Stochastic Variational Inference for Mixed Factor Analysers with Heterogeneous Data

Angela Moreno Martínez
Department of Signal Theory and Communications
Universidad Carlos III de Madrid, Spain

Abstract

The use of heterogeneous data is ubiquitous in the real-world applications of machine learning. Heterogeneous datasets typically contain mixed types of observations, i.e, continuous and discrete variables having different marginal distributions. The presence of missing values in datasets is another issue to overcome. Latent variable models allow us to find the hidden patterns in such scenario. A particular case of latent variable models, mixture models, are often used to capture the hidden classes that govern the data. However, these models suffer from a lack of expressiveness. In this work, we propose a mixture of factor analysers appropriate for learning from heterogeneous data. The proposed model handles marginal distributions for categorical, binary and real-valued data. We present an approximate posterior inference algorithm for fitting the model latent variables and parameters. Experimental results on an *eHealth* dataset are presented for demonstrating the hypothesis of the proposed method.

1 Introduction

In machine learning (ML) applications to real-world scenarios, data features are usually high-dimensional and heterogeneous, i.e. from different statistical type. In particular, many healthcare datasets contain heterogeneous features. For instance, Electronic Health Records (EHRs) [9] generally contain personal information about patients such as the sex, demographics or medications, which are discrete data features, or information like vital signs or laboratory data that are continuous data features. Another frequent challenge presented in real-world datasets is the presence of missing values, which may result in an unreliable analysis of the observations. Having both issues results in the need of a generative probabilistic model capable of impute or estimate missing data in an heterogeneous scenario.

Latent variable models [5] assume that there exist unobserved patterns where complex observed data relies on. There is an extensive literature on latent variable models for factor discovery, dimensionality reduction or even missing imputation [3]. We can differentiate between this type of models having discrete or continuous latent variables.

Mixture models are among the simplest forms of latent variable models broadly used for clustering problems. The main assumption is that the data is clustered and each observation is drawn from a posterior distribution [1]. Thus, the data is assumed to be represented by a discrete latent state, $\mathbf{z}_i \in [1, 2, .., K]$. For example, the circadian routine of an individual can be clustered into groups represented by a discrete variable. However, the use of mixture models in an heterogeneous scenario typically turns into a biased problem, due to the lack of expressiveness of the discrete latent variable..

Another sort of latent variable models are factor analysis (FA) models, which are important as a component in more complicated models [5]. Typically, factor analysers differ from mixture models in using real-valued latent variables, $\mathbf{z}_i \in \mathbb{R}^L$. These latent variables are not longer mutually exclusive as in mixture models. Then, factor analysis models have more representational power than using discrete latent variables [10].

To overcome the limitations of mixture models in our setting, we propose the use of a set of latent variable models as components of a more complex model for more complex data. Particularly, we present a mixture of factor analyzers (MFA) that assumes that the data belongs to a low curved dimensional manifold, which can be approximated by a piecewise linear manifold [10].

There are other approaches to tackle this issue, such as the transformation of discrete data into Gaussian-distributed *pseudo-observarions* [9, 3] or the use of deep generative models where the learning of continuous and discrete data is detached in different blocks

[2, 1].
Regarding the inference problem, we propose an scalable method based on stochastic variational inference.

2 Problem formulation

In mixture models, the discrete latent variable, $\mathbf{z}$, defines the assignments of data points to specific components of the mixture. In such a way, point-estimates of the posterior distribution, $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$, can be obtained to assign each data point to a particular class or component of the mixture, indexed by K. The relationship between model parameters, latent and observable variables is depicted in Fig. 1, where we define the categorical, real-valued and binary observations as x^c,x^r and x^b, respectively.

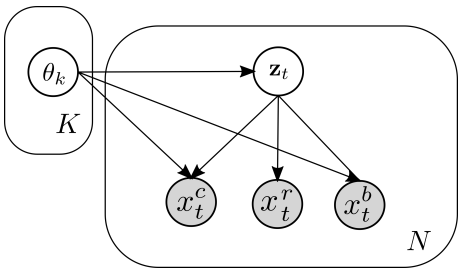

Figure 1: Graphical model for an heterogeneous mixture model with real, binary and categorical data.

