross validation we can claim that it outperforms, in terms of computational demands, the methods which have such requirements (i.e. the maximum likelihood-based methods).

A. Robust Automatic Model Size Selection

Let us first illustrate the model size selection procedure using a synthetic data set. Consider a three-component mixture of bivariate Gaussian distributions in equal mixing proportions, $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$ with means and covariance matrices, $\boldsymbol{\mu}_1 = (0, 3)^T, \boldsymbol{\mu}_2 = (3, 0)^T, \boldsymbol{\mu}_3 = (-3, 0)^T$ and

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 0.1 \end{pmatrix}, \boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{pmatrix}$$

respectively. From each one of the mixture component densities we draw 800 data points, resulting in a data set comprising 2400 simulated points. To evaluate the performance of our model under the presence of observation noise and outliers, we further add to the above derived data set 600 noise points (outliers) drawn from a uniform distribution over the range $[-\delta, \delta]$ on each variable; a similar synthetic data set has been used in [3] for the evaluation of SMMs. Using this data set we train a VB-MSFA model and a VB-MFA model with unitary factor space dimension, $Q = 1$, obtaining the number of component analyzers required by the Student's-t model and the Gaussian model to represent these data. We also evaluate the ML-MSFA model and ML-MFA models, as well as the methods MPPCA and tPPCA. For these latter methods, the number of mixture components is determined using the integrated classification likelihood (ICL) criterion [3]. We repeat this experiment for different noise ranges,

Table I
SYNTHETIC DATA MODELING: OBTAINED MODEL SIZES FOR DIFFERENT NOISE RANGES

Model	Model Size (g)			
	$\delta = 0$	$\delta = 5$	$\delta = 10$	$\delta = 20$
VB-MSFA	3	3	3	3
VB-MFA	3	3	5	5
ML-MSFA	3	3	3	4
ML-MFA	3	3	5	7
tPPCA	3	3	3	5
MPPCA	3	3	5	7

$\delta = 0$ (i.e. no outliers), $5, 10, 20$. The obtained model sizes are depicted on Table I. Concerning the Gaussian mixture-based models, we notice that without outliers the competition performs well, but as the noise distribution range increases, the estimated number of component analyzers increases too. On the contrary, the Student's-t based models, prove to be extremely robust to outliers, providing a correct estimation of the model size in each (VB-MSFA) or the majority (ML-MSFA, tPPCA) of the considered cases.

B. Classification of spontaneous electroencephalogram during mental tasks

Classification of electroencephalogram (EEG) signals has been widely applied in Brain-Computer Interfaces [30]. Nevertheless, EEG signals consist of a superposition of a large number of simultaneously active brain sources that are typically distorted by artifacts and even subject to nonstationarity. Outliers and artifacts can strongly distort the classifier performance [30] yielding bad generalization. This fact in conjunction with the high dimensionality of the formulated feature space motivates the application of the VB-MSFA model in this application.

For reference purposes, we report EEG classification results on a portion of the data set presented in [31]. Our data set comprises EEG signals obtained by four subjects performing two different mental tasks during which they had to (a) relax and think of nothing in particular (baseline measurement) (b) solve nontrivial multiplication problems (mental multiplication). The obtained EEG signals consisted of 6 channels, and were recorded for 10 seconds during each task, at a sampling rate of 250 Hz. Initially, each EEG signal was divided into 1/4 sec. windows overlapped by 1/8 sec. Further, the signals in each window were analyzed by separately estimating the coefficients of an order 6 autoregressive (AR) model from each one of their constituent 6 channels. This way, a 36×1 feature vector modeling each signal window was obtained. The coefficients of the estimated order 6 AR models were

Table II
EEG SIGNAL CLASSIFICATION: ERROR RATE PER CLASS (MENTAL TASK) AND ON AVERAGE FOR OPTIMAL MODEL CONFIGURATION

Model	g	Q	Baseline Measurement	Mental Multiplication	Average
VB-MSFA	3	22	1.428%	8.575%	5.001%
VB-MFA	4	29	2.860%	11.420%	7.140%
ML-MSFA	3	18	4.57%	25.64%	15.10%
ML-MFA	4	18	9.15%	32.32%	20.74%
tPPCA	3	18	5.14%	25.68%	15.86%
MPPCA	4	18	8.88%	36.54%	22.71%
Anderson et al. [33]	-	-	-	-	18.7%

computed using the Yule-Walker approach [32].

To conduct our experiment, we divide our data set into a training set comprising the 30% of the available data of each class (i.e. of each mental task), consisting of 200 samples per class, and a test set comprising the rest 70%, consisting of 465 samples per class. To evaluate the considered algorithms, we train one model per class, using each one of these algorithms, and, further, we evaluate the trained models as classifiers, under a maximum a posteriori probability (MAP) classification fashion. In Table II we illustrate the obtained optimal model configuration and classification rate of the VB-MSFA and VB-MFA models, as well as of the ML-MSFA and ML-MFA models and the methods MPPCA and tPPCA. The latter results, obtained by using EM algorithm variants, are means over 30 runs of the EM algorithm with the model configuration selection based on the maximization of the classification performance of the treated models. Finally, we quote the performance obtained by Anderson et al. in [33].