This model is characterized by a set of parameters $\boldsymbol{\theta}$, which contains both the *mixture weights* and the heterogeneous likelihood parameters. In case of having heterogeneous mixture models, the likelihood distribution is given by

$$\begin{aligned} p\left(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}\right) &= p\left(\mathbf{x}_i^{\text{bin}}, \mathbf{x}_i^{\text{cat}}, \mathbf{x}_i^{\text{real}}, \mathbf{z}_i \mid \theta_k\right) \\ &= \prod_{k=1}^{K} \pi_k^{\mathbb{I}\{z_i=k\}} \prod_{j=1}^{Dr} p\left(\mathbf{x}_i^{\text{real}} \mid \theta_{jk}\right)^{\mathbb{I}\{z_i=k\}} \\ &\times \prod_{j=1}^{Db} p\left(\mathbf{x}_i^{\text{bin}} \mid \theta_{jk}\right)^{\mathbb{I}\{z_i=k\}} \\ &\times \prod_{j=1}^{Dc} p\left(\mathbf{x}_i^{\text{cat}} \mid \theta_{jk}\right)^{\mathbb{I}\{z_i=k\}}, \end{aligned} \tag{1}$$

where π_k represents the *mixture parameter* or the proportions of each k-*th* component in the mixture. The marginal likelihood functions are

$$p(\mathbf{x}_i^{\text{real}} \mid \theta_{jk}) = \mathcal{N}(\mathbf{x}_i^{\text{real}} \mid \boldsymbol{\theta}_k), \tag{2}$$

$$p\left(\mathbf{x}_i^{\text{bin}} \mid \theta_{jk}\right) = \text{Bernoulli}(\mathbf{x}_i^{\text{bin}} \mid \boldsymbol{\theta}_k), \tag{3}$$

$$p\left(\mathbf{x}_i^{\text{cat}} \mid \theta_{jk}\right) = \text{Categorical}(\mathbf{x}_i^{\text{cat}} \mid \boldsymbol{\theta}_k), \tag{4}$$

being each $\boldsymbol{\theta}$ the parameters of each type of likelihood function. That is $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ for the Gaussian, $\{\boldsymbol{\rho}_k\}$ for the Bernoulli and $\{\boldsymbol{\alpha}_k\}$ for the categorical distribution.
The observational values for each data type correspond to different domains. This fact leads to a *likelihood scale problem* since the values of the Bernoulli probability distribution are bounded in the range between zero and one, while the values of the Gaussian probability distribution function are unbounded, for instance. Thus, some marginal likelihood contribution may dominate the overall learning while others may be poorly-modeled [1, 2].
In this work we propose a model with a latent indicator, $\mathbf{s} \in [1, \ldots, K]$, that specifies which low-dimensional subspace generates each heterogeneous observation. Note that we denote our discrete latent variable as $\mathbf{s}$, differently to the usual notation in mixture models. Thereby, the data is assumed to belong to a lower dimensional manifold while the scale unbalance between likelihood values and parameters of different marginal distributions is reduced.
This problem has also been addressed by *A.Nazabal et al.* in [1], where they propose an heterogeneous decoder for each type of data. In an auto-encoder setup, *C.Ma et al.* propose a hierarchical strategy where they first fit a different VAE independently for each data type [2]. Then, they fit a multi-dimensional VAE to capture the dependencies among marginal VAEs of the first stage.

3 Mixture of factor analysers for heterogeneous data analysis

In this section, we describe the model for mixed (discrete and continuous) data that we call mixture of factor analyzers (MFA) for heterogeneous data. Particularly, this model combines two latent variable models in order to overcome the *marginals scale* problem detected in the heterogeneous mixture models. The graphical model is given in Fig. 2.
We assume that the data points are indexed in the range $i \in \{1, ..., N\}$. The heterogeneous data dimensions are indexed in $j \in \{1, ..., D\}$, where $D = \{D_r, D_c, D_b\}$ for continuous, categorical and binary data, respectively. We also denote the complete data point by $\mathbf{x}_i = \left[x_{ji}^r, ..., x_{D_r i}^r, x_{ji}^c, ..., x_{D_c i}^c, x_{ji}^b, ..., x_{D_b i}^b\right]^{\mathrm{T}}$.
Each high-dimensional observed data array $\mathbf{x}_i \in \mathbb{R}^D$ lie in a particular low-dimensional continuous latent factor $\mathbf{z}_i \in \mathbb{R}^L$ given by the latent class, $s_i \in [1, .., K]$. Thus, our generative model relies on two different latent variables, s_i and $\mathbf{z}_i$. Note that the notation has changed since $\mathbf{z}$ is not the discrete latent variable anymore but the continuous latent variable.

$s_i \rightarrow$ Latent class variable
$\mathbf{z_i} \rightarrow$ Latent continuous factor

In the MFA model, each type of observations is associated with a hidden vector of *weights*, *i.e.*, $\mathbf{W}$, with one weight for each mixture component K, data dimension D and latent variable dimension L. Thus, each $\mathbf{W}_k$ specified in the graphical model of Fig. 2 is a $D \times L$ factor loading matrix. In standard factor analysis models we assume that the likelihood has the form $p(\mathbf{x_i} \mid \mathbf{z_i}) = \mathcal{N}(\mathbf{x_i} \mid \mathbf{W z_i} + \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Thus, the likelihood parameters are given by a linear combination of the continuous latent variable, $\mathbf{z}$, and its corresponding factor loading matrix, $\mathbf{W}$. Note that it can be extended to other types of likelihoods that belong to the exponential families such as Bernoulli, Poisson, Categorical or Gamma distribution, among others. The only requirement is that the natural parameters should have the form $\mathbf{W z_i} + \boldsymbol{\mu}$.

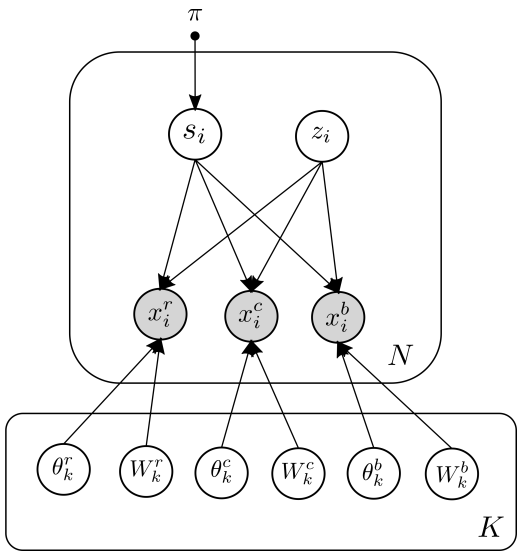

Figure 2: Graphical model of a mixture of factor analysers with heterogeneous observations

3.1 Generative process

Regarding the generative process of data, it begins with the *class-sampling* of the discrete latent variable, s_i, for each data point from a Categorical distribution with K-length parameter $\boldsymbol{\pi}$ generated by a Dirichlet distribution. It continues with the sampling of a continuous latent vector, $\mathbf{z}_i$, with length L, from a Gaussian distribution. These two steps are given by

$$p(\mathbf{s}_i, \mathbf{z}_i) = \text{Cat}(s_i \mid \boldsymbol{\pi}) \mathcal{N}(\mathbf{z}_i \mid \mathbf{0}, \mathbf{I}_L), \tag{5}$$

where $\mathbf{I}_L$ is the identity matrix.
Afterwards, the natural parameters of the marginals are obtained as a linear combination of the loading factor matrix $\mathbf{W}$ and the continuous latent vector $\mathbf{z}_i$, specified by the discrete class s_i. The model likelihood has the following form

$$\begin{aligned} p(\mathbf{x}_i \mid \mathbf{z}_i, s_i = k, \boldsymbol{\theta}) &= \prod_{m=1}^{M} \prod_{i=1}^{N} \prod_{k=1}^{K} p(\mathbf{x}_i^m \mid \mathbf{z}_i, \theta)^{\mathbb{I}\{s_i=k\}} \\ &= \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}_i^{\text{real}} \mid \mathbf{W}_k^r \mathbf{z}_i, \Sigma_k)^{\mathbb{I}\{s_i=k\}} \\ &\times \prod_{i=1}^{N} \prod_{k=1}^{K} \text{Ber}(\mathbf{x}_i^{\text{bin}} \mid \mathcal{S}(\mathbf{W}_k^b \mathbf{z}_i + \boldsymbol{\mu}_k^b))^{\mathbb{I}\{s_i=k\}} \\ &\times \prod_{i=1}^{N} \prod_{k=1}^{K} \text{Cat}(\mathbf{x}_i^{\text{cat}} \mid \mathcal{S}(\mathbf{W}_k^c \mathbf{z}_i + \boldsymbol{\mu}_k^c))^{\mathbb{I}\{s_i=k\}}, \end{aligned} \tag{6}$$

where m refers to the type of likelihood function and $\mathcal{S}(\eta)$ refers to the softmax function used in the discrete distributions to transform the natural parameters into their standard form. That is,

$$\mathcal{S}_i(\eta) = \frac{exp(\eta_i)}{\sum_k^K exp(\eta_{ik})}. \tag{7}$$

Note that the factor loading matrices for the k^{th} component of the mixture are $\mathbf{W}_k^m \in \mathbb{R}^{D_m \times L}$ for m different marginals. The set of model parameters is $\boldsymbol{\theta} = \left[\mathbf{W}_k^m, ..., \mathbf{W}_k^M, \Sigma_k, \boldsymbol{\mu}_k^b, \boldsymbol{\mu}_k^c, \boldsymbol{\pi}\right]$.

4 Approximate inference problem

Our general objectives can be outlined as unobserved structure discovering, new data predictions, dimensionality reduction and missing imputation. All these require computing the posterior distribution. Thus, the conditional distribution of the hidden variables given the observed data, $p(\mathbf{z}, s|\mathbf{x})$. In this section we present the posterior inference algorithm carried out in this work. Particularly, the posterior distribution has the form

$$p(\mathbf{z}, s \mid \mathbf{x}) = \frac{p(\mathbf{z}, s, \mathbf{x})}{\int p(\mathbf{z}, s, \mathbf{x}) dz ds}. \tag{8}$$

Computing the exact posterior is tractable in some basic models [5]. However, computing the marginal probability of the data and, in particular, the denominator in Eq. 8, may be intractable when the dimensionality of the latent variables is high. In our case, it means that we should marginalize out every possible combination of assignments of observations and mixture components. Furthermore, the complexity of integrating over all possible latent continuous factors grows exponentially.