We note that the proposed VB-MSFA model outperforms its Gaussian counterpart in terms of required model size and factor subspace dimensionality, as well as, in terms of the obtained classification performance, both on a per class basis and on average. We also notice that the applied ML approaches yield a significantly lower classification performance, obviously due to the instability of the EM algorithm in conjunction with the relatively small number of available training data. Finally, we underline that the obtained lower values for the factor subspace dimensionality, Q, parameter yielding the optimal classification performance of the considered ML-based approaches are obviously due to the overfitting proneness of the EM algorithm which, in conjunction with the limited number of available training data, was leading to a significant impair in the dependability of the EM-based parameter estimation procedure, and hence, the trained models classification performance, as the factor subspace dimensionality, and, subsequently, the number of parameters under estimation, would exceed some threshold.

Table III
AUDIO SIGNAL CLASSIFICATION: ERROR RATE PER CLASS AND ON AVERAGE FOR OPTIMAL MODEL CONFIGURATION

Model	g	Q	Music	Speech	Screams	Average
VB-MSFA	2	3	5.00%	2.5%	7.50%	5.00%
VB-MFA	2	4	7.50%	5.00%	12.50%	8.33%
ML-MSFA	2	3	7.30%	3.68%	11.27%	7.42%
ML-MFA	2	4	13.73%	8.45%	26.25%	16.14%
tPPCA	2	5	7.32%	3.54%	12.02%	7.62%
MPPCA	2	5	14.01%	8.21%	26.84%	16.35%

C. Audio signal classification

The significance of audio content in the semantic characterization of multimedia, has recently motivated the development of various techniques for content-based audio classification [34]. 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-MSFA model in audio classification based on content.

The data set used to carry out our tests, firstly presented in [34], consists of two hundred, 20 min. audio samples extracted by several movie genres. These samples have been segmented into semantically coherent audio segments (scenes). Each segment is represented by a 4×1 feature vector comprising the segment's spectral rolloff median (SRM), zero crossing rate (ZCR), spectral centroid (SC) and energy entropy. In this application, we focus on three audio content semantic types: music, speech and screams. We divide our data set into two subsets of equal size and use them as our training set and our test set, respectively. The trained models are evaluated as classifiers under a maximum a posteriori (MAP) classification notion. In Table III we illustrate the obtained optimal model configuration and classification rate obtained using the VB-MSFA and VB-MFA models. We also incorporate the results obtained by using the ML-MSFA and ML-MFA models, as well as the methods MPPCA and tPPCA. These latter results, obtained by using EM algorithm variants, are means over 30 runs of the EM algorithm with the model configuration selection based on the maximization of the classification performance of the treated models. As we notice, the proposed model clearly outperforms its competitors.

V. Conclusions

Mixture of factor analyzers (MFA) models are a common subspace modeling technique used in signal processing applications, based on Gaussian mixture models (GMMs). Nevertheless, the sensitivity of these popular GMM-based subspace modeling techniques to outliers and observation noise is a well-known problem that has not been completely tackled. Student's-t factor analysis (SFA) [15], where Student's-t distributions are used to model the *factor* and the *error* vectors, and, hence, the observation density of a factor analyzer, has been proposed recently as a promising solution towards the attenuation of these shortcomings. The so-obtained, mixture of Student's-t factor analyzers (MSFA) model can be, thus, viewed as a natural integration of Student's-t mixture models and factor analysis, allowing for the reduction of the degree of freedom of the covariance matrices while maintaining the recognition performance and being very robust to outliers and observation noise comparing to conventional MFA.

In this paper, we proposed a Bayesian treatment of the MSFA model using a variational approximation. The so-obtained, VB-MSFA model, provides significant advantages comparing to possible alternative maximum likelihood-based regards of the MSFA model (e.g. [15]): the model size and factor subspace dimensionality inference problems can be elegantly addressed in an effective and computationally efficient manner, singularities of the kind associated with maximum likelihood are absent, and surplus components revert to the prior distribution and play no role in the predictive density [14]. We also emphasize that the proposed variational approach imposes only a small computational overhead comparing to ML techniques, since the dominant computational costs in our approach arise from the evaluation and inversion of weighted empirical precision matrices, which is also the dominant cost in maximum likelihood approaches using variants of the EM algorithm.

In the experimental section of this paper we verified the merits of the proposed model by applying it in different signal processing applications from diverse domains. As we have showed, in each and every one of these applications, the proposed VB-MSFA model managed to gracefully outperform its competitors, in terms of statistical signal modeling quality and classification performance, without loss of computational efficiency. The diversity of the nature of the signals considered in the context of the aforementioned applications, indicate the generality of our observations on the superiority of the VB-MSFA model over competing latent subspace modeling techniques.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Stavros Perantonis and Prof. Sergios Theodoridis, as well as the anonymous reviewers for their useful comments.