This is the reason why we must approximate the posterior. The approximate method used in this work is variational inference. Particularly, we define a flexible and simple family of distributions over the hidden latent variables characterized by its variational parameters. Thus, we transform a complex inference problem into a high-dimensional optimization problem [4].
To fit the variational problem, Emtiyaz *et al.* in [3] use the Bohning bound, a variational bound over the *Log-Sum Exp* (LSE) function that allows them to obtain Gaussian-like *pseudo-observations* of discrete observations and to obtain closed form updates for expectation-maximization (EM) algorithm.

4.1 Stochastic variational inference

Regarding that our work aims to handle large datasets with high-dimensional data, we consider stochastic optimization [11]. Thus, we try to maximize the objective function by following noisy estimates of its gradient. Noisy estimates can be easily obtained by sub-sampling batches of observations and computing a scaled gradient on those batches [4]. Using batches of data, instead the complete dataset, helps to amortize the computational cost caused by the global parameters updates across more data points. Additionally, it may help the algorithm to find better local optima [4].
Our objective is to approximate the posterior distribution with a simpler variational function, $q(s, \mathbf{z}) \approx p(s, \mathbf{z} \mid \mathbf{x})$. This variational function can be factorized thanks to the mean-field approximation [10], where each hidden variable is independent.

$$q(s, \mathbf{z}) = q(s)q(\mathbf{z}). \tag{9}$$

In order to have the best possible approximation to the true posterior distribution, variational inference methods minimize the Kullback-Leibler (KL) divergence from the variational function to the true posterior. Similarly, it maximizes a lower bound on the log-marginal likelihood $\log p(x)$, widely known as the evidence lower bound (ELBO). The basic idea behind the ELBO is to restrict the form of the intractable posterior to a tractable class of distributions [15]. Before derivating the ELBO for our specific model, we first introduce the variational distributions considered.

$$q(\mathbf{z}) \sim \mathcal{N}(\mathbf{z} \mid \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \tag{10}$$

$$q(s) \sim \text{Categorical}(s \mid \boldsymbol{\pi}), \tag{11}$$

where the variational covariance matrix of $\mathbf{z}$ is assumed to be a diagonal matrix with variance σ^2. The set of variational parameters is $\boldsymbol{\phi} = \left[\boldsymbol{\mu}_i^z, \sigma_i^{2z}, \boldsymbol{\pi}_i^s\right]$.
To obtain the ELBO to the marginal likelihood, we begin with the logarithm of the marginal likelihood.

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x}, \mathbf{z}, s) d\mathbf{z} ds \\
&= \log \int p(\mathbf{x}, \mathbf{z}, s) \frac{q(\mathbf{z}, s)}{q(\mathbf{z}, s)} d\mathbf{z} ds \\
&= \log \left(\mathbb{E}_{q(\mathbf{z},s)} \left[\frac{p(\mathbf{x}, \mathbf{z}, s)}{q(\mathbf{z}, s)} \right] \right) \\
&\geq \mathbb{E}_{q(\mathbf{z},s)}[\log p(\mathbf{x}, \mathbf{z}, s)] \\
&\quad - \mathbb{E}_{q(\mathbf{z},s)}[\log q(\mathbf{z}, s)] = \mathcal{L}.
\end{aligned} \tag{12}$$

Note that we applied the Jensen's inequality in order to obtain the lower bound under the logarithm of the marginal likelihood [4].
Simultaneously, we can build the Kullback-Leibler divergence from the approximate distribution to the posterior distribution, which is a non-symmetric measure of disparity between both distributions.

$$\begin{aligned}
KL[q(\mathbf{z}, s) \,||\, p(\mathbf{z}, s \mid \mathbf{x})] &= \mathbb{E}_{q(\mathbf{z},s)}[\log p(\mathbf{z}, s)] \\
&\quad - \mathbb{E}_{q(\mathbf{z},s)}[\log p(\mathbf{z}, s \mid \mathbf{x})] \\
&= \mathbb{E}_{q(\mathbf{z},s)}[\log p(\mathbf{z}, s)] \\
&\quad - \mathbb{E}_{q(\mathbf{z},s)}[\log p(\mathbf{z}, s, \mathbf{x})] + \log p(\mathbf{x}) \\
&= -\mathcal{L} + \log p(\mathbf{x}).
\end{aligned} \tag{13}$$

Hence, maximizing the ELBO is equivalent to minimizing the KL divergence, which has a minimum value of zero when $q(\mathbf{z}, s) = p(\mathbf{z}, s \mid \mathbf{x})$.
We can already express the variational inference as an optimization problem where the objective function to be maximized is the evidence lower bound of our problem.