APPENDIX

Let us firstly provide the complete derivations for the variational posteriors over the VB-MSFA model variables. We begin with the update of the variational posterior over the factor vectors $\boldsymbol{y}_{ij}$. From (41), we have

$$\begin{aligned}
&\frac{\partial \mathcal{L}(q)}{\partial q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)} = 0 \Rightarrow q(z_{ij}=1) \int \mathrm{d}u_{ij} q(u_{ij}|z_{ij}=1) \left\{ \log p(\boldsymbol{y}_{ij}|u_{ij}) - \log q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \right. \\
&\left. + \int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i) \int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i) \log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, \boldsymbol{y}_{ij}, u_{ij}, z_{ij}=1) \right\} + \text{const.} = 0 \Rightarrow \\
&0 = \log p(\boldsymbol{y}_{ij}|u_{ij}) - \log q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) + \int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i) \int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i) \log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, \boldsymbol{y}_{ij}, u_{ij}, z_{ij}=1) \\
&\Rightarrow \log q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \propto -\frac{u_{ij}}{2} \left\{ \boldsymbol{y}_{ij}^T \boldsymbol{y}_{ij} + \left\langle \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i - \boldsymbol{\Lambda}_i \boldsymbol{y}_{ij}\right)^T \boldsymbol{\Psi}_i^{-1} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i - \boldsymbol{\Lambda}_i \boldsymbol{y}_{ij}\right) \right\rangle_{q(\boldsymbol{\mu}_i), q(\boldsymbol{\Lambda}_i)} \right\} \\
&\propto -\frac{u_{ij}}{2} \left\{ \boldsymbol{y}_{ij}^T \left(\boldsymbol{I}_Q + \left\langle \boldsymbol{\Lambda}_i^T \boldsymbol{\Psi}_i^{-1} \boldsymbol{\Lambda}_i \right\rangle_{q(\boldsymbol{\Lambda}_i)} \right) \boldsymbol{y}_{ij} - 2 \boldsymbol{y}_{ij}^T \left\langle \boldsymbol{\Lambda}_i^T \boldsymbol{\Psi}_i^{-1} \left(\boldsymbol{x}_j - \boldsymbol{\mu}_i\right) \right\rangle_{q(\boldsymbol{\mu}_i), q(\boldsymbol{\Lambda}_i)} \right\}
\end{aligned} \tag{68}$$

which implies (42). Concerning the variational posterior over the scaling variables, from (41) we obtain

$$\begin{aligned}
\frac{\partial \mathcal{L}(q)}{\partial q(u_{ij}|z_{ij}=1)} = 0 \Rightarrow \Big\{ &\int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i) \int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i) \int \mathrm{d}\boldsymbol{y}_{ij} q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i, \boldsymbol{\mu}_i, \boldsymbol{\Psi}_i, \nu_i, \boldsymbol{y}_{ij}, u_{ij}, z_{ij}=1) \\
&+ \int \mathrm{d}\boldsymbol{y}_{ij} q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \left[\log p(\boldsymbol{y}_{ij}|u_{ij}) - \log q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \right] \\
&+ \log p(u_{ij}) - \log q(u_{ij}|z_{ij}=1) \Big\} q(z_{ij}=1) + \text{const.} = 0
\end{aligned} \tag{69}$$

Using the result

$$\left\langle \frac{u_{ij}}{2} \left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}_{ij}}\right)^T \left(\boldsymbol{\Sigma}_i^y\right)^{-1} \left(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}_{ij}}\right) \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1)} = \frac{u_{ij}}{2} \mathrm{tr}\left[\left(\boldsymbol{\Sigma}_i^y\right)^{-1} \boldsymbol{\Sigma}_i^y / u_{ij} \right] = \frac{Q}{2} \tag{70}$$

(69) eventually yields

$$\begin{aligned}\log q(u_{ij}|z_{ij}=1) \propto & -\frac{u_{ij}}{2}\left\langle \boldsymbol{y}_{ij}\right\rangle^T_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)}\left\langle \boldsymbol{y}_{ij}\right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} - \frac{u_{ij}}{2}\left\langle (\boldsymbol{x}_j-\boldsymbol{\mu}_i)^T\boldsymbol{\Psi}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)\right\rangle_{q(\boldsymbol{\mu}_i)} \\ & - \frac{u_{ij}}{2}\mathrm{tr}\left[\boldsymbol{\Psi}_i^{-1}\mathrm{tr}\left(\left\langle \boldsymbol{y}_{ij}\right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)}\left\langle \boldsymbol{y}_{ij}\right\rangle^T_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)}\left\langle \boldsymbol{\Lambda}_i^T\boldsymbol{\Lambda}_i\right\rangle_{q(\boldsymbol{\Lambda}_i)}\right)\right] \\ & + u_{ij}\left(\boldsymbol{x}_j - \left\langle\boldsymbol{\mu}_i\right\rangle_{q(\boldsymbol{\mu}_i)}\right)^T\boldsymbol{\Psi}_i^{-1}\left\langle\boldsymbol{\Lambda}_i\right\rangle_{q(\boldsymbol{\Lambda}_i)}\left\langle \boldsymbol{y}_{ij}\right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} \\ & + \frac{P}{2}\log u_{ij} + \left(\frac{\nu_i}{2}-1\right)\log u_{ij} - \frac{\nu_i}{2}u_{ij}\end{aligned} \tag{71}$$

which implies (45). In the same fashion, regarding the posterior over the label indicator vectors, from (41) we have

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial q(z_{ij}=1)} =& 0 \Rightarrow \int \mathrm{d}\boldsymbol{\pi} q(\pi)\log p(z_{ij}=1|\boldsymbol{\pi}) - \log q(z_{ij}=1) + \int \mathrm{d}u_{ij}q(u_{ij}|z_{ij}=1)\left[-\log q(u_{ij}|z_{ij}=1)\right. \\ & + \int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i)\int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i)\int \mathrm{d}\boldsymbol{y}_{ij}q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)\log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i,\nu_i,\boldsymbol{y}_{ij},u_{ij},z_{ij}=1) \\ & \left. - \int \mathrm{d}\boldsymbol{y}_{ij}q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)\log q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)\right] + \text{const.} = 0\end{aligned} \tag{72}$$

which, using (70), yields (48). For the posteriors of the mixing proportions vector $\boldsymbol{\pi}$, eq. (41) yields

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial q(\boldsymbol{\pi})} &= \log p(\boldsymbol{\pi}|\boldsymbol{a}) + \sum_{i=1}^{g}\sum_{j=1}^{n} q(z_{ij}=1)\log p(z_{ij}=1|\boldsymbol{\pi}) - \log q(\boldsymbol{\pi}) + \text{const.} = 0 \Rightarrow \\ \log q(\boldsymbol{\pi}) &\propto \sum_{i=1}^{g}\left[(a_i-1) + \sum_{j=1}^{n} q(z_{ij}=1)\right]\log\pi_i\end{aligned} \tag{73}$$

whence we obtain (51). For the update of $q(\phi_{ik})$, from (41) and using (30) we have

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial q(\phi_{ik})} &= \log p(\phi_{ik}|\gamma_0,\omega_0) + \int \mathrm{d}\boldsymbol{\lambda}_{ik}q(\boldsymbol{\lambda}_{ik})\log p(\boldsymbol{\lambda}_{ik}|\phi_{ik}) - \log q(\phi_{ik}) + \text{const.} = 0 \Rightarrow \\ \log q(\phi_{ik}) &\propto (\gamma_0-1)\log\phi_{ik} - \omega_o\phi_{ik} + \frac{1}{2}\left[P\log\phi_{ik} - \phi_{ik}\left\langle\boldsymbol{\lambda}_{ik}^T\boldsymbol{\lambda}_{ik}\right\rangle_{q(\boldsymbol{\lambda}_{ik})}\right]\end{aligned}$$

which implies that the precision is Gamma distributed as given by (53). Concerning the variational posterior over the factor loading matrices, $q(\boldsymbol{\Lambda}_i)$, from eq. (41) we obtain

$$\begin{aligned}\frac{\partial\mathcal{L}(q)}{\partial q(\boldsymbol{\Lambda}_i)} =& \sum_{j=1}^{n} q(z_{ij}=1)\int \mathrm{d}u_{ij}q(u_{ij}|z_{ij}=1)\int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i)\int \mathrm{d}\boldsymbol{y}_{ij}q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)\log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i,\nu_i, \\ & \boldsymbol{y}_{ij},u_{ij},z_{ij}=1) + \int \mathrm{d}\boldsymbol{\phi}_i q(\boldsymbol{\phi}_i)\log p(\boldsymbol{\Lambda}_i|\boldsymbol{\phi}_i) - \log q(\boldsymbol{\Lambda}_i) + \text{const.} = 0\end{aligned} \tag{74}$$