$$\begin{aligned}
\mathcal{L} &= \mathbb{E}_{q(\mathbf{z},s)} \left[\log \frac{p(\mathbf{x}, \mathbf{z}, s)}{q(\mathbf{z}, s)} \right] = \mathbb{E}_{q(\mathbf{z},s)} \left[\log \frac{p(\mathbf{x}, \mathbf{z}, s)}{q(z)q(s)} \right] \\
&= \mathbb{E}_{q(\mathbf{z})q(s)} \left[\log p(\mathbf{x} \mid s, \mathbf{z}) \right] + \mathbb{E}_{q(\mathbf{z})q(s)} \left[\log \frac{p(s)p(\mathbf{z})}{q(s)q(\mathbf{z})} \right],
\end{aligned} \tag{14}$$

where the second term can be decomposed as

$$\begin{aligned}
&\mathbb{E}_{q(z)q(s)} \left[\log \frac{p(s)p(z)}{q(s)q(z)} \right] = \\
&= \mathbb{E}_{q(z)} \left[\log \frac{p(z)}{q(z)} \right] + \mathbb{E}_{q(s)} \left[\log \frac{p(s)}{q(s)} \right] \\
&= -KL[q(z) \,||\, p(z)] - KL\left[q(s) \,||\, p(s)\right]
\end{aligned} \tag{15}$$

The first term of the ELBO in Eq. 14 can be derived

as

$$\begin{aligned}
&\mathbb{E}_{q(z,s)}\left[\log \frac{p(x,s,z)}{q(z,s)}\right] \\
&= \int_z \int_s q(s)q(z) \log p(x \mid z,s) ds dz \\
&= \int_z q(z) \left(\int_s q(s) \log p(x \mid z,s) ds \right) dz \\
&= \mathbb{E}_{q(z)}\left[\mathbb{E}_{q(s)}\left[\log p(x \mid z,s)\right]\right] \\
&= \mathbb{E}_{q(z)}\left[\int_s q(s) \log p(x \mid \hat{z},s) ds\right] \\
&= \mathbb{E}_{q(z)}\left[\sum_{k=1}^{K} q(s=k) \log p(x \mid \hat{z}, s=k)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{k=1}^{K} \pi_k \sum_{d=1}^{D} \log p^d(x^d \mid \hat{z}_m, s=k)\right], \quad (16)
\end{aligned}$$

where d refers to an specific data-type and the values of the expectation over z is estimated through Monte Carlo (MC) sampling, $\hat{z}_m \sim q(z)$. As the expectation cannot be derived analytically, we approximate it by a finite sum under the corresponding distribution [10]. Then, the complete expression for the variational objective function is

$$\begin{aligned}
ELBO = \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{k=1}^{K} \pi_k \sum_{d=1}^{D} \log p^d(x^d \mid \hat{z}_m, s=k)\right] \\
- KL(q(z) \,||\, p(z)) - KL(q(s) \,||\, p(s)), \quad (17)
\end{aligned}$$

where the first term evaluates how the data fit into the likelihood model while the second and third terms act as regularizers. Our prior latent variables distributions are

$$\begin{aligned}
p(\mathbf{z}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}), \\
p(s) &= \text{Categorical}\left(\frac{1}{K}\right), \quad (18)
\end{aligned}$$

with K being the number of classes of the mixture. The KL divergences between two Gaussian and Categorical distributions can be calculated as

$$KL[q(z) \,||\, p(z)] = \frac{1}{2}\left[\boldsymbol{\mu}_q^T \boldsymbol{\mu}_{\boldsymbol{q}} + \text{tr}\left\{\Sigma_q\right\} - k - \log|\boldsymbol{\Sigma}_q|\right], \quad (19)$$

$$KL[q(s) \,||\, p(s)] = \sum q_i \log q_i - \sum q_i \log p_i, \quad (20)$$

where $\Sigma_q = \sigma_q \mathbf{I}$ with σ being the standard deviation in Eq. 19.

4.2 Stochastic Variational EM algorithm

The objective function of our optimization problem is described in Eq. 17. Our goal is to optimize it with respect to the variational parameters $\boldsymbol{\phi}$ and the model parameters $\boldsymbol{\theta}$. Particularly, we maximize it via coordinate ascent. Thus, we iteratively optimize the variational parameters, fixing the others. Then, we optimize the model parameters, holding the variational parameters fixed. This is also known as *ensemble* or variational Bayes (VB) algorithm. In order to maximize the variational objective, we use an *stochastic gradient descent* (SGD) method with MC sampling.

Algorithm 1 Stochastic Variational EM algorithm

1: Initialization
2: Set of step-size η_t of optimizer
3: **repeat**
4: Sample batch of observations from the dataset $x_{batch} = [x^{real}_{batch}, x^{cat}_{batch}, x^{bin}_{batch}]$
5: Sample $\hat{z}_m \sim q(z)$
6: Compute its local variational parameter
7: Update the current estimate of global parameters
8: Calculate ELBO (Eq. 17)
9: **until** convergence

The gradient-based optimization algorithm introduced in our experiments is Adagrad [14], which adapts the learning rate to the parameters based on the past computed gradients

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} \odot \nabla_\theta \mathcal{L}\left(\theta_t\right), \quad (21)$$

where ϵ is a vector of small numbers to avoid the zero-division, η is the learning rate, which is initially set to 0.01, and G_t is the sum of the squared gradient estimates.

5 Experimental results

In this section we expose different databases analyzed in our work. Moreover, the experiments developed during this work will prove the performance in solving the heterogeneous unbalance with the use of stochastic variational inference.

5.1 Databases and data pre-processing

In this stage it is important to first evaluate the performance of our work with toy and real datasets.

5.1.1 Toy dataset

We build a toy dataset to prove the correct convergence and address the model order selection. To this effect, we simulate the generative process of our model. After choosing the order of the mixture or the number of categories (K), we construct the proportions for

each class of the mixture and the values of such discrete distribution given the generated proportions

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha)$$
$$s_i \sim \text{Categorical}(\boldsymbol{\pi})$$

Then, we choose the dimension of the continuous latent variable, L, to have a certain value and construct both the Gaussian-distributed samples, $\mathbf{z}$, and the $\mathbf{W}$ matrices with $K \times D \times L$ dimension, as well as the parameters for each heterogeneous likelihood. In this moment we are able to build the toy observations from each type of distribution (Gaussian, Bernoulli, etc.) with its corresponding generated parameters.

5.1.2 Mobile health dataset

Mobile devices have a pervasive presence in our daily life. Thus, our behavioral patterns may be projected into the digital data that we generate. Regarding the mental health, there are plenty of mental disorders whose symptoms directly affects to the daily functioning of the sufferer. For example, some studies claim that some relapses in bipolar disorders are related to a decrease of social and physical activity of the sufferer. On this basis, we can analyze the mobile data of patients with mental disorders in order to discover hidden behavioral patterns that help to improve and support the conventional treatments. We obtain mobile data directly from the *eHealth* start-up Evidence-based Behavior (*eB2*)[1]. In Table 1 the different data features are presented. These features widely shape the daily mobility, social activity, physical activity and sleep of an individual.

Feature	Domain	Distribution
Steps	$\mathbb{R}$	Gaussian
Distance	$\mathbb{R}$	Gaussian
Games app usage	$\mathbb{R}$	Gaussian
Social app usage	$\mathbb{R}$	Gaussian
Sleep category	[1, 2, 3]	Categorical
Time walking	$\mathbb{R}$	Gaussian
Sport activity	[0,1]	Bernoulli
Time still	$\mathbb{R}$	Gaussian

Table 1: Mobile data features and their assumed probabilistic distribution.

Sleep duration feature is obtained after grouping the quality depending on the time sleeping. Sport activity feature indicates if a patient practices sport in a certain day. The data is structured in a matrix of size $N \times D$, where N refers to the total number of days and D refers to the dimension of a daily observation.

[1]More information about eB2 can be found at `https://eb2.tech/`

Due to the data heterogeneity, the daily data dimensionality can be separated in *Dr*, *Dc* and *Db* for real, categorical and binary data, respectively. The range of numerical values of Gaussian-distributed features differs significantly. For instance, the travelled distance is measured in meters while the duration of mobile phone usage is measured in minutes. Then, we normalize the numerical values in order to avoid that some of those numerical values dominate the gradient evaluations of the objective function and to reduce the instability of such gradients [1]. Furthermore, categorical data is transformed into *one-hot* encoded vectors before feeding the model. A graphical representation of the data of a patient can be found in the Fig. 3, which has approximately 100 days of heterogeneous mobile data available to be analyzed.

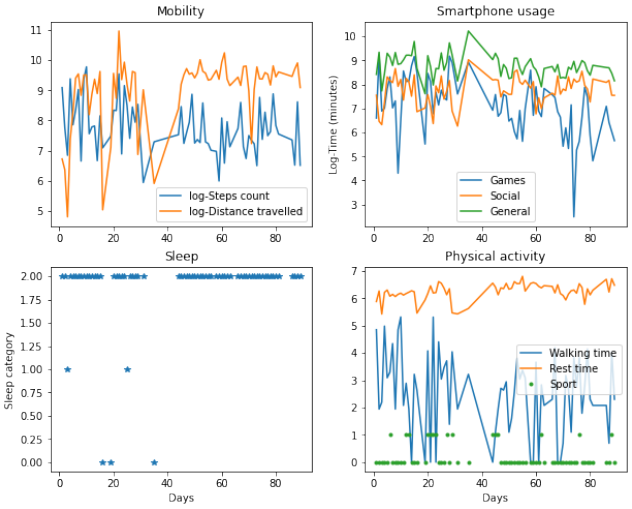

Figure 3: eB2 data features. The sub-figures are divided depending on the behavioral domain of each feature. Those are mobility, smartphone usage, sleep and physical activity.