Using (32) and denoting as $(\boldsymbol{M})_{m,n}$ the (m,n) element of a matrix $\boldsymbol{M}$, as $(\boldsymbol{v})_l$ the l-th element of a vector $\boldsymbol{v}$, and as $\boldsymbol{\lambda}_{il}$ the l-th row of the i-th factor loading matrix, (74) yields

$$\begin{aligned}\log q(\boldsymbol{\Lambda}_i) \propto \sum_{l=1}^{P} \Bigg\{ -\frac{1}{2}\boldsymbol{\lambda}_{il}^T \left(\left(\boldsymbol{\Psi}_i^{-1}\right)_{ll} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij})} \left\langle \boldsymbol{y}_{ij} \boldsymbol{y}_{ij}^T \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} \right) \boldsymbol{\lambda}_{il} \\ -\frac{1}{2}\boldsymbol{\lambda}_{il}^T \left\langle \mathrm{diag}(\boldsymbol{\phi}_i) \right\rangle_{q(\boldsymbol{\phi}_i)} \boldsymbol{\lambda}_{il} + \boldsymbol{\lambda}_{il}^T \left(\boldsymbol{\Psi}_i^{-1}\right)_{ll} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij})} \left[(\boldsymbol{x}_j)_l - (\left\langle \boldsymbol{\mu}_i \right\rangle_{q(\boldsymbol{\mu})})_l \right] \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} \Bigg\}\end{aligned} \tag{75}$$

whence it follows that the variational posterior $q(\boldsymbol{\Lambda}_i)$ is eventually given by (56)-(57). Finally, regarding the expression of the variational posterior over the factor analyzer means, eq. (41) yields

$$\begin{aligned}\frac{\partial \mathcal{L}(q)}{\partial q(\boldsymbol{\mu}_i)} =& \sum_{j=1}^{n} q(z_{ij}=1) \int \mathrm{d}u_{ij} q(u_{ij}|z_{ij}=1) \int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i) \int \mathrm{d}\boldsymbol{y}_{ij} q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1) \log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i,\nu_i, \\ & \boldsymbol{y}_{ij},u_{ij},z_{ij}=1) + \log p(\boldsymbol{\mu}_i|\boldsymbol{m}_0,\boldsymbol{s}_0) - \log q(\boldsymbol{\mu}_i) + \text{const.} = 0 \Rightarrow \\ \log q(\boldsymbol{\mu}_i) \propto & \boldsymbol{\mu}_i^T \left[\mathrm{diag}(\boldsymbol{s}_0)\boldsymbol{m}_0 + \boldsymbol{\Psi}_i^{-1} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij}|z_{ij}=1)} \left(\boldsymbol{x}_j - \left\langle \boldsymbol{\Lambda}_i \right\rangle_{q(\boldsymbol{\Lambda}_i)} \left\langle \boldsymbol{y}_{ij} \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)} \right) \right] \\ & - \frac{1}{2}\boldsymbol{\mu}_i^T \left[\mathrm{diag}(\boldsymbol{s}_0) + \boldsymbol{\Psi}_i^{-1} \sum_{j=1}^{n} q(z_{ij}=1) \left\langle u_{ij} \right\rangle_{q(u_{ij}|z_{ij}=1)} \right] \boldsymbol{\mu}_i\end{aligned} \tag{76}$$

which implies eq. (60). Based on these results, the analytical expression of the lower bound of log evidence, given by (41), can be written as

$$\begin{aligned}\mathcal{L}(q) =& \left\langle \log p(\boldsymbol{\pi}|\boldsymbol{a}) \right\rangle_{q(\boldsymbol{\pi})} - \left\langle \log q(\boldsymbol{\pi}) \right\rangle_{q(\boldsymbol{\pi})} + \sum_{i=1}^{g} \Big\{ \left\langle \log p(\boldsymbol{\phi}_i|\gamma_0,\omega_0) \right\rangle_{q(\boldsymbol{\phi}_i)} - \left\langle \log q(\boldsymbol{\phi}_i) \right\rangle_{q(\boldsymbol{\phi}_i)} + \left\langle \log p(\boldsymbol{\Lambda}_i|\boldsymbol{\phi}_i) \right\rangle_{q(\boldsymbol{\Lambda}_i),q(\boldsymbol{\phi}_i)} \\ & - \left\langle \log q(\boldsymbol{\Lambda}_i) \right\rangle_{q(\boldsymbol{\Lambda}_i)} + \left\langle \log p(\boldsymbol{\mu}_i|\boldsymbol{m}_0,\boldsymbol{s}_0) \right\rangle_{q(\boldsymbol{\mu}_i)} - \left\langle \log q(\boldsymbol{\mu}_i) \right\rangle_{q(\boldsymbol{\mu}_i)} + \sum_{j=1}^{n} q(z_{ij}=1) \Big[\left\langle \log p(z_{ij}=1|\boldsymbol{\pi}) \right\rangle_{q(\boldsymbol{\pi})} \\ & - \log q(z_{ij}=1) + \left\langle \log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i,\nu_i,\boldsymbol{y}_{ij},u_{ij},z_{ij}=1) \right\rangle_{q(\boldsymbol{\theta})} + \left\langle \log p(u_{ij}) \right\rangle_{q(u_{ij}|z_{ij}=1)} \\ & - \left\langle \log q(u_{ij}|z_{ij}=1) \right\rangle_{q(u_{ij}|z_{ij}=1)} + \left\langle \log p(\boldsymbol{y}_{ij}|u_{ij}) \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1),q(u_{ij}|z_{ij}=1)} \\ & - \left\langle \log q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1) \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1),q(u_{ij}|z_{ij}=1)} \Big] \Big\}\end{aligned} \tag{77}$$