5.2 Analysis of convergence

In Fig.4 we present the evolution of the objective function or ELBO at each epoch. Note that the optimization method aims to maximize the evidence lower-bound. We can observe how the categorical term of the objective function suffers from gradients stochasticity. At the end, every term converges and jointly collaborate to the learning process.

5.3 Data imputation

One of the main applications of generative models in real-world problems is the imputation of missing data. Machine learning algorithms not only have to face heterogeneous and high-dimensional data but also missing data. In Table 2, we evaluate the performance of our

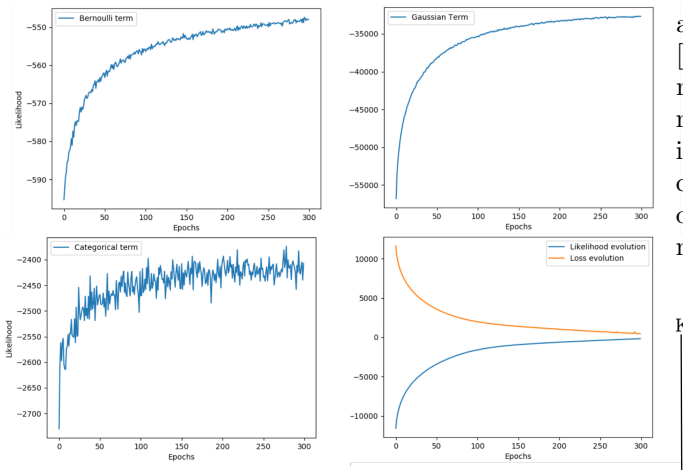

*Figure 4: Clockwise: Convergence of Bernoulli term in the objective function, convergence of Gaussian term in the objective function, total model convergence and convergence of Categorical term in the objective function. Experiments with **eB2 dataset**.*

model in data imputation for binary and real features. For each dataset we randomly remove a percentage of 20% of the data. The error metrics computed are the Accuracy Error $(1 - ACC)$ for binary data imputation and the RMSE for real data imputation. In order to select the best model easily, we compute the mean error of both metrics as:

$$\text{(Mean) error} = \frac{1}{2}\left[\text{Accuracy error} + \text{RMSE}\right] \quad (22)$$

These experiments mainly show the performance of data imputation regarding the dimension of the continuous latent variable, denoted as L. We observe that a smaller dimension of the continuous latent variable tends to improve the performance.

5.4 Model order selection

Selecting the dimensionality of the latent variables is a milestone in machine learning. The selection of such hyperparameters has to be tuned due to the unsupervised nature of our problem. Even so, we can select concrete values of those hyperparameters by comparing the likelihoods for each model as well as their data imputation error metrics. This comparison is shown in the Figure 5, where greener colors are linked to better values. Thus, the higher the likelihood and lower the imputation error, the better the model is. After some analysis in Figure 5 we select the model with higher K and lower L values. It means that the model better represents the data with a high dimension of the discrete latent variable and a low dimension of the continuous latent variable, respectively. First, the better approximation with a higher K-value is penalized by a method called Bayesian information criteria (BIC) [10], which adjusts the model complexity and performance. Second, a lower dimensional latent variable, $\mathbf{z}$, means that the better performance tends to approximate to a mixture model, which makes sense if we consider that a higher dimensionality would tend to deep models where the complexity of the model is significantly larger.

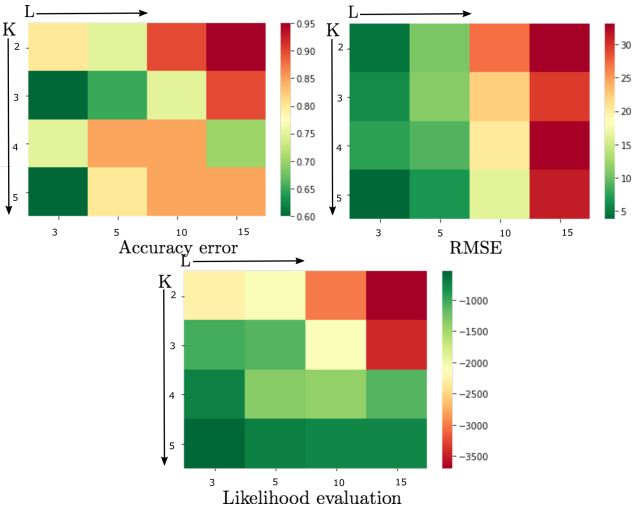

Figure 5: Accuracy error, RMSE and likelihood evolution with respect to L and K values.

5.5 Comparison with heterogeneous mixture model

This work is focused on solving the *likelihood scale* problem previously observed in heterogeneous mixture models. For this reason, we compare our work (hetMFA) with a mixture model (hetMM) [13], which is a particular case of our model when $L = 0$. Note that our hetMFA model outperforms the hetMM in the mean error when comparing the data imputation error metrics in Table 3.

	HetMFA	HetMM
ACC error	0.775±0.006	$\mathbf{0.4 \pm 0.016}$
RMSE	$\mathbf{5.584 \pm 0.135}$	51.701±33.91
Mean error	$\mathbf{3.179 \pm 0.070}$	26.050±16.963

Table 3: Data imputation error metrics for HetMFA and HetMM comparison using mobile health dataset.

However, the hetMM outperforms our model on the imputation of discrete data. This may be related to approximate inference issues that cause a reduction of accuracy in comparison with the tractable inference method used in hetMFA model.

	Mobile health dataset			Toy data		
	Accuracy error	RMSE	Mean error	Accuracy error	RMSE	Mean error
L=3	0.775±0.006	5.584±0.135	$\mathbf{3.179 \pm 0.070}$	0.390±0.009	8.786±3.969	$\mathbf{4.588 \pm 1.988}$
L=5	0.78±0.008	8.329±0.190	4.554±0.099	0.493±0.004	12.203±2.725	$\mathbf{6.348 \pm 1.364}$
L=10	0.712±0.002	18.91±4.621	9.811±2.131	0.542±0.009	13.125±5.589	6.834±2.799
L=15	0.812±0.010	29.309±4.611	15.06±2.310	0.593±0.005	17.269±19.5231	8.931±9.764

Table 2: Missing data imputation error for binary and real features in Toy and Mobile Health datasets. 20% of missing data.

6 Discussion and Future work

In this work we have proposed a stochastic variational inference method for fitting a mixture of factor analysis models with heterogeneous data. This algorithm is based in the variational EM method, an iterative algorithm that leads us to the maximization of the objective function. Thanks to the use of a continuous latent variable, we gain expressiveness and flexibility in the model while solving the *likelihood scale* problem existing in discrete latent variable models.
This work is a first approach to detect latent structures in heterogeneous observations. Further research options include deep generative models approximation, which may increase the computational efficiency and introduce the possibility of working with a larger data base. Another further possibility is to assume that the observed and latent variables are distributed in the exponential family, as M.Wedel *et al.* in [12] so that we could map each likelihood parameters into their natural parameters, which parametrize the *natural form* of the exponential family.

Acknowledgements

Our thanks to Pablo Moreno-Muñoz for sharing the source code of the hetMM to make our own experiments.

References

[1] A. Nazabal, P.M. Olmos, Z.Ghahramani and I.Valera. Handling incomplete heterogeneous data using VAEs. In *Pattern recognition*. Vol. 107, 2020.

[2] C.Ma, S.Tschiatschek, J.M. Hernandez-Lobato, R. Turner and C.Zhang. VAEM: a deep generative model for heterogeneous mixed type data. In *NeurIPS*,2020.

[3] M.Emtiyaz, G.Bouchard, B.M. Marlin, K.P. Murphy. Variational bounds for mixed-data factor analysis. In *NeurIPS*, 2010.

[4] M.D. Hoffman, D.M.Blei, C. Wang, J. Paisley. Stochastic Variational Inference. In *Journal of Machine Learning Research*, 2013.

[5] D.M. Blei. Build, compute, critique, repeat: Data analysis with latent variable models. In *Annual Review of Statistics and Its Application.* Vol. 1:203-232, 2014.

[6] J. Marino, Y.Yue and S.Mandt. Iterative inference models. In *ICML*, 2018.

[7] S.L. Clark, B. Muthen, K. Kaprio, B.M. D'Onofrio, R.Viken and R.J. Rose. Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. In *PMC*, 2014.

[8] K. Nasserinejad, J. van Rosmalen, W. de Kort and E.Lesaffre. Comparison of criteria for choosing the number of classes in bayesian finite mixture models, 2017.

[9] I. Valera, M.F. Pradier, M.Lomeli and Z. Ghahramani. General Latent Feature Models for Heterogeneous Datasets. *arXiv preprint arXiv: 1706.03779*, 2018.

[10] K.P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[11] H.Robbins and S.Monro. A stochastic approximation method. In *Annals of Mathematical Statistics*. Vol. 22, No. 3, 400-407, 1951.

[12] M. Wedel, W.A. Kamamura. Factor analysis with (mixed) observed and latent variables in the exponential family. In *Psychometrika*. Vol.66, No. 4, 515-530, 2001.

[13] P. Moreno-Muñoz, D. Ramirez, A. Artes-Rodriguez. Change-point detection on hierarchical circadian models. arXiv preprint arXiv:1809.04197, 2018.

[14] I. Goodfellow, Y. Bengio and A. Courville. Deep learning. MIT press, 2016.

[15] M. Emtiyaz Khan. *Variational learning for latent gaussian models of discrete data*. PhD dissertation, University of British Columbia, 2012.