where,

$$\langle \log p(\boldsymbol{\pi}|\boldsymbol{a})\rangle_{q(\boldsymbol{\pi})} = \log\Gamma(\sum_{i=1}^{g} a_i) + \sum_{i=1}^{g}[(a_i - 1)\langle \log\pi_i\rangle_{q(\boldsymbol{\pi})} - \log\Gamma(a_i)] \tag{78}$$

$$\langle \log q(\boldsymbol{\pi})\rangle_{q(\boldsymbol{\pi})} = \log\Gamma(\sum_{i=1}^{g} \hat{a}_i) + \sum_{i=1}^{g}[(\hat{a}_i - 1)\langle \log\pi_i\rangle_{q(\boldsymbol{\pi})} - \log\Gamma(\hat{a}_i)] \tag{79}$$

$$\langle \log\pi_i\rangle_{q(\boldsymbol{\pi})} = \psi(\hat{a}_i) - \psi(\sum_{i=1}^{g} \hat{a}_i) \tag{80}$$

$$\langle \log p(\boldsymbol{\phi}_i|\gamma_0,\omega_0)\rangle_{q(\boldsymbol{\phi}_i)} = \sum_{k=1}^{Q}\Big[(\gamma_0 - 1)\langle \log\phi_{ik}\rangle_{q(\boldsymbol{\phi}_i)} + \gamma_0 \log\omega_0 - \omega_o\langle \phi_{ik}\rangle_{q(\boldsymbol{\phi}_i)} - \log\Gamma(\gamma_0)\Big] \tag{81}$$

$$\langle \log q(\boldsymbol{\phi}_i)\rangle_{q(\boldsymbol{\phi}_i)} = \sum_{k=1}^{Q}\Big[(\gamma_{ik} - 1)\langle \log\phi_{ik}\rangle_{q(\boldsymbol{\phi}_i)} + \gamma_{ik}\log\omega_{ik} - \omega_{ik}\langle \phi_{ik}\rangle_{q(\boldsymbol{\phi}_i)} - \log\Gamma(\gamma_{ik})\Big] \tag{82}$$

$$\langle \phi_{ik}\rangle = \frac{\gamma_{ik}}{\omega_{ik}} \tag{83}$$

$$\langle \log\phi_{ik}\rangle_{q(\boldsymbol{\phi}_i)} = \psi(\gamma_{ik}) - \log\omega_{ik} \tag{84}$$

$$\langle \log p(\boldsymbol{\Lambda}_i|\boldsymbol{\phi}_i)\rangle_{q(\boldsymbol{\Lambda}_i),q(\boldsymbol{\phi}_i)} = \sum_{l=1}^{P}\left\{-\frac{1}{2}\langle \boldsymbol{\lambda}_{il}^T \mathrm{diag}(\boldsymbol{\phi}_i)\boldsymbol{\lambda}_{il}\rangle_{q(\boldsymbol{\Lambda}_i),q(\boldsymbol{\phi}_i)} - \frac{Q}{2}\log 2\pi + \frac{1}{2}\langle \log|\mathrm{diag}(\boldsymbol{\phi}_i)|\rangle_{q(\boldsymbol{\phi}_i)}\right\} \tag{85}$$

$$\langle \log q(\boldsymbol{\Lambda}_i)\rangle_{q(\boldsymbol{\Lambda}_i)} = \sum_{l=1}^{P}\left\{-\frac{1}{2}\left\langle (\boldsymbol{\lambda}_{il} - \boldsymbol{m}_{il}^*)^T (\boldsymbol{S}_{il}^*)^{-1}(\boldsymbol{\lambda}_{il} - \boldsymbol{m}_{il}^*)\right\rangle_{q(\boldsymbol{\Lambda}_i)} - \frac{Q}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{S}_{il}^*|\right\} \tag{86}$$

$$\langle \log p(\boldsymbol{\mu}_i|\boldsymbol{m}_0,\boldsymbol{s}_0)\rangle_{q(\boldsymbol{\mu}_i)} = -\frac{1}{2}\left\langle (\boldsymbol{\mu}_i - \boldsymbol{m}_0)^T \mathrm{diag}(\boldsymbol{s}_0)(\boldsymbol{\mu}_i - \boldsymbol{m}_0)\right\rangle_{q(\boldsymbol{\mu}_i)} - \frac{P}{2}\log 2\pi + \frac{1}{2}\log|\mathrm{diag}(\boldsymbol{s}_0)| \tag{87}$$

$$\langle \log q(\boldsymbol{\mu}_i)\rangle_{q(\boldsymbol{\mu}_i)} = -\frac{1}{2}\left\langle (\boldsymbol{\mu}_i - \boldsymbol{m}_i)^T \boldsymbol{S}_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{m}_i)\right\rangle_{q(\boldsymbol{\mu}_i)} - \frac{P}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{S}_i| \tag{88}$$

$$\langle \log p(z_{ij} = 1|\boldsymbol{\pi})\rangle_{q(\boldsymbol{\pi})} = \langle \log\pi_i\rangle_{q(\boldsymbol{\pi})} \tag{89}$$

$$\langle \log p(u_{ij})\rangle_{q(u_{ij}|z_{ij}=1)} = \frac{\nu_i}{2}\log\frac{\nu_i}{2} - \log\Gamma\left(\frac{\nu_i}{2}\right) + \left(\frac{\nu_i}{2} - 1\right)\langle \log u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)} - \frac{\nu_i}{2}\langle u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)} \tag{90}$$

$$\langle \log q(u_{ij}|z_{ij} = 1)\rangle_{q(u_{ij}|z_{ij}=1)} = (\alpha_{ij} - 1)\langle \log u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)} + \alpha_{ij}\log\beta_{ij} - \beta_{ij}\langle u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)} - \log\Gamma(\alpha_{ij}) \tag{91}$$

$$\langle u_{ij}\rangle = \frac{\alpha_{ij}}{\beta_{ij}} \tag{92}$$

$$\langle \log u_{ij} \rangle_{q(u_{ij}|z_{ij}=1)} = \psi(\alpha_{ij}) - \log\beta_{ij} \tag{93}$$

$$\begin{aligned}\left\langle \log p(\boldsymbol{y}_{ij}|u_{ij}) \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1),q(u_{ij}|z_{ij}=1)} =& -\frac{\langle u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)}}{2}\left\langle \boldsymbol{y}_{ij}^T\boldsymbol{y}_{ij}\right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1),q(u_{ij}|z_{ij}=1)} - \frac{Q}{2}\log 2\pi \\ &+ \frac{Q}{2}\langle \log u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)}\end{aligned} \tag{94}$$

$$\left\langle \log q(\boldsymbol{y}_{ij}|u_{ij}, z_{ij}=1) \right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1),q(u_{ij}|z_{ij}=1)} = -\frac{Q}{2} - \frac{Q}{2}\log 2\pi + \frac{Q}{2}\langle \log u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)} - \frac{1}{2}\log|\boldsymbol{\Sigma}_i^y| \tag{95}$$

$$\begin{aligned}\left\langle \log p(\boldsymbol{x}_j|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i,\nu_i,\boldsymbol{y}_{ij},u_{ij},z_{ij}=1)\right\rangle_{q(\boldsymbol{\theta})} =& -\frac{P}{2}\log 2\pi + \frac{P}{2}\langle \log u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)} - \frac{1}{2}\log|\boldsymbol{\Psi}_i| \\ &- \frac{\langle u_{ij}\rangle_{q(u_{ij}|z_{ij}=1)}}{2}\mathrm{tr}\left[\boldsymbol{\Psi}_i^{-1}\left\langle \left(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij}\right)\left(\boldsymbol{x}_j-\boldsymbol{\mu}_i-\boldsymbol{\Lambda}_i\boldsymbol{y}_{ij}\right)^T\right\rangle_{q(\boldsymbol{\mu}_i),q(\boldsymbol{\Lambda}_i),q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)}\right]\end{aligned} \tag{96}$$

and $\psi()$ is the digamma function.

Finally, concerning the expression of the predictive density estimation, $\log p(X'|X)$, from (67) we yield

$$\log p(X'|X) \geq \mathrm{pred}(X') \tag{97}$$

where $\mathrm{pred}(X')$ is the predictive density estimation (lower bound) and is given by

$$\begin{aligned}\mathrm{pred}(X') =& \sum_{j=1}^{n'}\sum_{i=1}^{g} q(z_{ij}=1)\left\{\int \mathrm{d}\boldsymbol{\pi} q(\boldsymbol{\pi})\log\frac{p(z_{ij}=1|\boldsymbol{\pi})}{q(z_{ij}=1)} + \int \mathrm{d}u_{ij} q(u_{ij}|z_{ij}=1)\left[\log\frac{p(u_{ij})}{q(u_{ij}|z_{ij}=1)} + \right.\right. \\ & \int \mathrm{d}\boldsymbol{y}_{ij} q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)\left(\int \mathrm{d}\boldsymbol{\Lambda}_i q(\boldsymbol{\Lambda}_i)\int \mathrm{d}\boldsymbol{\mu}_i q(\boldsymbol{\mu}_i)\log p(\boldsymbol{x}_j^{'}|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i,\nu_i,\boldsymbol{y}_{ij},u_{ij},z_{ij}=1)\right. \\ & \left.\left.\left. + \log\frac{p(\boldsymbol{y}_{ij}|u_{ij})}{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)}\right)\right]\right\} \Rightarrow\end{aligned} \tag{98}$$

$$\begin{aligned}\mathrm{pred}(X') =& \sum_{j=1}^{n'}\sum_{i=1}^{g} q(z_{ij}=1)\left[\langle \log p(z_{ij}=1|\boldsymbol{\pi})\rangle_{q(\boldsymbol{\pi})} - \log q(z_{ij}=1) - \left\langle \log q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)\right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)),q(u_{ij}|z_{ij}=1)}\right. \\ & - \langle \log q(u_{ij}|z_{ij}=1)\rangle_{q(u_{ij}|z_{ij}=1)} + \left\langle \log p(\boldsymbol{y}_{ij}|u_{ij})\right\rangle_{q(\boldsymbol{y}_{ij}|u_{ij},z_{ij}=1)),q(u_{ij}|z_{ij}=1)} + \langle \log p(u_{ij})\rangle_{q(u_{ij}|z_{ij}=1)} \\ & \left. + \left\langle \log p(\boldsymbol{x}_j^{'}|\boldsymbol{\Lambda}_i,\boldsymbol{\mu}_i,\boldsymbol{\Psi}_i,\nu_i,\boldsymbol{y}_{ij},u_{ij},z_{ij}=1)\right\rangle_{q(\boldsymbol{\theta})}\right]\end{aligned} \tag{99}$$

where the expressions of the used auxiliary quantities (mean values) are already derived above.

REFERENCES

[1] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[2] Z. Ghahramani and G. Hinton, "The EM algorithm for mixtures of factor analyzers," Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4, Tech. Rep. CRGTR- 96-1, 1997.

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

[4] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and J. Ye, "Time series classification using Gaussian mixture models of reconstructed phase spaces," *IEEE Trans. Knowledge and Data Engineering*, vol. 16, pp. 779–783, 2004.

[5] G. McLachlan, D. Peel, and R. Bean, "Modelling high-dimensional data by mixtures of factor analyzers," *Comput. Statist. Data Anal.*, vol. 41, pp. 379–388, 2003.

[6] L. Saul and R. M.G., "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 2, pp. 115–125, 2000.

[7] H. Yamamoto, Y. Nankaku, C. Miyajima, K. Tokuda, and T. Kitamura, "Parameter sharing and minimum classification error training of mixtures of factor analyzers for speaker identification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004*, vol. 1, 2004, pp. 29–32.

[8] B. Narayanaswamy and R. Gangadharaiah, "Extracting additional information from Gaussian mixture model probabilities for improved text-independent speaker identification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2005, pp. 621–624.

[9] M. Yang, N. Abuja, and D. Kriegman, "Face detection using mixtures of linear subspaces," in *Proc. 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 553–556.

[10] A. Kosinski, "A procedure for the detection of multivariate outliers," *Computational Statistics and Data Analysis*, vol. 29, pp. 145–161, 1999.

[11] S. Shoham, "Robust clustering by deterministic agglomeration EM of mixtures of multivariate t distributions," *Pattern Recognition*, vol. 35, no. 55, pp. 1127–1142, 2002.

[12] J. Zhao and Q. Jiang, "Probabilistic PCA for t distributions," *Neurocomputing*, vol. 69, no. 16-18, pp. 2217–2226, Oct. 2006.

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

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

[15] G. McLachlan, R. Bean, and L. B.-T. Jones, "Extension of the mixture of factor analyzers model to incorporate the multivariate t-distribution," *Comp. Stat. Data Analysis*, vol. 51, no. 11, pp. 5327–5338, 2007.

[16] 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.

[17] C. Archambeau, J. 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.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

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

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

[21] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. Jordan, Ed. Dordrecht: Kluwer, 1998, pp. 105–162.

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

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

[24] 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.

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

[26] 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.

[27] E. Fokoue, "Stochastic determination of the intrinsic structure in Bayesian factor analysis," Statistical and Applied Mathematical Sciences Institute, Tech. Rep. TR-2004-17, 2004.

[28] J. Winn and C. Bishop, "Variational message passing," *J. Machine Learning Research*, vol. 6, pp. 661–694, 2005.

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

[30] K.-R. Mueller, C. W. Anderson, and G. E. Birch, "Linear and nonlinear methods for brain-computer interfaces," *IEEE Trans. Neural Systems Rehabilitation Engineering*, vol. 11, no. 2, pp. 165–169, 2003.

[31] Z. Keirn and J. Aunon, "A new mode of communication between man and his surroundings," *IEEE Trans. Biomedical Engineering*, vol. 37, no. 12, pp. 1209–1214, 1990.

[32] S. M. Kay, *Modern Spectral Estimation: Theory and Application*. Prentice-Hall, 1988.

[33] C. W. Anderson, E. A. Stolz, and S. Shamsunder, "Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks," *IEEE Trans. Biomedical Engineering*, vol. 45, no. 3, pp. 277–286, 1998.

[34] T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," in *Advances in Artificial Intelligence*, 2006, pp. 502–507.